DOTTORATO di RICERCA in INGEGNERIA dell'INFORMAZIONE

SCUOLA di DOTTORATO in SCIENZE dell'INGEGNERIA

Università Politecnica delle Marche

Luca Virgili

got his Bachelor Degree in Computer Science and Auto+ mation Engineering in 2016 and his Master Degree in Computer Science and Automation Engineering (summa cum laude) in 2018 at the Polytechnic University of Marche. He attended the PhD course in Information Engineering at the Department of Information Enginee+ ring of the same University from 2018 to 2021.

The studies and research activities of Luca Virgili were carried out in cooperation with high experienced researchers belonging to the Polytechnic University of Marche, University of Calabria, University of Pavia and Daimler AG.

His main research field is the analysis of complex networks and social networks. His research activities involve the application of Complex Network Analysis and Social Network Analysis techniques to study heterogeneous scenarios. Modeling a problem using a graph-based representation allows people to obtain interesting results in many fields ,such as Social Networ+ ks, Internet of Things, Blockchain, Innovation Manage+ ment, Neurological Disorders and Data Lakes. Another research field of interest is the Artificial Intelligence one. In this case, Luca Virgili has contributed to the applica+ tion of Machine Learning models to fall detection, as well as to an approach for ontology alignment through Recursive Neural Networks. In the last months, he has also worked to a complex network-based representation of Convolutional Neural Networks (CNNs) in order to compress them and explain what happens under the hood of CNNs.

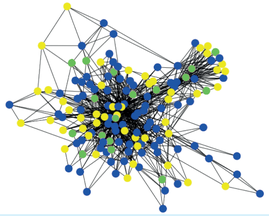
Luca Virgili is author of several scientific papers publi+ shed in International Journals (such as Information Sciences, International Journal of Information Manage+ ment, Future Generation Computer Systems, Expert Systems with Applications) and in the Proceedings of International Conferences.

He has been a Member of the Technical Program Com+ mittee and Session Chair for some International Confe+ rences. He was also a Member of the Organizing Com+ mittee of the International Winter School on Big Data in 2020.

He is a reviewer of scientific papers for international journals (such as IEEE Access, Multimedia Tools and Applications and Social Science Computer Review) and for international conferences and workshops.
He has participated to several research projects with research institutions and companies, and contributed to the realization of many prototype systems.

He has contributed to the didactic activities at the Polytechnic University of Marche. In particular, he has taught several classes in Mobile Programming, Big Data Laboratory and Data Science. He has been supervisor of several Bachelor Theses and some Master Theses. He has held several courses of Big Data Analytics, Machine Learning and Social Network Analysis to different companies.

Luca Virgili has participated to the Google Summer of Code program for 5 years where he had the opportunity to collaborate with DBpedia.



Nowadays, the amount and variety of scenarios that can benefit from techniques for extracting and managing knowledge from raw data have dramatically increased. As a result, the search for models capable of ensuring the representation and management of highly heterogeneous data is a hot topic in the data science literature. In this thesis, we aim to propose a solution to address this issue. In particular, we believe that graphs, and more specifically complex networks, as well as the concepts and approaches associated with them, can represent a solution to the problem mentioned above. In fact, we believe that graphs can be a unique and unifying model to uniformly represent and handle extremely heterogeneous data. Based on this premise, we show how the same concept and/or approach has the potential to address different open issues in different contexts.

Luca VIRGILI

Graphs behind data: a network-based approach to model different scenarios

DOCTORAL SCHOOL IN ENGINEERING SCIENCE
POLYTECHNIC UNIVERSITY OF MARCHE

DRINF

DOTTORATO di RICERCA in INGEGNERIA dell'INFORMAZIONE

SCIENTIFIC BOARD MEMBERS:

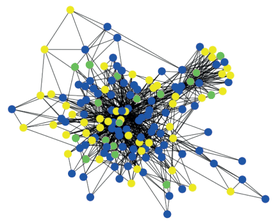
Franco CHIARALUCE (coordinator)
Marco BALDI
Andrea BONCI
Laura BURATTINI
Stefania CECCHI
Massimo CONTI
Claudia DIAMANTINI
Daniele EUGENIO
Sandro FIORETTI
Oriano FRANCESCANGELI
Emanuele FRONTONI
Ennio GAMBI
Donato IACOBUCCI
Gianluca IPPOLITI
Sauro LONGHI
Liana LUCCHETTI
Adriano MANCINI
Cristina MARCELLI
Valter MARIANI
Fabrizio MARINELLI
Davide MENCARELLI
Franco MOGLIE
Andrea MONTERIÙ
Antonio MORINI
Gianluca MORONCINI
Simone ORCIONI
Giuseppe ORLANDO
Valentina ORSINI
Luca PIERANTONI
Paola PIERLEONI
Ornella PISACANE
Domenico POTENA
Paola RUSSO
Luca SPALAZZI
Stefano SQUARTINI
Domenico URSINO
Primo ZINGARETTI

UNIVERSITÀ POLITECNICA DELLE MARCHE

Luca Virgili

GRAPHS BEHIND DATA: A NETWORK-BASED APPROACH TO MODEL DIFFERENT SCENARIOS

Supervisor: Prof. Domenico Ursino
S.S.D. ING-INF/05
XXXIV cycle



DOCTORAL SCHOOL IN ENGINEERING SCIENCE
POLYTECHNIC UNIVERSITY OF MARCHE

DEPARTMENT OF INFORMATION ENGINEERING
(DII)

PHD IN
INFORMATION ENGINEERING

S.S.D. ING-ING/05
XXXIV CYCLE

GRAPHS BEHIND DATA: A NETWORK-BASED APPROACH TO MODEL DIFFERENT SCENARIOS

CANDIDATE
Luca VIRGILI

SUPERVISOR
Prof. Domenico URSINO

COORDINATOR
Prof. Franco CHIARALUCE

ANCONA, FEBRUARY 2022

LUCA VIRGILI

GRAPHS BEHIND DATA: A NETWORK-BASED APPROACH TO MODEL DIFFERENT SCENARIOS

The Teaching Staff of the PhD course in

INFORMATION ENGINEERING

consists of:

Franco CHIARALUCE (coordinator)
Marco BALDI
Andrea BONCI
Laura BURATTINI
Stefania CECCHI
Massimo CONTI
Claudia DIAMANTINI
Daniele EUGENIO
Sandro FIORETTI
Oriano FRANCESCANGELI
Emanuele FRONTONI
Ennio GAMBI
Donato IACOBUCCI
Gianluca IPPOLITI
Sauro LONGHI
Liana LUCCHETTI
Adriano MANCINI
Cristina MARCELLI
Valter MARIANI
Fabrizio MARINELLI
Davide MENCARELLI
Franco MOGLIE
Andrea MONTERIÙ
Antonio MORINI
Gianluca MORONCINI
Simone ORCIONI
Giuseppe ORLANDO
Valentina ORSINI
Luca PIERANTONI
Paola PIERLEONI
Ornella PISACANE
Domenico POTENA
Paola RUSSO
Luca SPALAZZI
Stefano SQUARTINI
Domenico URSINO
Primo ZINGARETTI

Pensavamo che la vita funzionasse così,
che bastasse strappare lungo i bordi,
piano piano seguire la linea tratteggiata di ciò a cui eravamo destinati,
e tutto avrebbe preso la forma che doveva avere.
Negli anni, proviamo a convincerci che stiamo seguendo la linea tratteggiata
e, intanto, per paura che ci stiamo allontanando dalla guida
e che stiamo strappando a casaccio,
rimandiamo il momento in cui guardare il nostro foglio.
(Zerocalcare)

E' ora per me di guardare il mio foglio,
e vedere dove è finita la mia linea tratteggiata.

Foreword

It is absolutely not a case that our society is called "information society" and that knowledge is unanimously recognized as the new oil without which most of the activities that characterize everyday life would stop. Just to give an idea of the amount and variety of data produced every day, think that in one minute 500 hours of content are uploaded on YouTube, 700000 stories are shared on Instagram, about 70 million messages are sent via WhatsApp and Facebook Messenger, 28000 subscribers watch Netflix, 5000 downloads are made on TikTok and 200 million e-mails are sent worldwide. This is an amount and variety of data that are not even nearly comparable to what any previous generation has had to manage. And this trend is expected to grow even more impetuously with the advent of the Internet of Things. If previous generations had to deal with the lack of data, our own must address the opposite problem, i.e., an overabundance of the amount and variety of available data. This problem is equally difficult to manage, and the risk that data repositories will become data tombs is extremely high. To avoid this, scientific community has been performing studies and research for years, and these efforts have led to the emergence of new disciplines, such as Data Mining, Big Data Analytics, Machine Learning, Data Science, etc.

Luca Virgili's PhD thesis is set in this context and wants to provide a contribution in addressing these issues. It starts from the idea that graphs can be an extremely flexible, and, at the same time, very powerful model to represent very heterogeneous scenarios and data formats. At the same time, graph theory, as well as the complex network investigation and social network analysis that have their roots in it, represent a very mature body of knowledge, with rich and well tested results. As a consequence, many of the concepts, approaches and techniques defined by graph theory can be unique and unifying tools for successfully addressing several open issues related to possibly very different research areas. As a proof of this, in Luca Virgili's PhD thesis, six different areas are considered, namely Social Network Analysis, Internet of Things, Blockchain, Innovation Management, Neurological Disorders and

Data Lakes. For each of these areas, Luca Virgili's PhD thesis illustrates how graphs can be used for modeling the reference scenario. After that, it examines some open problems and describes how they can be successfully addressed by exploiting some well-known concepts (e.g., the ones of triad, clique, neighborhood, and centrality), as well as some approaches derived from graph theory.

Beyond the specifical technical merits, which the reader will be able to appreciate by proceeding with the reading of this thesis, Luca Virgili's approach has the characteristic of defining a uniform and unifying way of proceeding for handling very heterogeneous problems, which can be represented and managed through complex networks. For this reason, I believe that Luca Virgili's PhD thesis is an excellent piece of work. For each problem considered, it provides a complete description of the state of the art, clearly describes the proposed approach for its solution and presents an experimental campaign to evaluate its correctness and performance. The approach is methodologically and scientifically correct, as evidenced by the numerous papers already published by the author and his colleagues in several journals. I think that this thesis can be very useful for researchers who operate in the areas of Social Network Analysis, Internet of Things and Blockchain, as well as for practitioners working in the areas of Innovation Management, Neurological Disorder Analysis and Data Lakes.

In my role as advisor of Luca Virgili's PhD thesis, I had the privilege of being able to follow the entire development of the course of research that led the author to obtain the excellent results that this thesis describes. And here, writing this short preface, I have the pleasure to attest the quality, continuity, and passion that Luca Virgili has put, and continues to put, in his research activity. At the end of these three wonderful years, I certainly feel able to say that Luca Virgili has achieved all the goals we had set together beforehand, when this adventure has begun.

Prof. Domenico Ursino,
Università Politecnica delle Marche

Preface

This book is my PhD thesis and describes the research efforts I made at the Department of Information Engineering of the Polytechnic University of Marche from 2018 to 2021, under the supervision of Prof. Domenico Ursino.

During these three years, I had the opportunity to work with high experienced professors and researchers, such as Prof. Domenico Ursino himself, Prof. Antonino Nocera, Prof. Giorgio Terracina, Dr. Francesco Cauteruccio, Dr. Serena Nicolazzo and Dr. Alessia Amelio. I have learned so much from them and all the research findings I can present in this thesis are thanks to their ideas and advices.

My thesis starts from the observation that we have assisted to a huge growth of the available data in the last years. Every day we are flooded with more and more data. Think, for instance, of the weather forecasts, the routes recommended by navigators, news, data exchanged through social networks (consider that the average number of social media accounts is 8.4 per person in 2020), Internet of Things, and so forth. We are also harvesting for new data with the aim of optimizing any activity we make; think, for instance, of data provided by smart watches, fit bands, smart homes, and so on. Every day, we are overwhelmed by data, which makes it very difficult to extract only relevant information. For this reason, we need models and approaches able to handle huge amounts of data in order to extract only the most important information for a specific domain. For guaranteeing the efficiency and the effectiveness of the information and knowledge extraction from the data available, the necessity arises to represent it in a unique and unifying way. This unification and homogenization process is multi-dimensional because it regards the format, syntax and semantics of data involved.

This thesis aims at providing a contribution in this setting. Indeed, we propose a complex network-based model and some related approaches to uniformly represent and handle data in heterogeneous research scenarios. It is worth noting that, in each of them, we have not worked with tabular data, which focus on independent observations (i.e., rows of the table) containing information about the entities of a

domain. Instead, we have highlighted the importance of the connections between these entities and have represented them through complex networks. As a matter of fact, we can represent the domain entities as nodes and the entity connections as arcs. We can also attach labels, weights, and a set of features to these arcs for storing relevant information about these interactions (e.g. number of common posts in a Social Network, amount of money exchanged between two wallets in a blockchain, number of transactions in an Internet of Things scenario, etc.). Once a complex network representing a scenario is built, we can apply on it all the tools provided by Network Analysis, such as centrality measures, to derive the most important entities, cliques, to determine the presence of strongly connected components, and so forth. Following this reasoning, we are able to deal with any scenarios of interest through a unique model and with only minor adjustments.

In order to prove the validity of our conjecture, we have used complex networks and defined several related approaches to model and handle data in six different research areas, namely: *(i)* Social Networks, *(ii)* Internet of Things, *(iii)* Blockchain, *(iv)* Innovation Management, *(v)* Neurological Disorders, and *(vi)* Data Lakes.

As one might expect, these scenarios are very heterogeneous and each of them presents its peculiarities and issues to address. However, complex networks and the associated concepts and approaches have the intrinsic capability of uniformly representing and handling very different contexts. In this way, the same concept and/or approach has the potential to address different open issues in different contexts.

It is important for me to thank the people who have helped me during these three years. First of all, I want to express my gratitude to my supervisor Domenico Ursino, who has always believed in me. His advices have been useful and wise and I hope to continue my research path with him in the next years.

I would also like to thank my colleagues Enrico Corradini and Gianluca Bonifazi, with whom I shared offices, open spaces, labs, and any space with enough desks. Our discussions were fundamental to design experiments and propose new ideas to develop.

A special thanks goes to my family: my dad Domenico, my mum Giuseppina and my sister Sofia, who have supported me during this period. They were always ready to help me in any situation, both sentimentally and pragmatically. I hope that this thesis will make them proud of me, so that I can repay all the efforts they made for me.

I would like to thank my girlfriend Anna Lisa, with whom I shared my best moments. She has encouraged me many times during these three years. She is a fundamental part of my life and has played a key role to the writing of this thesis.

Last, but not least, I want to thank my friends: Roberto, Paolo, Maddalena, Beatrice and Camilla. Of course it is important to make progresses in our research, but it is worth resting and going out with our best friends as well. Unfortunately, I do not have the actual number of the beers and dinners shared with them, but I am sure that they have been a valuable part of my PhD period.

Finally, I would like to thank myself for never giving up. I have grown a lot during these years, I can say that I am a completely different man compared to who I was at the beginning of this journey. It was not easy for many reasons, but I always tried to fight and improve myself. I want to continue my research for leaving a little piece of me in the world, whether through papers or classes, whether words written somewhere or the memories (hopefully positive) of a student.

November 2021 Luca Virgili

Contents

Part I Social Networks

List of Figures

List of Tables

1

Introduction

This chapter is devoted to introducing the motivations and the general characteristics of the modeling approach proposed in this thesis. In particular, the plan of the chapter is as follows: in the first section, we illustrate the motivations which led to the definition of the proposed approach. The second section aims at presenting the complex networks as a unifying model, capable of representing and handling heterogeneous scenarios. The third section illustrates the contributions of complex network models in several heterogeneous contexts. Finally, in the fourth section, we provide an outline of the thesis organization.

1.1 Motivations

Data is everywhere and is constantly changing the way we live. Every day we are flooded with more and more data: think, for instance, of the weather forecasts, the routes recommended by our navigator, news, all the social networks (the average number of social media accounts is 8.4 per person in 2020), and so forth. We are also harvesting for new data with the aim of optimizing any activity we make: think, for instance of smart watches, fit bands, and smart homes.

In this scenario, we are overwhelmed by data and it is difficult to extract only relevant information. For this reason, we need solid models and approaches capable of managing huge amounts of data in such a way as to highlight the important peculiarities of the domains of interest. Furthermore, it would be useful to exploit these models and approaches in a unique and unifying way in different scenarios, because this would provide us with a general methodology and a set of tools to address a new domain never seen before.

This thesis aims at providing a contribution in this setting. Indeed, first we propose a complex network-based approach to uniformly extract knowledge and support decision making in heterogeneous research scenarios. Then, we apply the proposed model and several approaches based on it into four areas, namely: *(i)* Social Networks, *(ii)* Internet of Things, *(iii)* Blockchain, and *(iv)* Further Areas. This last

comprises Innovation Management, Neurological Disorders, and Extraction of Semantic Relationships among Concepts in Data Lakes.

In all these areas, we highlight the importance of the connections between the entities of the domains and investigate them. Then, we show that complex networks are a natural and, at the same time, powerful way to represent such connections. As a matter of fact, we can represent the domain entities as nodes, and the entity connections as arcs. We can also attach labels, weights, and a set of features to these arcs in order to deeply describe interactions (e.g. number of common posts in a Social Network, amount of money exchanged between two wallets, number of transactions in an Internet of Things scenario, etc.). This way of proceeding can be replicated in all the scenarios of interest with a small fine-tuning. Once the complex network representing a scenario is built, we can apply all the tools provided by Network Analysis, such as centrality measures, to derive the most important entities, cliques, to determine the presence of strongly connected entities, and so forth.

In this thesis, we start to apply this approach to the Social Network domain, specifically to two well-known social platforms, i.e., Reddit and Yelp. In both cases, and generally speaking in all social networks, the best way to model them is through the construction and the analysis of the corresponding complex networks. As for Reddit, we have obtained interesting results thanks to the co-posting network, in which we represented users and their activity of publishing posts. We verified that users tend to be connected to other ones with similar characteristics, which proved the existence of the homophily property in the network (which specializes to assortativity property in this case). As for Yelp, the usage of complex networks allowed us to highlight the friendship and review relationships between users, which paved the way to study the behavior of negative users and introduce a new kind of users, namely k-bridges, representing people interested to different business categories that can strongly influence users belonging to the same categories.

Another application of complex networks that is relatively new, but has already provided innovative results, regards the Internet of Things (i.e., IoT) domain. In this case, more and more research efforts are made for studying the behavior of smart objects in such a way as to derive their profiles and social interactions like if these were humans. Social Internet of Things (i.e., SIoT [70]) and Multiple Internet of Things (i.e., MIoT [82]) are only two of the latest architectures following this reasoning. In this representation model, smart objects are represented by network nodes, whereas network arcs could denote any type of relationship (e.g. distance, possibility to communicate, etc.). A model with these characteristics allows the definition of approaches for addressing most common issues of this domain, such as computing

trust and reputation, identifying anomalies and their impact, studying information flow.

The high generalizability intrinsic in complex networks model allows its applications also to blockchains. These have gained a lot of attention, especially thanks to cryptocurrencies, which rely on this technology. Indeed, all of us can remember the speculative bubble during the years 2017 and 2018, which enormously increased the prices of cryptocurrencies, and then exploded leading the same prices to decrease dramatically. In this context, an important factor that only few studies have considered is the social one. Indeed, a blockchain can be modeled through a complex network whose nodes denote blockchain addresses (each corresponding to a cryptocurrency wallet) and whose arcs represent transactions performed between two wallets. This representation can be analyzed through the tools provided by Social Network Analysis to identify the most important nodes in the network and how wallets tend to link to each others into strongly connected groups while exchanging a certain amount of cryptocurrency.

In this thesis, we will focus mainly on the three areas specified above. However, we will also show how this way of proceeding can be fruitfully adopted in several other areas, even if we will not describe the consequences of this application into detail. Specifically, the further areas we will consider are Innovation Management, Neurological Disorders, and Extraction of Semantic Relationships among Concepts in Data Lakes. In all these scenarios, complex networks play a key role in the knowledge representation and knowledge extraction issues. Specifically, as for Innovation Management, we modeled the peculiarities of patent citations, which are slightly different from the citations of scientific papers, and require suitable approaches to investigate them. A similar way of proceeding was adopted for the diagnoses of several neurological disorders. One way to investigate this kind of disorder is based on the usage of ElectroEncephaloGram (EEG, in short), which detects the brain activity through some electrodes attached to a human scalp. Starting from the EEG signals, we can build a suitable complex network and define metrics to evaluate the brain connectivity. Last, but not least, we applied our way of proceeding in a data lake scenario for managing the semantic relationships linking concepts stored in the corresponding data source. These last are presumably very heterogeneous from both the structural and the semantic viewpoints.

Summarizing, in this thesis, we want to show that complex networks and the associated concepts and approaches already defined in the past have the intrinsic capability of uniformly representing and handling very different scenarios. In this way, the same concept and/or approach has the potential to address different open issues in different contexts. The rest of this thesis is devoted to proving the correct-

ness of this intuition. In the next parts of this section, we present the six areas where we will apply our way of proceeding.

1.1.1 Social Networks

Online Social Networks (OSNs, in short) facilitate connections among people based on shared interests, values or memberships to specific groups. Nowadays, there are several OSNs with different aims and scope. For instance, LinkedIn is a professional network, in which people connect for work-related purposes, while Facebook is more devoted to people entertainment.

From a research perspective, OSNs are gold mines. Here, we can study all the nuances of human behavior, from their creation of posts and their comment activity, to their publication of business reviews, and so on and so forth. This type of knowledge has a lot of potential applications. Just think of the fact that OSNs have created a new figure, i.e., the influencer. On average, an influencer has many followers and she is able to make advertisements much more effectively than traditional ways. Another application regards the identification of the best targets for a marketing campaign, which could improve its effectiveness significantly. In the past literature, there are several approaches to advertise some products to only people that could be interested to them (such as a new running shoes for a runner, a laptop for a programmer, a book for a student, etc.). A further application regards the choice of new products or services to launch in the market thanks to the analysis of user needs or behaviors.

In order to extract this knowledge, we represent the OSNs domain as a complex network. Indeed, in this way, we have the capability of modeling this scenario using different kinds of node and arc, as well as of emphasizing user interactions in OSNs. This way of proceeding allows us to apply some of the techniques provided by Social Network Analysis for extracting knowledge from data. For instance, we can compute the centrality measures of users in order to identify the most influential and connected ones. Furthermore, we can investigate the structures formed by some users and identify if there are recurrent patterns and/or ways through which they tend to strongly connect with each other.

1.1.2 Internet of Things

In the last few years, we are experiencing a huge growth of the Internet of Things paradigm. Indeed, we can see the enormous increase of the number of sensors and devices, which are pervasive in our daily life. Roughly speaking, Internet of Things (i.e., IoT) consists of the interconnection of smart objects via the Internet, enabling them to send and receive data. At the time of the writing, the number of IoT devices

is more than 10 billion, and it is expected to reach more than 25.4 billion IoT devices in 2030[1]. Along with the increase of their number, devices are also developing smart and social skills. More and more researchers are beginning to study the behavior of things, to talk about their profiles and their social interaction [213], and to manage objects almost as if these were humans. As a result, several architectures implementing these ideas have been proposed, and are currently being proposed, in the literature. Social Internet of Things (i.e., SIoT [70]), Multiple IoT Environment (i.e., MIE [81]) and Multiple Internet of Things (i.e., MIoT [82]) are only three of the latest architectures with these characteristics. These architectures and analogous ones could be the foundation for dealing with the challenges posed by the IoT.

One of them regards the preservation of privacy and security of smart objects and their owners. According to anti-virus and computer security service provider Kaspersky[2], IoT cyberattacks more than doubled during the first half of 2021. The main issues regard: *(i)* the insecure communications that a device can establish with a potential attacker, and *(ii)* the storage of data containing the performed transactions.

Another IoT challenge is the identification of device anomalies. Indeed, IoT devices produce massive amounts of data continuously from numerous applications. Examining these collected data to detect suspicious events can reduce threats and avoid issues that can cause applications downtime. Some contexts possibly benefiting from this fact are healthcare, smart homes, self-driving cars, and so on. In all these cases, it is necessary to identify and address the anomalies in order to avoid severe consequences.

A last challenge that we mention in this area is network optimization. It comprises the tools and techniques that help to maintain, improve or maximize the communication performance across a network. For instance, we can improve the communication among devices thanks to the creation of virtual views of IoT, based on the content exchanged during transactions. In this way, we can study information flow and optimize communication paths. Another challenge could be the identification of the potential bottlenecks in network of smart objects thanks to suitable application of betweenness centrality.

[1] https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[2] https://www.iotworldtoday.com/2021/09/17/iot-cyberattacks-escalate-in-2021-according-to-kaspersky/

1.1.3 Blockchains

Since the creation of Bitcoin in 2008 from Satoshi Nakamoto, cryptocurrencies have been increasingly popular. This popularity has led to the speculative bubble that exploded during the end of 2017 and the beginning of 2018. Indeed, the Bitcoin price went up almost 800% during the last months of 2017 and then fell by 80% in the first few weeks of 2018. Of course, this price fall led to a huge gain for a few people and a big loss for the majority of investors. This period of huge growth and deep fall of price has been classified as a speculative bubble, similar to the tulipans' and stock market ones. A speculative bubble is an extreme event that has several consequences for both the economy and technology itself. These are the reasons why it is important to study these events in order to prevent (or at least face) them. Fortunately, market has been recovered since that event, and blockchains have proved to be a solid technology and not only a market manipulation.

In this overall scenario of blockchains, we think that social aspect has received a limited attention. For instance, in order to participate to a cryptocurrency network, a user has to create her own wallet, and then she can start to make transactions with other wallets in the same network. This process can be described by a complex network with the aim of studying the interactions between wallets. In this way, we can identify the most important ones in terms of centrality measures and network structures typical of network analysis.

In this thesis, we focus on the Ethereum blockchain and, thanks to the definition of a complex network, examine the behavior of its wallets during the speculative bubble period comprising the years 2017 and 2018, which we divided in pre-bubble, bubble post-bubble phases.

1.1.4 Innovation Management

Patents and other results of the collaboration among researchers have been largely investigated in the past especially in the scientometrics and bibliometrics research context. The impressive development of innovations in all the R&D fields and the attention we are paying to evaluate the performances of researchers, universities, and institutions are growing at a very rapid rate. One key aspect that has been intensively studied over the years is the interactions among researchers across firms and countries [612, 473, 130], which has led to interesting results. Indeed, research efforts have been made to understand whether international knowledge and investment flows from developed countries to less-developed ones have some positive effects. Others investigate the impact of international knowledge flows by focusing on R&D collaborations and inventions [202, 601, 122].

In this context, complex network analysis-based approaches are extremely promising due their capability of highlighting the interactions between the main actors of the domain. This approach is also motivated by a peculiarity of patents that we do not find elsewhere and that is hard to handle otherwise. Indeed, if a patent p_i cites a patent p_j, then p_i loses a part of its value. If we report this reasoning to the network analysis context, we have that, for a node, having incoming arcs is extremely positive; by contrast, having outgoing arcs is negative.

In this thesis, we propose a general methodology for the extraction of several knowledge patterns about innovation geography that can be applied on any country of interest. To this end, we introduce some novelties in the key metrics typical of Social Network Analysis in order to make them suitable to the patent domain.

1.1.5 Neurological Disorders

Thanks to the modern medicine and technology, life expectancy has grown in the last years. Estimates suggest that, in a pre-modern world, life expectancy was around 30 years in all the regions; since 1900 the global average life expectancy has more than doubled reaching 80 years. While the increase of life expectancy is an amazing result of human evolution, one of the drawbacks is the incidence of neurological disorders due to the fact that the population is aging in most countries. This has led to an increase of the efforts in designing approaches capable of determining and monitoring these disorders. In the meantime, the tools supporting neurologists in their activities are becoming much more complex and sophisticated (think, for instance, of the ElectroEncephaloGrams (EEGs, for short) with 256 electrodes, instead of the classical ones with 19 electrodes). This also means that we have to deal with huge amounts of data that experts have difficulty to analyze manually. For this reason, automatic tools helping them to analyze data are becoming mandatory. Among the many diagnostic tools available to neurologists, EEG is one of the least invasive, and it is adopted to support the analyses of neurological disorders.

In many neurological investigations, the key role is played by the connections between the brain areas. For instance, studies have found widespread underconnectivity, local overconnectivity, and, more in general, disrupted brain connectivity as a potential neural signature of autism [465]. An EEG and the data it provides can be easily modeled as a complex network, which can represent the interactions between brain areas in detail, and can provide an environment in which we can investigate the brain connectivity in order to help an expert in her diagnosis. In this complex network, nodes represent electrodes while arcs describe the relationship between two electrodes, derived, after several processing steps, by the voltage difference between them. Once we have the complex network, we can leverage some concepts of

network analysis (such as centrality measures, cliques, k-cores, etc.) to evaluate the most active brain areas and the corresponding connection levels.

In this thesis, we aim at proposing a complex network-based approach extracted from the EEG signals to help experts to investigate two neurological disorders, namely Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD).

1.1.6 Extraction of Semantic Relationships among Concepts

Metadata means "data about data". This expression summarizes their purpose, which consists of enriching data with additional information making it easier to find, use and manage. One example could be the information written on a letter envelope to help a letter getting correctly delivered. Metadata have a huge potential; indeed they have always played a key role in the cooperation of heterogeneous data sources. This role has become much more crucial with the advent of data lakes. A data lake is a centralized repository storing both structured and unstructured data. It allows us to store data as-is, and then run any task of analytics (e.g., creation of dashboards, real-time stream analytics, machine learning, etc.). In this scenario, metadata represent the only possibility to obtain an effective and efficient management of data source interoperability. Think, for instance, of a given application requiring to query only a subset of the data sources present in a data lake; it could process metadata to determine the portion of the data lake to examine.

Following this reasoning, we argue that, due to the heterogeneity of data lake sources, the necessity arises of flexible and powerful models and paradigms to support the metadata representation and management in a data lake. Our model starts from the considerations and the ideas proposed by data lake companies (in particular, it starts from the general metadata classification also used by Zaloni [519], a leader company in the data lake scenario), and then provide new contributions leveraging the potential of network-based and semantics-driven representation of metadata. As a result, it allows a large number of sophisticated tasks that most currently adopted metadata cannot guarantee. Specifically, it allows the definition of a structure for unstructured data and enables the extraction of thematic views from data sources. This task consists of the construction of views on one or more topics of interest to the user, obtained by processing data from different sources.

1.2 Complex Networks as a unique and unifying model

In this section, we provide an overview of our complex network-based model. As will be clear in the following, this model is able to uniformly handle data sources characterized by heterogeneous formats for extracting knowledge and supporting

decision making. Indeed, many phenomena can be represented thanks to a complex network. The main actors of these phenomena can be represented by means of the nodes of the network (think, for instance, of objects in an Internet of Things scenario, users in a social network, wallets in a blockchain, and so forth). Moreover, we can add information to these nodes thanks to suitable data structures (such as arrays, lists, dictionaries, etc.) in order to store relevant data that could not be mapped to a node or an edge, but is useful to detect specific patterns. Then, the connections between two actors are represented by the edges of a complex network. This way of modeling highlights the importance of the interactions between the entities of a domain and allows us to investigate them. As for nodes, we can add suitable information to edges for deeply representing an interaction between two entities. This information could be stored by means of arrays, lists and/or dictionaries (think, for instance, of the number of common posts in a Social Network, the amount of money exchanged between two wallets, the number of transactions in an Internet of Things scenario, and so on). These combined features allow us to manage any scenarios of interest with a small fine-tuning. After the construction of the complex network, the next step regards the application of the tools provided by Network Analysis (such as centrality measures, to derive the most important entities, cliques, to determine the presence of strongly connected entities, and so forth) to the constructed complex network.

In this section, we report the most important concepts and features of this model that we have employed for knowledge extraction. We provide a general overview of the concept of complex network and the Network Analysis tools; afterwards, we show how these concepts can be easily applied to different domains with only few adjustments.

1.2.1 Model definition

A complex network is a graph with non-trivial topological peculiarities that are not present in simple networks (e.g., grids or random graphs), but often occur in networks representing real systems. Some examples of complex networks are biological networks, technological networks, brain networks, climate networks, social networks, and so forth. One key feature of complex networks regards their "scale-free" property. It defines that the characteristics of the network are independent of the number of its nodes. A network is scale-free if the distribution of the number of arcs against nodes follows a power law, which means that we observe a small number of very highly connected nodes and a huge number of poorly connected ones. A direct consequence of this behavior is that the underlying structure remains the same when the network size grows.

Our complex network-based approach can deal with any scenario consisting of entities that interact with each other through one or more kinds of relationship. Formally speaking, it can be represented as:

$$\mathcal{N} = \langle V, A \rangle$$

Here, V is the set of nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to an entity, e.g. an object in an Internet of Things scenario, a user in a social network, a wallet in a blockchain, a patent, and so forth.

A is the set of arcs of $\mathcal{N}$. Each arc a_{ij} connects the nodes v_i and v_j and can be represented as:

$$a_{ij} = (v_i, v_j, w_{ij})$$

The arc between two nodes could represent many types of relationships. It could be a communication path between two objects in the Internet of Things, a friendship between two users in a social network, a transaction performed from a wallet to another one in a blockchain, a patent citation, etc. These arcs might be weighted. The weight w_{ij} is a measure of the connection strength between v_i and v_j. Considering the peculiarities of the different areas we are dealing with, our model is orthogonal to the different distance measurements that can be used. In the next chapters, we will employ different kinds of weight. In some scenarios, the weight is part of the input (e.g., the PDI in the EEG), while, in other cases, it is computed by pre-processing the input data (think, for instance, of the number of comments exchanged between two users in a social network or the number of transactions performed between two wallets).

In some cases, in order to perform the investigation of the issue of our interest, we must build projections of the networks involved, for instance by removing a type of node or arc. This allowed us to make our model more "user-friendly" and "expressive" and, at the same time, more capable of discriminating strong and weak connections between different network areas.

As an example, a network $\mathcal{N}_\pi$, being a projection of a network $\mathcal{N}$, can be obtained from this last one by removing the arcs with a "low" weight and by "coloring" the others based on their weight. As a matter of fact, if the arc weights represent closeness, the arcs with a "low" weight identify weak connections between the corresponding nodes and can be removed. The remaining arcs can be, instead, colored based on their weight. In particular, blue arcs denote strong connections, red arcs represent intermediate ones and, finally, green arcs indicate weak connections. We can formalize a network projection as follows:

$$\mathcal{N}_\pi = \langle V, A_\pi \rangle$$

Here, the nodes of $\mathcal{N}_\pi$ are the same as the ones of $\mathcal{N}$. To define A_π, we consider the distribution of the weights of the arcs of $\mathcal{N}$. Specifically, let max_A (resp., min_A) be the maximum (resp., minimum) weight of an arc of A. Starting from max_A and min_A, it is possible to define a parameter $step_A = \frac{max_A - min_A}{10}$, which represents the length of a "step" of the interval between min_A and max_A. We can define $d^k(A)$, $0 \leq k \leq 9$, as the number of the arcs of A with weights that belong to the interval between $min_A + k \cdot step_A$ and $min_A + (k+1) \cdot step_A$. All these intervals are closed on the left and open on the right, except for the last one that is closed both on the left and on the right. A_π consists of all the arcs of A belonging to $d^k(A)$, where $k \geq th_{min}$. Now, we can "color" the arcs composing A_π. Specifically, $A_\pi = A_\pi^b \bigcup A_\pi^r \bigcup A_\pi^g$. Here:

- $A_\pi^g = \{a_{ij} \in A | a_{ij} \in \bigcup_{th_{min} \leq k < th_{rg}} d^k(A)\}$
- $A_\pi^r = \{a_{ij} \in A | a_{ij} \in \bigcup_{th_{rg} \leq k < th_{br}} d^k(A)\}$
- $A_\pi^b = \{a_{ij} \in A | a_{ij} \in \bigcup_{th_{br} \leq k \leq th_{max}} d^k(A)\}$

As will be clear in the following, the projection technique described above, and therefore the corresponding network $\mathcal{N}_\pi$, represent a powerful tool for defining a uniform approach capable of handling knowledge in heterogeneous scenarios.

1.2.2 Network Characteristics

After the general definition of a complex network, we briefly introduce several measures describing it. One of them is *density*, which represents the proportion of the possible arcs in the network that are actually present. It is defined as:

$$density = \frac{2 \cdot |A|}{|N| \cdot (|N| - 1)}$$

The density value ranges between 0 and 1, with the lower limit corresponding to networks with no arcs and the upper limit representing networks with all possible arcs. The closer the value to 1, the denser the network and the more cohesive the nodes in it. Density can help to understand how much connected the network is, compared to how much connected it might be. When comparing two networks with the same number of nodes and the same type of relationships, it can provide us information about the connection differences between those networks.

Another important measure is *clustering coefficient* that describes the tendency of nodes in a network to cluster together. Clustering coefficient has both a local and a global definition. Before introducing them, we firstly have to define the concepts of neighborhood and node degree. The neighborhood of a node n in a network $\mathcal{N}$ is the sub-network of $\mathcal{N}$ induced by all the nodes adjacent to n, along with the corresponding arcs. The degree of a node n in $\mathcal{N}$ is the number of arcs connected to it.

The local clustering coefficient refers to the computation of the clustering coefficient of a node, while the global clustering coefficient is the mean of all the local clustering coefficients of the network nodes. The local clustering coefficient of a node $n \in \mathcal{N}$ is also the fraction of possible closed triads (i.e., node triangles) existing through that node. Formally speaking:

$$c_{local}(n) = \frac{2 \cdot T(n)}{d(n) \cdot (d(n) - 1)}$$

where $T(n)$ is the number of times the node n belongs to a triangle, and $d(n)$ is the degree of n. The global clustering coefficient of a network is defined as:

$$c_{global} = \frac{1}{|N|} \sum_{n \in N} c_{local}(n)$$

Both c_{local} and c_{global} belonging to the real interval between 0 and 1. The lower limit defines the case when there is no connection in the neighborhood, while the higher one denotes the scenario in which the neighborhood is fully connected. In a large complex network, it is difficult to interpret the global clustering coefficient, while the local one has a straightforward meaning. Indeed, if the neighborhood of a node n is dense and with a lot of mutual trust, n has a high clustering coefficient.

Density and clustering coefficient are very used in the next chapters, since they give us an initial overview of the features of the complex network into examination.

1.2.3 Network Structures

In this section, we report some of the most important and recurrent structures in a complex network, which are useful to identify the way its nodes are connected. Each structure has a computational cost for its processing and provides several insights about the network.

1.2.3.1 Ego Network

Ego networks are subnetworks centered on a certain node. This node is known as the ego and all the other nodes directly connected to it are called the alters. The computation of ego networks is performed by running a breath-first search limiting the depth of the search to a small value (usually 1). Ego networks are useful to derive interesting information from the most important nodes of a complex network. For instance, we can leverage this network structure to study the neighborhood of influencers in a social network or analyze the neighborhood of the wallets with the highest number of transactions in a blockchain.

We employed ego networks for evaluating the presence of backbones in the wallets of a blockchain in Chapter 9.

1.2.3.2 Clique, k-truss, k-core

A clique is defined as a maximal complete sub-network of a given network. It represents a group of nodes such that each of them is directly connected to the other ones. If a node is added to a clique, it is necessary to add arcs linking it to all the other nodes of the clique. Clique is an important structure to find in a network, because it describes a strong connection among a set of nodes. However, it requires very strict conditions to meet. For this reason, it is really hard to find clique in real life networks and its computation requires to solve a NP-hard problem. For these reasons, researchers often employ other network structures derived from the relaxation of the clique definition, such as k-truss and k-core.

A k-truss is a sub-network such that every arc is supported by at least k-2 other arcs that form triangles with that particular arc. In other words, every arc in a truss must be part of k-2 triangles made up of nodes that are part of the structure. By requiring each arc to include at least k-2 triangles, the k-truss computation achieves a great reduction of complexity, while still preserving the capability of identifying clusters of nodes. Indeed, the concept of k-truss is heavily used in the detection of communities in a complex network.

Finally, a k-core is a relaxation of both clique and k-truss concepts. In this case, a k-core is a sub-network in which every node has a degree greater than or equal to k. Conditions are less strict than the two other structures, which means that the computation complexity is much lower. Also in this case, k-core could be useful as an indicator for the presence of backbones among a subset of nodes.

One application of cliques within our work is the identification of the most connected brain areas. Furthermore, we have employed k-truss and k-core for extracting the nodes of strong connected backbones in the context of blockchains.

1.2.4 Centrality measures

Centrality measures aim at identifying the key nodes in a network. There are four basic centrality measures [402].

The first (and simplest) one is the *degree centrality*, which uses the number of arcs incidents to a node as an indicator of the "power" of that node. The advantage of this centrality is the fact that the results obtained through it are relatively easy to interpret and communicate.

The second centrality measure is the *closeness centrality* and is based on the idea that nodes having a short distance to other nodes, and consequently being able to disseminate information on the network effectively, have a power position in it. A node having a high closeness centrality requires from few to none intermediaries for

reaching other nodes, and, thus, is structurally relatively independent. The computation of this centrality includes the computation of the length of the shortest paths from a node to all the other ones in the network. The closeness centrality of a node n_i in a network $\mathcal{N}$ is:

$$CC(n_i) = \frac{|N| - 1}{\sum_{j=1, j \neq i}^{|N|} distance(n_i, n_j)}$$

where $distance(n_i, n_j)$ is the length of the shortest path between n_i and n_j.

The third centrality measure is *betweenness centrality* that considers the power of a node to control information flow in network. It is defined as the ratio between the number of all shortest paths between nodes in the network including the node into consideration and the number of all the shortest paths in the network. The betweenness centrality of a node n_i in a network $\mathcal{N}$ is:

$$BC(n_i) = \sum_{s,t \in N} \frac{\sigma(s, t|n_i)}{\sigma(s, t)}$$

where $\sigma(s, t)$ is the number of the shortest paths between s and t, and $\sigma(s, t|n_i)$ is the number of those paths passing through n_i.

The last basic centrality measure is the *eigenvector centrality*. It is based on the idea that a node is centrally involved in the network if it is directly connected to other nodes that are in turn well-connected. To compute the eigenvector centrality, we have to develop an iterative process, where, at each step, the centrality of a node is updated depending on the centrality of its neighbors. Given a complex network $\mathcal{N}$, let Adj be the adjacency matrix, i.e., $Adj[n_1, n_2] = 1$ if the node n_1 is linked to n_2, and $Adj[n_1, n_2] = 0$ otherwise. The eigenvector centrality of a node n_i can be defined as:

$$EC(n_i) = \frac{1}{\lambda_{max}} \sum_{j=1}^{|N|} Adj[n_j, n_i] \cdot v_j$$

where $v = (v_1, \cdots, v_n)^T$ refers to an eigenvector for the maximum eigenvalue λ_{max} of the adjacency matrix Adj.

A particular case of eigenvector centrality is PageRank. It was introduced by Google which has used it for indexing web pages. It can be applied only on directed networks. The PageRank of a node depends on the number of the links it receives, as well as on the centrality and the link propensity of the linkers. Formally speaking, the PageRank of a node n_i in a network $\mathcal{N}$ is defined as:

$$PR(n_i) = (1 - \gamma) + \gamma \cdot \sum_{n_j \in in(n_i)} \frac{PR(n_j)}{out(n_j)}$$

where γ is the damping factor, $in(n_i)$ returns the set of the neighbors pointing to n_i, and $out(n_j)$ returns the number of arcs outgoing from n_j. As the eigenvector centrality, also the PageRank is computed thanks to an iterative process, which eventually converges at a certain stable state.

Centrality measures are one of the key tools for this thesis, since they allow us to study the most important nodes in a complex network. For this reason, we have employed them in all the next chapters.

1.2.5 Assortativity

Assortativity is a property denoting that nodes with many connections tend to connect to each other [503]. Assortativity is strictly related to the concept of homophily. This property says that individuals in a social network tend to associate and link to similar others [468]. In Social Network Analysis, assortativity is a particular case of homophily. However, it can also be applied in other forms of complex networks, such as biological networks. Actually, it was shown that several existing complex networks are assortative, whereas other ones are disassortative. In this last, case high degree nodes tend to link to low degree ones. In the past literature, it was proved that social networks are often assortative, while technological and biological networks tend to be disassortative [503].

The concept of assortativity has implications for network resilience, since it was found that the connectivity of many networks can be destroyed by the removal of just a few of the highest degree nodes. This result may have many applications; one of the most interesting ones regards vaccination strategies. Indeed, in assortative networks the removal of high degree nodes is a relatively inefficient strategy for destroying network connectivity, because these nodes tend to be clustered together in the core group, so that removing them is redundant. On the other hand, in a disassortative network, attacks on the highest degree nodes are much more effective, because these nodes are broadly distributed over the network and presumably form links on many paths between other nodes.

These considerations are extremely relevant when the networks that we might want to break up are assortative, and therefore resilient against simple attacks involving only the highest degree nodes. Analogously, the same consideration plays a key role when the networks that we might want to protect are disassortative; in this case, we must consider that they are particularly vulnerable to attacks targeted to high degree nodes.

Assortativity is mostly computed based on to the degree centrality of nodes; however it is not out of place to employ the other centrality measures.

In this thesis, we have adopted the assortativity property to study both Safe For Work and Not Safe For Work posts in Reddit.

1.3 Problem Statements and Contributions

In this section, we describe the issues we want to address within each domain and present the contributions of our investigations. We report all the corresponding details in the next chapters of this thesis.

1.3.1 Social Networks

There are many social networks available online, such as Facebook, Twitter, Reddit, Yelp, etc. However, the literature has plenty of works on Facebook and Twitter, which also poses too many limits for accessing their data. Two social networks that have a great popularity but have received less attention by the research community are Reddit and Yelp, which are the focus of our approaches.

In both cases, we model these social media as a complex network, in which a node represents a user. The relationships between two nodes could represent any activity performed by the corresponding users in the social network (such as friendship, review of the same business, comment on the same post, etc.). Our model could have more than one type of relationships; in this case, we apply the correct projection to obtain the complex network suitable for a task.

As for Reddit, we downloaded the data for analyses from the website `https://pushshift.io`, which contains all the posts with the corresponding statistics and comments. This dataset allows us to define the user and subreddit stereotypes and to model co-posting activities. Co-posting denotes that two users publish a post in the same community. In this network, we have verified the assortativity property characterizing its users. Another aspect of Reddit worth to be analyzed involves NSFW (Not Safe For Work) posts. This term refers to user-submitted content not suitable to be viewed in public or in professional contexts. In this case, we investigate the possible differences between SFW (Safe For Work) and NSFW posts and between the users publishing them. Then, we create complex networks representing NSFW and SFW posts and users and exploit them to study the assortativity or disassortativity of these kinds of user.

As for Yelp, we downloaded data from its official site[3], and then, created the corresponding complex network. Here, a user is represented by a node, and the relationship between two users could be friendship or co-review. Depending on the phenomenon to investigate, we compute a projection of the complex network to get only the part that we need. Starting from it, we define a new category of users, i.e., k-bridge, who is a user publishing reviews to at least k types of different businesses. We study the influence of k-bridges in the network and propose some applications that

[3] `https://www.yelp.com/dataset`

could leverage this concept. Furthermore, thanks to the complex network approach, we are able to derive some user stereotypes in Yelp and define the characteristics of negative influencers. Finally, we investigate the influence of their negative reviews in their corresponding neighborhoods.
Some important contributions we have found in the social network domain are:

- the definition of subreddit and author stereotypes in Reddit;
- the evaluation of the assortativity of the co-posting activity in Reddit;
- the evaluation of the assortativity of users publishing NSFW and SFW posts in Reddit;
- the definition of a k-bridge user and its applications;
- the definition of negative influencer stereotypes and their impact in Yelp.

1.3.2 Internet of Things

The starting point of our investigation in this domain is the Multiple Internet of Things (MIoT) paradigm. Roughly speaking, a MIoT can be seen as a set of smart objects connected to each other by relationships of any kind and, at the same time, as a set of related IoTs, one for each kind of relationship. The MIoT model also introduces the concept of instance of a smart object in an IoT, which represents a virtual view of that object. The nodes of each IoT represent the instances of the smart objects participating to it. As a consequence, a smart object can have several instances, one for each IoT to which it participates. The existence of more instances for one smart object plays a key role in the MIoT paradigm, because it allows the definition of the cross relationships among the different IoTs of the MIoT. In such a scenario, IoTs are interconnected thanks to those nodes simultaneously belonging to two or more of them. We define cross nodes (c-nodes) these nodes and inner nodes (i-nodes) all the other ones. Then, a c-node connects at least two IoTs of the MIoT and plays a key role in favoring the cooperation among i-nodes belonging to different IoTs.

Basically, the classical MIoT paradigm models the Internet of Things as a complex network. In some cases, this representation fulfills the requirements necessary for our investigations. However, in other cases, we must introduce some novelties to the classical MIoT architecture.

In this thesis, we address the following issues: *(i)* analysis and optimization of the communication between smart objects; *(ii)* evaluation of the reliability of these interactions; *(iii)* safeguard of the privacy and security; and *(iv)* anomaly detection.

As for the analysis and optimization of the communication between smart objects, we provide three contributions. The first is the introduction of a new betweenness centrality measure that captures the peculiarities of the MIoT. Indeed, in this

case, the nodes in the complex network are not all equal, because c-nodes presumably play a more important role than i-nodes for supporting the activities in a MIoT. This important distinction between nodes is not considered in the classical betweenness centrality. The second contribution is the definition of a smart object profile, which allows us to introduce the concept of virtual IoT networks. They represent a view of the devices that exchange transactions with a particular content. Thanks to the focus on specific transaction contents and the analysis of the information diffusion in these networks, we are able to optimize the communication among devices. The third contribution is the definition of the neighborhood of a smart object in a MIoT and the possibility to define different neighborhood levels. In its turn, this allows us to define the concepts of scope and influence of a smart object in a MIoT.

As for the evaluation of reliability, we leverage the profile of a smart object previously mentioned and propose a new approach to compute trust and reputation in a MIoT. Thanks to the well-structured organization of the MIoT model, we are able to define trust and reputation at different levels. In fact, we can represent the trust between instances, between objects and between IoTs. Finally, we can compute the reputation of an instance or an object in an IoT as well as the reputation of an IoT in the MIoT.

As for the safeguard of privacy and security, we define a framework to mask the communications between devices thanks to the creation of heterogeneous groups (i.e., IoT networks), which hide the features and services provided by the smart objects belonging to them. In this case, we borrow some concepts from database anonymization (such as k-anonymity and t-closeness) for building the groups of smart objects.

Finally, we propose a new methodological framework for anomaly detection and classification in a MIoT. This framework models anomalies by means of three orthogonal taxonomies. Each combination of these taxonomies defines a specific kind of anomaly to study. Then, we perform two distinct investigations on anomalies: the former analyzes the impact of an anomaly in the MIoT, while the latter detects the source of an anomaly based on its overall effects on objects and connections.
Some important contributions we have found in the IoT domain are:

- the definition of an approach to determine virtual IoTs from the real ones, based on the content exchanged among smart objects, along with the definition of several applications possibly benefiting from them;
- the definition of a new centrality measure that captures the peculiarities of the MIoT paradigm;
- the definition and evaluation of the communication scope of the smart objects in a MIoT;

- the definition of an approach to compute trust and reputation of smart objects and communities of smart objects in a MIoT;
- the definition of a framework to preserve the privacy and security of the features and/or services provided by the smart objects in a MIoT;
- the definition of a taxonomy of the anomalies in a MIoT and of an approach to detect them.

1.3.3 Blockchains

The dataset we used for our analysis is based on Ethereum, which is a second generation blockchain and represents the technological framework behind the cryptocurrency Ether (ETH). We downloaded the transactions made on Ethereum from January 1^{st}, 2017 to December 31^{st}, 2018. This time interval corresponds to a speculative bubble period. Specifically, we divided this time interval in three phases, namely pre-bubble, bubble and post-bubble ones.

Starting from this dataset, we focus on four categories of users, namely: *(i)* Power Addresses, *(ii)* Survivors, *(iii)* Missings, and *(iv)* Entrants.

Then, we create the corresponding complex network. In this case, a node represents a wallet address, an arc between two nodes denotes a transaction between wallet addresses. Finally, the weight of an arc represents the number of transactions performed between the corresponding wallet addresses.

For each user category, we compute centrality measures and ego networks in order to characterize them. Furthermore, we check the possible existence of backbones linking the users of a certain category, which could reveal the possible existence of a form of homophily among them.

Finally, given a certain period (i.e., pre-bubble, bubble), we define an approach for predicting who will be the main actors in the next ones (i.e., bubble, post-bubble), based on some parameters.
Some important contributions we have found in the blockchain domain are:

- the definition of four categories of users and the detection of the main features characterizing them;
- the existence of backbones among users;
- the prediction of the main actors of the next period.

1.3.4 Innovation Management

Data regarding patents adopted in our analyses has been taken from PATSTAT-ICRIOS database [199]. PATSTAT (i.e., EPO worldwide PATent STATistical database) is a database storing raw data about patents. It was constructed by the European

Patent Office (EPO) in cooperation with the World Intellectual Property Organization (WIPO), the Organization for Economic Co-Operation and Development (OECD) and Eurostat. It stores data about all patents, from 1978 to the current year, coming from about 90 patent offices worldwide, comprising the most relevant ones, such as EPO and United States Patent and Trademark Office (USPTO).

Starting from the raw data of PATSTAT, we create our complex network model. Here, network nodes represent patents, while an arc from the node p_i to the node p_j denotes that p_i cites p_j. Furthermore, each node p_i has associated the set of the countries of the inventors of p_i. Clearly, this network is directed, since the arc direction is crucial for patent evaluation.

In order to model the citations impact, we propose two centrality measures, namely Naive Patent Degree and Refined Patent Degree. Both of them are based on the reasoning that having incoming arcs is extremely positive for a node, while having outgoing arcs is negative.

Our new definitions of centrality measures are then employed for the identification of the lifecycle and the scope of a patent, which indicate the width and strength of the influence of a patent on the other ones present in its neighborhood.
Some important contributions we have found in the patent domain are:

- the definition of an approach to evaluate the scope of a patent;
- the extraction of knowledge regarding the lifecycle of a patent;
- the definition of new metrics specifically conceived to evaluate the innovation level of each country, based on patent data.

1.3.5 Neurological Disorders

Our approach for investigating neurological disorders receives the ElectroEncephaloGrams (EEGs) of the patients to analyze and models them through a complex network, in which nodes represent electrodes and arcs denote connections between electrodes. Each arc has associated a weight representing a measure of the connection level between the brain areas covered by the corresponding electrodes.

In this application context, the EEGs to perform our investigation were provided by different Italian centers (i.e., University "Magna Graecia" of Catanzaro, Neurologic Institute "Carlo Besta" of Milano, Istituto Bonino-Pulejo and Neurologic Institute of the University of Catania). They regard a group of patients with neurological disorders, such as Mild Cognitive Impairment (MCI, for short) and Alzheimer's Disease (AD, for short).

In order to study the brain connectivity, we observe that, in complex networks, cliques play a key role to determine the network connection level, and, then, the

portion of networks most connected. Indeed, the higher the number and the dimension of available cliques in a network and the higher the corresponding connection level. Starting from this reasoning, we built a suitable data structure, called clique network, and an indicator of the connectivity level of the brain areas, called connection coefficient. The latter, when applied to the EEGs of patients with Cognitive Impairment allows patients with MCI to be distinguished from patients with AD. A further indicator, called conversion coefficient, which quantifies connection loss, has proven to be particularly useful in helping experts to understand if a patient with MCI is converting to AD.

In addition, our approach aims at verifying the possible existence of network motifs (i.e., specific sub-networks or network patterns), which are very frequent in one kind of patient and absent, or very rare, in the other. Also for this issue, we have obtained interesting results, since we have found some motifs characterizing patients with MCI from patients with AD.
Some important contributions we have found in the neurological disorders domain are:

- the definition of a coefficient supporting experts to distinguish patients with MCI from patients with AD;
- the definition of a coefficient for supporting experts to evaluate whether a patient is converting from MCI to AD;
- the definition of network motifs supporting experts to distinguish patients with MCI from patients with AD.

1.3.6 Extraction of Semantic Relationships among Concepts

In this field, we use complex networks to represent and handle the metadata of a data lake and to support an approach for extracting semantic relationships among the concepts represented in the data lake sources. This approach was developed having in mind two characteristics, namely: *(i)* the capability of handling unstructured sources; *(ii)* the lightweightness.

As for the former, our approach works with the metadata repository of a data lake and has a preliminary step for associating a certain structure to unstructured sources. For this purpose, it assumes that each unstructured source (e.g. a video, an audio, an image, a text) has associated a list of keywords describing its content; this list is just the foundation of the structuring process. After that, it computes the semantic similarities between the keywords of a source, and thus extract their corresponding relationships. At the end of these steps, we obtain a complex network providing a structured, yet flexible, representation of the metadata associated with data lake sources.

Observe that any approach operating in a big data scenario must take scalability into account [426, 423]. Now, a data lake is thought to handle numerous, large and heterogeneous data sources. As a consequence, an approach operating therein must be scalable. Our approach for the extraction of semantics relationships presented in this thesis presents this property.
Some important contributions we have found in the data lake domain are:

- the definition of an approach to create a structured representation of a natively unstructured data source;
- the definition of an approach to extract interschema properties and complex knowledge patterns from a data lake consisting of a huge number of heterogeneous data sources.

1.4 Outline of the thesis

This thesis aims to explore the possibility of using complex networks as a unique and unifying model to represent heterogeneous scenarios and to solve various open problems in each of them. It consists of five parts.

In Part I called *"Social Networks"*, we investigate the possibility of representing and handling new knowledge from two of the most important social networks, namely Reddit and Yelp. In particular, in Chapter 2, we define the subreddit and user stereotypes of Reddit, and evaluate the assortativity of co-posting users. Furthermore, we analyze the peculiarities of Not Safe For Work posts and their corresponding authors. In Chapter 3, we focus on Yelp, where we define a new type of users, namely k-bridges, along with an approach to detect them. Then, we investigate the negative reviews on Yelp and define an approach to identify negative influencers and to evaluate their impact on their neighbors in Yelp.

In Part II, called *"Internet of Things"*, we focus on the representation and management of smart objects in IoT scenarios. In particular, in Chapter 4, we report some preliminary concepts on the Multiple Internet of Things (i.e., MIoT) scenario, which is the foundation of the next approaches. In Chapter 5, we investigate the possibility of improving the communication between objects in a MIoT thanks to the definition of virtual views and the introduction of a MIoT-oriented betweenness centrality. In Chapter 6, we describe an approach to measure the trust, the reputation and the communication scope of the smart objects in a MIoT in order to assess their reliability. In Chapter 7, we propose a framework to preserve the privacy of features and services provided by the smart objects in a MIoT. In Chapter 8, we define a taxonomy of the possible anomalies in a MIoT and describe an algorithm to detect them.

In Part III, called *"Blockchains"*, we focus on the representation and management of blockchains. This part consists of only Chapter 9, where we study the speculative bubble occurred during the years 2017 and 2018. As for it, we investigate the possible existence of speculators.

In Part IV, called *"Further Areas"*, we apply our ideas and approaches to patents, neurological disorders and data lakes. In particular, in Chapter 10, we study the patent citations network, and propose two centrality measures able to capture the peculiarities of this setting. In Chapter 11, we analyze the connectivity of the different brain areas and, then, study the evolution of patients with Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD). In Chapter 12, we propose a model for an effective management of data lakes, in which we fuse both network-based and semantics-driven representations of metadata.

Finally, in Part V, called *"Closing Remarks"*, we draw our conclusions concerning the work presented in this thesis (Chapter 13), and mention some possible developments of our ideas (Chapter 14).

Part I

Social Networks

In this part, we apply our complex network-based approach to model the social network scenario. We investigate the behavior of users in two big social networks and derive useful knowledge patterns for several applications. This part is organized as follows: in Chapter 2, we focus on the definition of subreddit and user stereotype, the evaluation of the assortativity of the authors of posts, and a thorough investigation on Not Safe For Work (i.e., NSFW) posts and the users publishing them. In Chapter 3, we describe our work carried out on Yelp, aimed to define a new type of users (i.e., k-bridge), and investigate the impact of negative reviews and negative influencers on this social network.

2

Reddit

In recent years, Reddit has attracted the interest of many researchers due to its popularity all over the world. In this chapter, we show that, thanks to a complex network-based approach, we are able to extract useful information and make a contribution to the knowledge of this social medium. We first investigate several stereotypes of both subreddits and authors. This analysis is coupled with the definition of three possible orthogonal taxonomies that helps us to classify stereotypes in an appropriate way. Then, we investigate the possible existence of author assortativity in this social medium, paying our attention on co-posters, i.e., authors who submitted posts on the same subreddit. Afterwards, we focus on the Not Safe For Work (i.e., NSFW) posts, which are a real peculiarity of Reddit. We highlight three findings on the main differences between NSFW and SFW posts in Reddit, which allow us to better understand the dynamics (authors, subreddits, readers) behind NSFW posts. It becomes clear that this is a niche world where authors are strongly cohesive.

The material present in this chapter is taken from [165, 166, 208].

2.1 Investigating subreddit and author stereotypes and evaluating author assortativity

2.1.1 Introduction

Reddit[1] is a heterogeneous crowd-sourced news aggregator and online social platform, originally self-declared as "the front page of Internet". It was founded in 2005 and, in few years, has become an ecosystem of 430M+ average monthly active users[2]. At the time of writing, it ranks 19^{th} and 5^{th} in the Alexa's top 500 global and US websites, respectively[3]. Reddit is built on the concept of *subreddit*, which is an interest-based community where users can post and comment contents. A subreddit

[1] https://www.reddit.com

[2] https://www.redditinc.com

[3] https://www.alexa.com/topsites

is identified by a name and is referred to using the /r/ prefix within Reddit, such as /r/science and /r/cats. Currently, there are more than 1.9M subreddits[4]. They are mainly topical, although more general cases exist.

In Reddit, users can submit contents in the form of texts, images and links to external resources. Submitted contents (also simply called posts) can be read by other users and discussed via comments. Users can subscribe to multiple subreddits in order to receive the latest posted contents on their front pages. An important feature of Reddit is *voting*, which represents the mechanism affecting the visibility and the ranking of both posts and comments. In fact, users are allowed to *upvote* or *downvote* posts of other users, so that each submission has a *score*. This is a metric based on the difference between the number of upvotes and the number of downvotes, and it significantly affects the order through which posts and comments are shown to users. However, the exact numbers of upvotes and downvotes are not shown publicly.

Due to the great expansion of Reddit in the latest years, many researchers all over the world have been attracted by this social platform [469, 611, 143, 393, 603, 265, 404]. An overview of the studies on Reddit can be found in [469], whereas an interesting longitudinal analysis on the evolution of this social medium is presented in [611]. Authors have analyzed, and are continuously analyzing, many aspects of Reddit, ranging from community structures and interactions [636, 218, 265] to user behavior [143, 393], from the analysis of the structure and content of subreddits, posts and comments [603] to the analysis of the structural properties of Reddit when it is seen as a social network [265]. Other specific topics, such as text classification [404], user migration [501], political and ideological aspects [308], have been also studied.

In this chapter, we aim at providing a contribution in the knowledge of Reddit by investigating subreddit and author stereotypes and by evaluating author assortativity in this social platform.

The term "stereotype" comes from the combination of two Greek words, namely "stereos" (i.e., solid) and "typos" (i.e., impression). It is adopted to indicate a popular belief about specific groups of individuals. This term first appeared in the press at the end of the 18^{th} century. Later, it was introduced into modern psychology at the beginning of the 20^{th} century by Walter Lippman [430]. The tendency to classify people into groups and to associate each group with a "general idea", a "label" (and, ultimately, a stereotype) is intrinsic to the human mind. As a result, many (both positive and negative) stereotypes have been defined in the history of humanity, in the most disparate areas. Think, for instance, of the stereotypes coined in sport, art, literature, and so on. With the capillary spread of the Web, the practice of coining

[4] https://redditmetrics.com/history

and using stereotypes has extended from real life to Cyberspace [263, 399]. As the Web became increasingly interactive, with the transition to the Web 2.0 and, above all, with the appearance of social networks, the adoption of stereotypes in the Cyberspace become more and more evident [712, 538, 216, 625, 562, 138]. For example, in Facebook, one can encounter stereotypes like "Lime-Lighters", "Emo's", "Philosophy Majors", "Hopeless Romantics", "Ghosts", "Stalkers", "Addicts", and so forth [7]. Similarly, Instagram also presents a wide range of stereotypes [6]. We argue that stereotypes do not necessarily have a negative meaning, as it often happens in real life. On the contrary, they can be extremely useful in everyday communications and interactions in social networks. Here, we want to go one step further; in fact, we claim that it is possible to define "scientific" stereotypes that could be used in scientific applications. We also believe that Reddit fits well for our goal and that, in this context, besides defining stereotypes for the authors of Reddit, it is possible to also introduce stereotypes for subreddits.

The concept of "assortativity" or "assortative mixing" in a social network was introduced in a famous paper of Newman [502]. It is strictly related to the concept of homophily [468] and indicates a network node's predilection to relate to other nodes that are somewhat similar. Several possible similarities could be considered in assortativity, but the most investigated one is node degree. Newman focused on degree assortativity and defined a network as assortative if its nodes having many connections tend to be connected to other nodes with many connections. He showed that social networks are often assortatively mixed, whereas technological and biological networks tend to be disassortative. After Newman, some authors investigated assortativity in several social networks, such as Facebook [140], Twitter [137], Cyworld, Orkut and MySpace [26]. We extend the assortativity analysis to Reddit, which was only marginally considered in the past studies about this topic. We first consider degree assortativity because it is the most studied one in the past. Then, we also analyze eigenvector assortativity. We show that Reddit is assortative with respect to both these centralities, which confirms that also this social platform follows the hypotheses of Newman concerning the existence of assortative mixing in social networks.

The significance and value of our contribution concern both the theoretical and the application viewpoints. From the theoretical point of view, this is the first study on the concept of stereotype in Reddit; actually, approaches for the characterization and identification of *specific* traits of users have been independently presented in different scientific works: users showing multi-community engagement [636], antisocial behaviors [218], community opposers [400], "answer-persons" [143], and "explorers" [327] are some examples. It is also the first research effort to analyze the concept of assortativity in Reddit. Instead, as far as the application point of view

is concerned, we highlight that the knowledge patterns on stereotypes and author assortativity can be employed in a large variety of contexts. Just to cite a few of them, we mention: *(i)* the definition of some guidelines to follow in order to make a subreddit successful; *(ii)* the definition and realization of different categories of recommender systems for Reddit; *(iii)* the definition of an algorithm that finds subreddits to merge or, at least, to integrate; *(iv)* the detection of possible targets for an advertising campaign; *(v)* the definition and implementation of different categories of recommender systems; *(vi)* the definition of an algorithm that builds blacklists of users based on author stereotypes.

The outline of this chapter is as follows. In Section 2.1.2, we describe related literature. In Section 2.1.3, we describe the dataset adopted in our experiments, and we define the stereotypes of both subreddits and authors. Then, in Section 2.1.4, we evaluate the author assortativity, verify a possible correlation between subreddits and authors stereotypes, and present some possible real-world applications of them.

2.1.2 Related Literature

The study of social networks has rapidly become a core research field, thanks to its interdisciplinary aspects [447, 206, 236, 37, 206, 135, 158]. Indeed, many researchers of different disciplines, such as computer scientists, sociologists and anthropologists, exhibited a huge interest in social network analysis [466, 142, 188]. In this context, Reddit is an invaluable source of information, insights and research possibilities. Indeed, it is a prosperous environment, where users share contents and interact with each other. The heterogeneous nature of Reddit, together with the openness and the richness of its data, encouraged scientific community to explore the twists and turns of this platform.

The swift increase of scientific literature related to Reddit has produced a discrete number of papers with several goals and methodologies. In [469], the authors present an overall survey on Reddit, which illustrates several studies on this social network, spanning in time from 2005 to 2018. An interesting longitudinal analysis on the evolution of Reddit is presented in [611].

As pointed out in the Introduction, one of the main theoretical contributions is the study of the concept of author stereotype in Reddit, and the definition and characterization of several stereotypes of interest. As a matter of fact, in past literature, approaches for the characterization and identification of *specific* traits of users have been presented in different papers. Some of the considered traits are: users presenting multi-community engagement [636], anti-social behaviors [218], community opposers [400], “answer-persons” [143], and “explorers” [327]. The main contribution of our work with respect to these proposals is a systematic study of several traits

of users, which are summarized in a wide spectrum of stereotypes and in a suitable classification of them.

In more detail, the "multi-community interaction" trait is studied in [636], where the authors analyze the evolution of communities in which users post in their Reddit "life". They find out that, actually, Reddit users continually post in new communities; in fact, those who leave a community are intended to do so from the very early beginning of their history. Social and anti-social behaviors are analyzed in [218], where the authors apply a definition that extends Brunton's construct of spam in order to separate norm-compliant behaviors from norm-violating ones. This approach also investigates inter-community conflicts by associating social and anti-social homes to users. Conflicts between users are also studied in [400], but from a different point of view. Here, the authors analyze inter-community interactions across 36,000 communities and focus on cases where users of one community, driven by a negative sentiment, submit comments in another community. They highlight how such conflicts actually emerge from a very small number of communities and discuss on strategies for predicting conflicts and mitigating their negative impacts. The presence of users showing the trait of "answer-person" in Reddit is explored in [143], where the authors define an automated method based on user interactions for identifying this role, yet avoiding expensive content analysis. Finally, in [327], the authors present a study regarding highly related communities; in this analysis, they define the characteristics of explorers and non explorers by adopting a specific taxonomy.

The studies and approaches outlined above have been developed considering several communities and subreddits. In [393], a specific subreddit about online User Experience (/r/userexperience) is studied. Here, members socialize and learn together. The authors of this study identify five distinct social roles, namely the "knowledge broker" (i.e., a member that introduces knowledge to the community by sharing links), the "translator" (i.e., a member that offers her academic knowledge into the community), the "conversation facilitator", the "experienced practitioner", and the "learner". Even if the contribution of [393] is particularly interesting because it considers several facets of users' characterization (and, for this feature, it is similar to our work) these classes are specific and valid for the analyzed community only. On the contrary, author stereotypes introduced in our approach cover a wide range of possible facets of users' behavior, with no limitation on the kind and amount of subreddits the users interact with.

As a final remark about stereotyping in the literature, it is worth observing that our proposal introduces both author and subreddit stereotypes. To the best of our

knowledge, the definition of subreddit stereotypes received no attention in the literature and, consequently, it represents a step forward in the research on Reddit.

As far as this last aspect is concerned, we pointed out in the Introduction that one of the main potential applications of subreddit stereotyping is the definition of guidelines in order to make a subreddit successful. With respect to this topic, some papers studied how to predict the success of a subreddit or, more generally, of a community from different perspectives. In particular, the authors of [212] investigate the success and group dynamics of online communities, focusing on Reddit ones. In detail, they identify four success measures desirable for most communities, spanning from the growth of the numbers of members to the volume of activities within the community, and capturing different kinds of success. They also investigate the prediction of the final success of a new community. Furthermore, the authors of [679] present a broad exploration of posts, with a particular interest to comments. Here, they aim at fulfilling three different tasks. The first is analyzing a comment thread by looking at its topical structure and evolution; the second consists of exploiting comment threads to enhance web search; the third aims at distilling useful features to predict the final score of a comment. Finally, in [603], the authors investigate both the behavioral context of user posting and the polarization of user responses.

The main difference between the above mentioned approaches and the stereotyping activity proposed here is that the former observe communities evolution and, possibly, predict their success, whereas the latter could be used to provide *guidelines* for promoting *specific actions* to obtain the desired success. From a data analytics point of view, the former focuses on descriptive and predictive analytics, whereas the latter also performs diagnostic and prescriptive one.

As pointed out in the Introduction, another contribution is the study of assortativity in Reddit. While this topic has been analyzed with reference to other social platforms [140, 26, 137], only few works marginally analyzed it on Reddit. In particular, in [317], the authors focus on studying loyal communities, finding that they tend to be less assortative as long as their interaction level increases. In this case, assortativity is studied on monthly interaction networks, where users are considered connected if they submit a comment in the same comment chain with a gap of at most two comments. The authors also carry out a comparison with a null model and find that the difference between loyal communities and their random counterparts disappears. This result implies that users in loyal communities tend to interact with dissimilar users as a consequence of the community's activity. Actually, in [317], assortativity is used as a tool for characterizing loyal communities, studying single chains of comments. On the contrary, we study assortativity from a more general point of view, in order to provide an overall characterization of Reddit users across

several subreddits and comments. Furthermore, we study both degree assortativity and eigenvector assortativity.

Another work marginally related to our study on assortativity in Reddit is presented in [265]. Here, the authors discuss the rise of new trends in complex networks by looking at vertices that "shine" (i.e., high-degree vertices), also called network stars. They study the evolution of some complex networks, with Reddit among them. They analyze the temporal dynamics of the networks by looking at how different features, such as density and average clustering coefficient, change over time. Clearly, [265] and our approach are quite different. Indeed, differently from what happens in [265], our assortativity definition does not allow the analysis of temporal dynamics, that is the main goal of [265]. On the other side, it helps to characterize the tendency of users to associate with each others.

Other works, marginally related to our proposal, focus on the study of specific aspects of subreddits or user behaviors. For instance, in [404], the authors use text classification and computational critical discourse analysis to distinguish and interpret ideological differences between subreddits. In [713], the authors present a study regarding a quantitative, language-based typology of communities' identity, revealing how several social phenomena manifest across communities. The introduced taxonomy is based on two aspects of community identity, i.e., distinctiveness and dynamicity. User migration is studied in [501]. Here, Reddit is examined during a period of community unrest in order to identify the motivations for this kind of behavior. Political and ideological aspects emerging in Reddit are discussed in [308, 55, 302, 616]. Finally, in [264], the authors present a mixed-method study of 100,000 subreddits and their rules in order to define effective mechanisms for community governance.

2.1.3 Methods

2.1.3.1 Dataset description

The dataset required for our activity was downloaded from the `pushshift.io` website, which is one of the most known Reddit data sources. Our dataset contains all the posts published on Reddit from January 1^{st}, 2019 to September 1^{st}, 2019. All the posts wrote in a month were added to the dataset at the end of the next month. The number of posts available for our investigation was 150,795,895. For each post, we considered the following set of attributes: `id`, `subreddit`, `title`, `author`, `created_utc`, `score`, `num_comments` and `over_18`.

In order to carry out our experiments, we used a server equipped with 16 Intel Xeon E5520 CPUs and 96 GB of RAM with the Ubuntu 18.04.3 operating system.

We adopted Python 3.6 as programming language, its library Pandas to perform ETL operations on data, and its library NetworkX to perform operations on networks.

During the ETL phase, we observed that some of the available posts referred to authors that had left Reddit. We decided to remove these posts from our dataset. At the end of this last activity the number of posts at our disposal was 122,568,630.

We computed the number of authors who submitted these posts; it was equal to 12,464,188. Then, we found the number of the subreddits which they referred to; it was equal to 1,356,069.

Now, we describe some preliminary investigations on Reddit, concerning posts, comments, and authors.

Investigation on posts

We started this investigation by performing the following analyses on posts:

- distribution of subreddits against posts (Figure 2.1); it follows a power law with $\alpha = 1.651$ and $\delta = 0.014$;
- distribution of authors against posts (Figure 2.2); it follows a power law with $\alpha = 1.431$ and $\delta = 0.016$;
- distribution of posts against scores (Figure 2.3); it follows a power law with $\alpha = 1.600$ and $\delta = 0.005$.

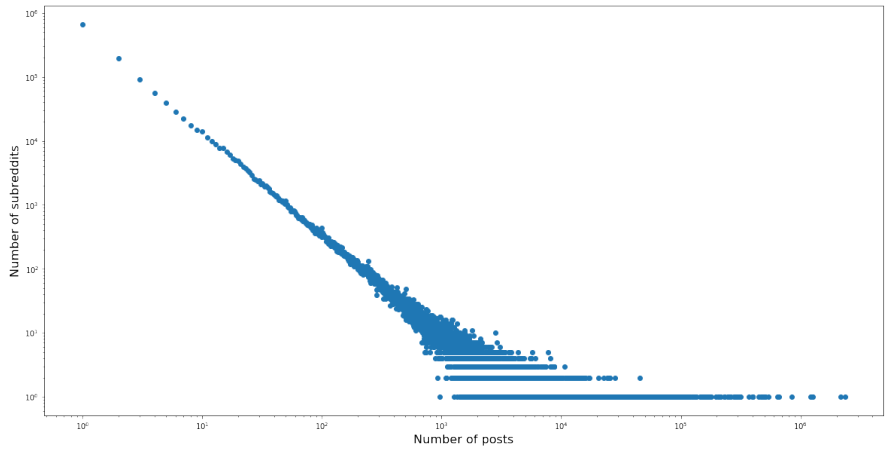


Fig. 2.1: Distribution of subreddits against posts (log-log scale)

The maximum number of posts with the same score is 51,721,824. Interestingly, these posts have associated a score equal to 1. Instead, the number of posts with a score equal to 0 or 2 is much smaller. This trend can be explained considering that a post submitted on Reddit starts with a score of 1. As a consequence, when no other author upvotes or downvotes it, the final score of the post is 1.

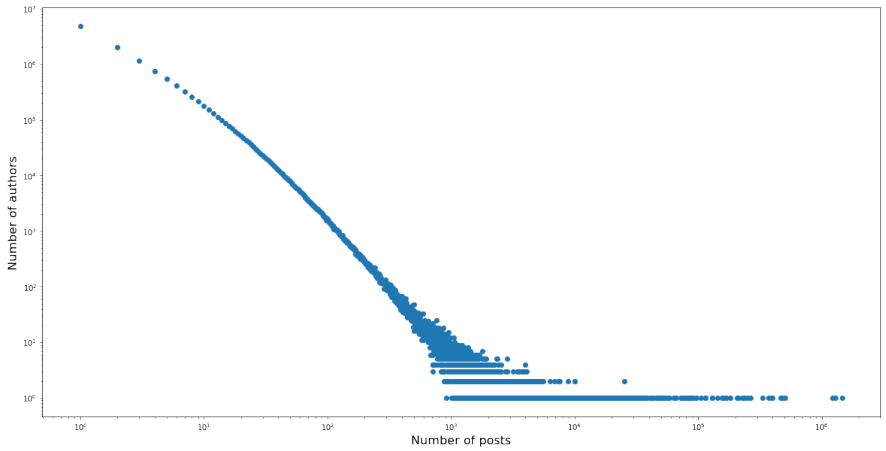


Fig. 2.2: Distribution of authors against posts (log-log scale)

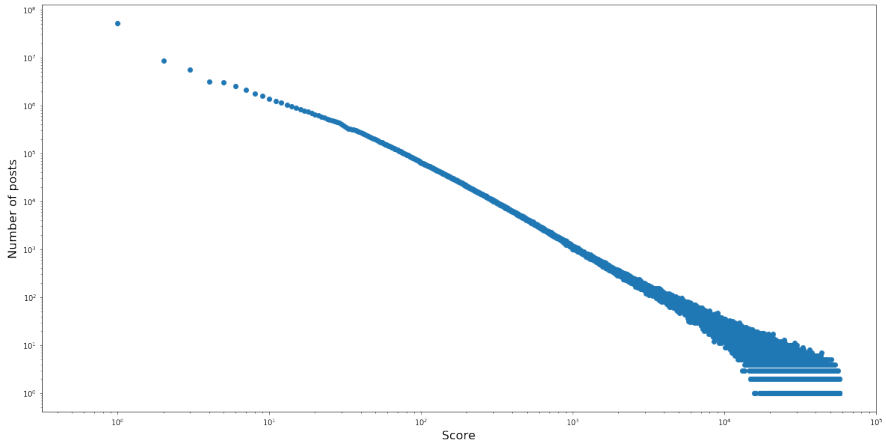


Fig. 2.3: Distribution of posts against scores (log-log scale)

We also observe that no post has a negative score. This fact is due to Reddit that shows and returns a score equal to 0 for a post whenever the number of downvotes is higher than the number of upvotes, i.e., also when the real score of the post is negative. So, posts with a score equal to 0 are to all intents and purposes intended as "negative" posts.

At this point, we also computed:

- the distribution of authors against negative posts (Figure 2.4); it follows a power law with $\alpha = 2.274$ and $\delta = 0.030$.
- the distribution of authors against positive posts (Figure 2.5); it follows a power law with $\alpha = 2.074$ and $\delta = 0.014$.

As for these two distributions, we found that the number of positive posts is about 16 times the number of negative ones.

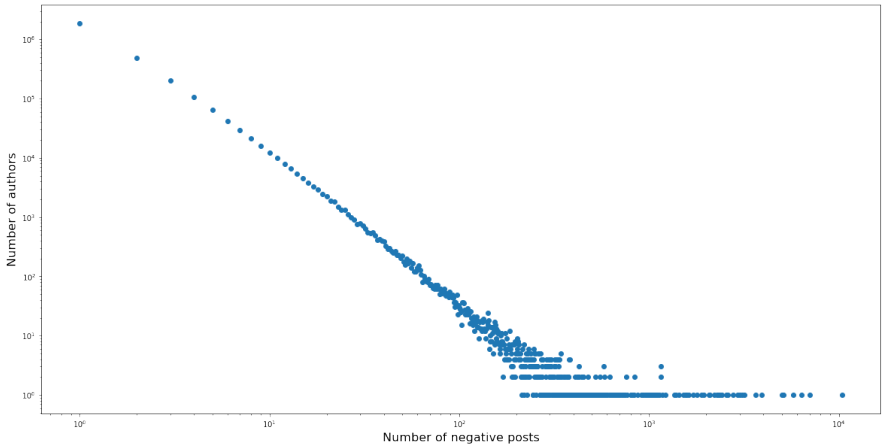


Fig. 2.4: Distribution of authors against negative posts (log-log scale)

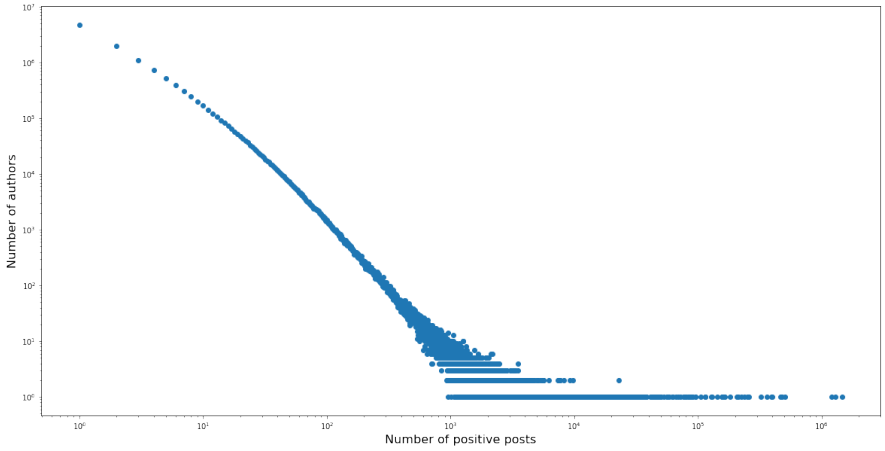


Fig. 2.5: Distribution of authors against positive posts (log-log scale)

Analysis of positive and negative posts for SFW and NSFW cases

In the previous section, we have observed that each post has a score, initially equal to 1, which can increase or decrease based on the upvotes or downvotes of users. Actually, Reddit does not report the posts with a negative score in its database. For this reason, the values of the scores both in Reddit and in `pushshift.io` range in the interval $[0, +\infty)$. In this setting, posts with a score equal to 0 are particularly relevant, because they are the only ones that have been rated negatively by at least one user, or have received more downvotes than upvotes.

We computed the distributions of authors against negative posts for both SFW and NSFW posts. In both cases, we have found that they follow a power law. We report the main parameters of these distributions in Table 2.1.

A Wilcoxon rank sum test showed that the number of authors of Jan-Feb SFW negative posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 5.1 \cdot 10^{-4}, p < 0.01$).

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of authors	66,162 (92.31%)	24,607 (74.86%)	61,254 (91.98%)	24,172 (73.87%)
Number of authors of the 99 percentile	40,028	11,606	40,024	11,598
Maximum number of posts	133 (9.64%)	460 (14.38%)	103 (8.98%)	399 (13.76%)
Number of posts of the 99 percentile	126	369	122	370
Average number of authors	1,666	505	1,691	544
Average number of posts	32	49	28	47
α (power law parameter)	1.4360	1.4349	1.5512	1.4360
δ (power law parameter)	0.0615	0.0,0616	0.0543	0.0616

Table 2.1: Parameters of the distributions of authors against negative posts

These conclusions, although interesting, must be intertwined with those regarding positive posts, to better characterize the features of negative ones. For this reason, we computed the distributions of authors against positive posts. Also in this case, the distributions follow a power law similar to the previous ones. We report the values of the main parameters of these distributions in Table 2.2.

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of authors	522,540 (79.66%)	124,054 (56.56%)	519,774 (79.54%)	126,602 (56.89%)
Number of authors of the 99 percentile	9,083	4,346	9,080	4,352
Maximum number of posts	18,684 (11.88%)	16,383 (5.77%)	16,481 (10.67%)	15,564 (5.73%)
Number of posts of the 99 percentile	5,165	4,638	5,160	4,641
Average number of authors	2,018	418	1,944	394
Average number of posts	483	541	493	514
α (power law parameter)	1.4318	1.5145	1.4855	1.5498
δ (power law parameter)	0.0311	0.0263	0.0275	0.0291

Table 2.2: Parameters of the distributions of authors against positive posts

A Wilcoxon rank sum test indicated that the number of authors of Jan-Feb SFW positive posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 1.1 \cdot 10^{-4}, p < 0.01$).

We now compare Tables 2.1 and 2.2 to extract the features characterizing negative posts versus positive ones. There are no significant differences between positive and negative posts in the maximum and average number of authors of NSFW and SFW posts. The same is true for the average number of posts and the trends of the power law distributions. However, there is a very interesting aspect that differentiates negative posts from positive ones. Indeed, the maximum number of negative posts is much higher for NSFW posts than for SFW ones. This trend is not found in positive posts.

The explanation behind this result is the same as the one seen previously.

Investigation on comments

As for this investigation, we computed:

- The distribution of subreddits against comments (Figure 2.6); it follows a power law with $\alpha = 1.730$ and $\delta = 0.015$.
- The distribution of the average number of comments against the scores of the posts they refer to (Figure 2.7). Interestingly, in this case, we have a roughly Gaussian distribution, whose mean is at a score near to 50,000. The distribution presents several outliers. For instance, for a score equal to 79,470, we have a post with a number of comments equal to 71,225.
- the distribution of posts against comments (Figure 2.8); it follows a power law with $\alpha = 1.455$ and $\delta = 0.011$.

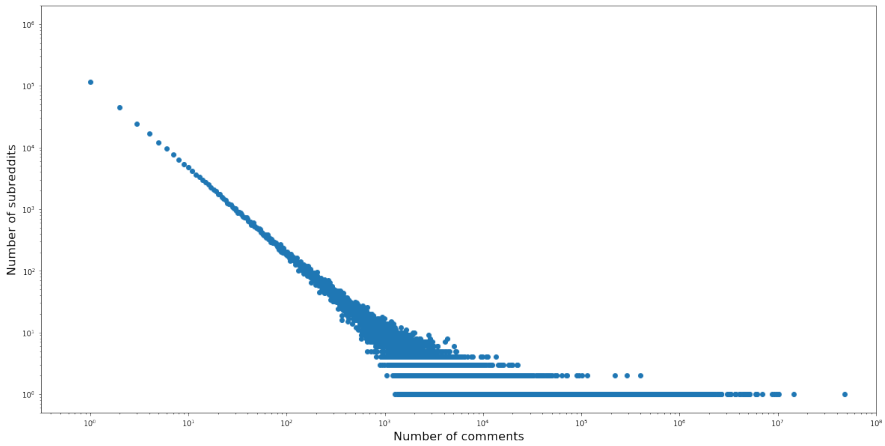


Fig. 2.6: Distribution of subreddits against comments (log-log scale)

Finally, we considered the 150 posts with the highest number of comments and the subreddits they were submitted to. We obtained only 31 subreddits. Then we computed the average number of comments for *all* the posts submitted in each of these subreddits. The results obtained are reported in Figure 2.9. From the analysis of this figure, we can observe that the distribution is very irregular. It decreases quickly for the first three subreddits, very slowly for the next 13 subreddits, quickly for the next 9 subreddits and, finally, it suddenly drops and becomes almost zero.

Investigation on authors

First, we determined the distribution of authors against subreddits (Figure 2.10). It follows a power law with $\alpha = 1.702$ and $\delta = 0.081$.

Afterwards, we selected the 150 posts with the highest number of comments and the corresponding authors. Interestingly, we had only 26 authors for all the

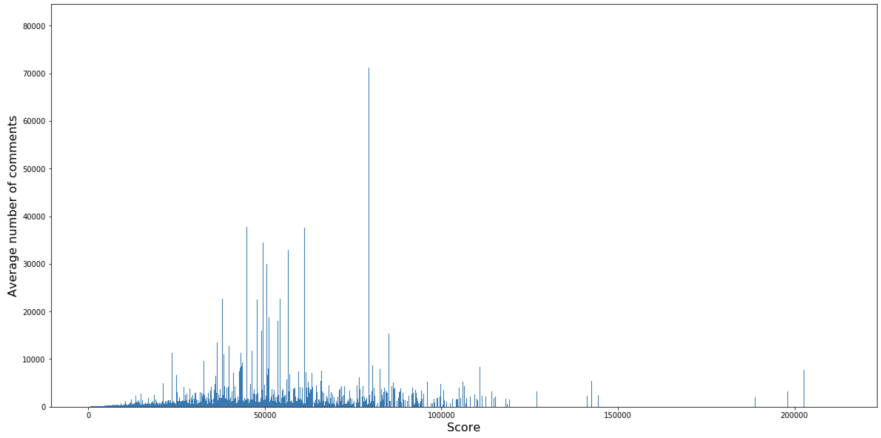


Fig. 2.7: Distribution of the average number of comments against the scores of the posts they refer to

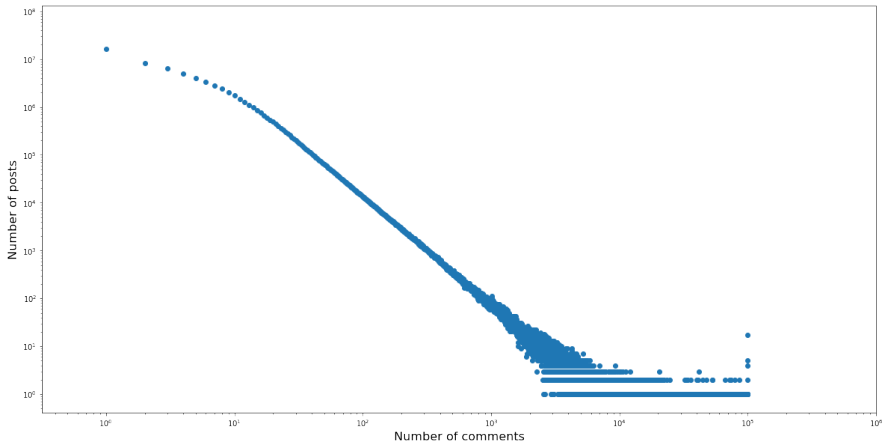


Fig. 2.8: Distribution of posts against comments (log-log scale)

150 posts. These can be considered as the most commented authors in Reddit and, maybe, they are influencers. Then, we computed the average number of comments for *all* the posts each author submitted. The results obtained are reported in Figure 2.11. From the analysis of this figure we can observe that the decrease of the distribution is roughly stepwise.

2.1.3.2 Stereotyping subreddits

In order to determine some possible stereotypes of subreddits, we start investigating the subreddit lifespan. As a first step, we considered the subreddits created in January 2019 and then verified the month when they performed their last activity (and, therefore, presumably died). The results obtained are reported in Figure 2.12. Here, an activity level of 1 implies that the subreddit died in the same month it was born, an activity level of 2 suggests that it died one month after it was born, and so on. An

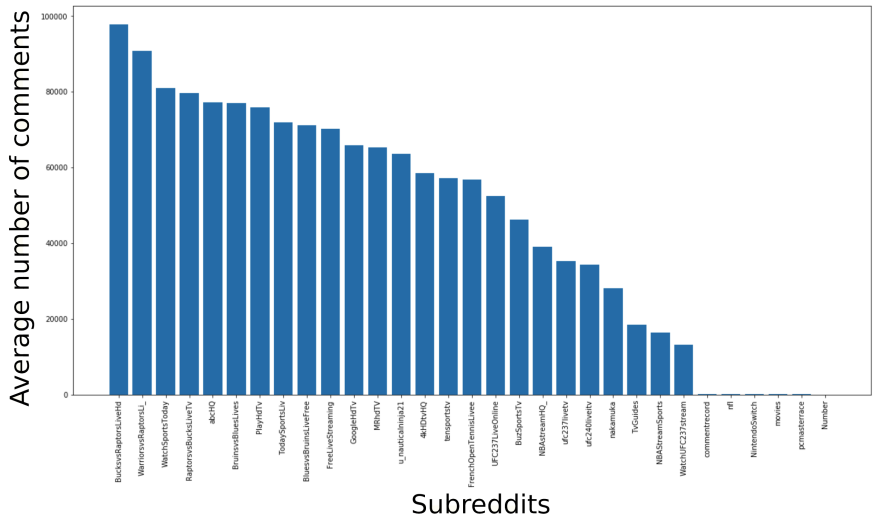


Fig. 2.9: Distribution of the average number of comments submitted to the subreddits receiving the 150 most commented posts

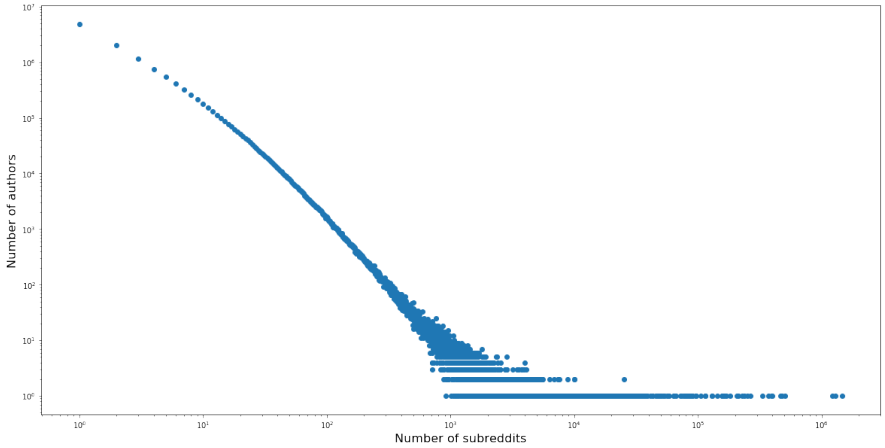


Fig. 2.10: Distribution of authors against subreddits (log-log scale)

activity level of 8 indicates that it is still alive (we recall that our dataset comprises data from January 1^{st}, 2019 to September 1^{st}, 2019). We proceeded in the same way for the subreddits created in February, March, and so forth. For instance, in Figure 2.13, we report the trends of the subreddits created in February 2019 and in March 2019.

After this, we focused on those subreddits died in the same month they were born. We analyzed their corresponding lifespan and we observed that almost all of them died in the same day they were born. For instance, in Figure 2.14, we report the trends of the subreddits born and died in February 2019 and in March 2019.

Then, we decided to deeply investigate those subreddits died in the same day they were born. We computed their distribution against the number of their posts. Figure 2.15 shows what happens for January 2019; the same trend can be observed

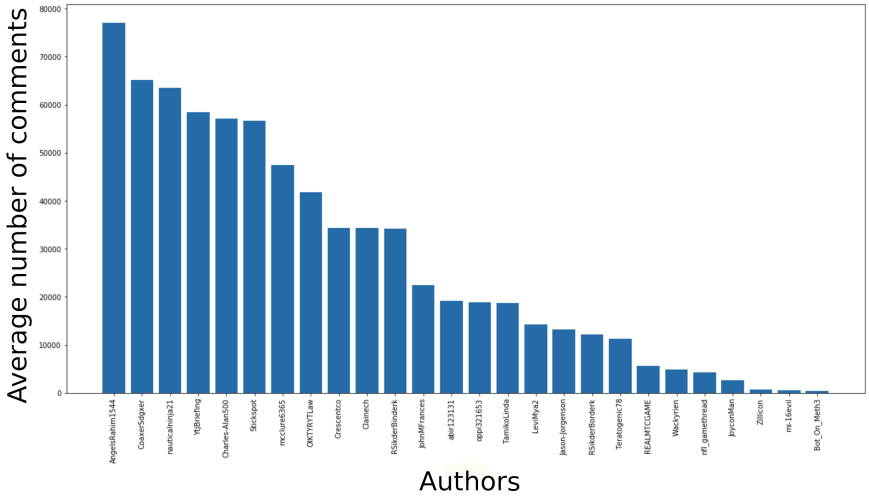


Fig. 2.11: Distribution of the average number of comments received against the authors submitting the 150 most commented posts

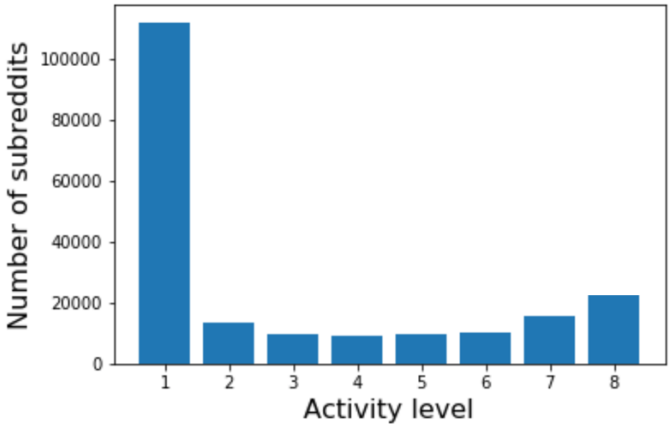


Fig. 2.12: Lifespan of the subreddits created in January 2019

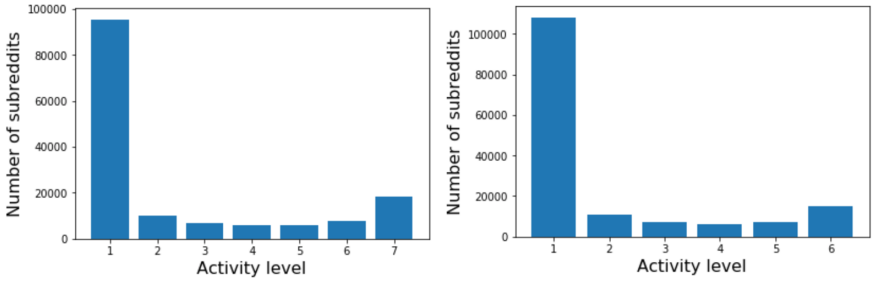


Fig. 2.13: Lifespan of the subreddits created in February 2019 (at left) and March 2019 (at right)

for the other months of this year. Clearly, this distribution follows a power law, a trend that can be observed also for similar subreddits born in the other months. From its analysis we observe that most of the subreddits, which died in the same day they

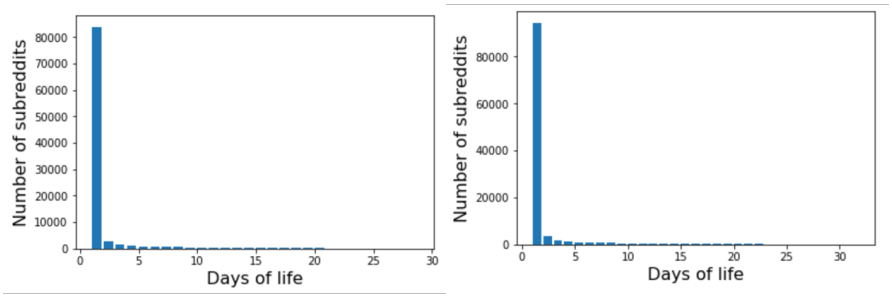


Fig. 2.14: Lifespan of the subreddits born and died in February 2019 (at left) and March 2019 (at right)

were born, have only one post. At this point, we computed the distribution of these subreddits against the number of comments. In Figure 2.16, we show the subreddits of January 2019, even if the same trend can be observed for the other months of this year. From the analysis of this figure we can note that this distribution follows a power law. Furthermore, most of these subreddits have no comments.

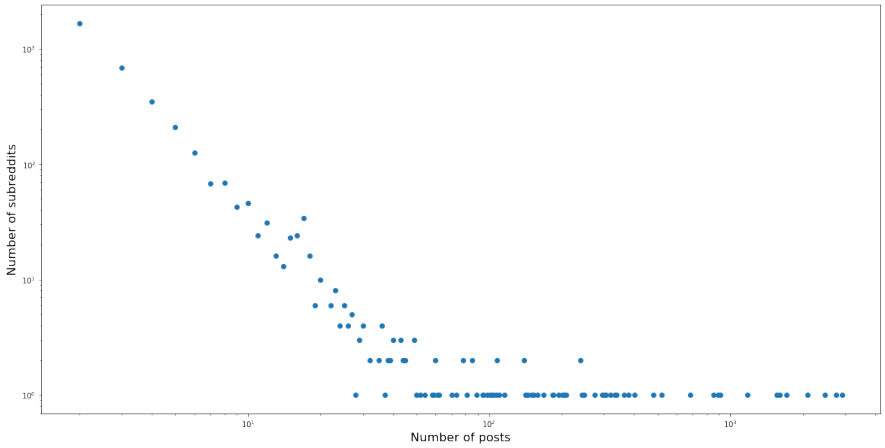


Fig. 2.15: Distribution of the subreddits of January 2019 died in the same day they were born against the number of their posts

Next, we examined a second class of subreddits, similar to the previous one. In fact, we selected all those subreddits that died one day after they were born. Again, we first computed their distribution against the number of posts. In Figure 2.17, we show what happens for the subreddits of January 2019; again, the same trend was found for all the other months. This distribution follows a power law, which was expected. The unexpected thing was that the minimum number of posts was 2 and not 1. Even more unexpectedly, this trend is also confirmed for the subreddits with the same features born in the other months. After that, we computed the distribution of these subreddits against the number of comments. In Figure 2.18, we show it for

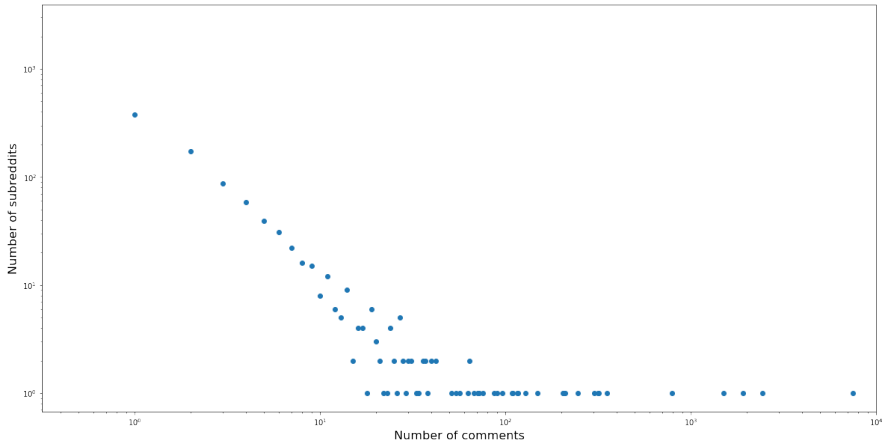


Fig. 2.16: Distribution of the subreddits of January 2019 died in the same day they were born against the number of their comments

the subreddits of January 2019; the same trend can be observed for all the other months. From the analysis of this figure, we note that this distribution follows a power law. Furthermore, most of these subreddits have no comments.

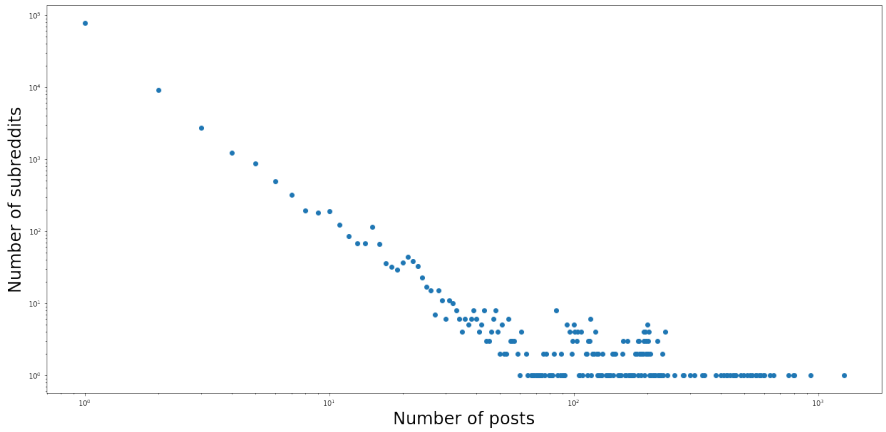


Fig. 2.17: Distribution of the subreddits of January 2019 died one day after they were born against the number of their posts

Note that the two classes of subreddits above have a proper characterization that differentiates them from all the other classes of subreddits (for instance, the ones that survived for some months). They also have few features distinguishing them from each other. However, the number of their similarities is much higher than the number of their differences. As a consequence, both these two classes can be considered as a "macro-category" of stereotypes that we call "dead in crib". At this point, by deepening what we have found previously, we have determined the following

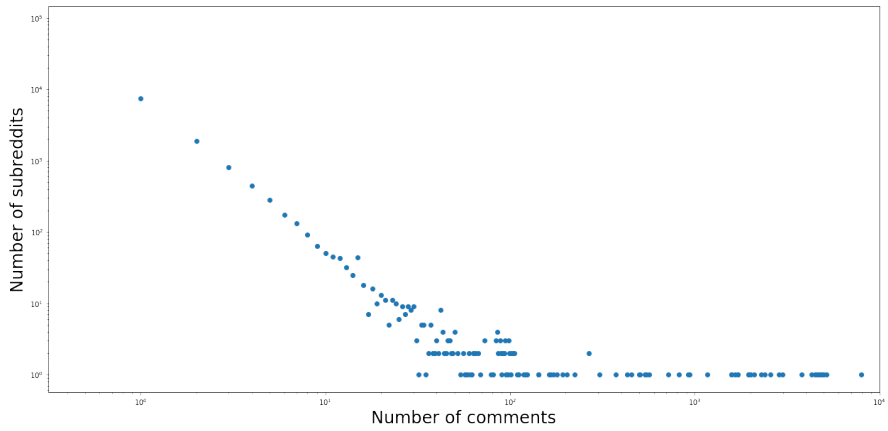


Fig. 2.18: Distribution of the subreddits of January 2019 died one day after they were born against the number of their comments

stereotypes characterizing the subreddits "dead in crib" (i.e., those subreddits who died at most one day after they were born):

- *User Profile*: it is associated with a user profile.
- *Unsuccessful Subreddit*: it initially stimulated several interactions. However, after few hours, these interactions finished and it quickly died.
- *Comment Grabber*: it had at least one post capable of stimulating a debate, even if minimal.
- *Private Community*: it requires an invitation to be accessed. It is often associated with a specific event of interest for a specific community.
- *Banned Subreddit*: it was banned probably because it was associated with a spammer.
- *Bot*: it can be recognized because its posts are always similar and consist of links and comments with links.

In order to characterize these stereotypes, and all the others that we will consider in the following, we have defined three possible orthogonal taxonomies. These are based on:

- the number of posts; we considered two possible classes, i.e., few posts and many posts;
- the number of comments; we considered two possible classes, i.e., few comments and many comments;
- the number of authors; we considered two possible classes, i.e., few authors and many authors.

Taking these three taxonomies into consideration, the previous stereotypes can be classified as shown in Tables 2.3 and 2.4.

Observe that a stereotype can often belong to both the classes of a taxonomy. This implies that it cannot be "categorized" based on that taxonomy. For instance, *Comment Grabber*, in presence of many comments and many authors, can be found with both few posts and many posts. This implies that this stereotype can be characterized only by the number of comments and the number of authors, but not by the number of posts. Analogously, in presence of many posts, *Banned Subreddit* cannot be characterized by the number of comments or the number of authors. By contrast, in presence of few posts, *Banned Subreddits* is characterized by few comments and few authors.

	Few Authors	**Many Authors**
Few Comments	User Profile Unsuccessful Subreddit Banned Subreddit	Unsuccessful Subreddit
Many Comments	Unsuccessful Subreddit Comment Grabber User Profile	Private Community Bot Unsuccessful Subreddit Comment Grabber

Table 2.3: Classification of stereotypes concerning the subreddits "dead in crib" - Few posts case

	Few Authors	**Many Authors**
Few Comments	User Profile Unsuccessful Subreddit Banned Subreddit	Unsuccessful Subreddit Bot Banned Subreddit
Many Comments	User Profile Banned Subreddit	Private Community Banned Subreddit Unsuccessful Subreddit Comment Grabber

Table 2.4: Classification of stereotypes concerning the subreddits "dead in crib" - Many posts case

After having investigated the stereotypes of the subreddits "dead in crib", we focused on the opposite category of subreddits, i.e., those survived for all the months of reference for our dataset. We collectively call them "survivors" in the following. We applied the same reasoning and tasks that we have made for the subreddits "dead in crib" and we obtained the following stereotypes:

- *User Profile*, *Bot*: these are the same ones we have seen for the subreddits "dead in crib".

- *Cringe / NSFW Subreddit*: it contains strange or strong-content posts, submitted by only one user, or, alternatively, it is an NSFW subreddit.
- *Niche Subreddit*: its topics are niche ones, and it draws the attention of users interested in them.
- *Successful Subreddit*.
- *Big Comment Grabber*: almost all the posts submitted in it stimulate a debate.
- *Utility Subreddit*: it is conceived to support a specific activity (think, for instance, of a subreddit where users ask for a translation).

Based on the three taxonomies defined above, the previous stereotypes can be classified as shown in Tables 2.5 and 2.6.

	Few Authors	**Many Authors**
Few Comments	User Profile Bot Cringe /NSFW Subreddit Niche Subreddit	Successful Subreddit Niche Subreddit
Many Comments	Successful Subreddit Niche Subreddit Big Comment Grabber	Big Comment Grabber Successful Subreddit Niche Subreddit

Table 2.5: Classification of stereotypes concerning the subreddits “survivors” - Few posts case

	Few Authors	**Many Authors**
Few Comments	Niche Subreddit	Cringe / NSFW Subreddit Niche Subreddit
Many Comments	Big Comment Grabber Utility Subreddit	Successful Subreddit

Table 2.6: Classification of stereotypes concerning the subreddits “survivors” - Many posts case

After these analyses on the stereotypes belonging to the two extreme categories “dead in crib” and “survivors”, we decided to apply the same reasonings and tasks to investigate a third category of stereotypes, intermediate between the two previous ones. Specifically, we focused on those subreddits that lived five months after their creation and, then, died. We call this category “undelivered promises” and we obtained the following stereotypes for it:

- *User Profile*, *Niche Subreddit*, *Bot*, *Cringe / NSFW Subreddit*, *Private Community*, *Banned Subreddit*: these are the same ones we have seen for the previous categories.

- *Unsuccessful Boomer*: it was successful for a while, but died after a period of decline.
- *Unsuccessful Zombie*: it was born without praise or blame, managed to survive for a while in a gray way and, finally, died.

Based on the three taxonomies that we defined above, the previous stereotypes can be classified as shown in Tables 2.7 and 2.8.

	Few Authors	**Many Authors**
Few Comments	User Profile Niche Subreddit Bot	Bot Cringe / NSFW Subreddit Niche Subreddit Unsuccessful Boomer
Many Comments	User Profile Private Community Unsuccessful Boomer Niche Subreddit	Niche Subreddit Private Community Unsuccessful Boomer

Table 2.7: Classification of stereotypes concerning the subreddits "undelivered promises" - Few posts case

	Few Authors	**Many Authors**
Few Comments	User Profile Cringe / NSFW Subreddit Bot Unsuccessful Zombie	Private Community Banned Subreddit Niche Subreddit
Many Comments	User Profile Bot Cringe / NSFW Subreddit	Cringe / NSFW Subreddit Banned Subreddit Unsuccessful Boomer

Table 2.8: Classification of stereotypes concerning the subreddits "undelivered promises" - Many posts case

2.1.3.3 Stereotyping authors

In order to determine the possible author stereotypes, we proceeded in a way analogous to what we have done to define subreddit stereotypes. In fact, also for authors, we found three macro-categories of stereotypes, namely "very positive", "neutral" and "very negative" authors. To better understand the reasoning underlying these categories, we recall that, in Section 2.1.3.1, we have found that the number of positive posts is about 16 times the number of negative ones in Reddit. As a consequence, it is possible to use this result as a baseline for a preliminary author classification. Specifically, we considered an author as "very positive" if the number of positive

posts submitted by her is at least $2 \cdot 16 = 32$ times the number of negative ones, which means at least twice the typical number of positive posts submitted for each negative one by a user. Instead, we considered an author as "neutral" if the number of positive posts submitted by her is between 1 and 16 times the number of negative ones. Finally, we considered an author as "very negative" if the number of negative posts submitted by her is at least 16 times the number of positive ones. Clearly, this classification is not exhaustive and it is also empirical because it derives from our observation on the behaviors of users in Reddit. However, we feel that it is useful to provide a first definition of three macro-categories of author stereotypes possibly interesting for application scenarios.

Analogously to what we have done for subreddit stereotypes, we have defined two possible orthogonal taxonomies, namely:

- the number of posts: the possible classes are few posts and many posts;
- the number of comments: the possible classes are few comments and many comments.

Afterwards, we determined the following stereotypes characterizing the "very positive" authors, proceeding in a way analogous to the one we adopted for subreddit stereotypes:

- *Unsuccessful Author*: she submits posts but she is never capable of stimulating interactions with other authors.
- *Fame Seeker*: she submits (and/or she is still submitting) an impressive amount of posts in order to reach fame in Reddit.
- *Cringe / NSFW Author*: she often submits cringe / NSFW posts.
- *FBG Publisher* (Few But Good Publisher): she does not publish a very high number of posts; however, her posts are generally appreciated by other users.
- *Content Creator*: she creates and submits contents for people.
- *Successful Author*: she submits many posts that receive many positive comments and are appreciated by other users.
- *Reposter*: she simply re-submits posts of other authors.

Based on the two taxonomies that we defined above, the previous stereotypes can be classified as shown in Table 2.9.

After the "very positive" authors, we focused on the opposite macro-category of author stereotypes, i.e., the "very negative" ones. We obtained the following stereotypes, applying the same reasoning and performing the same tasks that we made for "very positive" authors:

	Few Posts	**Many Posts**
Few Comments	Unsuccessful Author	Fame Seeker Cringe / NSFW Author
Many Comments	FBG Publisher Content Creator	Successful Author Reposter

Table 2.9: Classification of the stereotypes concerning "very positive" authors

- *Unsuccessful Author*: this stereotype is the same as we have seen for "very positive" authors.
- *Spammer*: she is an author submitting a lot of spam posts evaluated negatively by other users.
- *Hatred Sower*: she is a user whose goal is attacking minority groups with hate posts or comments.
- *Instigator*: she is an author using every opportunity to make herself known. For her, it is not important how she is judged, but the fact that one speaks of her.

Based on the two taxonomies defined above, the previous stereotypes can be classified as shown in Table 2.10.

	Few Posts	**Many Posts**
Few Comments	Unsuccessful Author	Spammer
Many Comments	Hatred Sower	Instigator

Table 2.10: Classification of the stereotypes concerning "very negative" authors

After having analyzed the stereotypes belonging to the two extreme categories, i.e., "very positive" and "very negative" authors, we decided to investigate "neutral" authors as representative of a third macro-category, intermediate between the two previous ones. We obtained the following stereotypes, applying the same reasoning and tasks that we made for the other two macro-categories:

- *Unsuccessful Author* and *Fame Seeker*: these stereotypes are the same ones we have seen for the previous macro-categories.
- *PP Author* (Private Purpose Author): she often creates subreddits for private purposes, for instance to talk about specific topics of interest for a particular community. Often, her subreddits require an invitation for being accessed.
- *Bot*: it is a bot; it can be recognized because it always submits similar posts consisting of links and comments with links.
- *Moody Author*: she creates subreddits and submits posts whose topics, expressed positions, and evaluations apparently swing without a logic.

- *Comment Grabber*: she occasionally submits posts capable of stimulating a debate, even if minimal.
- *Big Comment Grabber*: almost all the posts submitted by her stimulate a debate.

Based on the two taxonomies defined above for authors, the previous stereotypes can be classified as shown in Table 2.11.

	Few Posts	**Many Posts**
Few Comments	Unsuccessful Author	Fame Seeker Bot
Many Comments	PP Author Comment Grabber	Moody Author Big Comment Grabber

Table 2.11: Classification of the stereotypes concerning "neutral" authors

2.1.4 Results

2.1.4.1 Evaluating author assortativity

In the past, assortativity has been largely analyzed in several social media [140]. In this section, we aim at checking if a form of assortativity exists in Reddit; in particular, we focus on co-posters, i.e., authors submitting posts on the same subreddit.

In order to perform our analyses, we define a support network $\mathcal{P}$, which we call co-post network. Formally speaking:

$$\mathcal{P} = \langle N, E \rangle$$

Here, N is the set of the nodes of $\mathcal{P}$; there is a node $n_i \in N$ for each author a_i who submitted at least one post. There is an edge $(n_i, n_j, w_{ij}) \in E$ if the authors a_i and a_j (associated with the nodes n_i and n_j, respectively) submitted at least one post in the same subreddit. w_{ij} indicates the number of subreddits having at least one post of a_i and, simultaneously, at least one post of a_j.

The number of nodes of $\mathcal{P}$ is equal to the number of authors in our dataset, i.e., 12,464,188. The number of arcs of $\mathcal{P}$ is about 925 billion. The density of this network is 0.00596, whereas the average clustering coefficient is 0.43753.

First of all, we computed the degree centrality of the nodes of $\mathcal{P}$. In Figure 2.19, we report the corresponding distribution. This figure shows that degree centrality follows a power law, even if disturbed. This result is in line with the theory regarding this kind of centrality [647]. The maximum value of degree centrality is 1,820,412, while the minimum value is 0.

We sorted the corresponding authors in a descending order, based on their degree centrality, to verify the possible presence of a degree assortativity in Reddit. Then,

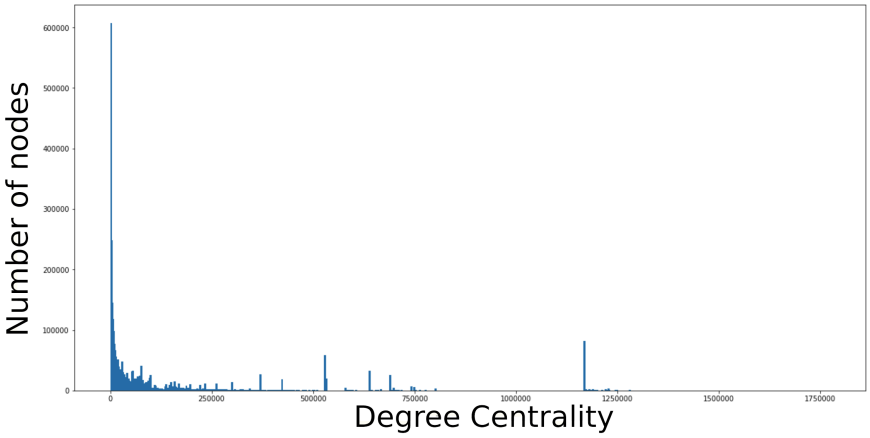
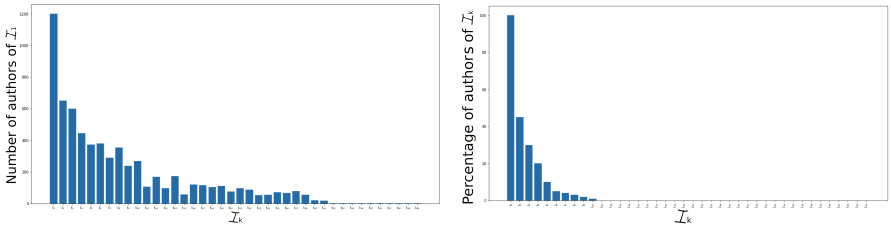


Fig. 2.19: Distribution of degree centrality for the nodes of $\mathcal{P}$

we divided the sorted list into intervals of authors. In particular, we considered equi-width intervals $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_{40}\}$, each consisting of 312,500 authors[5]. As a consequence, the interval $\mathcal{I}_k$, $1 \leq k \leq 39$, contained the authors of the sorted list comprised in the interval $(312,500 \cdot (k-1), 312,500 \cdot k]$, open at left and closed at right. The interval $\mathcal{I}_{40}$ contained the authors comprised in the interval $(12,187,500\ ,\ 12,464,188]$.

First of all, we considered the first interval $\mathcal{I}_1$ and, for each interval $\mathcal{I}_k$, $1 \leq k \leq 40$, we determined how many authors of $\mathcal{I}_1$ are connected to at least one author of $\mathcal{I}_k$. The results obtained are reported in Figure 2.20(a). Then, we computed the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_1$. The results obtained are reported in Figure 2.20(b). From the analysis of Figure 2.20, it is clear that a strict correlation (i.e., a sort of backbone) exists among the authors with the highest degree centrality.



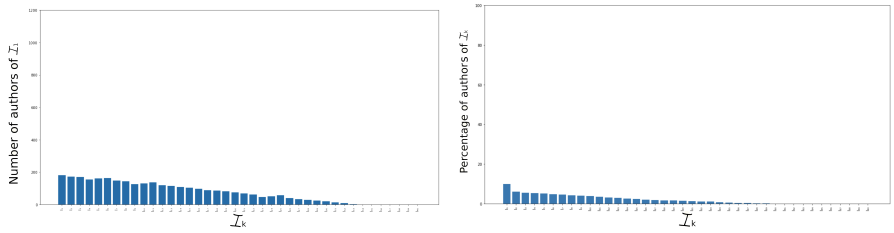
(a) **(b)**

Fig. 2.20: (a) Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$

In order to prove the statistical significance of our results, we generated a null model to compare our findings with the ones obtained in an unbiasedly random sce-

[5] Actually, the last interval had a width slightly lower than the other ones.

nario. Specifically, we built our null model shuffling the arcs of $\mathcal{P}$ (that, in our case, represent co-posting relationships) among the nodes of this network. In this way, we left unchanged all the original features of $\mathcal{P}$ with the exception of the distribution of co-posting tasks, which became unbiasedly random in the null model. After that, we repeated the previous analyses on the null model. The results obtained are reported in Figure 2.21. Comparing this figure with Figure 2.20, we can see that the distributions represented therein are similar, in a way that many of the intervals with the highest values in Figure 2.20 continue to reach the highest values in Figure 2.21. However, in this last case, the values are much smaller. Therefore, we can conclude that the behavior observed in Figure 2.20 (and the consequent possible degree assortativity revealed by them) is not random but it is intrinsic to Reddit.



(a) **(b)**

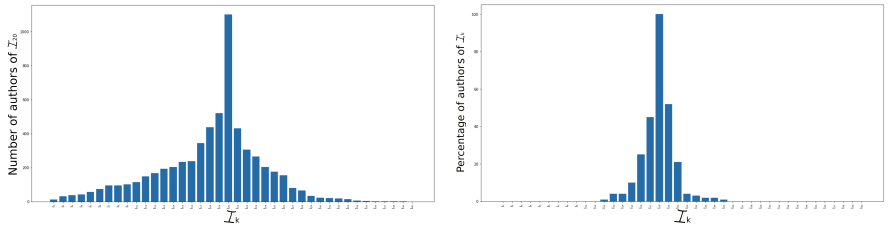
Fig. 2.21: (a) Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ in the null model - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$ in the null model

However, this is not sufficient to conclude that there is a degree assortativity for authors in Reddit. In fact, we must check if this trend is also confirmed for the authors with an intermediate degree centrality and for those with a low degree centrality.

Clearly, for an exhaustive analysis, we should repeat the tasks we have previously done for $\mathcal{I}_1$ for all intervals. Due to space constraints, we limit our analysis to the interval $\mathcal{I}_{20}$, representative of intermediate degree centrality intervals, and $\mathcal{I}_{39}$, representative of the low degree centrality intervals[6].

Figure 2.22(a) reports the number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$, whereas Figure 2.22(b) shows the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{20}$. From the analysis of this figure, it emerges a strict correlation between the authors with an intermediate degree centrality.

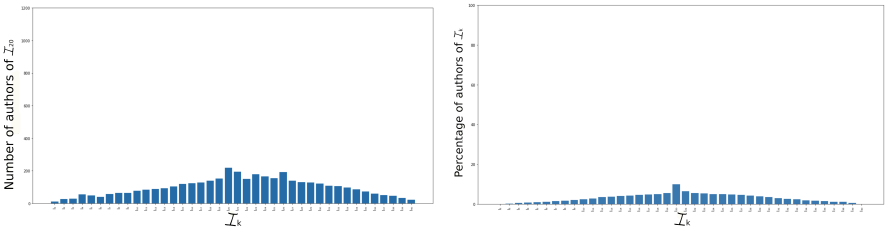
[6] We did not choose $\mathcal{I}_{40}$ because the number of its authors is less than the ones of the other intervals.



(a) **(b)**

Fig. 2.22: (a) Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$

Also in this case, we compared these findings with the ones obtained in the null model. These last ones are reported in Figure 2.23. Looking at these results and the ones represented in Figure 2.22, we can conclude that, again, the behavior observed in these last figures is not random but it is a property of Reddit.

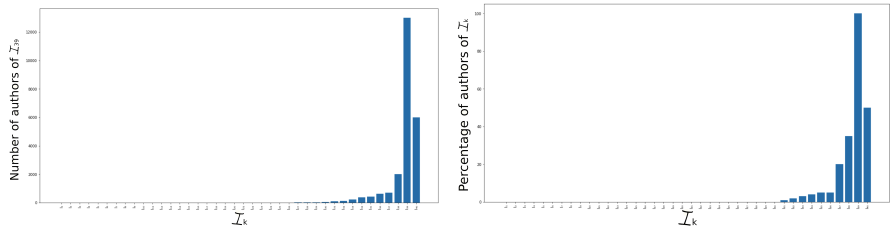


(a) **(b)**

Fig. 2.23: (a) Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ in the null model - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$ in the null model

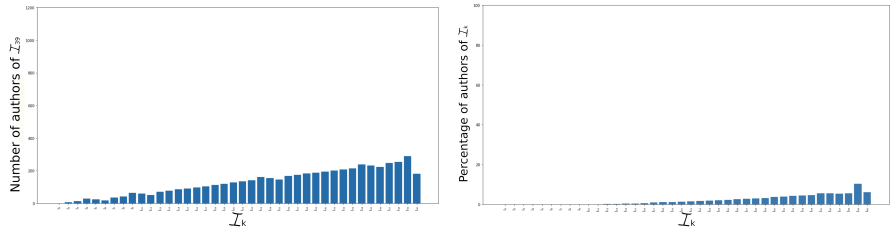
Finally, Figure 2.24(a) reports the number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$, whereas Figure 2.24(b) shows the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{39}$. Again, there is a strict correlation between authors with a low degree centrality. Also for this last case, we compared the results obtained with the ones returned using the null model. We report these last ones in Figure 2.25. The comparison of these figures confirms that the behavior observed in them is a property intrinsic to Reddit.

Having verified that there exists a sort of backbone among the authors with a high (resp., intermediate, low) degree centrality, we can conclude that actually Reddit is assortative with respect to degree centrality, as far as the co-posting relationship is concerned.



(a) **(b)**

Fig. 2.24: (a) Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$



(a) **(b)**

Fig. 2.25: (a) Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ in the null model - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$ in the null model

This important result can be explained considering the concept of karma and the posting rules in Reddit. Indeed, in this platform, each user has associated a karma, which is a score taking her past "reputation" into account. Generally, users with high karma are very active and, often, submit a lot of appreciated posts. As a consequence, it is presumable that they have a high degree centrality. In other words, a direct correlation between karma and degree centrality can be recognized for authors. Now, the posting rules of Reddit state that each subreddit has associated a minimum threshold of karma [470, 491, 57] so that only the authors with a karma higher than this threshold can submit a post on it. This threshold is dynamic and changes over time. Clearly, when it is low, all the authors can submit their posts on the subreddit. When it grows, the authors with a low karma (and, presumably, with a low degree centrality) cannot submit posts on it. Finally, when it becomes high, only the authors with a high karma (and, presumably, a high degree centrality) can submit posts on it. This way of proceeding tends to segment users into groups having homogeneous degree centralities.

Having verified the assortativity of Reddit with respect to degree centrality, it is natural to wonder whether this property depends on the type of centrality or is intrinsic in this social platform. As a premise to this investigation, it is worth un-

derlying that each form of assortativity is a unique history *per se*. Therefore, it is impossible to define a general rule. Nevertheless, it is possible to verify if a trend exists, and we have operated in this direction.

To this end, we have chosen a second form of centrality (i.e., the eigenvector centrality) and we have repeated for it all the steps previously seen for degree centrality. The results obtained are shown in Figures 2.26 - 2.28

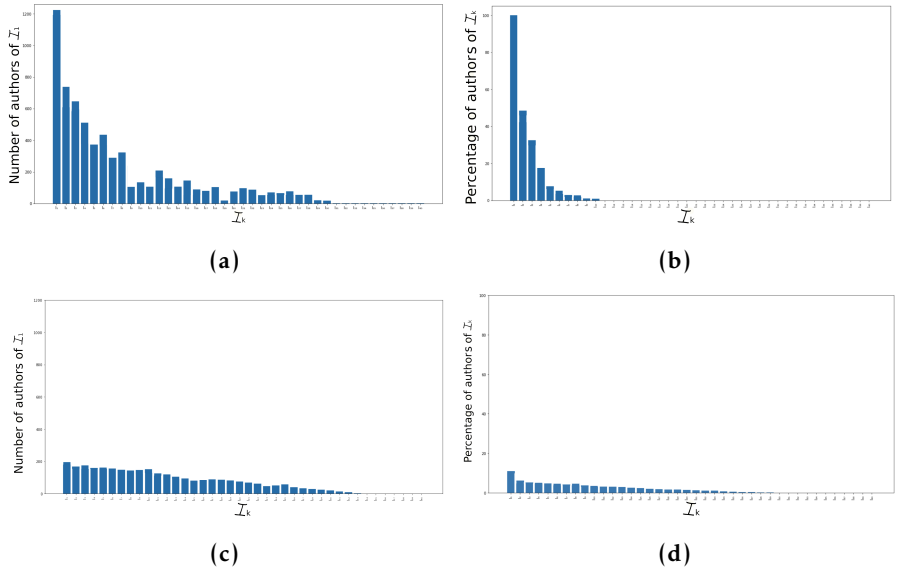


Fig. 2.26: (a) Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$ - (c) Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ in the null model - (d) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$ in the null model

They confirm that there is an assortativity among the authors of Reddit also with respect to the eigenvector centrality. As a consequence, we can conclude that the assortativity of Reddit authors is not limited to degree centrality but represents a trend characterizing this social platform beyond the form of centrality taken into consideration.

2.1.4.2 Correlation between subreddits and author stereotypes

First of all, we observe that, although in principle subreddit stereotypes and author stereotypes are two orthogonal concepts, in practice there are strong correlations between them. In fact, certain subreddit stereotypes are the ideal and perfectly tailored places for certain user stereotypes, and vice versa.

Let us now examine these correlations more closely. In the following of this section, for more clarity and to avoid heavy speech, we use the Successful Subreddit

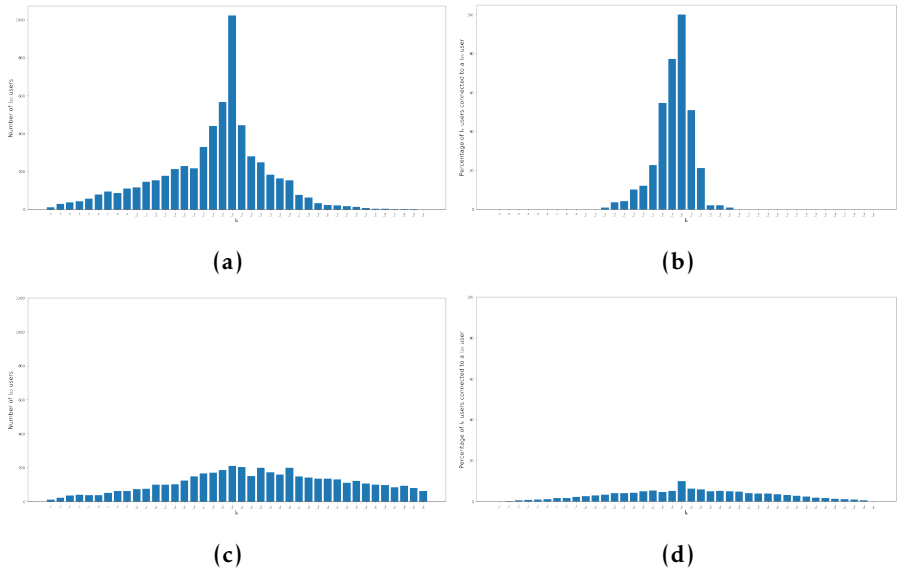


Fig. 2.27: (a) Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$ - (c) Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ in the null model - (d) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$ in the null model

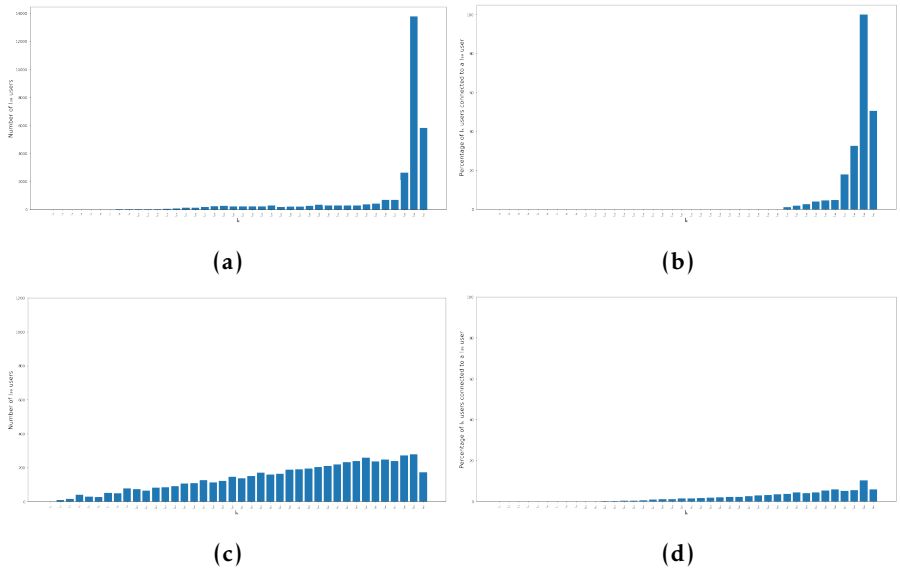


Fig. 2.28: (a) Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ - (b) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$ - (c) Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ in the null model - (d) Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$ in the null model

notation to indicate the name of a subreddit stereotype, whereas we adopt the *Successful Author* notation to denote an author stereotype.

User Profile is a fairly generic subreddit stereotype and can be related, at least partially, to various author stereotypes. Surely, a *Fame Seeker* can create a User Profile subreddit to advertise her profile. A similar argument probably applies to a *Content Creator* and a *Successful Author*.

Unsuccessful Subreddit could be at least partially related to *Unsuccessful Author* because if a subreddit was not successful then its posts did not attract Reddit users. Clearly, the authors of those posts, if this fact happens several times, would tend to become unsuccessful authors.

Clearly, there are very strong and direct correlations between Comment Grabber and the homonymous author stereotype, between Big Comment Grabber and *Big Comment Grabber*, between Private Community and *PP Author*, between Bot and the homonymous author stereotype, and between Cringe / NSFW Subreddit and *Cringe / NSFW Author*.

There is at least a partial relationship between Banned Subreddit and *Spammer* and *Hatred Sower*, because it is very likely that subreddits with many authors of those two categories are banned. Similarly, there is a correlation between Successful Subreddit and *Successful Author*; in fact, it is likely that if many successful authors write in a subreddit, then that subreddit will be successful.

A less obvious, but extremely interesting correlation exists between Niche Subreddit and *FBG Publisher*.

Again, Unsuccessful Boomer may be related to *Fame Seeker*, *Cringe / NSFW Author*, *Hatred Sower* or *Investigator*. In all these cases, the authors of these subreddits may have initially succeeded in stimulating the attention of other Reddit users but, after a while, this attention was lost.

Finally, there is a quite evident correlation between Unsuccessful Zombie and *Unsuccessful Author*, in the sense that if an author activates subreddits that become Unsuccessful Zombie, in the long run she risks to become an *Unsuccessful Author*. Finally, Unsuccessful Zombie could have a slightly subtler and hidden correlation with *Moody Author* because, if in a subreddit many posts of moody authors are published, it is likely that this subreddit will not attract people and eventually will become an Unsuccessful Zombie.

2.1.4.3 Considerations about author stereotypes and assortativity

After having examined the correlation between subreddit stereotypes and author stereotypes, we continue our discussion by examining the correlations between the results obtained for author stereotypes and those concerning assortativity. In Section 2.1.4.1, we found that there is a degree (resp., eigenvector) assortativity between Reddit authors. This implies that authors with similar degree (resp., eigenvector)

centrality tend to form a backbone. Keeping in mind the definition and properties of these two forms of centrality, it is possible to make some interesting deductions.

The first one is that *Fame Seekers*, who generally have a high degree centrality, tend to form a backbone and, therefore, to support each other. An analogous reasoning can be imagined for *Successful Authors* and *Reporters*, who are also characterized by a very high degree centrality. Continuing in this direction, even many authors characterized by negative stereotypes tend to support each other; in particular, this happens for *Spammers*, *Hatred Sowers* and *Investigators*. In these cases, a post published by one of them tends to provoke the reaction of the others, giving rise to very long discussions that often involve a huge number of people. A similar situation, even if with a neutral and not negative connotation, can concern the *Big Comment Grabbers*. Even these authors tend to form communities in which large discussions take place; however, unlike the previous cases, these discussions are not necessarily harmful.

As far as eigenvector centrality is concerned, in addition to all the communities mentioned above, the presence of backbones between *FBG Publishers* or *Content Creators* appears possible. In fact, these authors, who tend to use Reddit as a utility tool, may be strongly attracted by subreddits created by authors with the same intentions and, therefore, may tend to form communities. It is interesting to highlight that these types of figures (a sort of "grey cardinals") are the classical ones having a high eigenvector centrality and, as far as we are concerned, a high eigenvector assortativity.

A final discussion concerns the results on assortativity described in this chapter and the ones on assortativity in social networks described in the past literature. As previously pointed out, Newman's seminal work showed that social networks are generally assortative, unlike other types of networks, such as technological and biological ones, which are disassortative [502].

Next, the authors of [26] demonstrated that: *(i)* Cyworld is slightly disassortative with respect to degree centrality on a network built taking users and their friendships into account, while it is strongly assortative with respect to degree centrality on a network built considering users and the "testimonial" relationships (a kind of relationship specific of this social network) existing between them; *(ii)* Orkut is assortative with respect to degree centrality on a network built starting from users and their friendships; *(iii)* MySpace is neutral (that is neither assortative nor disassortative) with respect to degree centrality on a network that takes users and their friendships into account.

The authors of [137] showed that Twitter is strongly assortative with respect to degree centrality on a network that takes the sharing of interest among users into

account. Furthermore, the authors of [140] studied assortativity in Facebook and showed that such a social network is assortative with respect to the tendency of a bridge (i.e., a user joining more social networks) to communicate with other bridges.

Finally, in [317], the authors considered Reddit and investigated the concept of assortativity but for a very particular aspect, i.e., loyal communities. In particular, they showed that loyal communities are not assortative with respect to the activity level of the users belonging to them, while assortativity exists in the case of non-loyal communities. The lack of assortativity in loyal communities implies that users belonging to them are willing to communicate with all the other users of the same community, regardless the corresponding activity level. By contrast, the presence of assortativity in non-loyal communities implies that the corresponding users tend to partition themselves into subgroups based on their activity level. Indeed, a user with a certain activity level tend to communicate only with users having similar activity levels.

As said before, we want to provide a contribution in the study of assortativity in social networks. First, besides degree centrality, it also considers eigenvector centrality. Furthermore, it focuses on the study of assortativity in Reddit, a social platform that was not analyzed in the past as far as this feature is concerned, except for the investigations described in [317]. However, in this last paper, the main topic of the author investigation was not assortativity but loyalty, while assortativity simply served as a feature to assess whether loyal and non-loyal communities could be partitioned into smaller groups. Therefore, compared to the general studies on assortativity presented in [26, 137, 140], the analysis of [317] can be considered of niche. As a proof of this, we can observe that, contrary to all studies on assortativity proposed in the past, in [317] the presence of assortativity among the nodes of a network is seen as a negative factor (leading highly active users to disregard little active and new ones), rather than a positive feature.

Compared to [317], our approach aims at bringing the study of assortativity into Reddit in the general mainstream of the study of assortativity in social networks, analyzing this feature by itself, independently from other features, such as loyalty. As a matter of fact, the results we found are in line, and even strengthen, the trends on assortativity in social networks hypothesized by Newman and next found by most of the other authors.

2.1.4.4 Applications of stereotypes

This section presents two possible applications of the stereotypes previously investigated. The first regards the usage of subreddit stereotypes to make a subreddit

successful. The second concerns the exploitation of particular types of author stereotypes to improve the content quality of subreddits.

Application of subreddit stereotypes

In Section 2.1.3.2, we defined several subreddit stereotypes belonging to three macro-categories, namely "dead in crib", "survivors" and "undelivered promises". A first application of this research can be the definition of some guidelines to follow in order to make a subreddit successful. Indeed, knowing how a subreddit became successful (resp., unsuccessful) can lead to the characterization of "positive" (resp., "negative") actions that can influence the "lifespan" of a new subreddit. For instance, consider the subreddit /r/meme. It started during 2008 and, at the time of writing, has about 806,000 users. Certainly, it represents an example of a successful subreddit. Here, the authors post high quality and engaging contents. This kind of behavior could be registered as a "best practice" in the guidelines. On the other hand, a subreddit containing only few contents from few authors is an example of an unsuccessful subreddit. This failure could be caused by a lack of engaging contents posted in it. Clearly, what said above provides just an idea of what these guidelines could contain.

Another possible application of subreddit stereotypes could regard the definition and realization of recommender systems for Reddit. These systems would aim at recommending to a user subreddits with the same stereotype (or the same content) as the ones characterizing the subreddits accessed by her in the past. In any case, the recommender system should avoid "dead in crib" subreddits or, more generally, unsuccessful ones. On the other hand, the same system should suggest to a user successful subreddits, subreddits currently expanding their community and/or subreddits characterized by contents in line with her profile.

A further example of possible usage of subreddit stereotypes could be the definition of an algorithm that finds subreddits to merge or, at least, to integrate. For instance, consider two zombie subreddits with related topics, where authors are posting contents that were not able to attract other users. These two subreddits are surviving, but their interactions with users are so low that they can actually be considered dead. If they would be merged or integrated into a unique subreddit, they could have more chances of becoming successful. Joining together two, or even more, subreddits having the same (or related) topics/characteristics brings more visibility and more contents to them. These contents would be, otherwise, dispersed in different unsuccessful subreddits. Even if the new integrated subreddit is made up of past zombies, it could become so successful to attract authors and co-posters from other communities.

Application of author stereotypes

In Section 2.1.3.3, we defined some possible author stereotypes. Some of them are strictly related to the homonymous or corresponding subreddit stereotypes. Other ones, instead, are intrinsic to human behavior and, in particular, to the concept of author. For example, consider "Fame Seekers" and "Content Creators". These users could represent the target of a proposal of an advertising campaign aiming at promoting them. Take, for instance, a painter or a digital artist, who has been classified as "Fame Seeker". An advertising company can easily persuade her to give it an engagement to promote her image.

Another possible usage of author stereotypes is the definition and implementation of different categories of recommender systems. A first category could help bootstrapping a subreddit. Consider, for instance, a newborn subreddit where authors post comics strips created by them. Knowing successful authors of comics strips and being able to convince them to become "Content Creators" in the new subreddit could help this last one to get visibility. Complementary to this case, a second category of recommender systems could be used for talent scouting. In this case, a "Fame Seeker", who is also a creator of comics strips, could be recommended to successful subreddits if her contents are high-quality ones.

The last application we present in this overview is the definition of an algorithm that builds blacklists of users based on author stereotypes. As an example, we can define a "dangerousness level" of an author for one subreddit, a set of subreddits or all subreddits. For instance, in such a scenario, "Hatred Sowers" can be automatically banned from subreddits attended by sensitive people. This way of proceeding could certainly maintain the discussion in these subreddits clean, thus avoiding their visitors being harassed by fake news and cyberbullying.

2.2 Investigating Not Safe For Work posts

2.2.1 Introduction

Reddit[7] is currently one of the most active social media. It has been extensively studied by researchers in the past [469, 611]. Many papers have focused on specific aspects of this social network, concerning, for example, community structures and interactions [636, 218, 265], user behavior [143, 393, 424], structure and content of subreddits, posts and comments [603], structural properties [265, 324, 723], text classification [404], user migration [501], political and ideological aspects [308, 687].

One aspect of Reddit worth to be analyzed involves NSFW (Not Safe For Work) posts. This term refers to user-submitted content not suitable to be viewed in public or in professional contexts. The phenomenon of NSFW posts in Reddit has been very little investigated, although it is very common in this social medium. In fact, only a very small number of authors have analyzed it [464, 496]. The term "NSFW" has been proposed since 1998, and is one of the oldest acronyms of the Internet. Since its first appearance, many social media, such as Twitter, WhatsApp and Reddit, have adopted it to indicate certain sections or contents. In addition, several authors have focused on the analysis of this phenomenon in other social networks. The study about the role of images and selfies in NSFW content of `tumblr.com`, presented in [641], and the analysis of the anonymity level of NSFW content in both Twitter and Whisper, described in [209] are two examples.

In this chapter, we give a contribution in this setting investigating the phenomenon of NSFW posts in Reddit and describing the whole context (authors, subreddits and readers) behind it. For this purpose, we consider a dataset that includes all the posts published in Reddit from January 1^{st}, 2019 to December 31^{st}, 2019.

During our investigation, we carried out three types of analysis, namely:

- *Descriptive Analysis*, to study the distributions of the entities involved in the phenomenon (e.g., the distribution of NSFW posts against subreddits, authors, score and comments).
- *Social Network Analysis*, to study the co-posting phenomenon, and therefore the interactions between authors of NSFW posts.
- *Assortativity Analysis*, to extend and deepen the previous analyses to discover and study whether possible forms of assortativity [502] exist among the authors of NSFW posts. Recall that assortativity is a particular case of homophily in social networks [468], which indicates the tendency of a node to cooperate with nodes having similar characteristics.

[7] `https://www.reddit.com`

These analyses allowed us to extract three findings regarding NSFW posts, NSFW authors and NSFW subreddits, respectively. Throughout our analysis, in most of the cases, we compare each finding on NSFW posts with the corresponding one on SFW (Safe For Work) posts. Some of the questions these findings provide an answer to are the following:

- What can be said about the spread of NSFW posts in the subreddits?
- What can be said about the quantity of posts an NSFW author usually submits?
- What can be said about the score of NSFW posts?
- What can be said about the number and the score of comments to NSFW posts?
- What can be said about the level of interconnection between authors of NSFW posts?
- Is there a backbone among experienced authors of NSFW posts? In other words, do they tend to interact only with their peers (i.e., authors with the same level of experience), or are they open to collaborations with new authors who have just started publishing NSFW posts?

Finally, we suitably combine the knowledge represented by the three findings in order to describe the dynamics behind the phenomenon of NSFW posts in Reddit.

The rest of this chapter is organized as follows: in Section 2.2.2, we present related literature. In Section 2.2.3, we describe the adopted dataset and investigate the data distributions involving NSFW posts and the comments on these NSFW posts. Then, in Section 2.2.4, we describe the co-posting activity of the authors of NSFW posts, evaluate the assortativity of these authors, and, finally, we summarize our contributions in order to define an overall picture of this phenomenon.

2.2.2 Related literature

The term "NSFW" was first proposed in 1998 and it is one of the oldest acronyms of the Internet. It refers to content that is not suitable to be viewed in a working environment. Since then, different online systems, like Twitter, WhatsApp, many forums, and Reddit, have adopted this term to label sections with posted content not adequate for everybody and, in general, not suitable for public and professional contexts. Specifically, Reddit has introduced a dedicated group of contents called NSFW to separate posts suitable to be enjoyed in any context from those that should be watched in private environments.

Even if the contents of NSFW posts are considered side-contents to be kept separated from front-end contents, several researchers have started to study the characteristics of these contents, as well as the communities underneath them [464, 153, 641, 722, 209, 299].

From a high-level analysis of the research efforts in the context of NSFW content, we may distinguish two main directions. The former focuses on understanding the main characteristics of people publishing or viewing such materials, as well as the features of the NSFW content itself. The latter, instead, uses features of NSFW content to build content detection and filtering solutions, often with the objective of enabling/disabling the visualization of this material for users.

In particular, the work described in [641] is an example of the first research direction. Here, the author investigates the role of images and selfies in NSFW contents of `tumblr.com`. NSFW contents, having the explicit NSFW label assigned by their authors, are extracted from Tumblr blogs. Then, the described analysis focuses on images and reactions (interactions) surrounding them. The aim of this study is understanding the different roles that people assign to images and selfies, leading to the creation (or breaking) of trust relationships between users. Furthermore, the author provides evidence that different opinions about the membership of an image to the NSFW category may lead to violations of assumed trust between two individuals, thus causing the dissolution of a community.

Another contribution in the first research direction is the one reported in [209]. In this paper, the authors try to understand both the nature of the content posted in anonymous social media and the difference between NSFW content posted in these media and in non-anonymous ones (like, e.g., Twitter). To do this, they define an anonymity sensitivity metrics measuring how much users think that a post should be anonymous. Then, they use this metrics, in conjunction with a human annotator, to identify NSFW posts with the same level of anonymity sensitivity in Whisper (an anonymous media system) and Twitter. Hence, they carry out a deep comparative analysis of the two sets of posts and find that, actually, there is a strong difference between them, especially when it comes to the shades or levels of anonymity and their linguistic features.

Even if its main focus is slightly different from the one defining this first research direction, the work described in [464] gives a mentionable contribution in this setting. Indeed, the author considers a particular protest carried out by moderators of Reddit in 2015, when participants disabled their subreddits to block posting activities. In this context, the author studies the different behavior of NSFW and SFW subreddit moderators. The results show that, even if several confounding factors could be considered to understand the underlying dynamics, NSFW subreddit moderators were more inclined to join the protest and block posting activities.

In the second research direction mentioned above, several works have been published in recent scientific literature [496, 224, 108, 722, 201]. For instance, the work described in [496] focuses on the protection of minors accessing the Internet from

the exposure to unwanted and harmful contents. The proposed system can be seen as both an active content filtering solution, which protects the access of minor users to NSFW content, and a watchdog constantly monitoring and moderating websites to avoid the diffusion of unwanted content.

The problem of classifying video content as NSFW is faced in [224]. In this paper, the authors exploit Convolutional Neural Networks (ConvNets) for extracting audio-video patterns from NSFW videos. Specifically, they first extract separated audio and video features and then merge the two feature sets to obtain a single feature vector. After that, they provide this vector in input to some baseline classifiers. Even if the approach is naive, the achieved results outperform those of other methods, thus proving the adequacy of this proposal.

Similarly, the approach of [108] makes use of a deep neural network-based solution to identify content belonging to the NSFW category. This approach is based on a residual network, which returns a value specifying the probability that a given content belongs to NSFW category. Moreover, it allows the computation of the degree of explicitness of the analyzed content, which can be used to feed a filtering system. Finally, it is capable of labeling media content with tampered extension to warn users about the potential risk of suspiciously unwanted material. The experiments show very interesting performance for this approach, which reaches an accuracy of about 96% also on image and video contents.

Still in this context, also the approach described in [722] makes use of a fast Convolutional Neural Network (CNN) for the detection of both NSFW and SFW images. Specifically, this proposal deals with the design of a neural network-based solution to detect pictures with nudity in NSFW contents. After that, it defines picture filtering strategies for online media services.

Finally, the approach described in [201] strives to build a classifier for detecting NSFW content by looking at images and visual material in the post. The proposed solution uses a weighted sum of the results of multiple deep neural network models. The weighted combination is obtained by learning a linear regression model through Ordinary Least Squares. The authors prove that their solution outperforms the state-of-the-art solutions based on single CNN models. For this purpose, they present a deep comparison on a manual labeled dataset.

Our approach is somehow near to the studies belonging to the first research direction introduced above. However, these approaches only study the content of NSFW posts and none of them focus on the structural network-based properties of NSFW and SFW posts and authors. Instead, we want to study such differences between the two categories with a comparative approach and typical Social Network Analysis methodologies.

The identified findings can be fundamental to improve existing techniques for content detection, parental control or content filtering solutions, such as the ones mentioned above. To the best of our knowledge, no similar studies have been conducted in social media platforms. Our work aims at providing a first contribution in this setting using Reddit as reference social network. However, as we will see below, our investigation strategy is general and can be specialized to other social media [158].

2.2.3 Methods

2.2.3.1 Dataset description

The dataset used for our analysis has been downloaded from the website `pushshift.io` [89], one of the main Reddit data sources. In particular, we extracted all the posts published on Reddit from January 1^{st}, 2019 to September 1^{st}, 2019[8]. The number of posts available for our analysis was 150,795,895. In Reddit, an NSFW post must be marked as such by its author. Therefore, there is no need for automatic labeling by Reddit or manual labeling by third-parties. If the user specifies that a post she/he is publishing is NSFW, Reddit puts a red label when displaying it and sets the value of the `over_18` field in its database to `true`. We used the value of this field to separate NSFW posts from SFW ones in our analyses.

We performed a preliminary ETL (Extraction, Transformation and Loading) activity on our dataset. In Data Analytics, this activity is typically carried out prior to any data analysis campaign. It aims at cleaning the data in the dataset, removing any errors and inconsistencies, integrating any data from different sources, and transforming the cleaned and integrated data into a single format chosen for the next data analysis tasks [514].

During the ETL phase, we observed that some of the available posts were made by authors who had left Reddit. We decided to remove these posts from our dataset. At the end of this activity, the number of available posts was 122,568,630. NSFW posts were 11,908,377, equivalent to 9.72% of them.

As pointed out in the Introduction, the goal of our study is to understand the characteristics of NSFW posts and their authors, comparing them with the SFW posts and their authors. For this reason, we decided to extract from the dataset described above two sub-datasets, with the same number of posts each. Both of them are limited to January and February 2019. The first dataset $\mathcal{D}$ contains only SFW posts, while the second, called $\overline{\mathcal{D}}$, stores only NSFW posts. We randomly selected

[8] Actually, only for stability analysis, we considered all the posts from January 1^{st}, 2019 to December 31^{st}, 2019 (see Section 2.2.3.4).

1,250,000 posts for each of them to reduce the datasets' size and the computation time. It should be noted that this number is absolutely in line with the number of posts generally used in the analyses of Reddit [679, 501, 616, 308]. However, we repeated all the analyses on two other datasets $\mathcal{D}'$ and $\overline{\mathcal{D}}'$ to verify the stability of our results. The set $\mathcal{D}'$ (resp., $\overline{\mathcal{D}}'$) consists of 1,250,000 SFW (resp., NSFW) posts published in March and April 2019, randomly selected from the original dataset. In addition, we carried out a deeper stability check evaluating all posts of 2019 month by month.

As a preliminary analysis, we focused on the "context" of SFW and NSFW posts. Here, we use the term "context" of a post to denote its author, its comments and the subreddits in which it was published. In this analysis, we wanted to verify if the context of SFW posts and the one of NSFW posts are the same or not. To answer this question, we calculated the values of some parameters on $\mathcal{D}$ and $\overline{\mathcal{D}}$ and, then, on $\mathcal{D}'$ and $\overline{\mathcal{D}}'$. The results obtained are shown in Table 2.12.

Parameter	$\mathcal{D}$ *and* $\overline{\mathcal{D}}$	$\mathcal{D}'$ *and* $\overline{\mathcal{D}}'$
Number of authors who published at least one SFW post	59,465	58,561
Number of authors who published only SFW posts	58,801	57,891
Percentage of authors publishing SFW posts who published only posts of this type	98.88%	98.52%
Number of authors who published at least one NSFW post	36,758	36,461
Number of authors who published only NSFW posts	36,094	36,131
Percentage of authors publishing NSFW posts who published only posts of this type	98,19%	99.09%
Number of subreddits containing at least one SFW post	89,360	92,445
Number of subreddits containing only SFW posts	82,050	85,157
Percentage of subreddits containing SFW posts that contain only posts of this type	91.82%	92.12%
Number of subreddits containing at least one NSFW post	41,365	45,910
Number of subreddits containing only NSFW posts	34,055	38,622
Percentage of subreddits containing NSFW posts that contain only posts of this type	82.33%	84.13%

Table 2.12: Parameters about the authors and the subreddits of SFW and NSFW posts - $\mathcal{D}$ (resp., $\overline{\mathcal{D}}$) stores SFW (resp., NSFW) posts of January and February 2019, while $\mathcal{D}'$ (resp., $\overline{\mathcal{D}}'$) stores the same kind of post but for March and April 2019

This table shows that the reference contexts for SFW and NSFW posts are basically independent. In fact, more than 98% of authors writing SFW posts do not write NSFW posts, and vice versa. In addition, more than 91% of subreddits containing SFW posts do not contain NSFW posts, and more than 82% of subreddits containing NSFW posts do not contain SFW posts. Another important result is that all the computations are stable over time because the values obtained for January and February 2019 (Jan-Feb, for short) are very similar to the ones returned for March and April 2019 (Mar-Apr, for short).

2.2.3.2 Investigating the NSFW posts

In this section, we present some analyses directly involving NSFW and SFW posts. In particular, we study the distribution of subreddits and authors against posts and the distribution of posts against the scores assigned to them by Reddit users.

Firstly, we computed the distributions of the subreddits against NSFW and SFW posts for the datasets $\mathcal{D}$ and $\overline{\mathcal{D}}$. The results obtained are reported in Figure 2.29.

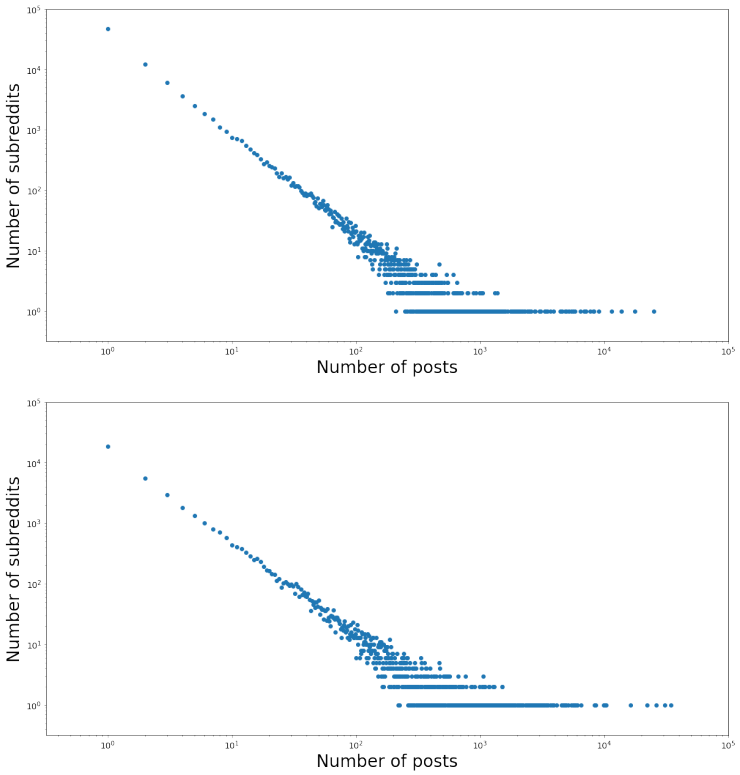


Fig. 2.29: Log-log plots of the distributions of subreddits against SFW posts (on top) and NSFW posts (on bottom) - Datasets regarding January and February 2019

This figure shows that the two distributions follow a power law. We also computed some parameters for the two power law distributions; they are shown in the second and third columns of Table 2.13. To verify the stability of results found, we made the same computations on $\mathcal{D}'$ and $\overline{\mathcal{D}}'$ datasets. They are shown in the fourth and fifth columns of Table 2.13.

From this table, we can observe that the maximum and the average numbers of subreddits for SFW posts is more than twice the value obtained for NSFW posts. The maximum and the average numbers of NSFW posts in a subreddit are slightly higher than SFW posts. There are no significant differences in the α and δ parameters of the two power law distributions. Indeed, both of them are very steep. The comparison of the second and the third columns of Tables 2.13, on the one hand, and the fourth

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of subreddits	47,480 (53.13%)	18,332 (44.31%)	49,502 (53.24%)	21,034 (45.02%)
Number of subreddits of the 99 percentile	1,095	571	1,101	569
Maximum number of posts	25,006 (4.62%)	34,424 (4.57%)	26,650 (4.98%)	31,329 (4.76%)
Number of posts of the 99 percentile	7,719	9,862	7,721	9,859
Average number of subreddits	126	54	137	57
Average number of posts	767	981	768	905
α (power law parameter)	1.6539	1.6974	1.6767	1.6859
δ (power law parameter)	0.0266	0.0364	0.0306	0.0432

Table 2.13: Parameters of the distributions of subreddits against posts

and fifth columns of the same table, on the other hand, also tells us that the trends obtained are stable over time, because their variations between Jan-Feb and Mar-Apr are not significant.

Although the two curves show almost identical trends, as confirmed by the similar values of α and δ, we found interesting the differences in the maximum and average values. In other words, the curve shapes are similar but the ranges of values are different. To confirm these results we compared the two distributions through the *Wilcoxon rank sum test* [682].

This test indicated that the number of subreddits in which Jan-Feb SFW posts were published was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 2.8 \cdot 10^{-4}, p < 0.01$).

This result can be explained taking into account the intrinsic nature of NSFW posts, whose content is certainly less suitable for the general public than the one of SFW posts.

Then, in Figure 2.30 we show the distributions of authors against SFW and NSFW posts for the datasets $\mathcal{D}$ and $\overline{\mathcal{D}}$. From the analysis of this figure we can see that both distributions follow a power law.

In Table 2.14, we report the main parameters of these two power law distributions for the datasets $\mathcal{D}$ and $\overline{\mathcal{D}}$, on one hand, and $\mathcal{D}'$ and $\overline{\mathcal{D}}'$, on the other hand.

A Wilcoxon rank sum test showed that the number of authors of Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 1.2 \cdot 10^{-4}, p < 0.01$).

This result can also be explained taking into account the topics of NSFW posts. Indeed, these are more specific than those involving SFW posts. Differently from SFW posts that can be written by anyone, the authors who generally publish NSFW posts are a small circle of people almost exclusively dedicated to this type of post. Consequently, while it is true that NSFW posts are much fewer than SFW posts, it

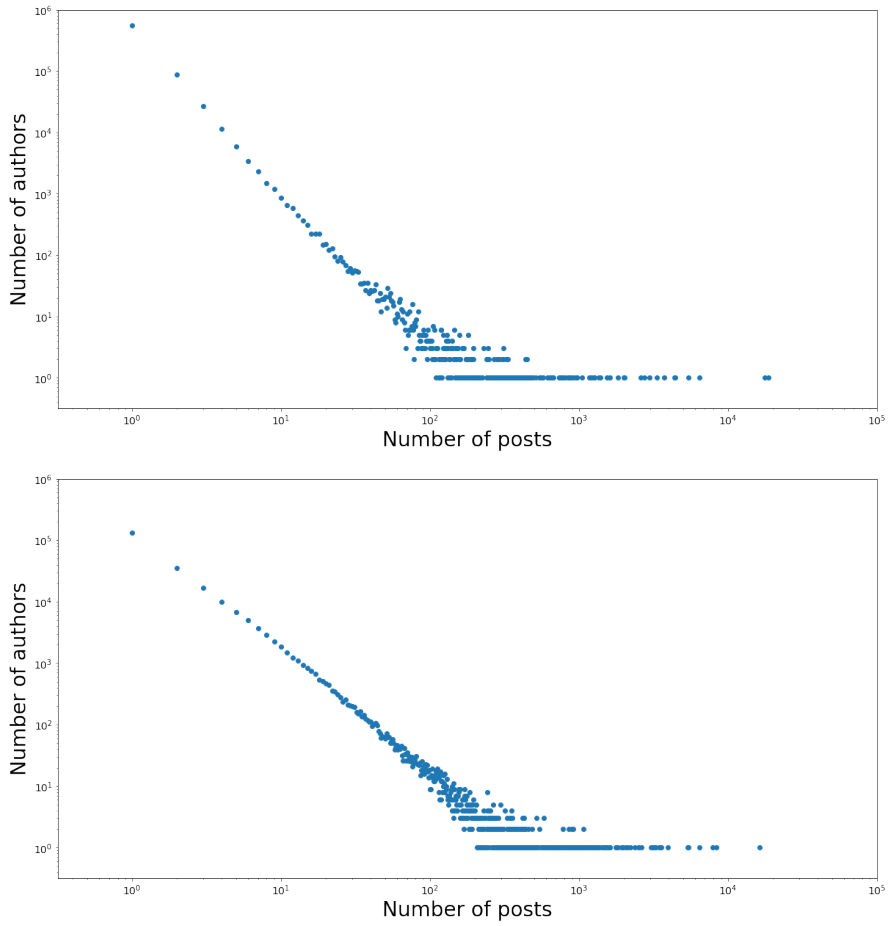


Fig. 2.30: Log-log plots of the distributions of authors against SFW posts (on top) and NSFW posts (on bottom) - Datasets regarding January and February 2019

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of authors	555,854 (79.06%)	131,070 (56.43%)	551,863 (78.97%)	133,594 (57.01%)
Number of authors of the 99 percentile	11,471	5,055	11,469	5,052
Maximum number of posts	18,724 (11.85%)	16,383 (5.70%)	16,513 (10.98%)	15,674 (5.48%)
Number of posts of the 99 percentile	5,426	5,393	5,424	5,393
Average number of authors	2,190	439	2,083	416
Average number of posts	491	543	491	521
α (power law parameter)	1.4631	1.5566	1.4505	1.5435
δ (power law parameter)	0.0473	0.0353	0.0304	0.0287

Table 2.14: Parameters of the distributions of authors against posts

is also true that they are published by an extremely limited number of authors. This explains the result.

Now, we want to evaluate the distribution of posts and their relative scores. A newly submitted post on Reddit has a score of 1. A user can upvote (resp., downvote) the post, increasing (resp., decreasing) its score by 1. We have computed the

distributions of SFW and NSFW posts against scores for the datasets $\mathcal{D}$ and $\overline{\mathcal{D}}$, and, then, for $\mathcal{D}'$ and $\overline{\mathcal{D}}'$, on the other hand. For the sake of simplicity, in Table 2.15, we report the main parameters of these distributions, which again follow a power law.

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum score	183,453 (57.98%)	106,947 (47.26%)	191,864 (61.87%)	112,830 (49.62%)
Number of score of the 99 percentile	4,746	3,645	4,825	3,275
Average score	9,881	4,191	8,809	3,819
α (power law parameter)	1.5998	1.5140	1.6061	1.5165
δ (power law parameter)	0.0197	0.0366	0.0154	0.0355

Table 2.15: Parameters of the distributions of posts against scores

A Wilcoxon rank sum test showed that the score of Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 0.00109, p < 0.01$).

Once again, this result can be explained by the type of contents that generally characterizes NSFW posts.

Finally, we computed the distributions of subreddits against the authors of SFW and NSFW posts. In both cases, we saw that they follow a power law similar to those shown in the previous figures. We report the values of the most important parameters in Table 2.16.

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of subreddits	62,839 (70.32%)	29,798 (72.03%)	65,861 (71.12%)	33,963 (72.01%)
Number of subreddits of the 99 percentile	932	538	930	533
Average number of subreddits	151	87	161	101
Maximum number of authors	20,285 (5.70%)	11,161 (4.70%)	21,801 (5.64%)	11,326 (4,59%)
Number of authors of the 99 percentile	6,435	4,627	6,431	4,635
Average number of authors	604	499	601	481
α (power law parameter)	1.7143	1.7992	1.6944	1.7343
δ (power law parameter)	0.0302	0.0.0382	0.0288	0.0362

Table 2.16: Parameters of the distributions of subreddits against authors

A Wilcoxon rank sum test showed that: *(i)* the number of subreddits of Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts; *(ii)* the number of authors of Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 6.3 \cdot 10^{-4}, p < 0.01$).

The explanation behind this result is essentially related to the fact that NSFW posts have particular contents that are of interest to a minority of people. Therefore, they are published in a limited number of subreddits.

In the next analyses, to save space, we will avoid highlighting those cases where the values α and δ of power law distributions are similar, as well as those cases where the parameter values are stable when switching from Jan-Feb to Mar-Apr. Only if one or both of these conditions are not valid in some analysis, we will explicitly highlight this situation.

2.2.3.3 Investigating the comments to NSFW posts

In this section, we analyze the comments to NSFW posts investigating their authors, the scores they get and the subreddits they are submitted to. Firstly, we present the distributions of comments against SFW posts and NSFW posts, which follow a power law. Table 2.17 shows the values of the main parameters of these distributions.

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of posts	499,068 (2.29%)	667,942 (5.79%)	522,477 (2.94%)	676,606 (5.81%)
Number of posts of the 99 percentile	8,257	10,707	8,362	10,719
Maximum number of comments	41,478 (39.93%)	28,227 (53.43%)	36,283 (40.01%)	23,485 (51.32%)
Number of comments of the 99 percentile	10,582	21,983	9,985	22,735
Average number of comments	1,237	771	1,402	656
α (power law parameter)	1.4836	1.3990	1.4779	1.4353
δ (power law parameter)	0.0178	0.0304	0.0160	0.0291

Table 2.17: Parameters of the distributions of comments against posts

A Wilcoxon rank sum test showed that the number of comments of Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 8.68 \cdot 10^{-5}, p < 0.01$).

As a further investigation on this topic, we considered both the top 150 most commented SFW and NSFW posts. As a first analysis, we observed that SFW (resp., NSFW) posts have been submitted by 141 (resp., 130) authors in 55 (resp., 77) different subreddits. This result highlights that there is no author or subreddit able to monopolize post comments. Indeed, the phenomenon is highly distributed.

Then, we computed the distributions of the number of these comments against subreddits. They are reported in Figure 2.31. Plots (a) and (b) of this figure show that the two distributions follow a power law. We computed the parameter values of these power laws and we obtained $\alpha = 3.41$ and $\delta = 0.075$ for SFW post comments, and $\alpha = 3.53$ and $\delta = 0.07$ for NSFW post comments. A Wilcoxon rank sum test indicated that the number of comments associated with the subreddits containing Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 0.16493, p < 0.01$).

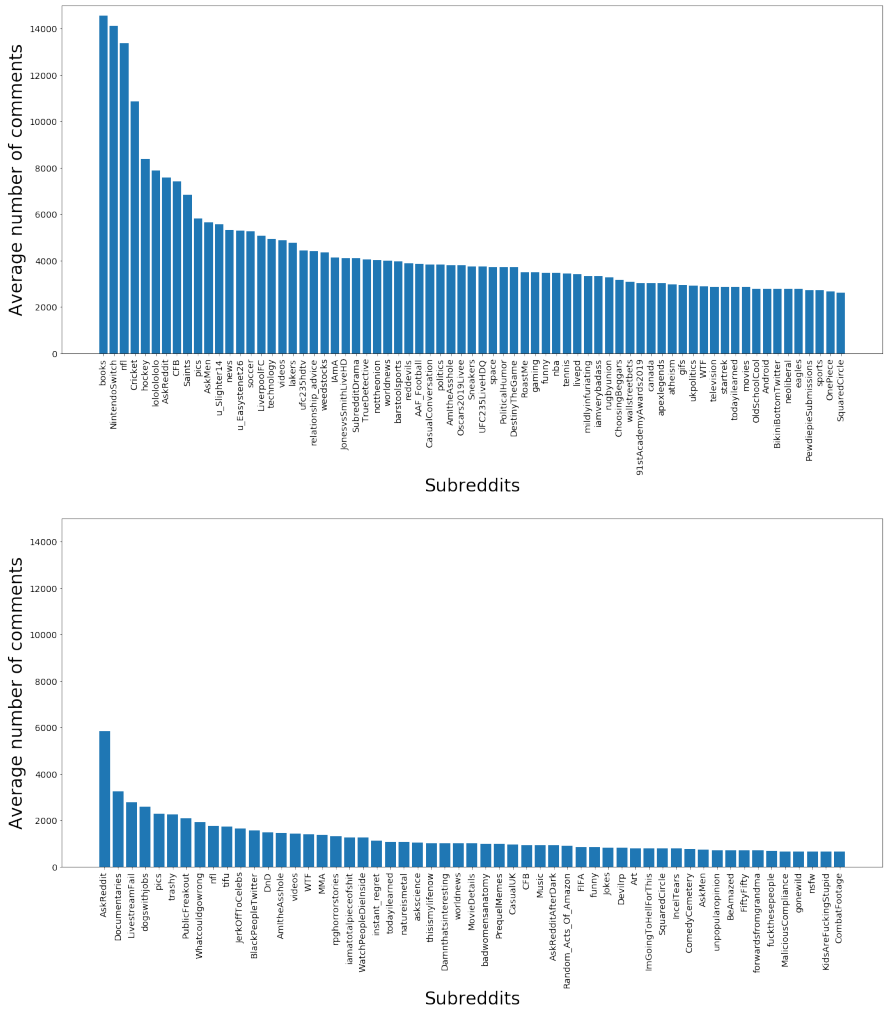


Fig. 2.31: Distributions of comments to the top 150 most commented SFW posts (on top) and NSFW posts (on bottom) against subreddits - Datasets regarding January and February 2019

Finally, we computed the distribution of the number of these comments against authors. Also in this case, we found that it follows a power law. The values of the corresponding parameters are $\alpha = 3.06$ and $\delta = 0.03$ for SFW post comments and $\alpha = 2.20$ and $\delta = 0.03$ for NSFW post comments. The conclusions about the trend and the values are analogous to the previous ones.

A Wilcoxon rank sum test indicated that the number of comments for Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 0.34951$, $p < 0.01$).

The motivations behind this result are the same as those related to the distribution of the subreddits against authors.

We then computed the distributions of subreddits against the comments to SFW and NSFW posts. In both cases we obtained that they follow a power law and show

trends similar to those shown in the previous figures. The main parameters of these distributions are reported in Table 2.18.

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Maximum number of comments	484,792 (5.45%)	301,040 (9.17%)	462,415 (5.41%)	244,912 (9.73%)
Number of comments of the 99 percentile	47,590	25,056	47,698	28,635
Average number of comments	3,942	2,607	3,800	2,391
α (power law parameter)	1.8025	1.7659	1.7981	1.7507
δ (power law parameter)	0.0236	0.0235	0.0217	0.0310

Table 2.18: Parameters of the distributions of subreddits against comments

A Wilcoxon rank sum test showed that the number of comments associated with the subreddits containing Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 6.34 \cdot 10^{-6}, p < 0.01$).

Once again, the motivations behind this result are the same as those related to the distribution of the subreddits against authors.

Moreover, we computed the distributions of comments to SFW and NSFW posts against scores. They are reported in Figures 2.32 and 2.33 for the datasets $\mathcal{D}$ and $\overline{\mathcal{D}}$. These figures show that the corresponding distributions do not follow a power law, and this is the first case. As we can see from figures, the distributions are irregular, even if both of them seem having a Gaussian trend.

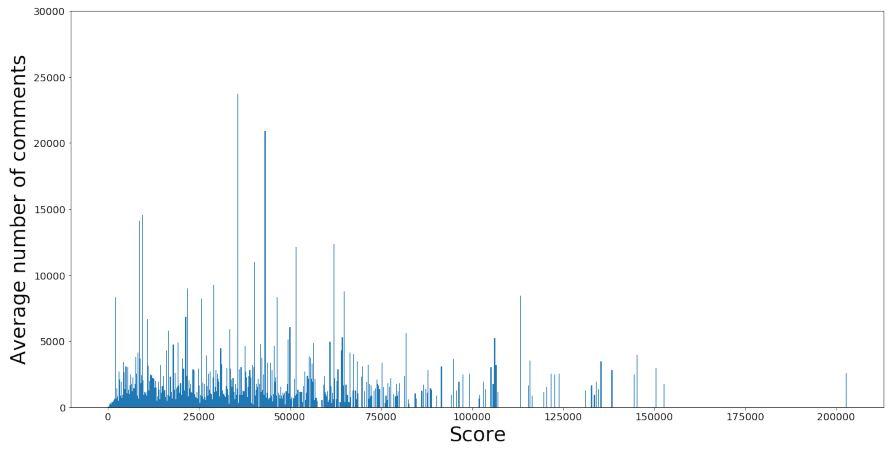


Fig. 2.32: Distribution of comments to SFW posts against scores - Datasets regarding January and February 2019

Also in this case, we computed some parameters for the two distributions. They are shown in Table 2.19.

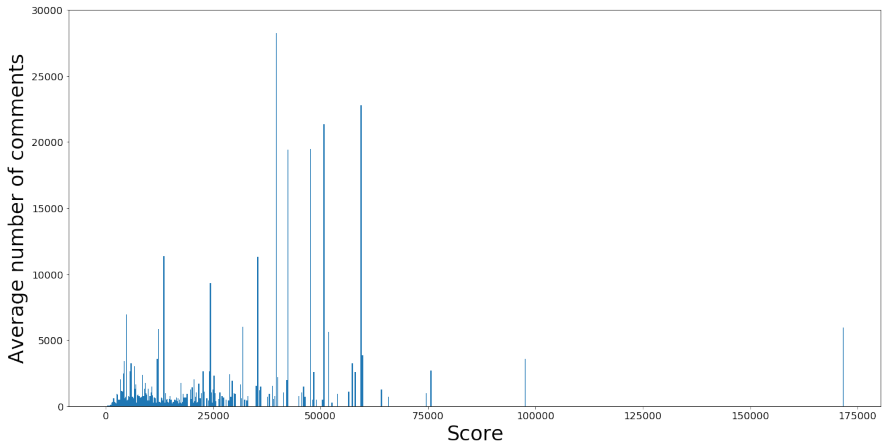


Fig. 2.33: Distribution of comments to NSFW posts against scores - Datasets regarding January and February 2019

Parameter	*SFW posts Jan-Feb*	*NSFW posts Jan-Feb*	*SFW posts Mar-Apr*	*NSFW posts Mar-Apr*
Average score	9,881	4,191	8,809	3,819
Score of the last comment of the first quartile	2,035	1,157	1,993	1,215
Score of the last comment of the second quartile	4,686	2,357	4,551	2,484
Score of the last comment of the third quartile	11,106	4,486	9,953	4,667
Score of the last comment of the fourth quartile	202,696	69,591	209,154	71,566

Table 2.19: Parameters of the distributions of comments to posts against scores

A Wilcoxon rank sum test indicated that the score of comments for Jan-Feb SFW posts was statistically significantly higher than the corresponding one of NSFW posts ($\tau = 5.88 \cdot 10^{-5}, p < 0.01$).

The motivations behind this result are the same as those related to the distribution of the posts against scores.

2.2.3.4 A deeper analysis of the stability of the investigations

All the distributions we have seen so far are based on a data sample recovered from January 1^{st}, 2019 to September 1^{st}, 2019. Due to computational complexity reasons, we could not process the whole sample at the same time and, therefore, we divided it into bi-months, i.e. Jan-Feb and Mar-Apr. In all the distributions we have presented so far, we could verify that the Jan-Feb and Mar-Apr data led to very similar results. This is a strong remark of the stability of the results of our investigations.

However, before continuing with the next analyses, which will have an even higher computational complexity, we decided to carry out a further stability check. To this end, we considered all the posts published in Reddit from January 1^{st}, 2019 to December 31^{st}, 2019, and split them months by months. Then, for each month,

we computed several parameters previously seen for the two bi-months. The results obtained are shown in Table 2.20 for SFW posts, and in Table 2.21 for NSFW posts. The analysis of these tables fully confirms that the results of our investigations are stable.

Parameter	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*
GENERAL CHARACTERISTICS						
Number of authors who published at least one SFW post	391,898	387,458	365,785	389,154	387,562	374,531
Number of authors who published only SFW posts	380,261	374,564	359,851	378,582	377,423	365,751
Percentage of authors publishing SFW posts who published only posts of this type	97.03%	96.67%	98.37%	97.28%	97.38%	97.65%
Number of subreddits containing at least one SFW post	58,843	57,965	58,786	57,653	58,426	57,953
Number of subreddits containing only SFW posts	54,189	53,482	53,952	54,236	54,873	52,432
Percentage of subreddits containing SFW posts that contain only posts of this type	92.09%	92.22%	91.77%	94.07%	93.91%	90.47%
DISTRIBUTION OF SUBREDDITS AGAINST POSTS						
Maximum number of subreddits	47,480	47,116	47,996	49,502	48,294	47,733
Maximum number of posts	25,006	23,746	26,055	26,650	28,743	24,211
α (power law parameter)	1.6321	1.5806	1.7512	1.8358	1.6293	1.7024
δ (power law parameter)	0.0256	0.0238	0.0362	0.0357	0.0263	0.029
DISTRIBUTION OF AUTHORS AGAINST POSTS						
Maximum number of authors	555,854	559,602	566,139	540,511	551,863	541,585
Maximum number of posts	18,724	17,401	18,268	16,513	17,226	19,949
α (power law parameter)	1.4531	1.6718	1.3565	1.399	1.5478	1.3742
δ (power law parameter)	0.0465	0.0359	0.0545	0.0233	0.0428	0.0757
DISTRIBUTION OF POSTS AGAINST SCORES						
Maximum score	183,453	185,056	180,553	191,864	180,578	179,099
α (power law parameter)	1.5986	1.631	1.4672	1.6026	1.6507	1.5681
δ (power law parameter)	0.0189	0.0186	0.0198	0.0086	0.0179	0.0359
DISTRIBUTION OF SUBREDDITS AGAINST AUTHORS						
Maximum number of subreddits	62,839	65,934	70,585	65,861	63,087	62,325
Maximum number of authors	20,285	19,571	18,808	21,801	20,029	19,801
α (power law parameter)	1.7185	1.7064	1.6209	1.608	1.7013	1.7853
δ (power law parameter)	0.0298	0.0485	0.0315	0.02	0.0379	0.0327

Parameter	*Jul*	*Ago*	*Sep*	*Oct*	*Nov*	*Dec*
GENERAL CHARACTERISTICS						
Number of authors who published at least one SFW post	59,465	60,563	59,489	59,873	58,985	60,236
Number of authors who published only SFW posts	58,801	59,423	58,965	58,742	58,632	59,542
Percentage of authors publishing SFW posts who published only posts of this type	98.88%	98.11%	99.11%	98.11%	99.40%	98.84%
Number of subreddits containing at least one SFW post	89,360	87,953	89,236	88,462	87,932	88,167
Number of subreddits containing only SFW posts	82,050	82,587	85,496	83,647	83,146	84,963
Percentage of subreddits containing SFW posts that contain only posts of this type	91.82%	90.74%	93.68%	91.76%	91.7%	94.4%
DISTRIBUTION OF SUBREDDITS AGAINST POSTS						
Maximum number of subreddits	46,283	46,882	48,777	47,676	48,886	47,070
Maximum number of posts	22,261	19,071	23,642	29,330	26,346	28,419
α (power law parameter)	1.582	1.8481	1.7838	1.7313	1.5937	1.5125
δ (power law parameter)	0.0186	0.0305	0.0535	0.0329	0.0468	0.0154
DISTRIBUTION OF AUTHORS AGAINST POSTS						
Maximum number of authors	541,585	574,678	542,568	569,611	576,835	556,736
Maximum number of posts	16,823	19,320	18,692	18,460	16,499	17,766
α (power law parameter)	1.3323	1.406	1.4688	1.4054	1.3093	1.525
δ (power law parameter)	0.0713	0.0491	0.0561	0.0424	0.064	0.038
DISTRIBUTION OF POSTS AGAINST SCORES						
Maximum score	194,305	176,975	164,394	186,004	172,001	177,739
α (power law parameter)	1.5089	1.5785	1.4772	1.6389	1.4331	1.6354
δ (power law parameter)	0.0114	0.054	0.0245	0.0389	0.0226	0.0012
DISTRIBUTION OF SUBREDDITS AGAINST AUTHORS						
Maximum number of subreddits	59,963	57,573	59,898	52,885	62,111	63,232
Maximum number of authors	18,901	20,056	20,285	19,962	21,078	20,909
α (power law parameter)	1.7622	1.6287	1.4544	1.8174	1.5256	1.7388
δ (power law parameter)	0.0159	0.0263	0.043	0.0254	0.0184	0.0378

Table 2.20: Monthly trend of some parameters related to SFW posts

Parameter	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*
GENERAL CHARACTERISTICS						
Number of authors who published at least one NSFW post	36,758	35,452	36,542	36,874	36,863	36,453
Number of authors who published only NSFW posts	36,094	35,259	36,501	36,165	36,135	36,023
Percentage of authors publishing NSFW posts who published only posts of this type	98.19%	99.45%	99.88%	98.07%	98.02%	98.82%
Number of subreddits containing at least one NSFW post	41,365	40,985	41,298	41,547	41,235	40,958
Number of subreddits containing only NSFW posts	34,055	33,254	34,587	32,982	33,563	34,159
Percentage of subreddits containing NSFW posts that contain only posts of this type	82.33%	81.13%	83.74%	79.38%	81.39%	83.40%
DISTRIBUTION OF SUBREDDITS AGAINST POSTS						
Maximum number of subreddits	18,332	17,985	19,547	21,034	20,135	20,235
Maximum number of posts	34,424	32,547	31,854	31,329	30,896	32,541
α (power law parameter)	1.6896	1.6721	1.6874	1.6852	1.6796	1.6852
δ (power law parameter)	0.0258	0.0254	0.0251	0.0254	0.0214	0.0261
DISTRIBUTION OF AUTHORS AGAINST POSTS						
Maximum number of authors	131,070	130,152	131,250	133,594	131,452	132,654
Maximum number of posts	16,383	16,125	14,214	15,674	16,540	14,210
α (power law parameter)	1.5463	1.7985	1.6222	1.8407	1.9456	1.4833
δ (power law parameter)	0.03345	0.0233	0.0239	0.0639	0.0388	0.0458
DISTRIBUTION OF POSTS AGAINST SCORES						
Maximum score	106,947	146,561	75,657	112,830	105,566	66,095
α (power law parameter)	1.6062	1.5162	1.6933	1.8989	1.6951	1.4956
δ (power law parameter)	0.0145	0.0265	0.042	0.0611	0.0346	0.0139
DISTRIBUTION OF SUBREDDITS AGAINST AUTHORS						
Maximum number of subreddits	62,839	63,382	61,204	33,963	50,609	53,781
Maximum number of authors	20,285	17,549	19,347	11,326	18,495	19,324
α (power law parameter)	1.7156	1.7682	1.6166	1.9204	1.753	1.6321
δ (power law parameter)	0.0312	0.0241	0.0384	0.0236	0.0187	0.0418

Parameter	*Jul*	*Ago*	*Sep*	*Oct*	*Nov*	*Dec*
GENERAL CHARACTERISTICS						
Number of authors who published at least one NSFW post	37,165	35,986	36,432	36,540	36,354	36,589
Number of authors who published only NSFW posts	36,984	35,421	35,962	35,986	35,756	35,852
Percentage of authors publishing NSFW posts who published only posts of this type	99.51%	98.42%	98.77%	98.48%	98.35%	97.98%
Number of subreddits containing at least one NSFW post	41,542	40,986	41,246	41,258	40,983	41,496
Number of subreddits containing only NSFW posts	34,478	33,352	34,254	34,165	33,241	33,986
Percentage of subreddits containing NSFW posts that contain only posts of this type	82.99%	81.37%	83.04%	82.80%	81.10%	81.90%
DISTRIBUTION OF SUBREDDITS AGAINST POSTS						
Maximum number of subreddits	20,135	18,564	17,423	19,631	18,328	20,124
Maximum number of posts	30,451	32,598	30,125	29,874	34,210	32,021
α (power law parameter)	1.6236	1.6454	1.59874	1.6598	1.6432	1.6953
δ (power law parameter)	0.0265	0.0259	0.0298	0.0265	0.0264	0.0254
DISTRIBUTION OF AUTHORS AGAINST POSTS						
Maximum number of authors	130,254	134,250	133,247	132,478	136,587	131,489
Maximum number of posts	16,125	14,256	15,879	16,325	14,369	16,362
α (power law parameter)	1.6992	1.4551	1.5295	1.5527	1.5524	1.6091
δ (power law parameter)	0.0446	0.048	0.0201	0.0268	0.0031	0.0428
DISTRIBUTION OF POSTS AGAINST SCORES						
Maximum score	97,462	143,430	102,590	100,844	104,027	81,167
α (power law parameter)	1.6422	1.5874	1.4948	1.7059	1.7936	1.3969
δ (power law parameter)	0.040	0.028	0.0386	0.0324	0.0184	0.0354
DISTRIBUTION OF SUBREDDITS AGAINST AUTHORS						
Maximum number of subreddits	49,210	76,791	64,241	54,351	50,864	34,037
Maximum number of authors	17,425	20,605	23,952	20,608	18,613	16,594
α (power law parameter)	1.7653	1.7342	1.5258	1.9738	1.6143	1.5882
δ (power law parameter)	0.0317	0.037	0.0204	0.0371	0.0207	0.0401

Table 2.21: Monthly trend of some parameters related to NSFW posts

2.2.4 Results

2.2.4.1 Co-posting activity of NSFW posts authors

The goal of this analysis is to verify whether there is any correlation between the authors of NSFW posts. As shown previously, we will extract the information of interest and we will compare the behavior of authors of NSFW posts with the ones of SFW posts. In this activity, we will use a support data structure that we call *co-*

posting network. Having observed in all the previous experiments that the results obtained for the Jan-Feb datasets (i.e., $\mathcal{D}$ and $\overline{\mathcal{D}}$) are stable, from now on we will refer to these two datasets only, avoiding to report the analysis of Mar-Apr datasets, too. In addition, since most of the operations that we will perform on the co-posting network are computationally expensive, we randomly extracted a subset $\mathcal{D}^*$ (resp., $\overline{\mathcal{D}}^*$) of $\mathcal{D}$ (resp., $\overline{\mathcal{D}}$) consisting of 75,000 SFW (resp., NSFW) posts to work on.

As a first task of this analysis, we give a formal definition of the co-posting network $\mathcal{P}$ (resp., $\overline{\mathcal{P}}$) built from the authors of SFW (resp., NSFW) posts stored in $\mathcal{D}^*$ (resp., $\overline{\mathcal{D}}^*$).

Formally speaking,

$$\mathcal{P} = \langle N, E \rangle \qquad \overline{\mathcal{P}} = \langle \overline{N}, \overline{E} \rangle$$

Here, N (resp., $\overline{N}$) is the set of the nodes of $\mathcal{P}$ (resp., $\overline{\mathcal{P}}$). There is a node $n_i \in N$ (resp., $\overline{N}$) for each author a_i of SFW (resp., NSFW) posts of $\mathcal{D}^*$ (resp., $\overline{\mathcal{D}}^*$). There is an edge $(n_i, n_j, w_{ij}) \in E$ (resp., $\overline{E}$) if the authors a_i and a_j (associated with n_i and n_j, respectively) submitted at least one post in the same subreddit. w_{ij} is the number of subreddits having at least one SFW (resp., NSFW) post of a_i and, simultaneously, at least one SFW (resp., NSFW) post of a_j.

Then, we calculated some of the basic parameters of $\mathcal{P}$ and $\overline{\mathcal{P}}$; they are shown in Table 2.22. From the analysis of this table, we can deduce that:

- The number of co-posting authors of NSFW posts is smaller than the number of co-posting authors of SFW posts.
- The authors of NSFW posts are more interconnected with each other. This is shown by both the density of $\overline{\mathcal{P}}$ (which is about three times the one of $\mathcal{P}$) and the average degree of $\overline{\mathcal{P}}$ (which is much greater than twice the degree of $\mathcal{P}$). As we will see in the following, this can be explained considering that they are authors belonging to a niche context.
- The average clustering coefficient of $\overline{\mathcal{P}}$ is greater than the one of $\mathcal{P}$, but not as much as the density. This suggests that in $\overline{\mathcal{P}}$ fewer triads are closed than in $\mathcal{P}$. This implies that, probably, in $\overline{\mathcal{P}}$ there are more "bridge" authors than in $\mathcal{P}$. These authors tend to act as intermediaries between other authors who do not know each other. They could be expert authors who cooperate with many new authors initially unknown to each other.

After this, we computed the distribution of the nodes of $\mathcal{P}$ and $\overline{\mathcal{P}}$ against their degree centrality. The results obtained are reported in Figures 2.34 and 2.35.

From the analysis of these figures we can see that both distributions follow a power law. We computed the corresponding values of α and δ and obtained that

Parameter	$\mathcal{P}$	$\overline{\mathcal{P}}$
Number of nodes	59,465	36,758
Number of edges	3,164,169	5,398,082
Density	0.001789	0.007990
Maximum Degree	2,593	3,670
Average Degree	106.42	293.70
Average Clustering Coefficient	0.7388	0.7755

Table 2.22: Basic parameters of the co-posting networks $\mathcal{P}$ and $\overline{\mathcal{P}}$

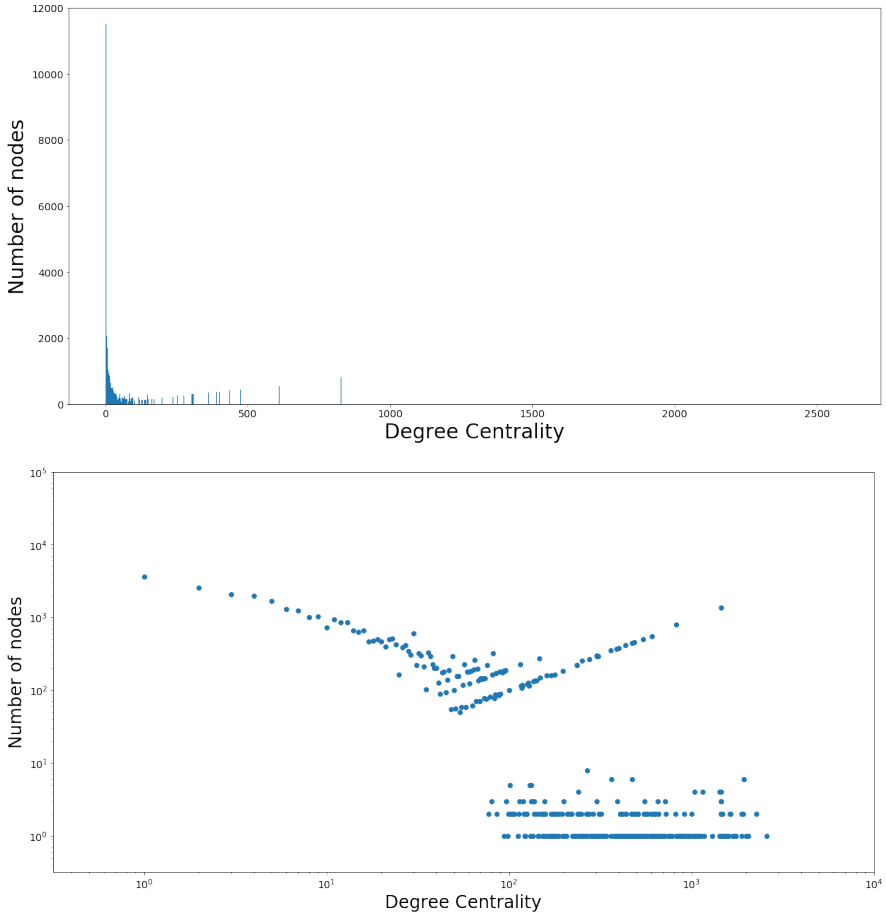


Fig. 2.34: Distribution of the nodes of $\mathcal{P}$ against their degree centrality - linear scale (on top) and log-log scale (on bottom)

$\alpha = 2.2929$ and $\delta = 0.0470$ for $\mathcal{P}$ and $\alpha = 2.6811$ and $\delta = 0.0678$ for $\overline{\mathcal{P}}$. These values tell us that the two distributions are similar.

Furthermore, looking carefully at the distributions in Figures 2.34 and 2.35, it emerges another unexpected, extremely peculiar, feature. In fact, we can observe some spikes. Excluding that these spikes are noise, they could be caused by the fact that the networks $\mathcal{P}$ and $\overline{\mathcal{P}}$ are actually disconnected and each network consists of a set of connected components. We found extremely interesting to check if this hy-

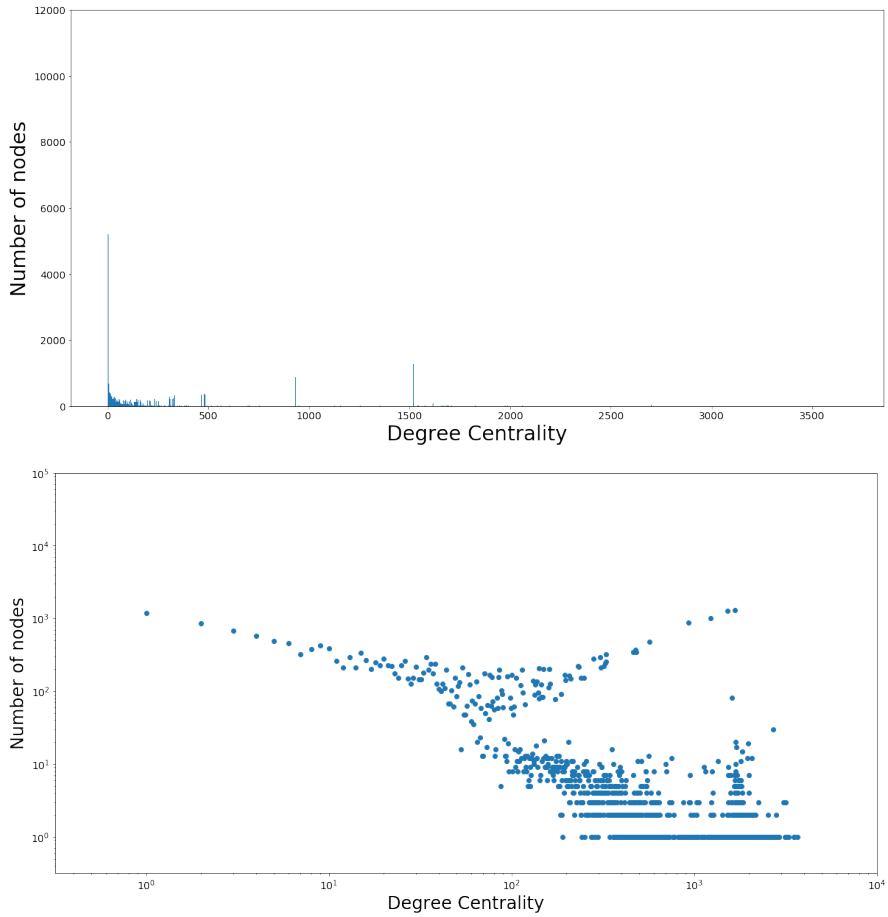


Fig. 2.35: Distribution of the nodes of $\overline{\mathcal{P}}$ against degree centrality - linear scale (on top) and log-log scale (on bottom)

pothesis was true. Therefore, we carried out this analysis and verified that, actually, we were right. In fact, we found that $\mathcal{P}$ consists of 15,952 connected components. Of these, 11,514 are made up of a single node. The maximum connected component includes 21,364 nodes (equal to 35,92% of the network nodes) and 2,909,206 arcs (equal to 91.94% of the network arcs). The distribution of the connected components against their size (i.e., the number of nodes they include) follows a power law with $\alpha = 1.562$ and $\delta = 0.060$. The network $\overline{\mathcal{P}}$ consists of 6,032 connected components, where 5,214 are made of a single node. The maximum connected component comprises 28,165 nodes (equal to 76.62% of the network's nodes) and 5,382,255 arcs (equal to 99.71% of the network's arcs). The distribution of the connected components against their size follows a power law with $\alpha = 1.548$ and $\delta = 0.065$.

The analysis of connected components strengthens some results obtained previously, in particular: *(i)* the number of co-posting authors of SFW posts is greater than the corresponding number of co-posting authors of NSFW posts; *(ii)* the authors of

NSFW posts are more connected to each other (probably due to the presence of the "bridge" users mentioned above) than the ones of SFW posts.

At this point, we wanted to investigate more on the behavior of the authors of SFW and NSFW posts. Specifically, we treated three activities, namely the writing of posts, the tendency to publish on many subreddits and the ability to attract interest. For each of these activities, we selected the top-ten authors from the maximum connected component of $\mathcal{P}$ and $\overline{\mathcal{P}}$ and we studied their behavior. In particular, Figure 2.36 (resp., 2.37 and 2.38) shows the top-ten authors who wrote the highest number of posts (resp., published in the largest number of subreddits, received the highest number of comments). The left part of this figure refers to the authors of SFW posts (belonging to the network $\mathcal{P}$), while the right part refers to the authors of NSFW posts (belonging to the network $\overline{\mathcal{P}}$).

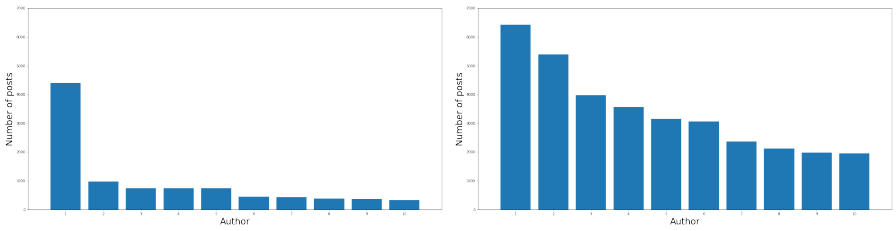


Fig. 2.36: Top-ten authors who submitted more posts - authors of SFW posts at left and of NSFW posts at right

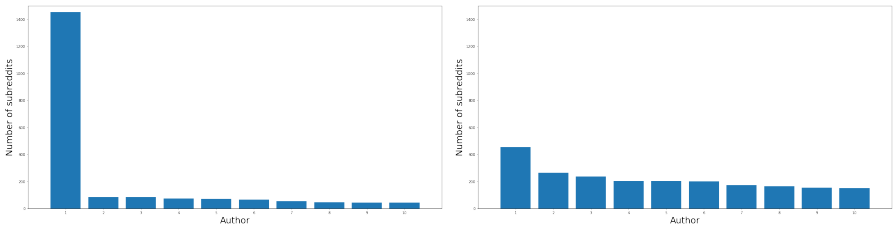


Fig. 2.37: Top-ten authors who published on more subreddits - authors of SFW posts at left and of NSFW posts at right

These figures altogether outline a very precise author behavior. In fact, it can be noted that, regardless of the activity considered, the authors of SFW posts show a power law distribution, while the authors of NSFW posts show a very slowly decreasing distribution. This allows us to conclude that there are few very active authors of SFW posts and many inactive ones in Reddit. By contrast, there are many quite active authors of NSFW posts. Once again, it seems that these last tend to "team up" much more than the ones of SFW posts.

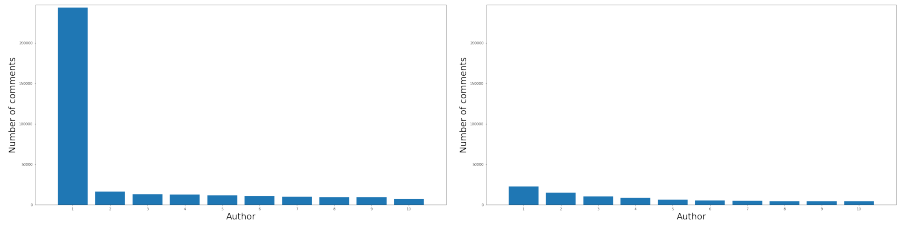


Fig. 2.38: Top-ten authors who received more comments - authors of SFW posts at left and of NSFW posts at right

These results can be explained considering that the phenomenon of NSFW posts is a niche one involving mostly particular kinds of user. These are very cohesive and form a fairly closed group. On the other hand, as we will see better in Section 2.2.4.3, all the knowledge extracted confirms this reasoning about the context behind NSFW posts.

2.2.4.2 Evaluating assortativity of NSFW posts authors

The concept of "assortativity", or "assortative mixing", in a social network points out the predilection of its nodes to be connected with other nodes that are somehow similar to them. This concept, introduced by Newman [502], can be seen as an evolution of the concept of homophily [468], typical of Social Network Analysis. Assortativity is orthogonal to node similarity metrics considered, even if most of the authors in the literature have studied it with respect to node degree. According to this definition of assortativity, the nodes of a social network tend to be linked with other nodes having a degree similar to their own.

Assortativity is considered an extremely important property to be investigated by social network researchers. So we decided to analyze it for the authors of SFW and NSFW posts in Reddit. We would also pinpoint that: *(i)* like in the previous analyses reported above, the goal is to characterize the assortativity of the authors of NSFW posts versus the one of the authors of SFW posts; *(ii)* the similarity property we decided to test for assortativity is node degree, because it is the most investigated one in the past literature on assortativity[9].

To carry out our assortativity analyses, we used the co-posting networks $\mathcal{P}$ and $\overline{\mathcal{P}}$ defined in Section 2.2.4.1. We showed the distributions of the nodes of these networks against degree centrality in Figures 2.34 and 2.35. As a first task, we sorted the authors of the two networks in descending order of degree centrality. After that, we split this ordered list into intervals. In particular, we considered 40 equi-width

[9] Actually, at the end of this section, for a further evidence of the results obtained, we also considered eigenvector centrality, beside degree centrality.

intervals $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_{40}\}$ for $\mathcal{P}$ and $\{\overline{\mathcal{I}}_1, \overline{\mathcal{I}}_2, \cdots, \overline{\mathcal{I}}_{40}\}$ for $\overline{\mathcal{P}}$. Since the number of nodes of $\mathcal{P}$ (resp., $\overline{\mathcal{P}}$) was 59,465 (resp., 36,578), each interval $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$) contained 1,487 (resp., 915) authors[10].

At this point, we considered the interval $\mathcal{I}_1$ (resp., $\overline{\mathcal{I}}_1$) and, for each interval $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$), we determined how many authors of $\mathcal{I}_1$ (resp., $\overline{\mathcal{I}}_1$) were connected to at least one author of $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$). The results obtained are shown in Figure 2.39(a) (resp., 2.39(c)). Next, we computed the percentage of the authors of $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$), who were connected to at least one author of $\mathcal{I}_1$ (resp., $\overline{\mathcal{I}}_1$). The results obtained are shown in Figure 2.39(e) (resp., 2.39(g)).

The analysis of Figures 2.39(a) and 2.39(e) shows a close correlation (i.e., a sort of backbone) between the authors of SFW posts with the highest degree centrality. On the contrary, the analysis of Figures 2.39(c) and 2.39(g) shows that this phenomenon does not occur for the authors of NSFW posts.

In order to evaluate the statistical significance of this result, we generated a null model to compare our outcomes with those of an unbiasedly random scenario. In particular, we built our null model shuffling the arcs of $\mathcal{P}$ (resp. $\overline{\mathcal{P}}$) among the nodes of this network. In this way, we left the original characteristics of $\mathcal{P}$ (resp. $\overline{\mathcal{P}}$) unchanged, except for the distribution of co-posting activities, which became unbiasedly random in the null model. The results obtained are shown in Figures 2.39(b), 2.39(d), 2.39(f) and 2.39(h).

Comparing Figures 2.39(b) and 2.39(f) with Figures 2.39(a) and 2.39(e) we can see that the represented distributions are similar. Indeed, many of the ranges with the highest values of Figures 2.39(a) and 2.39(e) continue to reach the highest values in Figures 2.39(b) and 2.39(f), too. However, these values are much smaller in the latter case. Therefore, we can conclude that the behavior observed in Figures 2.39(a) and 2.39(e) is not random, but intrinsic to $\mathcal{P}$ (and, therefore, to the authors of SFW posts in Reddit). On the contrary, if we consider Figures 2.39(c) and 2.39(g) (regarding the authors of NSFW posts in Reddit) and compare them with Figures 2.39(d) and 2.39(h), we can see that this phenomenon does not occur for the authors of $\overline{\mathcal{P}}$.

The above analysis suggests that there is a degree assortativity among the authors of SFW posts but not among the authors of NSFW posts. However, in order to confirm the assortativity of the authors of SFW posts, we need to verify whether this trend is still valid for the authors with an intermediate degree centrality and for those with a low degree centrality. If we want to make an exhaustive analysis, we should repeat the tasks previously performed for $\mathcal{I}_1$ (resp., $\overline{\mathcal{I}}_1$) for all the 40 intervals. For lack of space, we will limit our analysis to the intervals $\mathcal{I}_{20}$ (resp., $\overline{\mathcal{I}}_{20}$), as

[10] Actually, the last interval had a slightly smaller size equal to 1,472 (resp., 893) authors.

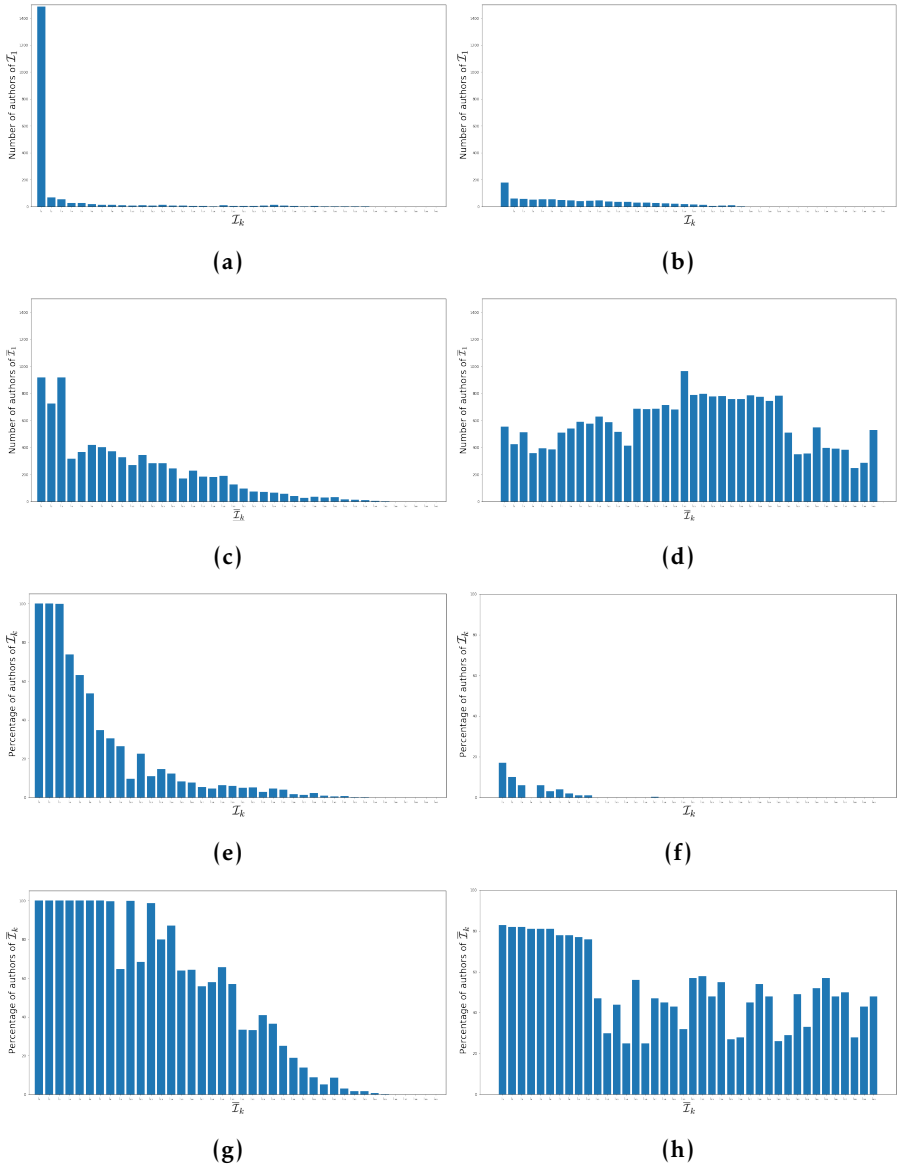


Fig. 2.39: Degree Assortativity of the authors of NSFW and SFW posts (high degree authors)

the representative of those with intermediate degree centrality, and $\mathcal{I}_{30}$ (resp., $\overline{\mathcal{I}}_{30}$), as the representative of those with low degree centrality[11].

Figure 2.40(a) (resp., 2.40(c)) shows the number of authors of $\mathcal{I}_{20}$ (resp., $\overline{\mathcal{I}}_{20}$) connected with at least one author of $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$), while Figure 2.40(e) (resp., 2.40(g))

[11] We did not choose the intervals $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$), $k > 30$, because, during the analysis of the connected components, we saw that there is a high number of isolated nodes in $\mathcal{P}$ (resp., $\overline{\mathcal{P}}$) - see Section 2.2.4.1. Clearly, these nodes belong to the highest intervals and, if considered, could represent a bias in our analysis. To avoid this bias, we chose to not consider the intervals where they reside, and to select $\mathcal{I}_{30}$ (resp., $\overline{\mathcal{I}}_{30}$) as the representative of the intervals with low degree centrality.

shows the percentage of authors of $\mathcal{I}_k$ (resp., $\overline{\mathcal{I}}_k$) connected with at least one author of $\mathcal{I}_{20}$ (resp., $\overline{\mathcal{I}}_{20}$). The analysis of these figures suggests the existence of a close correlation among the authors of SFW posts with an intermediate degree of centrality; this correlation does not exist for the authors of NSFW posts.

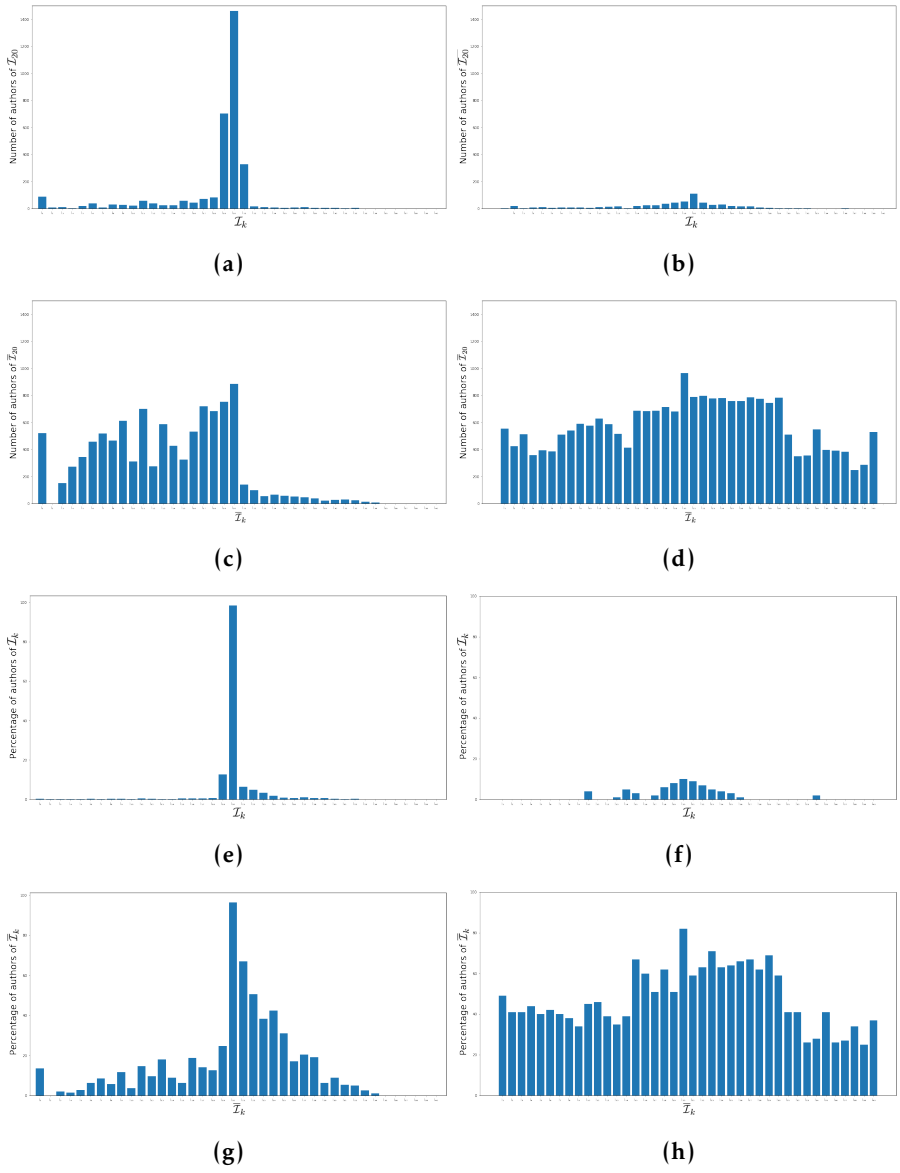


Fig. 2.40: Degree Assortativity of the authors of NSFW and SFW posts (medium degree authors)

Even in this case, we compared these findings with those obtained in the null model. The latter are shown in Figures 2.40(b), 2.40(d), 2.40(f) and 2.40(h). Looking at all the diagrams reported in Figure 2.40, once again we can conclude that the observed behavior is not random, but it is a property of Reddit.

In the light of the last observation and of the previous conclusions on authors with an intermediate and a high degree centrality, we can certainly assert that there is no degree assortativity for the authors of NSFW posts. Instead, the possibility that such assortativity exists for the authors of SFW posts remains open.

In order to verify this last possibility, we carried out a study on the authors of $\mathcal{I}_{30}$. Figure 2.41(a) shows the number of authors of $\mathcal{I}_{30}$ connected to at least one author of $\mathcal{I}_k$, while Figure 2.41(c) shows the percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{30}$. These figures reveal the presence of a close correlation between the authors of SFW posts with a low degree centrality.

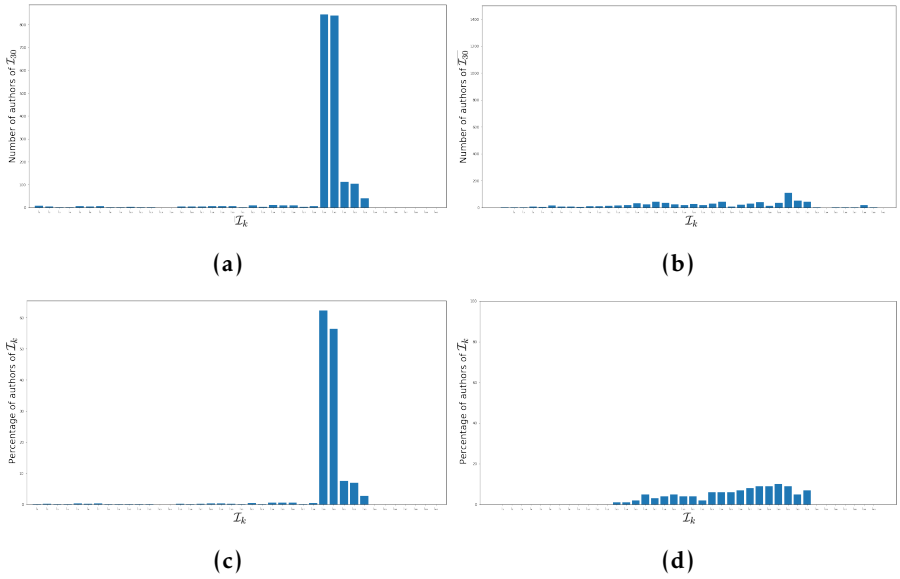


Fig. 2.41: Degree Assortativity of the authors of SFW posts (low degree authors)

Even in this case, we compared the results obtained with those returned using the null model. We report the latter in Figures 2.41(b) and 2.41(d). The comparison of these figures with Figures 2.41(a) and 2.41(c) confirms that the behavior observed for these authors is an intrinsic property of Reddit.

Having verified that there is a sort of backbone among the authors of SFW posts with high (resp., medium, low) degree centrality, we can conclude that there is a degree assortativity for the authors of SFW posts in Reddit. Instead, this property is absent for the authors of NSFW posts in Reddit.

A further interesting analysis is to check if the tendency of the authors of SFW posts to be assortative and the tendency of the authors of NSFW posts to be not assortative is general or strongly depends on the type of assortativity that is being considered (in this case, degree assortativity).

As a premise to this discussion, it should be pointed out that every form of assortativity is independent, so it is impossible to come to a *general rule*. However, the analysis previously mentioned could surely lead us to discover some *trends*.

Therefore, we chose a second form of centrality (in particular, the eigenvector centrality) and we repeated all the steps previously taken for degree centrality with this second one.

The results obtained are very similar to those we have seen for degree centrality, i.e., we found the existence of a strong eigenvector assortativity for the authors of SFW posts and a lack of eigenvector assortativity for the authors of NSFW posts. For space reasons, we cannot show all the results. However, in order to give an idea of them, in Figure 2.42, we report what happens for authors with high eigenvector centrality. Comparing this figure with Figure 2.39, we can observe a strong similarity in the authors behavior in the two cases. As a consequence, we can say that SFW authors *tend* to be assortative, while NSFW authors *tend* to be not assortative.

This result can be explained by the strong community sense of the authors of NSFW posts. They are so cohesive that they do not feel the need to split into groups of peers. The most active people are still willing to interact with everyone else and not only with other equally active people.

2.2.4.3 Knowledge findings on posts, authors and subreddits

Combining together all the previous results, we can define three main findings related to posts, authors and subreddits, respectively. Some of these findings are made up of several sub-findings.

The three findings are the following:

PF (Finding on NSFW posts)

1. NSFW posts are generally published in much fewer subreddits, have much lower scores and are much less commented than SFW posts.
2. The scores of comments to NSFW posts are much lower than the ones to SFW posts.

AF (Finding on NSFW authors)

1. NSFW authors tend: (i) to publish more posts, *(ii)* to publish in a fewer subreddits, *(iii)* to have a lower number of co-posting authors, *(iv)* to be more interconnected, active and "teamed" than SFW authors.
2. The maximum number of negative posts published by a single NSFW author is much higher than the corresponding one of a single SFW author.
3. Differently from what happens to SFW authors, there is no degree assortativity and no eigenvector assortativity among NSFW authors.

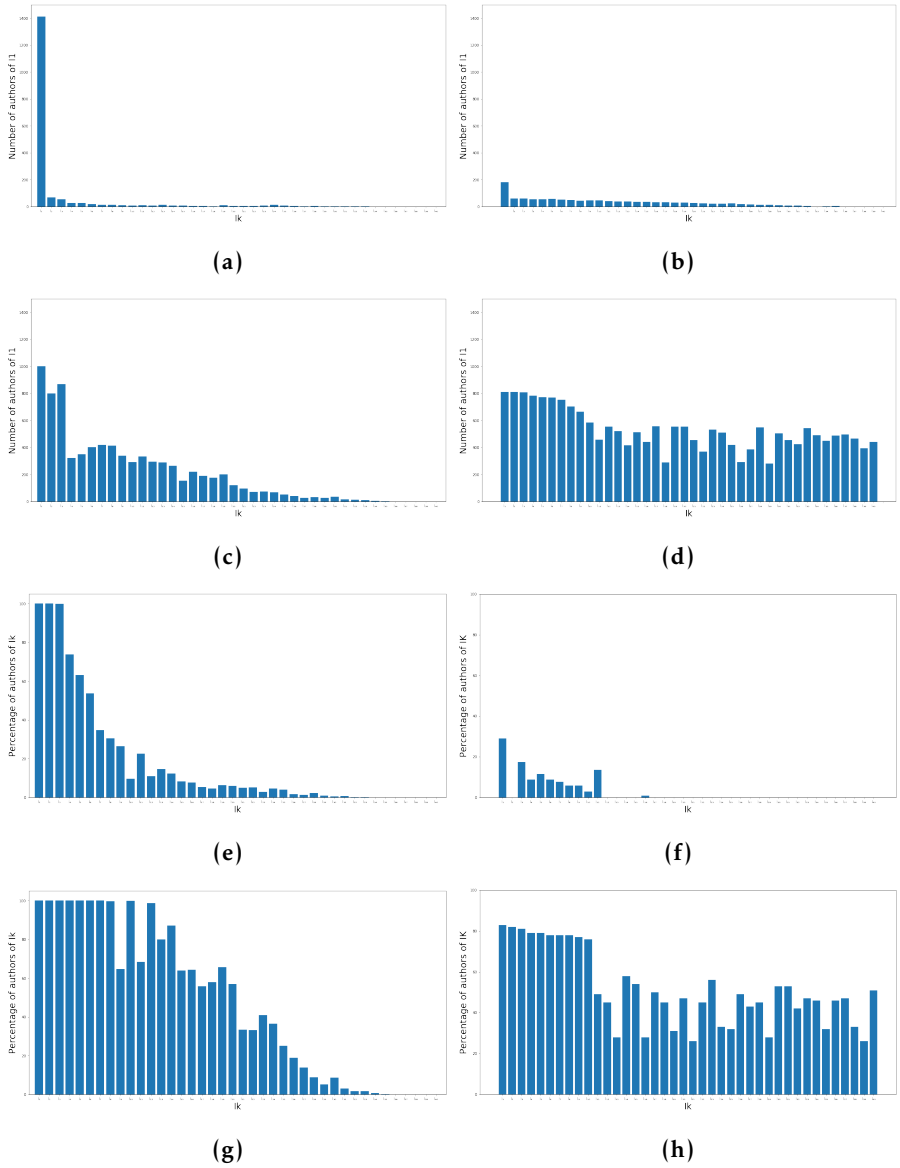


Fig. 2.42: Eigenvector Assortativity of the authors of NSFW and SFW posts (high degree authors)

SF (Finding on NSFW subreddits)

1. NSFW subreddits receive much fewer comments than SFW subreddits.

Now, we examine the previous findings in order to identify their correlations. This allows us to have a general view of the phenomenon of NSFW posts in Reddit.

The finding PF.1 tells us that an NSFW post is published in a limited number of subreddits. The finding AF.1 states that NSFW authors publish more than SFW ones. Now, since NSFW posts are fewer than SFW ones, we can conclude that NSFW posts have a much more limited number of authors. In addition, the combination of

PF.1 and AF.1 is also a justification to the claim that NSFW authors publish in fewer subreddits than SFW authors.

Combining the findings PF.1 and AF.1 we can conclude that the phenomenon of NSFW posts is a niche one.

The finding PF.1 also tells us that the NSFW posts are little appreciated; actually, this information was quite expected. The results expressed by the finding PF.1 are reinforced by the finding AF.2, which tells us that the maximum number of negative posts published by a single NSFW author is greater than the corresponding number of an SFW author. The finding AF.2 is also, in part, a direct consequence of the finding AF.1.

The finding SF.1, stating that the NSFW subreddits receive fewer comments than SFW ones, represents a further confirmation of what the findings AF.1 and PF.1 say about the fact that NSFW posts are a niche phenomenon.

The poor consideration for NSFW posts, expressed by the finding PF.1, is further confirmed by the finding PF.2, which tells us that not only NSFW posts, but even comments to these posts, receive a much lower score than the comments to SFW posts.

The finding AF.1 (which tells us that the number of co-posting NSFW authors is fewer than SFW authors and that NSFW authors are more interconnected, active and "teamed" than SFW ones) represents a further confirmation that the NSFW post phenomenon is a niche one, carried out by few authors. However, it also tells us that these authors are very active and very well interconnected, ready to play "team-work".

The last finding extracted, i.e., the finding AF.3, specifies that there is no degree or eigenvector assortativity for NSFW authors. In other words, the strong connection existing among NSFW authors is so widespread and compact that it does not let authors group into "narrow circles". In fact, the sense of cooperation between these authors is so high that the most active ones still collaborate with everyone else and do not limit their interactions to only those with their direct peers, as often happens in many other contexts.

3

Yelp

In this chapter, we apply our complex network approach to the popular social network Yelp. Initially, we introduce the concept of k-bridge (i.e., a user who connects k sub-networks of the same network or k networks of a multi-network scenario) and propose an algorithm for extracting k-bridges from a social network. Then, we analyze the specialization of this concept and algorithm in Yelp and derive several knowledge patterns about Yelp k-bridges. Furthermore, we define three stereotypes of Yelp users, along with their characteristics and the profile of negative influencers in Yelp. Regarding these lasts, we investigate their influence on their friends while doing negative reviews and the correlation between the centrality measures and being this kind of influencer.

The material present in this Chapter is taken from [169, 207].

3.1 Defining and detecting k-bridges

3.1.1 Introduction

Bridges, i.e., entities connecting different sub-networks of the same network or different networks of a multi-network scenario, attracted the interest of many researchers in several disciplines, ranging from sociology to telecommunication networks and transports. They also attracted the interests of researchers studying Online Social Networks, who considered them as users linking sub-networks of a single network [279, 606, 416, 95, 98, 689] or linking different networks in a multi-network context [134, 141, 136, 517].

In the past, all researchers focused on the bridge capability of connecting *two* communities. However, with the proliferation of social media, bridges currently tend to connect a higher number of sub-networks in a network or a higher number of networks in a multi-network scenario. Furthermore, we argue that their behavior and properties could vary against the number k of communities they connect. As a consequence, it appears interesting to introduce a new notion, that we call *k-bridge*. A

k-bridge is a user who connects k sub-networks of a network or k networks of a multi-network scenario. k-bridges are particular users capable of playing an important role in opinion transmission, user influence, etc. Indeed, they allow a person or a business in a community to be known in another one. This may have important applications in the dissemination of information, in the search for influencers, and in marketing, for example when a business, leader in one category, wants to expand in another related category.

In this chapter, we first present and formalize the notion of k-bridge and we show that it has interesting properties, such as the anti-monotone one. Then, we propose a k-bridge detection algorithm that exploits these properties. Afterwards, we extract several knowledge patterns about k-bridges.

In order to carry out these activities, we use Yelp as the main reference network. Yelp[1] is a platform that helps people find local businesses, like dentists, restaurants, hair stylists, and many more. It is a business directory service and a crowd-sourced review forum that provides its users with a web site (*Yelp.com*), a mobile app (*Yelp mobile app*), and a reservation service (*Yelp reservation*). In the second quarter of 2019, it reached a monthly average of 37 million visitors through its mobile application and 77 million visitors through its web site, along with a total of 192 million reviews.

The motivations underlying our choice to adopt Yelp as a main study platform are related to its pure crowd-sourced nature. This characteristic is very important in our investigations as users in Yelp are free to interact with the platform and write reviews without constraints. As a matter of fact, researchers have found in Yelp one of the main resources for studying user behavior in open-review platforms. Therefore, many works on Yelp have been focused on review and rate analysis, sentiment analysis, fake review and fake rate discovery, and recommendation analysis [145, 648, 493, 444, 669].

The definition of k-bridges in Yelp starts from the hypothesis of seeing this social platform as a set of sub-nets or communities, one for each of its macro-categories. Actually, the importance of studying Yelp categories has already been highlighted in recent scientific literature [187]. In this chapter, we want to go one step further and we consider that the communities associated with the macro-categories of Yelp are not independent from each other, because a user who reviews businesses of different macro-categories belongs to several communities.

Even if we performed our investigations of k-bridges and their characteristics in Yelp, we carried out some of the same experiments in two additional networks, i.e., Reddit[2] and the network of patent inventors derived from PATSTAT-ICRIOS [199],

[1] https://www.yelp.com

[2] https://www.reddit.com

a repository storing metadata of patents submitted in many countries (see below). The ultimate goal was to verify if the results we found in Yelp were generally valid for k-bridges.

As a last contribution, we present two possible use cases that could benefit from the knowledge and the exploitation of k-bridges. The former regards the engagement of k-bridges in Yelp to find the best targets of a market campaign, whereas the latter concerns the analysis of k-bridges' activities to infer new products/services in order to expand and improve the revenues of existing businesses.

The outline of this chapter is as follows: in Section 3.1.2, we present related literature. In Section 3.1.3, we propose a model for k-bridges along with an approach to extract them, and we investigate the k-bridge users properties. Then, in Section 3.1.4, we study the relationships between k-bridges and macro-categories in Yelp, validate their properties in other social networks (such as Reddit and the network of patent inventors), and present two use cases that could benefit from k-bridges and their properties.

3.1.2 Related Literature

Studying the behavior of users in social platforms is a fundamental aspect to understand the dynamics underlying the diffusion and the growth of these systems [365, 721]. A lot of research has been devoted to understanding how users interact in social media and how information diffusion takes place inside them [66, 661, 690, 100].

The interaction among users has been studied by leveraging several information available in these social systems, ranging from existing public friendship relationships to the posting of the same piece of information [588, 107, 15].

These studies have proved that there exist different categories of users, each participating to the platform with different levels of activity and heterogeneous contents [91, 455].

Of course, when dealing with user interactions, it is important to consider those that cannot be examined homogeneously [147]. This rises the necessity of analyzing data of each social medium by decomposing it in different networks of relations. Multi-relational networks have been largely investigated in the past [634, 223, 697, 717]. For instance, in [223], the authors focus on link prediction in an environment characterized by multiple relation types. Specifically, they present a probabilistically weighted Adamic/Adar measure for networks with heterogeneous relations. Moreover, they test their solution against three different real-world networks, characterized by heterogeneous relations, showing the performance of both supervised and unsupervised link prediction in such a multiple relation scenario.

Still in the context of predicting links in a multi-relation system, the authors of [697] focus on a co-authorship network and consider different types of link, namely: *(i)* co-author; *(ii)* co-participation to the same edition of a conference, and *(iii)* geographic proximity. They present a Multi-Relation Influence Propagation Model and demonstrate its usefulness in the link prediction task. Another interesting approach in the field of multi-relation networks is the one proposed in [719]. Here, the authors combine the analysis of the friendship network with a study of the author-topic network, both built from the information available in an Online Social Network. They use this knowledge to refine a community detection strategy and prove that the additional information coming from the author-topic network is fundamental to improve the overall performance of their strategy.

Considering each social medium as a set of overlapping relation networks also opens important consequences in the role of each user inside these platforms. Indeed, in [634] the authors perform a deep analysis of an Online Social Network derived by a community of online gamers. To study the multi-relation nature of this system, they consider three types of positive interactions (e.g., friendship) and three types of negative ones (e.g., enmity). First, they study each of these networks separately and find that those built on top of negative interactions have lower reciprocity, weaker clustering and fatter-tail degree distribution than those built on top of positive interactions. Then, they report a study about the tendency of users to be members of more networks and, hence, to play different roles inside the community.

Like the work described in [634], different studies have been devoted to analyzing the role of users in the creation of social communities. In particular, the authors of [371] demonstrate that users with a weak connection, bridging heterogeneous groups, have higher levels of community commitment, civic interest, and collective attention than the other users. Furthermore, they prove that Internet users, who bridge heterogeneous online communities by means of weak ties [298], have high social engagement, use the Internet for social purposes, and are prone to become members of new social communities.

The interest towards users serving as bridges among communities has increased over the years so that several studies have been performed to analyze the behavior and peculiarities of such users in complex networks [279, 606, 416, 54].

Studying nodes bridging communities together has been also a crucial research direction in the context of multi-relation networks [95, 98]. Here, the heterogeneity of the scenario is more evident because of the different nature of the relation considered. In particular, the authors of [95] report a complete analysis of bridge users among multi-relation networks. Specifically, they introduce a new class of parameters, namely Dimension Relevance, which measures the importance of different

dimensions for the user's capabilities of being a bridge. In order to prove the meaningfulness of their measures, they leverage real networks as well as null models and, then, they study the overlapping dimensions along with their effect on user connectivity.

In [98], instead, the authors focus on community discovery strategies taking the multi-relation structure of the network into account. Specifically, they define a new concept of community that groups together nodes sharing memberships to the same mono-relation communities and propose a community discovery algorithm based on frequent pattern mining in multi-relation networks. This algorithm is able to find multi-relation communities based on the analysis of frequent closed itemsets from mono-relation community memberships.

Still in the context of bridges among heterogeneous communities, several studies also analyzed the behavior of users serving as bridges among different social networks [134, 141, 136]. Here the concept of community is extended in such a way that a community is mapped to a whole social network. Specifically, in [134], the authors report a complete identikit of users bridging different social networks. They compare the behavior of this type of users with other members having different levels of activity and participation to the platforms. The results show that bridges are more active than average users but they still are not at the top of the tall head of the power law distribution that models user activities in these systems. Another study in this context is the one described in [141]. Here, the authors leverage the peculiarities of bridges to define a new crawling strategy to sample a multi-social network environment. Finally, the work of [136] performs a comparative study of users serving as bridges among two of the most famous social networks, namely Facebook and Twitter. Once again, the authors report that bridges have unique behaviors compared to normal users and that they tend to start new activities in social media. The authors also prove that this type of users is more aware of the functionalities provided by the online social platforms they are involved in. Interestingly, bridges are found to be also more cautious when it comes to their privacy and the security of the information released in social media.

All the works described above clearly highlight the importance of studying the peculiarities of users acting as melting pots among different social communities. The analysis performed follows this trend. Furthermore, it considers the different nature of the relations among users and investigates the role of bridges for each of them. Interestingly, to the best of our knowledge, our investigation is the first to study this type of users in Yelp. Actually, in recent years, Yelp has received a lot of attention from the scientific community. The corresponding works can be classified in the following groups, according to their goal: *(i) Rating Analysis*: It includes the inves-

tigations that analyze the dynamics describing how rates are assigned to businesses in Yelp [145, 342, 414, 630, 215, 614]; *(ii) Review Analysis*: It comprises the works focused on the analysis of reviews and of what events drive the users writing them [648, 632, 529, 530, 88, 307]; *(iii) Sentiment Analysis*: It also deals with the analysis of reviews, but with a specific focus on their content from a sentiment point of view [493, 582, 58, 306]; *(iv) Fake review and rate discovery*: It includes the proposals dealing with the detection of fake reviews and rates [444, 492, 456, 408]; *(v) Recommender Systems*: It comprises all the research works devoted to providing Yelp users with recommendations about suitable businesses, other users to interact with, and even text suggestions for new reviews [669, 395, 242, 187, 660].

Despite our work shares some similarities with several other ones described in this section, to the best of our knowledge, this is the first attempt to introduce a new concept, namely the k-bridge. This concept formalizes the idea that, in social networking, bridges with different level of strength exist, and that the strength of bridges represent an important dimension to investigate when analyzing their behavior in the environment which they operate on.

Given the new concept of k-bridge, this chapter provides several contributions to understand the main features of this kind of actors. In particular:

- It shows that k-bridges enjoy the anti-monotone property.
- Starting from this property, it proposes a new algorithm for the extraction of k-bridges from social networks.
- It provides a model for representing k-bridges in the social network they belong to.
- It presents three specializations of the concept of k-bridges for Yelp, Reddit and the network of patent inventors.
- It finds several important characteristics of k-bridges and shows that they are valid independently of the social network they refer to.
- It presents two use cases highly benefiting from bridges; the former regards the identification of the best targets of a market campaign, whereas the latter concerns the identification of new products/services to propose.

Our study strongly differs from the ones about Yelp presented above. Indeed, the purpose of our investigation is to provide a deep insight on the features of users acting as bridges among different Yelp macro-categories. The importance of studying Yelp categories has already been highlighted in recent scientific literature. For example, in [187] the authors argue about the importance of properly weighting features and information across categories when dealing with recommender systems. We start from this assumption and focus on users encouraging the interaction

among different Yelp macro-categories. The heterogeneous nature of Yelp macro-categories allows us to classify our work among those studying the peculiarities of users who act as bridges in heterogeneous online communities. In Yelp, the same pair of users can be linked by different kinds of relationship, for instance friendship and co-review. As a consequence, we can derive different network-based representations of a Yelp user, one for each kind of possible relationship type that can be defined among its users. Thanks to this, we can investigate k-bridges in Yelp from different viewpoints, one for each representation. Following a terminology similar to the one adopted in the approaches described above, this way of proceeding can be summarized by saying that we analyze Yelp as a multi-relation environment. The knowledge of previous works, along with the analogies and differences between the ideas reported therein and the objectives of our research, represents the base of our k-bridge model and our k-bridge extraction approach that we present in the next sections.

3.1.3 Methods

3.1.3.1 A model for k-bridges and an approach to extract them

In this section, firstly we propose a general model for k-bridges, and specialize it to several social networks and, then, we present an algorithm to extract k-bridges.

Defining and modeling k-bridges

Let $\mathcal{N}$ be a social network and let $\mathcal{CS}$ be the set of the communities of $\mathcal{N}$ of our interest:

$$\mathcal{CS} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_M\}$$

Given the community $\mathcal{C}_i$, $1 \leq i \leq M$, it is possible to define the corresponding user network $\mathcal{U}_i = \langle N_i, A_i \rangle$. N_i is the set of nodes of $\mathcal{U}_i$; there is a node n_{i_p} for each user u_{i_p} belonging to $\mathcal{C}_i$. A_i is the set of arcs of $\mathcal{U}_i$; there is an arc $a_{pq} = (n_{i_p}, n_{i_q}) \in A_i$ if there exists a relationship between the users u_{i_p} and u_{i_q}, corresponding to n_{i_p} and n_{i_q}, respectively.

Finally, it is possible to define the overall user network $\mathcal{U} = \langle N, A \rangle$ corresponding to $\mathcal{N}$. There is a node $n_i \in N$ for each user of $\mathcal{N}$. There is an arc $a_{pq} = (n_p, n_q) \in A$ if there exists a relationship between the users u_p and u_q, corresponding to n_p and n_q, respectively.

Here, and in the previous definition, we do not specify the kind of relationship between users. As we will see in the following, it is possible to define a specialization of $\mathcal{U}$ for each relationship we want to investigate. For instance, $\mathcal{U}^f$ is the specialization of $\mathcal{U}$ when we consider *friendship* as the relationship between users.

After having introduced our model, we can present our definitions of *k-bridge*, *bridge*, *non-bridge*, *strong bridge and very strong bridge*.

Definition 3.1. A ***k-bridge*** is a user of $\mathcal{N}$ belonging to exactly k different communities of this social network, $1 \leq k \leq M$. □

Definition 3.2. A ***non-bridge*** is a k-bridge such that $k = 1$, i.e., a user belonging to exactly one community. □

Definition 3.3. A ***bridge*** is a k-bridge such that $k \geq 2$, i.e., a user who belongs to at least 2 different communities of $\mathcal{N}$. □

Definition 3.4. A ***strong bridge*** is a k-bridge such that $k \geq th_s$. Here, th_s is a threshold such that $2 \leq th_s < M$. □

Definition 3.5. A ***very strong bridge*** is a k-bridge such that $k \geq th_{vs}$. Here, th_{vs} is a threshold such that $th_s < th_{vs} \leq M$. □

Observe that the definition of k-bridge is anti-monotone. This means that if a user is a k-bridge then she is also a h-bridge $1 \leq h \leq k-1$.

Finally, given a k-bridge $u_p^k \in \mathcal{U}$, there are k nodes $n_{1_p}, n_{2_p}, \cdots, n_{k_p}$ associated with her, one for each community of $\mathcal{N}$ it belongs to. Each node represents a sort of "avatar" of u_p^k in the network corresponding to this community.

An algorithm for k-bridge extraction

An important consequence of the anti-monotone property of k-bridges mentioned above is the possibility of designing an optimized algorithm to extract them, borrowing some ideas from the well-known Apriori approach [17]. Indeed, the anti-monotone property allows us to state that the search space to find k-bridges is reduced to the set of identified (k-1)-bridges, which can be obtained, in turn, starting from the set of identified (k-2)-bridges, and so forth. This observation strongly resembles the reasoning and the properties underlying the Apriori algorithm. In our case, due to the possible huge number of users who could be bridges, it is more convenient to revert the problem and extend our reasoning to communities. Indeed, according to the definition of bridges, we can derive a formal property for communities, as follows:

*Property 3.6 (*Anti-monotonicity of communities). *All the communities involved in the definition of k-bridges must also be involved in the definition of (k-1)-bridges.* □

Therefore, a possible algorithm to identify k-bridges from the communities of a social network consists of the following steps. First, for each community, the set of

the corresponding users is retrieved. Intuitively, in order to be consistent with its general definition, a community must have a minimum number of users joining it. We call this measure `support` and we impose that a community must have a `support` greater than a threshold min_sup. The result of this step is a set of communities called L_1.

To obtain 2-bridges, we start from L_1 and compute a set of community pairs, called P_1, joining L_1 with itself. Each pair of communities in P_1 represents a possible case in which at least a user acts as a bridge between them. Therefore, for each pair of communities in P_1, we compute the intersection of their users, and impose, once again, that its cardinality is greater than min_sup. The resulting filtered set of community pairs is called L_2. Observe that, for each community pair in L_2, the intersection among the corresponding users is also an outcome of this iteration as it contains all 2-bridges.

To compute 3-bridges, the algorithm proceeds by joining L_2 with itself; in this way, it obtains a set of community triplets, called P_2. Each triplet in P_2 contains the communities candidate to be simultaneously joined by 3-bridges. Once again, for each triplet in P_2, we compute the intersection of users among the three communities and impose that its cardinality is greater than min_sup. The resulting set is called L_3. Also in this case, the set of 3-bridges, which is the outcome of this iteration, is implicitly obtained in the intersection computed above for each element of L_3.

In general, this procedure can be extended to compute k-bridges starting from the set L_{k-1} used to computed (k-1)-bridges. Algorithm 1 reports a pseudo-code of our approach for extracting k-bridges from a social network.

As a final remark, we observe that our solution can be easily extended to a big data strategy (which is a realistic requirement in the social network context) by leveraging the advances available for Apriori in the scientific literature, because our algorithm follows a strategy very near to the one adopted by Apriori. For instance, it is possible to adapt our solution to work in a Map-Reduce based architecture following the studies described in [420, 702].

Specializing our k-bridge model to Yelp

In Yelp, businesses are organized according to a taxonomy consisting of four levels. Level 0 comprises 22 macro-categories. Each macro-category has one or more child categories, so that level 1 comprises 1002 categories. A category may have zero, one or more sub-categories, so that level 2 consists of 532 sub-categories. Proceeding with this reasoning, the final level, i.e., level 3, has only 19 sub-sub-categories; indeed, most sub-categories are not further categorized.

Input

- D, a dataset of a Social Network
- $\mathcal{CS}$, the set of communities of D
- min_sup, a suitable threshold for minimum support

Output

- L_k, the set of k-communities linked by k-bridges
- B_k, the set of k-bridges

Require: L_t, a temporary set; $\texttt{getN}(\mathcal{C}_i)$ a function returning the set of users of the community $\mathcal{C}_i$

$L_1 = \{\mathcal{C}_i \mid \mathcal{C}_i \in \mathcal{CS} \wedge |\texttt{getN}(\mathcal{C}_i)| > th_s\}$ *//the set of communities in the dataset having support greater than min_sup*
$P = L_1 \bowtie L_1$ *// $\bowtie$ is the join operator*
$j = 2$ *//start with 2-bridges*
while $j \leq k$ **do**
  if $P \neq \emptyset$ **then**
    //for each tuple of the communities in P
    for $<(\mathcal{C}_1),(\mathcal{C}_2),\cdots,(\mathcal{C}_j)> \in P$ **do**
      $I = \texttt{getN}(\mathcal{C}_1) \cap \texttt{getN}(\mathcal{C}_2) \cap \cdots \cap \texttt{getN}(\mathcal{C}_j)$
      //if the minimum support is satisfied for this intersection
      if $|I| > min\_sup$ **then**
        Add $<\mathcal{C}_1,\mathcal{C}_2,\cdots,\mathcal{C}_j>$ to L_t
        //in the last iteration, store the found bridges and the involved
        //communities into the output parameters B_k and L_k, resp.
        if $j == k$ **then**
          Add I to B_k
          $L_k = L_t$
        end if
      end if
    end for
    $P = L_t \bowtie L_t$ *//re-compute P for the next iteration*
    $j++$, $L_t = \emptyset$
  end if
end while
return L_k, B_k

Algorithm 1: k-bridges Extraction Algorithm

When we specialize our model to Yelp, we have that this social network can be modeled as a set of 22 communities, one for each macro-category:

$$\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \cdots, \mathcal{Y}_{22}\}$$

Given the macro-category $\mathcal{Y}_i$, $1 \leq i \leq 22$, and the corresponding user network $\mathcal{U}_i = \langle N_i, A_i \rangle$, there is a node n_{i_p} for each user u_{i_p} who reviewed at least one business of $\mathcal{Y}_i$. Based on the relationship that we want to model, $\mathcal{U}$ can be specialized into $\mathcal{U}^f$, obtained when we consider friendship as the relationship between users, and $\mathcal{U}^{cr}$, obtained when co-review (i.e., reviewing the same business) is the relationship between users.

Given a k-bridge $u_p^k \in \mathcal{U}$, the k nodes $n_{1_p}, n_{2_p}, \cdots, n_{k_p}$ associated with her represent u_p in the k macro-categories where she performed at least one review.

Specializing our k-bridge model to Reddit

In Reddit, a user can participate to several subreddits. In this social network, the number of both users and subreddits is huge. So, in specializing our model to it, we consider only a subset of subreddits, for instance those about a certain topic or those published in a certain time interval. We can consider all the users who published at least one post in a subreddit as a community. So, we can model this scenario as:

$$\mathcal{R} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_M\}$$

Given the subreddit $\mathcal{S}_i$, $1 \leq i \leq M$, and the corresponding user network $\mathcal{U}_i = \langle N_i, A_i \rangle$, there is a node n_{i_p} for each user u_{i_p} who submitted at least one post in $\mathcal{S}_i$. Based on the relationship that we want to model, $\mathcal{U}$ can be specialized into $\mathcal{U}^{cp}$, obtained when co-posting (i.e., contributing to the same subreddit) is the relationship between users.

Given a k-bridge $u_p^k \in \mathcal{U}$, the k nodes associated with her represent u_p in the k subreddits where she submitted at least one post.

Specializing our k-bridge model to the community of patent inventors (and/or applicants)

Patents are largely investigated in scientific literature because they provide a large amount of knowledge patterns on Research & Development sector [262, 236]. Patents can be grouped in several ways, for instance based on the country of their inventors and/or applicants or according to the International Patent Classification (IPC) class they belong to. According to this classification, they have associated a symbol of the form A01B 1/00. Here:

- The first letter denotes the "section" of the patent (for instance, A indicates "Human necessities").
- The following two digits denote its "class" (for instance, A01 indicates "Agriculture; forestry; animal husbandry; trapping; fishing").

Notation	Semantics
$\mathcal{N}$	a generic social network
$\mathcal{C}_i$	the i^{th} community of $\mathcal{N}$
M	the maximum number of communities of $\mathcal{N}$
$\mathcal{U}_i$	the network representing the users of $\mathcal{C}_i$ and their relationships
N_i	the set of nodes of $\mathcal{U}_i$
A_i	the set of arcs of $\mathcal{U}_i$
u_i^p	the p^{th} user of the community $\mathcal{C}_i$
n_i^p	the node of $\mathcal{U}_i$ corresponding to u_i^p
$\mathcal{U}$	the overall user network corresponding to $\mathcal{N}$
n_i	a node of $\mathcal{U}$
$\mathcal{U}^r$	the specialization of $\mathcal{U}$ to the relationship r
th_s	the threshold for defining strong bridges
th_{vs}	the threshold for defining very strong bridges
$\mathcal{Y}_i$	the i^{th} community of Yelp
$\mathcal{S}_i$	the i^{th} subreddit of Reddit
$\mathcal{I}_i$	the set of inventors who filed at least one patent belonging to the i^{th} IPC class
$\mathcal{U}^f$	the specialization of $\mathcal{U}$ by taking the friendship relationship in Yelp
$\mathcal{U}^{cr}$	the specialization of $\mathcal{U}$ by taking the co-review relationship in Yelp
$\mathcal{U}^{cp}$	the specialization of $\mathcal{U}$ by taking the co-posting relationship in Reddit
$\mathcal{U}^{ci}$	the specialization of $\mathcal{U}$ by taking the co-inventory relationship in PATSTAT-ICRIOS
$\mathcal{M}$	the "macro-category" network of Yelp
$\mathcal{M}^{X\%}$	the subset of $\mathcal{M}$ whose macro-categories have been reviewed by at least X% of users

Table 3.1: The main notations used throughout this chapter

- The next letter indicates the "subclass" (for instance, A01B represents "Soil working in agriculture or forestry; parts, details, or accessories of agricultural machines or implements, in general").
- The next one-to-three-digit number represents the "group".
- Finally, the other two digits denote the "main group" or "subgroup".

A patent examiner assigns classification symbols to each patent according to the above rule, at the most detailed level which is applicable to its content.

After having chosen a level of the IPC classification, for instance the "class" level, the set of patent inventors (or, alternatively, the set of patent applicants), taken from a world patent metadata repository, for example PATSTAT-ICRIOS, can be represented as:

$$\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_M\}$$

Given the IPC class i, the corresponding set of inventors $\mathcal{I}_i$ (i.e., the set of inventors who filed at least one patent belonging to this class), $1 \leq i \leq M$, and the corresponding user network $\mathcal{U}_i = \langle N_i, A_i \rangle$, there is a node n_{i_p} for each inventor u_{i_p} who filed at least one patent of the class $\mathcal{I}_i$. $\mathcal{U}$ can be specialized into $\mathcal{U}^{ci}$, obtained when co-inventing (i.e., filing the same patent) is the relationship between inventors.

After having defined a model for k-bridges and an approach to extract them, after having specialized it to Yelp, Reddit and the network of patent inventors, in

the next section, we will focus on k-bridge properties. To help the reader understand the concepts of this chapter, in Table 3.1, we report the main notations introduced.

3.1.3.2 Investigating k-bridge properties

In this section, we analyze k-bridge properties. We carried out this task focusing on Yelp, which is our reference network. However, in the next paragraphs, we present some experiments on Reddit and the network of patent inventors devoted to verifying if the results on k-bridges found in Yelp are general or specific for this social network.

Overview of Yelp dataset

The data required for the investigation activities was downloaded from the Yelp website at the address `https://www.yelp.com/dataset`.

In order to extract information of interest from this data, we needed a preliminary analysis. As a first insight, we found 10,289 businesses that belong to a category not referable to any of the macro-categories, and 482 businesses that belong to no category at all. Since the total number of businesses was 192,609, we considered these data as noise and so we discarded it.

After this task, we analyzed the distribution of the categories in the macro-categories. The result obtained is shown in Figure 3.1. From the analysis of this figure, we can observe that the "Restaurants" macro-category has a much larger number of categories than the other macro-categories.

Note that, in Yelp, a business can belong to more macro-categories. Therefore, as a preliminary step, it seemed us particularly interesting to analyze how many times two macro-categories appeared simultaneously in the same business. The total number of businesses with at least two macro-categories is 59,086. The top 20 pairs of macro-categories that appear several times together in one business of Yelp are shown in Table 3.2. As we can see from this table, there are two pairs of macro-categories (i.e., ⟨ "Restaurants", "Food" ⟩ and ⟨ "Restaurants", "Nightlife" ⟩) that appear together a much higher number of times than the other pairs.

After that, we considered the total number of Yelp users who made at least one review and we saw that it is equal to 1,637,138. The distribution of their reviews is shown in Figure 3.2. We can observe that this distribution follows a power law. This result is perfectly in line with the ones of numerous studies about Online Social Networks and communities [484]. These studies highlight that the well-known social theory, according to which human activities usually follow a power law distribution, is still valid also in online communities. As a consequence, also in this kind of community, a few numbers of individuals (typically 10-20% of members)

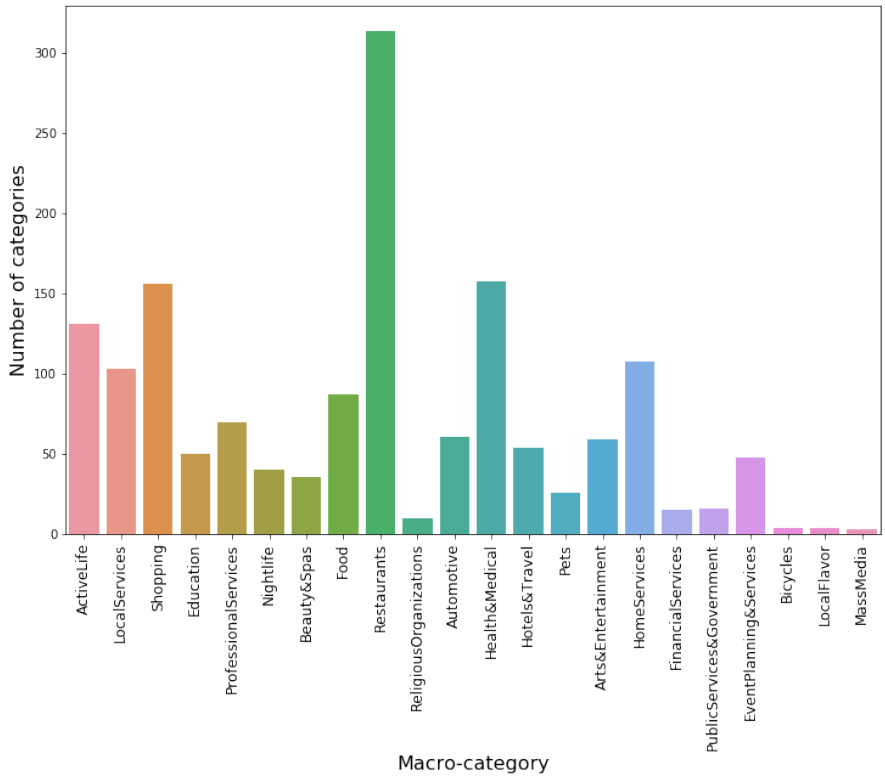


Fig. 3.1: Distribution of categories inside the macro-categories of Yelp

Pair of macro-categories	Count		Pair of macro-categories	Count
Restaurants, Food	11094		Restaurants, EventPlanning&Services	1051
Restaurants, Nightlife	5566		HomeServices, ProfessionalServices	758
Health&Medical, Beauty&Spas	2544		Automotive, Food	736
Shopping, LocalServices	2315		Shopping, EventPlanning&Services	708
HomeServices, LocalServices	1998		Arts&Entertainment, Nightlife	589
Hotels&Travel, EventPlanning&Services	1964		LocalServices, ProfessionalServices	579
Shopping, HomeServices	1883		ActiveLife, Health&Medical	527
Shopping, Beauty&Spas	1711		ActiveLife, Shopping	484
Shopping, Food	1470		FinancialServices, HomeServices	445
Shopping, Health&Medical	1384		Shopping, Arts&Entertainment	434

Table 3.2: The top 20 pairs of macro-categories that appear simultaneously in one business of Yelp

perform the majority of the activities (around 80-90% of the overall activities) [698]. Our experiment confirms that this trend also persists in the review tasks in Yelp.

The non-bridges are 530,411. All the other users are bridges. In order to start a deeper investigation of the k-bridge phenomenon, we computed the distribution of k-bridges against k. This is shown in Figure 3.3. An examination of this figure reveals that also this distribution follows a power law.

A last interesting, although partially expected, result that we found concerns the average number of reviews made by users. This is equal to 5.493 for bridges and

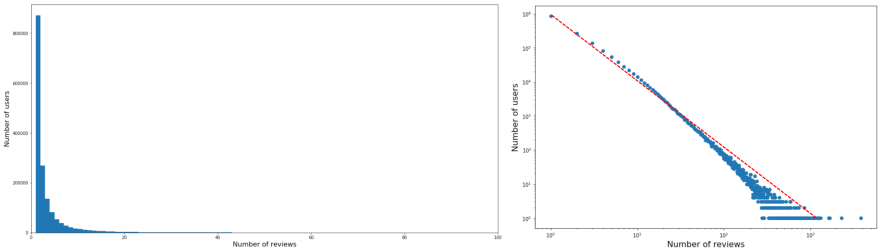


Fig. 3.2: Distribution of user reviews in Yelp - Linear scale (on the left) and Logarithmic scale (on the right)

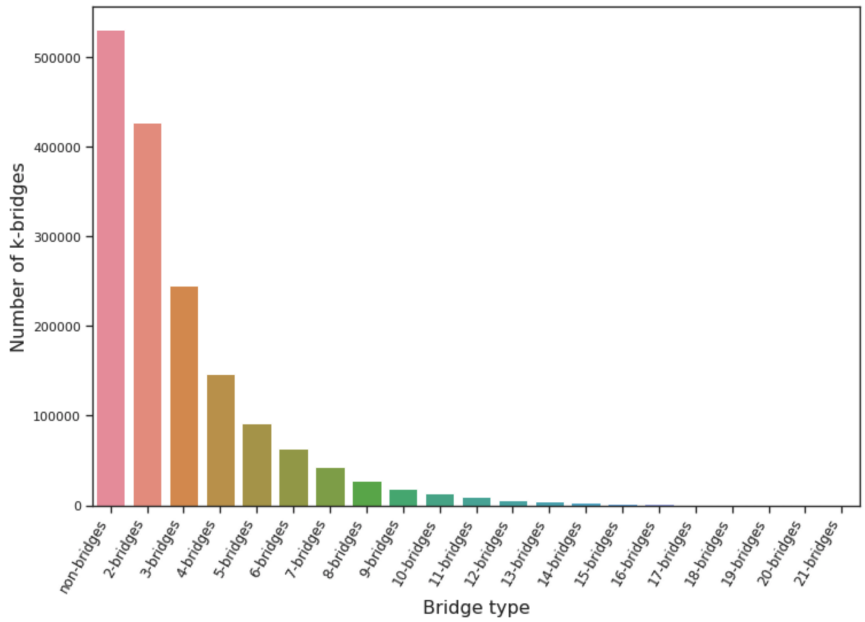


Fig. 3.3: Distribution of the k-bridges against k in Yelp

1.143 for non-bridges. This result confirms that a bridge tends to carry out more reviews than a non-bridge. It is also interesting to observe the corresponding standard deviations. In fact, the one for bridges is 17.69 whereas the one for non-bridges is 0.486. Such a high standard deviation for bridges confirms that this category of users is very varied, since it includes users who perform a huge number of reviews alongside users who perform few reviews. This is not the case, instead, for non-bridges, who always make few reviews.

k-bridges in the Yelp Friendship network

We began to verify the possible existence of a backbone among the bridges in $\mathcal{U}^f$. In order to have a connected network to study, we performed a pre-processing activity during which we eliminated the unconnected nodes from $\mathcal{U}^f$, corresponding to users who had no friendship relationship. The number of users having at least one friend

(and, therefore, the number of network nodes) is 948,076. Specifically, 676,445 of these were bridges, while 271,631 were non-bridges.

After that, for each bridge (non-bridge), we measured the fraction of her friends who were bridges (non-bridges). The results obtained are shown in Table 3.3. From the analysis of this table, we can see that there are no significant differences in the fraction of bridges in the neighborhoods of bridges and non-bridges. The same applies to the fraction of friends of non-bridges. In light of this, we can conclude that there is no backbone among the bridges in $\mathcal{U}^f$.

	Fraction of friends that are bridges	**Fraction of friends that are non-bridges**
Bridges	0.9618	0.0382
Non-bridges	0.9633	0.0367

Table 3.3: Types of friends for bridges and non-bridges in $\mathcal{U}^f$

Then, we analyzed whether there was any form of correlation between being a bridge and having friends. For this purpose, we computed the fraction of bridges (non-bridges) having at least one friend and the fraction of bridges (non-bridges) having no friends. The result obtained is reported in Table 3.4. From the analysis of this table, we can see that bridges have a higher tendency to have friends than non-bridges. However, the extent of this phenomenon is not extremely evident.

	Fraction of users with friends	**Fraction of users without friends**
Bridges	0.6113	0.3887
Non-bridges	0.5121	0.4879

Table 3.4: Fractions of users with and without friends in $\mathcal{U}^f$

At this point, we focused on investigating the possible influence that bridges exert on their neighborhoods. This investigation requires the usage of the strong and the very strong bridges. To detect them, it is necessary to specify the values of th_s and th_{vs} (see Section 3.1.3.1). To perform this task, we considered the distribution of the k-bridges against k in Yelp and we observed that it follows a very steep power law. As a consequence, according to the general trend of power law distributions, in particular of those showing a steep trend [698], it appeared us reasonable to choose th_s in such a way that only 10% of bridges are strong. Applying an analogous reasoning, we chose th_{vs} in such a way that only 10% of strong bridges are very strong. This way of proceeding led us to obtain that $th_s = 6$ and $th_{vs} = 12$.

After having determined the values of th_s and th_{vs}, we computed the fraction of strong and very strong bridges in the neighborhoods of bridges and non-bridges,

respectively. The result is shown in Table 3.5. Differently from what emerges from Table 3.3, where there is a little difference between the *fraction of bridges* in the neighborhoods of bridges and non-bridges, in Table 3.5 it is evident that there is a big difference on the *strength of bridges* in the neighborhoods of bridges and non-bridges. In fact, the fraction of very strong bridges is more than double in the neighborhoods of bridges compared to the neighborhoods of non-bridges.

	Fraction of strong bridges	**Fraction of very strong bridges**
Bridge neighborhoods	0.41	0.12
Non-bridge neighborhoods	0.27	0.05

Table 3.5: Fraction of strong and very strong bridges present in the neighborhoods of bridges and non-bridges in $\mathcal{U}^f$

As a further verification of this trend, we computed:

- The ratio of the number of *non-bridges* in a bridge's neighborhood to the number of non-bridges in a non-bridge's neighborhood. This is equal to 2.50.
- The ratio of the number of *bridges* in a bridge's neighborhood to the number of bridges in a non-bridge's neighborhood. This is equal to 5.23.
- The ratio of the number of *strong bridges* in a bridge's neighborhood to the number of strong bridges in a non-bridge's neighborhood. This is equal to 7.27.
- The ratio of the number of *very strong bridges* in a bridge's neighborhood to the number of very strong bridges in a non-bridge's neighborhood. This is equal to 10.97.

This analysis fully confirms the fact that, in the neighborhoods of bridges, it is much more frequent to find strong or very strong bridges than in the neighborhoods of non-bridges.

As a final analysis on neighborhoods, we computed the distribution of bridges and non-bridges present in the neighborhood of a bridge and a non-bridge, respectively. These two distributions are illustrated in Figures 3.4 and 3.5. These figures show that both of them follow a power law distribution. Looking at the values of these distributions, we can observe that the difference between the values of non-bridges and weak bridges is not very evident. Instead, this difference becomes evident for strong and very strong bridges. This is a third confirmation of the trends seen previously.

k-bridges in the Yelp Co-review network

After the analysis done on the friendship network $\mathcal{U}^f$, we investigated the co-review network $\mathcal{U}^{cr}$. We started by verifying the existence of a backbone among the bridges

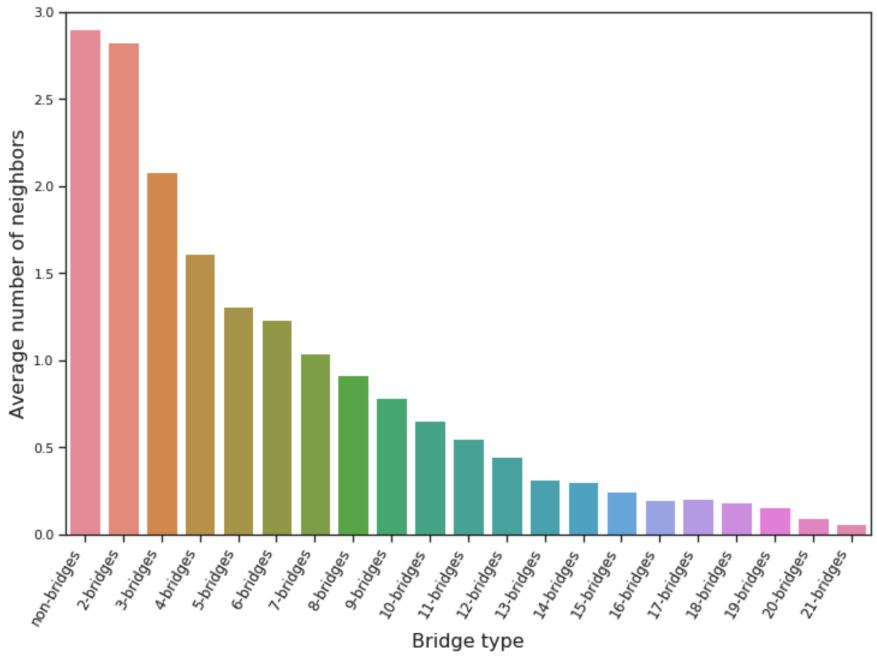


Fig. 3.4: Distribution of the neighbors of *bridges* in $\mathcal{U}^f$

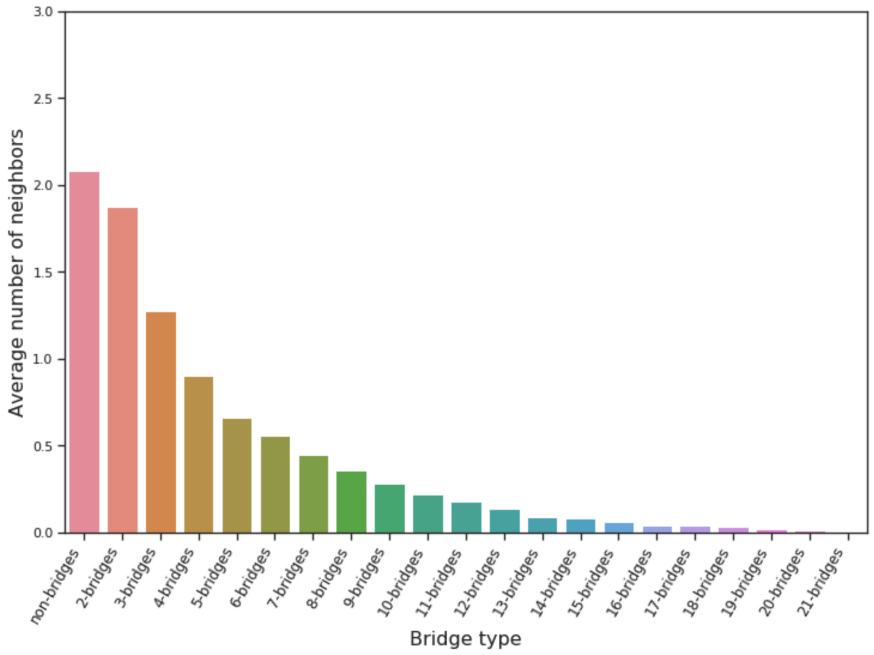


Fig. 3.5: Distribution of the neighbors of *non-bridges* in $\mathcal{U}^f$

in this network. Preliminarily, we removed those nodes corresponding to users who reviewed businesses not belonging to any macro-category of Yelp. As a consequence, the number of users (and, therefore, the number of nodes) who composed this network was equal to 1,634,547. Specifically, 1,037,484 of these were bridges while 597,063 were non-bridges.

The first analysis we made concerned the distribution of reviews with respect to users. The result obtained is shown in Figure 3.6. From the analysis of this figure, we can see that the distribution follows a power law. As a further analysis, we observe that $\mathcal{U}^{cr}$ is much denser than $\mathcal{U}^{f}$. In fact, the average degree of its nodes is equal to 1426.34, while, in $\mathcal{U}^{f}$, it is equal to 82.92.

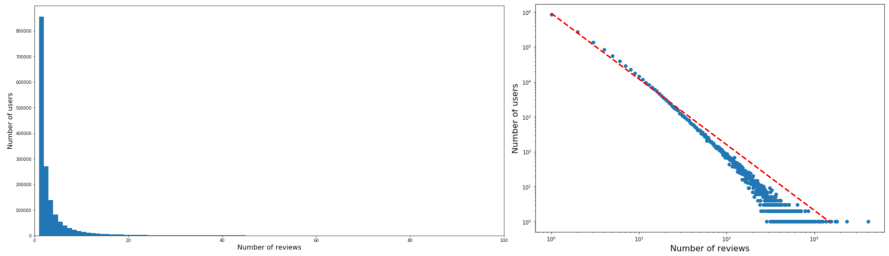


Fig. 3.6: Distribution of reviews for users in $\mathcal{U}^{cr}$ - Linear scale (on the left) and Logarithmic scale (on the right)

As a first analysis, we verified if there is a backbone among the bridges in $\mathcal{U}^{cr}$. Similarly to what we did for $\mathcal{U}^{f}$, for each bridge (non-bridge) we considered the fraction of co-reviewers that were bridges (non-bridges). The results obtained are shown in Table 3.6. From the analysis of this table we can see that there are significant differences in the percentage of co-reviewers that are bridges between a bridge and a non-bridge. The same applies to the percentage of co-reviewers that are non-bridges. In light of this, we can conclude that there is a backbone among the bridges in $\mathcal{U}^{cr}$.

	Fraction of co-reviewers that are bridges	**Fraction of co-reviewers that are non-bridges**
Bridges	0.9456	0.0543
Non-bridges	0.7451	0.2548

Table 3.6: Types of co-reviewers for bridges and non-bridges in $\mathcal{U}^{cr}$

As a further analysis of the neighborhoods of bridges and non-bridges in $\mathcal{U}^{cr}$, we computed the distribution of bridges and non-bridges present in the neighborhoods of bridges and non-bridges, respectively. These distributions are shown in Figures 3.7 and 3.8. These figures fully confirm the previous results about $\mathcal{U}^{cr}$. In fact, we can observe how the presence of bridges in the distribution of the neighbors of a bridge is very evident. The same happens for the presence of non-bridges in the distribution of the neighbors of non-bridges. These results represent a confirmation of the presence of a backbone among the bridges in the co-review network.

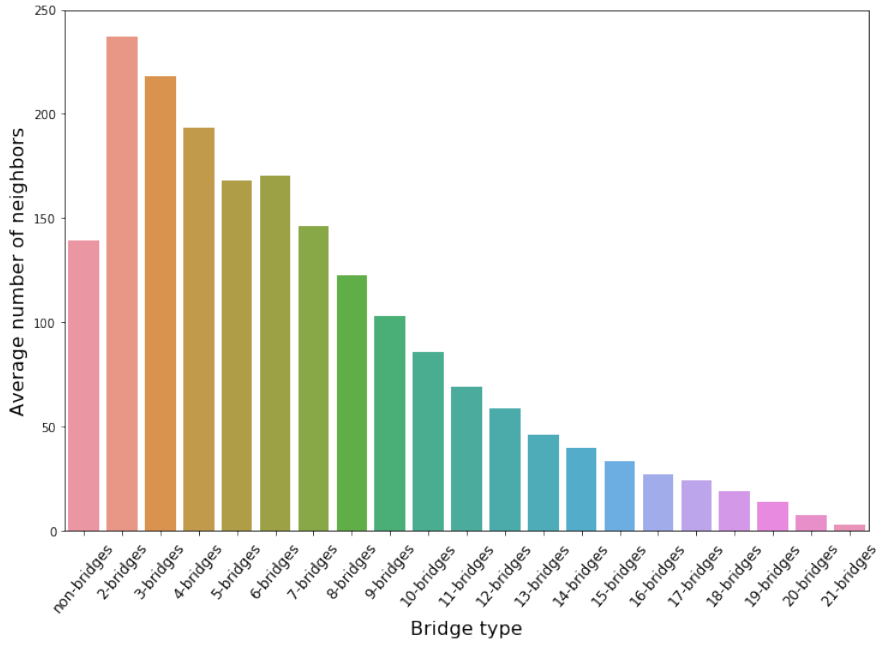


Fig. 3.7: Distribution of the neighbors of *bridges* in $\mathcal{U}^{cr}$

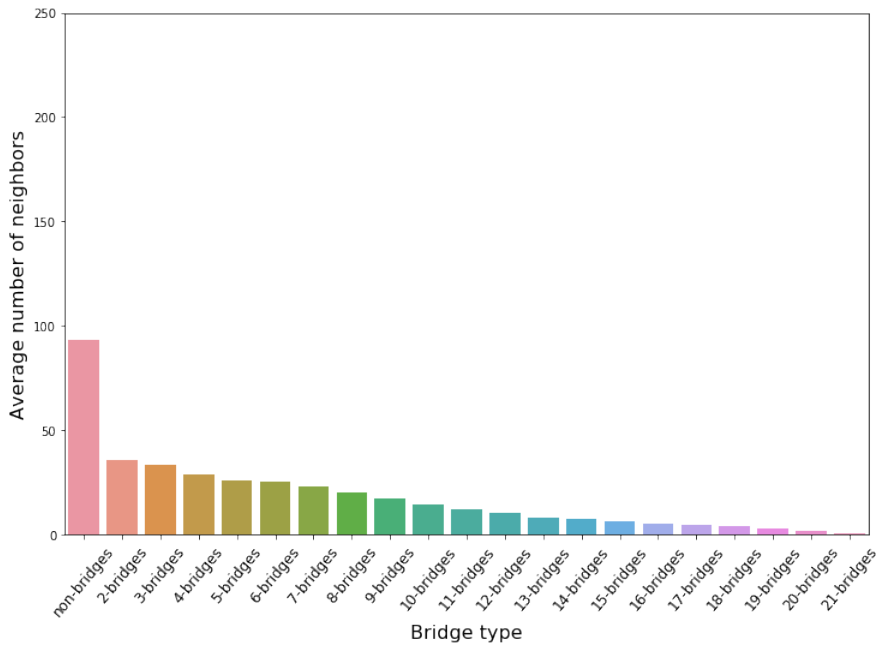


Fig. 3.8: Distribution of the neighbors of *non-bridges* in $\mathcal{U}^{cr}$

As a next analysis, we focused on the investigation of the possible influence that bridges can exert on their co-reviewers. For this objective, we computed the fraction of strong and very strong bridges present in the neighborhoods of bridges and non-bridges, respectively. The result is shown in Table 3.7. From the analysis of this table we can see that, differently from what happens in $\mathcal{U}^f$, in $\mathcal{U}^{cr}$ the fraction of strong and very strong bridges present in the neighborhoods of bridges is almost identi-

cal to the corresponding fraction relative to the neighborhoods of non-bridges. This means that, while there exists a backbone linking bridges together, their evolution towards strong and very strong bridges does not depend on the support received by their neighbors.

	Fraction of strong bridges	**Fraction of very strong bridges**
Bridge neighborhoods	0.54	0.15
Non-bridge neighborhoods	0.57	0.18

Table 3.7: Fraction of strong and very strong bridges present in the neighborhoods of bridges and non-bridges in $\mathcal{U}^{cr}$

As a further verification of this trend we computed:

- The ratio of the number of *bridges* in the neighborhood of a bridge to the number of bridges in the neighborhood of a non-bridge. This is equal to 12.83.
- The ratio of the number of *strong bridges* in the neighborhood of a bridge to the number of strong bridges in the neighborhood of a non-bridge. This is equal to 12.19.
- The ratio of the number of *very strong bridges* in the neighborhood of a bridge to the number of very strong bridges in the neighborhood of a non-bridge. This is equal to 10.73.

This analysis fully confirms the previous one, i.e., the fact that there is no strong correlation between the strength of a bridge and being or not neighbor to another bridge in $\mathcal{U}^{cr}$.

The presence of a backbone among the bridges in $\mathcal{U}^{cr}$ and the absence of an analogous backbone among the bridges in $\mathcal{U}^{f}$ led us to consider $\mathcal{U}^{cr}$ more interesting than $\mathcal{U}^{f}$ for further analyses on k-bridges. Therefore, we decided to perform all the next investigations only on $\mathcal{U}^{cr}$.

Analysis of the possible correlation between k-bridges and power users in the co-review network

Firstly, we verified if there is a correlation between k-bridges and power users or, in other words, between k-bridges and degree centrality. To this end, we computed the distribution of the number of arcs for non-bridges, bridges, strong and very strong bridges. The results obtained are shown in Figure 3.9. As we can see from this figure, all distributions follow power laws; their corresponding coefficients α and δ are reported in Table 3.8. However, we observe that as k grows, the power law distributions move to the right and flatten out. It implies that, as k grows, the degree cen-

trality of the corresponding k-bridges grows. This allows us to conclude that there is a correlation between the strength of k-bridges and degree centrality.

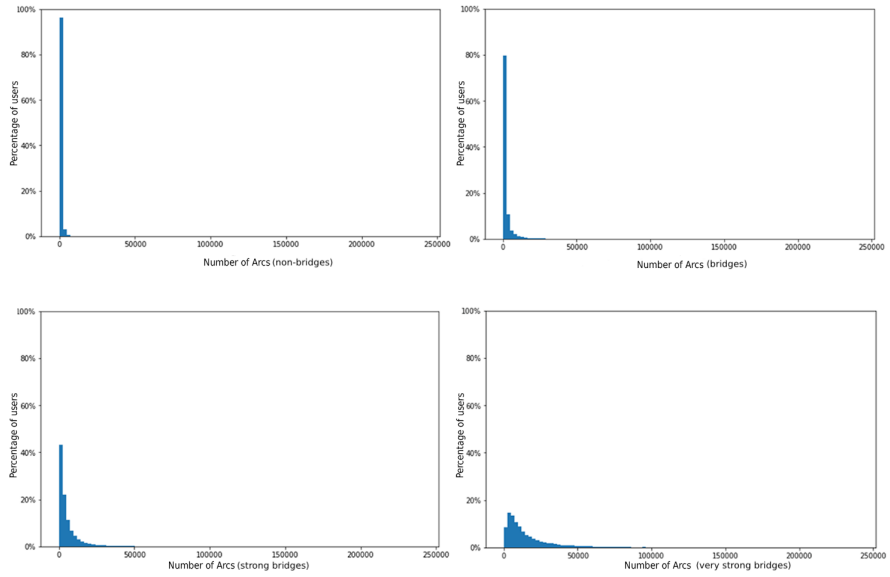


Fig. 3.9: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges

	α	δ
Non-bridges	1.203	0.177
Bridges	1.403	0.066
strong bridges	1.290	0.077
Very strong bridges	1.322	0.113

Table 3.8: Coefficients α and δ for the power law distributions of Figure 3.9

As a second analysis, we selected the top 1% of power users (corresponding to the top 1% of the nodes of $\mathcal{U}^{cr}$ with the highest degree) and determined how these were distributed between k-bridges (with k varying). We also repeated this analysis for the top 5%, the top 10%, the top 15%, the top 20% and, finally, for all users. The results obtained are shown in Figure 3.10. The analysis of this figure reveals that, as we select increasingly strong power users, the fraction of them that are strong bridges also increases, as the distribution moves to the right. This is a confirmation of the previous results regarding the existence of a correlation between k-bridges and power users.

As a final task, we repeated the previous analysis but we inverted k-bridges and power users. In particular, we selected the top 1% of k-bridges and determined the

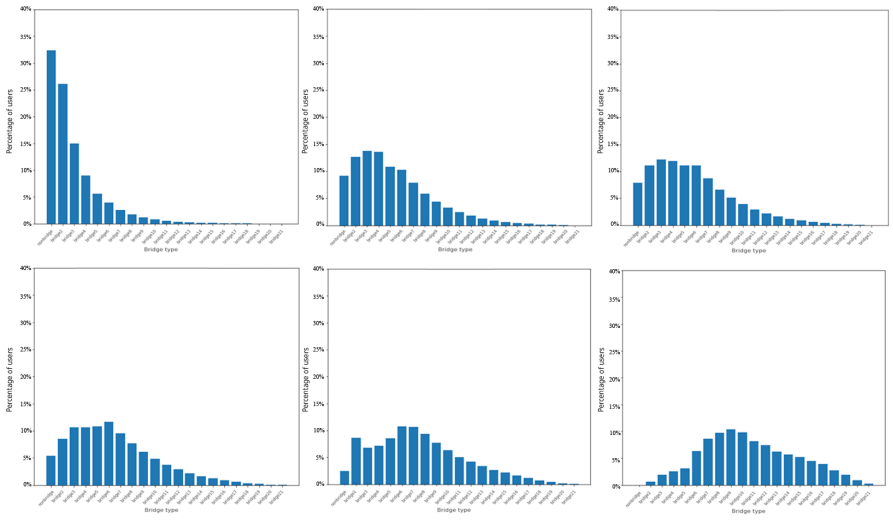


Fig. 3.10: Distributions of (power) users against the strength of bridges

distribution of their degree. We repeated this analysis for the top 5%, the top 10%, the top 15%, the top 20% of k-bridges and, finally, for all users. The results obtained are shown in Figure 3.11. From the analysis of this figure, we can see that the distribution moves to the right. This implies that, as we select stronger and stronger bridges, the fraction of them with higher and higher degree increases too. This represents a third confirmation of the previous results and, ultimately, allows us to say that there is a strong correlation between k-bridges and power users.

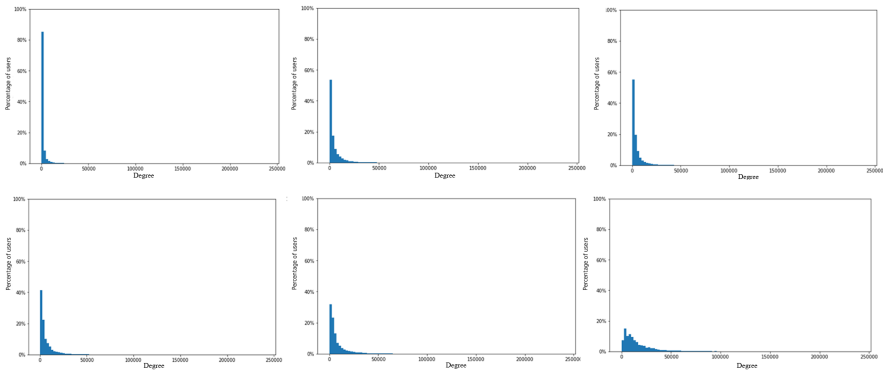


Fig. 3.11: Distributions of k-bridges against their degree

After having investigated the main properties of k-bridges, we focus on Yelp more deeply by analyzing the possible correlations between k-bridges and Yelp macro-categories.

3.1.4 Results

3.1.4.1 Analysis of k-bridges and macro-categories in Yelp

In this section, we aim at deepening our study of the correlations between k-bridges and Yelp macro-categories.

First of all, we considered the macro-categories which the reviews made by Yelp users refer to. The corresponding distribution is shown in Figure 3.12. From the analysis of this figure we can see that the "Restaurants" macro-category has a much higher number of reviews than all the other ones.

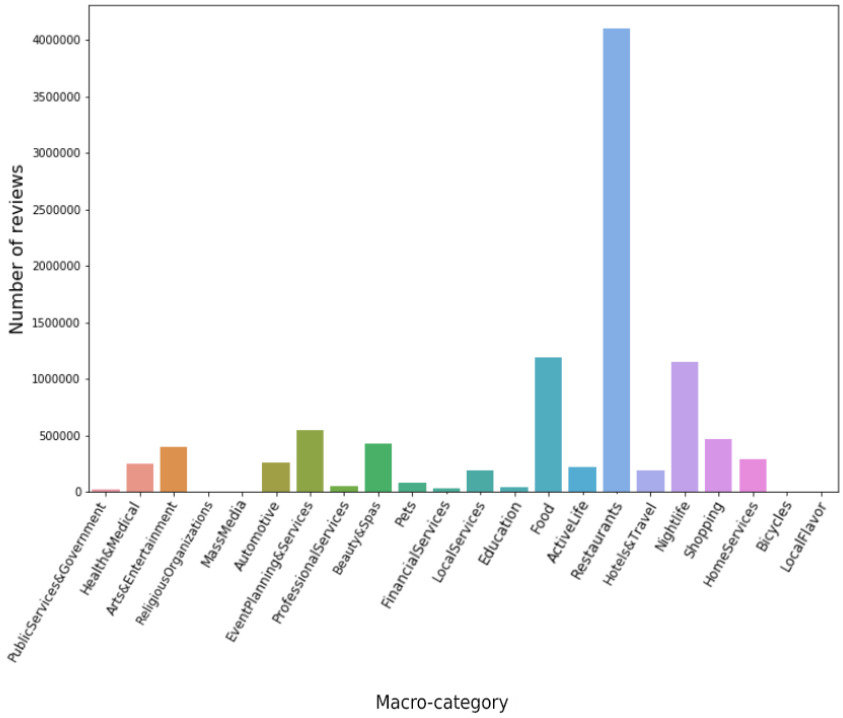


Fig. 3.12: Distribution of the reviews of Yelp users against the Yelp macro-categories

Once again, we are interested in investigating the co-review mechanism and the role of k-bridges as possible pioneers in this context. In order to carry out this study, we created a new network, which we call "macro-category network" and denote it with $\mathcal{M} = \langle N, E \rangle$. N represents the set of nodes of $\mathcal{M}$. In particular, there is a node $n_j \in N$ for each macro-category $\mathcal{Y}_j$ in Yelp. E is the set of edges of $\mathcal{M}$; in particular, there is an edge $e_{jh} \in E$ if both the macro-categories $\mathcal{Y}_j$ and $\mathcal{Y}_h$ have been reviewed by a fraction of users greater than or equal to a threshold $X\%$. Clearly, as X varies, we have different networks $\mathcal{M}^{X\%}$. Based on these definitions, we constructed the networks $\mathcal{M}^{1\%}$, $\mathcal{M}^{5\%}$, $\mathcal{M}^{10\%}$ and $\mathcal{M}^{15\%}$. These are shown in Figures 3.13 - 3.16.

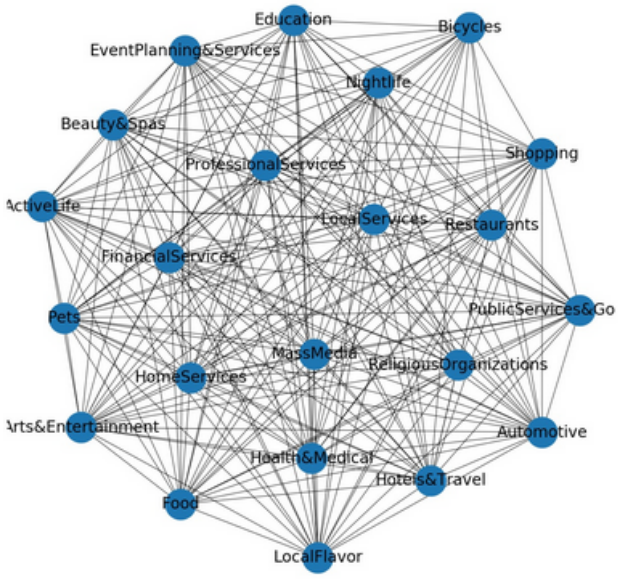


Fig. 3.13: The network $\mathcal{M}^{1\%}$

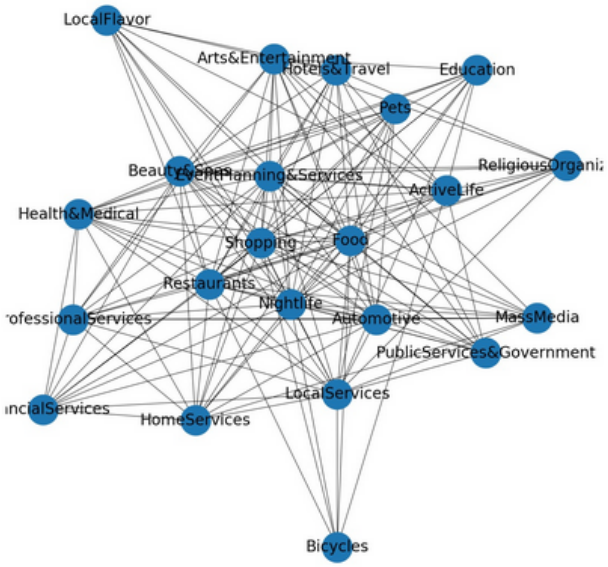


Fig. 3.14: The network $\mathcal{M}^{5\%}$

The corresponding density and average clustering coefficient are reported in Table 3.9. Figures 3.17 and 3.18 present the variation of the values of the density and the average clustering coefficient when X increases. As shown in these figures, it is very likely to find two macro-categories that are co-reviewed by a small number of users. In fact, 98.1% of the possible combinations of categories are co-reviewed by at least 1% of the users. However, if we are more demanding on the fraction of users

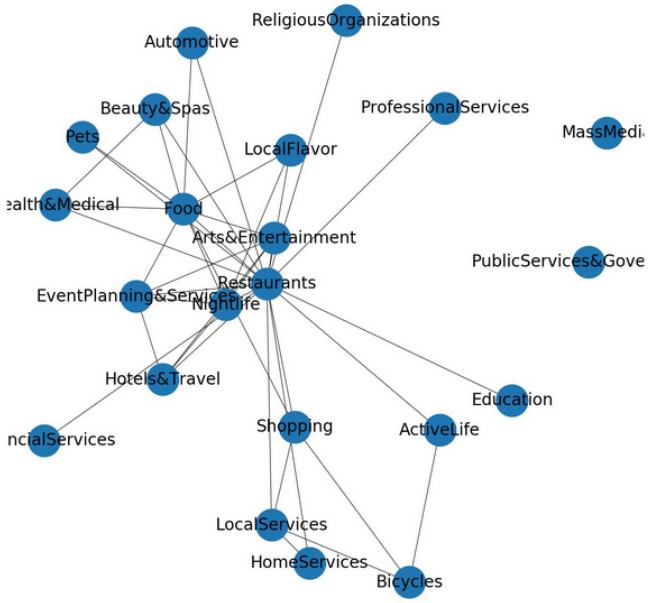


Fig. 3.15: The network $\mathcal{M}^{10\%}$

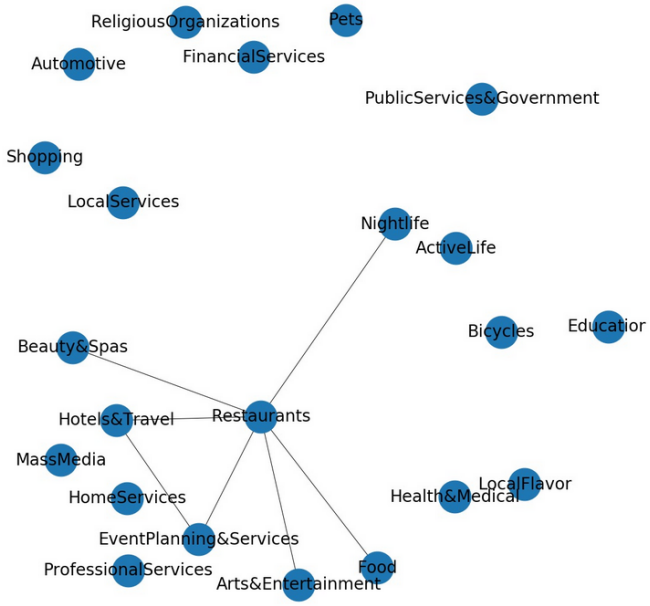


Fig. 3.16: The network $\mathcal{M}^{15\%}$

that co-review the same macro-category, we can see from the figures that the trend of co-reviews varies rapidly. In fact, even if the possible combinations of co-reviewed macro-categories is quite high with at least 5% of co-reviewing users, this number decreases rapidly when we further increase the value of X.

Table 3.10 shows the maximum and sub-maximum values of the degree centrality for the networks of Figures 3.13 - 3.16, along with the macro-categories

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Density	0.978	0.680	0.173	0.030
Average Clustering Coefficient	0.981	0.833	0.514	0.094

Table 3.9: Values of the density and the average clustering coefficient for the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

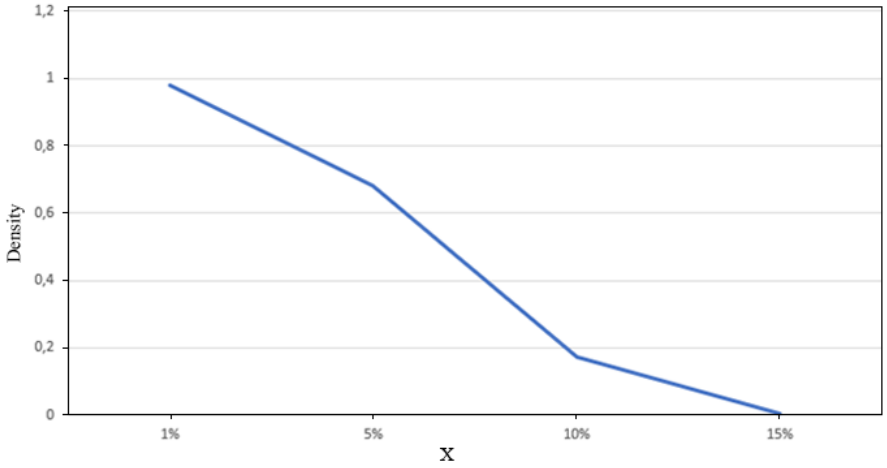


Fig. 3.17: Variation of the density of the macro-category networks $\mathcal{M}^{X\%}$ against the increase of X

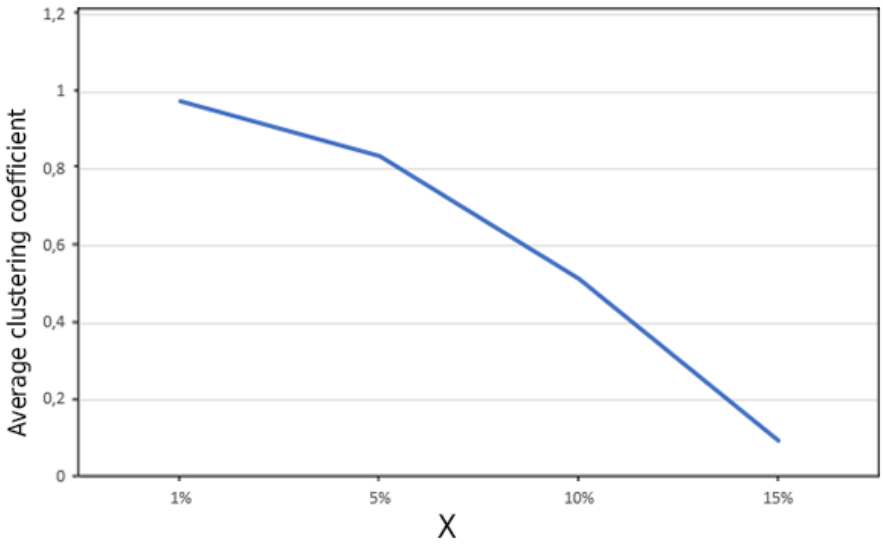


Fig. 3.18: Variation of the average clustering coefficient of the macro-category networks $\mathcal{M}^{X\%}$ against the increase of X

which they refer to. The objective is to identify which macro-categories tend to have more co-reviews with other ones. From the analysis of this table we can observe that the two macro-categories most present with maximum or sub-maximum values are “Restaurants” and “Food”. Actually, this result was quite obvious, given

the distribution of the reviews in Yelp (see Figure 3.12). Instead, the fact that the macro-categories "Beauty&Spas" and "Hotels&Travel" are present as maximum or sub-maximum is particularly interesting. In fact, these two macro-categories have a much lower number of reviews not only than "Restaurants" and "Food" but also than several other macro-categories not present in Table 3.10.

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Maximum value and associated macro-category	1 (Beauty&Spas)	1 (Food)	0.857 (Restaurants)	0.286 (Restaurants)
Sub-maximum value and associated macro-category	1 (Food)	1 (Nightlife)	0.476 (Food)	0.095 (Hotels&Travel)

Table 3.10: Maximum and sub-maximum values of degree centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

Table 3.11 shows the maximum and sub-maximum values of the closeness centrality for the networks of Figures 3.13 - 3.16. We do not present this table for the semantics of closeness centrality in this application context. Instead, we want to highlight that, unlikely what generally happens in Social Network Analysis, where the nodes having the highest degree centrality and the highest closeness centrality are generally different [647], the macro-categories that have the highest values of closeness centrality are exactly the same as the ones having the highest values of degree centrality.

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Maximum value and associated macro-category	1 (Beauty&Spas)	1 (Food)	0.86 (Restaurants)	0.286 (Restaurants)
Sub-maximum value and associated macro-category	1 (Food)	1 (Nightlife)	0.614 (Food)	0.171 (Hotels&Travel)

Table 3.11: Maximum and sub-maximum values of closeness centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

Table 3.12 shows the maximum and sub-maximum values of the betweenness centrality for the networks of Figures 3.13 - 3.16. As we can notice, in $\mathcal{M}^{1\%}$ all the values of the betweenness centrality are very low. This is not surprising because this network is almost totally connected. The maximum and sub-maximum values of the betweenness centrality grow, albeit slightly, in $\mathcal{M}^{5\%}$. Once again, this is understandable because, if we look at Figure 3.14, we can see that this network is still very connected. The most interesting situation for this kind of centrality happens in $\mathcal{M}^{10\%}$. In fact, in this case, we have that the maximum and sub-maximum values

of betweenness centrality are high. These values are associated with "Restaurants" and "Food". Now, looking at Figure 3.14, we can see how "Restaurants" and "Food" are actually two nodes from which we must pass to go from a node located in the top sub-net to a node located in the bottom one. Finally, as far as the betweenness centrality is concerned, the network $\mathcal{M}^{15\%}$ is not very significant, since it is almost completely disconnected.

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Maximum value and associated macro-category	0.001 (Arts&Entertainment)	0.049 (Food)	0.627 (Restaurants)	0.067 (Restaurants)
Sub-maximum value and associated macro-category	0.001 (LocalServices)	0.049 (Nightlife)	0.614 (Food)	0 (Beauty&Spas)

Table 3.12: Maximum and sub-maximum values of betweenness centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

Table 3.13 shows the maximum and sub-maximum values of the eigenvector centrality for the networks of Figures 3.13 - 3.16. We can observe that the maximum and sub-maximum values correspond to those of the degree centrality and the closeness centrality. Once again the two macro-categories with the highest values are "Restaurants" and "Food".

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Maximum value and associated macro-category	0.217 (Arts&Entertainment)	0.279 (Food)	0.525 (Restaurants)	0.665 (Restaurants)
Sub-maximum value and associated macro-category	0.217 (LocalServices)	0.279 (Nightlife)	0.397 (Food)	0.395 (Hotels&Travel)

Table 3.13: Maximum and sub-maximum values of eigenvector centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

The analysis of the distributions and the ones of all the different forms of centrality show that "Restaurants" is an extremely dominant macro-category. Therefore, it is interesting to verify whether or not most of the properties we have previously found depend exclusively on "Restaurants".

To perform this verification, we removed all references to the macro-category "Restaurants" from the reviews. Then, we computed again the number of k-bridges and the distribution of users. In particular, the number of k-bridges decreased from 1,106,727 to 813,146, while the number of non-bridges increased from 530,411 to 823,992.

The distribution of users is shown in Figure 3.19. From the analysis of this figure, we can observe that, in this case, the distribution follows a much steeper power law. This is understandable because those nodes that were previously non-bridges continue to be so now. At the same time, all the nodes that were previously 2-bridges and that referred to "Restaurants" become non-bridges. More in general, all nodes that where k-bridges ($k \geq 2$) and referred to "Restaurants" become $(k-1)$-bridges.

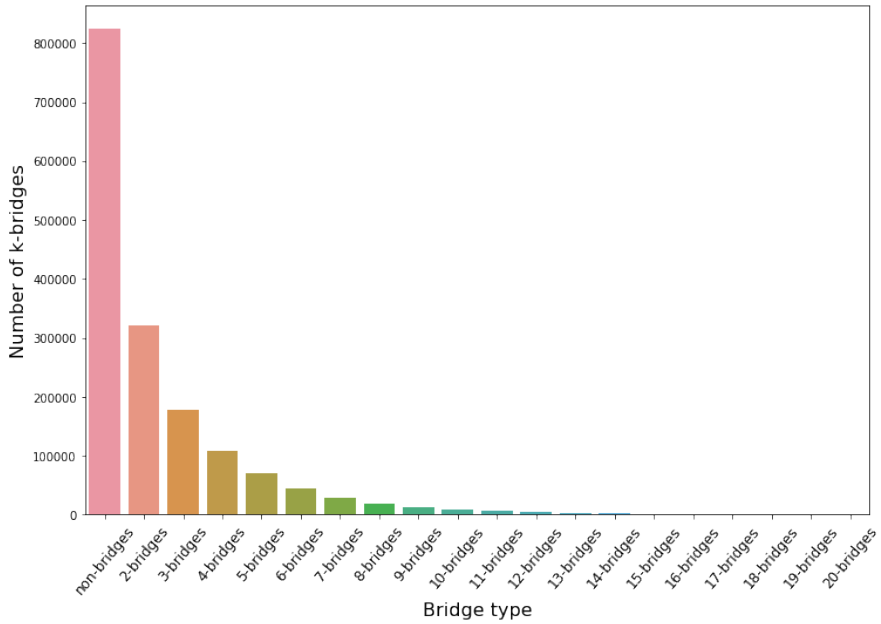


Fig. 3.19: Distribution of the k-bridges against k in Yelp after the removal of "Restaurants"

Then, we computed again the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$. They are shown in Figure 3.20. From the analysis of this figure, we can observe that the connection level of these networks slightly decreases compared to the corresponding networks with "Restaurants", albeit this trend remains the same from a qualitative viewpoint. This can also be deduced from the values of the density and the average clustering coefficient shown in Table 3.14.

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Density	0.976	0.719	0.176	0.024
Average Clustering Coefficient	0.979	0.846	0.452	0

Table 3.14: Values of the density and the average clustering coefficient for the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$ after the removal of "Restaurants"

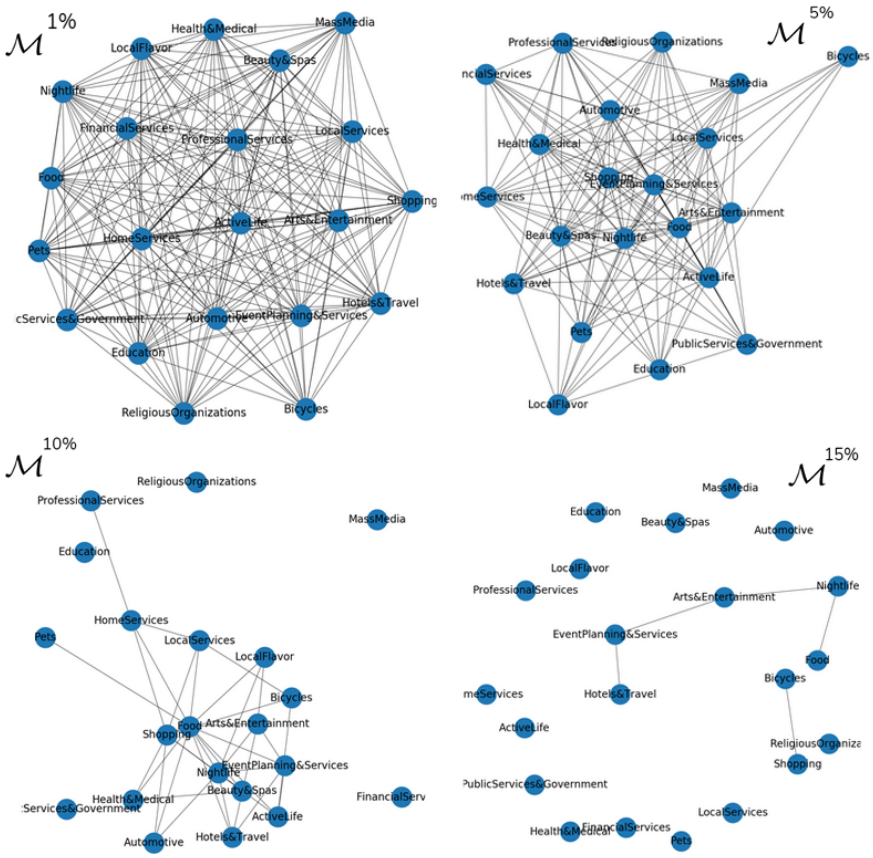


Fig. 3.20: The networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$ after the removal of "Restaurants"

Finally, we computed the maximum and sub-maximum values for all centrality measures for the new networks obtained after the removal of "Restaurants". The results are reported in Table 3.15. From the analysis of this table, we can observe that the values are slightly lower than before, but the trend is confirmed. This allows us to conclude that the trends and features related to co-reviews in Yelp are intrinsic to this social medium and are not biased by the presence of "Restaurants". This macro-category certainly contributes to strengthen these trends but it does not upset them.

Clearly, in absence of "Restaurants", the macro-category that plays the main role in the co-reviews is "Food". Instead, different macro-categories often alternate in the role of sub-maximum for the centrality measures into consideration.

After having performed a deep analysis on the features of k-bridges in Yelp, in the following section, we verify if some results on k-bridges found in this social network are general or specific to it.

3.1.4.2 Validation of k-bridge properties in other networks

This section is devoted to validating the k-bridge properties mentioned above in other networks. Actually, due to space constraints, we limit our analysis to only some

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
Maximum Degree Centrality	1 (Beauty&Spas)	1 (Food)	0.65 (Food)	0.1 (Nightlife)
Sub-maximum Degree Centrality	1 (Food)	1 (Nightlife)	0.45 (Nightlife)	0.1 (EventPlanning&Services)
Maximum Closeness Centrality	1 (Beauty&Spas)	1 (Food)	0.662 (Food)	0.133 (Arts&Entertainment)
Sub-maximum Closeness Centrality	1 (Food)	1 (Nightlife)	0.511 (Shopping)	0.114 (EventPlanning&Services)
Maximum Betweenness Centrality	0.002 (Beauty&Spas)	0.044 (Food)	0.271 Food)	0.021 (Arts&Entertainment)
Sub-maximum Betweenness Centrality	0.002 (Food)	0.044 (Nightlife)	0.074 (HomeServices)	0.016 (Nightlife)
Maximum Eigenvector Centrality	0.223 (Beauty&Spas)	0.273 (Shopping)	0.49 (Food)	0.577 (Arts&Entertainment)
Sub-maximum Eigenvector Centrality	0.223 (Food)	0.273 (Nightlife)	0.403 (Nightlife)	0.5 (Nightlife)

Table 3.15: Maximum and sub-maximum values of the various centrality measures and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$ after the removal of "Restaurants"

of the properties found above. We verify their validity first in Reddit and, then, in the network of patent inventors.

Validation of k-bridge properties in Reddit

We downloaded all the data for the investigation activity from the `pushshift.io` website, one of the most known Reddit data sources. Our dataset contains all the posts published on Reddit from January 1^{st}, 2019 to February 1^{st}, 2019. The number of posts available for our investigation was 485,623.

As a first task, we selected the 30 subreddits with the highest number of posts. According to our model, as described in Section 3.1.3.1, all the authors of a subreddit represented a community in our model, and the authors who submitted one or more posts in at least two subreddits represented bridges. Specifically, a k-bridge is an author who posted in exactly k subreddits.

As a first experiment, we computed the distribution of k-bridges against k in Reddit. It is shown in Figure 3.21. From the analysis of this figure, we can see that it follows a power law. This result is in total agreement with the one obtained for Yelp and reported in Figure 3.3.

As a second experiment, we considered the co-posting network $\mathcal{U}^{cp}$, defined in Section 3.1.3.1. We recall that, in this network, there is a node for each user who submitted at least one post in at least one of the 30 subreddits into consideration, and there is an arc between two users if both of them contributed to the same subreddit. The co-posting network in Reddit corresponds to the co-review network in Yelp. In that case, we had found that there is a backbone among the bridges of this network. Therefore, it appears interesting to verify whether this property exists also in $\mathcal{U}^{cp}$.

For this purpose, for each bridge (non-bridge), we considered the fraction of co-posters that were bridges (non-bridges). The results obtained are shown in Table 3.16. They denote that there is a backbone among bridges in $\mathcal{U}^{cp}$. They also confirm what we had obtained for Yelp in Table 3.6.

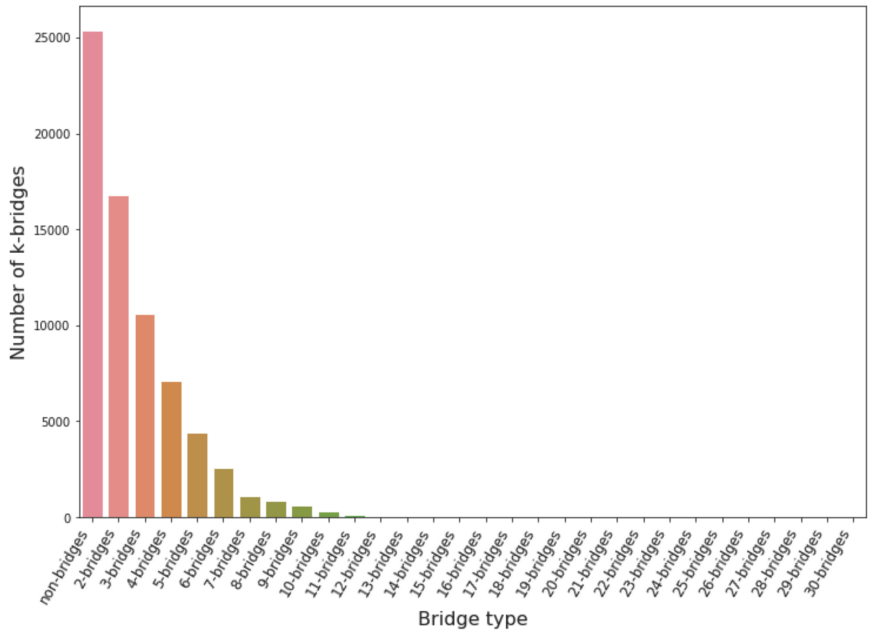


Fig. 3.21: Distribution of the k-bridges against k in Reddit

	Fraction of co-posters that are bridges	**Fraction of co-posters that are non-bridges**
Bridges	0.9234	0.0585
Non-bridges	0.7531	0.2243

Table 3.16: Types of co-posters for bridges and non-bridges in $\mathcal{U}^{cp}$

Finally, we verified if there is a correlation between k-bridges and power users. For this purpose, we computed the distribution of the number of arcs for non-bridges, bridges, strong and very strong bridges. Preliminarily, by applying the same approach described in Section 3.1.3.2 for Yelp, we found that, in Reddit, the thresholds for strong bridges and very strong bridges are $th_s = 5$ and $th_{vs} = 9$, respectively.

Afterwards, we computed the distribution of the number of arcs for non-bridges, bridges, strong and very strong bridges. The results obtained are shown in Figure 3.22. This figure reveals that, as k grows, the power law distributions move to the right and flatten out. This result confirms the one in Figure 3.9 obtained for Yelp and tells us that also for Reddit there is a correlation between the strength of k-bridges and their degree centrality.

Validation of k-bridge properties in the network of patent inventors

Data about patents adopted in our analyses has been taken from the PATSTAT-ICRIOS database. It stores data about all patents from 1978 to the current years coming from about 90 patent offices worldwide. The number of patents taken into consideration is 9,605,147 and the number of inventors is, instead, 23,637,883.

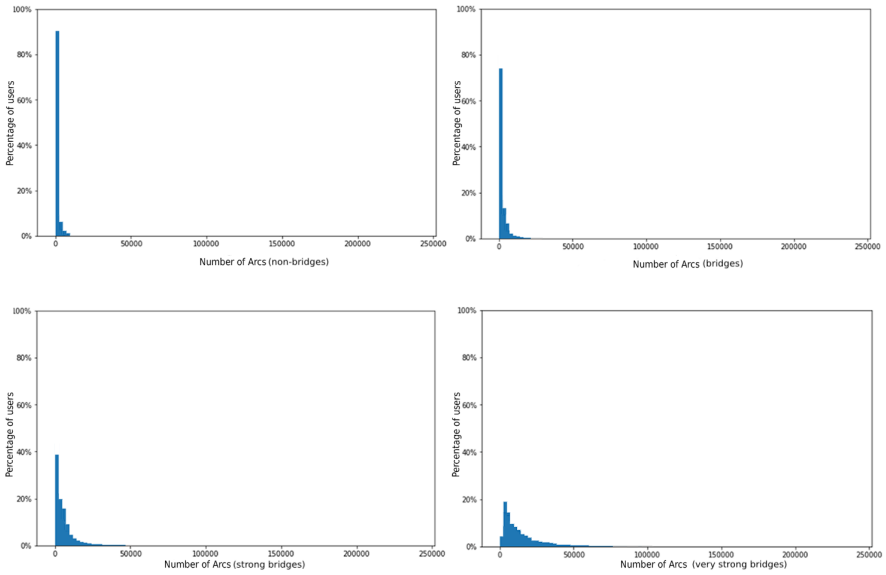


Fig. 3.22: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges in Reddit

According to our model, as described in Section 3.1.3.1, the set of inventors who filed at least one patent in an IPC class represents a community. Therefore, we have 127 communities. In this setting, the authors who filed patents in at least two IPC classes represent bridges. A k-bridge is an author who filed patents that, in the whole, cover exactly k IPC classes.

Also in this case, we computed the distribution of k-bridges against k. We report it in Figure 3.23. From the analysis of this figure, we can see that it follows a power law. This result is in line with what we have seen for Yelp and Reddit.

After this, we considered the co-inventing network $\mathcal{U}^{ci}$, defined in Section 3.1.3.1. Here, there is a node for each inventor and there is an arc between two inventors if both of them filed at least one patent together. Clearly, the co-inventing network strictly corresponds to the co-posting network of Reddit and the co-review network of Yelp.

In order to verify if there exists a backbone among the bridges of this network, for each bridge (resp., non-bridge), we considered the fraction of co-inventors that were bridges (resp., non-bridges). The results, reported in Table 3.17, clearly denote the existence of a backbone among the bridges in $\mathcal{U}^{ci}$, analogous to the ones found in $\mathcal{U}^{cr}$ for Yelp and in $\mathcal{U}^{cp}$ for Reddit.

Finally, we verified if there is a correlation between k-bridges and power users also in $\mathcal{U}^{ci}$. In this case, a reasoning analogous to the one described in Section 3.1.3.2 allowed us to find that, in the network of patent inventors, the threshold th_s for strong bridges is 5 whereas the threshold th_{vs} for very strong bridges is 10.

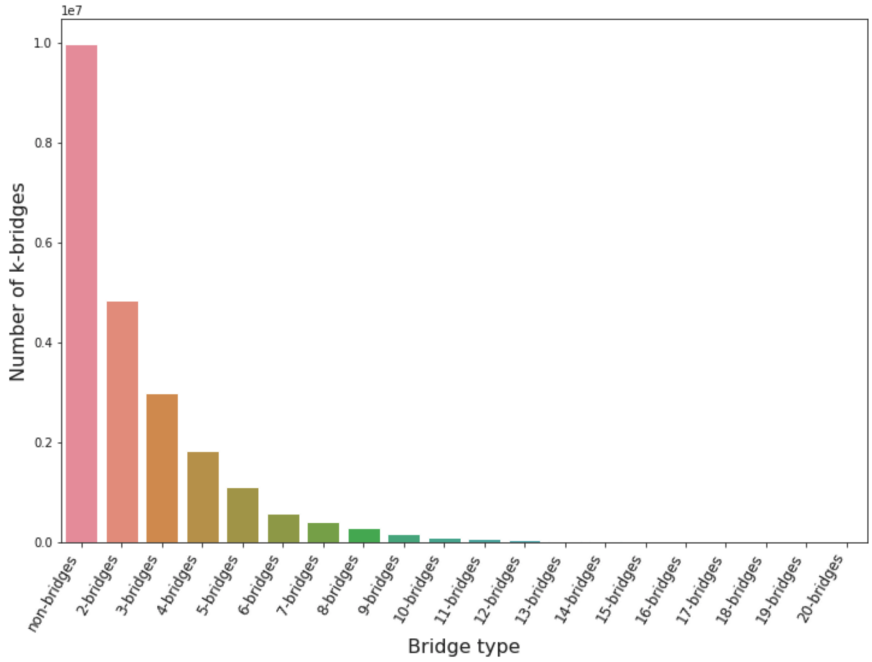


Fig. 3.23: Distribution of the k-bridges against k in the network of patent inventors

	Fraction of co-inventors that are bridges	**Fraction of co-inventors that are non-bridges**
Bridges	0.9632	0.0563
Non-bridges	0.7924	0.2356

Table 3.17: Types of co-inventors for bridges and non-bridges in $\mathcal{U}^{ci}$

We computed the distribution of the number of arcs for non-bridges, bridges, strong and very strong bridges. The results are reported in Figure 3.24. They denote that, as k grows, the power law distributions move to the right and flatten out. This result is a further confirmation of the ones reported in Figure 3.9 for Yelp and in Figure 3.22 for Reddit, i.e., that also in the network of patent inventors there is a correlation between the strength of k-bridges and the degree centrality.

After having verified that the main properties of k-bridges are intrinsic to this concept and not specific to only Yelp, in the next section, we present two use cases that could highly benefit from the knowledge of k-bridges.

3.1.4.3 Applications of k-bridges

The social networking phenomenon has completely changed the way people conceive interaction with each other and consume information. Several studies have investigated the consequences of the massive proliferation of Online Social Networks that we are observing in these years.

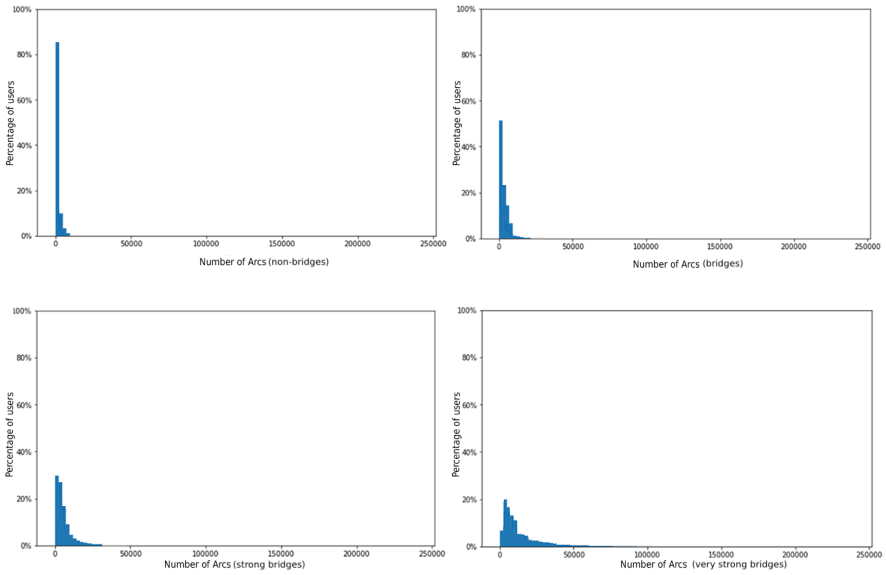


Fig. 3.24: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges in the network of patent inventors

From a consumer point of view, social networks bring impressive benefits, such as richer and more participative information, a broader selection of products, more competitive pricing, and cost reduction. Instead, in the industry context, 81% of firms plan to invest in social networking sites, and more than 50% of them consider digital advertising and marketing as a priority area of investment [640]. Actually, several online services, like Yelp (but also TripAdvisor[3], and, in a certain sense, Booking[4], Airbnb[5], etc.), have been conceived just to encourage this kind of interaction. Of course, in this scenario, obtaining a very large number of positive reviews is crucial for businesses. Therefore, designing ad-hoc marketing and advertising campaigns is extremely important. In the next paragraphs, we describe in detail two case studies related to this concept, which massively exploit k-bridges to conduct marketing campaigns and support business decisions in Yelp.

Finding the best targets for a marketing campaign

This first case study refers to a scenario in which a business is planning to expand its activities including services that belong to new Yelp categories, along the ones already covered. The business already performed an internal evaluation analysis with the goal of identifying the best services, possibly referring to new categories, to improve its revenues. The next step concerns the design of a goal-oriented marketing

[3] `https://www.tripadvisor.com`
[4] `https://www.booking.com`
[5] `https://www.airbnb.com`

campaign to foster the diffusion of the new services among new potential customers. Of course, a naive flooding approach of advertising messages appears not convenient, as it would not be possible to properly target the advertising campaign based on customer features. Moreover, it would lead to an excessive amount of unwanted messages from a user point of view.

For these reasons, the knowledge derived from the identification of k-bridges, who are already customers of both the original categories of interest for the business and the new ones it intends to embrace, plays a crucial role. Indeed, these bridges can be considered as links among the different communities they belong to and, hence, they can be "engaged" as convenient diffusion points to properly target the marketing campaign.

Now, let us consider a simple example scenario where a business, which already provides services belonging to the *Restaurant* category of Yelp, decides to include new services belonging to two new related categories, namely *Nightlife* and *Hotel&Travel*. In this case, according to the reasoning above, the following steps can be performed to obtain a very effective marketing campaign.

First, 3-bridges are identified as the most correct typology of users to involve. Indeed, 3-bridges can potentially link together all and only the three categories of interest. Actually, more powerful bridges (e.g., 4-bridges or higher) could have been also considered; however, this would lead to the inclusion of other categories not interesting for the business, which in turn would lead to a reduction of the campaign effectiveness.

After that, among all the available 3-bridges, the ones belonging to just the three categories of interest are selected.

Now, considering that the campaign success strongly depends on the capability of k-bridges to promote the new services, a metric to measure it must be introduced. This metric should consider the inclination of a bridge to review businesses, her proneness to create an articulated friend network, and her constant activity level over time. In Equation 3.1, we report a possible simple implementation of such a metric (clearly, future research efforts could be made to define a more sophisticated metric):

$$\mu_i = \frac{nr_i \cdot nf_i}{nd_i} \tag{3.1}$$

Here, nr_i represents the number of reviews performed by the 3-bridge u_i, nf_i denotes the dimension of the network of her friends, and, finally, nd_i indicates the number of days u_i is enrolled in the platform. Here, nr_i directly measures the activity level of u_i; however, this is not sufficient because early adopters of the platform typically make a very high number of reviews in a very short amount of time, but not

all of them remain active over time. For this reason, we consider two other important factors, i.e., the number of friends and the time interval in which they performed their activities. As the creation of a strong and rich network of friends requires time, nf_i allows us to exclude early adopters who left the platform too soon. Instead, nd_i acts as a weight and allows the estimation of the real activity level over time.

Now, the business can use the metric above to sort the set of 3-bridges according to their capability of promoting its services. Finally, it selects the top bridges as the target for its marketing campaign. The fact that the selected 3-bridges are members of all the three categories of interest increases the possibility that they can help the business to be known in the new communities.

The solution above, sketched for the simple example considered, can be easily extended and generalized for any similar application scenario with any number of involved categories. The overall process is described by Algorithm 2.

Input

- D, a dataset of a Social Network
- k, the number of communities of interest for the marketing campaign

Output

- $\overline{B_k}$, the k-bridges to consider for the marketing campaign

Require: `getInfo`(u_i), a function returning a DataFrame containing information about the number of reviews, the number of friends, and the days of enrollment in the platform of a user u_i; `bridgeExtraction(k)`, a function implementing Algorithm 1 and returning the set of k-bridges; S_k, a set of scores

B_k = `bridgeExtraction(k)`
for $u_i \in B_k$ **do**
  $info_{u_i}$ = `getInfo`(u_i)
  $nr_i = info_{u_i}[\text{“}reviews\text{”}]$, $nf_i = info_{u_i}[\text{“}friends\text{”}]$, $nd_i = info_{u_i}[\text{“}days\text{”}]$
  $\mu_i = (nr_i \cdot nf_i)/nd_i$
  add μ_i to S_k
end for
$\overline{B_k}$ = sort B_k by S_k
return $\overline{B_k}$

Algorithm 2: Algorithm for finding the best targets of a marketing campaign

Finding new products/services to propose

This second case study is strictly related to the previous one. However, it deals with a situation in which a business is still conducting a market analysis to identify new

services, belonging to new categories, that it can propose. In this context, the knowledge acquired by analyzing k-bridges can be used to know the most popular categories related to the ones already covered by the business. Indeed, in this scenario, the review activities of k-bridges implicitly encode association rules among categories. Such rules can be represented as:

$$review(\mathcal{C}_k) \Rightarrow \bigwedge_{i=1}^{k-1} review(\mathcal{C}_i)$$

Here, the term $\bigwedge_{i=1}^{k-1} review(\mathcal{C}_i)$ represents the logic conjunction of a sequence of reviewing activities in $k-1$ different categories.

Intuitively, the larger k the more disparate are the different categories included in the conjunction. For this reason, it is first necessary to identify the optimal value of k in the extraction of meaningful association rules among categories. For this purpose, it is possible to adopt a modified version of the Elbow-method [377], a very common strategy to identify the correct number of clusters in a typical clustering scenario. The basic idea underlying our approach to perform this task is to carry out an iterative task. At each iteration:

1. the value of k is increased;
2. Algorithm 2 is used to identify k-bridges;
3. k-bridges being members of the original category of the business are selected;
4. all the additional categories (involved by the identified k-bridges) are considered;
5. their average semantic distance with respect to the starting ones is estimated.

This procedure ends when, during an iteration, the average estimated distance for the new categories is considered too high with respect to the marketing objectives of the business.

At this point, by analyzing the k-bridges involving the original categories and the closest ones identified during the iterations, it is possible to identify a set of association rules between the original categories of the business and the new ones. For each rule, it is possible to estimate the corresponding *support* and *confidence*[6]. The obtained information can be used by the business to decide which new categories are more suitable for its development.

[6] Observe that, borrowing some ideas from the association rules theory, in our scenario, support can be defined as a measure of how frequently the new categories and the old ones appear together in k-bridges; instead, confidence quantifies how often the new categories appear in those k-bridges where the original categories appear too.

3.2 Investigating negative reviews and negative influencers

3.2.1 Introduction

Yelp[7] is a business directory service and a crowd-sourced platform designed to help users find businesses like restaurants, hotels, pet stores, spas, and many more. It is one of the most widely used review platforms on the Web. It ranks 9^{th} on the RankRanger list of the top 100 leading websites by traffic[8], with approximately 800 million visits per month. In addition of being a business search and review platform, Yelp is also a social network, because it allows its users to specify their friendships. Finally, it is also a business directory, because it groups businesses into categories and sub-categories.

The success of Yelp has prompted many researchers to investigate this platform [16, 64, 425, 515, 311].

A phenomenon that represents a hot topic for both Yelp and all review platforms is the analysis of negative reviews [94]. This topic is extremely important not only for the consequences it has in practice, but also from a more theoretical point of view. In fact, it is well known that the Likert scale, which the Yelp reviews and the corresponding scores are based on, is positively biased [41, 537, 104]. As a consequence, the presence of negative reviews is a really important problem indicator for a business and, consequently, a valuable piece of information [398, 418]. Indeed, negative reviews can provide much more information, knowledge and improvement possibilities than positive ones [178]. For this reason, many researchers have already investigated the role of ratings and reviews on businesses, along with their social implications [642, 443].

Despite the numerous studies on Yelp that have been presented in the past literature, to the best of our knowledge, no paper has proposed a multi-dimensional model capable of best capturing the specificity of Yelp to be at the same time a review platform, a social network and a business directory. Moreover, no paper has proposed a study focused entirely on negative reviews on Yelp that, starting from a representative model of them, could define several stereotypes of users and, hence, build the profile of negative influencers. For this reason, we aim at filling this gap.

Specifically, we first define a multi-dimensional social network-based model for Yelp and then use this model to study negative reviews and build a profile of negative influencers in this social medium. We decided to adopt this model because it perfectly fits the specificities of Yelp mentioned above. In fact, our model represents Yelp as a set of 22 communities, one for each macro-category of this social plat-

[7] https://www.yelp.com

[8] https://www.rankranger.com/top-websites

form (modeling Yelp as a business directory). At the same time, it represents Yelp as a social network, whose nodes indicate users and whose arcs denote the relationships between them. These can be of different types. For example, they can denote friendships between users (modeling Yelp as a social network), or the action of co-reviewing the same business (modeling Yelp as a review platform). Through the concepts and techniques of Social Network Analysis applied to our multi-dimensional model, our approach defines three stereotypes of Yelp users, namely the bridges, the double-life users and the power users. These stereotypes can help the detection of the negative influencers in Yelp and the definition of a profile for them. Both our model and the user stereotypes represent our theoretical contributions. These last are completed by a Negative Reviewer Network, which allows us to investigate the main characteristics of the negative influencers in Yelp.

Among the possible questions that can be answered thanks to our approach, here we focus on the following ones: *(i)* What about the dynamics leading a Yelp user to publish a negative review? *(ii)* How can the interaction of these dynamics increase the "power" of negative reviews and people making them? *(iii)* Who are the negative influencers in Yelp?

The practical implications of negative reviews and influencers have a large variety of real-world applications. First of all, it was proved that negative reviews have a stronger effect on businesses than positive ones [18]. Furthermore, influencers play a crucial role for the successful placement of products in a social network. So, it is important to know who are the negative influencers that could damage a business, in order to strive to turn them into neutral, or even positive, influencers [703, 714]. Finally, gaining trust through online reviews can help a business gather venture capitals for its growth [266, 398]. As a matter of fact, reviews are consumer opinions, unfiltered by traditional media, more sincere and imperfect [18, 192]. For this reason, a proper coverage of positive reviews can attract more financiers [18, 193, 385]. On the other hand, negative reviews and influencers can drive potential investors away from investing in a company [445].

The outline of this chapter is as follows. In Section 3.2.2, we present related literature and highlight the main novelties of our approach with respect to the past ones. In Section 3.2.3, we describe the Yelp model, the stereotypes of negative influencers and develop five hypotheses to verify. In Section 3.2.4, we investigate the correctness of the Hypothesis H1-H5. Finally, in Section 3.2.5, we propose a discussion and a synthesis of them, their real-world implications.

3.2.2 Related Literature

Over the years, researchers have focused on Yelp as a reference platform for studying how users interact with each other and build cooperative social groups. Their research efforts have also been supported by the social medium itself, which has made available a complete snapshot of its data to foster comprehensive analyses on it [211]. Many authors have used this snapshot to investigate the role of ratings and reviews on businesses and their social implications [642, 443]. Researchers have also analyzed how people search for information on Yelp [328] and what aspects (including uses and rewards) lead them to employ this platform.

Several authors have investigated Yelp using Social Network Analysis (SNA, for short) [555, 556]. For instance, the authors of [556] rely on the concept of homophily [468] to study the social influence possibly existing between users and, in particular, between friends. Starting from the results obtained, they propose the construction of the profile of an influencer in Yelp. The authors of [555] focus on the role of friendship in this social medium. Specifically, they investigate the impact of social relationships from the consumer's side and find that these relationships exert a significant impact in those consumers having at least one common purchase.

As for the analysis of social relationships, several studies have been conducted in both Yelp and other social platforms to understand how users perceive their social contacts and how they influence their acquaintances [425, 515, 311, 494, 604, 363, 725, 718]. For example, the authors of [494] propose an approach to analyze a large set of brand associations obtained from social tags for marketing research. They apply well-known text mining techniques to understand consumers' perceptions of brands starting from social tagging data. The authors of [192] analyze a dataset obtained from `OpenRice.com`, a crowd-sourced social medium for restaurant reviews in Hong Kong and Macau. The authors of [270] show that online community members rate reviews containing descriptive identity information more positively. Indeed, a disclosure of personal information on an online review system leads to a greater volume of sales. The authors of [604] aim at understanding how online reviewers compete to acquire the attention, typically scarce, of users. They propose a theory explaining the strategies adopted by online reviewers in choosing the right product and the right rate when posting reviews. As far as Yelp is concerned, the authors of [425] investigate the effects of the review rate, the reviewer profile, and the receiver familiarity with the platform, on the credibility of a review on this social medium. Moreover, the authors of [515] find a strong correlation between the moral attitude of a community of users and their tendency to express low rates and negative reviews in case some moral foundation is violated. As for the investigations of social relationships in social media, another interesting topic concerns information diffu-

sion [66, 690, 382, 109, 428]. In the analysis of this topic, an increasing number of researchers are studying the role not only of classic and direct relationships, such as friendship, but also several other ones, such as co-posting or homophily of interests (i.e., having interest in the same topics) [588, 107].

In all previous approaches, the reviews considered are general (i.e., they could be positive or negative). However, to our end, negative reviews and reviewers are worth a special attention. The importance of negative reviews in the analysis of social platforms has been investigated in the recent scientific literature by highlighting their impact in social contexts, along with the mechanisms leading users to make them [493, 266, 595, 64, 16]. In these studies, researchers point out that dealing with negative reviews is a fundamental task in review-based platforms for business operators [398, 418]. In fact, it was empirically shown that answers and justifications to negative rates contribute to the increase of trust between users and businesses [266], and that users tend to perceive reviews confirming their initial beliefs as more helpful [703]. Several studies focus on the key factors making a review helpful [593, 266], while others show that negative reviews are more useful and can influence user opinions more than positive ones [87, 152]. In this perspective, the authors of [714] propose a model to identify the key elements leading customers to make their decisions; this model was empirically tested with 191 users of an existing online review site. Furthermore, the authors of [18] use the VentureExpert database to gain knowledge on a sample of famous businesses. The authors of [333] formalize a metric, called disconfirmation, measuring the discrepancy between the expected evaluation of a product and the one assigned by experts or other people. The authors of [266] study a set of variables to evaluate the users' intention of employing Yelp, as well as their behavior in using a service or purchasing a product after reading Yelp reviews. Finally, the authors of [64] analyze the reviews made by hospital patients in order to identify a common language correlated with negative and positive reviews.

An important aspect to consider when using Social Network Analysis for evaluating reviews and reviewers is the fact that user relationships in a social network are often heterogeneous [147]. For this reason, many studies have proposed to decompose social media into different networks of relationships. Indeed, multi-relationship networks have been extensively studied in the past [223, 697, 719]. For example, the authors of [719] combine the analysis of the friendship network and the author-topic one, both constructed starting from the information available in an online social network. Instead, the authors of [697] focus on a co-authorship network and consider different types of relationships, i.e., co-authorship, co-participation to the same edition of a conference, and geographic proximity.

In multi-relationship networks, the classical definition of influencer is extended because the role of such users is not bound to communities derived from a single category of relationships. Instead, it also includes the capability of providing information diffusion channels among different networks, one for each type of relationships. To refer to this extended definition of influencer, the term "bridge" is often adopted. In the past literature, several studies have been devoted to investigating the role of bridges in the formation of social communities. For instance, the authors of [371] show that users with a weak connection bridging heterogeneous groups have higher levels of community commitment, civic interest, and collective attention than the other ones. Furthermore, the authors of [298] prove that Internet users, who bridge heterogeneous online communities by means of weak ties, have a high social engagement, use the Internet for social purposes, and are prone to become members of new social communities. The interest towards users serving as bridges among communities has increased over the years and, indeed, several studies have been done to analyze the behavior and peculiarities of such users in complex networks [279, 606, 416, 95, 98].

Some studies have also analyzed the behavior of users serving as bridges among different social networks [134, 141, 136]. Here, the concept of community is brought to the edge, because it is mapped to a whole social network. Specifically, the authors of [134] report a complete identikit of users bridging different social networks. The authors of [141] leverage the peculiarities of bridge users to define a new crawling strategy to sample a multi-social network environment. Finally, the authors of [136] perform a comparative study of users serving as bridges among two of the most famous social networks, namely Facebook and Twitter.

From the above description, it can be seen that, in the literature, there is an impressive number of papers dealing with issues similar to those analyzed here. However, none of them proposed a multi-dimensional social network-based model for Yelp, capable of representing the specificity of this social platform of being simultaneously a review platform, a social network and a business directory. The presence of this model would allow us to answer the following research question: What about the dynamics leading a Yelp user to publish a negative review? Furthermore, no paper proposed a study focused entirely on negative reviews and reviewers in Yelp, which, starting from a social network-based model representing them, could define a set of stereotypes of users publishing negative reviews. Having all this available would allow us to answer the following research question: How can the interaction of the dynamics driving negative reviewers increase their "power" and the one of their reviews? Finally, no past paper built a profile of a negative influencer in Yelp. Reaching this result would allow us to answer the following research question: Who

are the negative influencers in Yelp? Here we aim at filling this gap and answer the three research questions mentioned above.

We draw inspiration from the research strands mentioned previously. First of all, our multi-dimensional social network-based model of Yelp can be employed to handle different relationships (e.g., friendship, co-review). In particular, it is possible to define an occurrence of the model for each relationship. This way of proceeding falls within the context of multi-relationship networks, but in a new way. In fact, differently from past multi-relationship models, ours does not require the prior and static definition of the relationships to represent, but allows a dynamic choice of them, based on the analysis to be performed. For example, we have chosen friendship and co-review between Yelp users. Furthermore, the choice of including in our model the macro-categories in which the businesses are grouped in Yelp represents an additional feature of it. It makes possible a definition of the bridge concept perfectly fitted on Yelp, which, in turn, allows for the definition of three user stereotypes for this social platform. Therefore, the multi-dimensionality of our model enables an analysis of Yelp users and their relationships from multiple orthogonal viewpoints, acting simultaneously and influencing each other.

Our multi-dimensional social network-based model makes our definition of bridge possible. Starting from that definition, and operating on the model itself, we define three user stereotypes, namely: *(i)* the k-bridge, i.e., a person who reviewed businesses belonging to k different Yelp macro-categories; *(ii)* the power user, i.e., a person very active in all the macro-categories in which she is interested; *(iii)* the double-life user, i.e., a person exhibiting different behaviors in the different macro-categories in which she operates. Compared to the generic stereotypes presented in the past literature [139], those we identified are tailored to Yelp and, therefore, can provide a more specific contribution in the definition of the profile of negative influencers in this social medium.

Having the multi-dimensional model, the three stereotypes and the Negative Reviewer Network at disposal, our approach can investigate negative reviews and reviewers and can build a profile of negative influencers. These tasks are very important because it was shown that the effect of negative reviews and reviewers is much greater than the one of positive reviews and reviewers [18]. Furthermore, negative reviews and reviewers are not very common because people tend to give high ratings to businesses [104, 550]. But for this very reason, the information they bring is extremely valuable. Indeed, consumers and businesses are prone to rely on negative reviews and reviewers to understand the reasons for possible dissatisfaction caused by a product, a service or a business [64, 16].

Compared to the works on negative reviews and reviewers described above, our approach is more focused on the issue of influence, more specifically on negative influence. In this context, it offers a first important contribution thanks to the definition of the Negative Reviewer Network. This tool allows the exploitation of Social Network Analysis techniques to investigate the influence of a negative reviewer on other users. We point out that the Negative Reviewer Network is general and can be used to investigate the same issue in other review platforms. Starting from it and the multi-dimensional model introduced here, which is instead specific to Yelp, our approach provides a second important contribution, i.e., it constructs the profile of a negative influencer in Yelp. Such a profile is perfectly fitted on this social platform because it takes into account both the partitioning of Yelp into macro-categories and the possibility to specify user friendships, provided by this platform.

3.2.3 Methods

3.2.3.1 Definition of Yelp model

Our multi-dimensional investigation of negative reviews and detection of negative influencers in Yelp is possible thanks to a new multi-dimensional social network-based model of Yelp. This model starts from the observation that, in this social medium, businesses are organized according to a taxonomy consisting of four levels. Level 0 includes 22 macro-categories. Each macro-category has one or more child categories; therefore, level 1 includes 1002 categories. A category may have zero, one or more sub-categories; as a consequence, level 2 comprises 532 sub-categories. Finally, level 3, has only 19 sub-sub-categories; indeed, most sub-categories are not further categorized. Our model represents Yelp as a set of 22 communities, one for each macro-category:

$$\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_{22}\}$$

Given the macro-category $\mathcal{C}_i$, $1 \leq i \leq 22$, a corresponding user network $\mathcal{U}_i = \langle N_i, A_i \rangle$ can be defined. N_i is the set of the nodes of $\mathcal{U}_i$; there is a node n_{i_p} for each user u_{i_p} who reviewed at least one business of $\mathcal{C}_i$. A_i is the set of the arcs of $\mathcal{U}_i$; there is an arc $a_{pq} = (n_{i_p}, n_{i_q}) \in A_i$ if there exists a relationship between the users u_{i_p}, corresponding to n_{i_p}, and u_{i_q}, corresponding to n_{i_q}.

Finally, an overall user network $\mathcal{U} = \langle N, A \rangle$ corresponding to $\mathcal{Y}$ can be defined. There is a node $n_i \in N$ for each Yelp user. There is an arc $a_{pq} = (n_p, n_q) \in A$ if there exists a relationship between the users u_p, corresponding to n_p, and u_q, corresponding to n_q.

In the definition of $\mathcal{U}$ (and, consequently, of $\mathcal{U}_i$), we do not specify the kind of relationship between u_p and u_q. Actually, it is possible to define a specialization of

$\mathcal{U}$ for each relationship we want to investigate. Here, we are interested in two relationships existing between Yelp users, namely friendship and co-review. As a consequence, we define two specializations of $\mathcal{U}$, namely $\mathcal{U}^f$ and $\mathcal{U}^{cr}$. $\mathcal{U}^f$ is the specialization of $\mathcal{U}$ when we consider friendship as the relationship between users, whereas $\mathcal{U}^{cr}$ denotes the specialization of $\mathcal{U}$ when co-review (i.e., reviewing the same business) is the relationship between users.

Starting from this model, it is possible to define some Yelp stereotypes, namely: *(i) the k-bridge*, i.e., a person operating in k categories of Yelp; *(ii) the power user*, i.e., a person very active in all the categories that she is interested in; *(iii) the double-life user*, i.e., a person showing different behaviors in the different categories she attends. Her different behaviors can regard the activity level (*access-dl-user*) or the severity of her reviews (*score-dl-user*). These stereotypes can lead to the detection of negative influencers in Yelp.

3.2.3.2 Definition of negative influencer stereotypes

As we have seen above, our methodology starts from the multi-dimensional social network-based model, formulates some hypotheses and aims at verifying them using an inferential campaign based on social network analysis. This campaign makes use of a number of concepts, stereotypes and definitions that we introduce in this section. Instead, the way they are exploited to prove the hypotheses and, more in general, to extract useful knowledge is described in Section 3.2.4.

The first concept we introduce is a stereotype, namely the *k-bridge*. Specifically, a *k-bridge* is a Yelp user who reviewed businesses belonging to exactly k different macro-categories of Yelp. A user who reviewed businesses of only one macro-category is a *non-bridge*. Finally, we use the generic term *bridge* to denote a k-bridge such that $k > 1$. Given a k-bridge u_p of $\mathcal{U}$, where $\mathcal{U}$ is the overall user network corresponding to Yelp, there are k nodes $n_{1_p}, n_{2_p}, \cdots, n_{k_p}$ associated with her, one for each macro-category containing at least one review performed by her.

After having introduced the *k-bridge*, we present some other stereotypes, namely the power user and the double-life user. More specifically, let $\mathcal{C}_i \in \mathcal{Y}$ be one of the macro-categories of Yelp.

Let rn_i be the average number of reviews of $\mathcal{C}_i$. Let b_p be a Yelp bridge and let $CSet_p$ be the set of the macro-categories that received reviews from b_p. Then:

- b_p is defined as a *power user* if, for each macro-category $\mathcal{C}_j \in CSet_p$, the number of her reviews is greater than or equal to $2 \cdot rn_j$.
- b_p is defined as a *(x,y) access double-life user* (*access-dl-user*, for short) if both the following conditions hold:

- for a subset $CSet_{p_x} \subset CSet_p$ of x macro-categories, the number of reviews of each $\mathcal{C}_j \in CSet_{p_x}$ is greater than or equal to $2 \cdot rn_j$;
- for a subset $CSet_{p_y} \subset CSet_p$ of y macro-categories, such that $CSet_{p_x} \cap CSet_{p_y} = \emptyset$, the number of reviews of each $\mathcal{C}_k \in CSet_{p_y}$ is less than or equal to $\frac{1}{2} \cdot rn_k$.

Double-life users play an extremely interesting role because they are very rare. Therefore, we deepen our investigation on them and introduce a second kind of double-life users. Specifically, let b_p be a Yelp bridge. Then b_p is defined as a (x, y) *score double-life user* (*score-dl-user*, for short) if both the following conditions hold:

- for a subset $CSet_{p_x} \subset CSet_p$ of x macro-categories, the average number of stars that b_p assigned to the corresponding businesses is higher than or equal to 4;
- for a subset $CSet_{p_y} \subset CSet_p$ of y macro-categories, such that $CSet_{p_x} \cap CSet_{p_y} = \emptyset$, the average number of stars that b_p assigned to the corresponding businesses is lower than or equal to 2.

In order to make our inferential campaign on negative reviews and reviewers complete, we need to introduce a further network that we call *Negative Reviewer Network* $\overline{\mathcal{U}} = \langle \overline{N}, \overline{A} \rangle$. $\overline{N}$ is the set of nodes of $\overline{\mathcal{U}}$. There is a node $n_i \in \overline{N}$ for each Yelp user who made at least one negative review. There is an arc $a_{pq} = (n_p, n_q)$ if there exists a friendship relationship between the user u_p, corresponding to n_p, and the user u_q, corresponding to n_q.

3.2.3.3 Hypothesis definition

Starting from this theoretical background, we aim at answering the three questions mentioned in the Introduction. In particular, we use the above model and stereotypes to design and perform a social network analysis-based campaign aiming at evaluating some hypotheses that we synthesize in the following:

- First of all, the review mechanism of Yelp is based on a scale from 1 to 5 stars. This is similar to the review mechanisms encountered in several other social media. In this context, we formulate the following:

  Hypothesis 1 (H1) - The star-based review system of Yelp is positively biased.

  In the scale adopted by Yelp, 1 means "absolutely bad" and 5 means "fantastic". A review with 2 stars is still negative, but 3 stars already denote a positive review. In other words, the review mechanism of Yelp makes it more probable that users release positive reviews. Unless the experience was really bad, the review will almost always be positive. This is confirmed by how Yelp itself labels the stars (1

- "Eek! Methinks not"; 2 - "Meh. I've experienced better"; 3 - "A-OK"; 4 - "Yay! I'm a fan"; 5 - "Woohoo! As good as it gets!").

On the other hand, if we consider this review mechanism from a more formal and theoretical viewpoint, we can observe that it is based on a Likert scale, which was already shown to be asymmetric and positively biased [41, 537, 104].

- We think that the stereotypes introduced above can help very much in evaluating negative reviews and influencers. As for a specific kind of stereotype, i.e., the double-life users, we formulate the following:

  Hypothesis 2 (H2) - access-dl-users and score-dl-users play a key role in negative reviews.

  To understand the reasoning behind this hypothesis, consider score-dl-users. Clearly, they can be partitioned into two sets. The former is made up of users who mainly write positive reviews and few negative reviews. These are basically positive users who, for some reasons, had a bad experience with some businesses. So, what drove them to write negative reviews, considering that they are keen to write positive ones? A user assigns a 1-star score to a business when her expectations were not satisfied. This was already investigated in literature (see, for instance, [333]), where it was proved that a high discrepancy between the others' opinions and the experience of a user is the main driver for her to write a negative review.

  The latter set of access-dl-users is much more peculiar. It comprises those users who generally write negative reviews but, in some cases, release positive ones. These users have probably developed very severe criteria for evaluating businesses, leading them to be satisfied only rarely.

- We have already discussed about the multi-dimensionality of our model. One of its main dimensions is friendship. Actually, it is well known that this relationship plays a key role in social networks [109, 588, 107]. Starting from these results, it is reasonable to formulate the following:

  Hypothesis 3 (H3) - A user has a strong influence on her friends when doing negative reviews.

  This could seem obvious. In past literature it has been proved that users are influenced by others when writing reviews. In particular, it has been found that users tend to have a positive opinion of a product/service if it has been positively commented by other users [192].

  In addition, people generally trust more those users sharing their personal profile on online review platforms [270]. It was found that a personal information disclosure is crucial for the spread of positive comments about a product/service, because the possibility of associating information with a particular person gives

a boost in the overall perceived confidence. All of this is amplified when users share a common geographical location. This reasoning can also be applied to relationships like friendship, because personal information is certainly disclosed between friends.

Here, we hypothesize that the influence exerted by friends is valid not only for positive reviews but also for negative ones, possibly leading to a phenomenon of negative influence between friends.

- Another stereotype introduced above that could play an important role as negative influencer is the bridge one. As for it, we formulate the following:

  Hypothesis 4 (H4) - Bridges have a much greater influence power than non-bridges.

  If Yelp can be modeled as a network of different communities, each corresponding to a given business macro-category, it is immediate to think of bridge users as special ones, capable of facilitating information diffusion from a community to another. Bridge users have a position of power in the network, and this power can even be measured [373]. If we look at classical centrality measures in social network analysis, it is easy to argue that bridge users have a high betweenness centrality value. On the other hand, if we look at reviews, it is plausible that a bridge could expand the negative conception of a brand from a category to another which both the bridge and the brand belong to.
- The previous reasoning about the correlation between bridges and betweenness centrality paves the way to think that centralities play a key role in the diffusion of negative reviews. In particular, it is reasonable to make the following hypothesis:

  Hypothesis 5 (H5) - There is a correlation between degree and/or eigenvector centrality and the capability of being negative influencer.

  Degree centrality tells us which nodes have the highest number of relationships in a network. These are probably power users, if we consider our stereotypes. They certainly are important users, because they are densely connected. On the other hand, eigenvector centrality can help us to identify influential users, who do not like to appear as such (the so called grey eminences or grey cardinals). Those kinds of users are often connected to few nodes, each having a high number of relationships with the other users [454]. These two centrality measures can be useful to find negative influencers in Yelp.

3.2.3.4 Preliminary analysis of negative influencers stereotypes

We collected the data necessary for the activities connected with our inferential campaign from the Yelp website at the address `https://www.yelp.com/dataset`. In or-

der to extract information of interest from available data, we had to carry out a preliminary analysis. A first result concerns the presence of 10,289 businesses whose category did not belong to any of the Yelp macro-categories, and 482 businesses that did not have any category associated with them (recall that in Yelp a business can belong to one or more categories). Since the total number of businesses was 192,609, we decided to discard these two kinds of businesses, because the amount of data removed was insignificant while their presence would have led to procedural problems.

At this point, we analyzed the distribution of the categories among the macro-categories. We report the result obtained in Figure 3.25. As we can see from this figure, the macro-category "Restaurants" has a much greater number of categories than the other ones.

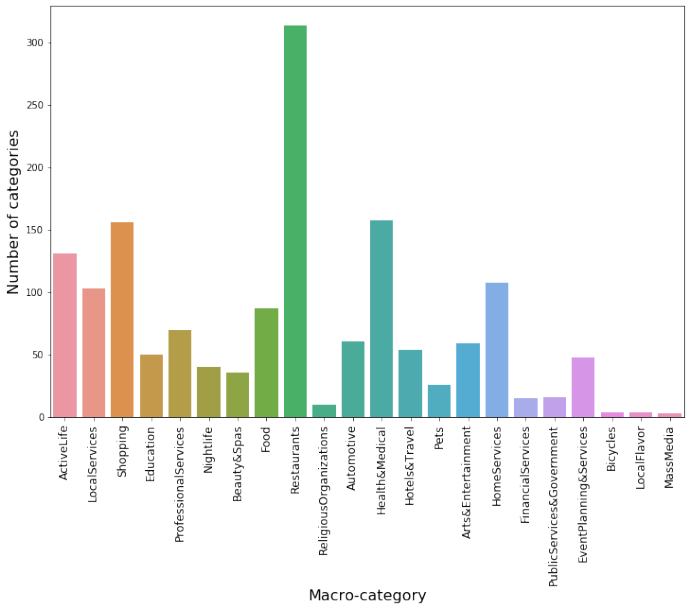


Fig. 3.25: Distribution of the categories inside the Yelp macro-categories

Figure 3.26 shows the average number of reviews per user for each macro-category. As we can see, the three macro-categories with the highest average number of reviews are "Restaurants", "Food" and "Nightlife". Furthermore, in Figure 3.27, we show the same distribution for bridges only. We can see that the three macro-categories with the highest number of reviews are always the same. However, the average number of reviews is generally higher for bridges than for normal users. Therefore, we can conclude that bridges not only tend to review businesses of different macro-categories (and this happens by definition of bridge itself) but also to do more reviews than non-bridges.

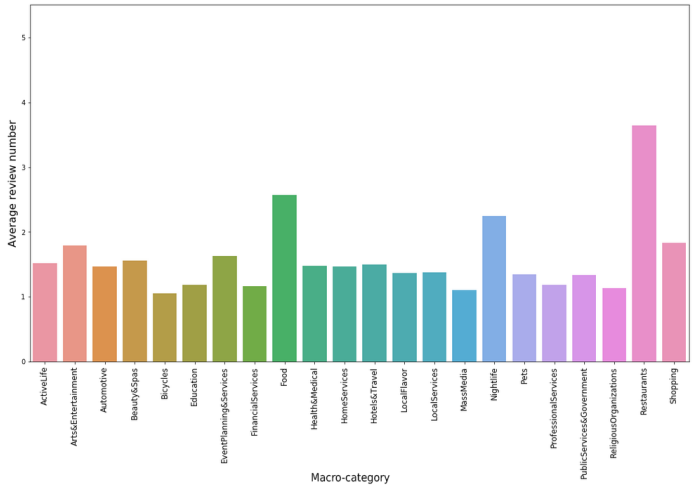


Fig. 3.26: Average number of business reviews made by Yelp *users* for each macro-category

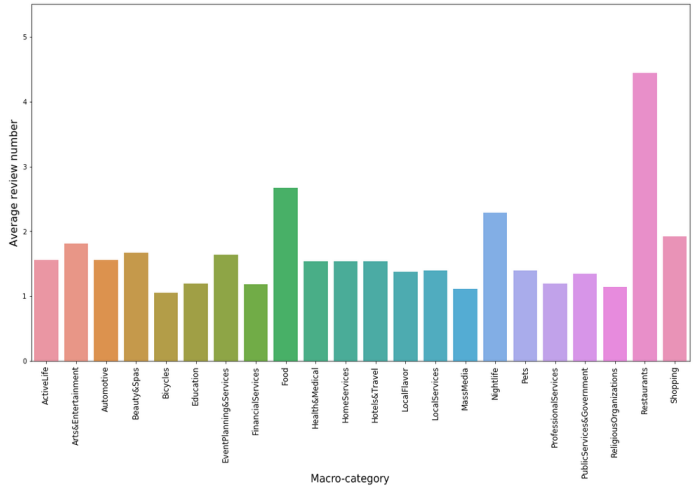


Fig. 3.27: Average number of business reviews made by Yelp *bridges* for each macro-category

In Figure 3.28, we report the distribution of access-dl-users against k. From the analysis of this figure, we observe that the number of access-dl-users is already very high for $k = 2$ and further increases for $k = 3$; then, it decreases very quickly and becomes almost negligible for $k > 4$.

We start looking at the access-dl-users corresponding to the simplest case of bridges, namely 2-bridges. Table 3.18 shows the total number of 2-bridges, the number of (1,1) access-dl-users and the number of power users, together with their corresponding percentage of the overall number of 2-bridges. This table shows that (1,1) access-dl-users and power users represent very small fractions of the overall set of 2-bridges.

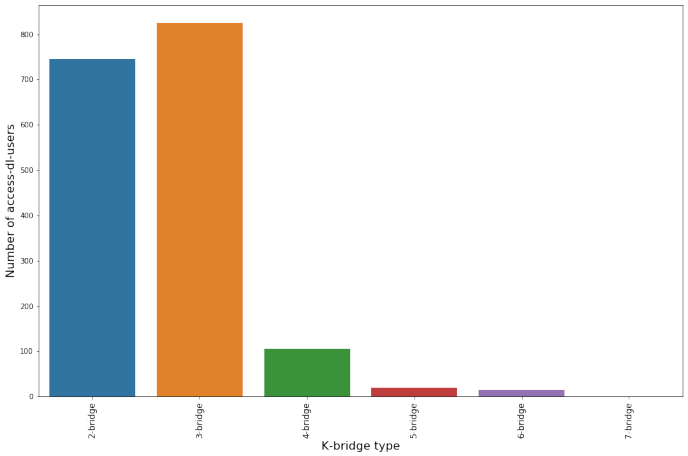


Fig. 3.28: Distribution of access-dl-users against k

Type of users	*Number and percentage*
2-bridges	427130 (100%)
(1,1) access-dl-users	745 (0.17%)
power users	375 (0.087%)

Table 3.18: Numbers and percentages of 2-bridges, access-dl-users and power users in Yelp

We continue by examining all the k-bridges as k grows, until at least one of them is an access-dl-user or a power user. We can observe that this condition occurs for $k \leq 6$. The corresponding numbers and percentages are shown in Tables 3.19 - 3.22. From the analysis of these tables, we can see how the number of k-bridges decreases as k increases, but the decrease is not fast. On the other hand, the number of access-dl-users decreases very rapidly, about one order of magnitude at each step. The number of power users decreases more slowly.

Type of users	*Number and percentage*
3-bridges	245123 (100%)
(1,2) access-dl-users	450 (0.18%)
(2,1) access-dl-users	374 (0.15%)
power users	200 (0.081%)

Table 3.19: Numbers and percentages of 3-bridges, access-dl-users and power users in Yelp

Type of users	*Number and percentage*
4-bridges	147101 (100%)
(1,3) access-dl-users	19 (0.013%)
(2,2) access-dl-users	59 (0.040%)
(3,1) access-dl-users	28 (0.019%)
power users	35 (0.023%)

Table 3.20: Numbers and percentages of 4-bridges, access-dl-users and power users in Yelp

Type of users	*Number and percentage*
5-bridges	91680 (100%)
(1,4) access-dl-users	6 (0.007%)
(2,3) access-dl-users	11 (0.012 %)
(3,2) access-dl-users	3 (0.003%)
(4,1) access-dl-users	0 (0%)
power users	14 (0.015%)

Table 3.21: Numbers and percentages of 5-bridges, access-dl-users and power users in Yelp

Type of users	*Number and percentage*
6-bridges	63708 (100%)
(1,5) access-dl-users	0 (0%)
(2,4) access-dl-users	0 (0%)
(3,3) access-dl-users	1 (0.002%)
(4,2) access-dl-users	2 (0.003%)
(5,1) access-dl-users	11 (0.017%)
power users	11 (0.017%)

Table 3.22: Numbers and percentages of 6-bridges, access-dl-users and power users in Yelp

3.2.4 Results

3.2.4.1 Investigating the Hypothesis H1

A user can assign a number of stars between 1 and 5 to a business in Yelp. The higher the number of stars, the better her rating is. Therefore, we decided to study the reviews of users focusing on the number of stars that they assigned to businesses.

Figure 3.29 shows the average number of stars that users assigned to the businesses of each macro-category. As we can see from this figure, this number is very high as it is always greater than 3. As previously pointed out, this is actually not very surprising because the mechanism based on stars follows a Likert scale and, in literature, it is well known that this scale is generally positively biased [41, 537, 104].

In Table 3.23, we report the mean, standard deviation and mode of the number of stars assigned by bridges and non-bridges to all businesses. As we can see from

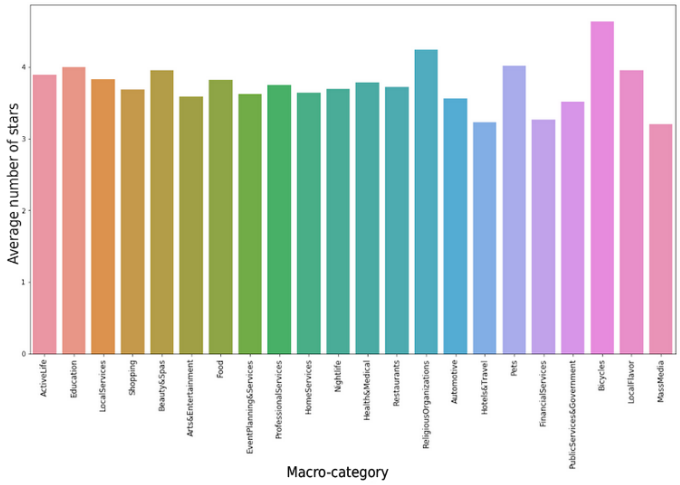


Fig. 3.29: Average number of stars for each macro-category of Yelp

this table, there is no substantial difference in this type of behavior between bridges and non-bridges.

Statistical Parameter	*Bridges*	*Non-bridges*
Mean	3.73	3.57
Standard Deviation	1.44	1.72
Mode	5	5

Table 3.23: Values of mean, standard deviation and mode of the number of stars assigned by bridges and non-bridges to all businesses

From the results of Table 3.23, it is clear that it makes no sense to talk about power users in the star-based analysis, because almost all users have the same behavior and assign a high number of stars to almost all businesses. All these tests allow us to define the following:

> Implication 1: The star-based review system of Yelp is positively biased. Indeed, almost all users assign a high number of stars to almost all businesses.

Implication 1 is clearly a confirmation of the correctness of the Hypothesis H1.

3.2.4.2 Investigating the Hypothesis H2

In Figure 3.30, we report the distribution of score-dl-users against k. From the analysis of this figure we note that it follows a power law. If we compare this figure with Figure 3.28, we observe that for $k = 2$, the number of score-dl-users is much smaller

than the one of access-dl-users. However, the decrease of the number of score-dl-users when k increases is much smaller because they are different from 0 until to $k = 14$.

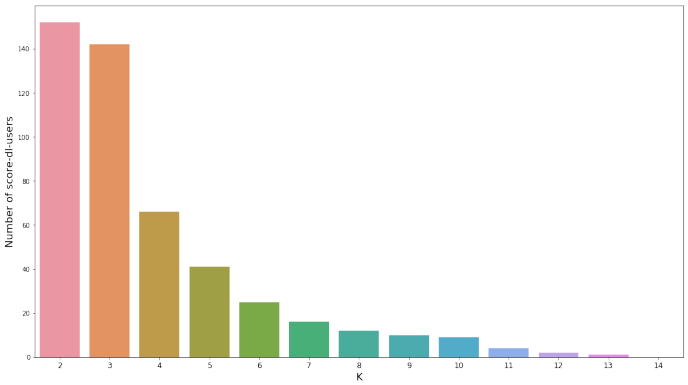


Fig. 3.30: Distribution of score-dl-users against k

We continued our analysis by verifying whether score-dl-users and access-dl-users were the same people or not. We carried out this analysis with $k = 6$, because we had no access-dl-users with higher values of k. In this case, we could see that the intersection of the two sets was empty.

To better understand the main features of score-dl-users we considered those corresponding to 7-bridges. These users were 16 (see Figure 3.30), a number that allowed us to examine in detail each review carried out by them. During this analysis we found several interesting knowledge patterns. More specifically, we observed that (1,6) and (6,1) score-dl-users show a completely different behavior from the other 7-bridges. In fact, in this case, each (1,6) score-dl-user assigned positive scores to all the business of the only macro-category that she positively reviewed. Similarly, each (6,1) score-dl-user assigned negative values to all the businesses of the only macro-category that she negatively reviewed. This can be justified thinking that users have a strong interest in that macro-category and so they developed more accurate and stable evaluation criteria for the businesses belonging to it.

As for the other 7-bridges, we found that (2,5), (3,4), (4,3) and (5,2) score-dl-users show a less extreme behavior, in the sense that they do not tend to give always positive or always negative ratings to all the businesses of a given macro-category.

We then repeated the previous analyses for the last category of access-dl-users that we had available, namely the 6-bridges, to verify if the particular behavior of score-dl-users was typical of this kind of double-life user or if it was something common. Actually, 6-bridge access-dl-users were 13; therefore, we were able to make a detailed analysis of each review performed by each user also in this case. We exam-

ined (1,5), (2,4), (3,3), (4,2) and (5,1) access-dl-users and we did not find substantial differences in the behavior of these five categories of users. This appeared as a confirmation of the singularity of the behavior observed for (1,6) and (6,1) score-dl-users. The previous analyses suggest the following:

Implication 2: *(a)* Score-dl-users play a key role in negative reviews. *(b)* They are very keen on negatively judging the macro-category they mostly attend.

Implication *2(a)* confirms the correctness of our Hypothesis H2. But there is much more. In fact, Implication *2(b)* was an unexpected result that prompted us to carry out a further experiment to have a confirmation. In it, we considered k-bridges, with $3 \leq k \leq 8$, and computed the percentage of them who negatively reviewed the macro-category of businesses they attended the most. Afterwards, we computed the same percentage taking into account only k-bridges that were score-dl-users. The results obtained are shown in Table 3.24. They represent an extremely strong confirmation of the previous qualitative analysis.

k	*Percentage of k-bridges*	*Percentage of score-dl-users k-bridges*
3	4.35%	91.5%
4	4.03%	79%
5	3.65%	61%
6	2.40%	63%
7	2.11%	56%
8	1.55%	33%

Table 3.24: Percentages of k-bridges and score-dl-users k-bridges who negatively reviewed the macro-category they mostly attended

As we have seen, the definition and behavior of score-dl-users are based on the number of stars assigned by a user to a business during a review. We have already said that this type of score is based on a Likert scale and, therefore, it is positively biased [41, 537, 104]. In order to overcome this problem, in the literature authors suggest evaluating the text of the reviews and to make a sentiment analysis on it [372, 369]. We carried out this activity using two well-known sentiment analysis tools. The first is TextBlob[9], which, given a text, specifies if the corresponding polarity is positive, negative or neutral. We applied TextBlob to users' review texts. The results obtained are reported in Table 3.25. From the analysis of this table we can see that the difference between the score based on stars and the polarity based on sentiment analysis is equal to 15%.

The second sentiment analysis tool we considered is Vader [350]. Also in this case, we applied it to the users' review texts. The results obtained are shown in Table

[9] https://textblob.readthedocs.io

Parameters	*Value obtained by applying TextBlob*
Reviews	6,685,902
Reviews with a number of stars less than or equal to 2 (negative reviews)	1,544,553
Reviews classified as negative by TextBlob	847,359
Reviews with a number of stars greater than or equal to 3 (positive reviews)	5,141,347
Reviews classified as positive by TextBlob	5,781,007
Reviews classified as neutral by TextBlob	57,536
Negative reviews classified as positive	823,414
Positive reviews classified as negative	154,176
Positive reviews classified as neutral	30,914
Negative reviews classified as neutral	26,620

Table 3.25: Comparison between the review score based on stars and the review polarity obtained by applying TextBlob

3.26. The analysis of this table confirms the very low difference between the score of the star-based reviews and the polarity of the review texts (in fact, in this case, this difference is equal to 14%).

Parameter	*Value obtained by applying Vader*
Reviews	6,685,902
Reviews with a number of stars less than or equal to 2 (negative reviews)	1,544,553
Reviews classified as negative by Vader	982,102
Reviews with a number of stars greater than or equal to 3 (positive reviews)	5,141,347
Reviews classified as positive by Vader	5,649,489
Reviews classified as neutral by Vader	54,311
Negative reviews classified as positive	724,241
Positive reviews classified as negative	184,557
Positive reviews classified as neutral	31,542
Negative reviews classified as neutral	22,767

Table 3.26: Comparison between the review score based on stars and the review polarity obtained by applying Vader

This allows us to conclude that score-based evaluations are generally confirmed by the sentiment analysis performed on the corresponding reviews.

3.2.4.3 Investigating the Hypothesis H3

At this point, we analyzed how users influence each other with regard to negative reviews. We took into consideration the network of friendships $\mathcal{Y}^f$ since it is easier for a user to have characteristics more similar to her friends than to people she does not know, due to the principle of homophily [468]. Therefore, the ability to influence someone and/or to be influenced by her is presumably greater with friends than with others.

As a first analysis, for each macro-category, we considered the percentage of users such that they, and at least one of their friends, reviewed the same business negatively. The results obtained are shown in Figure 3.31. From the analysis of this figure we can see how the percentages are extremely low. The macro-category with the highest percentage is "Restaurant", followed by "Nightlife" and "Food". This result can be explained taking into account that a person often attends restaurants or nightclubs with her friends. Therefore, it is not unlikely that her negative judgement of a business may lead to (or, on the contrary, may be caused by) a negative judgement of one or more of her friends.

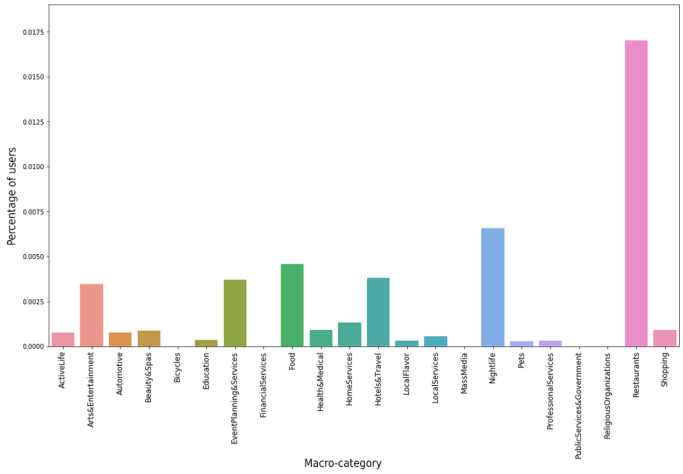


Fig. 3.31: Percentages of *users* such that they, and at least one of their friends, reviewed the same business negatively

We repeated the analysis by distinguishing bridges from non-bridges. The corresponding results are shown in Figures 3.32 and 3.33. From the analysis of these figures we observe higher values for bridges than for non-bridges. For example, the value of "Nightlife" for bridges is more than 4 times the value for non-bridges. Similarly, "Food", in case of bridges, has a percentage more than 7 times higher than for non-bridges.

To prove the statistical significance of our results we adopted a null model to compare our findings with those obtained in an unbiasedly random scenario. Specifically, we built our null model by shuffling the negative reviews among users in our dataset. In this way, we left unaltered all the original features with the exception of the distribution of negative reviews, which became unbiasedly random in the null model. After that, we repeated our analysis on the null model. The results obtained are reported in Figure 3.34. Comparing this figure with Figure 3.31, we can see that there is a certain similarity in the distributions; indeed, many of the macro-

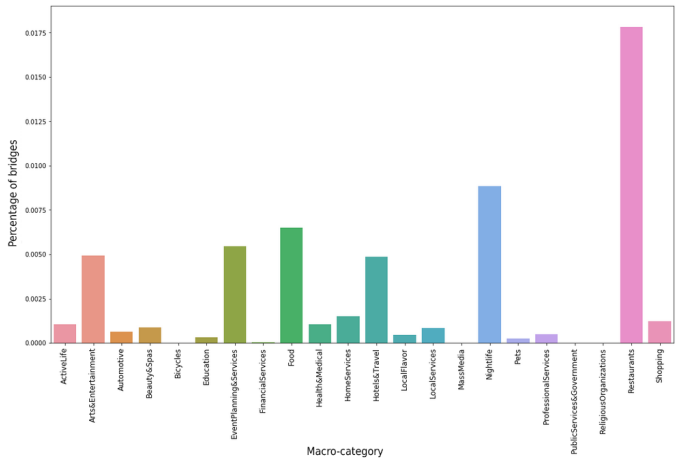


Fig. 3.32: Percentages of *bridges* such that they, and at least one of their friends, reviewed the same business negatively

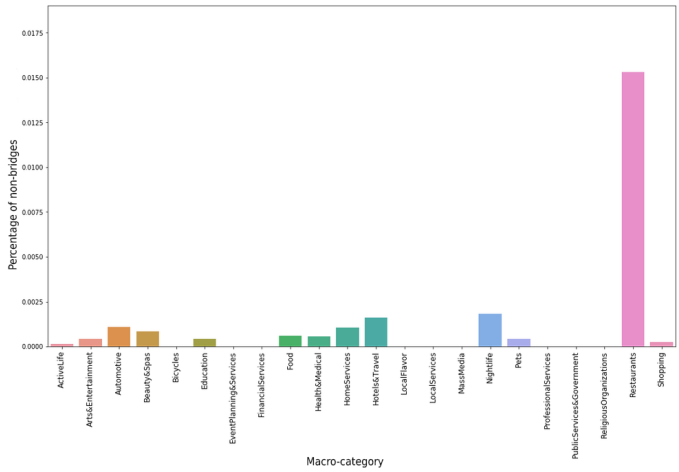


Fig. 3.33: Percentages of *non-bridges* such that they, and at least one of their friends, reviewed the same business negatively

categories that had the highest values in Figure 3.31 continue to have the highest values in Figure 3.34. However, in this last case, the values of the percentages are several orders of magnitude smaller. Therefore, we can conclude that the behavior observed in Figure 3.31 is not random but it is the result of the reference context.

At this point, for each macro-category, for each user who reviewed a given business negatively, we computed the percentage of her friends who, having reviewed the same business, made a negative review. The results obtained are shown in Figure 3.35. As we can see from this figure, the percentage values are very high for almost all macro-categories.

Figures 3.36 and 3.37 show the same distributions, but for bridges and non-bridges. From the analysis of these figures, it can be observed that the phenomenon is

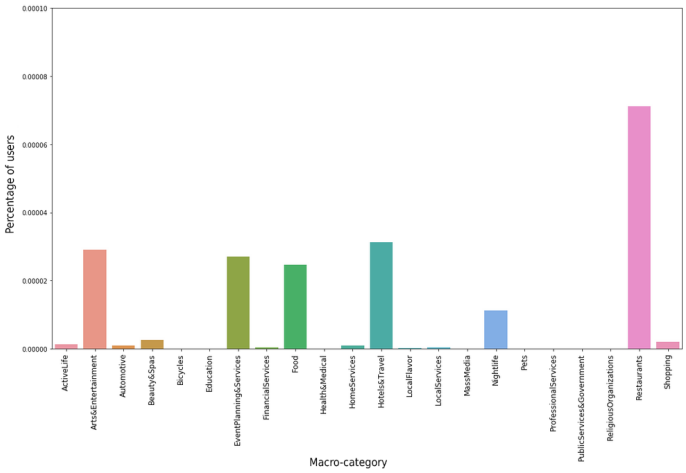


Fig. 3.34: Percentages of *users* in the null model such that they, and at least one of their friends, reviewed the same business negatively

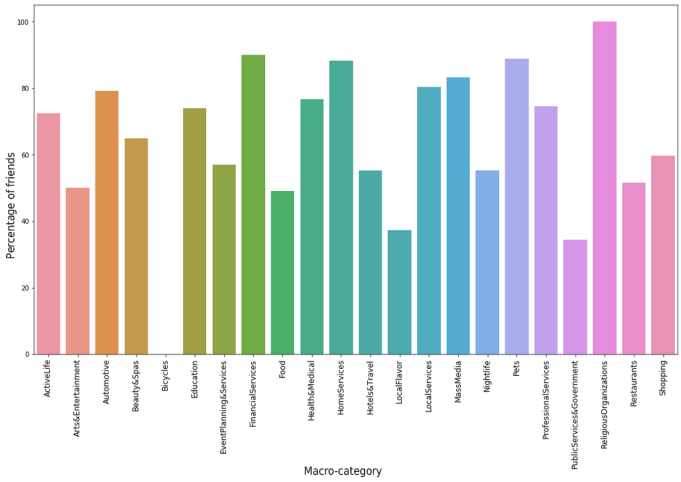


Fig. 3.35: Percentages of friends who, having reviewed the same business as a *user* who reviewed a business negatively, also provided a negative review

always strong, regardless of whether or not a user is a bridge. An interesting knowledge pattern to observe is that there is a strong polarization on the macro-categories especially in the case of non-bridges. In fact, the percentages of friends influenced by them are either above 90% or null.

All the results shown above allow us to deduce the following:

Implication 3: A user has a very high influence on her/his friends when doing negative reviews.

This implication represents a confirmation of the correctness of our Hypothesis H3.

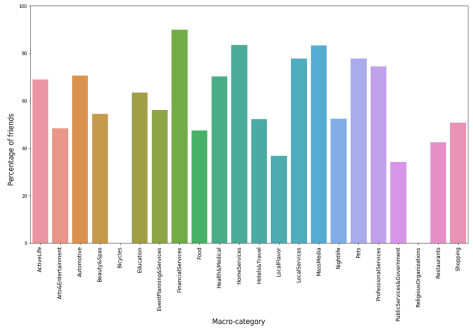


Fig. 3.36: Percentages of friends who, having reviewed the same business as a *bridge* who reviewed a business negatively, also provide a negative review

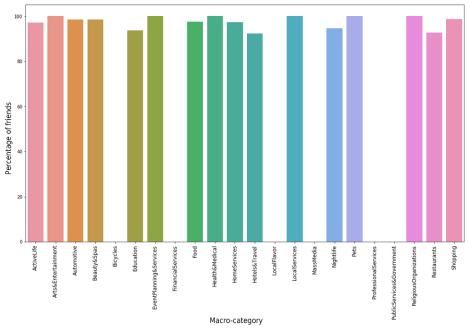


Fig. 3.37: Percentages of friends who, having reviewed the same business as a *non-bridge* who reviewed a business negatively, also provide a negative review

3.2.4.4 Investigating the Hypothesis H4

In order to evaluate the Hypothesis H4, we started with the computation of the average percentage of users who, having made a negative review in a category, have at least $X\%_{\text{oo}}$ of their friends who negatively reviewed a business in the same category. The values of X that we considered are 1, 2, 3, 5, 10 and 100. As an example, in Figure 3.38, we report the results obtained in the case of $X = 5$. As we can see from this figure, the percentages are some orders of magnitude greater than the ones of Figure 3.34. The macro-categories with the highest values are the same as before, i.e., “Restaurants”, “Food” and “Nightlife”.

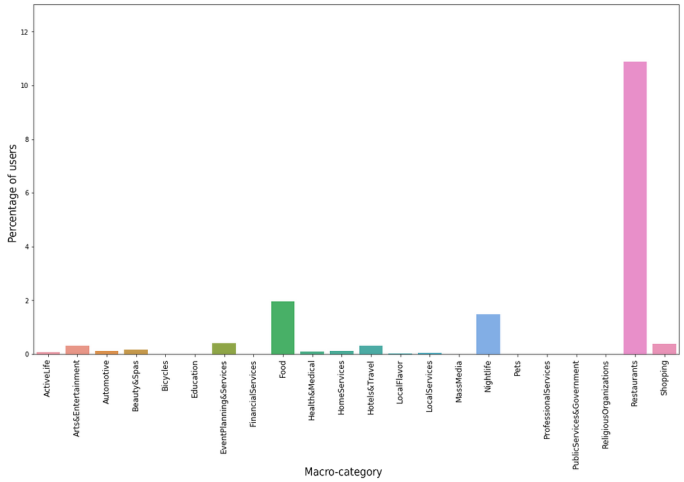


Fig. 3.38: Average percentages of *users* who, having made a negative review in a macro-category, have at least $X\%_{\text{oo}}$ of their friends who reviewed a business in the same macro-category negatively

As in the previous case, we distinguished bridges from non-bridges. The results of the corresponding analysis are shown in Figures 3.39 and 3.40. These figures, along with the previous ones involving bridges and non bridges, allow us to define the following:

Implication 4: Bridges have a much greater power of influence than non-bridges.

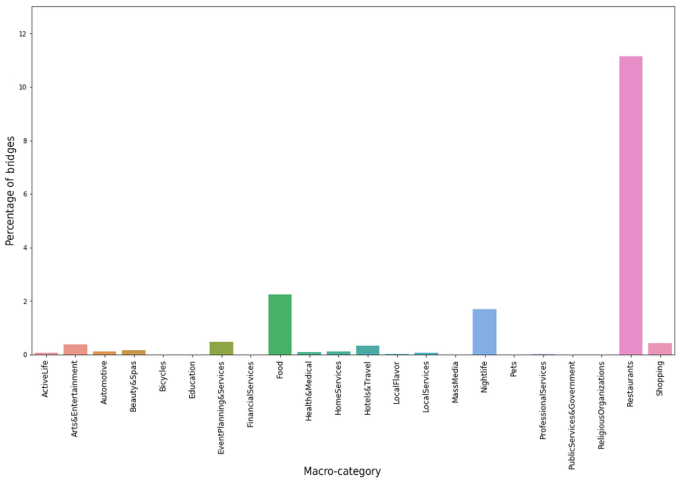


Fig. 3.39: Average percentages of *bridges* who, having made a negative review in a macro-category, have at least $X\%_{oo}$ of their friends who reviewed a business in the same macro-category negatively

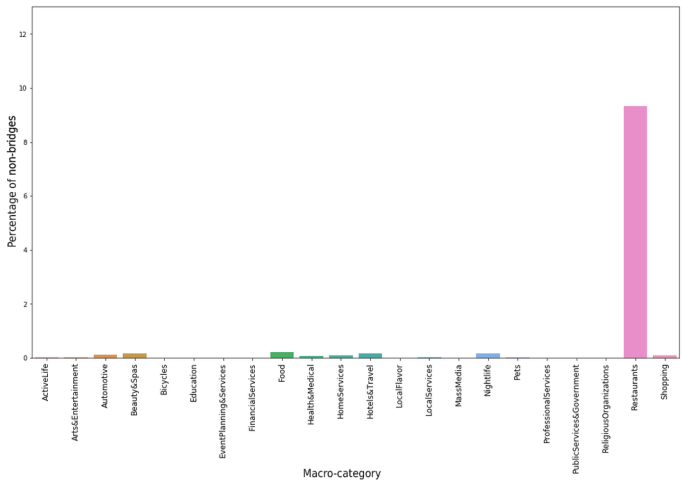


Fig. 3.40: Average percentages of *non-bridges* who, having made a negative review in a macro-category, have at least $X\%_{oo}$ of their friends who reviewed a business in the same macro-category negatively

Again, we made the comparison with the null model. The results obtained for $X = 5$ are reported in Figures 3.41, 3.42 and 3.43. From the examination of these figures, we can see how results obtained are not random but they are intrinsic to Yelp. Note that the non-randomness can be observed for *bridges* but generally not for *non-bridges*; this is important because it allows us to conclude that this property characterizes bridges against non-bridges.

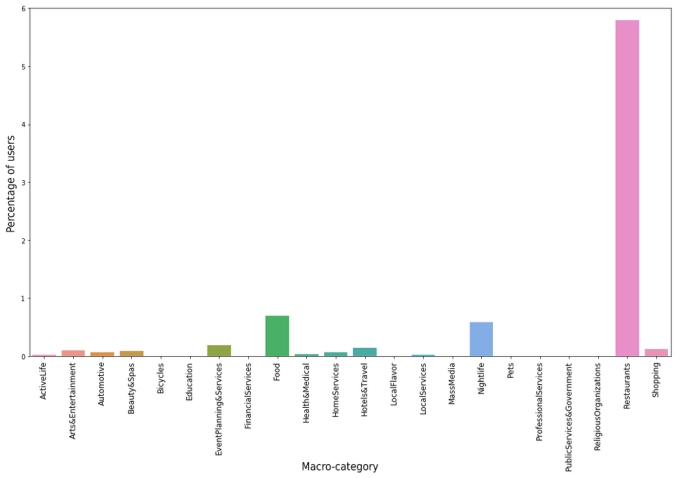


Fig. 3.41: Average percentages of *users* in the null model who, having made a negative review in a macro-category, have at least $X\%_{\text{oo}}$ of their friends who reviewed a business in the same macro-category negatively

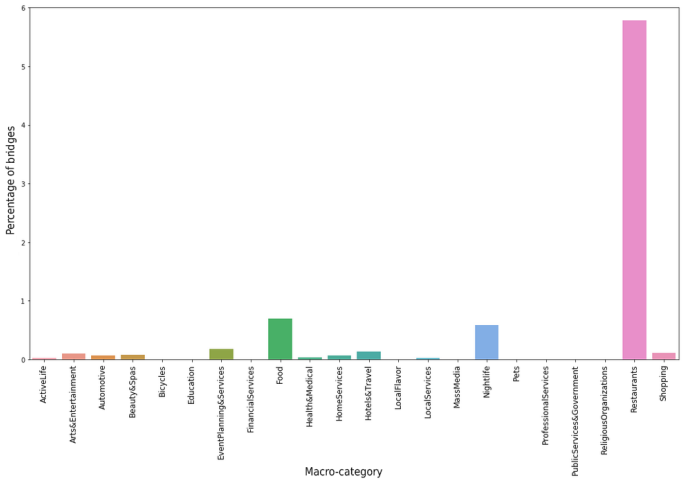


Fig. 3.42: Average percentages of *bridges* in the null model who, having made a negative review in a macro-category, have at least $X\%_{\text{oo}}$ of their friends who reviewed a business in the same macro-category negatively

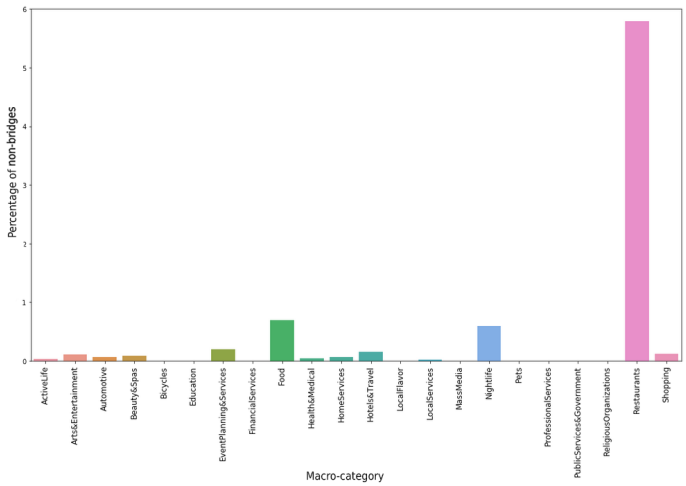


Fig. 3.43: Average percentages of *non-bridges* in the null model who, having made a negative review in a macro-category, have at least $X\%_{oo}$ of their friends who reviewed a business in the same macro-category negatively

Implication 4 represents a confirmation that our Hypothesis H4 was correct.

3.2.4.5 Investigating the Hypothesis H5 and defining a profile of negative influencers in Yelp

To investigate the correctness of the Hypothesis H5 we considered the *Negative Reviewer Network* $\overline{\mathcal{U}} = \langle \overline{N}, \overline{A} \rangle$ introduced in Section 3.2.3.2.

The analysis of this network allowed us to focus on users who reviewed some businesses negatively, because, as we saw in the previous analysis, they are uncommon. Firstly, we computed the number of nodes, the number of edges, the clustering coefficient and the density of $\overline{\mathcal{U}}$ and we compared them with the same parameters as $\mathcal{U}$. Results are shown in Table 3.27.

	$\mathcal{U}$	$\overline{\mathcal{U}}$
Number of nodes	1637138	743178
Number of edges	7392305	2199987
Average clustering coefficient	0.043	0.039
Density	0.00000551619	0.00000796645

Table 3.27: Characteristics of $\mathcal{U}$ and $\overline{\mathcal{U}}$

From the analysis of this table we can observe that the number of users who made at least one negative review is 45.39% of total users. As for the average clustering coefficient and the density, we found that their values do not present significant differences between $\mathcal{U}$ and $\overline{\mathcal{U}}$.

At this point, we computed the distribution of users for $\overline{\mathcal{U}}$; it is shown in Figure 3.44. As we can see from this figure, it follows a power law.

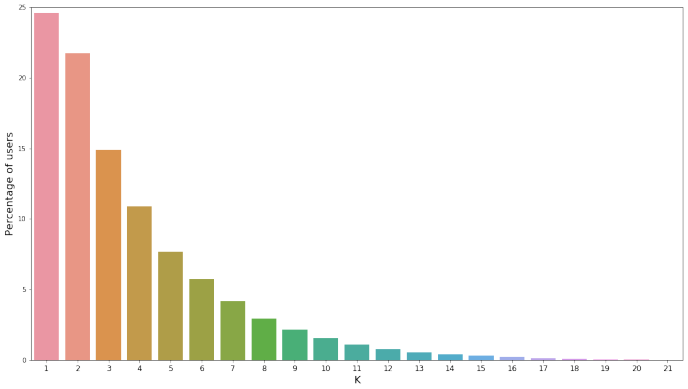


Fig. 3.44: Distribution of users of $\overline{\mathcal{U}}$ against k

After studying the basic parameters of $\overline{\mathcal{U}}$, we computed the degree centrality of the nodes of this network. In particular, we focused on the users with the highest values of degree centrality. More specifically, we considered the top X% users, $X \in \{1, 5, 10, 20\}$. Observe that as X decreases, the corresponding top users are increasingly central, i.e., increasingly strong. In Figure 3.45, we show the distributions against k for the top X% of users with the highest degree centrality. Note that for $X = 20$, the distribution follows a power law, even if it is flatter than the one of Figure 3.44, which referred to all users. As X decreases, we can see how the distribution becomes flatter and flatter, moving to the right and tending to a Gaussian shape. This allows us to conclude that more central users (i.e., those with the highest degree centrality) tend to be stronger also as k-bridges (i.e., characterized by an increasingly higher value of k).

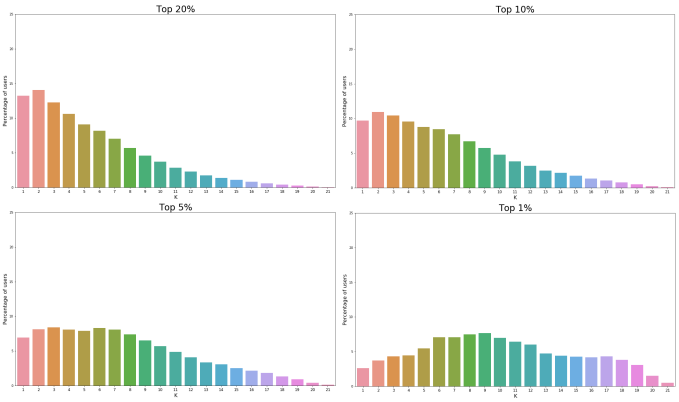


Fig. 3.45: Distributions of the top X% of users with the highest degree centrality against k

Instead, in Figure 3.46, we show the user distributions against k for the top X% of users with the highest eigenvector centrality. The trend of these distributions as X decreases is very similar to (although slightly less marked than) the one of the degree centrality.

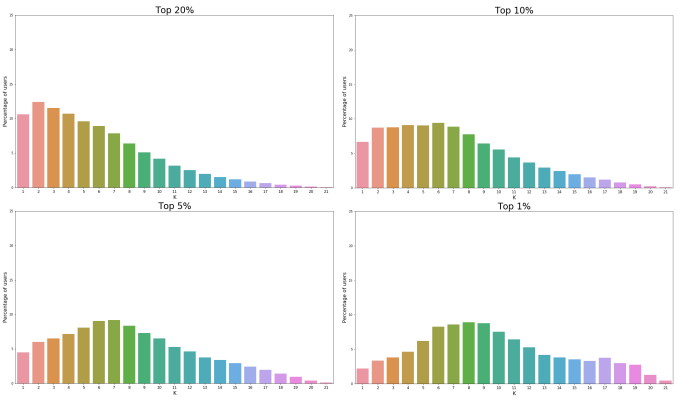


Fig. 3.46: Distributions of the top X% of users with the highest eigenvector centrality against k

Figure 3.47 shows the user distributions against k for the top X% of users with the highest PageRank. Also in this case, we have a similar trend, although the variations of the distributions as X decreases are much more attenuated, compared to the two previous cases. The last three figures allow us to define the following:

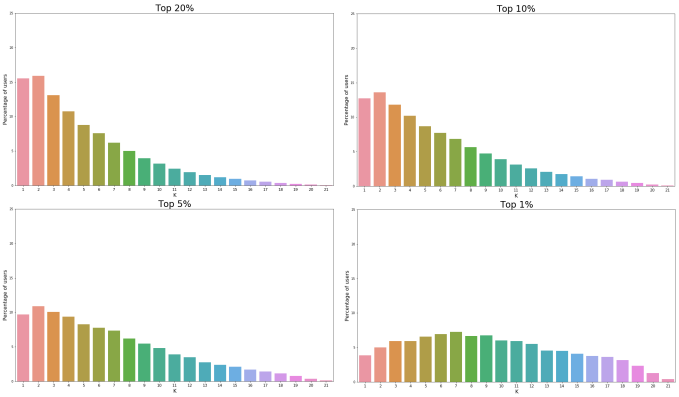


Fig. 3.47: Distributions of the top X% of users with the highest PageRank against k

Implication 5: There is a correlation between k-bridges and top central users.

Implication 5 is valid especially for the top central users based on degree centrality. This result, along with the previous ones, is extremely important because it

allows us to determine which are the main negative influencers in Yelp. In fact, we can define the following:

> ***Implication 6: The main negative influencers in Yelp are score-dl-users who simultaneously are top central users (according to degree and/or eigenvector and/or PageRank centrality measures).***

Implication 6 not only confirms the correctness of the Hypothesis H5, but goes much further. In fact, it defines a profile of the negative influencers in Yelp and, consequently, provides a way to detect them.

3.2.5 Discussion

3.2.5.1 Reference context

In the previous sections, we have investigated the phenomenon of negative reviews in Yelp and, then, we have characterized negative influencers in this social medium. In the past, different research papers have focused on the consequences that user-written reviews have on businesses and, generally, on the market. As a first step in this scenario, it is interesting to understand what makes customer reviews helpful to a consumer in her process of making a purchase decision. With regard to this, in [593], the authors first collect reviews made on `Amazon.com`. Then, they distinguish between two different product types, namely: *(i)* search goods, for which a consumer can obtain information on their quality before purchasing them; *(ii)* experience goods, which are products requiring a purchase before evaluating their quality. This product categorization plays a key role in understanding what a consumer perceives more from a review. Indeed, moderate reviews are more helpful than extreme (i.e., strongly positive or negative) ones for experience goods, but not for search goods. Furthermore, longer reviews are generally perceived as more helpful than shorter ones, but this effect is greater for search goods than for experience goods.

Another interesting contribution in this scenario is reported in [714], in which the authors introduce several factors that can influence the decision making process of consumers about their purchases. Indeed, the authors of [714] strive to understand the key elements that guide a user in the purchase of a certain product. They propose a model taking systematic factors (e.g., the quality of online reviews) and heuristic ones (e.g., the quantity of online reviews) into account. They test this model on 191 users and obtain interesting results. In fact, they identify important factors to care about; these are argument quality, source credibility, and perceived quantity of reviews. They empirically prove that consumers receiving reviews from credible

sources and perceiving the quantity of reviews as large tend to perceive the topics in online reviews as more informative and persuasive. This means that if consumers find review sources to be credible, their purchase intention is usually higher. Finally, they also show that consumers are more likely to purchase products with many online reviews rather than with few ones.

Several authors have investigated the impact of positive and negative reviews. For instance, the authors of [192] examine how a positive Electronic Word of Mouth (hereafter, eWOM) can affect other users' purchasing decisions. Indeed, eWOM is strictly related to the online reviews phenomenon, which can be regarded as a special case of it. Generally, eWOM is based on an analysis of costs and benefits. The authors investigate the psychological motivations beneath the spread of positive reviews. They take a sample dataset from the `OpenRice.com` platform, one of the most successful review platforms in Hong Kong and Macau. Through a questionnaire, they asked people who wrote reviews on this website their motivations. Starting from the received answers, they build a model based on different features, namely the eWOM intention of consumers, the reputation, the reciprocity, the sense of belonging, the pleasure to help, the moral obligation and the self-efficacy of knowledge. They show that their model is capable of representing the behavior of users when they share (positive) personal experiences on such online platforms.

The influence of positive reviews of businesses has been studied from many other points of view. For example, in [385], the authors analyze celebrity sponsorships in the context of for-profit and non-profit marketing. They actually find that famous people can influence the appreciation one has for a product or service, in a positive or negative direction. This suggests that it makes sense studying who negative influencers are, how they behave and how they can be detected in an online platform. Not limited to celebrities, people are more incline to follow users disclosing their personal information [270]. The members of an online community rate reviews containing descriptive identity information more positively, and the prevalence of identity information disclosure by reviewers is associated with increased subsequent sales of online products. In addition, the shared geographical location increases the relationship between disclosure and product sales.

Wrapping up these important results, we can say that buyers are influenced by positive eWOM, especially if it is performed by nearby identifiable users; even more, celebrities can change the appreciation that people have for a product or a service. But the consequences are not just limited to customers. Even internal decision-making processes of businesses can be influenced by online review systems [18]. The diffusion of personal opinions through the Internet has radically changed the concept of reviewing a product or a service that one has in traditional media. In fact,

online review platforms offer to users a space where they can express their *unfiltered* thoughts on products or services. In particular, eWOM encourages a two-way communication between a source and a reader, thus being more engaging. A very important result of [18] is that eWOM helps companies to obtain higher product and service evaluations and, if necessary, higher amounts of funding; furthermore, it influences the decision-making processes of companies, showing that its power is not limited only to buyers. The other important result of [18] is that the effect of negative eWOM is much greater than the one of positive eWOM.

Negative reviews open up many research issues. One of them is finding out what drives users to write negative reviews. Discontent, or "disconfirmation", with a product or service has been studied as a cause of this phenomenon. The authors of [333] define disconfirmation as the discrepancy between the expected evaluation of a product and the evaluation of the same product performed by experts. In particular, they find that a person is more likely to leave a review when the disconfirmation she encounters is great. They also find that the evaluation published by a person may not reflect her post-purchase evaluation in a neutral manner; indeed, the direction of such polarization is in agreement with disconfirmation.

The authors of [703] introduce a theory about the initial beliefs of a consumer when she is looking for a product. According to this theory, a consumer forms an initial judgement about a product based on its summary rating statistics. This initial belief plays a key role in her next evaluation of the review. To prove their conjecture, the authors of [703] collected the application reviews from Apple Store from July 1^{st} to August 31^{st}, 2013. By analyzing these reviews they show the existence of a confirmation bias, which outlines the tendency of consumers to perceive reviews confirming (resp., disconfirming) their initial beliefs as more (resp., less) helpful. This tendency is moderated by the consumer confidence in their initial beliefs. This bias also leads to a greater perceived helpfulness of positive reviews when the average product rating is high, and of negative reviews when the average product rating is low.

3.2.5.2 Main findings of the knowledge extraction process

In the Introduction, we specified that the main novelties concern: *(i)* the definition of the two social network-based models of Yelp; *(ii)* the definition of three Yelp user stereotypes and their characteristics; *(iii)* the construction of the profile of negative influencers in Yelp. We also pointed out that we aim at answering three research questions, namely: *(i)* What about the dynamics leading a Yelp user to publish a negative review? *(ii)* How can the interaction of these dynamics increase the "power" of negative reviews and people making them? *(iii)* Who are the negative influencers

in Yelp? In order to obtain these results and answer these questions, we conducted a data analytics campaign that allowed us to formulate six implications.

The first tells that "The star-based review system of Yelp is positively biased. Indeed, almost all users assign a high number of stars to almost all businesses.". It can be explained by taking into account that Yelp's review system is based on a Likert scale, and it is well known that this scale is positively biased [41, 537, 104]. This implication does not provide unexpected information, but still represents an important confirmation about the correctness of our knowledge extraction process.

The second implication tells that "Score-dl-users play a key role in negative reviews. They are very keen on negatively judging the macro-category they mostly attend.". Unlike the first one, it was not expected. Its explanation partially comes from the first implication. Indeed, if it is true that the Likert scale is positively biased, then a user must be particularly motivated to give a negative rating. Moreover, if such an evaluation is given by a double life user, then it means that it is provided by a person potentially balanced in her evaluations (indeed, she gave both positive and negative evaluations in the past). If a person with these characteristics gives a negative review, it is reasonable to assume that she did so because she had "something important to say". In that case, she probably provides some well founded justifications for her dissatisfaction. In order to do this, she must be competent in that macro-category, which explains the last part of the implication.

The third implication tells that "A user has a very high influence on her/his friends when doing negative reviews.". The first part of it represents an expected result, and is easily explained by the homophily principle [468]. The second part was unexpected and can be explained by considering that several studies in related literature show that negative reviews and reviewers are stronger than positive ones.

The fourth implication tells that "Bridges have a much greater power of influence than non-bridges.". It represents a partially expected result if we consider that bridges generally have a high betweenness centrality and, thus, have the ability to convey an idea, sentiment or opinion from one macro-category to another.

The fifth implication tells that "There is a correlation between k-bridges and top central users.". At first glance, it may appear an expected result, but actually this is not the case. In fact, in some contexts, for example in a Social Internetworking System, bridges connecting different social networks are not necessarily power users [134]. Actually, the more the communities involved in a (multi-) network scenario are integrated, the more likely a bridge is also a power user. Based on this reasoning, and considering that Yelp's macro-categories are closely related to each other, because both a user and a business can belong to more macro-categories simultaneously, the result obtained is reasonable and motivated.

Finally, the sixth implication tells that "The main negative influencers in Yelp are score-dl-users who simultaneously are top central users (according to degree and/or eigenvector and/or PageRank centrality measures).". It is certainly unexpected and is one of our major findings. It was obtained by appropriately integrating the previous five implications. For this reason, the justifications underlying it are those that allowed us to explain the implications from which it derives.

3.2.5.3 Theoretical contributions

Here, we provide several theoretical contributions to the literature on online review systems and eWOM. First of all, it introduces a new multi-dimensional social network-based model of Yelp. This model perfectly fits the category-based structure of this social medium. It represents Yelp as a set of 22 communities, one for each macro-category. At the same time, it models this social medium as a user network $\mathcal{U}$ where each node denotes a user and an arc between two nodes represents a generic relationship between the corresponding users. Our model can be used in several different scenarios, depending on the type of relationship one wants to represent. In our study, we have specialized it to two different types of relationships, namely the friendship between users (i.e., $\mathcal{U}^f$) and the co-review of the same business carried out by different users (i.e., $\mathcal{U}^{cr}$).

The usage of our model, together with a set of experiments performed on a Yelp dataset, allowed us to show that the star-based review mechanism of Yelp is positively biased. This fact implies that a user must have a strong motivation to write a negative review. In turn, this implies that all information about negative reviews and negative influencers in Yelp is extremely valuable.

After that, thanks to our multi-dimensional model, we were able to define different stereotypes of users in Yelp. In particular, we considered three different stereotypes, namely the bridges, the power users and the double-life users. Bridges are users connecting different communities in Yelp. They are crucial for the dissemination of information in this social platform. In fact, we have seen that the influence exerted by bridges is greater than the one exerted by non-bridges. Power users are very active in performing reviews in the categories of their interest. The amount of reviews they carry out makes them extremely important in the identification of potential influencers. Double-life users show different behaviors in the different categories in which they operate. They generally show a particular attention and severity in a category in which they are extremely experienced. This means that they can play a valuable role as influencers in this category.

We have defined our multi-dimensional model and these stereotypes with respect to Yelp. However, our model can be easily generalized to other online review

platforms, such as TripAdvisor, as well as to other types of social platforms. In case of online review platforms, the extension of our model is immediate. In fact, it is sufficient to know and report in our model the hierarchy of categories underlying the online review platform. In case of other types of social media, the extension is possible and quite simple. In fact, it is sufficient to specify a (possibly hierarchical) mechanism for dividing users into groups, as well as to identify the types of user relationships of interest. It seems quite obvious that friendship is a relationship of interest for any social platform. On the contrary, co-review does not always make sense and could be replaced by other types of relationships.

As for stereotypes, we observe that those considered here are not the only ones possible for an online review platform. In the future, we plan to identify other stereotypes and study their contribution to the extraction of useful knowledge from Yelp. At the same time, the three identified stereotypes can be directly extended to any other online review platform. The concept of power user can be easily extended to any social platform and any online social network too. The concept of bridge and double-life user can be extended only to those cases where users of a social platform can be organized into communities based on some parameters. In this case, a bridge is a user acting as a link between two communities, while a double-life user is a user having different behaviors in different communities.

The last theoretical contribution concerns the definition of the Negative Reviewer Network. This model plays an extremely important role in the study of negative reviews and, above all, in the identification of negative influencers, who correspond to nodes with high degree centrality and/or high eigenvector centrality, as we have seen in Section 3.2.4.5. Analogously to what happens for the other theoretical tools, the extension of this model to other online review platforms is immediate. Instead, its extension to other types of social platforms is much less simple than the other models and concepts seen above. In fact, by its nature, the Negative Reviewer Network is specifically designed to model negative reviews and reviewers. Therefore, its extension is only possible by identifying other negative behaviors that one wants to study and by defining a form of co-participation of multiple users to these behaviors.

3.2.5.4 Practical implications

Starting from the theoretical background, the hypotheses made and the implications confirming them, we can outline different applications of the knowledge here extracted to real life scenarios. In particular, we can identify two different perspectives, i.e., the business and the user ones.

The business perspective concerns all the possible actions that a company can take to expand its customer base, to improve its brand image or to extend the prod-

ucts/services it offers. In this context, the user identified stereotypes and the implications associated with them can be extremely useful. Let us consider, for example, k-bridges. We have seen the extremely important role that they play in disseminating information between different communities. In the previous sections, we have also seen that past literature highlights the strong impact that negative reviews can have. In this context, a k-bridge making a negative review could have a disruptive effect on a business image.

Therefore, the possibility of detecting k-bridges provided by our approach can become a valuable tool for a business, which can adopt a variety of policies aiming at improving their evaluation of its products/services from negative to neutral or, even, positive. Another extremely important policy in this sense could regard the promotion of a business to k-bridges who do not know it. This could favor the knowledge of this business in all the communities which the k-bridges belong to. In fact, a k-bridge belonging to a community where a business is well known and another community where this latter is unknown could become a promoter of the business from the former community to the latter one.

Another important application that could leverage k-bridges is the expansion of products/services offered by a business towards new categories, or even new macro-categories, of Yelp. One way to increase the chance of designing new products/services being of interest to users could be as follows. A business could identify all the k-bridges belonging to the categories in which it is already known and its products/services are highly appreciated. Then, it could determine the other categories of products/services where the identified k-bridges have performed revisions; in fact, the products/services of these last categories could be of interest for the potential customers of this business. The greater the number of k-bridges that have shown interest in these categories, the more likely customers belonging to them will be attracted by the business if it expands its offers towards these markets.

A further application of k-bridges, collateral to the one seen above, concerns advertising campaigns. In fact, knowing the most promising communities when proposing new products/services also implies being able to carry out advertising campaigns focusing on them. In this way, the effectiveness and efficiency of the advertisement activity in terms of time and costs are increased.

However, k-bridges are not the only identified stereotype having important practical applications. In fact, both power users and double-life users are equally important. Since the latter two stereotypes appear within the definition of negative influencers, we now see some possible applications of this last concept that subsumes the other two ones. Negative influencers have two important characteristics. The first concerns the high value of network centrality measures (degree centrality and/or

eigenvector centrality and/or PageRank), which makes them very influential in the communities where they operate. The second concerns their behavior in carrying out reviews. In fact, we have seen that a negative influencer, being a score-dl-user, tends to give positive reviews in the categories of lesser interest, while she is very demanding and severe in the categories in which she is more experienced and that interest her the most. This also assumes that such a user generally has a recognized leadership exactly in the category in which she is most severe. Therefore, it becomes crucial for a business in that category taking all possible actions to ensure that she takes a neutral, or hopefully a positive, attitude towards the products/services it offers. On the other hand, as we have seen for k-bridges, it is possible to think of targeted advertising and marketing actions on these users that, if successful, are characterized by a high level of efficiency and effectiveness.

So far we have seen the possible exploitations of our knowledge patterns from the business viewpoint. Now, we want to see how the same patterns can have practical implications for the user as well. In particular, we want to consider what benefits a user can get by looking at other relevant users (such as k-bridges, power users, influencers) in Yelp.

A first benefit can be obtained from the examination of the reviews of negative influencers in Yelp. Based on the knowledge we have extracted, we can assume that these users are very experienced in a certain category and very severe in exactly that category. Therefore, if these users have issued positive reviews on the products/services of a business in that category, it is very likely that they are of high quality.

A second benefit for a user concerns the knowledge of the features characterizing the profile of an influencer in Yelp. This knowledge becomes extremely useful if she wants to become an influencer in that social medium. In fact, based on the derived implications, the user knows that she has a better chance to become an influencer if she becomes a k-bridge. As a consequence, she will have to be active in making revisions in multiple categories. In addition, she should be a power user; therefore, she must have many friendship and co-review relationships (which implies she has a high degree centrality). Alternatively, she can have a limited number of friendship and co-review relationships as long as the users connected to her are, in turn, power users (which implies she has a high eigenvector centrality). Finally, she must identify one or more categories in which she wants to be an influencer and develop a high experience in them in order to give severe, but correct, reviews.

The knowledge here extracted can also be useful to define recommender systems for users who want to discover new products/services. This can be done, for example, by leveraging k-bridges. In fact, assume that a user follows some categories. It is possible to identify all the k-bridges of these categories and, for these k-bridges, to

consider the categories followed by them. In this way, it is possible to identify which categories are the most followed by these k-bridges. If one of these categories is not already followed by the user, it is possible to recommend it to her. This very general approach could be further refined by examining the proximity, in the Yelp hierarchy, of candidate categories to those already followed by the user. A further refinement could assign different weights to the different k-bridges, based on the similarity of their past evaluation to those of the user of interest on the same products/services, or based on the number of categories already followed by both them and the user of interest.

3.2.5.5 Limitations and future research directions

Our theoretical tools (i.e., the multi-dimensional social network-based model of Yelp, the stereotypes and the Negative Review Network), together with the hypotheses formulated and the implications confirming them, have allowed us to shed light on the phenomenon of negative reviews and negative influencers in Yelp. The tools proposed and the approach followed are sufficiently general to be extended directly to other online review platforms and, after some generalizations, to any social platform. However, they are to be considered simply as a first step in this direction, because they are not free from limitations, whose knowledge paves the way to new future research investigations.

The first limitation of our approach is that it is exclusively structural and does not take semantics into account. Actually, a more focused study on the contents of negative reviews would be necessary to understand the reasons that led users to formulate them. This would increase the effectiveness and efficiency of the applications of our approach discussed in Section 3.2.5.4. In fact, given a service/product receiving many negative reviews, we could strive to understand the main reasons for this fact and, therefore, make the appropriate improvements aimed at satisfying as many users as possible in the shortest time.

An in-depth semantic analysis of reviews would also be extremely useful to define one or more taxonomies of negative influencers. This would allow us to classify them based not only on the products/services they criticize, as in the present approach, but also on the main reasons for negativity (which would give us several indications on where intervening first or mainly). Semantic knowledge would also allow us to better evaluate negative influencers in order to understand who give plausible reasons and who, instead, are prevented, regardless it happens. As a matter of fact, a business could make an effective and efficient recovery work on the former category of influencers, while it could decide not to intervene on the latter one, because the possibility of making them neutral or positive is low.

Another limitation of our approach, which is, at the same time, a potential future development of our research concerns stereotypes. Here, we have presented three of them, namely the k-bridges, the power users and the double-life users. Their identification was driven by our research needs. However, we believe that several other stereotypes could be defined and that it could be even possible to go so far as to define a real taxonomy of stereotypes for both Yelp and other online (review) platforms. These would become a real toolbox available to decision makers when they need to make decisions regarding the products/services provided by their business (for instance, to determine those ones to be removed from catalogues, new ones to be proposed, existing ones to be modified for making them more in line with user needs and desires, etc.).

A third limitation of our approach, which is also linked to current technological limitations expected to become less impacting in the future, concerns the possibility of studying all these phenomena over time. In fact, our current approach is based on a temporal (albeit wide) photograph of the negative reviews of Yelp. It is not incremental and, if we want to study the evolution of a phenomenon over time, we should take more datasets referring to different times and study them separately. However, this does not allow us to have a continuous monitoring of the phenomenon, in order to capture any changes regarding it (for instance, any change of how some products/services are perceived by users) as soon as possible. The weight of this limitation (and, consequently, the relevance of overcoming it) is smaller in substantially stable socio-economic conditions, because user perceptions of products/services change very slowly over time in this scenario. Instead, it becomes crucial in historical periods characterized by sudden and disruptive phenomena (think, for instance, of the current COVID-19 pandemic), capable of upsetting all previous mental patterns of people's judgement. In this case, having the possibility of immediately understanding the changed perceptions of users about products/services and/or the appearance of new needs, with the consequent demand for new products/services, can allow a business to gain a huge advantage over its competitors. More importantly, this feature would allow the whole ecosystem of public and private product/service providers to be efficient and effective in responding to people demands.

Part II

Internet of Things

In this part, we model the Internet of Things (i.e., IoT) through our complex-network based approach and the Multiple Internet of Things (i.e., MIoT) paradigm already proposed in the past literature. This last allowed us to study the IoT as a set of device networks interacting with each other, which is the foundation for developing approaches to address some of the IoT common issues. This part is organized as follows: in Chapter 4, we introduce some preliminary concepts about the MIoT paradigm. In Chapter 5, we propose two solutions for improving the communication between the devices, thanks to the concept of topic-driven virtual IoTs and the new MIoT-oriented centrality measure and investigate the influence of these devices in the MIoT. Then, in Chapter 6, we define an approach to compute the trust and reputation of the devices. In Chapter 7, we describe a framework to ensure the privacy of the features and services provided by smart objects. Finally, in Chapter 8, we firstly model the possible device anomalies in a MIoT, and then we illustrate an approach to detect them.

4

Preliminary Concepts on Multiple Internet of Things

In this chapter, we report a brief introduction to the Multiple Internet of Things (i.e., MIoT) paradigm, which is the starting point for our next approaches. Specifically, we highlight the motivations behind the definition of the MIoT and formally introduce it. Then, we report an example of a MIoT and present its strengths with respect to the classical view of Internet of Things.

4.1 Introduction

The Internet of Things can be considered as an evolution of the Internet, based on the pervasive computing concept [73]. In the past, several strategies to implement the IoT paradigm and to guarantee ubiquitous computing have been proposed [310, 716, 233]. One of the most effective of them is based on the use of the social networking paradigm [70, 74, 71]. In this case, IoT is represented as a social network and, thanks to this association, Social Network Analysis-based models can be used to empower IoT. One of the most advanced attempts in this direction is SIoT (Social Internet of Things). In SIoT, things are empowered with social skills, making them more similar to people [70, 74]. In particular, they can be linked by five kinds of relationship, namely: *(i)* parental object relationship; *(ii)* co-location object relationship; *(iii)* co-work object relationship; *(iv)* ownership object relationship; *(v)* social object relationship. If: *(i)* a node is associated with each thing, *(ii)* an edge is associated with each relationship between things, and, finally, *(iii)* all the nodes and the edges linked by the same relationship are seen as joined together, SIoT can be modeled as a set of five pre-defined networks. Here, some nodes belong to only one network (we call them inner-nodes), whereas other ones belong to more networks (we call them cross-nodes).

The idea underlying SIoT is extremely interesting and, as a matter of fact, has received, and is still receiving, a lot of attention in the literature. However, we think that, in the next future, the number of relationships that might connect things could

be much higher than five, and relationships could be much more variegate than the ones currently considered by SIoT. As a consequence, we think that a new paradigm, taking into account this fact, is in order.

In [134, 514], we introduced the concept of Social Internetworking System (SIS, for short) as a system comprising an undefined number of users, social networks and resources. The SIS paradigm was thought to extend the Single Social Network paradigm by taking into account that: *(i)* a user can join many social networks, *(ii)* these joins can often vary over time, and *(iii)* the presence of users joining more social networks can favor the cooperation of users, who do not join the same social networks. We think that the key concepts of SIS can also be applied to things (instead of to users) and to relationships between things and so we propose the MIoT (Multiple Internets of Things) paradigm. The core of the SIS paradigm is modeling users and their relationships as a unique big network and, at the same time, as a set of related social networks connected to each other thanks to those users joining more than one social network. Here, we propose to extend the ideas underlying the concept of SIS to IoT. The MIoT paradigm arises as a result of this objective.

Roughly speaking, a MIoT can be seen as a set of things connected to each other by relationships of any kind and, at the same time, as a set of related IoTs, one for each kind of relationship. Actually, a more precise definition of MIoT would require the introduction of the concept of instance of a thing in an IoT. According to this concept, the instance of a thing in an IoT represents a virtual view of that thing in the IoT. Having this in mind, a MIoT can be seen as a set of related IoTs, one for each kind of relationship into consideration. The nodes of each IoT represent the instances of the things participating to it. As a consequence, a thing can have several instances, one for each IoT to which it participates. As will be clear in the following, the existence of more instances for one thing plays a key role in the MIoT paradigm because it allows the definition of the cross relationships among the different IoTs of the MIoT.

Differently from SIoT, in the MIoT paradigm, the number of relationships is not defined a priori. In a MIoT, there is a node for each thing; furthermore, there is an edge between two nodes if the corresponding things are linked by a relationship. If more kinds of relationship exist between two things, then more edges exist between the corresponding nodes, one for each kind of relationship. All the nodes linked by a given kind of relationship, together with the corresponding edges, form an IoT of the MIoT.

Observe that, under this MIoT definition, SIoT can be seen as a specific case of MIoT in which the number of the possible kinds of relationship is limited to 5 and these kinds are pre-defined. IoTs are interconnected thanks to those nodes corre-

sponding to things involved in more than one kind of relationship. We call *cross nodes* (*c-nodes*, for short) these nodes and *inner nodes* (*i-nodes*, for short) all the other ones. Then, a c-node connects at least two IoTs of the MIoT and plays a key role to favor the cooperation among i-nodes belonging to different IoTs. As a consequence, differently from SIoT, the nodes of a MIoT are not all equal: c-nodes will presumably play a more important role than i-nodes for supporting the activities in a MIoT.

Note that the MIoT paradigm can be seen as an attempt to address an open issue evidenced in [71] about some improvements that should be made on the SIoT paradigm.

From a more applicative point of view, having some IoTs that can "communicate" through c-nodes can lead to some beneficial synergies. For instance, assume that an environment-related IoT can communicate with a home-related IoT through a cross node. Assume that the former IoT evidences an abnormal presence of dioxin in a place located some kilometers away from the home (for instance, owing to a fire of a plastic deposit). Assume, also, that this IoT is evidencing that the wind direction is pushing the dioxin towards the home. The home-related IoT could be "informed" through a cross node about this fact and could close all windows before the arrival of the dioxin.

Once a MIoT has been defined, it is possible to apply Social Network Analysis-based techniques on it to extract powerful knowledge concerning its things, their relationships, the IoTs formed by them, etc.

This chapter is organized as follows: in Section 4.2, we present the MIoT paradigm. In Section 4.3, we present an example of a MIoT and, finally in Section 4.4, we present some reasoning behind the choice of the MIoT as a reference model.

4.2 MIoT paradigm

We define a MIoT $\mathcal{M}$ as a set of m Internets of Things (see Figure 4.1 for a schematic representation of it)[1]. Formally speaking:

$$\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$$

where $\mathcal{I}_k$ is an IoT.

Let o_j be an object of $\mathcal{M}$. We assume that, if o_j belongs to $\mathcal{I}_k$, it has an instance ι_{j_k}, representing it in $\mathcal{I}_k$. As pointed out in the Introduction, the instance ι_{j_k} indicates a

[1] The term "IoT" is intended according to the new trends that characterize this research field [71]. These trends suggest that, with the explosion of the number of available things, it is not realistic to talk about a unique Internet of Things. By contrast, it is more appropriate to consider several IoTs, each consisting of a (social) network of things.

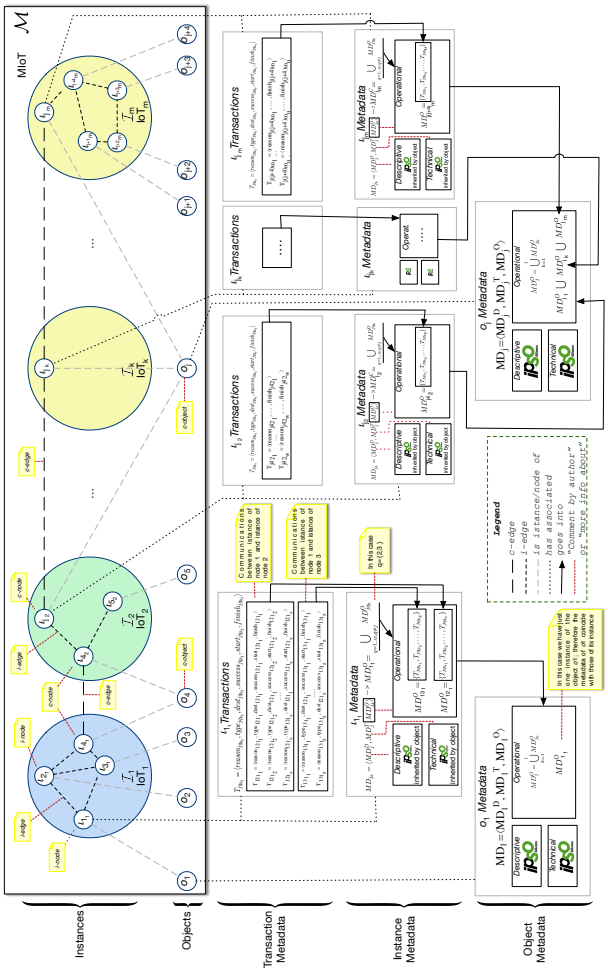


Fig. 4.1: Schematic representation of the proposed MIoT structure

virtual view (or, better, a virtual agent) representing o_j in $\mathcal{I}_k$. For instance, it provides all the other instances of $\mathcal{I}_k$, as well as the users interacting with $\mathcal{I}_k$, with all necessary information about o_j. Interestingly, this information is represented according to the format and the conventions adopted in $\mathcal{I}_k$.

In $\mathcal{M}$, *a set* MD_j *of metadata* are associated with an object o_j. We define a rich set of metadata of an object, because these play a key role in favoring the interoperability of IoTs and of their objects, which is the main objective of a MIoT. As a consequence, MD_j consists of three different subsets:

$$MD_j = \langle MD_j^D, MD_j^T, MD_j^O \rangle$$

Here:

- MD_j^D represents the set of *descriptive metadata*. It denotes the type of o_j. For representing and handling descriptive metadata, a proper taxonomy, such as the one defined by the IPSO Alliance [5], can be adopted.

- MD_j^T represents the set of *technical metadata*. It must be compliant with the object type. In other words, there is a different set of metadata for each object type of the taxonomy. Also in this case, the IPSO Alliance provides a well defined set of technical metadata for each object type. It is worth pointing out that, in principle, we could have allowed much richer descriptive and technical metadata. However, we did not make this choice because we preferred to relate our definition of metadata to an international IoT standard, such as the one defined by the IPSO Alliance. Furthermore, as will be clear in the following, our approach needs mainly operational metadata. As a consequence, making descriptive and technical metadata more complex would have added a useless level of complexity to our model.
- MD_j^O represents the set of *operational metadata*. It regards the behavior of o_j. The operational metadata of an object o_j is defined as the union of the sets of the operational metadata of its instances. Specifically, let $\iota_{j_1}, \iota_{j_2}, \ldots, \iota_{j_l}$, $l \leq m$, be the instances of o_j belonging to the IoTs of $\mathcal{M}$. Then:

$$MD_j^O = \bigcup_{k=1}^{l} MD_{j_k}^O$$

$MD_{j_k}^O$ is the set of the operational metadata of the instance ι_{j_k}. In order to understand the structure of $MD_{j_k}^O$, we first have to analyze the structure of $MD_{jq_k}^O$, i.e. the set of operational metadata between two instances ι_{j_k} and ι_{q_k}, of the objects o_j and o_q, in the IoT $\mathcal{I}_k$.

Specifically, $MD_{jq_k}^O$ is given by the set of metadata associated with the transactions between ι_{j_k} and ι_{q_k}. In particular:

$$MD_{jq_k}^O = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \ldots, T_{jq_{k_v}}\}$$

where $T_{jq_{k_t}}$, $1 \leq t \leq v$, represents the metadata of the t-th transaction between ι_{j_k} and ι_{q_k}, assuming that v is the current number of transactions between the two instances.

$T_{jq_{k_t}}$ can be represented as follows:

$$T_{jq_{k_t}} = \langle reason_{jq_{k_t}}, type_{jq_{k_t}}, inst1_{jq_{k_t}}, inst2_{jq_{k_t}}, success_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}} \rangle$$

where:

– $reason_{jq_{k_t}}$ denotes the reason causing the transaction, chosen among a set of default values.
– $type_{jq_{k_t}}$ indicates the transaction type (e.g., unicast, multicast, and so forth).

- $inst1_{jq_{k_t}}$ and $inst2_{jq_{k_t}}$ denote the two instances involved in $T_{jq_{k_t}}$. Observe that a transaction between ι_{j_k} and ι_{q_k} could be part of a longer path whose source and/or target nodes could be different from ι_{j_k} and ι_{q_k}. In principle, the source and/or the target nodes of a transaction could belong to an IoT different from $\mathcal{I}_k$. In this last case, it is necessary to reach $\mathcal{I}_k$ from the source, and/or to reach the target from $\mathcal{I}_k$, through one or more cross nodes, if possible.
- $success_{jq_{k_t}}$ denotes if the transaction succeeded.
- $start_{jq_{k_t}}$ is the timestamp associated with the beginning of the transaction.
- $finish_{jq_{k_t}}$ is the timestamp associated with the end of the transaction (its value is NULL if $T_{jq_{k_t}}$ failed).

In our model, the direction of a transaction is not considered. Furthermore, the parameter v, i.e., the number of transactions for each pair of instances, varies when moving from a pair of instances to another.

Observe that we have made our model powerful enough to represent and handle all the transactions between two instances of each IoT. Having all these detailed historical data at disposal could help the analysis of the real "social" behavior of each object. Furthermore, these data could be exploited in many applications; think, for instance, of the computation of the trust and reputation of each object, the investigation of objects with similar or complementary behaviors, and so forth. On the other hand, maintaining a full history of transactions may be very expensive and useless in many real life applications; in some cases, suitable data summarizations could be enough. As a consequence, when passing from the abstract model definition to real life applications, the transaction representation could be removed, extended or restricted on the basis of a tradeoff between costs and benefits for the current application.

We are now able to define the set of the operational metadata MD^O_{jk} of an instance ι_{j_k} of $\mathcal{I}_k$. Specifically, let $\iota_{1_k}, \iota_{2_k}, \ldots, \iota_{w_k}$ be all the instances belonging to $\mathcal{I}_k$. Then:

$$MD^O_{jk} = \bigcup_{q=1..w, q \neq j} MD^O_{jq_k}$$

In other words, the set of the operational metadata of an instance ι_{j_k} is given by the union of the sets of the operational metadata of the transactions between ι_{j_k} and all the other instances of $\mathcal{I}_k$.

Given an instance ι_{j_k}, relative to an object o_j and an IoT $\mathcal{I}_k$, we define the metadata MD_{j_k} of ι_{j_k} as:

$$MD_{j_k} = \langle MD^D_j, MD^T_j, MD^O_{j_k} \rangle$$

In other words, the descriptive and the technical metadata of an instance ι_{j_k} coincide with the ones of the corresponding object o_j. Instead, the operational metadata

of ι_{j_k} is a subset of the operational metadata of o_j that comprise only those ones regarding the transactions, which ι_{j_k} is involved in.

It is possible to associate a graph:

$$G_k = \langle N_k, A_k \rangle$$

with $\mathcal{I}_k$. Here, N_k indicates the set of the nodes of $\mathcal{I}_k$. There is a node n_{j_k} for each instance ι_{j_k} of an object o_j in $\mathcal{I}_k$. A_k denotes the set of the edges of $\mathcal{I}_k$. There is an edge $a_{jq_k} = (n_{j_k}, n_{q_k})$ if there exists a link between the instances ι_{j_k} and ι_{q_k} of the objects o_j and o_q in the IoT $\mathcal{I}_k$.

Also the overall MIoT $\mathcal{M}$ can be represented as a graph:

$$\mathcal{M} = \langle N, A \rangle$$

Here:

- $N = \bigcup_{k=1}^{m} N_k$;
- $A = A_I \cup A_C$, where:
  - $A_I = \bigcup_{k=1}^{m} A_k$;
  - $A_C = \{(n_{j_k}, n_{j_q}) | n_{j_k} \in N_k, n_{j_q} \in N_q, k \neq q\}$; observe that n_{j_k} and n_{j_q} are the nodes corresponding to the instances ι_{j_k} and ι_{j_q} of the object o_j in $\mathcal{I}_k$ and $\mathcal{I}_q$.

In other words, a MIoT $\mathcal{M}$ can be represented as a graph whose set of nodes is the union of the sets of nodes of the corresponding IoTs. The set A of the arcs of $\mathcal{M}$ consists of two subsets, A_I and A_C. A_I is the set of the inner arcs of $\mathcal{M}$ and is the union of the sets of the arcs of the corresponding IoTs. A_C is the set of the cross arcs of $\mathcal{M}$; there is a cross arc for each pair of instances of the same object in different IoTs. We call:

- *i-edge* an edge of $\mathcal{M}$ belonging to A_I;
- *c-edge* an edge of $\mathcal{M}$ belonging to A_C;
- *c-node* a node of $\mathcal{M}$ involved in at least one c-edge;
- *i-node* a node of $\mathcal{M}$ not involved in any c-edge;
- *c-object* an object having at least one pair of instances whose corresponding nodes are linked by a c-edge; clearly, any object with at least two different instances is a c-object.

It is worth pointing out that, as mentioned in the Introduction, there is a strict correlation between the MIoT paradigm and the concept of Social Internetworking System (hereafter, SIS) already presented in the literature [134]. In particular: *(i)* the concept of c-edges shares several features with the one of "me"-edge in a SIS; *(ii)* the concept of c-node is similar to the one of bridge in a SIS; *(iii)* a c-object corresponds to a user joining more social networks.

4.3 Example of a MIoT

Since the MIoT paradigm is new, in the Internet there is no known case study or real example about it yet. As a consequence, to provide the reader with an example, and, at the same time, to have a testbed for our experiments, we constructed a MIoT starting from some open data about things available on the Internet. In particular, we derived our data from *Thingful* [1]. This is a search engine for the Internet of Things, which allows us to search among a huge number of existing things, distributed all over the world. Thingful also provides some suitable APIs allowing the extraction of all the data we are looking for.

In order to construct our MIoT, we decided to work with 250 things whose data was derived from Thingful. Given the huge number of things available in Thingful, it could appear that the number of things composing our testbed is excessively limited. However, we observe that this was the first attempt to construct a real MIoT and, then, it was extremely important for us to have a full control of it in order to verify if we were proceeding well. A full human control with a much higher number of nodes was not possible.

We considered three dimensions of interest for our MIoT, namely:

a. *Category*: It specifies the application field which a given thing operates in. The categories we have chosen were five, namely *home*, *health*, *energy*, *transport*, and *environment*. Each category originated an IoT. Each thing was assigned to exactly one category.
b. *Coastal distance*: It specifies the coastal distance (i.e., the distance from any sea, lake or river) of each thing. The distance values we have set were:
   - *near*, for things distant less than 20 kilometres from the coast, for the categories *environment* and *energy*, and less than 5 kilometres, for the other three categories;
   - *mid*, for things whose minimum distance from the coast was between 20 and 105 kilometres, for the categories *environment* and *energy*, and between 5 and 25 kilometres, for the other three categories;
   - *far*, for things whose minimum distance from the coast was higher than 105 kilometres, for the categories *environment* and *energy*, and higher than 25 kilometres, for the other three categories.

   An IoT was created for each distance value. The different coastal distance values for *environment* and *energy*, on the one hand, and for the other three categories, on the other hand, have been determined after having analyzed the distribution of the involved categories of things against the coastal distance, in such a way

IoT	*Number of instances*
a.home	22
a.health	22
a.energy	22
a.transport	22
a.environment	22
b.near	14
b.mid	38
b.far	53
c.plain	44
c.hill	50
c.mountain	6

Table 4.1: Number of instances present in the IoTs of our MIoT

as to produce a uniform distribution of each category of things in the three IoTs related to the coastal distance dimension.

c. *Altitude*: it specifies the altitude of the place where the thing is located. The altitude values we have defined were: *plain* (corresponding to an altitude less than 500 meters), *hill* (corresponding to an altitude between 500 and 1000 meters), and *mountain* (corresponding to an altitude higher than 1000 meters). An IoT was created for each altitude value.

As a consequence, our MIoT consists of 11 IoTs. We associated an object with each thing; therefore, we had 250 objects. In principle, for each object, we could have associated an instance for each dimension. However, in order to make our testbed closer to a generic MIoT, representing a real scenario, where it is not said that all the objects have exactly the same number of instances, we decided not to associate three instances with each object. Instead, we associated only one instance (distributed uniformly at random among the three dimensions, and based on the features of the things of the IoTs of a given dimension) to 200 of the 250 objects. Analogously, we associated two instances (distributed by following the same guidelines mentioned above) to 35 of the 250 objects. Finally, we associated three instances, one for each possible dimension, to 15 of the 250 objects. At the end of this phase, we had 315 instances, distributed among the 11 IoTs of our MIoT as shown in Table 4.1.

To complete our MIoT and its network representation, we had to define a policy to create *i-edges*. In fact, it was clear that our MIoT should have had a node for each instance and a c-edge for each pair of instances referring to the same object. Therefore, the last decision regarded how to define i-edges. Given our scenario, it ap-

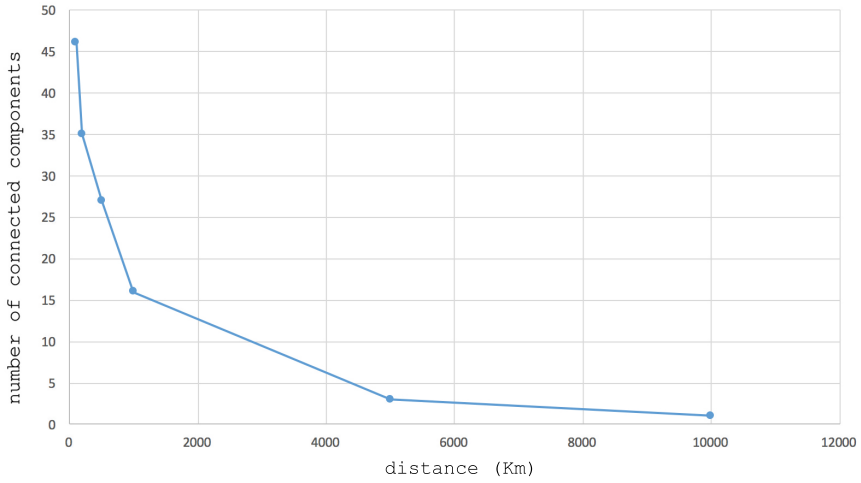


Fig. 4.2: Distribution of the number of connected components of the instances of our MIoT against distances

peared reasonable to consider distances among things as the leading parameter for the creation of i-edges. To carry out this last task, we have preliminarily computed the distribution of the number of connected components possibly created from our instances against the maximum possible distance. Obtained results are reported in Figure 4.2. Based on this figure, in order to obtain a balanced number of connected components, we decided to connect two instances of the same IoT if the distance of the corresponding things was lesser than 1000 kilometres.

After this last choice, our MIoT was fully defined. In order to help the reader to mentally portray it, in Figure 4.3, we provide a graphical representation. The interested reader can find the corresponding dataset (in the `.csv` format) at the address `www.barbiana20.unirc.it/miot/datasets/miot1`. The password to type is "`za.12&;1q74:#`".

4.4 MIoT strengths

In the Introduction, we have specified that the MIoT paradigm goes in the direction suggested by some authors, who observe that it is no longer possible to think of a single global Internet of Things [71].

In this section, we present a case study aiming at comparing the classical vision of a unique global Internet of Things with the new MIoT-based vision of several Internets of Things connected to each other through cross nodes and cross edges. In our opinion, this case study can help the reader to be convinced of the practicality of the MIoT paradigm.

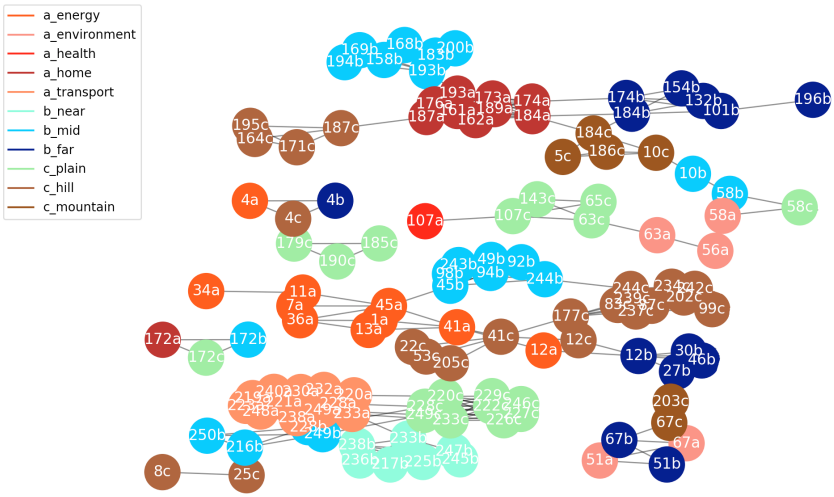


Fig. 4.3: Graphical representation of our MIoT

First, we must clarify that a slavish comparison between the previous vision of IoT and the MIoT-based vision is not possible, because this last paradigm associates more instances with the same object, one for each network joined by it. By contrast, the classical global IoT-based vision considers only objects and does not allow the existence of more instances of the same object. In other words, the global IoT-based vision returns a coarser model of the involved things and their relationships, incapable of verifying if the same object shows different features or behaviors in different subnetworks of the global network. Vice versa, this verification is not only possible, but also natural, in the MIoT paradigm. Indeed, it is sufficient to investigate the different features and behaviors of the various instances of the same object in the IoTs they belong to.

After having made this important premise, which already represents a justification of the usefulness of the MIoT paradigm, we start by presenting our case study by which we aim at showing that the global IoT-based vision can provide imprecise information about the features and the roles of the corresponding things.

Since the global IoT-based vision does not consider object instances, in this case study we assume that all the instances of a cross object have been merged in a unique c-node.

With these considerations in mind, let us consider Figure 4.4. Here, we report a set of nodes each associated with an object. If we consider the global IoT-based vision, all these nodes form a unique IoT where it is possible to distinguish two quite separated subnetworks, called S_1 and S_2 in the figure, connected only thanks to the object represented by Node 1. If we consider the MIoT-based vision, we have two IoTs connected, by means of the object represented by Node 1, to form a MIoT.

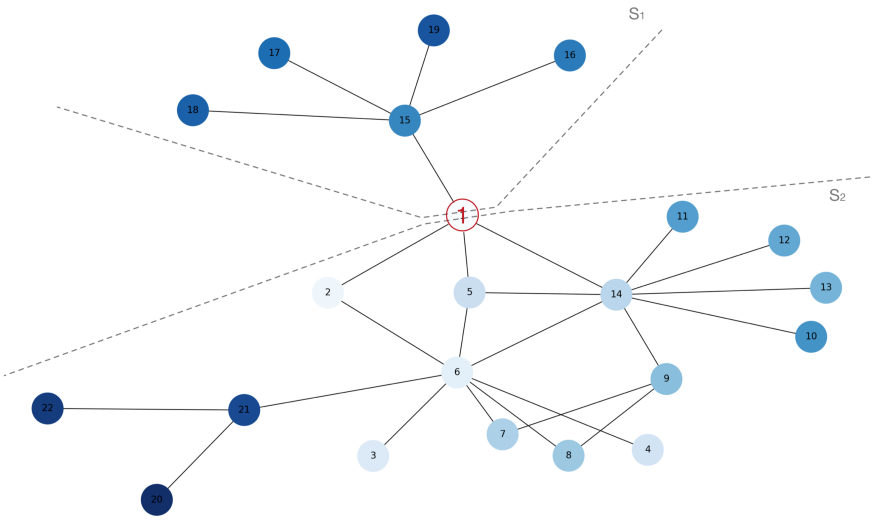


Fig. 4.4: Our case study

Let us focus our attention on this node. Clearly, it is the most important node of this scenario because it is the only one allowing the communication and the co-operation between the nodes of the subnetwork S_1 and the ones of the subnetwork S_2.

However, if we compute the classical centrality measures for the nodes of this network, we have that the rank of Node 1 is not very high in any centrality measure (see Table 4.2). In other words, if we adopt the global IoT-based vision, no centrality measure is capable of capturing the importance of this node. By contrast, the MIoT paradigm is capable alone of intrinsically evidencing the key role played by Node 1, without the need of computing any centrality measure.

With regard to this last observation, we are also aware that, in a real scenario, where the IoTs composing a MIoT are many and the number of c-objects is high, it could be extremely challenging to define a new MIoT-oriented centrality measure. This should be capable of determining the most relevant nodes in a MIoT taking also (but not exclusively) into account if they are c-nodes or not. We will address this issue in Chapter 5.2 where we propose a MIoT-oriented centrality measure.

Nodes	*Betweenness Centrality*	*Degree Centrality*	*Closeness Centrality*	*Eigenvector Centrality*
1	0.39 (3)	0.19 (4)	0.44 (4)	0.30 (4)
2	0.07 (6)	0.09 (8)	0.41 (5)	0.20 (6)
3	0.00 (11)	0.05 (11)	0.33 ()	0.13 (14)
4	0.00 (12)	0.05 (12)	0.33 ()	0.13 (15)
5	0.07 (7)	0.14 (6)	0.47 (3)	0.34 (3)
6	0.52 (1)	0.38 (1)	0.48 (2)	0.34 (2)
7	0.01 (9)	0.09 (9)	0.34 ()	0.19 (7)
8	0.01 (10)	0.09 (10)	0.34 ()	0.19 (8)
9	0.04 (8)	0.14 (7)	0.37 (6)	0.23 (5)
10	0.0 (13)	0.04 (13)	0.35 (9)	0.13 (10)
11	0.0 (14)	0.04 (14)	0.35 (10)	0.13 (11)
12	0.0 (15)	0.04 (15)	0.35 (11)	0.13 (12)
13	0.0 (16)	0.04 (16)	0.35 (12)	0.13 (13)
14	0.48 (2)	0.38 (2)	0.52 (1)	0.49 (1)
15	0.35 (4)	0.23 (3)	0.35 (7)	0.11 (16)
16	0.0 (17)	0.05 (17)	0.26 (17)	0.03 (19)
17	0.0 (18)	0.05 (18)	0.26 (18)	0.03 (20)
18	0.0 (19)	0.05 (19)	0.26 (19)	0.03 (21)
19	0.0 (20)	0.05 (20)	0.26 (20)	0.03 (22)
20	0.0 (21)	0.05 (21)	0.26 (21)	0.04 (17)
21	0.18 (5)	0.14 (5)	0.35 (8)	0.15 (9)
22	0.0 (22)	0.05 (22)	0.26 (22)	0.04 (18)

Table 4.2: Betweenness Centrality, Degree Centrality, Closeness Centrality and Eigenvector Centrality, and the corresponding ranks, for all the nodes of the case study of Figure 4.4

5

Communication and Influence Investigation

In the Internet of Things (i.e., IoT), we have thousands of devices that can connect with each other and exchange information. In the next years, we expect a further huge growth of the IoT and so we need to optimize the networks in order to decrease the time to reach a specific device (and also save some battery power). In this chapter, we report our contributions for this issue. First of all, we introduce the concept of profile of a thing in a MIoT. Then, we define the concept of topic-guided virtual IoT, along with two approaches to construct topic-guided virtual IoTs. As a second contribution, since the classical betweenness centrality is not able to correctly evaluate the centrality of nodes in a MIoT scenario, where several networks of smart objects cooperate with each other, we introduce a new betweenness centrality that is MIoT-oriented. Finally, as a third contribution, starting from the content exchanged in a transaction between two devices, we investigate the scope of a thing in a MIoT scenario. Specifically, we define the concept of scope and then, we propose two formalizations allowing its computation. Afterwards, we present two possible applications of scope and a set of experiments performed for its evaluation.

The material present in this chapter is taken from [434, 170, 163, 164].

5.1 Topic-driven virtual IoTs in a MIoT

5.1.1 Introduction

The Internet of Things (hereafter, IoT) is currently considered the new frontier of the Internet. As a matter of fact, a lot of research results, along with the continuous emergence of increasingly challenging issues to address, can be found in the literature [310, 585, 233, 533, 68, 303, 407].

One of the most effective ways to represent and handle the IoT scenario leverages social networking paradigm [62]. In this direction, several social network-based approaches to modeling and managing IoTs have been presented in the literature. Three of the most advanced ones are the SIoT (Social Internet of Things) [70, 259, 71, 581], the MIE (Multiple IoT Environment) [81] and the MIoT (Multiple

IoTs) [82] paradigms. The MIoT paradigm is the last of these proposals; it aims at extending both SIoT and MIE in such a way as to preserve their strengths and avoid their weaknesses [82]. Roughly speaking, a MIoT can be seen as a set of related IoTs, i.e., as a set of related networks of things. Actually, a more precise definition of MIoT requires the introduction of the concept of instance of a thing in an IoT. Specifically, the instance of a thing in an IoT represents a virtual view of that thing in the IoT. The nodes associated with a thing in a MIoT represent the instances of the same thing in the different IoTs of the MIoT. Indeed, a thing can have several instances, one for each IoT which it participates to. The existence of more instances for one thing plays a key role in the MIoT paradigm because it allows the definition of cross relationships among the different IoTs.

We adopted the MIoT paradigm as the reference model. There are several reasons which justify this choice. Indeed:

- The MIoT paradigm, like the SIoT and the MIE ones, introduces the idea that objects can show a social behavior in the environment where they operate. This feature allows several advantages, like the possibility of resource sharing (see [259, 71, 581] for a comprehensive idea of these advantages).
- Differently from SIoT, which introduces a social behavior of objects but still models IoT as one huge network of objects extended worldwide, MIE, and much more MIoT, allow the "breakdown" of the whole huge IoT into multiple networks of smart objects interconnected with each other. This way to proceed is analogous to the evolution of social networking into social internetworking [134]. In particular, MIoT allows the management of situations in which the same object shows different behaviors in different networks it joined. Furthermore, MIoT makes an object to act as a bridge between two objects allowing them to communicate even if they belong to different networks and, therefore, are not directly connected with each other.

Another important trend characterizing the current IoT scenario regards the existence of increasingly sophisticated and intelligent things. These are becoming increasingly smart and social, as well as more and more capable of performing computations and storage on their own. Furthermore, they are increasingly connected to each other through more and more complex and sophisticated frameworks, often based on cloud and edge computing [259, 71, 581]. The new smart and social capabilities of things and of the environments handling their interoperability paves the way to a sort of "humanization" of things, i.e., to apply to things concepts and ideas typically considered prerogative of humans. One of them is certainly the presence of a profile of a thing. Indeed, if a thing interacts with other things and exchange data

with them, it is possible to determine what are the most common concepts handled by it and, based on them, to construct a corresponding profile. Analogously to the profile of a human, the one of a thing depends on its past behavior and on the profile of the other things with which it interacts. As a consequence, it could be possible to think about both a content-based and a collaborative-filtering approach to handling thing profiles.

Furthermore, starting from the real IoTs of a MIoT, it is possible to construct virtual communities of things, based on common interests. Once again, this is an attempt to transfer behaviors typical of humans to things. As a matter of fact, in Social Network Analysis, it is well recognized that, accordingly to the homophily concept [468, 610], humans tend to group together in communities sharing the same interests.

In the literature, a lot of efforts have been made to investigate human profiles and virtual communities of people, especially (but not only) in Social Network Analysis [591, 547]. Instead, these topics have been little investigated in the Internet of Things.

Here, we aim at providing a contribution in this direction. First of all, we introduce the concept of profile of a thing. As the profile of a human, the one of a thing has two components. The former denotes its past behavior and can be used, for instance, to support content-based recommendations. The latter reflects its neighbors, i.e., the other things with which it most frequently comes into contact; it can be exploited, for instance, to support collaborative filtering recommendations.

After this, we introduce the concept of topic-guided virtual IoTs in a MIoT and we propose two approaches (one supervised and one unsupervised) to the construction of them in a MIoT. Differently from the real IoTs of a MIoT, which may encompass things with very heterogeneous profiles, topic-guided virtual IoTs should include all and only those things whose profile refers to specific topics. The supervised approach requires a user to provide a set of keywords of her interest. It aims at constructing a thematic IoT comprising all the keywords specified by the user. If such an IoT does not exists, it returns more thematic IoTs that, in the whole, comprise all the keywords specified by the user. She can choose whether to accept this set of virtual IoTs or to modify her query. The unsupervised approach tries to partition a MIoT into a set of virtual IoTs characterized by the maximum internal cohesion (in terms of topics present in the profiles of the corresponding things) and the minimum external coupling. Virtual IoTs in a MIoT provide a logic representation of the objects of a MIoT, which is not based on real links but on the content exchanged by them. As will be clear in the following, this can favor the effectiveness of information exchange, the construction of communities of objects (and, possibly, of the

corresponding users) sharing the same interests and the suggestions of the objects most adequate to a given exigency.

This chapter is organized as follows: in Section 5.1.2, we examine related literature. In Section 5.1.3, we introduce our definition of a thing profile, and we propose our approaches to construct topic-guided virtual IoTs in a MIoT. Finally, in Section 5.1.4, we present our testbed and several experiments devoted to verifying the performance of our approach.

5.1.2 Related Literature

Since its introduction some years ago, the term "Internet of Things - IoT" has been associated with a huge variety of concepts, technologies and solutions [68, 73, 481, 541]. In the latest years, with the advent of new technologies, such as big data and social networking, the very definition of this term is continuously changing. What IoT will become in the future depends on the evolution of these technologies [646] and their interaction with several other ones, such as Information Centric Networks [623, 716, 717, 50, 554, 51, 539] and Cloud [233, 638, 366]. As a matter of fact, the strengths of these last ones are exactly the features necessary to overcome the weaknesses of the current IoT concept [688]. Some examples of this combination can be already found in the literature [259, 303, 664, 663].

The first attempts to apply social networking to the IoT domain can be found in [309, 509, 397, 334]. In these papers, the authors propose to use human social network relationships to share services provided by a set of things. An important step forward is performed in [70], where the SIoT paradigm is introduced. Here, the authors propose an approach to creating relationships among things, without requiring the owner intervention. Thanks to this idea, things can autonomously crawl the network to find services and resources of their interest provided by other things. In [74], the same authors clearly highlight what are the main strengths of SIoT. Specifically: *(i)* the SIoT structure can be dynamically modified to ensure network navigability and to find new things; *(ii)* scalability is guaranteed, like in human social networks; *(iii)* a level of trustworthiness among things can be established; *(iv)* the past social network approaches can be redefined to solve problems typical of the IoT context [520].

One of the major drawbacks of the current IoT scenario is the presence of different technologies and solutions proposed by independent vendors to enable networking among objects. This poses the basis to a subsequent set of issues ranging from concept matching to technical compatibility, if heterogeneous smart-object-network solutions should be involved in the creation of a unique interoperable IoT [516, 615]. In this research context, different works partially addressing and solving these prob-

lems have been proposed. Specifically, [286] presents a study on how ontologies and semantic data processing can be used to improve interoperability across heterogeneous IoT platforms. The authors consider two use cases, namely *Health Care* and *Transportation and Logistics*, and, for each of them, provide a survey on the main ontologies available to describe and generalize concepts and relations.

In [417], instead, the authors focus their attention on the definition of a new framework for a fully functional mobile ad-hoc social network. In this paper, the term "mobile ad-hoc social network" refers to an IoT made of mobile devices. Of course, communication between this type of objects may happen in such a wide range of modes so that the referring scenario can be considered as a constellation of mobile networks interacting with each other. Concepts from real social networks are borrowed to define user profiles, which are built starting from the objects they own and the social network they belong to. One of the main contributions of this proposal is the definition of a profile-matching strategy based on semantics.

Another contribution in the context of interoperability is the one proposed in [626]. Here, the authors illustrate a novel architecture in which objects interact with each other by leveraging an open source cloud platform. The interaction among smart devices is information-and-service-driven and can be performed in both a centralized and a peer-to-peer mode. In [720], the authors propose *Acrost*, a system capable of retrieving data spread among heterogeneous IoT platforms by leveraging topics and semantics awareness. To build the metadata, Acrost uses two methodologies: the former exploits regular expression-based approaches, whereas the latter makes use of random fields-based strategies.

In order to address the issues arising when the interoperability among heterogeneous IoTs must be guaranteed, another research line proposes the extension of the results concerning Social Internetworking [134, 514] (instead of social networking) to the Internet of Things. By following this strategy, the MIE (Multiple IoT Environment) [81] and the MIoT (Multiple IoTs) [82] paradigms have been proposed.

In [232], the authors present an approach to constructing a virtual data mart on which several knowledge discovery tasks can be performed. Clearly the kinds of virtual source constructed in the approach of [232] and in our own are very different. However, the general ideas underlying the two approaches are similar.

In the past, a lot of efforts have been made to investigate human profiles and virtual communities of people, especially (but not only) in Social Network Analysis ([591, 547] provide two surveys about these topics). Instead, these issues have been little investigated in the Internet of Things. Specifically, to the best of our knowledge, a comprehensive, high-level abstraction approach to building and managing a profile of a thing, which also takes into account the content it exchanges during its

interactions with other things, has not yet been proposed. Instead, some approaches focusing on community detection in IoT have been presented in the very recent literature. Even if they are very different (both in their purposes and in their ways proceed) from the ones of our approach, in the following we present an overview of some of them.

The approach of [666] uses structural information derived from the complex graph of an IoT to extract communities. It exploits a neighbor-based strategy to detect also overlapping communities. The approach of [367] uses data produced by sensors to define a multi-dimensional clustering. The obtained clusters are then mapped to communities of nodes in the original IoT network. To cope with the size of the data graph, the authors leverage state-of-the-art community detection approaches. Finally, they present a new community detection approach that enhances the Girvan-Newman algorithm by using hyperbolic network embedding.

Other works, instead, use knowledge from social networks to refine their results. As an example, [486] proposes a community definition strategy combining both IoT information and structural data coming from the social network (relationship among users), which object owners belong to. This approach does not consider semantics and contents, but leverages only network structure. A similar method is proposed in [86], even though here the strategy works in the opposite way. In fact, first communities are derived from structural information of owners' social networks and, then, objects are seen as resources available inside each community.

Finally, the authors of [396] propose a new community detection algorithm working in a Social Internet of Things (SIoT) scenario. To achieve their objective, they make use of three metrics, namely social similarity, preference similarity and movement similarity. Social similarity is defined according to the concept of cooperativeness and community interest proposed in [512]. Preference similarity takes into account resource and service preferences of the involved things in the network. Finally, movement similarity specifies how much and how long two or more nodes are spatially close.

In [485], the authors propose a community detection approach working on an architecture capable of integrating the Internet of Things and social networking. This approach assumes that two nodes belong to the same community only if they are at most one hop apart and have at least two mutual friends. In order to construct communities, it exploits graph mining techniques.

5.1.3 Methods

5.1.3.1 Definition of a thing profile

As pointed out in the Introduction, analogously to what happens for human profiles, the profile of a thing can have two components. The former registers its past behavior and is extremely useful for content-based recommendations; for this reason, we call it "content-based component" in the following. The latter registers the main features of those things with which it mostly interacted in the past and can be used for collaborative filtering recommendations; for this reason, we call it "collaborative filtering component" in the following.

In this section, we present a model for representing and handling a thing profile. This model is based on the MIoT paradigm that we described in Chapter 4.

Given a MIoT $\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$, and two instances ι_{j_k} of o_j and ι_{q_k} of o_q in $\mathcal{I}_k$, we can define the set $tranSet_{jq_k}$ of the transactions from ι_{j_k} to ι_{q_k} as follows:

$$tranSet_{jq_k} = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \cdots, T_{jq_{k_v}}\} \tag{5.1}$$

A transaction $T_{jq_{k_t}} \in tranSet_{jq_k}$ is represented as:

$$T_{jq_{k_t}} = \langle reason_{jq_{k_t}}, source_{jq_{k_t}}, dest_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}}, success_{jq_{k_t}}, content_{jq_{k_t}} \rangle \tag{5.2}$$

Here:

- $reason_{jq_{k_t}}$ denotes the reason why $T_{jq_{k_t}}$ occurred, chosen among a set of predefined values.
- $source_{jq_{k_t}}$ indicates the starting node of the path followed by $T_{jq_{k_t}}$.
- $dest_{jq_{k_t}}$ represents the final node of the path followed by $T_{jq_{k_t}}$.
- $start_{jq_{k_t}}$ denotes the starting timestamp of $T_{jq_{k_t}}$.
- $finish_{jq_{k_t}}$ indicates the ending timestamp of $T_{jq_{k_t}}$.
- $success_{jq_{k_t}}$ denotes whether $T_{jq_{k_t}}$ was successful or not; it is set to `true` in the affirmative case, to `false` in the negative one, and to `NULL` if $T_{jq_{k_t}}$ is still in progress.
- $content_{jq_{k_t}}$ indicates the content "exchanged" from ι_{j_k} to ι_{q_k} during $T_{jq_{k_t}}$. In its turn, $content_{jq_{k_t}}$ presents the following structure:

$$content_{jq_{k_t}} = \langle format_{jq_{k_t}}, fileName_{jq_{k_t}}, size_{jq_{k_t}}, topics_{jq_{k_t}} \rangle \tag{5.3}$$

Here:

  - $format_{jq_{k_t}}$ indicates the format of the content exchanged during $T_{jq_{k_t}}$; the possible values are: "audio", "video", "image" and "text".

- $fileName_{jq_{k_t}}$ denotes the name of the transmitted file.
- $size_{jq_{k_t}}$ indicates the size in bytes of the content.
- $topics_{jq_{k_t}}$ indicates the set of the content topics; it consists of a set of keywords representing the subjects exchanged during $T_{jq_{k_t}}$. It can be formalized as: $topics_{jq_{k_t}} = \{(kw^1_{jq_{k_t}}, nkw^1_{jq_{k_t}}), (kw^2_{jq_{k_t}}, nkw^2_{jq_{k_t}}), \ldots, (kw^w_{jq_{k_t}}, nkw^w_{jq_{k_t}})\}$. In other words, the set of the topics of the t^{th} transaction from ι_{j_k} to ι_{q_k} consists of w pairs; each pair consists of a keyword and the corresponding number of occurrences.

Now, we can define the set $tranSet_{j_k}$ of the transactions performed by ι_{j_k} in $\mathcal{I}_k$. Specifically, let $Inst_k$ be the set of the instances of $\mathcal{I}_k$. Then:

$$tranSet_{j_k} = \bigcup_{\iota_{q_k} \in Inst_k, \iota_{q_k} \neq \iota_{j_k}} tranSet_{jq_k} \tag{5.4}$$

In other words, the set $tranSet_{j_k}$ of the transactions performed by an instance ι_{j_k} is given by the union of the sets of the transactions from ι_{j_k} to all the other instances of $\mathcal{I}_k$.

After having defined $tranSet_{j_k}$, we must introduce the following operators:

- $\uplus$: it receives a set $\{entitySet_1, entitySet_2, \cdots, entitySet_t\}$ of entity sets and performs their union not eliminating the duplicates but reporting the number of their occurrences. Therefore, this operator returns a set of pairs $\{(entity_1, ne_1), (entity_2, ne_2), \cdots, (entity_w, ne_w)\}$ in which the pair $(entity_r, ne_r)$ indicates the r^{th} entity and the number of its occurrences. In counting it, $\uplus$ takes the presence of synonymies and homonymies into account. These properties can be computed (for terms, images, etc.) by applying the classical approaches proposed in the past literature [102, 227].
- $avgFileSize$: it receives a set of files and computes their average size.

We are now able to define the profile $\mathcal{P}_{jq_k}$ of the relationship existing between two instances ι_{j_k} and ι_{q_k}, which performed a set $tranSet_{jq_k} = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \cdots, T_{jq_{k_v}}\}$ of transactions. As we will see in the following, this profile plays a crucial role in the definition of the content-based component of a thing's profile and is indirectly used also in the definition of the collaborative filtering component of it. Specifically:

$$\begin{aligned}\mathcal{P}_{jq_k} = \langle & reasonSet_{jq_k}, sourceSet_{jq_k}, destSet_{jq_k}, avgSzAudio_{jq_k}, avgSzVideo_{jq_k}, \\ & avgSzImage_{jq_k}, avgSzText_{jq_k}, successFraction_{jq_k}, topicSet_{jq_k}\rangle\end{aligned} \tag{5.5}$$

where:

- $reasonSet_{jq_k} = \uplus_{t=1..v}(reason_{jq_{k_t}})$;

- $sourceSet_{jq_k} = \biguplus_{t=1..v}(source_{jq_{k_t}})$;
- $destSet_{jq_k} = \biguplus_{t=1..v}(dest_{jq_{k_t}})$;
- $avgSzAudio_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}audio\text{”}\}$;
- $avgSzVideo_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}video\text{”}\}$;
- $avgSzImage_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}image\text{”}\}$;
- $avgSzText_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}text\text{”}\}$;
- $successFraction_{jq_k} = \frac{|\{T_{jq_{k_t}} | T_{jq_{k_t}} \in tranSet_{jq_k}, success_{jq_{k_t}} = true\}|}{v}$;
- $topicSet_{jq_k} = \biguplus_{t=1..v}(topics_{jq_{k_t}})$.

If we introduce the operator $\bigsqcup$, which compactly represents the set of operations for obtaining a profile of a pair of instances $\mathcal{P}_{jq_k}$ starting from the corresponding transactions, we can formalize the previous tasks by means of only one operation as follows:

$$\mathcal{P}_{jq_k} = \bigsqcup_{t=1..v} T_{jq_{k_t}} \tag{5.6}$$

Now, let ι_{j_k} be the instance of the object o_j in the IoT $\mathcal{I}_k$. Let $Inst_{j_k}$ be the set of the instances of $\mathcal{I}_k$ with which ι_{j_k} performed at least one transaction in the past. In this case, we can define the content-based component of the profile $\mathcal{P}_{j_k}$ of ι_{j_k} as:

$$\mathcal{P}_{j_k} = \bigsqcup_{\iota_{q_k} \in Inst_{j_k}} \mathcal{P}_{jq_k} \tag{5.7}$$

Finally, let o_j be an object and let $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_l\}$ be the set of the IoTs which it participates to. Let $ObjInst_j$ be the instances of o_j in the IoTs of the MIoT. We can define the content-based component of the profile $\mathcal{P}_j$ of o_j as:

$$\mathcal{P}_j = \bigsqcup_{\iota_{j_k} \in ObjInst_j} \mathcal{P}_{j_k} \tag{5.8}$$

After having defined the content-based component of an instance and an object, in order to present the corresponding collaborative filtering components, we must introduce the concept of neighborhoods of an instance ι_{j_k} in an IoT $\mathcal{I}_k$. Specifically, the structural neighborhood $sNbh(\iota_{j_k})$ of ι_{j_k} is defined as:

$$sNbh(\iota_{j_k}) = sNbh^{out}(\iota_{j_k}) \cup sNbh^{in}(\iota_{j_k}) \tag{5.9}$$

where:

$$sNbh^{out}(\iota_{j_k}) = \{\iota_{q_k} | (n_{j_k}, n_{q_k}) \in A_I\} \tag{5.10}$$

$$sNbh^{in}(\iota_{j_k}) = \{\iota_{q_k} | (n_{q_k}, n_{j_k}) \in A_I\} \tag{5.11}$$

Furthermore, we can also define the behavioral neighborhood $bNbh(\iota_{j_k})$ of ι_{j_k} as:

$$bNbh(\iota_{j_k}) = bNbh^{out}(\iota_{j_k}) \cup bNbh^{in}(\iota_{j_k}) \tag{5.12}$$

where:

$$bNbh^{out}(\iota_{j_k}) = \{\iota_{q_k} | \iota_{q_k} \in sNbh^{out}(\iota_{j_k}), |tranSet_{jq_k}| > 0\} \tag{5.13}$$

$$bNbh^{in}(\iota_{j_k}) = \{\iota_{q_k} | \iota_{q_k} \in sNbh^{in}(\iota_{j_k}), |tranSet_{qj_k}| > 0\} \tag{5.14}$$

In other words, $bNbh(\iota_{j_k})$ consists of those instances directly connected to ι_{j_k} from the structural viewpoint that shared at least one transaction with ι_{j_k}.

We are now able to present the collaborative filtering component $\mathcal{P}'_{j_k}$ of the profile of an instance ι_{j_k} in $\mathcal{I}_k$. It can be defined as follows:

$$\mathcal{P}'_{j_k} = \bigsqcup_{\iota_{q_k} \in bNbh(\iota_{j_k})} (\mathcal{P}_{q_k} \sqcup \mathcal{P}'_{q_k}) \tag{5.15}$$

Clearly, this definition is recursive and an accurate computation would require the resolution of a system with a number of equations and variables equal to the number of instances. In real situations, as there could be thousands or millions of instances in a MIoT, the time necessary to solve this system may easily become unacceptable. As a consequence, it appears reasonable to consider an approximate definition of $\mathcal{P}_{q_k}$ that is much simpler to handle. It is formalized as:

$$\mathcal{P}'_{j_k} = \bigsqcup_{\iota_{q_k} \in bNbh(\iota_{j_k})} \mathcal{P}_{q_k} \tag{5.16}$$

After having introduced the two components of the profile of an instance ι_{j_k} of $\mathcal{I}_k$, we can combine them for defining the overall profile $\overline{\mathcal{P}_{j_k}}$ of ι_{j_k}. It is defined as the union of the profiles $\mathcal{P}_{j_k}$ and $\mathcal{P}'_{j_k}$ performed by means of the operator $\sqcup$:

$$\overline{\mathcal{P}_{j_k}} = \mathcal{P}_{j_k} \sqcup \mathcal{P}'_{j_k} \tag{5.17}$$

Finally, we can define the overall profile of an object o_j as follows:

$$\overline{\mathcal{P}_j} = \bigsqcup_{k=1..l} \overline{\mathcal{P}_{j_k}} \tag{5.18}$$

5.1.3.2 Approach to build topic-guided virtual IoTs

Supervised approach

The supervised approach for the construction of topic-guided virtual IoTs in a MIoT requires the user to specify a query Q consisting of some keywords of her interest.

It tries to construct a thematic virtual IoT in such a way that each of its instances contains at least one keyword of Q in the content-based component of its profile. If such a virtual IoT does not exist, our approach returns a minimal set of thematic IoTs that, on the whole, contain, in the content-based component of the profile of their instances, all the keywords specified by the user. In this last case, she can choose whether to accept this set of IoTs or modify her query.

Before describing in detail this approach, we must introduce a new operator J^* that represents a modified Jaccard coefficient, as we will see below.

J^* receives two sets of topics[1] $topicSet = \{(kw_1, nkw_1), (kw_2, nkw_2), \cdots, (kw_p, nkw_p)\}$ and $topicSet' = \{(kw'_1, nkw'_1), (kw'_2, nkw'_2), \cdots, (kw'_p, nkw'_p)\}$ and computes the Jaccard coefficient between them. In carrying out this task, it considers the number of occurrences of each keyword and its possible synonyms.

More formally, first it computes the set:

$$\begin{aligned} commonTS = \{&(kw, nkw + nkw') | (kw, nkw) \in topicSet, \\ &(kw', nkw') \in topicSet', kw \text{ is identical to or synonymous of } kw'\} \end{aligned} \tag{5.19}$$

Then, it computes the final result as:

$$J^*(topicSet, topicSet') = \frac{\sum_{(kw,nkw) \in commonTS} nkw}{\sum_{(kw,nkw) \in topicSet} nkw + \sum_{(kw',nkw') \in topicSet'} nkw'} \tag{5.20}$$

After having introduced J^*, we can describe our approach. Specifically:

- It starts when a user specifies a query Q consisting of r keywords:

$$Q = \{kw_1, kw_2, \cdots, kw_r\} \tag{5.21}$$

  It searches for all the instances of the MIoT having at least one topic whose keyword is identical to, or synonymous of, at least one keyword specified in Q. These instances, as a whole, represent the set of candidate instances to be included in the new thematic view. We call this set $\mathcal{CI}$ (Candidate Instances).
- However, the fact that an instance $\iota \in \mathcal{CI}$ has a keyword in common with Q is necessary but not sufficient for it to be chosen. In fact, it is advisable that ι has more keywords in common with Q and, possibly, that the common keywords are among the ones of ι with the highest number of occurrences. This condition can be guaranteed by the usage of the operator J^*.

  In particular, our approach first constructs $Q' = \{(kw, 1) | kw \in Q\}$ in such a way as to make the application of J^* on the keywords specified by the user possible.

[1] We recall that, in our context, a topic is a pair (kw, nkw), where kw is a keyword and nkw is the corresponding number of occurrences.

Then, it constructs the set $\mathcal{RI}$ (Real Instances) of those instances of $\mathcal{CI}$ whose topics have a significant similarity with the keywords of Q:

$$\mathcal{RI} = \{\iota \in \mathcal{CI} | J^*(topicSet_\iota, Q') > th_J\} \tag{5.22}$$

Here, th_J is a suitable tuning threshold.

- Now, our approach can start to construct the thematic view $\mathcal{V}_Q$ corresponding to Q.
  - It first creates a node n_ι in $\mathcal{V}_Q$ for each instance ι of $\mathcal{RI}$. Let n_{ι_1} and n_{ι_2} be the nodes corresponding to two instances ι_1 and ι_2 belonging to $\mathcal{RI}$.
    - If an i-arc exists between the nodes corresponding to ι_1 and ι_2 in the MIoT $\mathcal{M}$, then an i-arc is also created between the nodes n_{ι_1} and n_{ι_2} in $\mathcal{V}_Q$.
    - Instead, if a c-arc exists between the nodes corresponding to ι_1 and ι_2 in $\mathcal{M}$, then n_{ι_1} and n_{ι_2} are merged in a unique node $n_{\iota_{12}}$ in $\mathcal{V}_Q$. This task is motivated by the fact that n_{ι_1} and n_{ι_2} represent different instances of the same object in different real IoTs, but they represent the same instance in the same virtual IoT; as a consequence, they must be merged and no cross arc can exist between them. The profile $\overline{\mathcal{P}_{12}}$ of $n_{\iota_{12}}$ is obtained by applying the operator $\bigsqcup$ on the profiles $\overline{\mathcal{P}_1}$ of ι_1 and $\overline{\mathcal{P}_2}$ of ι_2.
- Finally, our approach adds a disconnected node in $\mathcal{V}_Q$ for each keyword in Q such that there is no MIoT instance having at least one topic whose keyword is identical to, or synonymous of, it[2].
- At this point, two cases may occur. In particular:
  - It could happen that $\mathcal{V}_Q$ is connected. In this case, it is returned as the answer to the query Q submitted by the user.
  - If $\mathcal{V}_Q$ is not connected and if the number of its connected components is less than a certain threshold, our approach adds the minimum number of "fictitious" i-arcs necessary to make $\mathcal{V}_Q$ connected.
  - Otherwise, if the number of connected components of $\mathcal{V}_Q$ is higher than a certain threshold, our approach concludes that a unique thematic virtual IoT corresponding to the keywords specified by the user does not exist and returns the thematic views related to the connected components of $\mathcal{V}_Q$. At this point, the user can decide whether to accept these thematic views or to modify the query in such a way as to construct a unique thematic view by re-applying all the above mentioned steps starting from the new query.

[2] The rationale underlying this step will be clearer in the following.

Unsupervised approach

The unsupervised approach begins with the construction of a support network $\mathcal{N}$ starting from the MIoT $\mathcal{M}$. In particular:

- For each node n_{ι_k} of $\mathcal{M}$, a node $\overline{n_{\iota_k}}$ is added in $\mathcal{N}$.
- For each i-arc $(n_{\iota_{j_k}}, n_{\iota_{q_k}})$ in $\mathcal{M}$, an (unoriented) arc $(\overline{n_{\iota_{j_k}}}, \overline{n_{\iota_{q_k}}})$ is added in $\mathcal{N}$. The arcs of $\mathcal{N}$ are weighted. The weight of the arc $(\overline{n_{\iota_{j_k}}}, \overline{n_{\iota_{q_k}}})$ is obtained by applying the operator J^* on the topic sets $topicSet_{j_k}$ and $topicSet_{q_k}$ of ι_{j_k} and ι_{q_k}, respectively. Therefore, the weight of an arc in $\mathcal{N}$ belongs to the real interval $[0,1]$; the higher this weight the higher the semantic similarity between the topics of the profiles $\overline{\mathcal{P}_{j_k}}$ and $\overline{\mathcal{P}_{q_k}}$ of ι_{j_k} and ι_{q_k}, respectively.
- For each c-arc in $\mathcal{M}$, which relates two instances $n_{\iota_{j_k}}$ and $n_{\iota_{j_q}}$ of the same object o_j in two different IoTs $\mathcal{I}_k$ and $\mathcal{I}_q$, the two nodes $\overline{n_{\iota_{j_k}}}$ and $\overline{n_{\iota_{j_q}}}$ in $\mathcal{N}$, corresponding to the nodes $n_{\iota_{j_k}}$ and $n_{\iota_{j_q}}$ in $\mathcal{M}$, are merged into a unique node $\overline{n_{\iota_j}}$. This node inherits all the arcs of $\overline{n_{\iota_{j_k}}}$ and $\overline{n_{\iota_{j_q}}}$.

At the end of these steps, it could happen that two or more arcs relate the same nodes $\overline{n}$ and $\overline{n'}$ in $\mathcal{N}$. In this case, all these arcs must be merged into a single arc. Clearly, it is necessary to determine the weight of this arc. Here, it appears reasonable that it must be higher than or equal to the maximum weight of the merged arcs. To reach this objective, our approach operates as follows. Let $\{(\overline{n}, \overline{n'}, \overline{w^1}), (\overline{n}, \overline{n'}, \overline{w^2}), \cdots, (\overline{n}, \overline{n'}, \overline{w^s})\}$ be the arcs to merge, ordered by decreasing weight. The new arc $(\overline{n}, \overline{n'}, \overline{w})$ will have a weight equal to:

$$\overline{w} = min\left(1, \overline{w_1} + \alpha \sum_{k=2..s} \overline{w_k}\right) \tag{5.23}$$

In other words, in the computation of $\overline{w}$, the arcs with the maximum weight will contribute with all their weight. All the other arcs will contribute to a lesser extent, with a fraction of their weight. This last is determined by means of the coefficient α.

Once the construction of $\mathcal{N}$ has been completed, the thematic views are derived by applying on $\mathcal{N}$ a graph clustering algorithm among the ones already existing in the literature (see [590] for a survey on them).

Comparison between supervised and unsupervised approach

An important issue about the supervised and the unsupervised approaches to address regards their scalability or, better, the possibility to use them in MIoTs comprising thousands or even millions of nodes.

With regard to this issue, first of all we observe that both approaches aim at deriving virtual IoTs which are, then, exploited by users to perform their desired

tasks (such as querying). As a consequence, we can distinguish two moments in the life of a MIoT, namely: *(i)* the construction of virtual IoTs, which can be performed *offline*, and *(ii)* their usage, which is generally carried out *online*.

The first moment is computationally expensive because it involves several network operations in the supervised approach and a clustering activity in the unsupervised one. Clustering's computational cost is intrinsically exponential even if all the corresponding methods adopted in the reality are heuristic and most of them have a linear or a quadratic computational complexity. In any case, as pointed above, this task is performed offline and rarely because it is necessary only when many changes have been made in the MIoT.

The second moment is certainly less expensive; its cost depends on the size of the involved clusters; in fact, each user activity generally involves one or a few clusters. Concerning this aspect, it is important to verify: *(i)* if clustering is possible in presence of huge MIoTs, and *(ii)* how the size of clusters increases against the growth of the MIoT. As for the first point, we observe that, in the past, several algorithms have been specifically conceived to cluster a huge amount of elements [256]. Concerning the second point, instead, first we observe that the size of clusters can be determined by suitably tuning the parameters of the selected clustering algorithm. However, it could be interesting to verify how much the size of clusters increases if we maintain constant all the clustering algorithm parameters and the MIoT size increases. We decided to perform this experiment. It is described in detail in Section 5.1.4.6. Here, we evidence the obtained results, i.e., that when the MIoT size highly increases, the cluster size slightly grows, whereas the number of clusters increases very much. This is a positive result for our purposes because the parameter to monitor for investigating the performance obtained during the second moment is just cluster size.

Another important issue to investigate regards the possible existence of a unique framework handling all the objects of the MIoT and, therefore, in principle, thousands or millions of objects. With regard to this aspect, we evidence that, in the past, several attempts have been successfully performed in this direction (think, for instance, of the SIoT framework proposed in [70, 74]). Clearly, we understand that, in the future, the number of objects possibly belonging to a MIoT is enormously higher than the number of objects available in the past IoT frameworks. However, we point out that: *(i)* our approach needs to store only the metadata of the involved objects, and these are small; *(ii)* the real objects can operate in a distributed environment thanks to the new available technologies, such as cloud, edge and fog computing, which can ease the organization and the management of distributed contexts.

5.1.4 Results

In this section, we present the experimental campaign that we carried out to evaluate the performance of our approach from several viewpoints. Specifically, we describe our dataset in a subsection, whereas, in the next ones, we illustrate our tests, along with the underlying motivations and the obtained results.

5.1.4.1 Testbed

To perform our experiments, we had the necessity to create several MIoTs with different sizes, ranging from hundreds to thousands of nodes. Since, currently, real MIoTs with the size and the variety handled by our model do not exist yet, we had to realize a MIoT simulator, i.e., a tool that, starting from real data, is capable of simulating MIoTs with certain characteristics specified by the user.

The MIoTs created by our simulator follow the model described in Section 4. In order to perform its task, our simulator carries out the following steps: *(i)* creation of objects; *(ii)* creation of object instances; *(iii)* creation of instance connections; *(iv)* creation of instance profiles.

Our MIoT simulator is also provided with a suitable interface allowing a user to "personalize" the MIoT to construct by specifying the desired values for several parameters, such as the number of nodes, the maximum number of instances of an object, and so forth.

To make "concrete" and "plausible" the created MIoT, our simulator leverages a real dataset. It regards the taxi routes in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. It can be found at the address `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. Each route contains several Points of Interests corresponding to the GPS coordinates of the vehicle.

We partitioned the city of Porto in six areas and associated a real IoT with each of them. Our simulator associates an object with a given route recorded in the dataset and an object instance for each partition of a route belonging to an area. It creates a MIoT node for each instance and a c-arc for each pair of instances belonging to the same route. Furthermore, it creates an i-arc between two nodes of the same IoT if the length of the time interval between the corresponding routes is less than a certain threshold th_t. The weight of the i-arc indicates the length of this time interval. The value of th_t can be specified through the constructor interface. Clearly, the higher th_t the more connected the constructed MIoT.

As far as instance profiles are concerned, since there are no thing profiles available, we had to simulate them. However, we aimed to make them as real as possible. In order to increase the likelihood of constructed MIoTs, we performed a sentiment

analysis task for each of the six areas in which we partitioned the city of Porto and for each day which the dataset refers to. For this purpose, we leveraged IBM Watson on the social media and blogs it uses as default. Having this data at disposal, our simulator assigns to each instance the most common topics (along with the corresponding occurrences) discussed in that area in the day on which the corresponding route took place. The constructed MIoTs are returned in a format that can be directly processed by the cypher-shell of Neo4J (see below).

Some features of the constructed MIoTs are reported in Table 5.1. The interested reader can find the MIoTs adopted in the experiments described in this section at the address `http://daisy.dii.univpm.it/miot/datasets/virtualIoTs`.

MIoT (size)	Number of arcs	Mean in-degree	Mean out-degree	Number of i-arcs	Number of c-arcs
$\mathcal{M}_1$ (176)	1176	6.29	6.61	980	126
$\mathcal{M}_2$ (301)	2050	7.76	7.74	1709	341
$\mathcal{M}_3$ (485)	3756	8.80	8.54	3130	626
$\mathcal{M}_4$ (778)	5866	8.89	9.11	4895	971
$\mathcal{M}_5$ (946)	7624	8.64	8.84	6422	1202
$\mathcal{M}_6$ (1256)	9860	7.87	7.98	7917	1943
$\mathcal{M}_7$ (1725)	12263	7.94	8.18	9964	2299
$\mathcal{M}_8$ (2028)	15568	8.22	8.38	12857	2711
$\mathcal{M}_9$ (3544)	26428	8.36	8.42	22718	3710
$\mathcal{M}_{10}$ (5024)	38642	8.44	8.54	33724	4918

Table 5.1: Main features of the constructed MIoTs

We carried out all the tests presented in this section on a server equipped with an Intel I7 Quad Core 7700 HQ processor and 16 GB of RAM with Ubuntu 16.04 operating system.

To implement our approaches we adopted:

- Python, powered with the NetworkX library, as programming language;
- Neo4J (Version 3.4.5) as underlying DBMS; we also exploited some plugins of Neo4J to perform community detection and to compute clustering coefficients.

5.1.4.2 Cohesion of the obtained topic-guided virtual IoTs

Our first test started from the idea that if our approach aims at extracting virtual thematic IoTs, they should present both a structural and a semantic cohesion higher than the corresponding ones characterizing the original IoTs of the MIoT. This experiment was devoted to evaluating if this assumption is verified. We considered two well known structural cohesion parameters used in network analysis literature, namely *clustering coefficient* and *density* [647]. Both of them range in the real interval $[0,1]$; the higher their value the higher the corresponding network cohesion. In the

following, first we test the supervised approach and, then, we consider the unsupervised one.

Supervised approach

In this test, we run our supervised approach on ten MIoTs, $\mathcal{M}_1, \ldots, \mathcal{M}_{10}$, consisting of 176, 301, 485, 778, 946, 1256, 1725, 2028, 3544 and 5024 nodes. Clearly, the number of IoTs for each MIoT was equal to six, one for each area of the city of Porto that we have defined. For each MIoT, we submitted a set of 10 queries consisting of 1 (resp., 2, 4, 6, 8 and 10) word(s).

Each query returned a virtual thematic IoT for which we computed the corresponding clustering coefficient and density. Finally, we averaged the obtained results for each MIoT and for each set of queries, and we compared them with the average clustering coefficient and the average density of the corresponding real IoTs. The obtained results are reported in Tables 5.2 and 5.3.

MIoT (size)	Avg. clustering coeff. (real IoTs)	Avg. clustering coeff. (virtual IoTs)					
		$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.230	0.318	0.368	0.389	0.394	0.401	0.408
$\mathcal{M}_2$ (301)	0.272	0.343	0.388	0.419	0.424	0.434	0.446
$\mathcal{M}_3$ (485)	0.293	0.396	0.437	0.477	0.482	0.488	0.497
$\mathcal{M}_4$ (778)	0.353	0.447	0.478	0.503	0.508	0.511	0.517
$\mathcal{M}_5$ (946)	0.371	0.452	0.492	0.512	0.522	0.524	0.526
$\mathcal{M}_6$ (1256)	0.385	0.486	0.511	0.529	0.530	0.532	0.535
$\mathcal{M}_7$ (1725)	0.386	0.501	0.524	0.536	0.537	0.538	0.539
$\mathcal{M}_8$ (2028)	0.388	0.519	0.536	0.541	0.541	0.542	0.543
$\mathcal{M}_9$ (3544)	0.392	0.522	0.540	0.544	0.544	0.545	0.546
$\mathcal{M}_{10}$ (5024)	0.395	0.534	0.546	0.546	0.546	0.547	0.548

Table 5.2: Values of the clustering coefficient for real and virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

From the analysis of these tables, we can observe that, in almost all circumstances, the values of both clustering coefficient and density are higher or much higher for the virtual thematic IoTs than for the real ones. This is clearly a confirmation of the goodness of our supervised approach, which returns topic-guided IoTs more cohesive than the original ones. We also observe that when $|Q|$ increases, the values of both clustering coefficient and density increases. This can be explained by observing that, in processing Q, our approach takes the portions of networks containing at least one keyword of Q. When $|Q|$ increases, the portion of networks selected by our approach increases too, and the probability of selecting a very high number of edges (i.e., a number so high to lead to an increase of clustering coefficient and density) increases as well.

MIoT (size)	Average density (real IoTs)	Average density (virtual IoTs)					
		$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.348	0.260	0.264	0.280	0.289	0.296	0.301
$\mathcal{M}_2$ (301)	0.262	0.292	0.303	0.309	0.315	0.320	0.324
$\mathcal{M}_3$ (485)	0.274	0.390	0.395	0.400	0.402	0.405	0.408
$\mathcal{M}_4$ (778)	0.269	0.476	0.483	0.490	0.501	0.509	0.514
$\mathcal{M}_5$ (946)	0.276	0.492	0.509	0.521	0.536	0.534	0.556
$\mathcal{M}_6$ (1256)	0.284	0.547	0.556	0.567	0.572	0.576	0.581
$\mathcal{M}_7$ (1725)	0.278	0.582	0.582	0.594	0.598	0.598	0.601
$\mathcal{M}_8$ (2028)	0.273	0.609	0.610	0.620	0.626	0.630	0.639
$\mathcal{M}_9$ (3544)	0.269	0.626	0.628	0.630	0.634	0.636	0.637
$\mathcal{M}_{10}$ (5024)	0.262	0.636	0.636	0.638	0.638	0.640	0.642

Table 5.3: Values of the density for real and virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

Unsupervised approach

In this test, we run our unsupervised approach, powered with the Louvain graph clustering algorithm [114] as underlying engine, on the same MIoTs described in Section 5.1.4.2. For each MIoT, we computed the average clustering coefficient and the average density of real and virtual IoTs. The obtained results are reported in Table 5.4.

MIoT (size)	Average clustering coefficient		Average density	
	Real IoTs	Virtual IoTs	Real IoTs	Virtual IoTs
$\mathcal{M}_1$ (176)	0.230	0.473	0.348	0.315
$\mathcal{M}_2$ (301)	0.272	0.499	0.262	0.350
$\mathcal{M}_3$ (485)	0.293	0.500	0.274	0.375
$\mathcal{M}_4$ (778)	0.353	0.511	0.269	0.318
$\mathcal{M}_5$ (946)	0.372	0.509	0.276	0.316
$\mathcal{M}_6$ (1256)	0.385	0.506	0.284	0.314
$\mathcal{M}_7$ (1725)	0.386	0.522	0.280	0.328
$\mathcal{M}_8$ (2028)	0.388	0.535	0.273	0.360
$\mathcal{M}_9$ (3544)	0.394	0.547	0.271	0.364
$\mathcal{M}_{10}$ (5024)	0.398	0.562	0.269	0.368

Table 5.4: Values of both clustering coefficient and density of real and virtual IoTs against the size of MIoTs (unsupervised approach)

From the analysis of this table we can observe that, in this case, analogously to what happened for the supervised approach, the cohesion level of the virtual IoTs is higher or much higher than the corresponding ones of the real original IoTs. Interestingly, both clustering coefficient and density values obtained by the unsupervised approach are generally higher than those returned by the supervised one, at least when the MIoT size is small. Instead, when the MIoT size is large, they become lower

than the ones of the supervised approach. Actually, the increase of both clustering coefficient and density when the MIoT size increases is significant for the supervised approach, whereas it is more limited for the unsupervised one.

5.1.4.3 Analysis of merged c-nodes and node distribution in virtual IoTs

Another quality parameter for virtual IoTs returned by our approach regards the average number of merged c-nodes present in each of them. Indeed, the presence of merged c-nodes in an IoT is an indicator of the fact that this IoT is capable of connecting concepts coming from different real IoTs, and, therefore, from concepts whose relationships would have been uncaptured otherwise, or, in other words, that the knowledge it is presenting is new and did not exist previously. Clearly, the higher the fraction of merged c-nodes and the higher the fraction of different original IoTs they belong to, the higher the connecting capability of virtual IoTs.

Also for this experiment, we considered the ten MIoTs described in Section 5.1.4.2 and performed the same tasks illustrated therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 5.5, 5.6 and 5.7.

MIoT (size)	Average fraction of merged c-nodes					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.304	0.455	0.517	0.532	0.554	0.572
$\mathcal{M}_2$ (301)	0.380	0.515	0.608	0.627	0.652	0.679
$\mathcal{M}_3$ (485)	0.539	0.661	0.782	0.798	0.813	0.823
$\mathcal{M}_4$ (778)	0.690	0.786	0.860	0.874	0.883	0.892
$\mathcal{M}_5$ (946)	0.724	0.812	0.884	0.898	0.916	0.924
$\mathcal{M}_6$ (1256)	0.808	0.883	0.939	0.943	0.946	0.948
$\mathcal{M}_7$ (1725)	0.862	0.908	0.952	0.961	0.961	0.963
$\mathcal{M}_8$ (2028)	0.908	0.959	0.974	0.975	0.976	0.977
$\mathcal{M}_9$ (3544)	0.928	0.963	0.976	0.977	0.977	0.978
$\mathcal{M}_{10}$ (5024)	0.936	0.968	0.978	0.979	0.980	0.981

Table 5.5: Average fraction of merged c-nodes against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

From the analysis of these tables, we observe that both the supervised and the unsupervised approaches return satisfying results. As for the supervised approach, we can observe that the fraction of merged c-nodes increases when the size of MIoT increases. Furthermore, we can also observe a slight increase of this fraction when $|Q|$ increases. The same trends can be observed for the average fraction of involved real IoTs, even if, for this parameter, its increase against the increase of $|Q|$ is more pronounced. As for the unsupervised approach, we can observe that the average

MIoT (size)	Average fraction of involved real IoTs					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.373	0.467	0.488	0.476	0.452	0.448
$\mathcal{M}_2$ (301)	0.365	0.469	0.525	0.501	0.488	0.480
$\mathcal{M}_3$ (485)	0.482	0.477	0.448	0.442	0.435	0.432
$\mathcal{M}_4$ (778)	0.457	0.432	0.418	0.415	0.413	0.411
$\mathcal{M}_5$ (946)	0.455	0.482	0.624	0.628	0.647	0.644
$\mathcal{M}_6$ (1256)	0.453	0.514	0.805	0.864	0.917	0.924
$\mathcal{M}_7$ (1725)	0.482	0.577	0.815	0.872	0.917	0.924
$\mathcal{M}_8$ (2028)	0.514	0.672	0.833	0.898	0.917	0.924
$\mathcal{M}_9$ (3544)	0.584	0.704	0.844	0.905	0.924	0.926
$\mathcal{M}_{10}$ (5024)	0.624	0.727	0.888	0.911	0.928	0.934

Table 5.6: Average fraction of real IoTs involved in a virtual IoT against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average fraction of merged c-nodes	Average fraction of involved real IoTs
$\mathcal{M}_1$ (176)	0.227	0.361
$\mathcal{M}_2$ (301)	0.306	0.353
$\mathcal{M}_3$ (485)	0.309	0.357
$\mathcal{M}_4$ (778)	0.342	0.356
$\mathcal{M}_5$ (946)	0.334	0.359
$\mathcal{M}_6$ (1256)	0.326	0.361
$\mathcal{M}_7$ (778)	0.332	0.360
$\mathcal{M}_8$ (2028)	0.335	0.358
$\mathcal{M}_9$ (3544)	0.341	0.371
$\mathcal{M}_{10}$ (5024)	0.344	0.378

Table 5.7: Average fraction of merged c-nodes and average fraction of real IoTs involved in a virtual IoT against the size of MIoTs (unsupervised approach)

fraction of merged nodes is always very high, independently of the MIoT size. By contrast, in this case, the fraction of involved real IoTs is quite high even if lower than the ones generally observed for the supervised approach. Furthermore, its value does not significantly change when the MIoT size increases.

In order to deepen this investigation, for each virtual IoT, we compared the distribution of its nodes against the real IoTs they belong to. Indeed, if almost all the nodes of a virtual IoT derive from only one real IoT, the information contribution provided by the virtual IoT would be very small because it would be analogous to the one provided by the corresponding real IoT. By contrast, if the nodes of a virtual IoT homogeneously derive from several real IoTs, then the knowledge it provides is really new, and this knowledge would be uncaptured and lost if the new IoT had not been extracted. On the basis of this reasoning, we evaluated the heterogeneity of the provenance of the various nodes of each virtual IoT (see below). For this purpose, we adapted the Herfindahl Index [332] to our context. This index is very used in sev-

eral research fields of Economics from several decades; for instance, it is exploited to evaluate the concentration degree in an industry.

In order to adapt the Herfindahl Index to our scenario, consider a MIoT $\mathcal{M}$ consisting of s real IoTs $(\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_s)$. Consider, also, a virtual IoT $\mathcal{V}_j$ derived by either the supervised or the unsupervised approach. Let n_j be the number of nodes of $\mathcal{V}_j$ and let $\frac{n_{j_k}}{n_j}$, $1 \leq k \leq s$, be the fraction of the nodes of $\mathcal{V}_j$ belonging to $\mathcal{R}_k$ (i.e., the k^{th} real IoT of the MIoT). The Herfindahl Index H_j of $\mathcal{V}_j$ is defined as $\sum_{k=1}^{s} \left(\frac{n_{j_k}}{n_j}\right)^2$. H_j ranges in the real interval $\left[\frac{1}{s}, 1\right]$; the higher its value, the higher the concentration degree of the nodes of $\mathcal{R}_k$ in $\mathcal{V}_j$. Clearly, as previously pointed out, one property desired for our approach is the ability to construct virtual IoTs connecting nodes that belong to different real IoTs in such a way as to extract knowledge that would be lost otherwise. If we report this property to the Herfindahl Index, this implies to obtain a value of this index as lower as possible[3].

We computed the average Herfindahl Index of the thematic IoTs returned by both the supervised and the unsupervised approaches by considering the ten MIoTs described in Section 5.1.4.2 and performing the same tasks illustrated therein. The obtained results are reported in Tables 5.8 and 5.9.

MIoT (size)	Average Herfindhal Index					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.207	0.186	0.177	0.175	0.173	0.172
$\mathcal{M}_2$ (301)	0.204	0.183	0.174	0.173	0.172	0.171
$\mathcal{M}_3$ (485)	0.178	0.173	0.170	0.170	0.169	0.168
$\mathcal{M}_4$ (778)	0.172	0.172	0.170	0.170	0.169	0.168
$\mathcal{M}_5$ (946)	0.172	0.170	0.169	0.169	0.169	0.168
$\mathcal{M}_6$ (1256)	0.173	0.168	0.167	0.169	0.168	0.167
$\mathcal{M}_7$ (1725)	0.170	0.168	0.167	0.169	0.168	0.167
$\mathcal{M}_8$ (2028)	0.168	0.167	0.167	0.167	0.167	0.167
$\mathcal{M}_9$ (3544)	0.168	0.167	0.167	0.167	0.167	0.167
$\mathcal{M}_{10}$ (5024)	0.167	0.167	0.167	0.167	0.167	0.167

Table 5.8: Average Herfindahl Index of virtual IoTs against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

These tables evidence that also the analysis based on object distribution and Herfindahl Index returns very satisfying results that confirm and strengthen those obtained by examining the average fraction of merged nodes involved in a virtual IoT. Interestingly, as for this parameter, we observe that the supervised approach returns excellent results, very close to the best ones. By contrast, the unsupervised

[3] Consider that, since we have six real IoTs in our MIoTs, the minimum value of the Herfindahl Index is $\frac{1}{6} = 0.167$.

MIoT (size)	Average Herfindahl Index
$\mathcal{M}_1$ (176)	0.658
$\mathcal{M}_2$ (301)	0.543
$\mathcal{M}_3$ (485)	0.658
$\mathcal{M}_4$ (778)	0.636
$\mathcal{M}_5$ (946)	0.654
$\mathcal{M}_6$ (1256)	0.694
$\mathcal{M}_7$ (1725)	0.656
$\mathcal{M}_8$ (2028)	0.635
$\mathcal{M}_9$ (3544)	0.664
$\mathcal{M}_{10}$ (5024)	0.686

Table 5.9: Average Herfindahl Index of virtual IoTs against the size of MIoTs (unsupervised approach)

approach returns good results, even if those returned by the supervised approach are better.

5.1.4.4 Computation time

In this experiment, we aimed at evaluating the variation of the computation time of both the supervised and the unsupervised approaches against the variation of the size of the involved MIoT. Furthermore, as for the supervised approach, we also evaluated the variation of the computation time against the variation of the size of queries.

To perform this task, we considered the ten MIoTs described in Section 5.1.4.2 and carried out the same tasks illustrated therein. Finally, we measured the corresponding average computation times. The obtained results are reported in Figures 5.1, 5.2 and 5.3.

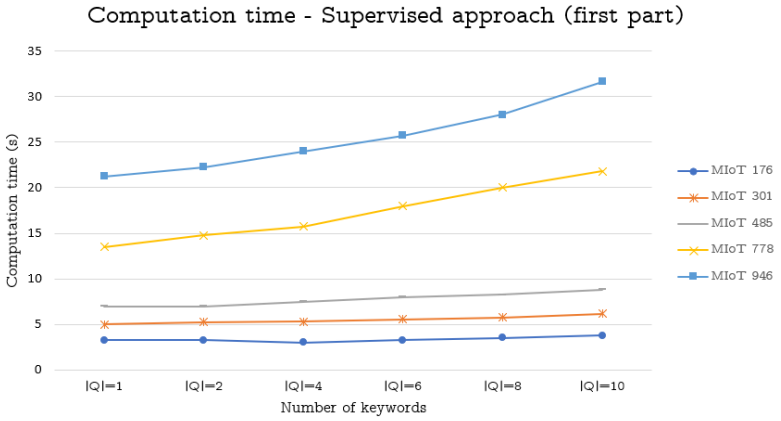


Fig. 5.1: Computation time (in seconds) against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach) - first part

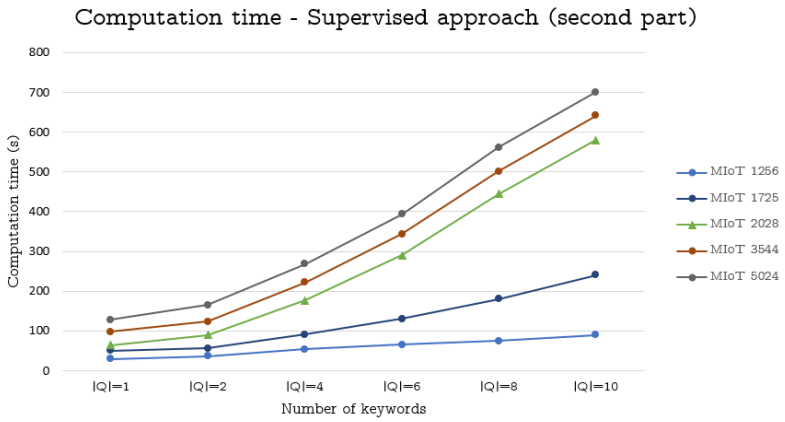


Fig. 5.2: Computation time (in seconds) against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach) - second part

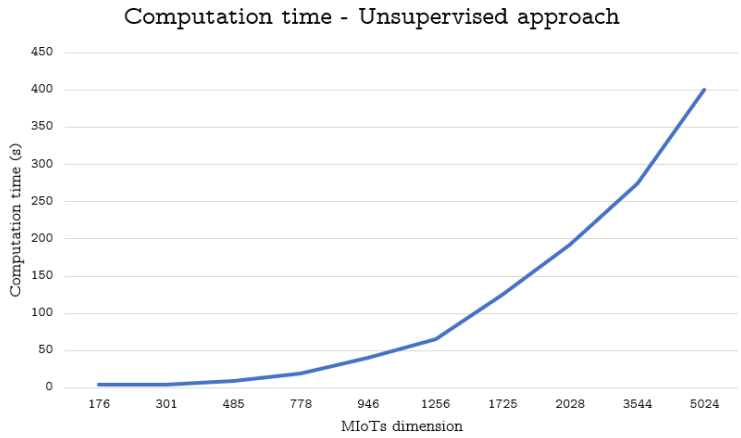


Fig. 5.3: Computation time (in seconds) against the size of MIoTs (unsupervised approach)

From the analysis of these figures, we can observe that our approaches obtain satisfying results. Specifically, as for the supervised approach, the computation time is always very low for MIoTs having at most 1256 nodes. Instead, for MIoTs with more than 2028 nodes, the computation time is low for $|Q| = 1$ or $|Q| = 2$. Then, it increases, even if it remains acceptable for $|Q| = 4$ and $|Q| = 6$, whereas it becomes excessive for $|Q| = 8$ and $|Q| = 10$. However, with regard to this fact, we must point out that queries consisting of 8 or 10 keywords are very uncommon[4].

[4] It is worth pointing out that the topics considered by our approach for constructing a thing's profile are extremely generic and heterogeneous. As a consequence, in our scenario, a query with 8 or 10 keywords would encompass a great number of different topics and, as such, it would not be generally able to capture a clear and specific desire of a user.

As for the unsupervised approach, its computation time is still acceptable also for 2028 nodes. It starts to become excessive with MIoTs consisting of at least 10000 nodes.

5.1.4.5 Analysis of the efficiency of information dissemination

This experiment was devoted to measuring the efficiency of both supervised and unsupervised approaches. The rationale underlying this experiment is that if some information must be transferred from a source object o_s to a target one o_t, the number of objects to be contacted for this task should be minimized. At the same time, if an object is involved in an information dissemination task, it would be desirable that the information it is transmitting is also useful for it (which, in our case, means that it is in line with the interests of its profile).

In order to perform this experiment, we randomly selected some pairs of (source, target) nodes from our MIoT. Let (n_s, n_t) be one of these pairs. We verified if there existed at least one virtual IoT comprising both n_s and n_t[5]. In the negative case, we discarded that pair. Let $\mathcal{V}$ be a virtual IoT comprising both n_s and n_t.

After this, we computed the number $num_{st}^{\mathcal{V}}$ (resp., $\widehat{num_{st}^{\mathcal{V}}}$) of MIoT nodes involved in the dissemination of information in presence (resp., absence) of the virtual IoT $\mathcal{V}$. Specifically, we computed $num_{st}^{\mathcal{V}}$ by performing the information dissemination task only through its nodes; instead, we obtained $\widehat{num_{st}^{\mathcal{V}}}$ by performing the same task on the whole MIoT. Finally, we computed: $f_{st} = \frac{num_{st}^{\mathcal{V}}}{\widehat{num_{st}^{\mathcal{V}}}}$. Clearly, the lower f_{st}, the higher the contribution of the virtual IoTs in reducing the number of nodes necessary for the information dissemination task and, consequently, the higher the contribution that our virtual IoT detection approach can provide to information dissemination.

We computed the average values of f_{st} by operating on the ten MIoTs introduced in Section 5.1.4.2 and by performing the same tasks described therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 5.10 and 5.11.

From the analysis of these tables we can observe that both the supervised and the unsupervised approaches really contribute to decrease the number of the nodes of a MIoT involved in the information dissemination, and, therefore, to increase the efficiency of this task. As for the supervised approach, we observe that the decrease of the number of involved nodes is always high. It becomes very high as the MIoT size and the number of keywords composing the query increase. As for the unsupervised approach, we observe that it leads to a decrease of the number of the MIoT nodes involved in the dissemination task. However, this decrease is minimum for small

[5] This is always true for the unsupervised approach, whereas it could not happen for the supervised one.

MIoT (size)	Average f_{st}					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	0.144	0.220	0.290	0.304	0.336	0.347
$\mathcal{M}_2$ (301)	0.126	0.170	0.177	0.175	0.178	0.179
$\mathcal{M}_3$ (485)	0.104	0.112	0.074	0.052	0.041	0.037
$\mathcal{M}_4$ (778)	0.057	0.051	0.028	0.038	0.047	0.049
$\mathcal{M}_5$ (946)	0.048	0.034	0.022	0.028	0.032	0.024
$\mathcal{M}_6$ (1256)	0.031	0.015	0.017	0.011	0.007	0.007
$\mathcal{M}_7$ (1725)	0.026	0.014	0.011	0.010	0.008	0.008
$\mathcal{M}_8$ (2028)	0.016	0.010	0.009	0.009	0.009	0.009
$\mathcal{M}_9$ (3544)	0.012	0.009	0.009	0.009	0.009	0.009
$\mathcal{M}_{10}$ (5024)	0.011	0.008	0.007	0.007	0.007	0.007

Table 5.10: Average values of f_{st} against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average f_{st}
$\mathcal{M}_1$(176)	0.904
$\mathcal{M}_2$(301)	0.722
$\mathcal{M}_3$(485)	0.635
$\mathcal{M}_4$(778)	0.584
$\mathcal{M}_5$(946)	0.580
$\mathcal{M}_6$(1256)	0.576
$\mathcal{M}_7$(1725)	0.516
$\mathcal{M}_8$(2028)	0.477
$\mathcal{M}_9$(3544)	0.452
$\mathcal{M}_{10}$(5024)	0.426

Table 5.11: Average values of f_{st} against the size of MIoTs (unsupervised approach)

MIoTs, whereas it becomes significant for large ones (i.e., for MIoTs with a number of nodes higher than 1256).

We performed a second experiment in this direction. Specifically, given a pair (n_s, n_t) of a MIoT such that information must be disseminated from n_s to n_t and there exists at least one virtual IoT $\mathcal{V}$ comprising both n_s and n_t, we computed the fraction $g_{st}^{\mathcal{V}}$ (resp., $\widehat{g_{st}^{\mathcal{V}}}$) of the nodes of the MIoT involved in the diffusion of information from n_s to n_t and having at least one content of the disseminated information registered in their profile (which implies that, in principle, they could benefit from the information they are required to disseminate). As in the previous experiment, we computed $g_{st}^{\mathcal{V}}$ by assuming the existence of $\mathcal{V}$ and, hence, by performing the information dissemination task through it; by contrast, we computed $\widehat{g_{st}^{\mathcal{V}}}$ by carrying out the information dissemination task through the whole MIoT. Finally, we computed $g_{st} = \frac{g_{st}^{\mathcal{V}}}{\widehat{g_{st}^{\mathcal{V}}}}$. Roughly speaking, it denotes how much the presence of the virtual IoT $\mathcal{V}$ can contribute to require information dissemination tasks only to nodes possibly benefiting of it. A value of this coefficient higher than 1 denotes a positive contri-

bution of $\mathcal{V}$; the higher this value the higher the contribution. As in the previous experiment, we computed the average values of g_{st} by operating on the ten MIoTs introduced in Section 5.1.4.2 and by performing the same tasks described therein for both the supervised and the unsupervised approaches. The obtained results are reported in Tables 5.12 and 5.13.

MIoT (size)	Average g_{st}					
	$\|Q\| = 1$	$\|Q\| = 2$	$\|Q\| = 4$	$\|Q\| = 6$	$\|Q\| = 8$	$\|Q\| = 10$
$\mathcal{M}_1$ (176)	4.018	2.792	2.223	1.918	1.331	1.321
$\mathcal{M}_2$ (301)	3.563	2.619	2.445	2.009	1.683	1.664
$\mathcal{M}_3$ (485)	3.269	2.370	1.426	1.528	1.626	1.674
$\mathcal{M}_4$ (778)	3.130	2.168	2.367	1.916	1.494	1.325
$\mathcal{M}_5$ (946)	3.232	2.102	1.864	1.712	1.461	1.391
$\mathcal{M}_6$ (1256)	3.467	1.979	1.378	1.412	1.438	1.452
$\mathcal{M}_7$ (1725)	3.476	2.224	1.414	1.444	1.494	1.492
$\mathcal{M}_8$ (2028)	3.496	2.669	1.489	1.491	1.521	1.545
$\mathcal{M}_9$ (3544)	3.507	2.712	1.612	1.624	1.631	1.632
$\mathcal{M}_{10}$ (5024)	3.517	2.926	1.783	1.841	1.864	1.874

Table 5.12: Average values of g_{st} against the size of MIoTs and queries used to generate the virtual IoTs (supervised approach)

MIoT (size)	Average g_{st}
$\mathcal{M}_1$ (176)	1.341
$\mathcal{M}_2$ (301)	1.269
$\mathcal{M}_3$ (485)	1.211
$\mathcal{M}_4$ (778)	1.177
$\mathcal{M}_5$ (946)	1.173
$\mathcal{M}_6$ (1256)	1.171
$\mathcal{M}_7$ (1725)	1.194
$\mathcal{M}_8$ (2028)	1.273
$\mathcal{M}_9$ (3544)	1.281
$\mathcal{M}_{10}$ (5024)	1.301

Table 5.13: Average values of g_{st} against the size of MIoTs (unsupervised approach)

The analysis of these tables is a further confirmation of the efficiency of our approach. Indeed, thanks to the presence of virtual IoTs, the fraction of nodes participating to the spreading of information that can also benefit from this task increases remarkably.

The results of Tables 5.10 and 5.11, along with the ones of Tables 5.12 and 5.13, agree to evidence that the discovery of virtual IoTs is highly beneficial in terms of efficiency for the information dissemination task in a MIoT. In this case, the contribution of $\mathcal{V}$ in increasing the efficiency of the spreading task, by limiting it mainly

to nodes that could benefit from the information they are disseminating, is very high for the supervised approach when $|Q| = 1$ or $|Q| = 2$. When $|Q|$ increases, this contribution decreases, even if it remains still significant. As for the unsupervised approach, the contribution of $\mathcal{V}$ can be always observed even if it is less evident than the one characterizing the supervised approach.

5.1.4.6 Analysis of the virtual IoTs

This last experiment makes sense only for the unsupervised approach. Through it we aimed at investigating how the number and the size of returned virtual IoTs (and, therefore, the number and the size of returned clusters) vary when the MIoT size increases. To make this experiment significant, we maintained constant all the parameters of the adopted clustering algorithm. We considered the MIoTs $\mathcal{M}_1 \cdots \mathcal{M}_{10}$ used in the previous experiments because, in this way, we had the possibility to investigate MIoT sizes ranging from 176 to 5024 nodes. We report the obtained results in Table 5.14.

MIoT (size)	Average size of virtual IoTs	Number of virtual IoTs
$\mathcal{M}_1$ (176)	22.44	10
$\mathcal{M}_2$ (301)	28.21	13
$\mathcal{M}_3$ (485)	36.64	16
$\mathcal{M}_4$ (778)	40.82	22
$\mathcal{M}_5$ (946)	44.66	24
$\mathcal{M}_6$ (1256)	46.74	30
$\mathcal{M}_7$ (1725)	48.12	39
$\mathcal{M}_8$ (2028)	50.24	45
$\mathcal{M}_9$ (3544)	50.46	78
$\mathcal{M}_{10}$ (5024)	50.64	105

Table 5.14: Average size and number of virtual IoTs against the increase of the MIoT size (unsupervised approach)

From the analysis of this table we can observe that the average size of virtual IoTs:

- increases when the MIoT size ranges from 176 to 946;
- slightly increases when the MIoT size ranges from 946 to 2028;
- remains essentially constant when the MIoT size is higher than 2028.

In the meantime, the number of clusters:

- slightly increases when the MIoT size ranges from 176 to 946;
- increases when the MIoT size ranges from 946 to 2028;
- highly increases when the MIoT size is higher than 2028.

The obtained results are extremely interesting because they confirm the soundness of the reasoning made in Section 5.1.3.2. In particular, this experiment confirms the scalability of our approach. As a matter of fact, after the virtual IoTs have been constructed offline, their usage for querying and for the other tasks of interest for the user can be performed online. Now, we observed that the number of available virtual IoTs highly increases when the MIoT size increases. However, because the size of each virtual IoT is only slightly impacted by the growth of the corresponding MIoT, and because user tasks generally involve one or at most a few of available virtual IoTs, we can conclude that our approach is scalable with respect to the size variation of the MIoT.

5.2 Redefining Betweenness Centrality in a MIoT

5.2.1 Introduction

The betweenness centrality of a node in a network is defined as the fraction of the shortest paths between all the pairs of nodes that pass through it. Betweenness centrality is well suited for measuring the influence of a node over the information spread through the network [85, 505], to identify boundary spanners (i.e., nodes acting as bridges between two or more subnetworks), and to measure the "stress" (in the sense of a higher usage) that a node must undergo during network activities [120, 121, 182, 280]. Due to its relevance in network analysis, betweenness centrality has been largely investigated in the past, and several extensions, tailored to specific contexts, have been proposed (see, for instance, [680, 254, 255, 96]). Also in the context of the Internet of Things (IoT), several approaches for the computation of betweenness centrality have been presented [354, 552, 410].

However, the classical betweenness centrality is not able to correctly evaluate the centrality of nodes in a multiple IoT scenario, i.e., a scenario where several networks of smart objects (SO) cooperate with each other. In such a scenario (known as Multi-IoT or MIoT in the literature [82, 271, 434, 650]), IoT (i.e., networks of SO) are interconnected thanks to those nodes simultaneously belonging to two or more of them. We call *cross nodes* (*c-nodes*) these nodes and *inner nodes* (*i-nodes*) all the other ones. Then, a c-node connects at least two IoT of the MIoT and plays a key role in favoring the cooperation among i-nodes belonging to different IoT. As a consequence, the nodes of a MIoT are not all equal: c-nodes will presumably play a more important role than i-nodes for supporting the activities in a MIoT. Here, the classical betweenness centrality is not able to distinguish c-nodes from i-nodes and to evidence the key role played by c-nodes in favoring communication and cooperation between SO belonging to different IoT of the MIoT.

Here, we aim at providing a contribution to address this problem. Indeed, we propose three new measures of betweenness centrality, well suited for a MIoT and, more in general, for a scenario consisting of a set of related IoT. These measures are called *Inner Betweenness Centrality* (IBC), *Soft Cross Betweenness Centrality* (SCBC) and *Hard Cross Betweenness Centrality* (HCBC). They have been designed to clearly distinguish the contributions of c-nodes and i-nodes and we show that they are able to reach this objective. In particular, IBC has been conceived for measuring the betweenness centrality with a focus on a single IoT of the MIoT and it privileges i-nodes over c-nodes. As will be clarified in the following, it does not coincide with the classical betweenness centrality because, differently from this last one, it also considers paths which connect two nodes of the same IoT but, at the same time, in-

volve nodes belonging to other IoT of the MIoT. By contrast, SCBC and HCBC are specialized to measure the betweenness centrality of nodes by privileging paths involving more IoT of the MIoT and, therefore, c-nodes over i-nodes. As it is indicated by their names, this privilege is more marked in HCBC than in SCBC.

This chapter is organized as follows. In Section 5.2.2, we provide an overview of related literature. In Section 5.2.3, we introduce our new betweenness centrality measures. In Section 5.2.4, we describe our testbed and experimental analysis.

5.2.2 Related Literature

As one of the most important centrality measures, betweenness centrality [280] has been the subject of in-depth studies in the literature [182, 129]. Recognizing high spreading power nodes is fundamental in social networks but, based on its definition, the cost for computing the betweenness centrality of a node is high. For this reason, several heuristic approaches, aiming at providing the closest possible value of the betweenness centrality of a node in a reasonable time, have been proposed in the past (see [128, 76, 290, 565], to cite a few).

As for the IoT, which is an example of a very dynamic and constantly evolving network, the approaches for the incremental computation of betweenness centrality are extremely interesting. Among these, we mention the ones described in [354, 552, 410]. Specifically, in [354], the authors propose iCENTRAL, which is well suited for large and evolving biconnected graphs. In [552], the authors illustrate an approach for a quick incremental computation of betweenness centrality. After a pre-processing phase, the computational cost of this approach is independent of the network size. In [410], the authors describe an approach that reduces the search space by finding a set of candidate nodes that are the only ones to be updated during the incremental computation of the betweenness centrality.

Surprisingly, despite the strong tie existing among betweenness centrality and information diffusion, there are very few studies concerning the role of betweenness centrality in IoT. To the best of our knowledge, the only approaches dealing with centrality in IoT have been proposed as part of methods for determining trustworthiness [513] or network navigability [476, 511] in IoT.

5.2.3 Methods

5.2.3.1 MIoT-oriented Betweenness Centrality

Recall that, the MIoT paradigm introduced in Chapter 4 is the reference model in which we redefine the betweenness centrality.

Given a node n_j of a graph $\mathcal{G}$, the classic definition of betweenness centrality is the following:

$$BC(n_j) = \sum_{n_s \in N, n_t \in N, n_s \neq n_j, n_t \neq n_j} \frac{\sigma_{n_s n_t}(n_j)}{\sigma_{n_s n_t}}$$

where $\sigma_{n_s n_t}$ is the total number of the shortest paths from n_s to n_t, whereas $\sigma_{n_s n_t}(n_j)$ is the number of those shortest paths passing through n_j.

If we apply BC to the graph G_k associated with an IoT $\mathcal{I}_k$ and consider $\mathcal{I}_k$ isolated from the MIoT, this formula involves shortest paths which only pass from nodes of $\mathcal{I}_k$. In order to consider also the potential shortest paths that connect nodes of G_k but pass through nodes of the other IoT of the MIoT, it should be applied to the graph G corresponding to the whole MIoT. However, in this way, it does not capture that a MIoT consists of *different autonomous* IoT cooperating with each other thanks to c-nodes, which play a key role that should be evidenced by any measure of centrality conceived for a MIoT. We argue that, owing to these weaknesses, BC could present several problems in a MIoT context, especially when it is necessary to compute a centrality measure, which privileges those nodes that allow the crossing from an IoT to another.

To address the challenges mentioned above, we define three new centrality metrics. The first of them is called *Inner Betweenness Centrality* (IBC) and is defined as follows.

Let $n_{j_k} \in N_k$ be the node corresponding to the instance ι_{j_k} of the object o_j in the IoT $\mathcal{I}_k$ of the MIoT $\mathcal{M}$. The Inner Betweenness Centrality $IBC(n_{j_k})$ is defined as:

$$IBC(n_{j_k}) = \sum_{n_{s_k} \in N_k, n_{t_k} \in N_k, n_{s_k} \neq n_{j_k}, n_{t_k} \neq n_{j_k}} \frac{\overline{\sigma}_{n_{s_k} n_{t_k}}(n_{j_k})}{\overline{\sigma}_{n_{s_k} n_{t_k}}}$$

where $\overline{\sigma}_{n_{s_k} n_{t_k}}$ is the total number of the shortest paths from n_s to n_t that involve also nodes of the MIoT not belonging to N_k, and $\overline{\sigma}_{n_{s_k} n_{t_k}}(n_{j_k})$ is the total number of these shortest paths that pass through n_{j_k}.

IBC can be considered as an evolution of BC, capable of evaluating inner central nodes taking into account the fact that the network $\mathcal{I}_k$ is not alone but it is part of a MIoT. As a consequence, if all the paths connecting n_{s_k} to n_{t_k} include at least one node belonging to networks different from $\mathcal{I}_k$ but inside the MIoT, then BC does not capture them and considers n_{s_k} and n_{t_k} unconnected. By contrast, in a more precise way, IBC considers that there may exist one or more connections between them in the MIoT, even if they require the intervention of nodes belonging to other networks.

The second betweenness centrality measure that we propose here is called *Soft Cross Betweenness Centrality* (SCBC) and is defined as follows. Let $n_{j_k} \in N_k$ be the

node corresponding to the instance ι_{j_k} of the object o_j in the IoT $\mathcal{I}_k$. The Soft Cross Betweenness Centrality $SCBC(n_{j_k})$ is defined as:

$$SCBC(n_{j_k}) = \sum_{n_{s_u} \in N_u, n_{t_v} \in N_v, u \neq v} \frac{\overline{\sigma}_{n_{s_u} n_{t_v}}(n_{j_k})}{\overline{\sigma}_{n_{s_u} n_{t_v}}}$$

In few words, $SCBC(n_{j_k})$ computes the centrality of a node by selecting only the shortest paths between nodes belonging to different networks. There is no constraint on the node n_{j_k} for which we are computing the SCBC. As a matter of fact, n_{j_k} could belong either to N_u or to N_v or, finally, to another IoT of the MIoT different from N_u and N_v.

SCBC can be considered as an evolution of BC capable of detecting central (in the betweenness centrality sense) c-nodes and i-nodes by taking into account that these nodes do not belong to a single-IoT scenario but that they are part of a MIoT, and this fact can influence the shortest paths considered in the computation of betweenness centrality.

The last betweenness centrality measure we are proposing here is called *Hard Cross Betweenness Centrality* (HCBC) and is defined as follows. Let $n_{j_k} \in N_k$ be the node corresponding to the instance ι_{j_k} of the object o_j in the IoT $\mathcal{I}_k$. The Hard Cross Betweenness Centrality $HCBC(n_{j_k})$ is defined as:

$$HCBC(n_{j_k}) = \sum_{n_{s_u} \in N_u, n_{t_v} \in N_v, k \neq u, k \neq v, u \neq v} \frac{\overline{\sigma}_{n_{s_u} n_{t_v}}(n_{j_k})}{\overline{\sigma}_{n_{s_u} n_{t_v}}}$$

In few words, analogously to $SCBC(n_{j_k})$, $HCBC(n_{j_k})$ computes the centrality of a node by selecting only the shortest paths between nodes belonging to different networks. Furthermore, differently from the definition of SCBC, the node n_{j_k} is constrained to belong to a network different from the ones of the source and the destination nodes of the path.

HCBC can be considered as an evolution of BC along the same direction as SCBC. The only difference between SCBC and HCBC is that the latter is capable of detecting central c-nodes and i-nodes linking *at least three* IoT.

IBC, SCBC and HCBC are capable of overcoming the limits characterizing the classic BC in a MIoT. We remark again that IBC is different from the classical BC because it considers that the corresponding IoT is not isolated but inside the MIoT. Given the complexity of a MIoT, such a specific study can be really useful for several applications.

By contrast, if we want to know the most central nodes in a MIoT, the most suitable choices are SCBC and HCBC. SCBC is capable of highlighting the most suitable nodes which allow the cooperation of nodes belonging to different IoT. The term "Soft" characterizing SCBC is due to the soft restrictions of its constraints.

HCBC, instead, is much more restrictive than SCBC. As a consequence, it detects few nodes presenting very high values of betweenness centrality. In fact, they ensure a high cooperation level in the MIoT because they are linked to a higher number of IoT than the other nodes.

The choice between SCBC and HCBC depends on the application context. For instance, if we consider information diffusion, SCBC is well suited for fast information diffusion. HCBC, instead, is a better choice for spreading information among many IoT, even though the diffusion process will be slower than the one guaranteed by SCBC, because of the reduced number of nodes with a high HCBC.

5.2.4 Results

5.2.4.1 Testbed

We derived our testbed from *Thingful*[6], a search engine for the Internet of Things supporting the search of data regarding a huge number of existing things, distributed all over the world. Thingful also provides some suitable APIs, which can be used for querying it through a software program and which we exploited for the construction of our testbed. In order to obtain our testbed, we needed to perform several tasks. They are described in detail in [82]. Here, we limit ourselves to illustrate the characteristics of our testbed thus allowing the reader to understand the presented experiments.

Our MIoT consists of 11 IoT, reported in the first column of Table 5.15. We associated an object with each thing. Since we had 250 things, we obtained 250 objects. 200 of these objects had associated only one instance; 35 of them had associated two instances; finally, 15 of them had associated three instances. As a consequence, we had 315 instances in our testbed, distributed among the 11 IoT of our MIoT, as shown in Table 5.15.

A (necessarily complex) visualization of our testbed is presented in Figure 5.4. The interested reader can find the corresponding dataset (in `.csv` format) at the address `www.barbiana20.unirc.it/miot/datasets/miot2`. The password to type is "`za.12&;1q74:#`".

5.2.4.2 Evaluating the MIoT-oriented betweenness centrality

In this section, we describe the tests that we carried out to evaluate the significance of our new betweenness centrality measures in a MIoT and to compare them with the classical betweenness centrality. In our test activity, we adopted the testbed illustrated in the previous section.

[6] Thingful: a Search Engine for the Internet of Things - `https://thingful.net`

IoT	*Number of instances*
a.home	22
a.health	22
a.energy	22
a.transport	22
a.environment	22
b.near	14
b.mid	38
b.far	53
c.plain	44
c.hill	50
c.mountain	6

Table 5.15: Number of instances present in each IoT of our MIoT

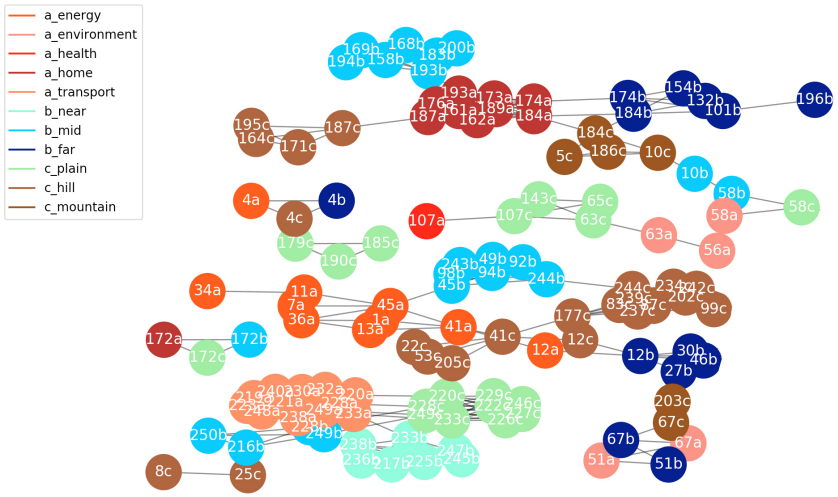


Fig. 5.4: A graphical representation of our MIoT

We started our experiments considering the top-12 central nodes returned by BC and verifying the rank of the same nodes when the other centrality measures are applied[7]. Obtained results are reported in Table 5.16.

From the analysis of this table we can clearly observe that BC and IBC return completely different results. In fact, 11 of the top-12 central nodes returned by BC have a rank higher than 200 in IBC. Instead, a good correspondence can be observed between the ranks of BC and SCBC, denoting that BC shows a good capability of finding the most "soft" central nodes in a MIoT. By contrast, there is a very loose correspondence between BC and HCBC. This denotes that BC is incapable of finding

[7] Recall that our MIoT consists of 315 nodes.

Nodes	*BC rank*	*IBC rank*	*SCBC rank*	*HCBC rank*
76b	1	208	1	1
76c	2	207	2	2
99b	3	202	3	48
99c	4	201	4	47
54b	5	2	158	98
12b	6	293	5	3
76a	7	209	6	4
41a	8	232	7	116
244c	9	245	8	143
244b	10	246	9	144
149c	11	288	10	258
12a	12	294	11	5

Table 5.16: IBC, SCBC and HCBC ranking of the top-12 central nodes returned by BC

the most central hard c-nodes. In conclusion, it seems that the BC's incapability of distinguishing between c-nodes and i-nodes and between c-edges and i-edges leads it to show a behavior (someway similar to the one of SCBC) intermediate between IBC and HCBC.

Then, we repeated the same evaluation for the top-12 central nodes returned by IBC. Obtained results are reported in Table 5.17. From the analysis of this table we can observe that the ranks returned by IBC and those returned by SCBC and HCBC are totally different. Actually, this was an expected result. However, it is interesting to observe that there is a weak correspondence between IBC and BC, because the top-12 central nodes returned by IBC have a rank between 5 and 95 in BC.

After this, we analyzed the top-12 central nodes returned by SCBC. Obtained results are reported in Table 5.18. Again, we observe a certain correspondence between SCBC and BC, a totally different behavior characterizing SCBC and IBC and a weak correspondence between SCBC and HCBC.

All the previous conclusions are confirmed by the analysis of the top-12 central nodes returned by HCBC, reported in Table 5.19. Observe, also, in this table the substantial difference between HCBC and SCBC, due to the restriction characterizing the definition of the former.

To further verify our previous conclusions and to quantify them, we decided to apply the Kendall Tau rank distance metric [375]. This is a metric aiming at measuring the differences between two different rankings by counting the number of

Nodes	*IBC rank*	*BC rank*	*SCBC rank*	*HCBC rank*
177c	1	37	248	224
54b	2	5	158	98
57b	3	55	156	94
33c	4	72	173	127
21c	5	74	208	172
211a	6	29	216	182
133c	7	76	289	277
91a	8	63	124	56
212c	9	65	215	181
156b	10	82	267	249
144c	11	94	277	265
142c	12	95	279	267

Table 5.17: BC, SCBC and HCBC ranking of the top-12 central nodes returned by IBC

Nodes	*SCBC rank*	*BC rank*	*IBC rank*	*HCBC rank*
76b	1	1	208	1
76c	2	2	207	2
99b	3	3	202	48
99c	4	4	201	47
12b	5	6	293	3
76a	6	7	209	4
41a	7	8	232	116
244c	8	9	245	143
244b	9	10	246	144
149c	10	11	288	258
12a	11	12	294	5
40c	12	13	233	117

Table 5.18: BC, IBC and HCBC ranking of the top-12 central nodes returned by SCBC

pairwise disagreements between them. More formally, it determines the number of swaps necessary to make the two ranks equal. The higher its value, the higher the distance between the two ranks.

We computed the Kendall Tau rank distance metric for all the possible pairs of ranks determined by considering the four metrics mentioned above. Obtained results are reported in Table 5.20. From the analysis of this table we can see that all

Nodes	*HCBC rank*	*BC rank*	*IBC rank*	*SCBC rank*
76b	1	1	208	1
76c	2	2	207	2
12b	3	6	293	5
76a	4	7	209	6
12a	5	12	294	11
191c	6	14	269	13
2c	7	20	237	19
191a	8	22	271	21
2a	9	26	239	25
12c	10	35	292	33
2b	11	38	238	35
184a	12	42	276	39

Table 5.19: BC, IBC and SCBC ranking of the top-12 central nodes returned by HCBC

τ_1	τ_2	$K(\tau_1, \tau_2)$
BC	IBC	18204
BC	$SCBC$	8489
BC	$HCBC$	24997
IBC	$SCBC$	27907
IBC	$HCBC$	30195
$SCBC$	$HCBC$	14816

Table 5.20: Values of Kendall Tau rank distance for all the possible pairs of Betweenness Centralities

of our previous conjectures about the metric characteristics and similarities are confirmed. In fact, we can see that IBC and HCBC are completely different. The same happens for IBC and SCBC. Quite a high difference can be observed for BC and HCBC. A certain (not very high) difference can be observed for BC and IBC and for SCBC and HCBC. Finally, BC and SCBC present the highest similarity.

5.3 Communication Scope in a MIoT

5.3.1 Introduction

When we throw a stone in a pond, we can see that the water moves, and small waves are created. These waves are higher in the proximity of the stone and, as we move away from it, they become smaller and smaller until they disappear. Generally, the heavier the stone, the higher the initial waves and the farther they arrive. This image, in our opinion, describes better than anything else what is meant by "scope". In the Concise Oxford Dictionary [8], *scope* is defined as "*the extent of the area or subject matter that something deals with or to which it is relevant*".

We can surely find several analogies between scope and some other concepts used in sociology; think, for instance, of centrality, reliability, power, reputation, influence, trust, diffusion, etc. [650, 508]. Actually, scope goes beyond these concepts and simultaneously embraces them and is influenced by all of them.

Scope has been investigated by social network researchers in the past [413, 374, 448, 449, 479, 569, 678]. In the meantime, social networks have become more and more complex, and social networking has evolved into social internetworking [517, 134]. In this new context, some social networks interact with each other thanks to some users, called bridges, each joining at least two social networks. Bridges play a key role in social internetworking because they allow users of different social networks to interact with each other.

Along with social internetworking, another key phenomenon we are experiencing in the last few years is the presence of increasingly smart and social objects [273]. This is deeply influencing the Internet of Things (hereafter, IoT) scenario [711]. As a consequence of this fact, an increasingly high number of authors have begun to investigate the behavior of smart objects and to analyze their profiles and social interaction [213]. As a matter of fact, several architectures performing these tasks have been recently proposed in literature; think, for instance, of the most recent ones, i.e., Social Internet of Things (hereafter, SIoT [70]), Multiple IoT Environment (MIE [81]) and Multiple Internets of Things (hereafter, MIoT [82, 434, 650]). MIoT is the most recent of them and, for this reason, considers the most recent results obtained by researchers on IoT. A MIoT can be modeled as a set of IoT, which interact with each other through those objects, called "cross-objects" (analogous to bridges in social internetworking scenarios), which belong to more IoT. From this definition it is clear that the MIoT paradigm is an attempt to extend the social internetworking ideas to IoT.

[8] Concise Oxford Dictionary - `https://en.oxforddictionaries.com`

In spite of the high number of researches on IoT performed in the latest years, to the best of our knowledge no investigation on the scope of an object in a MIoT, or at least in an IoT, has been yet proposed. Actually, some aspects presenting several relationships with scope have been analyzed in IoT or, in some cases, in the SIoT context (think, for instance, of [510, 580, 63, 726, 126]). However, none of them is as general as the investigation of the scope in a MIoT could be.

In this chapter, we contribute to fill this gap by introducing and analyzing the concept of scope of a smart object in a MIoT. Specifically, we present two formalizations of this concept. The former is called Naive; it is simple (because it considers only trust), but it does not take into account all the factors that could play a key role in this context. The latter is called Refined; it is quite complex, but it takes all the possible involved factors into account; in fact, it considers trust, proactivity, stimulation capability and security level.

After having introduced both these formalizations, we analyze them through a set of experiments devoted to understanding the pros and the cons of each of them. Furthermore, these experiments are conceived to highlight the relationships between centrality measures and scope, as well as the possible connection between this last parameter and network density. Moreover, we experimentally compare our definition of scope with two related concepts (i.e., diffusion degree and influence degree) proposed in past literature on IoT. This analysis reveals that scope provides a balanced assessment of the "power" of a smart object over its neighbors. Indeed, its assessment is intermediate between the one returned by diffusion degree (which is overly optimistic) and the one provided by influence degree (which is overly pessimistic). We also examine related literature to evidence the analogies and the differences between the previous proposals and the one illustrated here. Finally, we present two case studies (i.e., a smart city and a shopping center) where scope can play an important role.

The outline of this chapter is as follows: in Section 5.3.2, we present an overview of related literature. In Section 5.3.3, we describe the novelties introduced to the MIoT paradigm in order to model our scenario, illustrate the concept of scope and present two formalizations of it. In Section 5.3.4, we report our testbed and the set of experiments performed on it. Finally, in Section 5.3.5, we describe two typical use cases benefiting from this definition of scope.

5.3.2 Related Work

In this section, we provide a comparison between our approach and related literature. Before starting this discussion, a preliminary consideration about the MIoT model is in order, because it is the substrate which our definition of scope relies

on. Indeed, the MIoT model adopts an abstract perspective of IoT, different from a technical one. It does not aim at handling technological heterogeneities and other challenging technological issues. Instead, it aims at providing a high-level representation of interconnected IoT, which, thanks to the adoption of metadata, is independent from the underlying technology. The definition of a semantics-based representation of IoT is currently considered one of the main challenging issues in this research field [70]. Some preliminary attempts in this direction have been recently proposed in literature. One of the most known of these attempts is SIoT [70]. However, this model is still strictly related to technological issues because the forms of relationships between objects proposed by the authors, namely *(i)* parental object relationship; *(ii)* co-location object relationship; *(iii)* co-work object relationship; *(iv)* ownership object relationship; *(v)* social object relationship, are only partially semantic. Actually, the MIoT model captures different aspects w.r.t. SIoT. Indeed, it focuses on data-driven and semantics-based aspects and not on technological ones; as a matter of fact, it considers the contents exchanged by smart objects [277] during their transactions.

After this premise, we can start to overview related literature. In order to perform this activity better and to define some guidelines for comparing other approaches with ours, in Table 5.21 we provide an overview of the most important features that should characterize approaches conceived to evaluate scope or other related parameters in an IoT scenario. In particular, we consider the following features: *(i)* capability of handling a trade-off between quality of results and running time; *(ii)* capability of handling labeled networks; *(iii)* capability of handling multiple IoT or multiple complex networks; *(iv)* usage of content and relationship data within the approach; *(v)* usage of structural properties; *(vi)* usage of physical information concerning IoT, and *(vii)* application in recommendation services.

The classical IoT architectures share some similarities with the classical social networks, whereas social IoT paradigms (such as SIoT [70], MIE [81], and MIoT [82]) share some similarities with Social Internetworking Systems [134, 514]. Actually, to the best of our knowledge, no investigation about the scope in a multiple IoT scenario has been proposed in past literature, whereas very few approaches investigate concepts similar to the impact of smart objects in IoT. Furthermore, when this last investigation is performed, it is limited to a single IoT and no extension to multiple IoT is performed. As there is no past approach that simultaneously examines all the issues reported here, in the following, we will focus on single aspects of the overall analysis, such as the kind of interaction, the network complexity, the kind of exchanged information, and so forth.

	Management of a trade-off between quality of results and execution time	Management of labeled networks	Management of multiple IoT and/or networks	Data-driven approach	Usage of structural properties	Usage of physical information concerning IoT	Applicability in recommendation services
Our approach	✓	✓	✓	✓	✓	-	-
[510]	-*	-	-	✓	-	✓	-
[63]	-*	-	-	✓	✓	✓	-
[726]	-	✓	-	✓	-	✓	-
[499]	-	-	-	✓	-	-	-
[406]	-	-	-	✓	-	-	-
[699]	-	✓	✓	✓	✓	-	✓
[343]	✓	✓	-	-	✓	-	✓
[457]	-	-	-	-	✓	-	-
[672]	✓	-	-	-	✓	-	-
[269]	-*	-	-	✓	-	-	✓

Table 5.21: A taxonomy of approaches evaluating scope or related parameters in IoT. The symbol * denotes that the corresponding feature is not directly present, but may be re-constructed indirectly

In the context of social networks, many investigations focusing on the centrality of a node have been performed. The interested reader can see [217] for a survey on this topic. In [453], the authors investigate the evolution of the centrality of nodes in complex dynamic networks, where nodes and links may appear and disappear over time and may move over the network. In [715], the authors propose an analysis of customer engagement in complex social networks. It evidences that many important dimensions used to study customer engagement are similar to the ones that we consider for scope computation. In [628], the authors exploit the posts of users to analyze the information flow in a network. In [727], the authors propose an approach that generates a bipartite graph between users and contents; then, they employ it to measure the influence of users in the corresponding social network. In particular, this influence is computed by leveraging random walks on this graph, along with a related Markov chain model.

In [553], the authors define a new model where the influence of a user is based on her attractiveness, that is the number of other new users with whom she established relations over time. Another interesting concept introduced in the analysis of content sharing is the one of "information cascade". This term is used to denote the investigation of how diffusion protocols can affect the way information is diffused within a network. Understanding how information is disseminated among users can support the detection of the most influential ones in a network. This issue has been recently addressed in [190] in the context of complex networks. Information cascade shares some aspects with our concept of scope. However, there is an important difference between these two concepts in that the former aims at modeling the whole information flow in a network, whereas the latter focuses on the evaluation of the impact degree on the subnetwork of the MIoT coinciding with the ego network centered on the node whose scope we want to analyze.

Information diffusion and propagation have been also analyzed in IoT contexts at different levels [510, 580, 63, 726, 126, 686]. For instance, in [510], the authors investigate information diffusion in narrowband IoT with the goal of optimizing information flow at network level. In [63], the authors investigate the adoption of context-aware information diffusion to alert messages in 5G mobile social networks. Both [510] and [63] exploit IoT physical information, which is a feature not considered by our approach. However, several aspects covered by our proposal are not considered in these two approaches. For example, they do not consider the context of multiple IoT and handle a trade-off between quality of results and running time only partially. Finally, [510] does not exploit structural properties of networks.

An interesting approach to content dissemination in the Internet of Vehicles (IoV) is described in [726]. Here, the authors investigate how to combine the information coming from the physical layer with the one regarding the social layer to perform a rapid content dissemination in IoV networks. The approach of [726] exploits physical information, which is not considered by our approach. On the other side, differently from our approach, it does not address the multiple IoT context. Furthermore, it does not provide the possibility to tune a trade-off between quality of results and running time, which is a feature provided by our approach.

Significant research efforts have been devoted to studying the interaction between objects in complex IoT [499]. As an example, in [406], the authors present an IoT application in the context of smart cities, a scenario in which an IoT system can reach large scale dimensions. [406] also introduces the concept of IoT hub. The features of these two approaches are only marginally overlapping with our own. In fact, analogously to our approach, they are data driven. However, they do not con-

sider the structural properties of networks, do not handle a multiple IoT scenario, and do not manage a trade-off between quality of results and running time.

Another line of research on IoT regards the design of approaches to recommender systems and services in IoT contexts; an overview of these approaches is presented in [260]. As for this research line, in [269], the authors propose a multi-agent recommender system for IoT aiming at producing a set of significant suggestions for a user with specific characteristics. Here, smart objects are represented through bit vectors, called thing descriptors, managed by cyber-agents. Smart objects can be linked together and, then, can be managed by neighbor cyber-agents. The approach of [269] is more oriented to analyze recommendation processes than to investigate information diffusion, which our approach is centered on. Differently from our approach, the approach of [269] does not exploit structural properties, and does not handle multiple IoT. Finally, it manages a sort of trade-off, but this last regards the traffic load generated and the number of hops performed and, therefore, is completely different from the trade-off considered by our approach.

In [699], the authors propose an approach that integrates the concept of social network of users and IoT. It merges information coming from social networks of users and correlation networks of things by learning shared latent factors. To perform this task, it exploits a technique for probabilistic matrix factorization. The approach [699] addresses smart object recommendation in IoT, a feature not directly provided by our approach. On the other side, the concept of scope could be adopted in [699] as a further factor to determine relationships across heterogeneous smart objects in IoT. As a consequence, the two approaches can be considered orthogonal, even if they share several common features. In fact, both of them are able to deal with several IoT and labeled networks, and both of them exploit contents and relationships to address their tasks. Differently from our approach, the approach of [699] does not allow the management of the trade-off between quality of results and running time.

Beside the approaches regarding social networks or IoT, several related studies can be found when other forms of complex heterogeneous networks are considered. For instance, Heterogeneous Information Network (hereafter, HIN) is a graph model whose nodes and edges are annotated with types. A challenging issue in HINs is the computation of the closeness between two nodes, interpreted as the relevance of one of them for the other. In [343], the authors address this issue by introducing the concept of meta-structure. This is a directed acyclic graph of object types with edge types connecting in between. The approach of [343] shares several similarities with our own. Indeed, both of them use labeled networks and structural properties, and both of them are able to tune the quality of results and running time based

on some parameters. Differently from the approach of [343], our own considers a multiple IoT scenario and exploits data exchanged among objects. On the other side, the approach of [343] differs from ours because it studies the properties of meta-structures in the recommendation context, which is a feature we plan to address in the future.

In [457], the authors propose an analysis for detecting influential nodes in complex networks. To address this issue, they identify relevant graph substructures, called maximal k-trusses, conceived to characterize the ability of influential nodes better than the previously adopted measures, such as node degree, k-core index, etc. In [672], the authors present a new measure, called efficiency centrality, for identifying influential nodes. Like scope, this measure considers nodes and their neighbors. However, it ranks spreaders in the whole network by removing nodes and considering the changes in the degrees of the other nodes of the network after removal. Both [457] and [672] share with our approach the idea to study the influence of smart objects in a network using its structural properties. However, differently from [457] and [672], our approach also considers the data exchanged between smart objects and handles labeled networks. Moreover, it is specifically designed for a multiple IoT scenario. Finally, analogously to our approach, the one described in [672] can handle a trade-off between quality of results and running time.

In [439], the authors propose an extensive review of the identification of vital nodes in complex networks. The concept of vital node reflects a general property of a node that plays a critical role in some specific dynamical processes.

5.3.3 Methods

5.3.3.1 Extending the MIoT paradigm

In this section, we extend the MIoT paradigm introduced in Chapter 4 in order to make it capable of handling the concept of scope.

Consider a MIoT $\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$, where $\mathcal{I}_k$ is an IoT.

Let o_j be an object of $\mathcal{M}$. We assume that, if o_j belongs to $\mathcal{I}_k$, it has an instance ι_{j_k}, representing it in $\mathcal{I}_k$. ι_{j_k} has associated a Security Level λ_{j_k} whose possible values are: 1 = low, 2 = medium-low, 3 = medium, 4 = medium-high, 5 = high. It indicates how much the security requirements are tight for o_j in $\mathcal{I}_k$. Clearly, it depends on the nature of both o_j and $\mathcal{I}_k$, as well as on the role that o_j plays in $\mathcal{I}_k$.

The concept of neighborhood nbh_{j_k} of an instance ι_{j_k} in $\mathcal{I}_k$ is defined as:

$$nbh_{j_k} = out\_nbh_{j_k} \cup in\_nbh_{j_k}$$

where:

$$out\_nbh_{j_k} = \{n_{q_k} | (n_{j_k}, n_{q_k}) \in A_I, |tranSet_{jq_k}| > 0\}$$

and

$$in\_nbh_{j_k} = \{n_{q_k} | (n_{q_k}, n_{j_k}) \in A_I, |tranSet_{qj_k}| > 0\}$$

In other words, nbh_{j_k} comprises those instances directly connected to ι_{j_k} through an incoming or an outgoing arc, which shared at least one transaction with it performed in the past.

Given a pair of instances ι_{j_k} of o_j and ι_{q_k} of o_q in $\mathcal{I}_k$, the MIoT stores the set $tranSet_{jq_k}$ of the transactions from ι_{j_k} to ι_{q_k}. It is defined as:

$$tranSet_{jq_k} = \{T_{jq_{k_1}}, T_{jq_{k_2}}, \cdots, T_{jq_{k_v}}\}$$

A transaction $T_{jq_{k_t}} \in tranSet_{jq_k}$ is represented as follows:

$$T_{jq_{k_t}} = \langle req_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}}, success_{jq_{k_t}}, content_{jq_{k_t}} \rangle$$

Here:

- $req_{jq_{k_t}}$ denotes if ι_{j_k} started $T_{jq_{k_t}}$ as an answer to a specific request of ι_{q_k} or if it started $T_{jq_{k_t}}$ proactively.
- $start_{jq_{k_t}}$ denotes the starting timestamp of $T_{jq_{k_t}}$.
- $finish_{jq_{k_t}}$ indicates the ending timestamp of $T_{jq_{k_t}}$.
- $success_{jq_{k_t}}$ denotes whether $T_{jq_{k_t}}$ was successful or not; it is set to *true* in the affirmative case, to *false* in the negative one, and to `NULL` if it is still in progress.
- $content_{jq_{k_t}}$ indicates the set of the content topics considered by $T_{jq_{k_t}}$. Specifically, it consists of a set of w keywords:

$$content_{jq_{k_t}} = \{kw^1_{jq_{k_t}}, kw^2_{jq_{k_t}}, \ldots, kw^w_{jq_{k_t}}\}$$

Now, we can define the set $tranSet_{j_k}$ of the transactions activated by ι_{j_k} in $\mathcal{I}_k$. Specifically, let $\iota_{1_k}, \iota_{2_k}, \cdots, \iota_{w_k}$ be all the instances belonging to $\mathcal{I}_k$. Then:

$$tranSet_{j_k} = \bigcup_{q=1..w, q \neq j} tranSet_{jq_k}$$

In other words, the set $tranSet_{j_k}$ of the transactions of an instance ι_{j_k} is given by the union of the sets of the transactions from ι_{j_k} to all the other instances of $\mathcal{I}_k$.

From the above characterization, it clearly emerges that the MIoT paradigm deeply differs from the so called cross-domain IoT. They both deal with an interconnection of, often heterogeneous, IoT; however, the MIoT adopts an abstract perspective, while the cross-domain IoT a technical one. Indeed, the cross-domain IoT mainly addresses low-level concerns deriving from the technological heterogeneity

$\mathcal{M}$	a MIoT
$\mathcal{I}_k$	an IoT
o_j	an object of a MIoT
ι_{j_k}	an instance of an object o_j in $\mathcal{I}_k$
G_k	a graph associated with an IoT $\mathcal{I}_k$
N_k	the set of the nodes of G_k
A_k	the set of the arcs of G_k
nbh_{j_k}	the neighborhood of an instance ι_{j_k} in $\mathcal{I}_k$
$out\_nbh_{j_k}$	the instances connected to ι_{j_k} through an outgoing arc
$in\_nbh_{j_k}$	the instances connected to ι_{j_k} through an incoming arc
$tranSet_{jq_k}$	the set of the transactions from ι_{j_k} to ι_{q_k}
$T_{jq_{k_t}}$	a transaction of the set $tranSet_{jq_k}$
$reposted_{j_k}$	the set of the transactions received by ι_{j_k} and reposted by it
$elaborated_{j_k}$	the set of the transactions received by ι_{j_k} and whose contents it elaborated for its purposes
$requested_{j_k}$	the set of the transactions explicitly requested by ι_{q_k}
PD_{j_k}	the proactivity degree of an instance ι_{j_k}
π_{jq_k}	the minimum path from an instance ι_{j_k} to an instance ι_{q_k}
InD_{j_k}	the Inactivity Degree of an instance ι_{j_k}
TD_{qj_k}	the Trust Degree of ι_{q_k} in ι_{q_k}
NID_{j_k}	the Naive Impact Degree of an instance ι_{j_k}
RID_{j_k}	the Refined Impact Degree of an instance ι_{j_k}
NS	Naive Scope
RS	Refined Scope

Table 5.22: Main abbreviations used throughout this chapter

– typical of IoT belonging to different domains – and places the interoperability issue on the spotlight [276]. The MIoT, instead, is more abstract, yet more flexible, by providing a high-level, technology agnostic (i.e., metadata- and metamodel-based) representation of interconnected and heterogeneous IoT which, in addition, can also be implemented.

5.3.3.2 Scope definition

In this section, we present the definition of the scope of an instance ι_{j_k} in an IoT $\mathcal{I}_k$ and the scope of an object o_j in a MIoT $\mathcal{M}$. For this purpose, we must introduce some preliminary concepts. They are also reported in Table 5.22.

The first of them regards the *Proactivity Degree* PD_{j_k} of an instance ι_{j_k} in an IoT $\mathcal{I}_k$. PD_{j_k} ranges in the real interval $[0,1]$ and is set equal to the fraction of the transactions received by ι_{j_k} that it reposts to another instance of $\mathcal{I}_k$ or whose contents it elaborates for its purposes.

To formalize this concept, we must introduce:

- the set $reposted_{j_k}$ of the transactions received by ι_{j_k} and reposted by it;
- the set $elaborated_{j_k}$ of the transactions received by ι_{j_k} and whose contents it elaborated for its purposes.

PD_{j_k} can be formalized as follows:

$$PD_{j_k} = \frac{|tranSet_{j_k} \cap (reposted_{j_k} \cup elaborated_{j_k})|}{|tranSet_{j_k}|}$$

Now, we need to introduce the neighborhood of level t of an instance ι_{j_k} in its IoT $\mathcal{I}_k$. It is an extension of the concept of $out\_nbh_{j_k}$ presented in Section 5.3.3.1. It is defined as follows:

$$out\_nbh^t_{j_k} = \begin{cases} out\_nbh_{j_k} & \text{if } t = 0 \\ \{\iota_{r_k} | \iota_{r_k} \in out\_nbh_{q_k}, \iota_{q_k} \in out\_nbh^{t-1}_{j_k}, \iota_{r_k} \notin out\_nbh^w_{j_k}, 0 \leq w < t\} & \text{if } t > 0 \end{cases}$$

The concept of $out\_nbh^t_{j_k}$ will be extremely important later. In the meantime, we introduce a new concept, namely the minimum path π_{jq_k} from an instance ι_{j_k} to an instance $\iota_{q_k} \in out\_nbh^t_{j_k}$. π_{jq_k} is defined as the sequence of instances $\{\iota_{0_k}, \iota_{1_k}, \ldots, \iota_{t_k}\}$ such that $\iota_{0_k} = \iota_{j_k}$, $\iota_{t_k} = \iota_{q_k}$, $\iota_{w_k} \in out\_nbh_{(w-1)_k}$ for $1 \leq w \leq t$.

Afterwards, we introduce the definition of the *Trust Degree* TD_{qj_k} of an instance ι_{q_k} in the instance ι_{j_k} in $\mathcal{I}_k$. It can be defined as the fraction of the transactions sent by ι_{j_k} to ι_{q_k} that have been requested by ι_{q_k} or that ι_{q_k} has considered so interesting to repost or elaborate them[9]. In order to formalize TD_{qj_k}, we must preliminarily introduce the set $requested_{q_k}$ of the transactions explicitly requested by ι_{q_k}. Now, TD_{qj_k} can be expressed as:

$$TD_{qj_k} = \frac{|tranSet_{jq_k} \cap (requested_{q_k} \cup reposted_{q_k} \cup elaborated_{q_k})|}{|tranSet_{jq_k}|}$$

Starting from this definition and the concepts of $out\_nbh^t_{j_k}$ and π_{jq_k}, we can proceed with the transitive closure of TD_{qj_k}. In particular, the general definition of TD_{qj_k} is as follows:

$$TD_{qj_k} = \begin{cases} \frac{|tranSet_{jq_k} \cap (requested_{q_k} \cup reposted_{q_k} \cup elaborated_{q_k})|}{|tranSet_{jq_k}|} & \text{if } \iota_{q_k} \in out\_nbh_{jk} \\ \prod_{w=1}^{t} TD_{((w-1)w)_k} & \text{if } \iota_{q_k} \in out\_nbh^t_{j_k},\ t > 0, \pi_{jq_k} = \{\iota_{0_k}, \iota_{1_k}, \cdots, \iota_{t_k}\} \end{cases}$$

Intuitively, the Trust Degree TD_{qj_k} of ι_{q_k} is given by the base formula if ι_{q_k} is directly connected to ι_{j_k}; otherwise, it is obtained by the product of the trust degrees associated with the pairs of instances belonging to the minimum path from ι_{j_k} to ι_{q_k}.

The next step regards the definition of the concept of Impact Degree of an instance ι_{j_k} in $\mathcal{I}_k$. Actually, we can define two forms of Impact Degree. The first one is simple and immediate to compute; we call it *Naive Impact Degree* (hereafter, NID). The second one is more accurate and precise, even if computationally more expensive; we call it *Refined Impact Degree* (hereafter, RID).

We start by introducing the Naive Impact Degree NID_{j_k} of ι_{j_k} in $\mathcal{I}_k$. It is defined as the average of the Trust Degrees that all the instances belonging to $out\_nbh_{j_k}$ have in ι_{j_k}. It can be formalized as follows:

[9] Clearly, it might happen that an unrequested transaction of $tranSet_{jq_k}$ is not considered interesting by ι_{q_k}. In this case, ι_{q_k} neither posts nor elaborates it.

$$NID_{j_k} = \frac{\sum_{\iota_{q_k} \in out\_nbh_{j_k}} TD_{qj_k}}{|out\_nbh_{j_k}|}$$

After having defined the Naive Impact Degree, we can introduce the Refined Impact Degree. Its definition is based on the following considerations:

(C_1) Given an instance ι_{j_k}, the higher the number of transaction requests received by the other instances of $\mathcal{I}_k$, the higher its RID.

(C_2) Given an instance ι_{j_k}, the higher its capability of leading an instance ι_{q_k} with a low proactivity degree to send one of its transactions to a further instance of $\mathcal{I}_k$, the higher its RID.

(C_3) Given an instance ι_{j_k}, the higher its capability of receiving a transaction sent by an instance ι_{r_k} with a low proactivity degree, the higher its RID.

(C_4) Given an instance ι_{j_k}, the higher its capability of leading an instance ι_{q_k} with a high RID to repost its transactions, the higher its RID.

Observe that Consideration C_4 is very complex to handle because it implies that the RID of an instance ι_{j_k} depends on the RID of an instance ι_{q_k}. This means that, for the computation of the instance RIDs, it would be necessary to solve (at least in the most complex case) huge systems, characterized by hundreds, or even thousands, of equations and variables. As a consequence, the computation of RID appears difficult to handle without a heuristic. Taking this consideration into account, we have defined a heuristic for the computation of RID. In particular, we consider the NID of ι_{q_k}, instead of the RID of this instance, in the computation of the RID of ι_{j_k}.

Taking Considerations (C_1) - (C_4) into account, RID_{j_k} can be defined as:

$$RID_{j_k} = \frac{\alpha \cdot RID1_{j_k} + \beta \cdot RID2_{j_k} + \gamma \cdot RID3_{j_k} + \delta \cdot RID4_{j_k}}{\alpha + \beta + \gamma + \delta}$$

In other words, RID_{j_k} is obtained as a weighted mean of four components, each formalizing one of the considerations presented above.

$RID1_{j_k}$ is associated with Consideration C_1. It is defined as follows:

$$RID1_{j_k} = \frac{|reqTranSet_{j_k}|}{maxCardReqTranSet_k}$$

Here:

- $reqTranSet_{j_k}$ is the set of the transactions from ι_{j_k} to any instance of $\mathcal{I}_k$ originated after a specific request:

$$reqTranSet_{j_k} = \bigcup_{\iota_{j_k} \in out\_nbh_{q_k}} reqTranSet_{jq_k}$$

In the previous formula, $reqTranSet_{jq_k}$ is the set of the transactions from ι_{j_k} to ι_{q_k} originated after a specific request of ι_{q_k}:

$$reqTranSet_{jq_k} = \{T_{jq_{k_t}} | T_{jq_{k_t}} \in tranSet_{jq_k}, req_{jq_{k_t}} = true\}$$

- $maxCardReqTranSet_k = max_{\iota_{j_k} \in \mathcal{I}_k} |reqTranSet_{j_k}|$.

$RID2_{j_k}$ is related to C_2. It is defined as follows:

$$RID2_{j_k} = \frac{\sum_{\iota_{q_k} \in out\_nbh_{j_k}} \frac{InD_{q_k}}{InD_k^{max}} \cdot \frac{|tranSet_{jq_k} \cup reposted_{q_k}|}{|tranSet_{jq_k}|}}{|out\_nbh_{j_k}|}$$

Here:

- InD_{q_k} is the *Inactivity Degree* of ι_{q_k} and is defined as $InD_{q_k} = 1 - PD_{q_k}$;
- InD_k^{max} is the maximum Inactivity Degree of an instance of $\mathcal{I}_k$.

$RID3_{j_k}$ is associated with C_3. It can be defined as follows:

$$RID3_{j_k} = \frac{\sum_{\iota_{r_k} \in in\_nbh_{j_k}} \frac{InD_{r_k}}{InD_k^{max}} \cdot \frac{|tranSet_{rj_k}|}{|tranSet_{r_k}|}}{|in\_nbh_{j_k}|}$$

Finally, $RID4_{j_k}$ is related to C_4. Taking into account the aforementioned reasoning about the need to simplify its computation by substituting RID_{j_k} with NID_{j_k}, it can be defined as follows:

$$RID4_{j_k} = \frac{\sum_{\iota_{q_k} \in out\_nbh_{j_k}} \frac{NID_{q_k}}{NID_k^{max}} \cdot \frac{|tranSet_{jq_k} \cup reposted_{q_k}|}{|tranSet_{jq_k}|}}{|out\_nbh_{j_k}|}$$

Here, $NID_{j_k}^{max}$ is the maximum Naive Impact Degree of an instance of $\mathcal{I}_k$.

Having defined the Naive and the Refined Impact Degree, we have almost all parameters necessary to define the Naive and the Refined Scope. Indeed, we need to define only a last one. It is the *Security Requirement Degree* SRD_{qj_k} and takes the level of the security tightness of ι_{j_k} and ι_{q_k} into account. In particular, it is defined as:

$$SRD_{qj_k} = min\left(1, \frac{\lambda_{j_k}}{\lambda_{q_k}}\right)$$

The rationale underlying this formula is as follows: as we will see later, SRD_{qj_k} contributes, along with TD_{qj_k}, to weight the Impact Degree that ι_{j_k} has on ι_{q_k}. If $\lambda_{j_k} < \lambda_{q_k}$ then the Security Level of ι_{q_k} is tighter than the one of ι_{j_k}; this condition represents an obstacle to the propagation of the contents of ι_{j_k} towards ι_{q_k}. Vice versa, if $\lambda_{j_k} \geq \lambda_{q_k}$ then the Security Level of ι_{j_k} is higher than or equal to the one of ι_{q_k}. This implies that, from the security viewpoint, there is no obstacle for the propagation of the contents of ι_{j_k} towards ι_{q_k}.

Observe that, if an instance ι_{j_k} has a high Security Level λ_{j_k} (for instance, $\lambda_{j_k} = 5$), then SRD_{qj_k} is high; as a consequence, ι_{j_k} can propagate all its contents towards the other instances. This because having a high Security Level means being highly secure or, in other words, having highly verified contents. This represents a pass for

the other instances that trust to receive content sent by ι_{j_k}. Therefore, in this sense, having a high Security Level makes it easy having a high scope.

We are now able to define the *Naive Scope* $NS^t_{j_k}$ (resp., the *Refined Scope* $RS^t_{j_k}$) of level t of an instance ι_{j_k} in $\mathcal{I}_k$. It is obtained as the weighted sum of the Naive Impact Degrees (resp., Refined Impact Degrees) of the instances belonging to $out\_nbh^t_{j_k}$, where the weights are the trust and the security values that these instances have in ι_{j_k}. This sum is, then, averaged by the number of instances belonging to $out\_nbh^t_{j_k}$. Formally speaking:

$$NS^t_{j_k} = \frac{\sum_{\iota_{q_k} \in out\_nbh^t_{j_k}} TD_{qj_k} \cdot NID_{q_k} \cdot SRD_{qj_k}}{|out\_nbh^t_{j_k}|}$$

$$RS^t_{j_k} = \frac{\sum_{\iota_{q_k} \in out\_nbh^t_{j_k}} TD_{qj_k} \cdot RID_{q_k} \cdot SRD_{qj_k}}{|out\_nbh^t_{j_k}|}$$

Now, we can define the *Naive Scope* NS^t_j (resp., the *Refined Scope* RS^t_j) of level t of an object o_j in the MIoT. It is obtained by averaging the Naive Scopes (resp., the Refined Scopes) of level t of its instances in the corresponding IoT. Specifically, let $Inst_j = \{\iota_{j_1}, \iota_{j_2}, \cdots, \iota_{j_l}\}$ be the instances of o_j in the IoT of the MIoT. Then:

$$NS^t_j = \frac{\sum_{\iota_{j_k} \in Inst_j} NS^t_{j_k}}{|Inst_j|} \qquad RS^t_j = \frac{\sum_{\iota_{j_k} \in Inst_j} RS^t_{j_k}}{|Inst_j|}$$

Considerations

After having provided a formalization of Naive and Refined Scope, we now present some considerations that highlight the connection between the formalized concepts and the general definition of scope. In this discussion, we mainly focus on Refined Scope, because this is the most advanced definition. We observe that our formalization of Refined Scope makes it holistic, allowing it to take a large variety of aspects into consideration. As a matter of fact, the Refined Scope of an instance ι_{j_k} considers the trust that the other instances of $\mathcal{I}_k$ have on it, the impact exerted by it on the other nodes and the tightness and the severity of its security requirements. In turn, the impact of ι_{j_k} considers its capability of receiving transaction requests from the other instances of $\mathcal{I}_k$ and its ability to stimulate them to deliver its contents. The overall set of these features is well suited to model, in the multiple IoT scenario, the concept of scope intended as "the extent of the area or subject matter that something deals with or to which it is relevant", as reported in the Concise Oxford Dictionary.

Even if scope may seem similar to context-awareness at a first sight, it actually presents important differences. Indeed, context-awareness in IoT is defined as any implicit or explicit information – current location, identity, activity, and physical condition – about the involved service stakeholders [540, 156]. By contrast, Refined

Scope is a data-driven and transaction-oriented concept, dealing with the contents exchanged among nodes and not with physical aspects.

Finally, observe that Refined Scope also handles privacy aspects, even if indirectly, thanks to the usage of the concepts of trust and security. As a matter of fact, in several scenarios, it is possible to find a certain correlation between trust and privacy in that the higher the trust, the higher the availability to exchange information. Analogously, the higher the Security Level of an instance, the higher its reliability and the higher the interactions and information exchange stimulated by it.

At a first glance, some of the concepts, and especially some of the activities, described above could appear far away from the IoT context. Think, for instance, of the concept of proactivity of a smart object and of the posting and elaborating activities. Actually, especially in the SIoT context, several models proposing concepts and activities similar to ours have been presented in recent literature. Indeed, one of these models is described in [316], where the authors discuss the Adaptive Interest Forward strategy. Some of the ideas underlying this strategy are close to the Considerations $C_1 - C_4$ representing the bases for the definition of the RID parameter in Section 5.3.3.2. In fact, in [316], the authors take two kinds of device into account, namely high- and low-capability devices[10]. The Adaptive Interest Forwarding strategy proceeds by prioritizing forwarding tasks from the node with the highest capabilities, while constrained nodes can transmit only if they do not overhear packet transmission from their neighbors.

Even if the two policies leading smart objects to transmit are different, it is possible to observe a parallelism between them. In fact, being proactive and able to stimulate the interest in the information sent through a transaction plays, in our approach, the same role as having capabilities in the approach of [316].

Actually, the parallelism is even closer. Indeed, we recognize a high similarity between:

- the situation in our approach where a smart object must decide whether or not reposting (intended as forwarding to other linked smart objects) a transaction received from another smart object, and
- the situation in the approach of [316] where an Information Centric Networking (hereafter, ICN) node receiving an Interest must decide whether or not forwarding it towards the producer.

In the same way, we can recognize a high similarity between:

[10] For the sake of clarity, we outline that the capability considered in [316] regards mainly energy and storage.

- the situation in our approach where a smart object decides to elaborate the content of a transaction (which could mean, for instance, selecting a part of a text or reducing the quality or the length of a video before reposting it), and
- the situation in the approach of [316] where an ICN receiving an Interest can decide to cache the content and send it according to an Adaptive Interest Forwarding strategy considering the status of node resources.

5.3.4 Results

In this section, we present the experiments we carried out to evaluate the performance of our approach from several viewpoints.

5.3.4.1 Testbed

In order to perform our experiments, as real MIoT with the dimension and the variety handled by our model do not exist yet, we constructed a MIoT simulator. This tool starts from real data and returns simulated MIoT with certain characteristics specified by the user.

The MIoT created by our simulator follow the paradigm described in Section 5.3.3.1. Our simulator is also provided with a suitable interface allowing a user to "personalize" the MIoT to build by specifying the desired values for several parameters, such as the number of nodes, the maximum number of instances of an object, and so forth.

To make "concrete" and "plausible" the simulated MIoT, we had the necessity that our simulator was capable of returning MIoT having the characteristics specified by the user and being as close as possible to real-world scenarios. In the simulator design, and in the next construction of the MIoT to use for the experiments, we followed the ideas expressed in [73, 74], in which the authors highlight that one of the main factors used to build links in an IoT is node proximity. In order to reproduce the creation of links among objects, we decided to leverage information about real-life paths in a city. In fact, having this information at disposal, we may associate each path with an object and link two objects if their paths have been near enough for a sufficient time period. As for a dataset containing real-life paths in a city, we selected the one reported in `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. It regards taxi routes in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. Each route contains several Points of Interests corresponding to the GPS coordinates of the vehicles. As said above, our simulator associates an object with a given route recorded in the dataset. Furthermore, it creates an arc between two nodes if the distance between the corresponding routes is

less than a certain threshold th_d for a predefined time interval th_t. The value of th_d and th_t can be specified through the constructor interface. Clearly, the higher the value of th_d and the lower the value of th_t, the more connected the constructed MIoT. The interested reader can find the MIoT created in this phase at the address `http://daisy.dii.univpm.it/miot/datasets/scope`. This MIoT consists of 1256 nodes. The six IoT of the MIoT had 128, 362, 224, 280, 98, and 164 nodes, respectively. The constructed MIoT is returned in a format that can be directly processed by the cypher-shell of Neo4J.

We carried out all the tests presented in this section on a server equipped with an Intel I7 Quad Core 7700 HQ processor and 16 GB of RAM with the Ubuntu 16.04 operating system. To implement our approach, we adopted *(i)* Python, as programming language, and *(ii)* Neo4J (Version 3.4.5), as underlying DBMS. In Figure 5.5, we report the activity diagram describing the various tasks performed by our MIoT simulator, along with the underlying logic. Furthermore, the code of our simulator is open source; the interested reader can access it at the address: `https://github.com/lucav48/miot-simulator`.

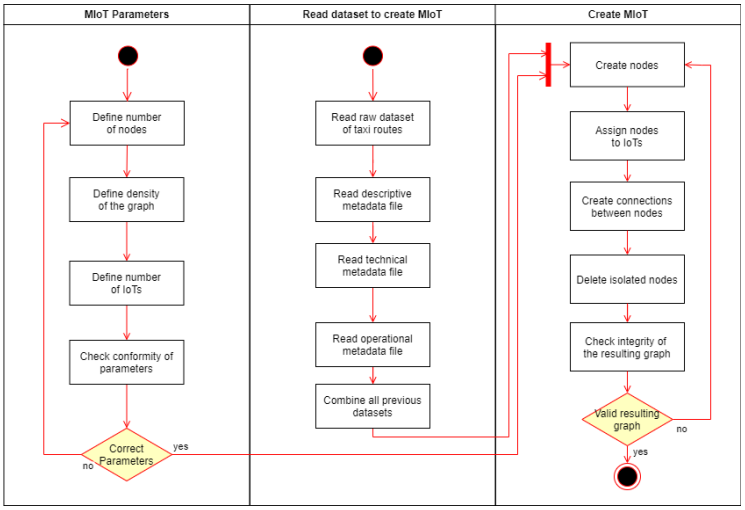


Fig. 5.5: Activity diagram of our MIoT simulator

5.3.4.2 Variation of the scope against the neighborhood level

In this experiment, we aimed at investigating the trend of the Naive Scope (hereafter, NS) and the Refined Scope (hereafter, RS) against the neighborhood level t (see Section 5.3.3.1). In particular, for each instance ι_{j_k} of the MIoT, we computed $NS^t_{j_k}$ and $RS^t_{j_k}$ when t increases from 1 to the diameter of $\mathcal{I}_k$. After this, we grouped the instances of our MIoT into clusters, based on some specific rationales, and we computed the variation of the average values of NS and RS for each group.

As a first task of this activity, we computed the variation of the average values of NS and RS for each IoT of the MIoT. This is equivalent to say that clusters coincided with IoT. The results obtained are reported in Figure 5.6. From the analysis of this figure, we can observe that, in each IoT, the values of NS and RS decrease quite quickly. As for NS, its value is extremely high when $t = 1$ in all the IoT. When $t = 2$, the value of NS is high for the largest IoT, whereas it is intermediate for the other ones. In any case, the values of NS become very low when t is greater than 3 for small IoT and when t is greater than 4 for large ones. As for RS, its trend against t is analogous to the one of NS. However, RS appears more capable than NS in distinguishing the neighborhoods with a high scope from those with a low one. In fact, in Figure 5.6, we can observe that the decrease from the high values of scope to the low ones is much steeper in RS than in NS. In our opinion, the capability of clearly discriminating the neighborhoods with high values of scope from the ones with low values of this parameter is an important feature for an approach aiming at formalizing the concept of scope.

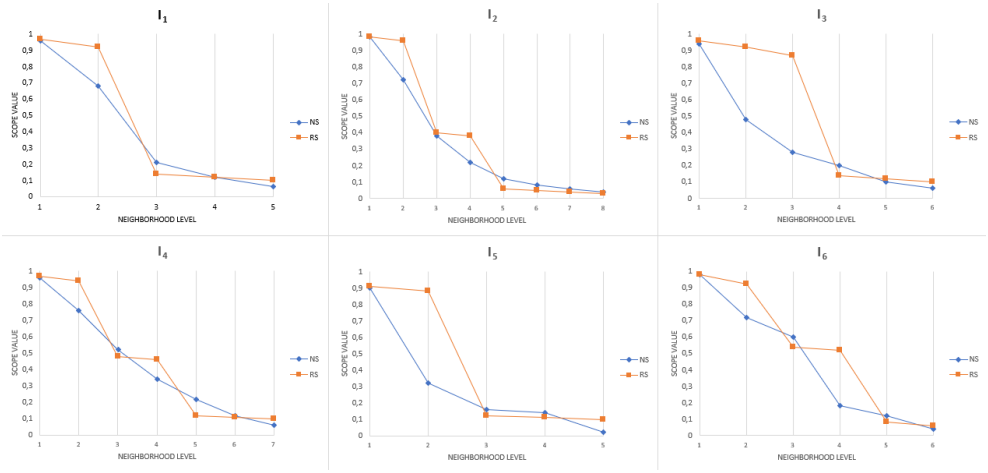


Fig. 5.6: Variation of the average values of NS and RS for each IoT of the MIoT against the neighborhood level

As a second task, we computed the variation of the average values of NS and RS for the whole MIoT. This is equivalent to say that we had a unique cluster coinciding with the MIoT. The results obtained are reported in Figure 5.7. From the analysis of this figure, we can conclude that NS (resp., RS) presents a trend similar to the one shown by it in the largest IoT of Figure 5.6. In particular, NS is very high for $t = 1$; it is high for $t = 2$; it has an intermediate value for $t = 3$, whereas it is low for $t > 5$. Instead, RS presents high values for $t = 1$ or $t = 2$; it shows intermediate values for $t = 3$ and $t = 4$ and low values for $t \geq 5$. Again, RS is more capable than NS in discriminating the neighborhoods with a high value of scope from the ones

characterized by an intermediate value of this parameter, and these last ones from the neighborhoods where RS has low values.

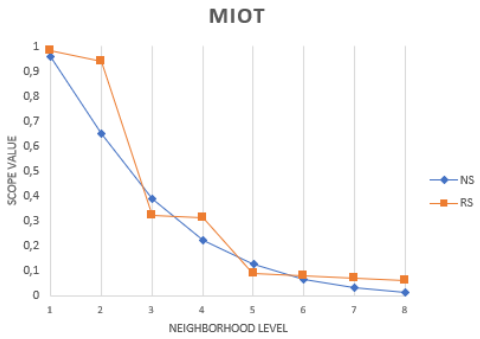


Fig. 5.7: Variation of the average values of NS and RS for the whole MIoT against the neighborhood level

As a final task, we grouped the available instances in two clusters containing c-nodes and i-nodes, respectively. Then, we computed the variation of the average values of NS and RS for the two clusters. The final goal of this task was to verify if i-nodes and c-nodes had different behaviors as far as their value of scope was concerned. The results obtained are reported in Figure 5.8. From the analysis of this figure we can observe that the values of NS decrease for both i-nodes and c-nodes. However, the corresponding trends are different. Indeed, the decrease is much smoother for i-nodes than for c-nodes. In particular, as for c-nodes, the decrease is very steep because the scope is less than 0.2 already for $t = 3$. As for RS, its trend for c-nodes is steeper than the one of NS; again, RS is more capable than NS in discriminating the neighborhoods with high, intermediate and low values of scope. Instead, the trend of RS for c-nodes is very similar to the corresponding trend of NS. Actually, this could have been expected because the trend of scope for NS was already very steep. The different trends of the values of scope for i-nodes and c-nodes can be explained by considering that, analogously to what was made in all the past approaches, our definition of neighborhood (which plays a key role in our definition of scope) considers as neighbors of a node only other nodes of the same IoT. In other words, it takes only i-arcs into account. Actually, we believe (and the results of Figure 5.8 confirm our belief) that it is worthwhile to investigate the role of c-arcs in the computation of the neighborhood of a node, and we plan to make this investigation in the future.

As for the analysis of the values of NS and RS for objects, we observe that they are obtained by averaging the values of NS and RS of the corresponding instances. As a consequence, it does not make sense to perform the first and the final tasks of the previous activity. The only task that makes sense is the second one; in this case,

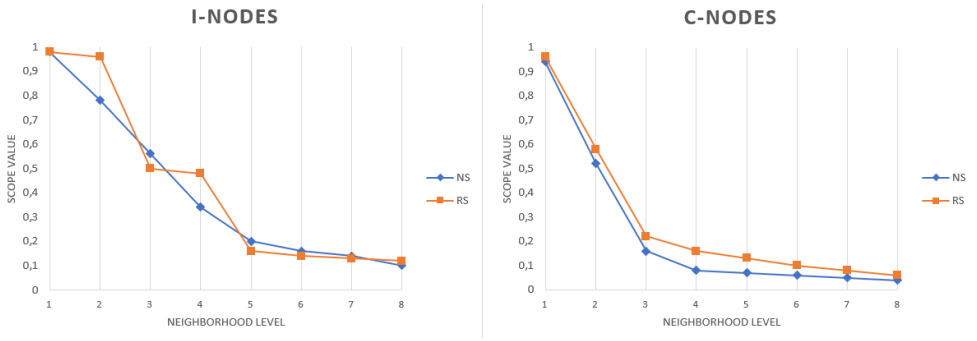


Fig. 5.8: Variation of the average values of NS and RS for the i-nodes and the c-nodes of the MIoT against the neighborhood level

the variation of the average values of NS and RS for the whole MIoT is reported in Figure 5.9.

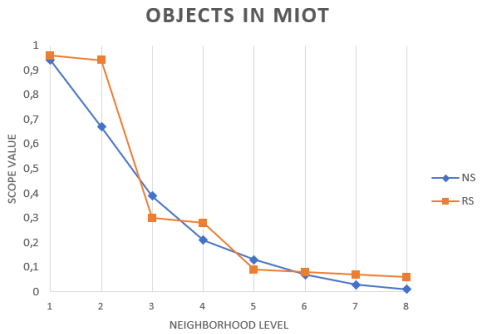


Fig. 5.9: Variation of the average values of NS and RS for the objects of the MIoT against the neighborhood level

As we could have expected, this trend is very similar (or, better, almost identical) to the one of Figure 5.7. This was not surprising for us; indeed, the value of NS and RS of an object is obtained by averaging the values of NS and RS of the corresponding instances. Therefore, it was to be expected that the trends of NS and RS for objects could not have been very different from the ones of NS and RS for instances.

5.3.4.3 Relationship between scope and centrality

In this second experiment, we aimed at investigating the relationships possibly existing between the scope and the main forms of centrality already considered in the literature. For this purpose, first we computed the degree, the closeness, the betweenness and the eigenvector centralities of all the instances of the MIoT. Then, we constructed the cluster $\mathcal{D}$ (resp., $\mathcal{C}$, $\mathcal{B}$ and $\mathcal{E}$) containing the 100 instances having the highest values of the degree (resp., closeness, betweenness and eigenvector) central-

ity. Finally, we computed the variation of the average values of NS and RS against the neighborhood level for the four groups. The results obtained are reported in Figure 5.10.

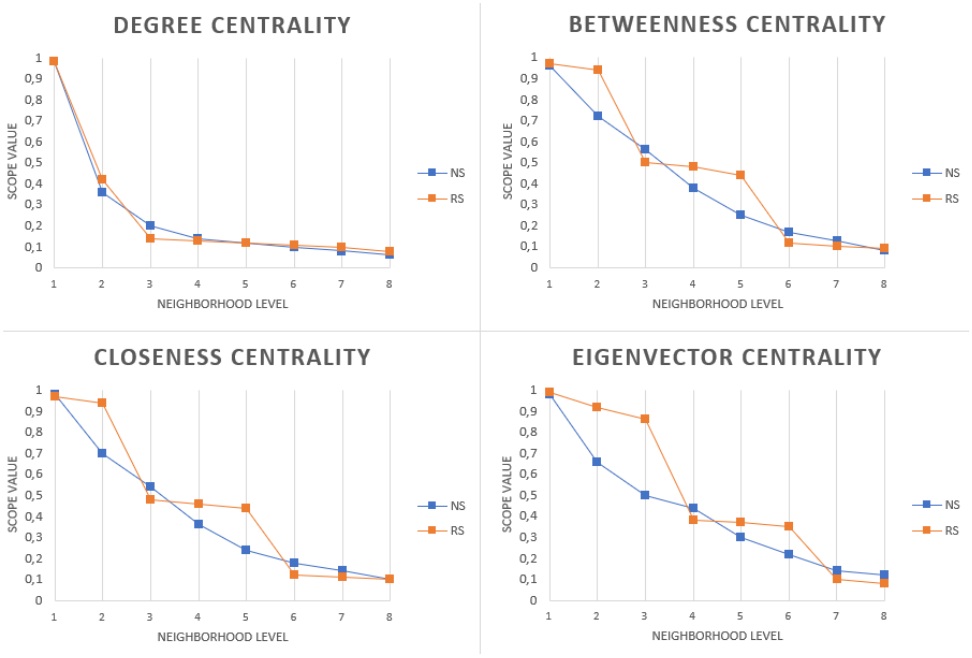


Fig. 5.10: Relationship between NS and RS, on the one side, and centrality measures, on the other side

From the analysis of this figure, we can draw very interesting considerations. Preliminarily, we observe that this experiment confirms the results of the previous one on the fact that RS is more capable than NS in distinguishing neighborhoods with high, intermediate and low values of scope. We can also observe that:

- The nodes with a high degree centrality present a very high value of scope in their closest neighborhoods, i.e., when $t = 1$. Already for $t = 2$ we observe a steep decay of scope. This parameter becomes very low for $t = 3$ and further decreases for $t \geq 4$. This trend can be explained by considering that degree centrality privileges nodes with a high number of outgoing arcs, which, thanks to this property, can easily have a high impact on their immediate neighbors. However, it is not guaranteed that the neighbors of the nodes with a high degree centrality have, in their turn, a high degree centrality. Rather, this does not generally happen because degree centrality follows a power law distribution, which implies that most of the nodes in the network have a low value of this parameter. As a consequence, already for $t = 2$, the value of scope rapidly decreases.

- The nodes with high values of closeness and/or betweenness centrality present high values of scope for $t = 1$. When t increases, the scope decays; however, this happens smoothly. This trend can be explained by considering that closeness and betweenness centralities privilege nodes that are, on average, close to the other ones or that are crucial to reach some other ones. In the past, it was shown [647] that these nodes rarely present a high outdegree; instead, most of them have an intermediate outdegree but, on the other side, they can reach a lot of nodes in few steps. As a further confirmation of the correctness of this result, we observe that, in the literature, it was found that, with these two centrality measures, the distribution of nodes tends to be gaussian, differently from what happens for degree centrality, whose distribution follows a power law.
- The nodes with a high eigenvector centrality present high values of scope for $t = 1$ and $t = 2$. These values become quite high for $t = 3$ and intermediate for $t = 4$. Afterwards, they rapidly decrease for $t \geq 5$. This trend can be explained by considering that nodes with a high value of eigenvector centrality are generally characterized by a high value of outdegree and are linked to other nodes that, in their turn, generally have the same characteristics. This feature allows them to have a high scope on the immediate neighborhoods (and this property is similar to the one characterizing the nodes with a high degree centrality). Furthermore, since also the nodes present therein have a high eigenvector centrality (and, therefore, a high outdegree), the impact of the original nodes can easily be preserved also in the neighbors of the neighbors, and so forth, for some steps. Clearly, when $t \geq 4$, this impact inevitably decreases, and this fact is intrinsic to the very concept of network.

5.3.4.4 Analysis of the approximation and the computation time of the Naive Scope w.r.t. the Refined Scope

This experiment aimed at evaluating the strengths and the weaknesses of NS and RS and at determining in which situations one should be preferred to the other. Actually, NS and RS are complementary because the strengths of the former represent the weaknesses of the latter, and vice versa. In particular, quickness is the main strength of NS, whereas accuracy is the main strength of RS.

The trends of NS and RS against the variation of the neighborhood level t in several circumstances have been reported in Figures 5.6 - 5.8. Starting from them, if we consider correct the values of RS, we can compute the approximation degree of NS w.r.t. RS by means of the formula:

$$\alpha_{j_k}^t = RS_{j_k}^t - NS_{j_k}^t$$

We computed the values of α_{jk}^{t} for all the circumstances considered in Figure 5.6 - 5.8. The corresponding results are reported in Figures 5.11 - 5.13.

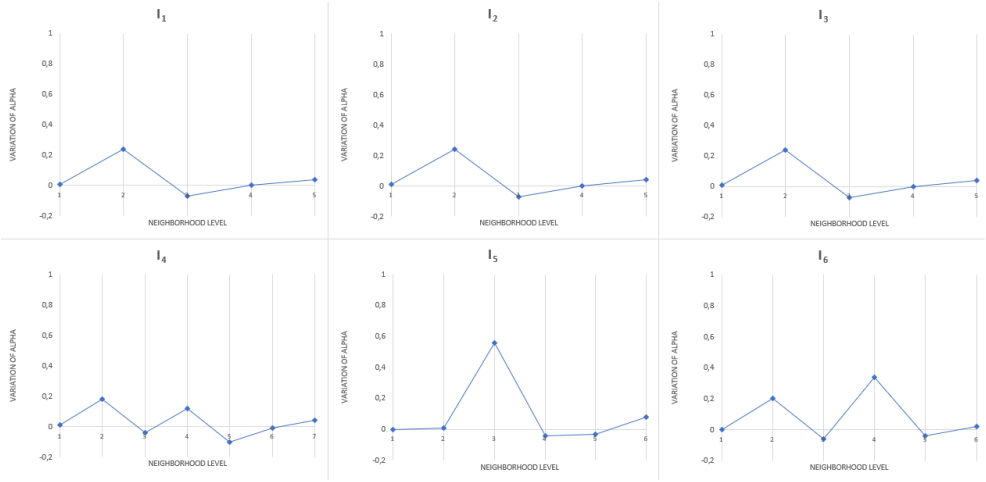


Fig. 5.11: Variation of α_{jk}^{t} for each IoT of the MIoT against the value of the neighborhood level

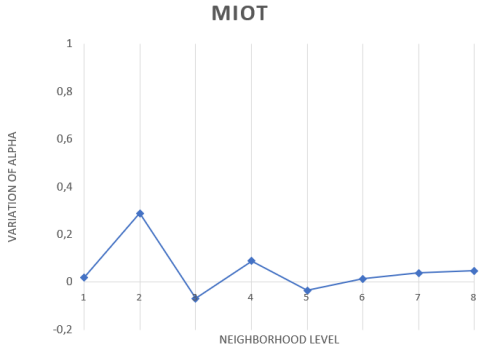


Fig. 5.12: Variation of α_{jk}^{t} for the whole MIoT against the value of the neighborhood level

From the analysis of these figures, we can observe that, for the neighborhoods in which scope is stably low, the value of α_{jk}^{t} is minimal. By contrast, when the values of scope are not stable (this, generally, happens for intermediate values and, in some cases, for high values of both the scope and the neighborhood level), the value of α_{jk}^{t} could become significant. These figures represent a further confirmation of the main feature characterizing RS and not present in NS, i.e., the capability of clearly distinguishing the neighborhoods with a high level of scope from the ones where the value of this parameter is low.

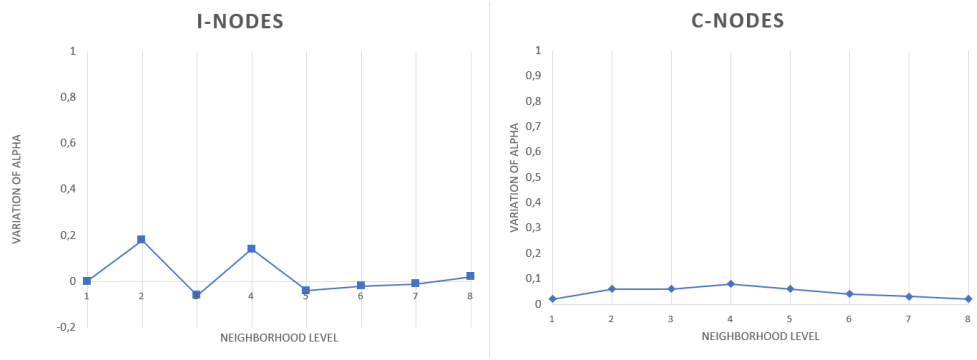


Fig. 5.13: Variation of α_{jk}^{t} for the i-nodes and the c-nodes of the MIoT against the value of the neighborhood level

Afterwards, we determined the computation time necessary to evaluate the average values of NS and RS on the whole MIoT (which, we recall, consists of 1256 nodes). The results obtained are reported in Table 5.23.

Parameter	Average computation time (s)							
	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8
NS	22	89	213	364	512	657	788	927
RS	45	124	246	420	670	884	993	1221

Table 5.23: Computation time (in seconds) necessary to evaluate the average values of NS and RS on the reference MIoT

This table evidences that the time necessary for computing RS is higher than the one required to compute NS. Furthermore, the difference between the two times increases when t increases and becomes more evident for $t \geq 6$. If we combine this result with the previous ones concerning the approximation of NS w.r.t. RS (Figures 5.11 - 5.13) and the values of NS and RS against t (Figures 5.6 - 5.9) we can define important guidelines on how to proceed for scope computation. In particular, when t has low or intermediate values (i.e., $t < 6$), it is better to adopt RS because it is more accurate and the time necessary for its computation is acceptable. Vice versa, when t has high values (i.e., $t \geq 6$) it is better to adopt NS because its computation is much less expensive and both the involved values and the corresponding approximations are negligible.

Actually, a complete and satisfactory analysis of the computation time can be performed only if we consider MIoT with different numbers of nodes. For this reason, we repeated the task described above for six different MIoT having 176, 301, 485, 778, 1256 and 2028 nodes, respectively. The results obtained are reported in Figure 5.14.

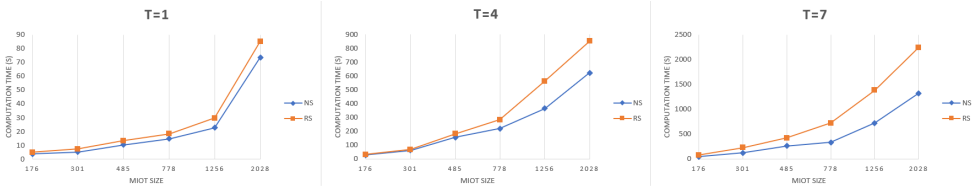


Fig. 5.14: Variation of the average computation time against the size of the MIoT

This figure fully confirms our previous conclusions. As a matter of fact, when $t = 1$, the differences between the computation time of NS and RS are negligible for MIoT with less than 1000 nodes, and very small in the other cases. When $t = 4$, these differences are very small for MIoT with less than 400 nodes; they are intermediate for MIoT with a number of nodes between 400 and 1000; finally, they become high for MIoT with more than 1000 nodes. When $t = 7$ the differences between the computation time of NS and RS are always significant, as we could have expected.

5.3.4.5 Relationship between scope and density

In this experiment, we aimed at investigating the relationship possibly existing between the scope and the average density of a MIoT. Here, we consider the average density of a MIoT as the weighted mean of the average densities of the IoT composing it. The weight of each IoT corresponds to the number of its nodes. We recall that, given an IoT $\mathcal{I}_k$, represented by means of a graph $G_k = \langle N_k, A_k \rangle$, the corresponding density δ_k is defined as:

$$\delta_k = \frac{|A_k|}{|N_k| \cdot (|N_k| - 1)}$$

In order to perform our investigations, we considered our reference MIoT and computed the corresponding density. Then, we decreased its value of 5%, 10%, 15%, 20%, 25% and 30%. We performed this task by randomly removing some previously existing arcs. For each of the six configurations thus obtained, we computed the corresponding values of NS and RS, averaged on the whole MIoT, for $t = 1$, $t = 3$ and $t = 6$. After this, we increased the original density of the MIoT of 5%, 10%, 15%, 20%, 25% and 30%. To obtain these new configurations, we randomly added new arcs to the original MIoT, along with a suitable set of transactions performed on them. Again, for each of these configurations, we computed the values of NS and RS, averaged on the whole MIoT, for $t = 1$, $t = 3$ and $t = 6$. In Figure 5.15, we report the results obtained.

From the analysis of this figure, we can observe that the correlation between density and scope is evident, at least in several cases. In particular, when density increases, scope increases too; instead, a decrease of density implies a decrease of

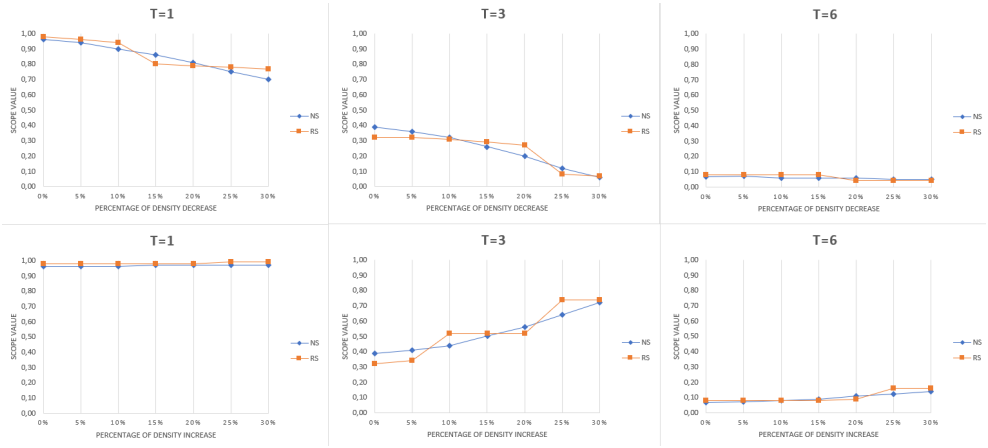


Fig. 5.15: Variation of the values of NS and RS of a MIoT against the variation of the corresponding density

scope. The correlation degree between density and scope depends on the value of t. Indeed, when t is low or t is high, the impact of density on scope is low. By contrast, when t has an intermediate value, this impact is high. These trends can be explained by considering the information diffusion theory in Social Network Analysis. In fact, the intermediate values of t correspond to those scenarios in which the critical mass has been reached and structural holes started to transform into closed triads [647].

5.3.4.6 Comparing scope with related concepts and other approaches

In this section, we compare our scope parameter and our approach to its computation with related concepts and approaches described in Section 5.3.2. As said in that section, to the best of our knowledge, the concept of scope has never been investigated in IoT. Therefore, an experimental comparison is only possible with other approaches working on IoT and proposing parameters related to scope, although different from it.

Proceeding in this way, we decided to compare the scope in a MIoT with: *(i)* the diffusion degree returned by the SIR model and used to test the approach of [457]; *(ii)* the influence degree introduced in Social Network Analysis [160] and, then, extended to the SIoT scenario [313]. Both these parameters are well known in past literature and have been adopted to investigate a large variety of phenomena belonging to very heterogeneous fields.

Comparing scope with diffusion degree

In this section, we compare the scope in a MIoT with the diffusion degree returned by the SIR model used to test the approach of [457].

Susceptible-Infected-Removed (SIR) is a well known model used to test spreading behaviors in several contexts. It describes the spreading of an infectious disease in a population of individuals. Originally proposed by Kermack & McKendrick [376], this model assumes that the population consists of three classes of individuals, namely Susceptible (S), Infective (I) and Recovered (R). The three variables S, I and R represent the number of individuals in each class. S is the number of individuals recovered, who are not infected but could become infected in the future; I is the number of individuals affected by the disease and capable of transmitting it to susceptible individuals; R is the number of individuals recovered, who cannot become infected again. The SIR model is defined by a set of differential equations and is governed by two parameters, namely β and γ, representing the infection rate and the recovery rate, respectively. At each time step, the infection rate β denotes the probability that infected nodes infect their susceptible neighbors; the recovery rate γ indicates the probability that infected nodes recover from the infection.

In our comparison, we are interested in the infection degree that can be derived from the model as the fraction of individuals who are currently infected.

We point out that the SIR model is used to investigate not only infections, but also several phenomena, such as information diffusion and spreading [483, 691, 457, 672], news and rumor modelling in social networks [358], attacks towards wireless networks [446], and so forth. It is exactly these types of phenomena (in particular, information diffusion and spreading) that is relevant in our experiments. Therefore, in the following, we will speak about *diffusion degree* to indicate the infection degree modeled by SIR when this model is applied to information diffusion in an IoT context. In particular, it indicates the fraction of smart objects reached by a given information sent by a node through a chain of transactions (see below).

Clearly, in order to be able to compare diffusion degree with scope, it is necessary to plan the experiment so that the two parameters are comparable.

For this purpose, we have considered the six IoT $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_6\}$ used in the experiment described in Section 5.3.4.2, because we want to take the variation of RS against the neighborhood level as the reference measure for scope evaluation.

Given an IoT $\mathcal{I}_k$ and a node n_{i_k}, we focused on computing the variation of the diffusion degree against the level t of the neighborhood $out\_nbh^t_{j_k}$ of n_{j_k}. Specifically, the diffusion degree of $out\_nbh^t_{j_k}$ at a certain time instant is equal to the fraction of its nodes reached by a certain information sent by n_{j_k} through a chain of transactions starting from it and reaching the nodes of that neighborhood. Recall that, in the SIR model, an infected node can heal, in which case it can no longer transmit the infection. From the information diffusion viewpoint, this scenario is equivalent to

the one of a node reached by a certain information that it no longer wants to transmit to its neighbors.

For the computation of the diffusion degree against t, we decided to operate as follows. First, we had to set the parameters of the SIR model. For this purpose, according to [457, 174], we set β to the so called epidemic threshold $1/\lambda_k$, where λ_k is the largest eigenvalue of the adjacency matrix of the IoT $\mathcal{I}_k$. Thus, we have a different value of β for each IoT. As for γ, following the guidelines in [457], we set it to 0.8.

Analogously to SIR, our model for the computation of the diffusion degree assumes that, at each time instant, a node can infect only its direct neighbors. As a consequence, at the first time instant ($\tau = 1$), a node n_{j_k} can infect only the nodes belonging to $out\_nbh^1_{j_k}$. At the second time instant ($\tau = 2$), n_{j_k} continues to infect other nodes of $out\_nbh^1_{j_k}$. In addition, the nodes of $out\_nbh^1_{j_k}$ can, in turn, infect their direct neighbors. As a result, at the time instant $\tau = 2$, the infection can reach the nodes belonging to $out\_nbh^2_{j_k}$.

At the third time instant, n_{j_k} continues to infect other nodes of $out\_nbh^1_{j_k}$ that have not been infected previously. The nodes of $out\_nbh^1_{j_k}$ infected at time $\tau = 1$ and not yet healed, may continue to infect other nodes of $out\_nbh^2_{j_k}$. At the same time, the nodes of $out\_nbh^2_{j_k}$ already infected at the time instant $\tau = 2$ may, in turn, begin to infect their direct neighbors, i.e., the nodes of $out\_nbh^3_{j_k}$. Usually, at the time instant $\tau = h$, an infected node n_{j_k} can spread its infection until to the nodes of $out\_nbh^h_{j_k}$. The infection process continues with the above rules but, as time goes by, many infected people heal and can no longer be infected.

In order to compare scope with diffusion degree, since the latter is dependent on the time instant τ considered, we decided to make our comparison with reference to a time instant τ_m equal to the maximum level of neighborhood associated with $\mathcal{I}_k$ in Figure 5.6 (and, therefore, $\tau_m = 5$ for $\mathcal{I}_1$ and $\mathcal{I}_5$, $\tau_m = 6$ for $\mathcal{I}_3$ and $\mathcal{I}_6$, $\tau_m = 7$ for $\mathcal{I}_4$ and $\tau_m = 8$ for $\mathcal{I}_2$). Moreover, given the neighborhood $out\_nbh^h_{j_k}$, $1 \leq h \leq \tau_m$, the diffusion degree of n_{j_k} for that neighborhood at the time instant τ_m will be equal to the fraction of its nodes reached by the information initially sent by n_{j_k}.

What we have described so far applies to the computation of the diffusion degree of a single node. We performed this task for all the nodes of the network and, then, averaged the corresponding values. After this, we compared the average value thus obtained with the one of RS shown in Figure 5.6.

The results obtained for the six IoT are shown in Figure 5.16 and the one regarding the whole MIoT are represented in Figure 5.17.

From the analysis of these figures we can observe that the trends of RS and DD are similar because both decrease as the neighborhood level grows. However, there are some differences in the way the decrease of the two parameters happens. In fact, DD

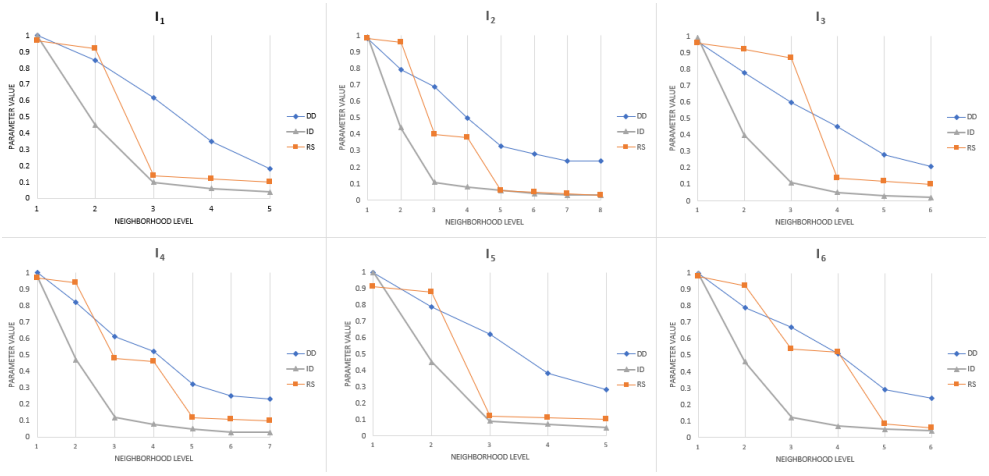


Fig. 5.16: Variation of the average values of the Diffusion Degree DD, Refined Scope RS and Influence Degree ID for each IoT of the MIoT against the neighborhood level

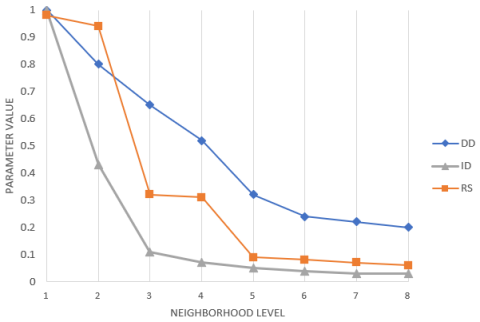


Fig. 5.17: Variation of the average values of the Diffusion Degree DD, Refined Scope RS and Influence Degree ID for the whole MIoT against the neighborhood level

decreases much more slowly than RS and its decrease is quite regular. Instead, RS decreases more quickly and its decrease has a rather irregular characteristic shape, with some steps when passing from one level to another (look, for instance, at the step present when passing from level 2 to level 3 in $\mathcal{I}_1$, or the steps present when passing from level 2 to level 3 and from level 4 to level 5 in the MIoT). In Section 5.3.4.2, we have seen that this trend is characteristic both of RS and NS and that it is to be considered a positive property of scope because it is able to clearly distinguish the neighborhoods in which a node exerts a "power" from those in which such a "power" is lacking.

In this section, we want to go one step further and try to understand the reasons for this trend and, ultimately, for this important property of RS. In Section 5.3.3.1, we have seen that each node of an IoT corresponds to a smart object. In Section 5.3.3.2, we have seen that the scope of a node depends on the number of transaction requests received by the smart object corresponding to that node, its ability to

stimulate not very proactive objects to repost its transactions or to activate transactions with it and, finally, its ability to stimulate smart objects with a high scope to repost its transactions. Ultimately, the scope of a node models its "power" on the other nodes of the network.

Both the experience with Online Social Networks and the theory related to Social Network Analysis reveal us that the "power" exerted by a node remains strong as long as we move towards its neighbors or the neighbors of its neighbors. As we move further away from the node, the possibility of finding a node on which the original node keeps its "power" intact decreases.

Now, the values of RS of a node for the different neighborhood levels in Figures 5.16 and 5.17 are average values obtained by considering all the nodes of the neighborhood. When we move from a neighborhood level to the next, the number of nodes at the new level increases, and this increase can also be significant if the network is very connected. If the "power" of the original node remains intact on all the elements of the new neighborhood, the average value of RS does not change significantly.

But if (as it happens from level 3 onwards) there is a significant decrease of the number of nodes on which the "power" of the original node remains intact, together with a significant increase of the number of nodes on which the "power" is considerably reduced, we have that the huge increase in the denominator of the average is no longer counterbalanced by an equal increase in the numerator. As a consequence of this fact, there is a collapse of the overall value, and therefore of the value of RS, in correspondence with the level of the new neighborhood.

This collapse leads to the characteristic stepped shape that can be observed on the trend of the scope against the neighborhood level almost always and, as far as it is concerned here, in Figures 5.16 and 5.17.

Comparing scope with influence degree

In this section, we compare our scope parameter with influence degree, which was initially proposed in Social Network Analysis [160] and later extended to Social IoT [313].

The influence degree of a node in a social network is an indicator of how much the information it sends to its neighbors appears so interesting that they in turn forward it to their neighbors. This definition of influence degree is based on the information delivered; however, it is possible to think of similar definitions taking into account services provided or other phenomena originating from the node whose influence degree is to be measured [647]. The most immediate way to extend the concept of influence degree of a node n_{jk} to our MIoT scenario is to consider the fraction of the

transactions activated by n_{j_k} that are, in turn, reposted by the nodes belonging to its neighborhoods.

To carry out this experiment, we started from the six IoT $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_6\}$ considered in all our experiments and, once again, we decided to take the variation of RS against the neighborhood level as the reference parameter for scope.

Given an IoT $\mathcal{I}_k$ and a node n_{j_k}, we focused on the variation of the influence degree against the neighborhood level. According to what stated above, influence degree was measured considering the fraction of the transactions activated by n_{j_k} and reposted by at least one node of the neighborhood level into consideration. We observe that the trend of the influence degree is anti-monotonous because the number of transactions originally sent by n_{j_k} reposted by the nodes of $out\_nbh^t_{j_k}$ can only be less than or equal to the corresponding number of transactions reposted by $out\_nbh^{t-1}_{j_k}$.

As in the experiment described in Section 5.3.4.6, also in this case we first computed the influence degree of each node n_{j_k} of an IoT $\mathcal{I}_k$ and, then, we averaged the values thus obtained. Finally, we compared the average influence degree with the average value of RS shown in Figure 5.6.

The results obtained for the six IoT are reported in Figure 5.16, while the results for the whole MIoT are presented in Figure 5.17. From the analysis of these figures, we can see that the trend of the scope and the one of the influence degree are similar because both these parameters decrease with the increase of the neighborhood level. However, we can observe differences in the way they decrease. In fact, the decrease of influence degree is steeper and more regular than the one of scope.

Considerations about the comparisons

To better understand the results of the comparison between Refined Scope RS, Influence Degree ID and Diffusion Degree DD, we must first keep in mind what is the goal of scope. Actually, this parameter was introduced to measure the "power" of a node versus the other nodes of its IoT or versus the nodes of the MIoT. Therefore, the ability of the scope to be a valid parameter for measuring the "power" of a node is closely related to its ability to correctly model what happens in real social networks about this phenomenon, also taking into account the results of past research on Social Network Analysis.

As we have seen in Section 5.3.4.6, Online Social Networks assume that generally the "power" of a person joining them extends to the neighbors of the neighbors and, thus, to the neighborhood of level 2. Moving from the neighborhood of level 2 to the neighborhood of level 3 there is a first significant decrease of this "power". This

decrease becomes very quick in subsequent neighborhood levels, until the "power" becomes almost null from the neighborhoods of level 4 or 5 onwards.

These assumptions made in Online Social Networks are confirmed by research on Social Network Analysis, in particular by the theory of six degrees of separation and the one of the Dumbar Pyramid [647]. The former tells us that, given two people totally unknown to each other and that, perhaps, are at the antipodes of our planet, there are at most six relationships of friendship to separate them. All this is confirmed by the theory of the Dumbar Pyramid, which sets the number of intimate contacts (i.e., friends or relative) of a person at about 20, and the maximum number of (even loose) contacts that a person can handle at about 150.

The above reasoning shows that the ideal parameter for measuring the "power" of a node in an IoT or a MIoT should have a high value for the neighbors of level 1 and 2, an intermediate value for the neighbors of level 3 and, possibly, for those of level 4; finally, it should have low values for the neighbors of level 5 onwards. Instead, a too optimistic parameter, which assumes significant values even for neighbors of level 4 or higher, is not a good indicator of the "power" of a node in a network. On the other hand, a too pessimistic parameter, which assumes low values even for the neighbors of levels 2 and 3, is not adequate for the opposite reasons.

Now, if we consider Figures 5.16 and 5.17, we can observe that Diffusion Degree DD is too optimistic while Influence Degree ID is too pessimistic. Although for opposite reasons, both of them are not accurate in modeling the trend of the "power" of a node in an IoT or a MIoT.

Conversely, the same figures show that RS has an intermediate behavior between DD and ID assuming high values for the neighbors of level 1 and 2, intermediate values for the neighbors of level 3 and very low values for the neighbors of level 5 onwards. This trend is totally in line with the behavior that both the Online Social Networks and the research on Social Network Analysis assume should characterize the "power" of a node in a network. This allows us to conclude that RS is actually the best parameter to model this phenomenon.

5.3.5 Use cases

In a scenario characterized by the pervasive diffusion of increasingly intelligent and social objects, our approach for the computation of scope can have a large variety of applications. To give an idea of real use cases that can benefit from our approach, in the next subsections, we examine two of them.

5.3.5.1 Scope in a MIoT for smart cities

As a first example case, consider some public areas (such as parks, squares, shopping centers, etc.) in a (smart) city, and assume that a group of people actively visits these areas. Each area is equipped with several smart objects for monitoring weather, air quality, traffic conditions, level of noise, etc., along with several actuators, such as smart lamps or information hubs provided as online services. Each person may have several smart devices, such as smartwatches, smartphones, other wearable devices, and so forth. People and places can interact with each other through their smart objects [157].

Such a scenario can be modeled through a MIoT $\mathcal{M}$ consisting of a set $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$ of IoT, each representing a public area. The set of the objects of $\mathcal{M}$ comprises the smart objects in the public areas and the set of personal devices of people visiting them. If an object o_j of the MIoT is active in the k^{th} public area, it has an instance ι_{j_k} in the IoT $\mathcal{I}_k$. Clearly, when a person with a smart object o_j moves around different public areas corresponding to different IoT, o_j will have different instances, one for each IoT.

Each visitor of an area is generally interested in a certain kind of activity; for instance, she could be a fitness runner. The final goal of the MIoT is supporting people to get the best experience from their activities. In this setting, scope can play a key role in reaching this objective. In the following, we report some possible usage scenarios.

Assume that a person wants to go out for a run. First, she needs to choose the best area for the run, based on weather conditions, traffic and other parameters that she considers relevant. To carry out her choices, she can contact, through her device, the sensors of each public area of her interest, the information hubs and the devices of other trusted runners in order to ask for weather, traffic and other conditions. The choice of the information sources to consult is usually related to the corresponding trustworthiness and the easiness of getting the desired information from it. These two properties are clearly strictly correlated to the scope of the source; indeed, this scope can be seen as a "summary" of these two parameters and some other related ones, such as accuracy, reputation, impact, etc. Once a person has performed her choice, she can decide to send this information to the MIoT in such a way as to serve, in her turn, as information provider for the community.

A similar activity flow may happen in several other circumstances in which there is a decision to make, e.g., when a user must choose the best shopping center where she can buy a given object, the best cinema where she can see a movie, etc.

In all these cases, data regarding the choices of a user can be coupled with those registered during the activities she performed as a consequence of these choices (e.g.,

data coming from personal smartwears) in such a way as to confirm the correctness of the choice or, on the contrary, to alert the other users of the evaluation errors. For instance, imagine a scenario in which a person verifies that the weather was actually too cold for the clothes she had selected (interestingly, this information could be automatically detected and sent by the sensors present in her smartwears). In this case, the scope of the smartwears is useful to understand how extended and how strong their capability is of influencing the decision of the other users. In other words, the scope of an object o_j in this scenario determines how many users are impacted by the data sent by it and how much strong this impact is.

It is worth pointing out the relevance of the scope in this context. As a matter of fact, the knowledge of the objects with the highest impacts in the MIoT allows the improvement of the efficiency and the effectiveness of the information disseminated through the network. At a higher abstraction level, some smart objects of the MIoT could assume the role of reliable information hubs for the whole MIoT if their scope is particularly high and extended.

Recall that scope depends not only on the Impact Degree but also on Trust Degree and Security Level. According to the definitions of Naive and Refined Scope in Section 5.3.3.2, a high value of scope (which is a condition for being an information hub) can be obtained only if all the three parameters defining the scope (i.e., impact, trust and security) are high.

Scope may also have an important role in the detection and the management of possible anomalies characterizing one or more devices in the network. As an example, assume that a weather sensor in a public area is malfunctioning; in this case, all the objects relying on its data will be affected by this anomaly. Knowing the scope of an object may help in the detection and management of its possible anomalies. For instance, in the previous case, if one or more other trustworthy weather devices are present in the same area, they could help the whole MIoT to determine the sensor malfunction, to avoid the propagation of its effects and, finally, to repair the anomalous device.

5.3.5.2 Scope in a MIoT for shopping centers

Another possible scenario where scope plays an important role is a big shopping center consisting of several buildings, each dedicated to specific product typologies, such as food, clothing, do-it-yourself, electronic devices, and so on. In this context, smart devices can be modeled by a MIoT $\mathcal{M}$ consisting of m IoT, one for each building. The set of the objects of $\mathcal{M}$ consists of the set of the intelligent sensors present in each building (including video surveillance, temperature sensors, fire sensors, pres-

ence sensors, etc.) and the set of personal devices of visitors (including smartphones, tablets, smartwatches, etc.).

Each object o_j that interacts with the ones of the k^{th} building has an instance ι_{j_k} representing it in $\mathcal{I}_k$. Clearly, when the owner of an object o_j, such as a smartphone, moves throughout the buildings of the shopping centers, o_j will have different instances associated with the different buildings of the center.

Here, an intelligent system of the shopping center could push offers to the enabled customer devices based on proximity, past preferences, habits, and so on. Analogously, a personal device can suggest its owner the most comfortable and promising places to visit during her stay in the shopping center based on the knowledge provided by the smart objects and the sensors dispersed in the shopping center.

In this scenario, each person connected to the MIoT is interested in a certain kind of activity, somehow related to shopping. Indeed, users can play several roles ranging from vendors, suppliers or customers. In this context, an innovative role is the one of the personal shopper, i.e., a person, who helps customers by giving them alerts or making them suggestions. Personal shoppers are often employed directly by stores and boutiques, but the number of freelancers or online personal shoppers is constantly growing.

While a customer visits the building of a shopping center, her device may constantly locate the nearest devices and query for interesting products or offers. In the meantime, it could query other objects of the customer (for instance, wearable devices) to measure her vital parameters in order to evaluate her pleasure in checking the products of a shopper. This can represent feedback information that the device supplies to the MIoT. Furthermore, it can act as a personal shopper. Indeed, it interacts with the other objects of the MIoT, considers the offers of the shops, elaborates this information through machine learning algorithms, makes some proposals to its customer, registers her feedbacks and transmits them to the other devices in such a way as to improve the quality of its recommendations.

Assume, now, that a customer wants to go out for shopping. First, she needs to locate the best building to start with. This activity can be carried out by contacting the preferred personal shopper or by checking the preferred destinations of "special" customers (for instance, the most influential ones) or, again, by detecting the most comfortable shops. All these activities can be done by her personal device that can contact the other ones of the MIoT for acquiring all necessary data. Once the desired knowledge has been obtained, the device can process it to make its suggestions. Clearly, once the customer has made her choices and has performed her shopping activities, she can share information about her experience. In this way, she and/or her devices can become information providers for other customers. Scope plays an im-

portant role in this scenario. Indeed, the scope of each smart object determines how many devices (and, ultimately, people) it can influence and how strong its influence is.

Again, this depends on its Impact Degree, its Trust Degree and its Security Level. The higher each of these parameters, the higher the corresponding scope and, consequently, the stronger its influence.

As in the previous scenario, an important issue to investigate and address is the presence of possible anomalies. The impact of an anomaly depends on several factors; the scope of the affected objects is certainly one of the most important. As an example, given an anomaly of the device acting as a personal shopper, for instance the loss of historical data on product prices, the corresponding suggestions might not be the most convenient ones for its owner. In this case, the anomaly will certainly have a high impact on the device's owner. Furthermore, it can have an impact, even if smaller, on all the other objects (and, ultimately, on the corresponding customers) that it can reach and influence. The extension and the strength of the impact of an object o_j on an object o_q depends on the value of the scope of o_j on o_q.

Assume, now, that an anomaly affects the system for the temperature detection of a building or, even, of the whole shopping center. Clearly, the scope of this system is much larger and stronger than the one of a personal device. Indeed, this anomaly impacts on all the customers present in the building or, even, in the shopping center because, due to it, the air conditioning system will determine an uncomfortable situation for all the people present therein. This last example allows us to draw a further conclusion, i.e., knowing the scope of the devices of a MIoT is also relevant to properly prioritize anomaly management.

6

Reliability

In the past research, trust and reputation have been investigated for communities of people, for organizations and multi-agent systems. As we have seen in Chapter 5, the thing in a MIoT has a profile that defines its behavior over time. If a thing can have a profile and a behavior like that, it is not out of place to extend the concept of trust and reputation to things and define ad-hoc approaches for their computation. In this chapter, we investigate trust and reputation of a thing in a MIoT scenario and propose a context-aware approach to evaluate them. We also report a running example in order to further explain our approach.

The material present in this chapter is taken from [649, 650].

6.1 Introduction

We experience the concepts of trust and reputation every day; for example, when we buy something on an online service provider, in most cases, we do not have enough information about the service and/or the provider. This forces us to accept a "risk of prior performance", like paying for services and goods before receiving them. This information asymmetry can be mitigated thanks to the concepts of trust and reputation. The term *trust* is reported in literature with different nuances; therefore, it could be difficult to understand what it really is. However, as stated in [359], there are mainly two ways of defining trust. The first one is called *reliability trust*. This type of trust can be defined as the subjective probability by which an individual A expects that another individual B performs a given action from which its welfare depends [285]. The second type of trust is called *decision trust*. In this case, there is one party that is willing to depend on something or somebody in a given situation with a feeling of relative security [467]. Both these definitions involve the notions of *dependence* and *reliability* on the trusted entity and a certain *risk* related to a misbehavior of service provider. Starting from the concept of trustworthiness, it is possible to define the one of *reputation*. According to the Concise Oxford Dictionary [4], reputa-

tion is “the beliefs or opinions that are generally held about someone or something”. Therefore, the concept of trust is based on personal and subjective events described through factors and evidences. Instead, reputation can be considered as a collective measure of trustworthiness, based on advices or ratings from members of a community.

In the past computer science research, the concepts of trust and reputation have been investigated for communities of people, for organizations, for wireless sensor networks, for vehicular ad-hoc networks, and for multi-agent systems, and a lot of relevant results have been obtained [359, 674, 574, 65, 289, 228, 226, 48].

In the last few years, things are becoming increasingly important in the Internet scenario [70, 81, 82, 32, 33, 34, 546, 61, 146, 560, 291] and, presumably, in the future, the number of objects connected to the Internet will be much higher than the corresponding number of people. As a matter of fact, the term “Internet of Things” is becoming more and more common and, based on it, increasingly complex architectures [275, 274], requiring things to show a smart and social behavior [587, 272], are continuously proposed in literature. Social Internet of Things (hereafter, SIoT [70]), Multiple IoT Environment (hereafter, MIE [81]) and Multi Internet of Things (hereafter, MIoT [82]) are only three of the latest architectures with these characteristics.

This chapter aims at providing a contribution in this setting. In the MIoT model, things are organized in networks called IoTs. A thing can belong to one or more IoTs. Things belonging to more IoTs behave as “bridges" and allow communication and interaction between different IoTs of the MIoT. Things interact with each other through suitable transactions. The analysis of the information content exchanged by a thing with the other ones of the MIoT allows the construction of the thing profile. The profile of a thing can be further enriched by considering the profiles of the things directly connected to it, according to the homophily principle characterizing social networks [468].

These are the same considerations that underlie the profiles of humans. As a consequence, most of the ideas and results about trust and reputation of a human in a community or of an agent in a multi-agent system can be extended and, possibly, redefined for a thing in a MIoT. Clearly, this extension is not immediate because it must consider all the peculiarities of a thing w.r.t. a human or an agent, and the specificities of a MIoT w.r.t. a community of people or a multi-agent system.

Investigating trust and reputation of things in a social context is extremely beneficial. Indeed, it has a lot of applications. Think, for instance, of the detection and isolation of a malicious object, the support of thing cooperation, the detection and the manipulation of thing reliability parameters, the evaluation of quality of ser-

vices, just to cite a few of them. The presence of a thing profile allows us to define context-aware notions of trust and reputation, along with suitable approaches for their computation. These notions are well suited to capture and address the complexity of the scenario we are investigating.

This chapter is organized as follows: in Section 6.2, we provide an overview of related literature. In Section 6.3, we explain the introduced novelties on the MIoT paradigm for modeling our scenario, define the concept of thing profile, and describe the proposed approach for computing trust and reputation. In Section 6.4, we illustrate the experimental campaign that we conducted to test it. In Section 6.5, we present two possible use cases of our approach. Finally, in Section 6.6, we discuss the implications and the possible exploitation of the results obtained through our experiments.

6.2 Related Literature

In computer science research, there is a plenty of papers addressing trust and reputation. Each of them proposes a model to handle these concepts from different points of view. However, as in most cases, the efficiency and the effectiveness of each model depend on the environment where it works.

As reported in [359], there are some features that allow the cataloguing of general trust and reputation models proposed in literature. These are: *(i)* trust classes, *(ii)* categories of trust semantics, *(iii)* reputation network architectures, and *(iv)* reputation compute engines. As for this last issue, there are several families of approaches to compute trust and reputation. For instance, some possible families could be based on: *(i)* the sum or the average of ratings, *(ii)* fuzzy operators, *(iii)* "flow" models computing trust and reputation scores through looped or long chains. For example, Google's PageRank [523] belongs to this last family.

Before deeping on trust and reputation in the IoT scenario, it is necessary to spend some words on the computation of these measures in an online service provisioning [693, 564], which is the first Internet context where these concepts have been applied. One of the most famous reputation systems is the eBay's one [564]. In this system, after each purchase, the buyer and the seller have the opportunity to rate each other as positive, negative or neutral. The architecture is centralized and the central authority computes the reputation score of each participant as the sum of positive and negative ratings. Even if this system is primitive and can be quite misleading, it seems to have a strong positive impact in the marketplace. On the other hand, there are experts' sites where pools of individuals are willing to answer

questions in their areas of expertise. For instance, AskMe is one of these sites; in this system, a participant has to pay a fee to take part to the corresponding network.

Another important contribution in trust management is reported in [171]. In this paper, the authors introduce a framework, called Socialtrust, aiming at analyzing the reliability of information exchanged in online social networks. In order to perform trust computation, Socialtrust considers three factors, namely: *(i)* the trust group feedback, *(ii)* the difference between the user's perceived quality and the trust concept, and *(iii)* the tracking of user behavior. The authors also describe the application of Socialtrust to MySpace profiles.

In [599], the authors propose an approach to find users in a social network, who are able to spread a specific information as far as possible. In particular, a company selects a set of people, who are willing to send advertisements to their friends in order to get discounts or free goods. If a user is highly respected by her friends, her advertisements will be probably spread over the social network.

Finally, another notable reputation system is the PageRank [523]. In this case, the collection of hyperlinks to a given page can be exploited to evaluate the reputation score of that page.

After a description of trust management in online service provisioning, we examine the transpositions of all these concepts to the IoT context. A well-defined reputation system is really relevant in IoT, because it is necessary to manage the services provided by objects. As previously described, the concept of trust encompasses factors like the goodness of a service or the reliability and the availability of an object.

The relevance of trust management in an IoT context is investigated in [694]. Here, the authors show how trust management can favor data fusion and mining, privacy and information security.

In literature, it is possible to catalogue trust and reputation approaches operating in IoT scenarios according to the type of architecture that the authors decided to develop. Based on it, three different kinds of model can be recognized, namely: *(i)* centralized, *(ii)* semi-centralized, and *(iii)* distributed ones.

An example of a centralized model is proposed in [578]. This model can evaluate the context in which objects work. In this architecture, there is a node called Trust manager, which handles all the information related to the trustworthiness of agents in the IoT. The evaluation approach consists of five phases, namely: *(i)* information gathering; *(ii)* entity selection; *(iii)* transaction; *(iv)* reward or punish, and *(v)* learning. Roughly speaking, when an object requires a service, it asks to the trust manager a list of trustable nodes offering that service. Then, after the transaction completion,

it sends a report to the trust manager that contains a positive score (reward) or a negative one (punish) regarding the service provider.

Another example of a centralized system is described in [194], where the authors propose a model called Trusted Resource Sharing (TRS). TRS has three main components, namely Trust, Usage and Relation. The Trust component, which is the one of interest, is developed through a centralized architecture, in which there is an entity managing trust and resource policies. Trust is evaluated for both objects and resources available in the network. However, IoT is expected to exponentially grow in the next future, so a central authority could be a bottleneck. Indeed, a failure on the Trust Manager can block the whole network. Furthermore, the Trust Manager could also be attacked to change the trust and the reputation of each participant.

A further centralized system is described in [645]. Here, the authors propose a trust model called "REK", developed in a SIoT context. In order to evaluate trustworthiness, REK leverages two indicators, namely "Experience" and "Reputation". Both these indicators are modeled using mathematical tools and are extracted from previous interactions among entities in the SIoT environment. The "Experience" component is modeled by means of PageRank [523].

As far as semi-centralized architectures are concerned, [644] developed a model operating on a SIoT. It introduces three new components into the SIoT, namely Trust Agent, Trust Broker and Trust Analysis and Management. The proposed model focuses on the social parameters of IoT. Specifically, the authors examine honesty, cooperativeness, and community-interest. These parameters contribute to the building of a knowledge (reported as Knowledge Trust Management - KTM) useful to evaluate the reputation of an agent. Knowledge grows in a huge way over time. In order to manage this big amount of data, the authors propose a fuzzy-based model, capable of representing attributes in vague terms, like "low" or "high", "bad", "acceptable" or "good". KTM is a good way to build up a consistent reputation system capable of adapting to different situations.

As for distributed architectures, there are several works proposing distributed models to compute trust and reputation. One of the first models belonging to this category is described in [184]. Here, the authors study a community of sensors in a Wireless Sensor Network. This model builds two types of reputation, namely direct reputation, computed through personal observations, and indirect one, based on the recommendations of other nodes. Analogously to the model of [644], the one of [184] is based on fuzzy theory. It represents a starting point for future developments in the computation of trust and reputation among IoT devices. However, it considers only a specific IoT environment, with a limited number of measures. The model proposed in [83] is an extension of the one introduced in [184]. Here, the authors

describe a dynamic protocol, aiming at addressing the limits presented in previous ones. The concept of trust is modeled through three elements: honesty, cooperativeness and community-interest (these are the same elements described in [644]). As it is a distributed model, each node updates the trust ratings of the other nodes by collaborating with them.

Another distributed system is described in [668]. Here, trust and reputation are computed by analyzing each device from three different viewpoints, namely sensor, core and application ones. Each of these layers has its own trust information. In the sensor layer, trust denotes which node must be contacted for a service. Instead, in the core layer, trust is used to select a set of networks and routes, through which data can be sent. Finally, in the application layer, trust is exploited to evaluate which candidate method of data processing or which storage service are trusted. These three trust scores are, then, composed through fuzzy logic.

A further interesting distributed system is described in [186]. Here, the authors develop a technique to compute trust on an IoT based on the Service Oriented Architecture. This technique aims at selecting feedbacks using rating similarities, communities of interest and social contacts. The whole model is based on a collaborative filtering approach. To guarantee scalability, each node saves trust information only about a subset of nodes of interest and performs a minimum computation to update these values.

Finally, [512] introduces two trust and reputation models for the SIoT environment. The former is called *subjective trustworthiness*. In this case, each node computes the trustworthiness of its neighbors based on its experience and the one of them. Trust computation considers five viewpoints, namely: *(i)* direct opinion; *(ii)* indirect opinion; *(iii)* long-term opinion; *(iv)* relationship factor; *(v)* direct opinion in the credibility. The latter is called *objective trustworthiness*. It is defined in a Peer-to-Peer (P2P) scenario, in which information of each node is visible and managed through special nodes, called Pre-Trusted Objects. In this case, trust computation considers four points of view, namely: *(i)* long-term opinion; *(ii)* short-term opinion; *(iii)* relationship factor in credibility; *(iv)* intelligence in credibility. In both models, the credibility of a node is used to evaluate the opinion of other nodes. Therefore, these models give a high weight to recommendations made by “good" friends and a low weight to feedbacks provided by “bad” friends.

6.3 Methods

6.3.1 Extending the MIoT paradigm

In this section, we extend the MIoT paradigm introduced in Chapter 4 in order to make it capable of handling the concepts of trust, reputation and reliability.

Given a MIoT $\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$ and an instance ι_{j_k} in $\mathcal{I}_k$, the neighborhood of ι_{j_k} is defined as:

$$nbh(\iota_{j_k}) = nbh^{out}(\iota_{j_k}) \cup nbh^{in}(\iota_{j_k})$$

where:

$$nbh^{out}(\iota_{j_k}) = \{\iota_{q_k} | (n_{j_k}, n_{q_k}) \in A_I, |tranSet_{jq_k}| > 0\}$$

and

$$nbh^{in}(\iota_{j_k}) = \{\iota_{q_k} | (n_{q_k}, n_{j_k}) \in A_I, |tranSet_{qj_k}| > 0\}$$

In other words, $nbh(\iota_{j_k})$ comprises those instances directly connected to ι_{j_k} through an incoming or an outgoing arc, which shared at least one transaction with ι_{j_k}.

Given a pair of instances ι_{j_k} of o_j and ι_{q_k} of o_q in $\mathcal{I}_k$, the MIoT stores the set $tranSet_{jq_k}$ of the transactions from ι_{j_k} to ι_{q_k}.

The set $tranSet_{jq_k}$ is defined as:

$$tranSet_{jq_k} = \{Tr_{jq_{k_1}}, Tr_{jq_{k_2}}, \cdots, Tr_{jq_{k_v}}\}$$

A transaction $Tr_{jq_{k_t}} \in tranSet_{jq_k}$ is represented as:

$$Tr_{jq_{k_t}} = \langle reason_{jq_{k_t}}, source_{jq_{k_t}}, dest_{jq_{k_t}}, start_{jq_{k_t}}, finish_{jq_{k_t}}, success_{jq_{k_t}}, content_{jq_{k_t}} \rangle$$

Here: *(i)* $reason_{jq_{k_t}}$ denotes the reason why $Tr_{jq_{k_t}}$ occurred, chosen among a set of predefined values; *(ii)* $source_{jq_{k_t}}$ indicates the starting node of the path followed by $Tr_{jq_{k_t}}$; *(iii)* $dest_{jq_{k_t}}$ represents the final node of the path followed by $Tr_{jq_{k_t}}$; *(iv)* $start_{jq_{k_t}}$ denotes the starting timestamp of $Tr_{jq_{k_t}}$; *(v)* $finish_{jq_{k_t}}$ indicates the ending timestamp of $Tr_{jq_{k_t}}$; *(vi)* $success_{jq_{k_t}}$ denotes whether $Tr_{jq_{k_t}}$ was successful or not; it is set to *true* in the affirmative case, to *false* in the negative one, and to `NULL` if $Tr_{jq_{k_t}}$ is still in progress; *(vii)* $content_{jq_{k_t}}$ indicates the content "exchanged" from ι_{j_k} to ι_{q_k} during $Tr_{jq_{k_t}}$.

In its turn, $content_{jq_{k_t}}$ presents the following structure:

$$content_{jq_{k_t}} = \langle format_{jq_{k_t}}, fileName_{jq_{k_t}}, size_{jq_{k_t}}, topics_{jq_{k_t}} \rangle$$

Here: *(i)* $format_{jq_{k_t}}$ indicates the format of the content exchanged during $Tr_{jq_{k_t}}$; the possible values are: "audio", "video", "image" and "text"; *(ii)* $fileName_{jq_{k_t}}$ denotes the name of the transmitted file; *(iii)* $size_{jq_{k_t}}$ indicates the size in bytes of this

content; *(iv)* $topics_{jq_{k_t}}$ denotes the set of the content topics; it consists of a set of keywords representing the subjects exchanged during $Tr_{jq_{k_t}}$. It can be formalized as: $topics_{jq_{k_t}} = \{kw^1_{jq_{k_t}}, kw^2_{jq_{k_t}}, \ldots, kw^w_{jq_{k_t}}\}$.

Now, we can define the set $tranSet_{j_k}$ of the transactions performed by ι_{j_k} in $\mathcal{I}_k$. Specifically, let $Inst_k$ be the set of the instances of $\mathcal{I}_k$. Then:

$$tranSet_{j_k} = \bigcup_{\iota_{q_k} \in Inst_k, \iota_{q_k} \neq \iota_{j_k}} tranSet_{jq_k}$$

In other words, the set $tranSet_{j_k}$ of the transactions performed by an instance ι_{j_k} is given by the union of the sets of the transactions from ι_{j_k} to all the other instances of $\mathcal{I}_k$.

Indeed, Figure 6.1 shows the overall MIoT architecture that we designed for supporting our approach. Each colored circle represents a distinct IoT of the MIoT. The stack composed by “Transaction metadata", “Instance metadata" and “Object metadata" handles the transaction, instance and object metadata described in Chapter 4.

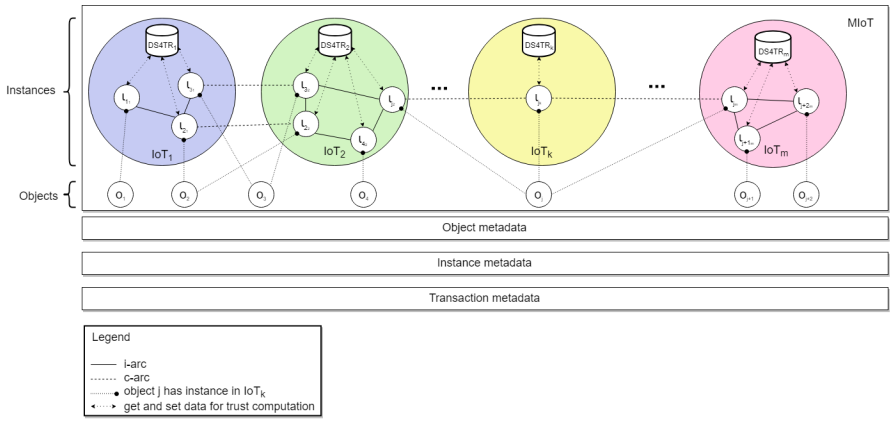


Fig. 6.1: Schematic representation of the proposed MIoT architecture

The aim of this architecture is to provide a scalable way to manage both the storage of support data and the tasks necessary for the computation of trust and reputation. We propose to add a repository, called *DS4TR* (Data Storage for Trust and Reputation) in each IoT, which stores the data necessary for the computation of trust and reputation. From a logical viewpoint, *DS4TR* is separated from the other objects of the corresponding IoT. Actually, from an implementation point of view, *DS4TR* could be either deployed through a cloud service or embedded in one of the objects operating in the IoT. As for security issues involving *DS4TR*, the following reasonings hold: the overall scenario consists of two different cases, one in which *DS4TR* is provided by a cloud service and another in which it is part of an IoT.

In the former case, *DS4TR* represents a trusted-third party. For the sake of space, and since this issue does not present the core of our approach, we do not describe this case in detail. We only refer to relevant techniques to protect trusted-third parties in a cloud environment [631, 728] already proposed in past literature.

In the latter case, which is a scenario typical of MIoT, each object is a peer of a P2P architecture and *DS4TR* is part of an IoT. In order to obtain a stable and reliable IoT, it is fundamental to define a good strategy to build *DS4TR*. For instance, we can select some reliable nodes from the IoT. To assure node reliability, we can exploit our trust and reputation model to compute a ranking of nodes, ordered by their reputation values. Then, we can choose the first τ ones to compose a *DS4TR*. Recall that, as a typical situation in a P2P scenario, each selected object stores only a part of the whole data repository. However, in order to maintain a certain level of fault-tolerance, some parts can be replicated on multiple objects. The parameter τ represents a tradeoff between reliability and performance of the network. The greater τ, the more trustworthy the network, but the lower its speed. Of course, the setting of τ depends on the context in which the overall model is developed. However, once a node is selected to be part of a *DS4TR*, it has to save a portion of the trust and reputation repository of its IoT.

Another interesting aspect to consider is how to check whether the nodes contributing to a *DS4TR* are properly working or not. Each transaction made by a node is part of our model, so that we can compute the trust and reputation values of these objects. To maintain a high level of reliability, we can set a threshold, say th_{rep}, which represents the minimum reputation value that a node must have to be part of a *DS4TR*. If a node of *DS4TR* obtains a reputation value lower than th_{rep}, it leaves the repository and another node is chosen (by following an approach similar to the one presented above) to replace it.

Beside *DS4TR*, cross-nodes play an important role in the computation of trust and reputation. Indeed, they allow every node of an IoT to ask for trust data regarding participants to other IoTs of the MIoT. Finally, after instances completed a transaction, each of them has to add a feedback about the other part. Obviously, this feedback has to be added in the *DS4TR* node(s) corresponding to the transaction participants.

This architecture can face scalability issues because each IoT has its own repository to save data. In this way, the problem of bottlenecks in the network is highly mitigated. A careful reader could point out that requests coming from different IoTs could overwhelm a *DS4TR* node. However, in the intrinsic architecture of a MIoT, there are much less transactions between two different IoTs than within an IoT.

MIoT Running Example

Consider a smart shopping center consisting of three buildings, one for each store. These last are a supermarket, an electronics store and a clothing store (see Figure 6.2). We can associate a MIoT $\mathcal{M}$ with this center. $\mathcal{M}$ consists of three IoTs:

$$\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$$

$\mathcal{I}_1$ (resp., $\mathcal{I}_2, \mathcal{I}_3$) connects all the instances of the smart objects of people accessing the supermarket (resp., electronics store, clothing store).

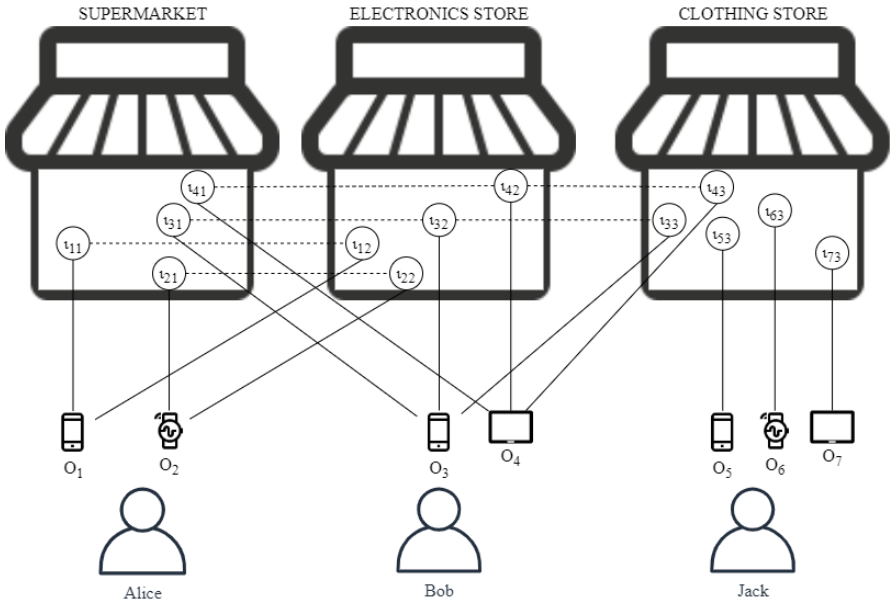


Fig. 6.2: An example of a MIoT associated with a smart shopping center

Consider three customers: *(i)* Alice, who owns a smartphone o_1 and a smartwatch o_2; *(ii)* Bob, who owns a smartphone o_3 and a tablet o_4; *(iii)* Jack, who owns a smartphone o_5, a smartwatch o_6 and a tablet o_7.

Alice visits the supermarket and the electronics store. ι_{1_1} (resp., ι_{1_2}) represents the instance of the Alice's smartphone in $\mathcal{I}_1$ (resp., $\mathcal{I}_2$); instead, ι_{2_1} (resp., ι_{2_2}) denotes the instance of the Alice's smartwatch in $\mathcal{I}_1$ (resp., $\mathcal{I}_2$).

Analogously, Bob visits the supermarket, the electronics store and the clothing store. ι_{3_1}, ι_{3_2} and ι_{3_3} (resp., ι_{4_1}, ι_{4_2} and ι_{4_3}) denote the instances of the smartphone (resp., tablet) of Bob in the three stores.

Finally, Jack visits only the clothing store. He has a smartphone o_5, a smartwatch o_6 and a tablet o_7; ι_{5_3}, ι_{6_3} and ι_{7_3} represent the instances of these smart objects in $\mathcal{I}_3$.

In Figure 6.2, the dashed line between ι_{1_1} and ι_{1_2} indicates that they are two instances of the same object o_1. An analogous semantics regards the dashed lines between ι_{2_1} and ι_{2_2}, ι_{3_1} and ι_{3_2}, ι_{4_1} and ι_{4_2}, ι_{3_2} and ι_{3_3} and, finally, ι_{4_2} and ι_{4_3}.

6.3.2 Definition of a thing profile

In this section, we present our definitions of instance and object profiles. They represent a preliminary knowledge, mandatory to fully understand the rest of our approach. To introduce them, we need to present the following operators:

- $\biguplus$: it receives a set $\{entitySet_1, entitySet_2, \cdots, entitySet_t\}$ of entity sets and performs their union, not eliminating duplicates but reporting the number of their occurrences. Therefore, this operator returns a set of pairs $\{(entity_1, ne_1), (entity_2, ne_2), \cdots, (entity_w, ne_w)\}$ in which the pair $(entity_r, ne_r)$ indicates the r^{th} entity and the number of its occurrences. In counting the number of occurrences, $\biguplus$ takes the presence of synonymies and homonymies into account. These properties can be computed (for terms, images, etc.) by applying the classical approaches proposed in past literature [102, 227].
- $avgFileSize$: it receives a set of files and computes their average size.

We are now able to define the profile $\mathcal{P}_{jq_k}$ of the relationship existing between two instances ι_{j_k} and ι_{q_k}, which performed a set $tranSet_{jq_k} = \{Tr_{jq_{k_1}}, Tr_{jq_{k_2}}, \cdots, Tr_{jq_{k_v}}\}$ of transactions. Specifically:

$$\mathcal{P}_{jq_k} = \langle reasonSet_{jq_k}, sourceSet_{jq_k}, destSet_{jq_k}, avgSzAudio_{jq_k}, avgSzVideo_{jq_k}, avgSzImage_{jq_k}, avgSzText_{jq_k}, successFraction_{jq_k}, topicSet_{jq_k} \rangle$$

where:

- $reasonSet_{jq_k} = \biguplus_{t=1..v}(reason_{jq_{k_t}})$;
- $sourceSet_{jq_k} = \biguplus_{t=1..v}(source_{jq_{k_t}})$;
- $destSet_{jq_k} = \biguplus_{t=1..v}(dest_{jq_{k_t}})$;
- $avgSzAudio_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}audio\text{”}\}$;
- $avgSzVideo_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}video\text{”}\}$;
- $avgSzImage_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}image\text{”}\}$;
- $avgSzText_{jq_k} = AvgFileSize_{t=1..v}\{fileName_{jq_{k_t}} | format_{jq_{k_t}} = \text{“}text\text{”}\}$;
- $successFraction_{jq_k} = \frac{|\{Tr_{jq_{k_t}} | Tr_{jq_{k_t}} \in tranSet_{jq_k}, success_{jq_{k_t}} = true\}|}{v}$;
- $topicSet_{jq_k} = \biguplus_{t=1..v}(topics_{jq_{k_t}})$.

If we introduce the operator $\bigsqcup$, which compactly represents the set of the operations described above, needed to obtain a profile of a pair of instances $\mathcal{P}_{jq_k}$ starting from the corresponding transactions, we can formalize the previous tasks with only one operation as:

$$\mathcal{P}_{jq_k} = \bigsqcup_{t=1..v} Tr_{jq_{k_t}}$$

Furthermore, let ι_{j_k} be the instance of the object o_j in the IoT $\mathcal{I}_k$. Let $Inst_{j_k}$ be the set of the instances of $\mathcal{I}_k$ with which ι_{j_k} performed at least one transaction in the past. In this case, we can define the profile $\mathcal{P}_{j_k}$ of ι_{j_k} as :

$$\mathcal{P}_{j_k} = \bigsqcup_{\iota_{q_k} \in Inst_{j_k}} \mathcal{P}_{jq_k}$$

Finally, let o_j be an object and let $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_l\}$ be the set of the IoTs it participates to. Let $ObjInst_j$ be the instances of o_j in the IoTs of the MIoT. We can define the profile $\mathcal{P}_j$ of o_j as:

$$\mathcal{P}_j = \bigsqcup_{\iota_{j_k} \in ObjInst_j} \mathcal{P}_{j_k}$$

Everything we have seen so far regards the profile of an instance from a "content-based" perspective (i.e., taking its past behavior into account). Beside this perspective, another one can be considered, i.e., the "collaborative filtering" perspective (i.e., based on the similarity of the behaviors of the instance neighbors). However, it is out of the scope of our approach.

6.3.3 Trust of an instance in another one of the same IoT

Let ι_{j_k} and ι_{q_k} be two instances of an IoT $\mathcal{I}_k$. We want to define the trust T_{jq_k} of ι_{j_k} in ι_{q_k}. Actually, this trust is not unique, because it depends on both the topic and the format of the data exchanged during the corresponding transactions. As a consequence, T_{jq_k} is a matrix and has a value for each possible combination of topics and formats. Since the possible formats are 4 (i.e., "audio", "video", "image", and "text"), T_{jq_k} is a $|topicSet_{jq_k}| \times 4$ matrix. The element $T_{jq_k}[u,v]$ of this matrix indicates the trust of ι_{j_k} in ι_{q_k} regarding the topic u delivered in the format v. This trust depends on several factors, namely: *(i)* the fraction of successful transactions; *(ii)* the overall number of transactions, which is an indicator of the robustness of the result; *(iv)* the size of exchanged files; *(v)* the timestamp of the last transaction, which is an indicator of the possible obsolescence of the relationship between ι_{j_k} and ι_{q_k}.

In order to define $T_{jq_k}[u,v]$, we must introduce some notions. Specifically:

- $tranSet_{jq_k}[u,v]$ is the subset of the transactions $tranSet_{jq_k}$ whose content presents the topic u in the format v at least once;
- $OKTranSet_{jq_k}[u,v]$ is the fraction of successful transactions in $tranSet_{jq_k}[u,v]$;
- $maxNumTranSet_k[u,v]$ is the maximum number of transactions, concerning the topic u in the format v, performed between two given instances of $\mathcal{I}_k$. It is defined as:

$$maxNumTranSet_k[u,v] = max_{\iota_{j_k} \in Inst_k, \iota_{q_k} \in Inst_k, \iota_{j_k} \neq \iota_{q_k}} |tranSet_{jq_k}[u,v]|$$

- $size_{jq_k}[u,v]$ is the size of contents concerning the topic u in the format v exchanged by ι_{j_k} and ι_{q_k}.
- $maxSize_k[u,v]$ is the maximum size of the overall contents concerning the topic u in the format v, exchanged between two instances of $\mathcal{I}_k$. It is defined as:

$$maxSize_k[u,v] = max_{\iota_{j_k} \in Inst_k, \iota_{q_k} \in Inst_k, \iota_{j_k} \neq \iota_{q_k}} size_{jq_k}[u,v]$$

We are now able to define $T_{jq_k}[u,v]$. It consists of a pair $(V_{jq_k}[u,v], LTS_{jq_k}[u,v])$. $V_{jq_k}[u,v]$ can be computed as a weighted mean of the parameters introduced above. Specifically:

$$V_{jq_k}[u,v] = \frac{\alpha \cdot OKTranSet_{jq_k}[u,v] + \beta \cdot \frac{|tranSet_{jq_k}[u,v]|}{maxNumTranSet_k[u,v]} + \rho \cdot \frac{size_{jq_k}[u,v]}{maxSize_k[u,v]}}{\alpha + \beta + \rho}$$

Here, α, β and ρ denote the weights of the three components of the mean. We have experimentally set $\alpha = 0.55$, $\beta = 0.35$ and $\rho = 0.10$ (see Section 6.4.1 for all details).

$LTS_{jq_k}[u,v]$ is the last ending timestamp concerning a transaction of $tranSet_{jq_k}[u,v]$.

Example (continued)

Consider the smart shopping center described in Section 6.5.2 and assume that Alice and Bob enter the electronics store. Assume, also, that Alice needs information about smart home products. In this case, the smartphone of Alice asks the other smart objects of the customers in the store if they have information about smart home products sold there (for example, in order to know the current promotions of the store). Assume that Bob had already visited the store several times in the last days for searching information and buying a smart home product. The smartphone of Bob (o_3) can answer the smartphone of Alice (o_1) and the corresponding instances ι_{1_2} and ι_{3_2} can start their interaction. Here:

- $tranSet_{13_2}$ denotes the transactions exchanged between o_1 and o_3 in the electronics store (i.e., between ι_{1_2} and ι_{3_2}). Assume that $tranSet_{13_2} = 150$.
- $tranSet_{13_2}[u,v]$ represents the transactions of $tranSet_{13_2}$ about the topic u (in our case, smart home products) in the format v (for instance, videos). Assume that $tranSet_{13_2}[u,v] = 90$.
- $OKTranSet_{13_2}[u,v]$ is the fraction of successful transactions of $tranSet_{13_2}[u,v]$. Suppose that some transactions of $tranSet_{13_2}[u,v]$ failed because the smartphone of Bob had connection problems with the WiFi of the store. Assume that the successful transactions of $tranSet_{13_2} = 85$ so that $OKtranSet_{13_2}[u,v] = \frac{85}{90} = 0.94$.

- $maxNumTranSet[u,v]$ is the maximum number of videos about smart home products exchanged between two smart objects in the electronics store. Assume that $maxNumTranSet[u,v] = 110$.
- $size_{13_2}[u,v]$ is the overall size of the videos concerning smart home products exchanged between the smartphones of Alice and Bob. Assume that $size_{13_2}[u,v] = 10\ MB$.
- $maxSize[u,v]$ is the maximum overall size of the videos about smart home products exchanged between two smart objects in the electronics store. Assume that $maxSize[u,v] = 12\ MB$.
- $V_{13_2}[u,v]$ is the value of the trust about videos on smart home products that the smartphone of Alice has in the smartphone of Bob in the electronics store. It is equal to:

$$V_{13_2}[u,v] = \frac{0.55 \cdot 0.94 + 0.35 \cdot \frac{90}{110} + 0.10 \frac{10}{12}}{0.55 + 0.35 + 0.10} = 0.89$$

- $LTS_{13_2}[u,v]$ is the timestamp of the last video about smart home products that the smartphone of Bob sent to the smartphone of Alice in the electronics store.

6.3.4 Trust of an object in another one of the MIoT

Let o_j and o_q be two objects of $\mathcal{M}$. Let $\mathcal{M}_{jq} = \{\mathcal{I}_1, \cdots, \mathcal{I}_l\}$ be the subset of the IoTs of $\mathcal{M}$ that simultaneously contain one instance of o_j and one instance of o_q. In this case, it is possible to define the trust $T_{jq}[u,v]$ of o_j in o_q regarding the topic u delivered in the format v. Also in this case, $T_{jq}[u,v]$ consists of a pair $(V_{jq}[u,v], LTS_{jq}[u,v])$. Here:

- $V_{jq}[u,v]$ is set to the average of the trusts that the instances of o_j have in the instances of o_q in the IoTs of $\mathcal{M}_{jq}$:

$$V_{jq}[u,v] = \frac{\sum_{k=1..l} V_{jq_k}[u,v]}{l}$$

- $LTS_{jq}[u,v]$ is set to the maximum ending timestamp of any transaction simultaneously involving one instance of o_j and one instance of o_q:

$$LTS_{jq}[u,v] = max_{k=1..l} LTS_{jq_k}[u,v]$$

Example (continued)

Consider the smart shopping center described in Section 6.5.2. Assume that Alice and Bob first enter the supermarket and then the electronics store. Consider the smartphone of Alice (o_1) and the one of Bob (o_3) and assume that they interact in both stores to help Alice find the smart home products she desires.

Assume that the value of the trust about videos on smart home products that the smartphone of Alice has in the smartphone of Bob in the supermarket is equal to 0.93[1]. In the example in Section 6.3.3, we have seen that the trust about videos on smart home products that the smartphone of Alice has in the smartphone of Bob in the electronics store was equal to 0.89.

As a consequence, the value of the overall trust about videos on smart home products that the smartphone of Alice has in the smartphone of Bob in the whole smart shopping center is:

$$V_{13}[u,v] = \frac{0.93+0.89}{2} = 0.91$$

Instead, $LTS_{13}[u,v]$, i.e. the ending timestamp of the last video on smart home products that the smartphone of Bob sent to the smartphone of Alice coincides with the value of $LTS_{13_2}[u,v]$ computed in the example of Section 6.3.3, because Alice ad Bob entered first the supermarket and then the electronics store.

6.3.5 Reputation of an instance in an IoT

Let $\mathcal{I}_k$ be an IoT and let ι_{j_k} be an instance of $\mathcal{I}_k$. The reputation $R_{j_k}[u,v]$ of ι_{j_k}, regarding the topic u in the format v, depends on the following factors: *(i)* the number of instances from which ι_{j_k} received transactions in the past; *(ii)* the trust that these instances have in ι_{j_k}; *(iii)* the reputation of these instances in $\mathcal{I}_k$; *(iv)* their oldness.

To formalize this type of dependencies, the classical approach involves the usage of the PageRank formula. To proceed in this direction, it is necessary to introduce the following preliminary definitions:

- $Age_{q_k}[u,v]$ is the number of days spent from the first transaction, regarding the topic u delivered in the format v, performed by the object o_q in the IoT $\mathcal{I}_k$.
- $Age_k^{max}[u,v]$ is the maximum number of days spent from the first transactions, regarding the topic u delivered in the format v, performed by an object in the IoT $\mathcal{I}_k$.
- $R_k^{max}[u,v]$ is the maximum reputation of an instance of $\mathcal{I}_k$, regarding the topic u delivered in the format v.

We are now able to define the formula for the computation of $R_{j_k}[u,v]$. In particular:

$$R_{j_k}[u,v] = \gamma + (1-\gamma) \cdot \frac{\sum_{\iota_{q_k} \in nbh^{in}(\iota_{j_k})} T_{qj_k}[u,v] \cdot R_{q_k}[u,v] \cdot \frac{Age_{q_k}[u,v]}{Age_k^{max}[u,v]}}{|nbh^{in}(\iota_{j_k})|}$$

[1] This value is obtained by proceeding in the same way as we did for the electronics store in Section 6.3.3.

In this formula, γ is the damping factor generally adopted in the PageRank. It determines the minimum absolute reputation assigned to an instance of $\mathcal{I}_k$. From a more abstract viewpoint, it allows us to tune the fraction of the absolute reputation that ι_{j_k} transmits to ι_{q_k}.

$R_{j_k}[u,v]$ belongs to the real interval $[\gamma,+\infty)$. In order to obtain a reputation value belonging to the interval $[0,1]$ and, at the same time, to normalize the reputations of the instances of the IoTs of the MIoT, we define the relative reputation $\widehat{R_{j_k}}[u,v]$ of ι_{j_k} in $\mathcal{I}_k$ as follows:

$$\widehat{R_{j_k}}[u,v] = \frac{R_{j_k}[u,v]}{R_k^{max}[u,v]}$$

Example (continued)

Consider the smart shopping center described in Section 6.5.2. We want to evaluate the reputation of the smartphone of Bob in the electronics store, i.e. the reputation of ι_{3_2} in $\mathcal{I}_2$. Here:

- $Age_{1_2}[u,v]$ (resp., $Age_{2_2}[u,v]$, $Age_{4_2}[u,v]$) is the number of days since the first transmission of a video on smart home products performed by the smartphone of Alice (resp., the smartwatch of Alice, the smartwatch of Bob) in the electronics store. Assume that $Age_{1_2}[u,v] = 20$, (resp., $Age_{2_2}[u,v] = 20$, $Age_{4_2}[u,v] = 70$).
- $Age_2^{max}[u,v]$ is the maximum number of days since the transmission of a video on smart home products performed by a smart object in the electronics store. Assume that $Age_2^{max}[u,v] = 75$.
- Assume that all the objects currently present in the electronics store are totally connected to each other. As a consequence, $nbh^{in}(\iota_{3_2}) = \{\iota_{1_2}, \iota_{2_2}, \iota_{4_2}\}$.
- $R_2^{max}[u,v]$ is the maximum reputation of a smart object transmitting a video on smart home products in the electronics store. Assume that $R_2^{max}[u,v] = 0.68$.
- γ is the minimum absolute reputation assigned to an object in the electronics store. Assume $\gamma = 0.30$.
- $T_{13_2}[u,v]$ (resp., $T_{23_2}[u,v]$, $T_{43_2}[u,v]$) is the value of the trust that the smartphone of Alice (resp., the smartwatch of Alice, the smartwatch of Bob) has in the smartphone of Bob, regarding videos on smart home products sent in the electronics store. Assume that $T_{13_2}[u,v] = 0.95$, $T_{23_2}[u,v] = 0.90$ and $T_{43_2}[u,v] = 1$.
- $R_{1_2}[u,v]$ (resp., $R_{2_2}[u,v]$, $R_{4_2}[u,v]$) is the value of the reputation of the smartphone of Alice (resp., the smartwatch of Alice, the smartwatch of Bob), regarding videos on smart home products sent in the electronics store. Assume that $R_{1_2} = 0.98$, $R_{2_2} = 0.93$ and $R_{3_2} = 0.96$.
- The reputation of the smartphone of Bob, regarding videos on smart home products sent in the electronics store, is obtained as:

$$R_{3_2}[u,v] = 0.30 + (1 - 0.30) \cdot \frac{0.95 \cdot 0.98 \cdot \frac{20}{75} + 0.90 \cdot 0.93 \cdot \frac{20}{75} + 1 \cdot 0.96 \cdot \frac{70}{75}}{3} = 0.62$$

- The normalized reputation $\widehat{R_{3_2}}[u,v]$ of the smartphone of Bob, regarding videos on smart home products sent in the electronics store, is obtained as:

$$\widehat{R_{3_2}}[u,v] = \frac{0.62}{0.68} = 0.91$$

6.3.6 Reputation of an object in a MIoT

Let o_j be an object of $\mathcal{M}$. Let $\mathcal{M}_j = \{\mathcal{I}_1, \cdots, \mathcal{I}_l\}$ be the subset of the IoTs of $\mathcal{M}$ containing one instance of o_j.

The reputation of o_j depends on both the trust that its instances receive in each IoT of $\mathcal{M}_j$ and the reputation of the object, which the instance providing this trust refers to. To formalize this concept, we can say that the reputation of o_j, regarding the topic u delivered in the format v, is defined as follows:

$$R_j[u,v] = \delta + (1 - \delta) \cdot \frac{\sum_{k=1..l} \sum_{\iota_{q_k} \in nbh^{in}(\iota_{j_k})} V_{qj}[u,v] \cdot R_q[u,v]}{l \cdot |nbh^{in}(\iota_{j_k})|}$$

As in the previous case, this formula is similar to the PageRank one. δ is the damping factor and its semantics is analogous to the one of γ seen in Section 6.3.5.

At this point, it is necessary to proceed with the normalization of $R_j[u,v]$. This task is performed in a way analogous to the one defined for the instance reputation in the previous section:

$$\widehat{R_j}[u,v] = \frac{R_j[u,v]}{R^{max}[u,v]}$$

Example (continued)

Consider the smart shopping center described in Section 6.5.2. The reputation $\widehat{R_3}[u,v]$ of the smartphone of Bob, when it sends videos on smart home products in the whole smart shopping center, can be computed in a way analogous to the computation of the reputation $\widehat{R_{3_2}}[u,v]$ of the smartphone of Bob in the electronics store, illustrated in the example of Section 6.3.5. For this reason, and due to space limitations, we do not report all details of the computation of $\widehat{R_3}[u,v]$ below.

6.3.7 Reputation of an IoT in a MIoT

The reputation of an IoT $\mathcal{I}_k$ in $\mathcal{M}$, regarding the topic u delivered in the format v, is given by the average of the reputations of the objects of $\mathcal{M}$ having one instance in $\mathcal{I}_k$.

If we introduce the set Obj_k of the objects having one instance in $\mathcal{I}_k$, the reputation $\widehat{\mathcal{R}^k}[u,v]$ of $\mathcal{I}_k$ in $\mathcal{M}$ can be formalized as follows:

$$\widehat{\mathcal{R}^k}[u,v] = \frac{\sum_{j \in Obj_k} \widehat{R}_{j_k}[u,v]}{|Obj_k|}$$

Example (continued)

Consider the smart shopping center described in Section 6.5.2. The reputation $\widehat{\mathcal{R}^2}[u,v]$ of the IoT associated with the electronics store, when the objects present therein send videos on smart home products in the smart shopping center, can be computed in a way analogous to the computation of the trust $T_{13}[u,v]$ of the smartphone of Alice in the smartphone of Bob, when this last sends video on smart home products in the smart shopping center, as illustrated in the example of Section 6.3.3. For this reason, due to space limitations, we do not report all details of the computation of $\widehat{\mathcal{R}^2}[u,v]$ below.

6.3.8 Trust of an IoT in another IoT

The trust $\mathcal{T}^{hk}[u,v]$ of an IoT $\mathcal{I}_h$ in an IoT $\mathcal{I}_k$, regarding the topic u delivered in the format v, consists of a pair $(\mathcal{V}^{hk}[u,v], \mathcal{LTS}^{hk}[u,v])$.

$\mathcal{V}^{hk}[u,v]$ is defined as the average of the trust values of any object of $\mathcal{I}_h$ in any object of $\mathcal{I}_k$, with which it performed at least one transaction.

To formally define $\mathcal{V}^{hk}[u,v]$, we must introduce the set $tranSet_{j_k}[u,v]$ of the transactions that any instance ι_{j_h} of $\mathcal{I}_h$ carried out with any instance of $\mathcal{I}_k$ and having in their content the topic u delivered in the format v. After having introduced $tranSet_{j_k}[u,v]$, we can define $\mathcal{V}^{hk}[u,v]$ as follows:

$$\mathcal{V}^{hk} = \frac{\sum_{j \in Obj_h} \sum_{q \in tranSet_{j_k}[u,v]} V_{jq}[u,v]}{\sum_{j \in Obj_h} |tranSet_{j_k}[u,v]|}$$

$\mathcal{LTS}^{hk}[u,v]$ is the last ending timestamp that can be found in a transaction involving any instance of $\mathcal{I}_h$ with any instance of $\mathcal{I}_k$.

Example (continued)

Consider the smart shopping center described in Section 6.5.2. The trust $\mathcal{T}^{21}[u,v]$ of the IoT associated with the electronics store in the IoT associated with the supermarket, when the objects in this last network send videos on smart products, can be computed in a way analogous to the computation of the trust $T_{13_2}[u,v]$ of the smartphone of Alice in the smartphone of Bob, when it sends videos on smart home products in the electronics store, illustrated in the example of Section 6.3.3. For this reason, due to space limitations, we do not report all details of the computation of $\mathcal{T}^{21}[u,v]$ below.

6.3.9 Trust of an object in an IoT

Let o_j be an object of $\mathcal{M}$ and let $\mathcal{I}_k$ be an IoT of $\mathcal{M}$. Again, the trust $\mathcal{T}_j^k[u,v]$ of o_j in $\mathcal{I}_k$, regarding the topic u delivered in the format v, consists of a pair $(\mathcal{V}_j^k[u,v], \mathcal{LTS}_j^k[u,v])$. In the computation of $\mathcal{T}_j^k[u,v]$ we must distinguish two cases, namely:

- o_j has one instance ι_{j_k} in $\mathcal{I}_k$. In this case, let $Inst_{j_k}[u,v]$ be the set of the instances of $\mathcal{I}_k$ with which ι_{j_k} carried out at least one transaction involving the topic u delivered in the format v in the past. $\mathcal{V}^{hk}[u,v]$ is defined as the average of the trusts of ι_{j_k} in all the instances of $Inst_{j_k}[u,v]$. More formally:

$$\mathcal{V}_j^k[u,v] = \frac{\sum_{q \in Inst_{j_k}[u,v]} V_{jq_k}[u,v]}{|Inst_{j_k}[u,v]|}$$

  $\mathcal{LTS}_j^k[u,v]$ is the last ending timestamp that can be found in a transaction involving ι_{j_k} and any instance of $Inst_{j_k}[u,v]$.
- o_j has no instance in $\mathcal{I}_k$. In this case, the trust of o_j in $\mathcal{I}_k$ is equal to the sum of the trusts of the instances of o_j in the IoTs it belongs to, weighted by the trust of the corresponding IoT in $\mathcal{I}_k$. More formally, let $\mathcal{M}_j = \{\mathcal{I}_1, \cdots, \mathcal{I}_l\}$ be the set of the IoTs of $\mathcal{M}$ containing one instance of o_j. In this case:

$$\mathcal{V}_j^k[u,v] = \frac{\sum_{h=1..l} \mathcal{V}_j^h[u,v] \cdot \mathcal{V}^{hk}}{l}$$

  $\mathcal{LTS}_j^k[u,v]$ is the maximum LTS among the ones associated with $\mathcal{T}_j^h[u,v]$, $1 \leq h \leq l$. Formally speaking:

$$\mathcal{LTS}_j^k[u,v] = max_{h=1..l} \mathcal{LTS}_j^h[u,v]$$

Example (continued)

Consider the smart shopping center described in Section 6.5.2. The trust $\mathcal{T}_3^2[u,v]$ that the smartphone of Bob has in the IoT associated with the electronics store, when the objects in this last network send videos on smart home products, can be computed in a way analogous to the computation of the trust $T_{13_2}[u,v]$ of the smartphone of Alice in the smartphone of Bob, when it sends videos on smart home products in the smart shopping center. We have illustrated the computation of $T_{13_2}[u,v]$ in the example of Section 6.3.3. For this reason, due to space limitations, we do not report all details of the computation of $\mathcal{T}_3^2[u,v]$ below.

6.4 Results

In this section, we present the set of experiments that we carried out to evaluate the performance of our approach from several viewpoints. First of all, in Subsec-

tion 6.4.1, we describe how we experimentally set the weight of α, β and ρ in the computation of the trust of an instance into another one of the same IoT, in order to give an example of how we set the weights in our approach. Then, we describe our testbed in Subsection 6.4.2, whereas, in Subsections 6.4.3 - 6.4.6, we illustrate our tests, along with the underlying motivations and the results obtained. Finally, in Subsection 6.4.7, we present an experiment to evaluate the accuracy of our approach.

6.4.1 Setting of weights

In this experiment, we aimed at determining the values of α, β and ρ in the computation of the value of the trust of an instance in another one of the same IoT (see Section 6.3.3). First of all, we observe that, roughly speaking, α represents the weight of the fraction of the correct transactions between ι_{j_k} and ι_{q_k}, β denotes the significance of the number of transactions existing between ι_{j_k} and ι_{q_k}, whereas ρ indicates the weight of the size of the content exchanged between ι_{j_k} and ι_{q_k}.

To perform this experiment, we initially selected 100 smart objects and we connected them to form an IoT. Then, we let them to perform 100,000 transactions through which they exchanged data. At the end of this task, we computed the values of trust for each pair of objects by applying the formulas reported in Section 6.3.3.

Afterwards, we forced some fictitious wrong behaviors in the transactions between the smart objects of the network. The entity of the error was varying; it was evaluated by some domain experts as *null*, *small*, *medium*, *high* and *very high*. In particular, we made sure that 20% of the transactions had a *null* (resp., *small*, *medium*, *high*, *very high*) error.

After this, for each pair of smart objects, we asked the human experts to evaluate the overall perturbation caused in their transactions by the wrong behavior induced in the experiment. The possible evaluations were *negligible*, *small*, *medium*, *high* and *very high*.

At this point, for each pair of smart objects, we recomputed the value of trust with the perturbed transactions. Then, we compared the size of the change of the trust values against the values themselves. We considered as *negligible* (resp., *small*, *medium*, *high* and *very high*) the perturbation caused in a trust value if its change was less than 20% (resp., between 20% and 40%, between 40% and 60%, between 60% and 80%, more than 80%) of the original value.

We repeated this last part of the experiment with different combinations of values of α, β and ρ. In particular, the adopted combinations are the ones reported in the first three columns of Table 6.1.

Finally, for each weight combination, we computed the percentage of times the evaluation of trust perturbations performed by our approach and the one carried

α	β	ρ	*Percentage of errors*
0.35	0.25	0.05	35.67 %
0.35	0.25	0.10	30.34 %
0.35	0.25	0.15	25.86 %
0.35	0.35	0.05	25.56 %
0.35	0.35	0.10	20.87 %
0.35	0.35	0.15	15.12 %
0.35	0.45	0.05	15.76 %
0.35	0.45	0.10	10.95 %
0.35	0.45	0.15	5.56 %
0.45	0.25	0.05	25.38 %
0.45	0.25	0.10	20.54 %
0.45	0.25	0.15	15.37 %
0.45	0.35	0.05	15.54 %
0.45	0.35	0.10	10.48 %
0.45	0.35	0.15	5.83 %
0.45	0.45	0.05	5.69 %
0.45	0.45	0.10	5.25 %
0.45	0.45	0.15	5.58 %
0.55	0.25	0.05	15.28 %
0.55	0.25	0.10	10.94 %
0.55	0.25	0.15	5.59 %
0.55	0.35	0.05	5.93 %
0.55	0.35	0.10	0.50 %
0.55	0.35	0.15	5.73 %
0.55	0.45	0.05	5.28 %
0.55	0.45	0.10	10.62 %
0.55	0.45	0.15	15.28 %
0.65	0.25	0.05	5.74 %
0.65	0.25	0.10	5.34 %
0.65	0.25	0.15	5.79 %
0.65	0.35	0.05	5.93 %
0.65	0.35	0.10	10.48 %
0.65	0.35	0.15	15.52 %
0.65	0.45	0.05	15.28 %
0.65	0.45	0.10	20.58 %
0.65	0.45	0.15	25.92 %
0.75	0.25	0.05	5.36 %
0.75	0.25	0.10	10.83%
0.75	0.25	0.15	15.28 %
0.75	0.35	0.05	15.27 %
0.75	0.35	0.10	20.74 %
0.75	0.35	0.15	25.94 %
0.75	0.45	0.05	25.38 %
0.75	0.45	0.10	30.19 %
0.75	0.45	0.15	35.18 %

Table 6.1: Setting of the weights α, β and ρ in the computation of the trust of an instance in another one of the same IoT

out by human experts coincided. The obtained values are reported in the fourth column of Table 6.1. From the analysis of this table, we can see that the optimal combination of values is $\alpha = 0.55$, $\beta = 0.35$ and $\rho = 0.10$. We can also observe that the combinations slightly differing from the previous one produce a small number of errors. On the other side, as long as the combinations differ from the optimal ones, the errors increase. This witnesses the optimal resilience of our approach that, however, is (correctly) sensitive to weight errors when these become high.

Interestingly, in this experiment, we were guided only by the semantics of the three measures weighted by α, β and ρ, and not on the nature of the scenario where it was performed. As a consequence, even if this experiment could be repeated each time we want to determine the values of α, β and ρ in the most disparate scenarios,

we are confident that the values obtained are general and do not depend on the application environment in which our approach is operating.

6.4.2 Testbed

In order to perform the next experiments, we had the necessity to create several MIoTs with different sizes, ranging from hundreds to thousands of nodes. Since, currently, real MIoTs with the size and the variety handled by our model do not exist yet, we constructed a MIoT simulator. This tool starts from real data and returns simulated MIoTs with certain characteristics specified by the user.

The MIoTs created by our simulator follow the paradigm described in Section 6.3.1. Our MIoT simulator is also provided with a suitable interface allowing a user to "personalize" the MIoT to construct by specifying the desired values for several parameters, such as the number of nodes, the maximum number of instances of an object, and so forth.

To make "concrete" and "plausible" the created MIoTs, our simulator leverages a real dataset. It regards the taxi routes in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. It can be found at the address `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. Each route contains several Points of Interests corresponding to the GPS coordinates of the vehicle.

We partitioned the city of Porto in six areas and associated a real IoT with each of them. Our simulator associates an object with a given route recorded in the dataset and an object instance with each partition of a route belonging to an area. It creates a MIoT node for each instance and a c-arc for each pair of instances belonging to the same route. Furthermore, it creates an i-arc between two nodes of the same IoT if the length of the time interval between the corresponding routes is less than a certain threshold th_t. The weight of the i-arc indicates the length of this time interval. The value of th_t can be specified through the constructor interface. Clearly, the higher th_t, the more connected the constructed MIoT.

As far as instance profiles are concerned, since there are no available thing profiles, we had to simulate them. However, we aimed at making them as real as possible. For this purpose, we performed a sentiment analysis task for each of the six areas in which we partitioned the city of Porto and for each day which the dataset refers to. To carry out this task, we leveraged IBM Watson on the social media and blogs available in it. Having this data at disposal, our simulator assigns to each instance the most common topics (along with the corresponding occurrences) discussed in that area in the day on which the corresponding route took place. The constructed MIoTs are returned in a format that can be directly processed by the cypher-shell of Neo4J.

The interested reader can find the MIoTs adopted in the experiments at the address `http://daisy.dii.univpm.it/miot/datasets/trustReputation`.

We carried out all the tests presented in this section on a server equipped with an Intel I7 Quad Core 7700 HQ processor and 16 GB of RAM with Ubuntu 16.04 operating system. To implement our approach, we adopted: *(i)* Python, as programming language; *(ii)* Neo4J (Version 3.4.5), as underlying DBMS.

6.4.3 Computation time

Our first test is devoted to evaluating the computation time of our approach. Indeed, since it could operate in large MIoTs, whose IoTs could consist of even hundreds of nodes, it is necessary to verify if, in these real cases, the time it needs to return a result is still acceptable.

6.4.3.1 Trust of an instance in another one of the same IoT and of an object in another one of the MIoT

In this experiment, we considered several MIoTs having a different number of nodes. Given a MIoT $\mathcal{M}$, we considered all its IoTs $\mathcal{I}_1, \cdots, \mathcal{I}_6$ (see Section 6.4.2). For each IoT, we computed the trust of each of its nodes in the others. At the first iteration, we set the value of the trust of a node in any other of the MIoT to 0.5. In other words, we decided to assume a "neutral" policy in order to not outweigh either positively or negatively on the trust of one node in another.

We performed a total number of 60,000 transactions in the MIoT and we recomputed all trust values every 10 transactions (in the following, we call *epoch* an interval of 10 transactions). We measured the time required by our approach to compute the trust of an instance in another one of the same IoT against the number of epochs for MIoTs having a different number of nodes (ranging from 10 to 1,000). Finally, we averaged the obtained computation times for all the instances of the MIoT. The results obtained are reported in Figure 6.3.

From the analysis of this figure we can observe that the computation time is very small when the numbers of epochs is less than 2,000, independently of the size of the MIoT. After 2,000 epochs, it starts to increase more quickly. In this case, if the number of the nodes of the MIoT is lower than 1,000, the computation time and the quickness of its growth are still acceptable. Instead, in presence of a MIoT with more than 1,000 instances, the computation time tends to become excessively high and quickly unacceptable.

As shown in Section 6.3.4, the trust of an object in another one of $\mathcal{M}$ is easily determined by computing the average values of the trust of its instances in the ones

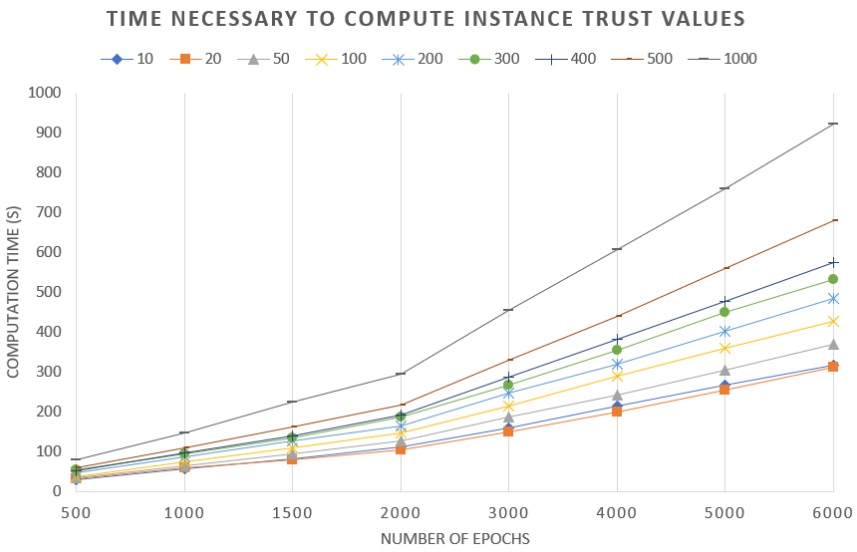


Fig. 6.3: Average time of the computation of the trust of an instance in another one of the same IoT against the number of epochs for MIoTs with different numbers of nodes

of the other object in the IoTs where both of them are present. The additional computations necessary to obtain it, once the trust values of the corresponding instances have been determined, are negligible. As a consequence, all the considerations about the computation time that we made for the trust of an instance in another one of the same IoT can be extended to the trust of an object in another one of the MIoT.

With regard to this result, we observe that the cases in which the computation time begins to become unacceptable regard scenarios that we do not currently find in real cases. In fact, in order to start having computational problems, we should be in presence of a MIoT consisting of more than 1,000 instances. If we consider that, in real cases, the number of IoTs in a MIoT is currently less than 10, we should have more than 100 objects which simultaneously want to interact in all the IoTs of the MIoT. This highly unlikely scenario could be still managed by our approach if the number of epochs used to compute the trust values is less than 2,000. Now, since an epoch corresponds to 10 transactions, this means that our approach starts to present an excessive computation time only in presence of about 100 objects wanting to simultaneously interact in 10 different IoTs of the MIoT and performing at least 20,000 simultaneous transactions.

Actually, the current MIoTs would consist of at least 3-5 IoTs. The number of objects in each IoT that want to simultaneously interact is less than 50. As a consequence, in real cases, a MIoT consists of at most 200-400 instances. Furthermore, not all the objects want to interact with all the other ones. In fact, the number of pairs

of objects wanting to interact with each other is very limited and do not exceed 400. In addition, in real cases, the overall number of transactions necessary to compute a stable value of trust for each pair of interacting instances does not exceed 20. With all the hypotheses above, our approach would need at most 8,000 transactions to determine stable values of trust for each pair of interacting objects. This number is much smaller than the limit value of 20,000 transactions. Clearly, we think that, in the future, the size and the density of MIoTs will increase; however, the computing power available in servers should increase too. In any case, if this increase would be not sufficient, we could adopt two countermeasures to make the computation time still reasonable. Indeed: *(i)* we could increase the number of transactions associated with an epoch (for instance, from 10 to 100 or to 1,000); *(ii)* we could use distributed and parallel processing to perform trust evaluation.

6.4.3.2 Reputation of an instance in an IoT and of an object in the MIoT

In this experiment, we considered the same MIoTs adopted in the previous one and, for each instance of an object, we computed its reputation in the corresponding IoT. Also in this case, at the first iteration, we set the initial reputation of each instance to 0.5. In this case, we performed a total number of 600,000 transactions.

Reputation is intrinsically much more static than trust. As a consequence, it appears more reasonable to assume epochs of 100 transactions, instead of 10. We measured the time required by our approach for computing the reputation of each instance in its IoT against the number of epochs for the MIoTs adopted in the previous experiment. Then, we averaged these values for all the instances of the MIoT. The results obtained are reported in Figure 6.4.

From the analysis of this figure, we can observe that the time necessary to compute the values of instance reputation is always low when the number of epochs is lower than, or equal to, 2,000 and the number of nodes is lower than, or equal to, 500. When the number of nodes is higher, the computation time increases, even if it is still acceptable for a number of epochs lower than, or equal to, 2,000. When the number of epochs is higher than 2,000, the computation time starts to rapidly increase. It tends to become unacceptable when the number of nodes is higher than 500 and the number of epochs is higher than 2,000. With regard to this result, we observe that all the reasonings about the computation of trust in real cases, which we have presented at the end of Section 6.4.3.1, can be extended here to the computation of reputation in real cases.

As illustrated in Section 6.3.6, the definition of the reputation of an object in a MIoT is structurally similar to the definition of the reputation of an instance in an

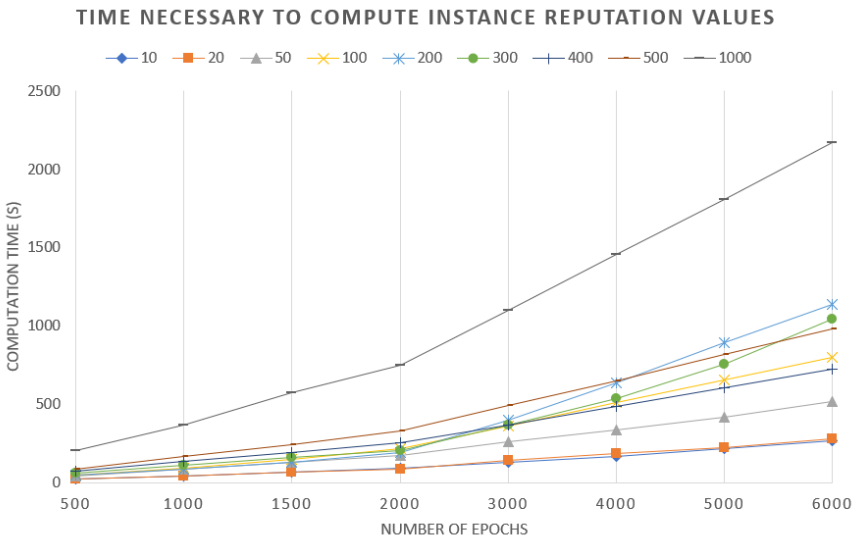


Fig. 6.4: Average time of the computation of the reputation of an instance in its IoT against the number of epochs for MIoTs with different numbers of nodes

IoT. As a consequence, all the considerations about the computation time that we have illustrated above can be easily extended to this last case.

6.4.3.3 Reputation of an IoT in the MIoT

As shown in Section 6.3.7, the reputation of an IoT in the MIoT is obtained by averaging the reputation of the objects having one instance in it. The computation of the average is negligible after that the reputation of the corresponding objects has been determined. Therefore, all the considerations about the computation time, which we made for the reputation of an object in the MIoT, can be extended to the reputation of an IoT in the MIoT.

6.4.3.4 Trust of an instance in another one of the same IoT

In order to investigate the features characterizing the trust of an instance in another one of the same IoT, we applied the guidelines described in Section 6.4.3.1 even if, this time, we focused on values and not on computation time. In Figure 6.5, we report the average values of this trust against the number of epochs for the same MIoTs we introduced in the previous section. From the analysis of this figure, we can observe that, initially, as the number of transactions increases, the trust values increase too. This fact is justified by considering that, as the number of correct transactions increases, the instances "have more confidence" in each other. This increase is particularly evident until to 1,000 epochs (i.e., 10,000 transactions). When the number of epochs ranges between 1,000 and 2,000, the value of the trust still slightly grows,

even if the speed of growth is much lower than before. Finally, after 4,000 epochs, the average trust reaches an approximately fixed value in some cases, whereas, in other ones, it grows very slowly.

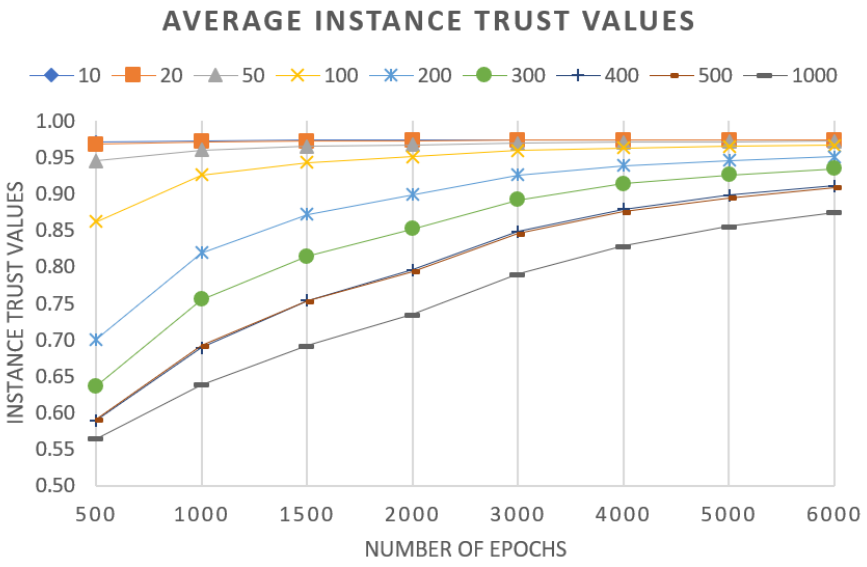


Fig. 6.5: Average values of the trust of an instance in another one of the same IoT against the number of epochs for MIoTs with a different number of nodes

Observe that, in Figure 6.5, the performance of our approach depends on the number of epochs and the number of instances. Clearly, the increase of the number of epochs always leads to an increase of the trust between instances or objects. On the other side, it requires a higher number of transactions and, ultimately, a higher computational and time cost. Clearly, a tradeoff is necessary between these two exigencies. In particular, in cases where computational costs are more important than accuracy, it is better to choose a number of epochs lower than 2,000. By contrast, whenever accuracy is extremely important and computational costs can be partially sacrificed, it is better to choose a number of epochs higher than 2,000. As for the number of instances, it depends on the scenario on which the MIoT is operating, and cannot be tuned by the operator.

We also computed the distributions of the trust values after 1,000, 2,000 and 3,000 epochs. We performed this computation for the MIoT with 300 instances adopted in the previous experiments. The results obtained are reported in Figure 6.6. From the analysis of this figure we can observe that the distribution shape moves to the right. This phenomenon is very evident when passing from 1,000 to 2,000 epochs, but it is still significant also when passing from 2,000 to 3,000 epochs.

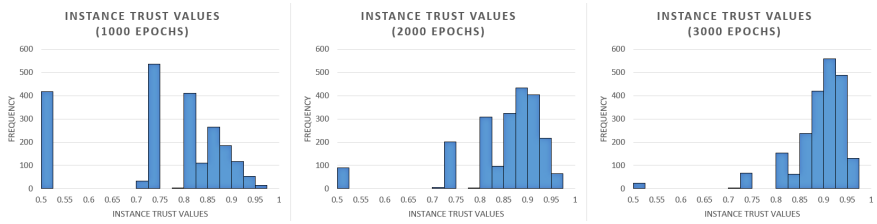


Fig. 6.6: Distribution of the trust of an instance in another one of the same IoT after 1,000, 2,000 and 3,000 epochs for the MIoT with 300 instances

A final parameter that we computed is the standard deviation after 1,000, 2,000 and 3,000 epochs. The values we obtained were 0.1385, 0.0934, and 0.0626, respectively. This result is extremely interesting; as a matter of fact, already after 1,000 epochs, the values of the standard deviation are acceptable. Furthermore, when passing from 1,000 to 2,000 and from 2,000 to 3,000 epochs, we can observe a quick decrease of the corresponding values. This denotes a high stability of the overall instance trusts that can be already observed after only 1,000 epochs.

6.4.4 Trust of an object in another one of the MIoT

In this experiment, we applied the guidelines described in Section 6.4.3.1, but we focused on trust values and not on computation time. The average values of the trust of an object in another against the number of epochs for the MIoTs introduced previously is reported in Figure 6.7. From the analysis of this figure, we can observe that the trend of the trust values for objects is analogous to the corresponding one for instances, discussed in Section 6.4.3.4. Actually, by carefully examining Figures 6.5 and 6.7, we can observe an "extremization" of some phenomena. For instance, in small MIoTs, the object trust is constantly equal to about 1. Furthermore, when the MIoTs are medium or large, the trust values increase very quickly until to 1,000 epochs. This increase is still significant from 1,000 to 3,000 epochs. Finally, it becomes very small after 3,000 epochs.

This conclusion can be also extended to the distribution of the object trust values, reported in Figure 6.8, for the usual MIoT with 300 instances. As for the analysis of the standard deviation, we obtained that, after 1,000, 2,000 and 3,000 epochs, its values are 0.1911, 0.1093, and 0.0716, respectively. We can observe a rapid decrease when passing from 1,000 to 2,000 and from 2,000 to 3,000 epochs. This is an indicator of the stability of the obtained values for object trust.

Observe that, analogously to Figure 6.5, also in Figure 6.7 the performance of our approach depends on the number of epochs and the number of instances. With regards to these two parameters, the same reasonings we have proposed for Figure 6.5 can be applied to this figure.

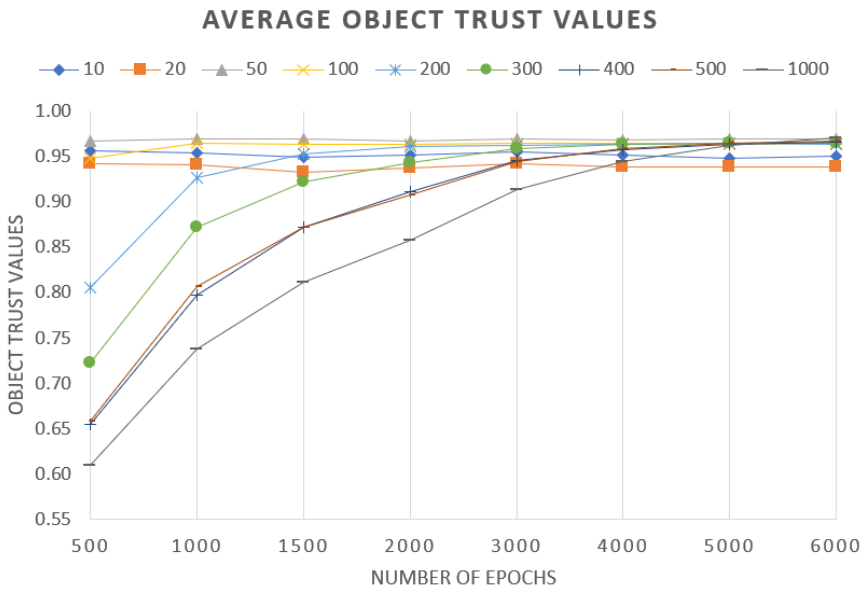


Fig. 6.7: Average values of the trust of an object in another one of the MIoT against the number of epochs for MIoTs with a different number of nodes

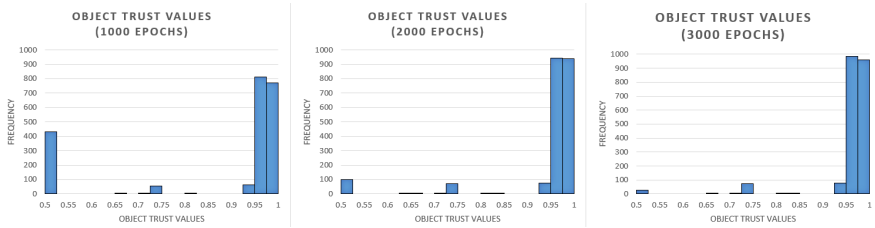


Fig. 6.8: Distribution of the trust of an object in another one of the same IoT after 1,000, 2,000 and 3,000 epochs for the MIoT with 300 instances

6.4.5 Reputation of an instance in an IoT, of an object in the MIoT and of an IoT in the MIoT

In order to investigate the variation of the reputation of an instance in an IoT against the number of epochs we applied the guidelines described in Section 6.4.3.2. The corresponding results are reported in Figure 6.9.

From the analysis of this figure, we can observe that the trend of the reputation values shows a continuous (even if slow) increase against the number of epochs. This can be explained by observing that, analogously to what happens to communities of people, as time passes and the number of transactions increases, objects tend to trust each other. As a consequence, the number of failed transactions decreases, which leads to an increase of the reputation of the objects performing them. Observe that the reputation value is higher for smaller networks. This reflects a general trend also observed in social networks of humans and, more in general, in communities of

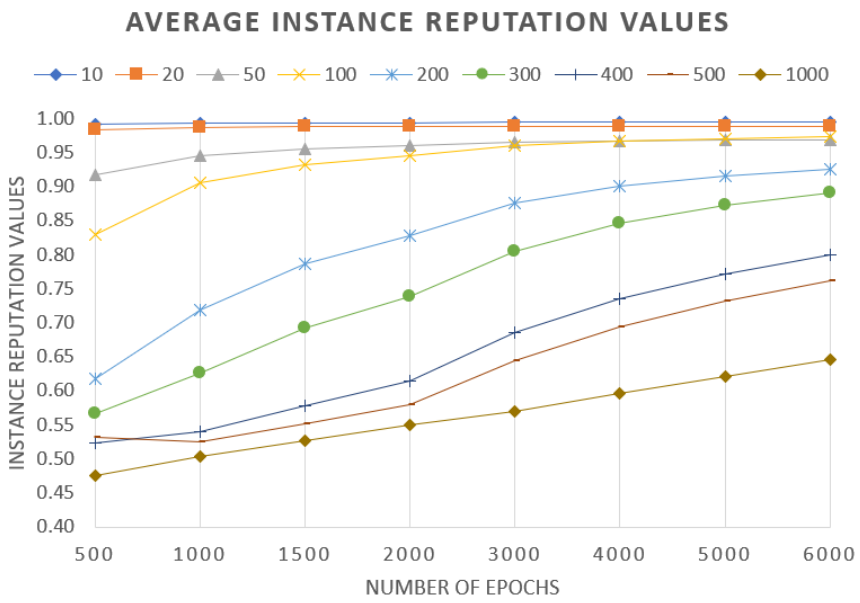


Fig. 6.9: Average values of the reputation of an instance in its IoT against the number of epochs for MIoTs with a different number of nodes

people. In fact, in a small community, the corresponding members tend to trust each other more.

The distribution of the corresponding values, for the usual MIoT with 300 instances, is reported in Figure 6.10. This figure represents a further confirmation of what we observed in Figure 6.9. Indeed, we can note that the shape of the distribution does not significantly change over time, but, as the number of epochs increases, the distribution values move to the right. This phenomenon is much more evident when passing from 1,000 to 2,000 epochs than when passing from 2,000 to 3,000 ones.

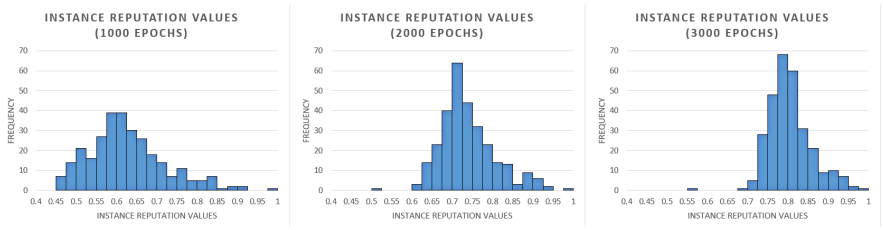


Fig. 6.10: Distribution of the reputation of an instance in its IoT after 1,000, 2,000 and 3,000 epochs for the MIoT with 300 instances

Finally, the values of the standard deviation of the instance reputation after 1,000, 2,000 and 3,000 epochs are 0.0950, 0.0682, and 0.0535, respectively. This ex-

tremely low and quite constant values evidence that the results obtained are acceptable and stable over time.

A similar procedure can be applied to object reputation, whose values against the number of epochs for the usual MIoTs are reported in Figure 6.11, and whose value distributions for the MIoT with 300 instances are shown in Figure 6.12.

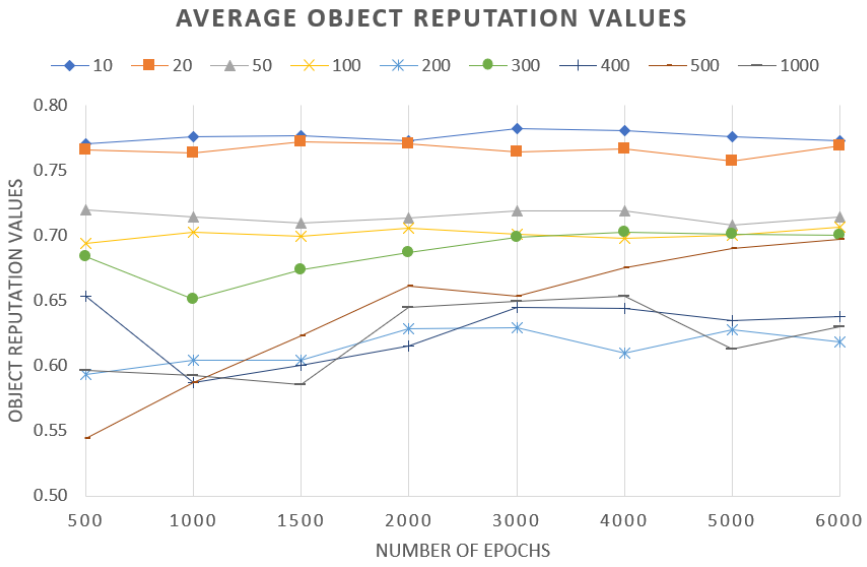


Fig. 6.11: Average values of the reputation of an object in its MIoT against the number of epochs for MIoTs with a different number of nodes

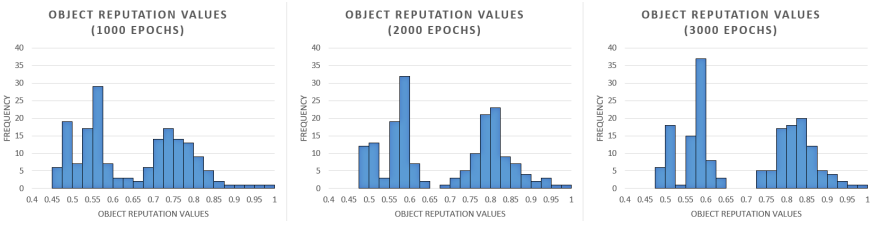


Fig. 6.12: Distribution of the reputation of an object in the MIoT with 300 instances after 1,000, 2,000 and 3,000 epochs

From the analysis of Figure 6.11, we can observe that the values of object reputation are often smaller than the corresponding ones of instance reputation, even if they are still acceptable. This is explained by the fact that object reputations refer to the whole MIoT and not to a single IoT, i.e., to a larger and more variegate scenario than the one characterizing the evaluation of instance reputations. In this context, it is clearly more difficult for an object to acquire and maintain trustworthiness.

We point out that the same observations and conclusions which we have drawn for Figures 6.5 and 6.7 can be extended to Figures 6.9 and 6.11.

The distributions of Figure 6.12, performed for the usual MIoT with 300 instances, confirm these observations. Indeed, in this case, we can observe that the distribution shape is roughly the same after 1,000, 2,000 and 3,000 epochs, but, differently from what happens for instance reputation values, it moves to the right only very slightly as the number of epochs increases. In other words, in this case, object reputation values show only a very small increase over time. The reasons are the same as the ones reported for Figure 6.11.

The value of the standard deviation after 1,000, 2,000 and 3,000 epochs are 0.1269, 0.1369 and 0.1403, respectively. These values are higher than the ones characterizing the standard deviation of instance reputation, even if they are still acceptable and quite constant over time. This evidences that the object reputation scenario is certainly more difficult to handle than the instance reputation one, even if it can be still maintained under control.

Finally, the reputation of an IoT in the MIoT is obtained by averaging the reputations of the objects having one instance in it. As a consequence, the corresponding values and distributions are very similar to the ones illustrated for the objects in the MIoT. Therefore, due to space limitations, we do not report them here.

6.4.6 Resilience

This experiment aimed at evaluating the robustness of our approach against the possible anomalies of the trust values assigned by an instance to another. We conducted it on the usual MIoT with 300 instances adopted for the previous experiment. In particular, we assumed the average value of the trust of an instance in another of the same IoT against the number of epochs for the MIoT with 300 instances (shown in Figure 6.5) as the "ground truth", i.e., as the case with no anomalies. After this, we considered two possible extreme anomalies. The former assumed that a fraction X% of instances constantly assigns a trust equal to 1 to all the other instances, independently of exchanged transactions, and all the transactions regarding these instances are reported as successful, independently of their real result (we call them "positive anomalies" in the following). The latter assumed an opposite behavior; therefore, it assumed that, independently of exchanged transactions, a fraction Y% of instances associates a value of 0 with the trust in all the other instances and all the transactions concerning these instances are reported as failed, independently of their real results (we call them "negative anomalies" in the following). We computed the average values of trust against the number of epochs for several fractions of positive or negative

anomalies (namely, 5%, 10%, 15%, 20%, 30%). The results obtained are reported in Figures 6.13 and 6.14.

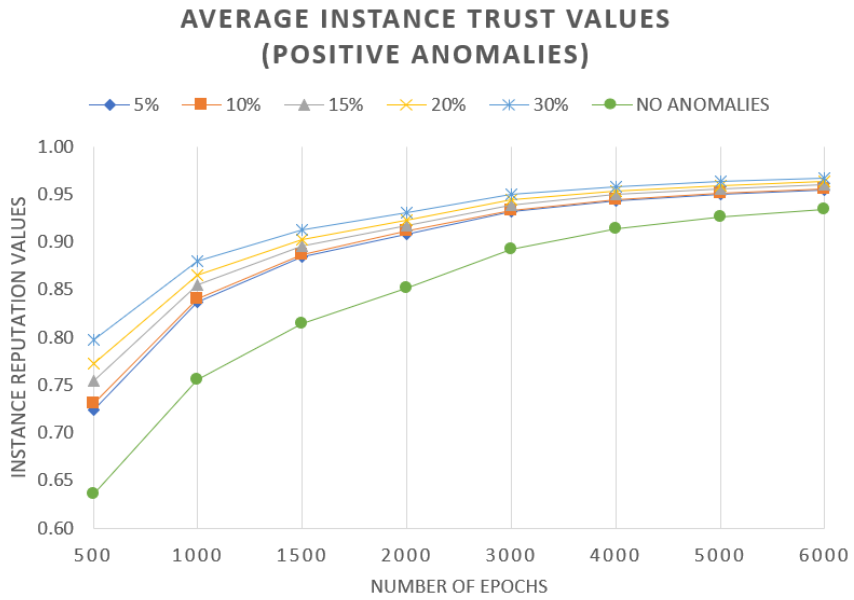


Fig. 6.13: Average values of the trust of an instance against the increase of positive anomalies for the MIoT with 300 instances

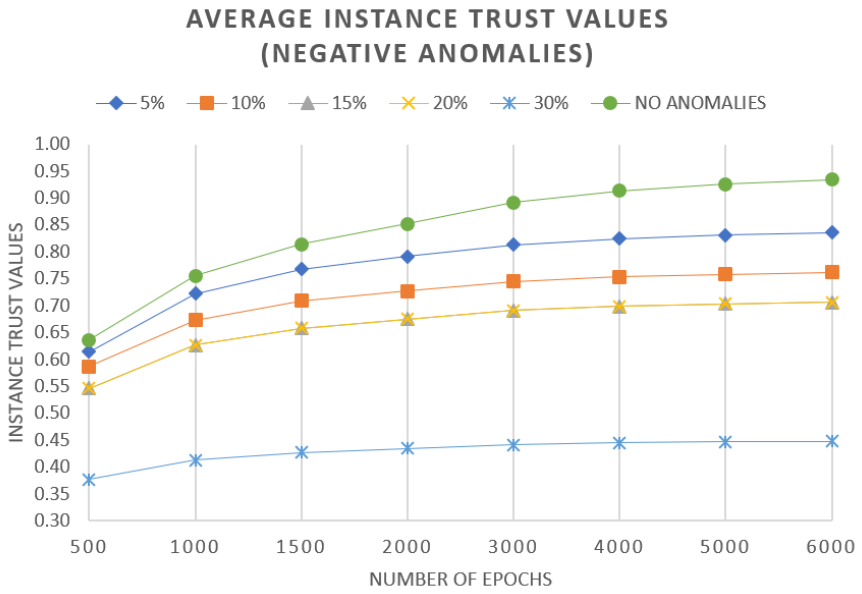


Fig. 6.14: Average values of the trust of an instance against the increase of negative anomalies for the MIoT with 300 instances

First, let us consider Figure 6.13, which refers to positive anomalies. From the analysis of this figure, we can observe that our approach is very resilient to this kind of anomaly. For example, the presence of 20% of positive anomalies leads to an increase of the trust values ranging from 10.24% at 500 epochs to 1.29% at 6,000 epochs.

After having examined positive anomalies, we analyze negative ones. They are reported in Figure 6.14. From the analysis of this figure, we can observe that our approach is sensitive to them. For instance, the presence of 20% of negative anomalies leads to a decrease of the trust values ranging from 22.06% at 500 epochs to 25.75% at 6,000 epochs, which is much higher than the corresponding one seen for positive anomalies. Even more interesting, when the fraction of negative anomalies reaches 30% of the MIoT instances, we can observe a strong fall of the trust values. In fact, its decrease ranges from 46.16% at 500 epochs to 52.87% at 6,000 epochs, which implies that the behavior of our approach is no longer acceptable.

The overall analysis of positive and negative anomalies allows us to conclude that our approach is very resilient to positive anomalies; perhaps, it is excessively resilient to them when they become high. An opposite behavior can be observed for negative anomalies. Our approach allows users to find them very easily; however, it is excessively sensitive to them when they are few.

An analogous reasoning can be drawn for the resilience of our approach to compute the reputation of an instance in an IoT. Analogously to what we have done for trust, we considered the average values of the reputation of an instance in its IoT against the number of epochs for the MIoT with 300 instances (shown in Figure 6.9) as the "ground truth", i.e., as the case with no anomalies. After this, we operated in the same way as we had operated for trust. In this case, the variation of the average reputation value against the number of epochs in presence of positive (resp., negative) anomalies is reported in Figure 6.15 (resp., 6.16). We computed it for several fractions of positive (resp., negative) anomalies (namely, 5%, 10%, 15%, 20%, 30%).

From the comparison of Figures 6.13 and 6.15, we can observe that, in presence of positive anomalies, the trend of the resilience for the computation of reputation is very similar to the one regarding the computation of trust; in this case, the increase of the average reputation is lower, as it ranges from 6.24% at 500 epochs to 1.07% at 6,000 epochs.

Analogously, by comparing Figures 6.14 and 6.16, we can observe that, in presence of negative anomalies, the trends of the resilience for the computation of reputation are similar (even if more mitigated) to the one regarding the computation of trust. Indeed, the decrease of the average reputation ranges between 14.17% at 500 epochs and 29.37% at 6,000 epochs. Interestingly, in this case, the fall of the reputa-

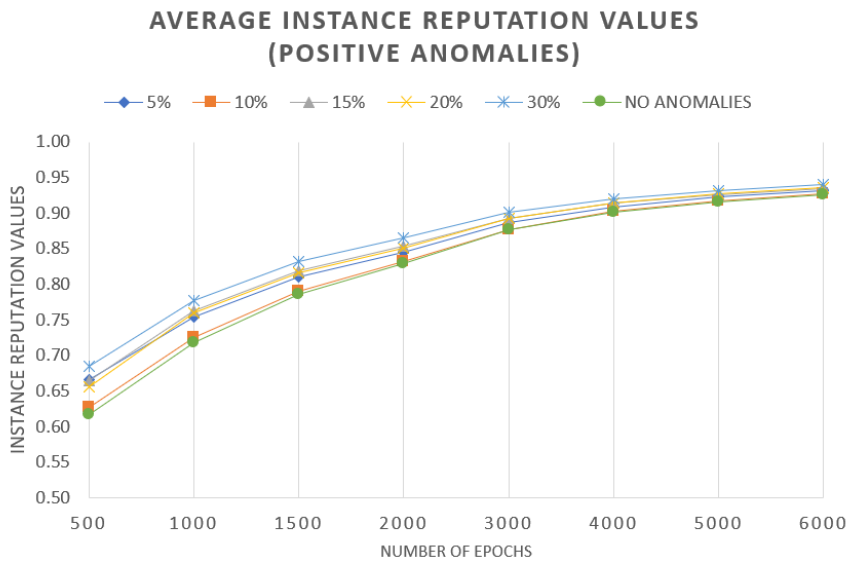


Fig. 6.15: Average values of the reputation of an instance against the increase of positive anomalies for the MIoT with 300 instances

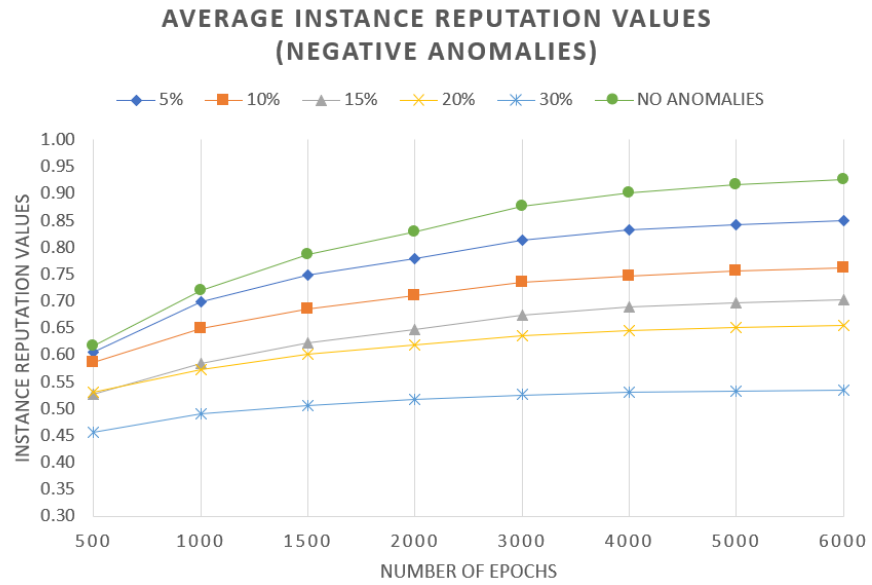


Fig. 6.16: Average values of the reputation of an instance against the increase of negative anomalies for the MIoT with 300 instances

tion values, when the percentage of negative anomalies passes from 20% to 30%, is less than the corresponding one observed for the trust values.

6.4.7 Accuracy

In order to measure the accuracy of our approach, we needed a ground truth regarding the trustworthiness of the smart objects involved in the MIoT. Unfortunately, the dataset used in the previous experiments did not have this information. As a

consequence, we had to construct a new dataset. This was obtained by drawing inspiration from the smart city scenario described in Section 6.5.1. In particular, we asked 30 students of our university, 15 males and 15 females, to wear a smartwatch and run in three different parks of our town. The first was near the city center; the second was in a suburb; the third was in a naturalistic area near the sea. In each park we put several smart sensors capable of measuring temperature, humidity and light intensity. During the run of each student in each park, her/his smartwatch communicated with the park sensors to evaluate the environmental quality of the park. Through these communications, the students' smartwatches and the park's smart sensors could interact with each other to evaluate their mutual trust. At the same time, the smart sensors in each park communicated with each other and, thanks to these communications, it was possible to measure the trust of a smart sensor in the other ones of the same park. All these trust values contributed to the computation of the reputation of each sensor of the park.

The interested reader can find this dataset at the address `http://daisy.dii.univpm.it/miot/datasets/trustReputation` clicking on the link regarding this section.

The distribution of the average reputation $\widehat{R_a}$ of all park sensors provided by our approach, against the number of exchanged transactions, is reported in the left part of Figure 6.17. This figure shows that the average reputation is quite high, and this result was actually not surprising taking the previous experiments into account.

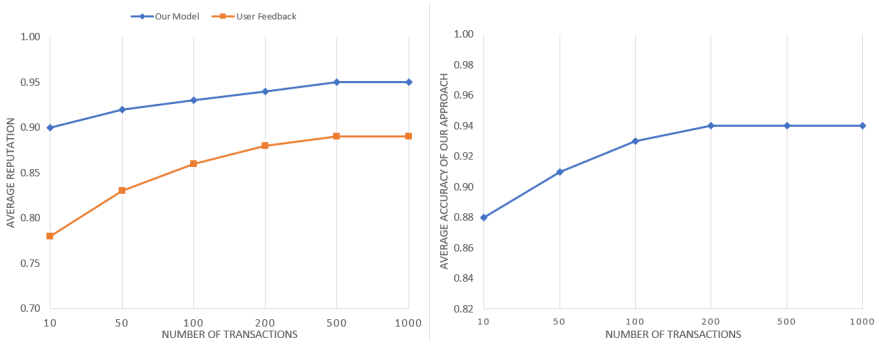


Fig. 6.17: Accuracy of our approach

In order to have the ground truth, we asked each student to provide her/his evaluation of the information provided by each park sensor. To our surprise, we observed that this evaluation was not constant and grew over time. We attributed this to the fact that, as time went by, the runner gradually adapted better to the environmental conditions of the park where she/he was running and, therefore, was more capable of objectively evaluating the information provided by sensors. The

average reputation $\widehat{R_r}$ of all park sensors provided by the runners is reported at the left of Figure 6.17.

At this point, we were able to compute the average accuracy $\mathcal{A}$ of our approach. Specifically:

$$\mathcal{A} = 1 - |\widehat{R_a} - \widehat{R_r}|$$

The accuracy values obtained by our approach are reported at the right of Figure 6.17. From the analysis of this figure, we can observe that: *(i)* the accuracy values are always very high; *(ii)* they initially tend to increase over time; *(iii)* after an initial phase, they tend to become very stable and very high. These results allow us to conclude that the accuracy of our approach is certainly very satisfying.

6.5 Use cases

6.5.1 Trust and reputation in a smart city

As a first example case, consider some public areas (such as parks, squares, shopping centers, etc.) in a smart city, and assume that a group of people actively visits them. Each area is equipped with several smart objects for monitoring weather, air quality, traffic conditions, level of noise, etc., along with several actuators, such as smart lamps or information hubs provided as online services. Each person may have several smart devices, such as smartwatches, smartphones, other wearable devices, and so on. People and places can interact with each other through their smart objects [274].

Such a scenario can be modeled through a MIoT $\mathcal{M}$ consisting of a set $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$ of IoTs, each representing a public area. The set of the objects of $\mathcal{M}$ comprises the smart objects in the public areas and the set of personal devices of people visiting them. If an object o_j of the MIoT is active in the k^{th} public area, it has an instance ι_{j_k} in the IoT $\mathcal{I}_k$. Clearly, when a person with a smart object o_j moves around different public areas, corresponding to different IoTs, o_j will have different instances, one for each IoT.

Each visitor of an area is generally interested in a certain kind of activity; for instance, she could be a fitness runner. The final goal of the MIoT is supporting people to get the best experience from their activities. In this setting, trust and reputation can play a key role in reaching this objective. In the following, we report some possible usage scenarios.

Assume that a person wants to go out for a run. First, she needs to choose the best area for the run, based on weather conditions, traffic and other parameters that she considers relevant. To carry out her choices, she can check data provided by the

sensors of each public area of her interest, the information hubs or other trusted runners. The choice of the information sources to consult is usually related to the trust and the reputation of the smart objects present therein. Once a person has performed her choices, she can decide to send this information to the MIoT in order to serve, in her turn, as information provider for the community.

A similar activity flow may happen in several other circumstances in which there is a decision to make, e.g., when a user must choose the best shopping center where she can buy a given object, the best cinema where she can see a movie, etc.

In all these cases, data regarding the choices of a user can be coupled with those registered during the activities she performed as a consequence of these choices (e.g., data coming from personal smartwears) in order to confirm the correctness of the choice or, on the contrary, to alert the other users of the evaluation errors. For instance, imagine a scenario in which a person verifies that the weather was actually much colder than the sensor in the public area seemed to indicate. In this case, the trust of the person in the sensors of that area decreases. This could also lead to a decrease of the overall reputation of these sensors, thus influencing the decision of the other users. In particular, the reputation decrease of the smart objects of the public area determines how many users are impacted by the negative experience of the user and how much strong this impact is.

It is worth pointing out the relevance of the smart object reputation in this context. As a matter of fact, some smart objects of the MIoT could assume the role of reliable information hubs for the whole MIoT if their reputation is particularly strong and durable over time.

Trust and reputation may also have an important role in the detection and the management of possible anomalies characterizing one or more devices in the network. As an example, assume that a weather sensor in a public area is malfunctioning; in this case, all the objects relying on its data will be affected by this anomaly. First of all, this leads to a decrease of its trust and reputation. Furthermore, if one or more other trustworthy weather devices are present in the same area, they could help the whole MIoT to determine the sensor malfunction, to avoid the propagation of its effects and, finally, to repair it.

6.5.2 Trust and reputation in a smart shopping center

Another possible scenario, where trust and reputation play an important role, is a big shopping center consisting of several buildings, each of them dedicated to specific product typologies, such as food, clothing, do-it-yourself, electronic devices, and so on. In this context, smart devices can be modeled by a MIoT $\mathcal{M}$ consisting of a set $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$ of IoTs, one for each building. The set of the objects of $\mathcal{M}$

consists of the set of the smart sensors present in each building (including video surveillance, temperature sensors, fire sensors, presence sensors, etc.) and the set of personal devices of visitors (including smartphones, tablets, smartwatches, etc.).

Each object o_j that interacts with the ones of the k^{th} building has an instance ι_{j_k} representing it in $\mathcal{I}_k$. Clearly, when the owner of an object o_j, such as a smartphone, moves throughout the buildings of the shopping center, o_j will have different instances associated with the different buildings of the center.

Here, a smart system of the shopping center could push offers to the enabled customer devices based on proximity, past preferences, habits, and so on. Analogously, based on the knowledge provided by the smart objects and the sensors dispersed in the shopping center, a personal device can suggest its owner the most comfortable and promising places to visit during her stay in the shopping center.

In this scenario, each person connected to the MIoT is interested in a certain kind of activity, somehow related to shopping. Indeed, users can play several roles ranging from vendors, suppliers or customers.

While a customer visits the building of a shopping center, her device may constantly locate the nearest ones and query for interesting products or offers. In the meantime, it could query other customers' smart objects (for instance, wearable devices) to measure her vital parameters in order to evaluate her pleasure in checking the products of a shopper. This can represent feedback information that the device supplies to the MIoT. Furthermore, a personal device of a customer can act as a personal shopper providing her with suitable suggestions. It interacts with the other objects of the MIoT, considers the offers of the shops, elaborates this information through machine learning algorithms, makes some proposals to its customer, registers her feedback and transmits them to the other devices in order to improve the quality of its recommendations.

Assume, now, that a customer wants to go out for shopping. First, she needs to locate the best building to start with. This activity can be carried out by contacting her devices that act as personal shopper or by checking the preferred destinations of "special" customers (for instance, the most influential ones) or, again, by detecting the most comfortable shops. Once the desired knowledge has been obtained, the device can process it to make its suggestions. After the customer has made her choices and has performed her shopping activities, she can share information about her experience. In this way, she and/or her devices can become information providers for other customers. In this scenario, trust and reputation play an important role. For instance, the reputation of each smart object determines how many devices (and, ultimately, how many people) it can influence and how strong its influence is.

As in the previous scenario, an important issue to investigate and address is the presence of possible anomalies. The impact of an anomaly depends on several factors; the reputation of the affected objects is certainly one of the most important of them. As an example, given an anomaly of the device acting as a personal shopper, for instance the loss of historical data on product prices, the corresponding suggestions might not be the most convenient ones for its owner. In this case, the anomaly will certainly have a high impact on the device's owner. Furthermore, it can have an impact, even if smaller, on all the other objects (and, ultimately, on the corresponding customers) that it can reach and influence. The extension and the strength of the impact of an object o_j on an object o_q depend on the value of the trust of o_q in o_j, on the overall reputation of o_j and on the decrease of the reputation of o_j.

6.6 Discussion

6.6.1 Considerations about the obtained results

The experiments have shown that our approach has an optimal resilience to the setting of weights and thresholds, even if it is (correctly) sensitive to weight errors when these become high (see Section 6.4.1). The results discussed in Section 6.4.1 also make us confident that the values obtained for weights are general and valid independently of the application environment in which our approach is operating. The experiments described in Section 6.4.3 revealed us that the time necessary to compute trust and reputation values is acceptable in all real cases. There are some theoretical situations in which this time could become unacceptable, but these cases are very far from the current real ones. Certainly, in the future, with the enormous development of IoTs, they could become possible, but we are confident that, in the meantime, the computation power of servers will simultaneously increase. In any case, we have specified some countermeasures that could be taken to face this problem, if it will happen in the future.

Section 6.4.3 revealed that it is possible to define a tradeoff between computation time and accuracy in determining the value of trust and reputation. Specifically, if computation time is the main factor to consider in this last activity, the number of epochs adopted to evaluate trust and reputation should not exceed 2,000 (which is a really huge number in the current real settings). If this number is not exceeded, there is no tradeoff to perform and, therefore, no need to sacrifice accuracy over computation time. By contrast, if the number of epochs is higher than 2,000, and there are more than 500-700 instances in the MIoT, in order to maintain an acceptable computation time, it is necessary to perform some actions that lead to partially sacrificing accuracy over computation time.

Section 6.4.6 shows that our approach is very resilient to positive anomalies and quite resilient to negative ones.

Last, but not the least, Section 6.4.7 reveals that: *(i)* the accuracy of our approach is always high; *(ii)* it tends to increase over time; *(iii)* after an initial phase, it tends to become very stable and high.

6.6.2 Possible usage of the extracted knowledge from a practical perspective

In Section 6.5, we have described two motivating examples, which illustrate two possible scenarios that could benefit from the approach presented here. The former regards a smart city scenario and describes how people can use the data exchanged by their smart objects and the ones of the city to improve the effectiveness and the efficiency of their activities. The latter concerns a smart shopping center and illustrates how customers can use the data exchanged between their smart objects and the ones of the shopping center to improve the effectiveness and the efficiency of their shopping activities. However, these are only two of the large amount and variety of scenarios that could benefit from our approach. Think, for instance, of the adoption of smart objects to best regulate transports, to improve predictive maintenance in manufacturing, to regulate the patient flow to a hospital during a health emergency, to regulate the visitor flow to an exposition, and so forth.

6.6.3 Generalization level of results from a practical point of view

Throughout the discussions on the figures presented in Section 6.4, we have seen that most of the experiments are general (think, for instance, of the ones for setting the values of α, β and ρ presented in Section 6.4.1) and do not depend on the application environment our approach is operating on.

Furthermore, in Section 6.3 and in the next ones, we have observed that the limits on the computation time that we could find in our approach regard theoretical cases that are currently very over-dimensioned w.r.t. the real scenarios. In fact, all current real scenarios are fully manageable by our approach, which can be considered general and not limited to some specific scenarios.

The generality of our approach, and its applicability to all real cases, also regard its resilience (as witnessed by the results and the discussion of Section 6.4.6) and its accuracy (as witnessed by the results and the discussion of Section 6.4.7).

All the reasonings above make us confident that our approach can be a precious support in a large variety and amount of practical situations, as the ones described in Section 6.5 and the other ones mentioned above.

6.6.4 Similarities and differences between communities of people and communities of objects

Sensors and devices are becoming increasingly smart. The amount of data that they can store and the computing power at their disposal are constantly increasing. If, in the past, they were passive entities, without any autonomy, currently they have become increasingly active.

In this scenario, it is not surprising that, for some years, researchers have started to discuss about Social Internet of Things and, if the social relationships investigated by them in the past were extremely simple and elementary, the ones analyzed in the current researches are increasingly rich, complex and variegate.

Smart objects start to have a profile and to show a behavior obtained by implementing artificial intelligence-based algorithms on them. As a consequence, the boundary between what can be done by communities of people and communities of objects becomes increasingly blurred [650].

Clearly, in this discussion, we are considering only the technical viewpoint. However, when we discuss on Social Networking and Social Network Analysis (both if they are applied to humans and if they are applied to smart objects), we must consider that there is also another viewpoint, more related to humanistic and sociological studies. It concerns the investigation of the intrinsic essence distinguishing humans from animals and humans from machines. As for this aspect, we think that the gap between humans and smart objects is still enormous and, in our opinion, it will never be fully filled. But, here, we would open a discussion which is not object of this chapter.

7

Privacy and Security

In this chapter, we propose a privacy-preserving approach to prevent feature disclosure in a MIoT scenario. Our approach is based on two notions derived from database anonymization, namely k-anonymity and t-closeness. They are applied to cluster the involved objects in order to provide a unitary view of them and their features. Indeed, the use of k-anonymity and t-closeness makes derived groups robust from a privacy perspective. In this way, not only information disclosure, but also feature disclosure, is prevented. This is an important strength of our approach because the malicious analysis of objects' features can have disruptive effects on the privacy (and, ultimately, on the life) of people.

The material present in this chapter is taken from [508].

7.1 Introduction

In the last few years, we are assisting to the enormous increase of the number of sensors and devices, which are becoming extremely pervasive and used in most contexts of daily life. At the same time, objects are developing awfully smart and social skills. All these aspects are revolutionizing the Internet of Things (hereafter, IoT) [711]. As a proof of this, more and more researchers are beginning to study the behavior of things, to talk about their profiles and their social interaction [213], and to manage objects almost as if these were humans. As a result of these investigations, several architectures implementing these ideas have been proposed, and are currently being proposed, in the literature. Social Internet of Things (hereafter, SIoT [70]), Multiple IoT Environment (hereafter, MIE [81]) and Multiple Internets of Things (hereafter, MIoT [82]) are only three of the latest architectures with these characteristics.

Such an evolution of the IoT scenario puts researchers in front of several issues that can become important opportunities if correctly addressed. A major example is the huge interest the researchers have shown in security and privacy in IoT. Indeed, in the recent years, many approaches to the definition of security solutions in the context of smart objects have been proposed, such as solutions for intrusion detec-

tion [45, 522], access control [34, 432] and privacy [36]. In the context of privacy in IoT, one of the most relevant challenges regards the capability of preserving the privacy of users, who are employing a set of smart objects connected with each other and, possibly, with objects belonging to other users. In such a scenario, characterized by the pervasive presence of smart objects, a lot of user's data can be produced by the smart objects she is using. This scenario appears even more complex if we consider that objects are becoming increasingly autonomous when they perform their tasks. Among these, one of the most important and crucial for the whole IoT is the interaction with other objects. In order to refine and improve this capability, objects may use and propagate information about the features they can provide. This information allows other objects to improve the selection of the preferred contacts and to enhance their querying capability. However, if properly combined with other data, it can provide sensitive information about the user, which she had no intention of disclosing. Knowing the features of more objects adopted by users, the amount of sensitive knowledge about her that can be derived dramatically increases.

To give an example of what we stated above, let us consider a scenario in which a person is in a hospital because she is suffering from gastrointestinal disorders. To carry out diagnosis, she must undergo several analyses in different departments of the hospital. The simplest and fastest of these analyses can be performed through smart objects. For example, in one department, the patient could be connected to an insulin meter, in another one she could be connected to a heart rate meter, and so forth. Knowing that a patient is connected to a specific device (for instance, the insulin meter) already discloses important sensitive information about her (in particular, that she could suffer from diabetes or some pancreatic disease). Knowing also that she is connected to more devices that are simultaneously used for the test of the condition of a specific organ (for example, the devices used to diagnose pancreatic disorders, such as diabetes, pancreatitis or pancreatic cancer), the amount of sensitive information disclosed becomes much more serious because we know in detail what are the possible diseases that doctors suspect may affect the patient.

As a second example, let us consider a patient that simultaneously undergoes three tests, one for the measurement of blood sugar, one for measuring the level of hemoglobin in the blood and one for measuring her respiratory function. Just knowing that she is carrying out only one of these tests, we can hypothesize several diseases from which she may suffer (for example, we may hypothesize that she is using glucose meter because she is suspected of suffering from diabetes). But if we know that she is carrying out these three tests simultaneously, we might conclude that doctors suspect she might have lung cancer, considering that some forms of lung cancer involve important variations in blood sugar and hemoglobin.

Here, we aim at addressing this issue by proposing a privacy-preserving approach to prevent feature disclosure in an IoT scenario. Our approach is not focused on specific queries. Instead, as said before, it aims at preventing the disclosure of sensitive information of a user that can happen simply by examining the features of the devices she is employing. Taking also into account that utility and privacy is a major trade-off for privacy-preserving techniques, our approach aims at preserving all existing information about user-object interaction. In fact, this information is extremely useful to support other applications and possible analyses on an IoT scenario. On the other hand, our approach is capable of protecting users' privacy by partially hiding objects' features still allowing their full exploitation in order to support objects' communication.

In more details, our approach leverages some traditional concepts from databases, such as k-anonymity [633] and t-closeness [419]. The basic idea of both these paradigms is to group data together so that the same piece of information is present in at least k records. This creates a sort of blurred cloud of data, in which it is not possible to successfully map the protected piece of information to a specific record among the k sharing it. Of course, when dealing with data distribution, it is possible to reduce the number of candidate records to be associated with a specific feature by exploiting the probability that a record contains that piece of information. The t-closeness paradigm overcomes this possibility by imposing criteria based on the probability distribution when selecting the admissible values used to k-anonymize a sensitive piece of information.

Our approach applies k-anonimity [633] and t-closeness [419] to build small conglomerates, hereafter referred as *groups* of objects, inside an existing network with the purpose of creating a single view of the objects present in each of them. The individuality of smart objects is preserved from a connectivity point-of-view, whereas their features are mixed inside each group. From the outside, a smart object presents itself by advertising the features available in the group it belongs to. Groups are built by solving a trade-off between privacy requirements and communication performance. k-anonymity and t-closeness are combined to make each group robust from a privacy perspective by properly selecting the number of features, their typology, and the number of objects as tuning parameters in order to meet the desired protection level.

Our approach is orthogonal to the existing strategies for the protection of communication channels and data exchange among objects, such as the ones described in [258, 567, 440, 238, 675]. Moreover, while many researchers have been developing frameworks to protect *object* interaction from both a security and privacy perspective, our approach focuses on the effects produced on the privacy of the *users* by the

direct observation of the objects (and the corresponding features) they are employing. As a matter of fact, with the evolution of smart objects, techniques to allow the automatic interactions among them based on proximity or homogeneity have been developed [304]. As stated above, such strategies can be improved by using object scopes and features; therefore, enabling feature advertising is an important point and a key aspect for improving object interactions in the IoT. This consideration, combined with the observation that the knowledge of object features is an important vehicle to privacy leakage, leads to the need of a stable solution that enables these interactions in a privacy-preserving way.

Our proposal refers to such a scenario and presents a solution in this setting. In its design we also take into account the most recent developments on IoT research. It has been proved that it is more realistic to model an IoT scenario as a set of connected networks, instead of only a unique network of objects. This is due to the number of involved objects, their smartness and social interaction capabilities, as well as the possibility that each portion of the object network may desire to hide part or most of data exchanged inside it [82]. The usage of a multi-network representation of our scenario is a key point in our proposal. Indeed, *(i)* each identified group corresponds to a network of the system; *(ii)* each object can be modeled by means of a node; *(iii)* relationships between objects of the same group can be represented by means of arcs inside the corresponding networks (they are called inner arcs); *(iv)* relationships between objects of different groups are modeled as arcs linking nodes of different networks (they are called cross-arcs). The possibility to have a direct, natural and immediate multi-network representation of our scenario allows for an easy mapping with properties, operations and concepts of multi-network contexts [134, 142, 434].

The outline of this chapter is as follows. In Section 7.2, we examine related literature. In Section 7.3.1, we describe the proposed model in detail, whereas in Section 7.3.2, we illustrate our privacy-preserving object grouping scheme. In Section 7.3.3, we describe our security model. Finally, in Section 7.4, we propose a discussion about the peculiarities of our approach.

7.2 Related Work

Like all the areas of networked computing, the IoT presents particular challenges to security and privacy, due to the interconnected nature of the Internet. It means that Internet resources can be attacked from everywhere at every moment. The threats that can affect IoT entities are numerous, such as attacks targeting communication channels, physical threats, denial of service, identity fabrication, and so on [75]. This has led several researchers to develop countermeasures for addressing security and

privacy issues specific to the IoT [12, 258, 656, 567, 521]. In particular, in [12], the authors present an overview of security principles, as well as of technological and security challenges; then, they propose countermeasures for securing the IoT. One of the main challenges in this research field is that proposed solutions must cope with the restrictions and limitations in terms of components, devices, computational and power resources characterizing the IoT [35]. On the one hand, the pervasive nature of this technology provides its users with more opportunities to enhance their interactions and to have access to advanced features fostering the creation and consolidation of social relationships. However, on the other hand, it poses new severe technical challenges [99, 72, 597, 39].

Many researchers have adopted Blockchain based strategies to overcome resource availability in the IoT and to propose solutions to privacy and security issues [238, 172, 559, 602]. Specifically, in [238], the authors propose an approach using Blockchain to build a decentralized security and privacy-preserving model. This approach has been thought for smart-home scenarios, in which there is the possibility of having a dedicated high-resource device playing the role of miner. The approach described in [172], instead, uses Blockchain to build a network of gateways, to which smart objects can connect. In this way, even though older devices can be not equipped with resources necessary to implement security and privacy-preserving protocols, they can communicate through the gateway network to overcome their limitations. A further step towards the protection of privacy in the IoT is described in [559]. Here, the authors address data confidentiality in the IoT by combining Attribute-Based Encryption (ABE) with Blockchain to achieve integrity, non-repudiation and confidentiality in IoT communications. Another interesting idea in this context is the one described in [602], in which an approach to build SVM models using data from the IoT, but preserving user privacy, is provided. To reach its goal, this approach uses a Blockchain-based solution in which data collected by smart sensors are first encrypted by means of a homomorphic cryptosystem. Then, each sensor shares encrypted data by using Blockchain as distributed public ledger. Finally, a modified SVM algorithm working on encrypted data is adopted to train a classifier using such data.

Still in the context of data protection in the IoT, many researchers propose applications leveraging Fog Computing. For instance, in [440], the authors describe an approach to protect privacy of users when data aggregation strategies leveraging Fog Computing delegation are adopted. The peculiarity of this approach, with respect to other well known solutions, such as those described in [441] and [605], relies on the capability of aggregating data from heterogeneous smart devices in a privacy-preserving way. The importance of investigating privacy and security issues when

delegating IoT services over Fog Computing solutions is discussed in [409] and [46]. Both these papers provide evidence of the high-impact issues brought about by the adoption of Fog Computing to improve IoT operability.

Other works focus on data confidentiality, i.e., on the objective that data is secure and available only to authorized users. In [258], the authors present an architecture for the IoT security, caring that sensors do not reveal collected data to neighboring nodes. They assure data confidentiality through encryption technologies, which prevent data stealing threats. Furthermore, the authors of [567] focus on how data will be managed, stating that, to ensure protection throughout the process, there must be policies on how to manage several kinds of data, as well as some policy-enforcement mechanisms.

Even though our approach shares some common aspects with the proposals described above, its objective is different. Indeed, most of the approaches above aim at protecting data and avoiding unauthorized access to it. To carry out this task, they operate on the communication channel among objects; some of them also provide facilities to perform privacy-aware data aggregation. Our proposal can be considered as an application on top of existing and consolidated strategies to obtain security and confidentiality in the physical communication channel among objects. Indeed, it focuses on a scenario in which objects directly advertise their capabilities and features (by using existing technologies to interact with other objects in a secure way) to foster the creation of new links in the network. Feature advertising is very common in networking as it is used by the network administrator to detect services running on a device, along with the corresponding versions. This strategy can be also investigated to improve the IoT by means of UPnP scans, through which objects can exchange their descriptions as a response to an HTTP request in an XML document, or by means of Banner Grabbing [79].

Feature description has been adopted in some application scenarios to improve the use of the IoT by exploiting the social-side of this network, in order to filter contents and contacts, thus evolving towards the concept of opportunistic IoT [313] and, therefore, to classify objects data and information for improving their interactions [60]. Also for these approaches, the knowledge of the features and the kind of information that an object can produce is a very important aspect and has been used in different applications, such as service discovery in the IoT [557].

It is worth underlying that, the impact to privacy of both service discovery and feature disclosure in the IoT has been already an important subject of study. Indeed, the recent scientific literature on the IoT includes numerous proposals of privacy protecting schemes in this context [685, 219, 568]. In particular, the authors of [685] illustrate a new approach to private authentication and service discovery

IoT	Internet of Things			SIoT	Social Internet of Things
MIE	Multiple IoT Environment			MIoT	Multiple Internets of Things
n_i	the i^{th} node			P_i	the profile of n_i
ϕ_i	the set of the features exposed by n_i			G_k	the k^{th} group
min_k	the minimum number of nodes of G_k			max_k	the maximum number of nodes of G_k
φ	a feature			NS_k	the set of the nodes of G_k
NS_k^P	the set of the nodes permanently associated with G_k			NS_k^T	the set of the nodes temporarily assigned to G_k
Φ_k	the set of the features exposed by G_k			WZ	the Welcome Zone
$\mathcal{M}$	a MIoT			N	the set of the nodes of $\mathcal{M}$
A	the set of the arcs of $\mathcal{M}$			A_I	the set of the i-arcs of $\mathcal{M}$
A_C	the set of the c-arcs of $\mathcal{M}$			$\mathcal{I}_k$	the k^{th} IoT of $\mathcal{M}$ corresponding to the group G_k
$\overline{\mathcal{I}}$	the IoT of $\mathcal{M}$ corresponding to the Welcome Zone			$\mathcal{G}_k$	a graph representing $\mathcal{I}_k$
N_k	the set of the nodes of $\mathcal{G}_k$			A_k	the set of the arcs of $\mathcal{G}_k$
σ_c	the score of the node n_c			π_c	the priority of the node n_c
τ_c	the time elapsed since n_c participated to its current group			i_c	the importance of n_c

Table 7.1: The main abbreviations used throughout this chapter

in the IoT. This approach ensures the mutual privacy for both the device delivering the service and the one exploiting it. It can also guarantee that the service is authentic (unforgeable service). In [219], the author proposes a solution to the problem of privacy-preserving service discovery and access control. This strategy is, then, successfully deployed in a smart-home scenario. Another interesting evaluation of the privacy and security flaws, when enabling distributed service discovery in the IoT, is presented in [568].

While all these approaches strive to protect the identity of both *the object* offering a service and the one receiving it, our approach focuses on a different privacy threat. Indeed, although, by adopting the strategies described in this section we could improve the security of object interactions and the protection of service delivery, an attacker can still have access to the basic information about which features and services are available. As explained in the Introduction, also this simple knowledge can lead to disruptive privacy threats as it can be used to infer information about the habit, behavior or status of the corresponding object owners. This is an important application-level privacy flaw that must be considered and faced, and, to the best of our knowledge, our approach is a first attempt in this direction.

7.3 Methods

7.3.1 Extending the MIoT paradigm

In this section, we illustrate the model that we adopt to represent and handle the actors operating in our approach. In order to increase the readability of this section and of the next ones, in Table 7.1, we report the main abbreviations used throughout this chapter.

Our model uses the following main concepts:

- *Node*. It represents a smart object and has a profile, which allows its interaction with other nodes in an anonymous way. The profile of a node consists of an identifier, which does not report information about the specific features of the object (in order to guarantee anonymity), and of the set of the features provided by the group it belongs to. A node has also associated all the information needed for the communication with other nodes (such as the MAC address, the IP address, etc.). Throughout this chapter, we will use the symbols n_i to denote a node and ϕ_i to indicate the set of the features exposed by it.

  Furthermore, since there is a biunivocal correspondence between a smart object and the corresponding node, in the following, we will use these two terms interchangeably.
- *Group*. It is a set of smart objects characterized by heterogeneous features to comply with the principle of t-closeness. A group has a minimum and a maximum number of nodes. In the following, we will use the symbols:
  - G_k, to denote the k^{th} group;
  - min_k and max_k, to represent the minimum and the maximum number of nodes of G_k;
  - NS_k, to indicate the set of the nodes of G_k;
  - Φ_k, to denote the set of the features exposed by G_k.

  In turn, NS_k consists of two subsets, namely:
  - NS_k^P, i.e., the set of the nodes permanently associated with G_k;
  - NS_k^T, i.e., the set of the nodes temporarily assigned to G_k.
- *Welcome Zone* (hereafter, WZ). It is a staging area where nodes are put during their startup phase, when they require to join our system. It can be seen as a special group of nodes in which no feature is exposed. Furthermore, it contains a reference to all the other groups operating in our system.
- *MIoT* (Multi-IoT, as described in Chapter 4). It represents the environment where smart objects operate and through which they exchange messages. From a physical viewpoint, a MIoT consists of a network of smart objects that can communicate with each other either directly (if there exists a direct link between them) or indirectly (if there is the need to pass through other intermediate nodes). The network handles two basic kinds of communication, namely:
  - *Point-to-point*: it consists of a private message between two nodes of the MIoT that cannot be accessed by any other node.
  - *Broadcast*: it consists of a public message delivered inside a group or inside the Welcome Zone that can be seen by all the corresponding nodes.

  From a logical viewpoint a MIoT can be modeled as a set of Internets of Things (hereafter, IoTs):

$$\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m, \overline{\mathcal{I}}\} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m, \mathcal{I}_{m+1}\}$$

Here, each IoT $\mathcal{I}_k$, $1 \leq k \leq m$, corresponds to a group, whereas $\overline{I} = I_{m+1}$ corresponds to the Welcome Zone. A graph $\mathcal{G}_k = \langle N_k, A_k \rangle$, $1 \leq k \leq m+1$, can be associated with each IoT of $\mathcal{M}$. Recall that, in a MIoT $\mathcal{M}$ there are two sets of arcs: A_I and A_C. A_I is the set of the inner arcs (hereafter, *i-arcs*) of $\mathcal{M}$; they link nodes belonging to the same group. A_C is the set of cross arcs (hereafter, *c-arcs*) of $\mathcal{M}$; they link nodes belonging to different groups and play an important role in our privacy-preserving protocol, as will be clear in the following. A node connected to at least one c-arc is called *c-node*; otherwise, it is called *i-node*. Actually, in our model, we can distinguish two main categories of c-nodes. The former refers to nodes that *temporarily* belong to a group G_k; indeed, just because they are not permanently assigned to G_k, they still continue to belong also to WZ[1]. The latter, instead, comprises nodes that have c-arcs towards nodes belonging to other groups.

As a final point, we observe that, while i-arcs are automatically built by our system once a group is formed, c-arcs are built by nodes. Specifically, c-arcs can be created either to connect a node of the WZ temporarily assigned to a group with the other nodes of this group, or to connect nodes belonging to different groups. Concerning this last aspect, it is worth underlying that, in our solution nodes can still interact with each other by using the classical strategies defined in the IoT literature, such as node proximity or node homogeneity [69].

7.3.2 Privacy-preserving object grouping scheme

The objective of our approach is to protect the privacy of the users of smart objects in a MIoT when feature advertising guides object interactions. As explained in the Introduction, to prevent privacy leakage, our approach borrows some concepts, namely k-anonymity [633] and t-closeness [419], from databases.

In our scenario, we implement these notions by creating groups of objects so that each object can participate to the MIoT by using the features of its group as a business card. Intuitively, any object can be a mean to reach the content available inside a group of objects if they can interact with each other. As a consequence, if all the communications happening inside the group are made anonymous, observers cannot know which nodes of the group can provide content related to a specific feature.

Our scheme consists of two main operation categories, namely *Node-level operations* and *Group-level operations*. The former includes the two fundamental actions

[1] Recall that, in our approach, WZ is modelled as an IoT of $\mathcal{M}$.

that a single smart object (i.e., a node in our model) can perform inside the MIoT, namely *join* and *leave*. The latter refers to operations performed by all the nodes of a group to preserve the MIoT liveness. In more details, it consists of the following actions: *Formation of a group*, *Remediation of a group* and *Resize of a group*.

As depicted in Figure 7.1, each node can enter our system by means of a join operation. Our system is equipped with a staging area, i.e., the Welcome Zone, in which nodes are welcomed. Nodes joining WZ send hello messages to advise other nodes of their presence in WZ.

To satisfy privacy requirements, we impose a minimum number of nodes in the WZ before group formation can start. When this constraint is satisfied (see Section 7.3.2.1 for further details), smart objects exchange messages about their features through the information delivery protocol proposed in Section 7.3.2.3. This was designed to guarantee the anonymity of the source of each available feature. A group can be formed if, in WZ, objects and their features comply with specific criteria. These are defined by taking both the k-anonymity and the t-closeness paradigms into account.

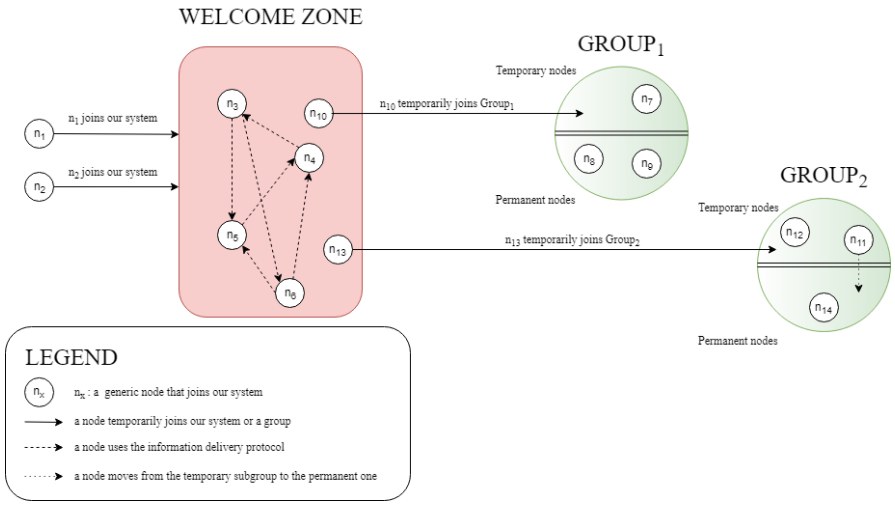


Fig. 7.1: Overview of our approach

Over time, new nodes can register to the system and join (even temporarily) existing groups or take part to the formation of new ones. Furthermore, a node can leave its current group and, eventually, the system. Once again, objects use protocol messages to communicate their intentions (e.g., leaving the current group); in this case, group-level operations (such as the remediation and the resize of a group) are triggered in response to them. These last operations have been conceived to manage the variation of the number of nodes inside groups over time.

As a final aspect, considering that rising messages with specific features as a subject can also lead to a privacy leakage, our approach provides a querying mechanism allowing for a privacy-preserving retrieval of information in such a complex system. It basically consists of two kinds of message, namely *Intra-group Query* and *Extra-group Query*, and of a communication protocol. Nodes can retrieve information from their group or from the MIoT network. The former task is achieved by using intra-group messages; the latter, instead, adopts special extra-group messages.

It is worth mentioning that group formation is only based on the arrival order of the nodes in WZ. Of course, this implies that a group can potentially contain heterogeneous nodes. However, the nodes of a group share a consistent number of features, because of the requirements of our privacy model. Anyway, the node homogeneity requirement is not crucial in our context; in fact, our objective is different and regards the creation of relatively small blurred clouds of nodes to protect the features exposed by each of them. From a technical point of view, the connections among nodes are handled by the MIoT, which provides the basic networking functionalities (private point-to-point communication and broadcast messages). Whenever a node joins the system, it actually registers its connectivity information (MAC address, IP address, etc.) to the MIoT. An important point is that we need to guarantee the possibility for nodes to directly interact with each other inside the group because we want to map each node to the features exposed in the whole group. For this reason, we impose the full connectivity of the nodes inside each group. Once again, all the communications (and, hence, the use of the corresponding connection links) is handled by the MIoT.

As a final point, group formation is the strategy adopted to implement our privacy model. However, we also preserve the original nature of an IoT by guaranteeing that nodes can still get in touch and interact according to existing strategies and links [304, 73, 74]. Indeed, as explained below, our solution also includes extra-group communication among nodes. Therefore, if two nodes are in proximity and, according to [304], a link can be established between them, two situations may happen, namely: *(i)* they belong to the same group and, hence, no further operation is necessary; *(ii)* they belong to different groups, in which case a *c-arc* will be created between them in such a way as to allow their (extra-group) communication.

In the next subsections, we provide a complete description of our protocol by examining node-level operations, group-level operations, and the delivery protocol in details.

7.3.2.1 Node-level operations

Node-level operations specify the tasks that a single node can perform in a MIoT. There are basically two operations, namely *join* and *leave*. We describe them in the next paragraphs.

Join of a node

A join operation is performed when a node n_i requires to join WZ or a group G_k of the MIoT.

In the former case, n_i sends a "hello message" (see Section 7.3.2.3) to the other nodes of WZ. These answer it by specifying the number ϵ of the nodes that already joined WZ without having communicated their features yet. As a matter of fact, in order to preserve the k-anonymity property, it is necessary that at least k new nodes simultaneously communicate their features. To reach this objective, ϵ is increased whenever a node joins WZ. When $\epsilon \geq k$, all the nodes in WZ communicate their features and ϵ is set to 0.

In case n_i joins a group G_k, it is necessary to distinguish two further subcases, namely permanent and temporary joins. The former represents the main form of membership of a node to a group; it is a stable situation in which the node can stay in the group and can participate to all the tasks involving the members of the group without time limitation, and, therefore, until a group no longer exists or the node spontaneously decides to leave the group. The latter, instead, has been conceived to face anomalous situations in which the conditions for the formation of new groups are not satisfied for a long time interval (this generally happens when there is a lack of a sufficient number of new nodes, see Section 7.3.2.2). In this case, the objects waiting in WZ are temporarily joined to existing groups if the features exposed by them make it possible. In this case, nodes can join groups but with some limitations (mainly related to the features they expose) until new groups tailored to their features can be built (see Section 7.3.2.2 for details about this operation). Specifically, a node can temporarily join a group if the intersection between the set of its feature and that of the group is not empty. It is worth underlying that, in this case, the node would conceal the additional features it may have with respect to the ones exposed by the group it is joining.

In case of a permanent join, n_i communicates the change of its state to the nodes of WZ so that they can remove it from their lists of contacts. In case of a temporary join, n_i simultaneously belongs to G_k and WZ. Indeed, in this last case, it still interacts with the nodes of WZ in order to create new groups or to participate to the remediation or to the resize tasks involving already existing groups (see Section

7.3.2). As a consequence, in this case n_i acts as a c-node, as pointed out in Section 7.3.1.

Leave of a node

A leave operation is performed when a node n_i requires to leave WZ or a group G_k of the MIoT. In the former case, it is sufficient that n_i informs the other nodes of WZ so that they will remove the arcs linking them to n_i. In the latter case, n_i must inform the nodes of both G_k and WZ, which will remove all the arcs linking them to it.

After this task, the process terminates if n_i is an i-node. On the other hand, i.e. n_i is a c-node, it is necessary to handle the arcs between it and the nodes of the other groups of the MIoT.

For each arc between n_i and a node n_l of another group G_q, two cases might happen:

- *the arc is recent and has been rarely used*; in this case, it can be removed;
- *the arc is old and has been frequently used*; in this case, it should be "inherited" by another node of G_k.

To distinguish these two cases, it is possible to introduce a parameter ρ measuring the relevance of an arc. ρ is defined as $\rho = \frac{\nu}{\lambda}$, where ν is the number of times in which the arc was used for a communication, whereas λ is the lifetime of the arc. If ρ is less than a threshold th_ρ, the arc can be removed; otherwise, it must be "inherited" by another node of G_k.

In this latter case, it is necessary to select the node that inherits the arc. For this purpose, first the set $CSet_k$ of the candidate nodes of G_k is determined. This set comprises all the c-nodes of G_k different from n_i. Then, each node n_c of $CSet_k$ must compute a score σ_c, which takes into account both its priority π_c and the compatibility σ_c between its features and the ones of G_q. Formally speaking:

$$\sigma_c = \omega \cdot \pi_c + (1 - \omega) \cdot J(\phi_c, \Phi_q)$$

Here, ω is a weight, belonging to the real interval $[0, 1]$, used to weigh the importance of priority against compatibility.

The priority π_c of n_c is a real number that takes into account the time τ_c elapsed since n_c participated to G_k and the importance ι_c of n_c in the MIoT:

$$\pi_c = \tau_c \cdot \iota_c$$

The value of ι_c belongs to the real interval $[0, 1]$ and is determined by the human expert in a friendly fashion. For instance, a device measuring a vital parameter (e.g., the heartbeat or the blood glucose) is generally more important than one measuring the brightness. The policy above tends to assign the arcs to the nodes with a higher

priority; it aims at minimizing the probability of new re-assignments of the same arc in the future. Indeed, since the priority of a node is computed as a combination of both the time elapsed from the moment it joined G_k and its importance (in terms of offered features), a node with a high priority is less probable to leave G_k.

J is the Jaccard coefficient between the features of n_c and the ones exposed by the group G_q, which n_l belongs to. We recall that the Jaccard coefficient measures the similarity between two sets and returns a value in the real interval $[0,1]$; the higher this value the higher the similarity [652].

The competition to inherit the arc is initialized by the leaving node. After all the candidate nodes of G_k have determined their score, they anonymously communicate it by using the anonymous broadcast communication of the information delivery protocol described in Section 7.3.2.3. Hence, the node with the highest score will be selected to inherit the arc left by n_i. It will inherit this arc in an anonymous way. When this happens, the value of ν, and consequently of ρ, for this arc is reset.

As previously pointed out, when n_i leaves G_k and the MIoT, it must also inform the nodes of WZ. In fact, all the nodes belonging to WZ, or temporarily assigned to other groups, must know all the changes in every group because these changes may activate resize or remediation operations that might involve them.

7.3.2.2 Group-level operations

Group level operations indicate those operations that can be carried out by a group in a MIoT. The possible operations are three, namely *Formation*, *Remediation* and *Resize*. We describe them in the next subsections.

Formation of a group

A new group is formed when all the following conditions are verified:

- The number of features currently present in WZ is higher than or equal to k, in such a way as to satisfy the k-anonymity property.
- At least k of these features belong to equivalence classes that satisfy t-closeness. We recall that an equivalence class satisfies t-closeness if the distance between the distribution of a sensitive attribute in this class and the one of the same attribute in the whole data sample is lower than or equal to a threshold t.
- Each of these features is present in at least $\eta > k$ nodes.

In other words, a new group can be formed if there are at least k features with a sufficiently similar distribution in WZ. It is not necessary that each feature is present in the same number of nodes; indeed, it is sufficient that it is present in at least η nodes.

Finally, a group can also have more than k features provided that the additional ones are present in at least η nodes and the sum of their distributions is not higher than the sum of the distributions of the first k features. This condition is justified by the fact that the k features must be characterizing for the group, and this does not happen if there are other ones more present than them therein. As a consequence of the previous reasoning, the number $|NS^P|$ of the permanent nodes of a new group must be higher than or equal to $k \cdot \eta$. There is also a threshold th_{max} for the maximum number of nodes (i.e., for the maximum value of $|NS^P| + |NS^T|$) of the new group. This threshold is linked to the performance of the routing algorithm and to the fact that the graph $\mathcal{G}$ corresponding to the new group is totally connected.

A final parameter that plays a key role in the formation of a new group is the priority π_c of the candidate nodes (see Section 7.3.2.1). In fact, if there are two or more candidate nodes, our approach selects the one with the highest priority.

Example 7.1. In Figure 7.2, we illustrate an example of the formation of a new group according to our strategy. Here, we consider a situation in which the WZ contains five nodes, namely $n_1..n_5$, whose features are reported in the legend of Figure 7.2. For the sake of simplicity, in this example, we set $k = 2$ and $\eta = 2$ and we assume that "energy" and "lighting" are two features belonging to an equivalence class. Therefore, because WZ contains at least 2 nodes with the features above, both the privacy requirements (i.e., $k = 2$ and $\eta = 2$) are satisfied. As a consequence, a new group, namely "$Group_x$", can be formed containing nodes n_1, n_2, n_3, and n_4. The set of features exposed by this group, and therefore by its members, will be: "energy", "lighting", and "cooling".

It is worth noting that, because the requirement on k is already satisfied by the presence of "energy" and "lighting", the feature "cooling" can be safely exposed as it satisfies the requirement on η.

Of course, n_5 cannot be part of this new group because it does not share any feature with the other nodes. □

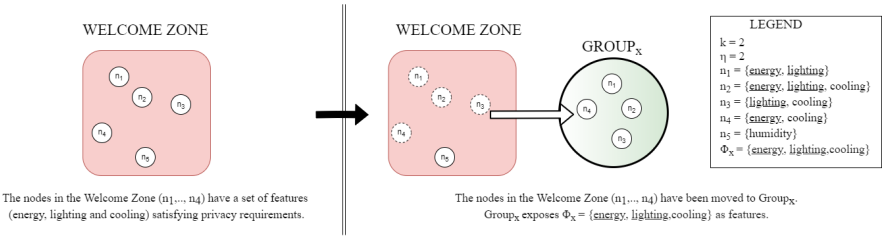


Fig. 7.2: Tasks carried out during the formation of a new group

Remediation of a group

In case the rate of arrival of new nodes in the MIoT is low, the overall dynamism of the MIoT can be reduced, and some degenerative situations may arise, in which the nodes remain a long time in WZ before being able to join any group. The temporary join of a node to a group has been thought just to address this issue. As a matter of fact, each group can temporarily accept some nodes (if the overall number of its permanent and temporary nodes is less than th_{max}) provided that their features are already exposed by that group. In any case, WZ keeps track of temporary joins because, if the set of the nodes belonging to it or temporarily assigned to a group satisfies the conditions necessary for the formation of a new group, this last activity is started.

However, in spite of the previous policies, it can happen that, owing to the arrival rate of new nodes in the MIoT, there exists a node n_i whose features are not exposed by any group yet, and, therefore, incapable of participating to the MIoT's life for a long time. To address this issue, our approach provides the remediation operation. It can be activated if there are at least two groups whose number of permanent and temporary nodes is less than th_{max}. Let G_h and G_l be two of these groups. Remediation starts by recalling the nodes of G_h and G_l in WZ. This task aims at constructing two new groups G'_h and G'_l starting from the nodes of G_h and G_l in such a way that one of the new groups can contain n_i[2].

The approach followed by the remediation plan leverages the fact that each node knows only the nodes of its group and, in case it is a c-node, some other ones of different groups.

Now, since in a group there are k characterizing features and each feature is exposed by η nodes ($\eta > k$), our remediation operation can guarantee that a feature exposed by the new node is "hidden" among the ones exposed by at least $(k \cdot \eta) - 1$ existing nodes in the corresponding group. As a consequence, the probability that a node of this group detects the node providing the new feature is less than $\frac{1}{(k \cdot \eta)-1}$ that, in turn, is less than $\frac{1}{k}$. This implies that our remediation operation can guarantee k-anonymity.

Example 7.2 (continued). Figure 7.3 shows a possible evolution of the previous example. Now, a new node, say n_7 is entering the WZ already containing nodes n_5 and n_6. Once again, the features of all nodes are reported in the legend of Figure 7.3. In this situation, a new group cannot be created as features of nodes in the WZ do not satisfy privacy requirements. However, while n_5 and n_6 do not share any feature

[2] Clearly, it is not sure that the features of n_i allow it to be a member of G'_h or G'_l. If this does not happen, n_i will remain in WZ.

with existing groups, n_7 has the feature "lighting" already exposed by "$Group_x$". Therefore, according to the remediation operation, n_7 could safely joins "$Group_x$" provided that it conceals the feature "alarm" not exposed by this group. □

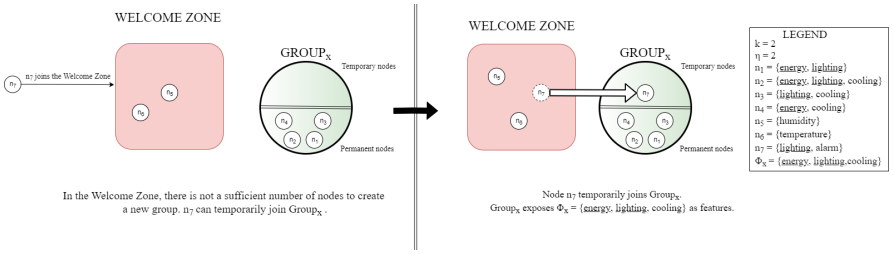


Fig. 7.3: Tasks performed during the remediation of a group

Resize of a group

A group resize operation is activated after that k permanent nodes performed a leave operation in a group. Waiting for k leave operations before carrying out this task is necessary to guarantee k-anonymity. In fact, we can reconstruct one or more features of a node leaving the group if we verify the corresponding impact on the set of features after each node leaves, at least in some cases. By contrast, waiting for k leave operations before verifying the features of a group allows our approach to guarantee that the possible impacts can be associated with k different nodes and, then, that k-anonymity is preserved.

When the resize of a group G_k starts, two different cases are possible, namely:

- all the features previously exposed by G_k are still present, but for at least one of them k-anonymity is not guaranteed;
- at least one feature previously exposed by G_k is no longer present and t-closeness is not guaranteed; the other features may or may not guarantee k-anonymity.

If one of the previous conditions is true, it is necessary to start a group restore task. Given a feature φ that does not currently guarantee k-anonymity, the resize task tries to perform one of the following countermeasures:

- C_1: if G_k contains a temporary node that exposes φ, then it is added to G_k as a permanent node.
- C_2: if G_k contains no node that exposes φ, but a suitable node is present in WZ, then it is added to G_k as a permanent node.
- C_3: if neither a temporary node in G_k nor a node in WZ exposes φ, but at least another group contains a temporary node exposing this feature, then this node is added to G_k as a permanent node. If more than one node exposing φ exists

in the MIoT, then one with the minimum priority is chosen to be added to G_k. This is justified by considering that priority depends on the time a node elapsed in the group and on its importance. Removing from a group G_l a node with a high priority (even if it has been assigned to G_l only temporarily) could imply removing from G_l a node important for it and/or a node that spent a certain amount of time in this group. This last condition could have led this node to construct several links and relationships that are broken if it is forced to change its group.

Of course, the operations described above are carried out for all the features that are not currently guaranteeing k-anonymity in such a way as to preserve node privacy. Actually, the verification of a group G_k is performed as a challenge between nodes permanent in G_k and external nodes. Analogously to what happens for group formation, the permanent nodes of G_k start by anonymously communicating their features to WZ. The other nodes that are listening to WZ (i.e., those nodes not assigned to a group yet, or those nodes temporarily assigned to a group) participate to the challenge by adding their features (still leveraging the anonymous broadcast) until G_k satisfies the privacy requirements again.

By following the algorithm above, in the resize of G_k, its temporary nodes are preferred to the free nodes of WZ that, in turn, are preferred to the temporary nodes of other groups. Each node independently estimates its contribution to G_k; in this task, it considers the priority of its category as a key aspect. Finally, if more suitable nodes exist in the same category, a priority-based approach, similar to the one discussed in Section 7.3.2.1, is adopted to select the one to be added to G_k.

This task terminates when:

- G_k exposes a set of features that guarantees both k-anonymity and t-closeness;
- G_k is in one of the two cases that do not guarantee k-anonymity and/or t-closeness and there exists at least one feature of G_k for which no countermeasure can be applied.

In the former case, G_k is restored, whereas, in the latter case, it must be dissolved, and the corresponding nodes must be re-assigned to WZ. Observe that these nodes will remain in WZ only until either G_k can be fully restored or they can join (even temporarily) another group G_l, such that the set $\Phi_{int} = \Phi_k \cap \Phi_l$ contains at least k features that belong to equivalence classes satisfying t-closeness.

Example 7.3 (continued).

In Figure 7.4, we illustrate another possible evolution of our running example. In this case, nodes n_1 and n_4 are leaving the system so that "$Group_x$" no longer

satisfies privacy constraints on feature "energy" ($\eta < 2$). Observe that, this feature also contributed to comply with the requirement on k (i.e., $k = 2$) as it belonged to an equivalence class together with the feature "lighting". In this case, the resize operation has to be executed for "$Group_x$".

According to Case A of Figure 7.4, the node n_8 is available in the WZ and because it has the feature "energy", it can safely join "$Group_x$". In this way, the privacy requirements for this group are restored and, hence, the group can remain alive.

In Case B, instead, no node, with the needed features, is available to join "$Group_x$". In this scenario, this group can no longer exist. Therefore, its nodes leave it to join WZ once again. □

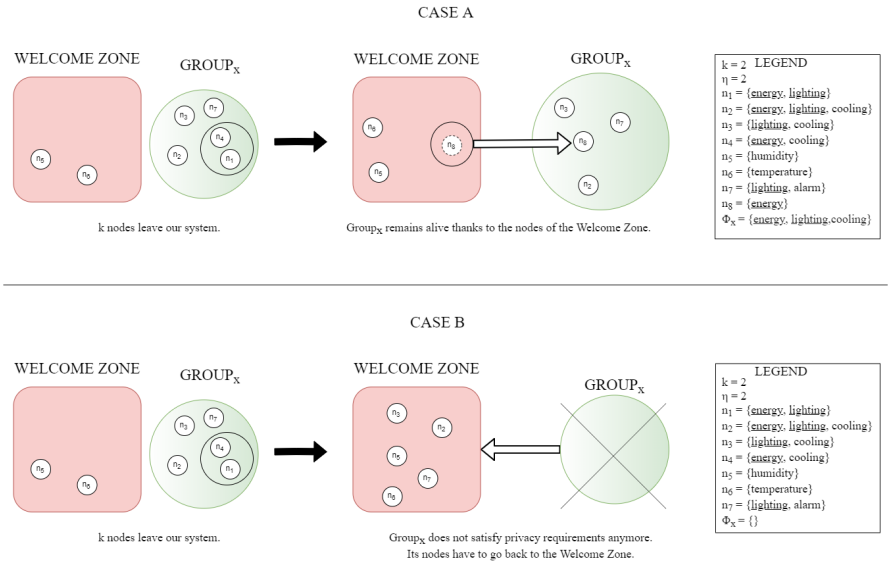


Fig. 7.4: Tasks performed during the resize of a group

7.3.2.3 Information delivery protocol

Our information delivery protocol is based on three kinds of message, namely point-to-point, broadcast and anonymous broadcast. The first two are directly derived from the corresponding functionalities provided by the network underlying the MIoT. Instead, the third is based on a combination of the first two; it will be illustrated below.

The objective of anonymous broadcast is the implementation of a mechanism to anonymously deliver a message to all the nodes of a group or of WZ. Actually, anonymous broadcast can be seen as a hybrid approach consisting of a preliminary set of point-to-point exchanges of the message to deliver, handled in a way analogous to what happens in mix-net networks [681, 283], followed by a broadcast delivery of the same message.

There are several techniques to implement this strategy (see, for instance, [681, 437, 708, 9, 296, 28, 339]). A naive (but, at the same time, efficient and effective) way of proceeding is as follows. When a node n_i receives a message m, it forwards m to another node n_j with a given probability p by using the point-to-point mode. Instead, with a probability equal to $1 - p$, it forwards m in broadcast mode to all the nodes of its group (or to WZ). The value of p must be chosen to guarantee a trade-off between the need to quickly deliver m to all the nodes of the group (in such a way as to avoid that m becomes obsolete) and the need to preserve privacy. When m is received in broadcast mode by a node n_i of a group, if n_i has arcs towards nodes of other groups that expose features characterizing m, it can use these arcs to deliver m to the corresponding groups in a point-to-point mode.

After having illustrated the three possible message modes, we now examine the possible message types provided by our information delivery protocol. They can be grouped in three categories, namely *join*, *leave* and *query*. We illustrate all of them in the following subsections.

Join Messages

The messages belonging to this category are the following:

- *WZ Hello*. This message has the form ⟨'`Hello`', '`WZ`'⟩. It is sent in broadcast mode by a node n to WZ when n requires to join the MIoT.
- *WZ Answer*. This message has the form ⟨'`Welcome`', $\epsilon + 1$⟩. It is sent in broadcast mode by WZ as an answer to the corresponding *WZ Hello* message previously sent by a new node n to WZ. $\epsilon + 1$ is an integer denoting the number of nodes (including n) present in WZ after the join of n.
- *Temporary Hello*. This message has the form ⟨'`Hello`', '`T`'⟩. It is sent in broadcast mode by a node n to a group G when n requires to temporarily join G.
- *Permanent Hello*. This message has the form ⟨'`Hello`', '`P`'⟩. It is sent in broadcast mode by a node n to a group G when n requires to permanently join G.
- *Feature Set*. This message has the form ⟨'`Feature Set`', ϕ⟩. It is sent in anonymous broadcast mode by a node n to the nodes of WZ. ϕ denotes the set of the features exposed by n. In order to preserve the privacy of n, this message can be sent when at least $\epsilon \geq \eta$ nodes are present in WZ. It represents the first step for the formation of a group.

Leave Messages

The messages belonging to this category are the following:

- *Temporary Leave*. This message has the form ⟨'Bye', 'T'⟩. It is sent in broadcast mode by a node n, which has been temporarily assigned to a group G, when it decides to leave G. When this happens, n is assigned to WZ.
- *Permanent Leave*. This message has the form ⟨'Bye', 'P'⟩. It is sent in broadcast mode by a node n, which has been permanently assigned to a group G, when it decides to leave G. When this happens, n is assigned to WZ.
- *WZ Leave*. This message has the form ⟨'Bye', 'WZ'⟩. It is sent in broadcast mode by a node n, which is assigned to WZ, when n decides to leave WZ and, consequently, the MIoT.
- *Score Communication*. This message has the form ⟨'Score', 'Sc'⟩. It is an anonymous broadcast message sent by each node during a challenge for selecting a candidate to participate to a group or to inherit an arc (see Section 7.3.2.1).

Query Messages

The messages belonging to this category are used by a node when it requires a certain feature. They are the following:

- *Intra-group Query*. This message has the form ⟨'Intra Query', ⟨content⟩, φ⟩. Here, ⟨content⟩[3] denotes the message payload, whereas φ represents the feature the message refers to. It is delivered in anonymous broadcast mode by a node n to the nodes of its group.
- *Extra-group Query*. This message has the form ⟨'Extra Query', ⟨content⟩, φ⟩. Here, ⟨content⟩ denotes the message payload, whereas φ represents the feature the message refers to. It is delivered in anonymous broadcast mode by a node n to the nodes of its group G. If G contains any c-node toward another group G', whose features match those in φ, then the c-node delivers the message to its contact in G'. However, if, in turn, G' has c-nodes, the message is not further delivered to other groups. This choice has been made to avoid the traffic overloading in the network underlying the MIoT.

7.3.3 Security Model

7.3.3.1 Attack Model

As a preliminary assumption, we consider a realistic situation in which a sufficient number of nodes is available so that our approach can be implemented successfully. Therefore, we will not consider anomalous situations, in which the number of the

[3] Observe that no constraint is put on the content to handle, in such a way as to guarantee data confidentiality and integrity.

nodes available in the system is less than the minimum number necessary to guarantee, at least in principle, privacy (i.e., $k \cdot \eta$).

Furthermore, our approach focuses on the protection of node information and does not deal with attacks on the protocol, such as sinkhole or DoS attacks [563, 651]. Indeed, these threats are common for most of the communication protocols and the strategies for preventing them are orthogonal to our proposal. In our case, it is possible to adopt any of these strategies, such as the ones presented in [123, 195, 696], in such a way as to make our approach robust also to these kinds of attack.

Given this basic assumption, we now identify the security properties of our approach. They are:

- *SP1* - The definition of the groups' features ensures the privacy of nodes.
- *SP2* - Our approach is resistant to attacks exploiting group resize operation.
- *SP3* - Our approach is resistant to timing attacks exploiting cross-feature interview.
- *SP4* - The jeopardizing of the routing protocol does not have impact on the privacy of nodes.
- *SP5* - Our anonymous broadcast delivery protocol is resistant to classical attacks (e.g., the timing and the routing ones).
- *SP6* - Our approach is resistant to attacks based on historical data concerning join and leave operations.

In the analysis of the security properties described above we will consider the following assumptions:

- *A1* - An attacker cannot control a whole group of nodes.
- *A2* - The underlying network provider is not interested in violating node privacy.
- *A3* - The basic features delivered by the MIoT system (point-to-point communication, etc.) are robust to attacks.
- *A4* - All the features considered in our approach are not related to geographic positions.
- *A5* - At most t nodes can collude to break the security properties of our protocol.
- *A6* - The attacker has no additional knowledge derived from any direct physical access to nodes.

In the following, we will investigate the security properties mentioned above. To perform this analysis, we needed a reference scenario. To model it, and to test our approach, we constructed a prototype. Furthermore, as real MIoTs with the size and the variety handled by our model do not exist yet, we constructed a MIoT simulator.

To make "concrete" and "plausible" the simulated MIoTs, we had the necessity that our simulator was capable of returning MIoTs having the characteristics specified by the user and being as close as possible to real-world scenarios. In the simulator design, and in the next construction of the MIoTs to use for the experiments, we followed the ideas expressed in [304, 73, 74], in which the authors highlight that one of the main factors used to build links in an IoT is node proximity. In order to reproduce the creation of links among objects, we decided to leverage information about real-life paths in a city. In fact, having this information at disposal, we may associate each path with an object and link two objects if their paths have been near enough for a sufficient time period. As for a dataset containing real-life paths in a city, we selected the one reported in `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. It regards taxi routes in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. Each route contains several Points of Interests corresponding to the GPS coordinates of the vehicle. As said above, our simulator associates an object with a given route recorded in the dataset. Furthermore, it creates an arc between two nodes if the distance between the corresponding routes is less than a certain threshold th_d for a predefined time interval th_t. The value of th_d and th_t can be specified through the constructor interface. Clearly, the higher this value the more connected the constructed IoT. The interested reader can find the IoTs created in this phase at the address `http://daisy.dii.univpm.it/miot/datasets/privacy`.

Regarding the MIoT construction, since group creation depends on the sequence of subscriptions of the nodes to our system (which, for the sake of simplicity, can be assumed as random) and on their features, we reproduced it by simulations, as will be clear in the following. When we defined the distribution of the features among the nodes, we leveraged scientific literature and used the corresponding results to properly tune our simulator. In particular, we adopted the values reported in [301].

Some statistics about our dataset are reported in Table 7.2.

Parameter	*Value*
Number of nodes	1000
Number of relationships	6860
Mean outdegree	6.995
Mean indegree	7.002
Number of distinct features	20
Maximum number of features for an equivalence class	10
Maximum number of features for a node	3

Table 7.2: Parameter values for our simulator

7.3.3.2 Security Analysis

SP1 - The definition of the groups' features ensures the privacy of nodes

This property is fundamental in our approach because it guarantees that, inside a group, nodes are protected against attacks to their privacy. Our approach uses a combination of k-anonymity and t-closeness to ensure this property. Indeed, k-anonymity alone fails because, in real life, features are not uniformly distributed among smart objects. Therefore, an attacker, near a node, may take advantage of the probability distribution function to perform a statistical attack and to improve the guessing probability.

For this reason, our algorithm takes into account the distributions of the features that characterize a new group when it selects k features. In accordance with the t-closeness paradigm, the characterizing features of a group must belong to an equivalence class when it comes to their probability distribution. This ensures that an attacker cannot exploit the background knowledge on the popularity of features among smart objects in such a way as to exclude the least probable ones, thus increasing the probability of mapping a feature to an object.

Furthermore, as for group formation, our protocol exploits, once again, the notion of k-anonymity to allow nodes to freely exchange information about features without being identified. Indeed, each node inside WZ waits until $\epsilon > k$ nodes are available before adopting the anonymous broadcast protocol to communicate its features. Now, in absence of collusion attacks, ϵ can be set to k. In this way, an attacker can only observe that there are some features among those k nodes, without having further advantages in mapping them to the right objects. Moreover, in this case, t-closeness is not needed because the attacker is dealing with a set of k nodes each having exactly the same probability to own the specified properties. As a final observation, in accordance with Assumptions *A1* and *A5*, an attacker can only control t nodes simultaneously. Therefore, to block a collusion attack, it is possible to set $\epsilon = k + t$ in such a way as to preserve the k-anonymity property.

SP2 - Our approach is resistant to attacks exploiting group resize operation

The aim of this property is to protect our system from attacks based on the observations of resize operations. Indeed, during each resize operation, the structure of groups may change in terms of both the number of involved nodes and, possibly, the number of available features. An attacker can evaluate the features proposed by a group by either interacting in proximity with a node of that group or by being a member of the group itself.

Our approach adopts two countermeasures to this kind of attack. The former consists in forcing the resize algorithm in such a way that it can be activated only when k leave operations have been recorded. Due to Assumption *A1*, the attacker cannot control a group and, hence, cannot control which nodes leave the system and when it happens. Moreover, as a further security measure, we require that, for each feature, there are at least η nodes owning it. The combination of these countermeasures inhibits the attacker from detecting which feature was owned by the leaving nodes (the probability of guessing it will be the same as the one of guessing the features of any other node in the group). In this way, our approach prevents the attacker from being able to detect a reduction of the number of the available features included in the group.

SP3 - Our approach is resistant to timing attacks exploiting cross-feature interview

A common attack typical of scenarios similar to the ones proposed here is based on the statistical observation of the response time of nodes to external events. In our case, this attack can be executed by querying a node about information related to a predefined set of features and by comparing response times. Fast answers can be associated with features owned by the node, whereas slow answers (or empty ones) can be mapped to features owned by other nodes of the same group that the attacked node must contact to provide its answer.

To prevent this kind of attack, each node adopts a pattern recognition algorithm and enters a protection mode each time it recognizes a suspect querying pattern. Basically, whenever a target node receives a suspect sequence of consecutive cross-feature queries from a source node, say n_a, it starts by adding a random delay in its answers to n_a. This delay ranges from 0 to the maximum answering time detected by it in any previous communications[4]. Furthermore, if the node is not able to answer two consecutive cross-feature queries, it will stop answering any next query from n_a for a certain time interval.

These two countermeasures, when combined with Assumption *A4*, prevent the attacker from gaining advantages by maliciously interviewing any node of our system. Indeed, Assumption *A4* states that the attacker cannot leverage information about specific geographic positions (for instance, to isolate a small set of devices) when she formulates her queries. Without this assumption, an attacker can construct, and then submit, queries whose answer can be provided only by devices lo-

[4] Observe that no countermeasure is adopted in case of consecutive queries referring to the same feature. Indeed, in this case, it can be assumed as a normal interaction between two nodes.

cated in a specific geographic position. Of course, this is a local attack that, in order to have success, requires a contemporary physical attack allowing the malicious user to isolate a small set of devices to detect the features owned by them. For this reason, we have assumed that geolocalized features are out of the scope of our approach.

SP4 - The jeopardizing of the routing protocol does not have impact on the privacy of nodes

This property guarantees that an attacker cannot gather information about the properties of nodes by tampering the communication protocol. Indeed, she can try to force any communication of a group to pass through it. Although this cannot be achieved for intra-group communications, because the corresponding path is randomly chosen by the nodes inside a group, it can be tried for inter-group communications. Indeed, an attacker may tamper the protocol during the leave of nodes and may promote itself as the node with the highest score, in such a way as to inherit all the arcs towards other groups. This is a variant of the sinkhole attack. The result is that the group will be potentially isolated and its nodes cannot use external arcs without involving the attacker.

Of course, this is an unwanted situation, which can cause issues to the communication protocol. However, no harm is done to nodes' privacy, as each node will still continue to communicate with each other leveraging the anonymous delivery protocol described in Section 7.3.2.3. Therefore, even though the attacker may force itself in the middle of all the communications towards external groups, it cannot reveal any information about the nodes being the sources of these communications.

As stated above, our approach does not directly deal with sinkhole attacks when it comes to damages to the communication protocol. Actually, the adoption of well-known countermeasures for these attacks proposed in the scientific literature (such as the ones described in [123, 195, 696]) can help preventing them.

SP5 - Our anonymous broadcast delivery protocol is resistant to classical attacks

This property aims at guaranteeing the robustness of the anonymous broadcast delivery protocol described in Section 7.3.2.3. First, observe that, thanks to Assumption *A3*, the basic communication functionalities, such as the private point-to-point communication mechanism among nodes, are assumed to be robust against attacks. Therefore, the anonymous delivery protocol can be built on top of these basic features by directly adapting any anonymous broadcast communication protocol proposed in the scientific literature, whose security has been already proved [681, 437, 708, 9, 296, 28, 339].

Having said these premises, let us consider the naive method to address this goal already described in Section 7.3.2.3. To achieve an anonymous broadcast delivery, this approach leverages a random sequence of private point-to-point messages among nodes to obfuscate the source of a message before broadcasting it. This strategy somehow resembles the one adopted in mix-net solutions, whose security level and possible flaws are investigated in [681]. However, because of its simplicity, this approach can be effective and efficient in low-severity scenarios, in which more advanced solutions, like the ones mentioned above, are not necessary.

Due to Assumption *A3*, an attacker cannot have access to point-to-point messages exchanged between generic pairs of nodes. To guess the original source, she can only observe broadcast messages and the point-to-point ones sent to her. As each node sends a message to another one in a point-to-point fashion with a probability p and the same message in broadcast with a probability $1-p$, the length of the communication path will be strongly variable and unpredictable a priori. Furthermore, the next node in the communication path will be chosen randomly and there is no limit to the path length. If all these features are combined with Assumption *A1*, it is possible to conclude that our approach prevents an attacker from being able to trace back the message source and, ultimately, from having advantages in guessing its features.

SP6 - Our approach is resistant to attacks based on historical data concerning join and leave operations

This property aims at guaranteeing the robustness of our approach against attacks exploiting the knowledge of historical data, which examine join and leave operations from groups to disclose the features of an object.

Although nodes can freely join and leave groups, re-join operations involving different groups are, in principle, insidious. Indeed, in this case, nodes can drastically change the exposed set of features. This would allow an attacker to intersect the previously exposed features with the currently exposed ones to determine the real subset of them owned by the attacked node.

However, in our approach, a re-join task only happens when a node leaves a group and joins another one during the resize operations (see Section 7.3.2.2 for all details). In any case, this issue is addressed by the condition specified in Section 7.3.2.2 according to which a node can re-join the same group it belonged to (even temporarily) in the past or a new one if the intersection of the features exposed by the two groups contains at least k features belonging to equivalence classes that satisfy t-closeness. This countermeasure, along with Assumptions *A1* and *A6*, contrasts this kind of attack.

Another situation to be investigated regards the case in which a node permanently leaves the MIoT (and not simply a group) and, then, re-joins it. Also in this case, historical data can lead to advantages for an attacker. Actually, this issue is not directly considered by our approach. However, a simple protection strategy can be adopted to address it. Indeed, it is sufficient to require that the nodes, which re-join a MIoT after a permanent leave, should restore information about the last group they belonged to during the previous interaction with it. In this way, it is possible to apply the countermeasures for the other re-join situation described above.

7.4 Results

7.4.1 Solving the trade-off between privacy requirement and network performance

In this section, we aim at investigating the configuration of the privacy parameters, namely k and η, in such a way as to achieve the desired privacy level. Indeed, the more severe privacy requirements, the greater the impact on the network performances.

According to our protocol, a more demanding privacy requirement leads to an increase of the group size. The communication among nodes is influenced by both the presence of groups and the anonymous broadcast protocol, which requires the involvement of a random number of nodes inside each group before reaching the desired destination. As a consequence, both intra-group and inter-group communications are strongly dependent on the group size; specifically, the bigger the groups the higher the number of involved nodes. This has two direct implications on the network performance: *(i)* the overall load of the network increases; *(ii)* the average length of the paths among nodes grows (leading to higher average communication delays). For this reason, a first experiment is devoted to simulating the behavior of our system and to monitor the creation of groups.

The metrics we adopted for this investigation are: *(i)* the variation of the group size against different privacy settings (i.e., different configurations of k and η); *(ii)* the variation of the length of the communication paths among nodes after the application of our privacy model.

For simulation, we considered different values of both k and η. Specifically, as for k, we selected the range $[2, 8]$, with a step of 1; as for η, instead, we considered a multiple of k; in particular, its range was $[k, 2k]$.

As a first investigation, we measured the metric *(i)*. For this purpose, we simulated a random subscription to our system (i.e., a random arrival order in the Welcome Zone) of the 1000 nodes of the original IoT graph considered in this exper-

iment. We applied our algorithm for group formation and measured the average number of nodes inside each group, as well as the average number of nodes not involved in a group and, hence, waiting in WZ. In this experiment, we did not consider temporary joins that can be adopted to minimize the number of nodes not assigned (either temporarily or permanently) to any group.

To consider different configurations of node arrivals, we repeated the experiment 250 times and averaged the corresponding results. In Figure 7.5, we report the average percentage of all the nodes of the MIoT that are present in a group against the increase of k and η. Instead, Figure 7.6 shows the average percentage of all the nodes of the MIoT that remain in WZ against the increase of k and η.

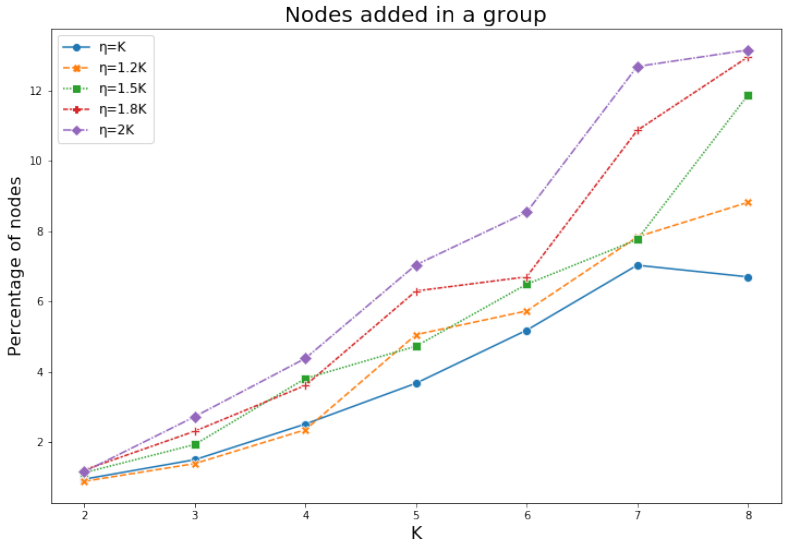


Fig. 7.5: Percentage of nodes present in a given group against the increase of k and η

By analyzing the obtained results, we can observe that the percentage of nodes in a group grows linearly with the increase of both k and η. Interestingly, even with the most demanding privacy requirement (i.e., $k = 8$ and $\eta = 2 \cdot k$), it does not exceed 12.5% of the whole set of nodes. Of course, as proved in [249], higher values of k do not provide additional benefits, once the desired privacy requirement has been reached. With regard to this reasoning, we point out that there is no best practice in the estimation of the right value of k. Typical values adopted in the literature range from 2 to 5. As for η, this is a security mechanism introduced to maintain the full operation of a group also in presence of node leaves. However, since our approach for group resize is executed each time k permanent nodes leave a group, to preserve its robustness, we need to have the k-anonymity property guaranteed in the interval from the leave of the first node to the leave of the k^{th} one (after which the group size will be fixed by our approach). At a first analysis, we may affirm that,

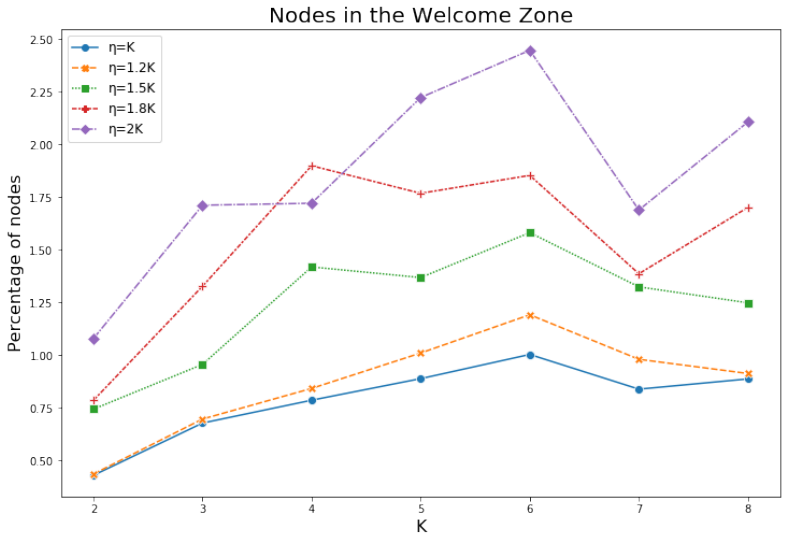


Fig. 7.6: Percentage of nodes waiting in the Welcome Zone against the increase of k and η

if η is equal to $2 \cdot k$, no issues will arise before the resize procedure will be executed. This setting is the most preserving one but, as a contrast, it requires a very high number of nodes for each feature. However, if we consider a limit case in which all the leave operations involve nodes owning only one of the available features without repetition, we could safely set $\eta = k + 1$ to ensure the k-anonymity property and the operability of the group during leave operations. These considerations are crucial to properly tune η. Indeed, we can conclude that its right value should range from $k+1$ to $2 \cdot k$.

As a further observation, keeping η to the minimum values strongly reduces the number of nodes still waiting in WZ after the formation of groups. Indeed, if we set $k = 4$ and $k < \eta = 1.2 \cdot k$, the average percentage of nodes waiting in WZ after the execution of the algorithm for the formation of groups is about 0.08%. Also the number of nodes in each group is low and equals to 2.2% of the nodes of the original graph on average.

The second experiment aims at measuring the metric *(ii)*. To perform this measurement, we applied the same logic adopted in the previous experiment to simulate the formation of groups, but we preserved the original links in the graph built from our dataset for inter-group connections. Observe that this choice is compliant to what should happen in a real life scenario because inter-group connections rise in accordance with proximity events among nodes belonging to different groups, which is exactly how links have been established in the original IoT graph. Now, given a pair of nodes (n_i, n_j) such that $n_i \in G_i$, $n_j \in G_j$, $G_i \neq G_j$ and there exists a path from n_i to n_j in the original graph, Figure 7.7 reports the ratio of the length

of the path between n_i and n_j in our system to the length of the path between the same nodes in the original graph. We call "Cost of the Protocol" (hereafter, CoP) this parameter. The values reported in this figure are averaged on 1000 pairs of nodes satisfying the requirements above.

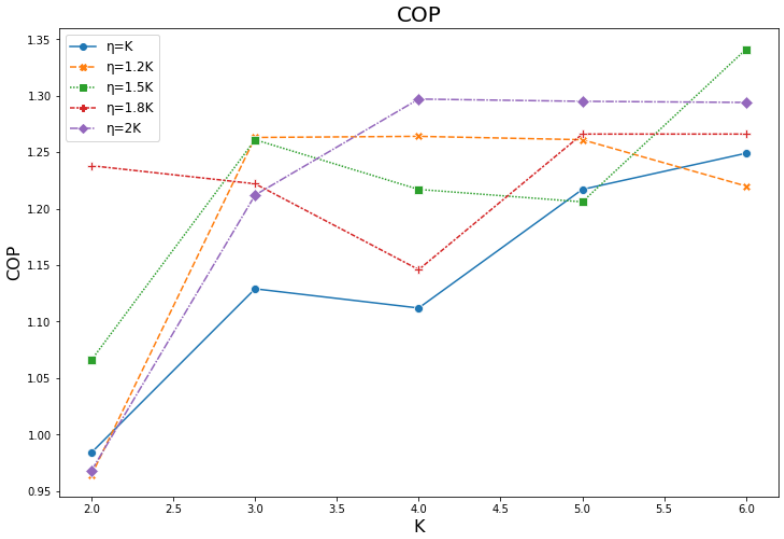


Fig. 7.7: Value of CoP against the increase of k and η

The obtained results show that, if we keep $k \leq 4$ and $\eta = 1.2 \cdot k$, CoP reaches a maximum value of 1.263, meaning that the length of the path among the pairs of nodes obtained by applying our approach increases to a maximum of about 26% with respect to the length of the original path.

7.4.2 Comparison with other approaches

As pointed out in the Introduction, to the best of our knowledge, our approach is the first one conceived to prevent feature disclosure in a multiple IoT scenario. As a consequence, a direct comparison between our approach and a strictly related one is not possible. Nevertheless, it is possible to perform an "indirect" comparison with another approach which, even if conceived for a different objective, shares some similarities with ours in both the reference scenario (i.e., smart devices and IoT) and the adopted methodology.

To carry out this task, from the scientific literature, we identified the work described in [45]. It presents an intrusion detection system aiming at protecting smart devices in vehicular networks. In this approach, the main idea is to group nodes into "clusters" in order to build protected zones where nodes collaborate to improve their security. We remark, again, that the goal of the approach of [45] is different from the objective of our approach. However, both of them define a security model conceived

to operate on smart devices and IoT, and their strategy is centered on the presence of groups and clusters of objects.

Interestingly, the authors of [45] measure the delay introduced by their solution to the communication time. In Section 7.4.1, we carried out a similar analysis but we evaluated another performance parameter, namely the increase of the average path length caused by our privacy preserving solution. In order to allow a comparison between our approach and the one of [45], we decided to measure the communication delay introduced by our approach. We defined it as the average difference, in terms of time to delivery, between a scenario in which our approach is used and another in which it is not adopted. As done in [45], we measured such a variation against the size of groups. To estimate communication time, we leveraged a global ping service available at the address `https://wondernetwork.com/pings`. In Figure 7.8, we report both our results and the ones of the approach described in [45].

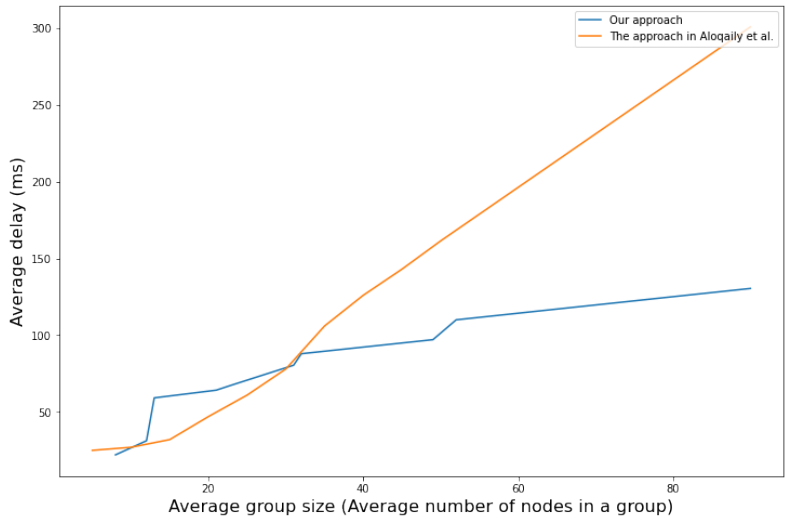


Fig. 7.8: Average delay in the objects' communication introduced by our approach against the group size

From the analysis of this figure we observe that the average delay introduced by our approach ranges from 22 *ms* to 130 *ms*, whereas the average delay of the approach of [45] ranges from 24 *ms* to 170 *ms*. The outcome of this experiment shows that the performance of our approach is comparable with the one of other solutions, already present in the scientific literature, addressing security issues in the context of smart devices and IoT. This encourages us to state that our approach achieves pretty satisfactory results, still preserving the overall IoT usability to values considered acceptable by the scientific community in this application scenario.

7.5 Discussion

7.5.1 Privacy features

We start by analyzing the two features adopted in this chapter, namely: *(i)* k-anonimity, and *(ii)* t-closeness. k-anonimity is a very old notion that, in principle, can avoid information disclosure in a database as long as sufficiently "noisy" tables (i.e., tables guaranteeing k collisions) can be generated [243]. However, it was also proved that, when dealing with value distributions of attributes, an attacker can take advantages by comparing the distribution in the noisy dataset with the real-world attribute distribution to bypass such a privacy mechanism [450]. Therefore, even if k-anonimity can protect against identity disclosure, it cannot protect against attacks based on attribute disclosure. In this last case, an attacker can leverage the disclosure of the value of a confidential attribute associated with an external identified individual to violate k-anonimity features. In real-life scenarios, the risk of such an attack is very high and, therefore, the only application of k-anonimity appears inadequate for our privacy objectives.

t-closeness was widely studied in the scientific literature [419]. It was conceived as an evolution of k-anonimity that also protects against attribute disclosure. The scenario of our interest is very close to the ones t-closeness was designed for. Indeed, our aim is concealing the features (or attributes) of an object behind a group of heterogeneous and equivalent ones (in terms of probability distributions). For this reason, in our approach, we leverage t-closeness to enhance k-anonimity with the capability of protecting against attribute disclosure, assuming that object attributes (or features, in our case) have specific and measurable distributions.

Interestingly, our solution also recalls the concept of ϵ-differential privacy [244]. This kind of privacy solution aims at limiting the knowledge gain between datasets that differ in one individual. It originally focused on the protection of the outcomes of queries performed in a database. Then, other papers extended this concept to non-interactive scenarios (i.e., cases in which it is not necessary to protect a specific query or set of queries). These solutions often deal with specific classes of generic queries (typically, count ones) [115, 321]. Interestingly, it was proved that t-closeness and ϵ-differential privacy are somehow related to each other [234]. Indeed, the authors of [235] proved that, in a dataset in which t-closeness holds, differential privacy is guaranteed on the projection over the confidential attributes.

7.5.2 Applicability and limitations

As for the applicability of our proposal to real-world scenarios, we highlight that our strategy is in-line with the new trend of improving the independence of nodes

in an IoT. Specifically, several papers focused on the definition of approaches aiming at identifying links between objects with a reduced human intervention [557, 313]. Other papers, instead, focused on the definition of models to uniformly handle data coming from heterogeneous smart objects [60]. Our solution finds a direct application in this context because the knowledge of the features characterizing objects and the services provided by them is fundamental for improving the efficiency of links in an IoT. For this purpose, it is important to filter the contacts of an object according to the usefulness of the information that these contacts can provide. Of course, as stated throughout this chapter, the knowledge of the features of an object has serious impacts on the privacy of its user.

Clearly, due to the extremely high dynamics of the considered scenario, our approach has some limitations that must be taken into account. Indeed, as stated in Assumption *A4* , our solution does not cover the protection of features related to specific geographic positions. Indeed, without this assumption, it is not possible to guarantee the security property *SP3*. To clarify this concept, consider the case in which an attacker can isolate a node in a specific location. Furthermore, assume that some of the exposed features can be related to the object position; think, for instance, of the temperature of a room. In this case, the attacker can evaluate whether the node is capable of correctly answering a query about the temperature of the zone controlled by it. Either a positive or a negative answer results in a privacy leakage, as the attacker is able to identify one of the features of the object for reducing the admissible set. In addition to Assumption *A4* , this security property also requires a pattern recognition solution to detect anomalous cross-feature interviews. Of course, a naive and very conservative solution can be obtained by forcing each node to label as suspect (and, hence, to apply the countermeasure described in Section 7.3.3.2 to it) each direct interaction with a node that submits queries related to more than two features. A more sophisticated and refined solution can be obtained by adopting any existing approach for anomalous pattern recognition [362]; however, it requires a base knowledge to model the normal behavior of nodes.

-

8

Anomaly Detection

In this chapter, we report a first attempt to investigate anomalies in a MIoT scenario. First, we propose a new methodological framework and three orthogonal taxonomies, in which each combination of these taxonomies defines a specific type of anomaly to study. Then, in the context of anomaly detection in a MIoT, we define the so-called "forward problem" and "inverse problem". The definition of these problems allows the investigation of how anomalies depend on inter-node distances, the size of IoT networks, and the degree centrality and closeness centrality of anomalous nodes. The proposed approach is applied to a smart city scenario, which is a typical MIoT. Here, data coming from sensors and social networks can boost smart lighting in order to provide citizens with a smart and safe environment.

The material present in this chapter is taken from [161].

8.1 Introduction

In the Concise Oxford Dictionary[1], *anomaly* is defined as *"something that deviates from what is standard, normal, or expected"*. If regularities allow investigating the general characteristics of a complex system, anomalies allow the uncover and analysis of unexpected features that might not be otherwise discovered. For this reason, the detection of anomalies has become very important in data analytics, and is widely investigated both in statistics and machine learning [23, 22, 25]. The relevance of anomaly detection is universally acknowledged, since data anomalies are at basis of significant events and patterns. Example application domains include: privacy and cybersecurity [707, 673]; fault detection [347]; ecological disturbances [181]; communication networks [665]; social media life [183, 596, 627, 706]; and gene regulation [378, 380].

In recent years, anomalies have been widely investigated in social networks to detect fraudulent individuals [586, 30], spammers [607, 257], malicious behavior,

[1] Concise Oxford Dictionary - `https://en.oxforddictionaries.com`

and so forth. Even more recently, anomaly detection has been analyzed in contexts where more social networks interact with each other [134], thus going from social networking into social internetworking.

Social internetworking is certainly one of the frontiers of social network analysis, since people tend to have multiple social network accounts and can, thus, become "social bridges". Furthermore, all sorts of networked objects are getting increasingly smart and social, giving rise to the so-called Smart Objects (SOs) and revolutionizing both the Internet of Things (IoT) and the Social Internet of Things (SIoT) [70]. Also, several SIoTs and IoTs cooperate with each other through "bridge" objects, thus generating new architectures, referred to in the literature as Multiple IoT (MIoT) [82].

The detection of anomalies in a single-IoT environment has been widely investigated [90, 710, 80, 421, 167], and many results involving privacy, security and fault detection have been found. However, to the best of our knowledge, no investigation on anomalies and their possible detection in a MIoT has been performed so far.

Here, we aim at filling this gap by proposing a new methodological framework for anomaly detection and classification in MIoTs. Our framework models anomalies and the corresponding issues in a MIoT by providing a multi-dimensional view, based on three orthogonal taxonomies: *(i)* presence anomalies vs success anomalies; *(ii)* hard anomalies vs soft anomalies; and *(iii)* contact anomalies vs content anomalies. Each combination of the possible values of these dimensions gives rise to a specific type of anomaly to investigate, for instance the *Presence-Hard-Contact* anomalies. Furthermore, anomaly definitions are orthogonal to specific anomaly detection approaches, past or future, which may be applied (and will be combined) in the context of our framework.

Together with the multi-dimensional taxonomy, another main component of our framework is the extension of conventional methodological frameworks to the MIoT case. Our framework has been conceived to address two problems, known as the "forward problem" and the "inverse problem", respectively. In the forward problem, we aim to analyze the effects that multiple anomalies have onto the MIoT. On the other hand, in the inverse problem, which is traditionally more complex, we aim at detecting the source of the anomalies (i.e., the objects that have generated them) based on the effects that these have on the objects or their connections.

In order to show the possible usage of our framework, we present a case study centered around a smart city. Furthermore, in order to evaluate our framework and extract knowledge, we have conducted a series of tests. These allowed us to find several important knowledge patterns about anomalies and their effects in a MIoT. Our most important findings may be summarized as follows: *(i)* the effects of the anomalies of a node rapidly decrease as the distance from the node itself increases; *(ii)*

anomalies are less evident in a MIoT than in a single IoT; *(iii)* the number of anomalous nodes increases as the number of IoTs increases, in a roughly linear way; *(iv)* the outdegree of anomalous nodes has a great impact on the spread of the anomaly over the MIoT; *(v)* closeness centrality is even more important than degree centrality in the spread of anomalies; *(vi)* the computation time necessary for the detection of anomalous nodes is polynomial against the number of MIoT nodes; *(vii)* the time necessary for evaluating the effects of anomalies in a MIoT is quadratic against the number of its nodes.

The rest of this chapter is organized as follows. In Section 8.2, we examine related literature. In Section 8.3.1, we extend the MIoT paradigm. In Section 8.3.2, we present our multi-dimensional taxonomy of anomalies in a MIoT context. In Section 8.3.3, we introduce the specialization of the forward and the inverse problems for MIoTs. Finally, in Section 8.4 we describe a use case, and in Section 8.5 we illustrate our experiments.

8.2 Related Work

Anomaly detection has been largely investigated in past literature. Here, anomalies have been defined in very different ways, based on the reference domain and data model. A widely accepted definition of anomaly is the one proposed by Hawkins in [323], where an anomaly is defined as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". A definition of anomaly specific for social networks can be found in [110], where the authors define anomaly as "an observation which appears to ignore interactions and relationships between individuals and their peers". In [176], anomalies are referred to as "patterns in data that do not conform to a well–defined notion of normal behavior".

Anomaly detection is an issue largely investigated in past literature. The corresponding research studies can be grouped in several ways. One approach distinguishes these studies into: *(i)* surveys and taxonomies, *(ii)* approaches for anomaly detection in generic networks, *(iii)* approaches for anomaly detection in social networks, and *(iv)* other approaches.

If we consider this classification, our approach belongs to class *(iii)*. In this context, we introduce two main novelties, in that: *(i)* we focus on networks of objects instead of networks of people; *(ii)* we focus on multiple network scenarios instead of single networks. In addition, our methodological framework introduces two further novelties, namely: *(i)* the definition of three new taxonomies specific for anomaly detection in MIoTs; and *(ii)* the investigation of the so called forward and inverse

problems in this research context. Moreover, the study we are presenting is orthogonal to other approaches for anomaly detection in network-based data, since we do not aim at proposing a specific approach to address this last issue.

In the following, in order to give a better overview of the literature, we first examine the four classes of research studies on anomalies and, then, present a table comparing our approach to methods introduced in the literature.

Surveys and taxonomies

Recently, several surveys have proposed structured and comprehensive overviews of anomalies to cope with the need of providing usable taxonomies. A first classification of anomalies can be found in [176], which is considered a pioneering paper in this sense. Besides a formal definition of different kinds of anomalies, the authors highlight the challenges related to anomaly detection. In particular, for each class of anomalies introduced, they focus on existing techniques and application domains. Based on their nature, anomalies have been also classified as Point, Contextual and Collective anomalies. Some applications related to these categories are reported in [24, 427, 583, 394].

A significant amount of work has been carried out on anomaly detection in individual IoTs, as captured by a number of survey papers [90, 653, 80]. On the contrary, to the best of our knowledge, no investigation or categorization of possible anomalies in the context of networks and layered networks (mostly related to MIoTs) has been proposed so far. Works presenting relevant aspects are described in the following.

In [31, 24], the authors investigate anomalies in graph-based environments. Specific analyses of this topic can be found in [56] for social networks, in [287, 295, 360, 312] for intrusion detection, in [600] for traffic modelling, and in [378, 380] for gene regulation.

We characterize anomalies as being either static or dynamic, and as being labelled or unlabeled. In [586], the authors survey the state-of-the-art related to the detection of different types of anomalies in social networks. Here, they show that anomalous users' behaviors in social networks are due to a change in their patterns of interaction or in their ways of interacting with the network, which markedly differ from the ones of their peers. The impact of this anomalous behavior can be observed in the resulting structure, allowing anomalies to be characterized as static or dynamic, labelled or unlabeled. For instance, fraudulent individuals may create a network of collaborations to enhance their reputation in a social network. However, when individuals behave in this way, they show an increased level of interaction in the network and tend to form highly interconnected sub-regions therein.

Anomalies in generic networks

In [607], the authors analyze the detection of e-mail spam in a static, unlabeled network context. In particular, they note that spam and other viral materials are typically sent from a single malicious individual to many targets. As a consequence, detecting a specific star-like structure in a network can be a symptom of malicious behavior. Another approach to spam detection is proposed in [257]. In [30], the authors show that both near-stars and near-cliques are indicators of anomalous behaviors in networks. They focus on anomaly detection in weighted graphs. Their approach can be applied to different contexts, such as intrusion detection, spammer detection, anomalies in social networks, and so forth. They also address the problem of anomaly detection in static, labeled networks. In this context, they consider some ego-networks, each one centered on an individual and, when the sum over a particular label is disproportionately high with respect to the number of edges in the network, they conclude that the corresponding individual has a potentially anomalous behavior. In [338], a universal coding method for unlabeled graphs is introduced and is adopted for anomaly detection in static, unlabeled graphs.

In [189], the authors propose an approach to anomaly detection in dynamic networks. This exploits the analysis of sub-structures, such as maximal cliques, for detecting community-based anomalies, i.e., unexpected variations of communities. In this work, a community coincides with a maximal clique. This approach considers grown, shrunken, merged, split, born and vanished communities, respectively.

In [478], an approach to detect anomalies on dynamic labeled networks in a big data context is presented. Big data is usually equipped with significant amounts of metadata. This approach exploits both raw data and metadata to detect anomalous events. It is based on the probability of an edge to occur between any two nodes. This probability is a function of the linear combination of node attributes.

Anomalies in social networks

In recent years, social networks have been able to attract the interest of many researchers, who have started to study them from many points of view. A recent guide to research methods, applications and software tools related to social network analysis can be found in [148], while a review of social network analysis problems (including anomaly detection) and related applications is presented in [151]. A review of research methods for figurative language analysis in social networks can be found in [13], while the application of social network analysis to extract critical information after a disaster is considered in [383]. Plenty of applications and software tools are also available on this topic. For example, [655] discusses the integration of het-

erogeneous social networks; [351] analyzes the search of opinion leaders in social networks; while [52] investigates recommendation techniques in this context.

Recently, some authors have started to study scenarios in which several social networks interact with each other to allow their users to achieve certain goals [134]. In past literature, different terms have been used to refer to this context, including multilayer social networks [110], cross platform online social networks [598], multi social networks [461], and Social Internetworking Scenarios [134]. This is a highly investigated field, since the number of users who simultaneously interact with multiple social networks is constantly growing. For instance, in [110], new forms of anomalies emerging in multi-layer social networks are investigated. In [598], the authors propose an approach that exploits an intelligent-sensing model for analyzing behavioral variations in multiple social networks. In it, controlled faulty data, referred to as cognitive tokens, are intentionally introduced in the information flow for attracting anomalous users. The authors show that the same approach could also be applied to a *single* IoT scenario.

The MIoT environment represents the extension to smart objects and the IoTs of social internetworking scenarios [82]. Indeed, users joining multiple social networks can be assimilated to objects belonging to different IoTs, although the data type and nature, and the kind of issues to be addressed, are rather different.

Other approaches

Several recent approaches on anomaly detection exploit classification through machine learning-based and/or neural network-based engines [524, 44, 119, 653, 507, 288, 500]. Due to the intrinsic nature of these engines, the corresponding approaches do not construct an explicit model of anomalies. This way of proceeding is complementary and dual with respect to the one adopted in our approach which, indeed, aims at modeling anomalies in new MIoT scenarios.

Classification of our approach

After having examined the literature about anomalies, we can compare our approach with the most related ones, which have been introduced above. For this purpose, we consider some comparison properties, namely: *(i)* the ability of handling more networks; *(ii)* the usage of a unified scheme; *(iii)* the ability of managing labeled networks; *(iv)* the ability of handling dynamic networks; *(v)* the exploitation of additional metadata; and *(vi)* the usage of structural properties. Based on these features, our approach compares to the and the most related studies, as shown in Table 8.1.

	Capability of handling more networks	Usage of a unified scheme	Capability of managing labeled networks	Capability of handling dynamic networks	Exploitation of additional metadata	Usage of structural properties
Our approach	✓	✓	✓	✓	✓	✓
[607]	-	✓	-	-	-	✓
[30]	-	-	✓	-	-	✓
[338]	-	-	-	-	-	✓
[189]	-	-	✓	✓	-	✓
[478]	-	-	✓	✓	✓	-
[110]	✓	-	-	-	-	✓
[598]	✓	-	✓	✓	-	-

Table 8.1: Comparison between our approach and the most related ones

8.3 Methods

8.3.1 Extending the MIoT paradigm

In this section, we extend the MIoT paradigm introduced in Chapter 4 in order to make it capable of representing and handling anomalies.

Given a MIoT $\mathcal{M} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$, and pair of instances ι_{j_k} of o_j and ι_{q_k} of o_q in $\mathcal{I}_k$, the MIoT saves the set TrS_{jq_k} of the transactions from ι_{j_k} to ι_{q_k}. It is defined as:

$$TrS_{jq_k} = \{Tr_{jq_{k_1}}, Tr_{jq_{k_2}}, \cdots, Tr_{jq_{k_v}}\} \tag{8.1}$$

A transaction $Tr_{jq_{k_z}} \in TrS_{jq_k}$ is represented as follows:

$$Tr_{jq_{k_z}} = \langle st_{jq_{k_z}}, fh_{jq_{k_z}}, ok_{jq_{k_z}}, ct_{jq_{k_z}} \rangle \tag{8.2}$$

Here:

- $st_{jq_{k_z}}$ denotes the starting timestamp of $Tr_{jq_{k_z}}$.
- $fh_{jq_{k_z}}$ indicates the ending timestamp of $Tr_{jq_{k_z}}$.
- $ok_{jq_{k_z}}$ denotes whether $Tr_{jq_{k_z}}$ was successful or not; it is set to `true` in the affirmative case, to `false` in the negative one, and to `NULL` if it is still in progress.
- $ct_{jq_{k_z}}$ indicates the set of the content topics considered by $Tr_{jq_{k_z}}$. Specifically, it consists of a set of w keywords:

$$ct_{jq_{k_z}} = \{kw^1_{jq_{k_z}}, kw^2_{jq_{k_z}}, \ldots, kw^w_{jq_{k_z}}\} \tag{8.3}$$

An important subset of TrS_{jq_k} is $TrOkS_{jq_k}$, which stores the successful transactions of TrS_{jq_k}. It is defined as:

$$TrOkS_{jq_k} = \{Tr_{jq_{k_z}} | Tr_{jq_{k_z}} \in TrS_{jq_k}, ok_{jq_{k_z}} = \texttt{true}\} \tag{8.4}$$

In other words, this set comprises all the transactions through which ι_{q_k} gave a positive answer to a request of ι_{j_k}, thus providing this last one with services, information or data it required.

Now, we can define the set TrS_{j_k} of the transactions activated by ι_{j_k} in $\mathcal{I}_k$. Specifically, let $\iota_{1_k}, \iota_{2_k}, \cdots, \iota_{w_k}$ be all the instances belonging to $\mathcal{I}_k$. Then:

$$TrS_{j_k} = \bigcup_{q=1..w, q \neq j} TrS_{jq_k} \tag{8.5}$$

This means that the set TrS_{j_k} of the transactions of an instance ι_{j_k} is given by the union of the sets of the transactions from ι_{j_k} to all the other instances of $\mathcal{I}_k$.

We should note that, herein, we have reported only those aspects of the MIoT paradigm that are strictly necessary for our aim. The interested reader can find further details in [82].

We can now introduce the concept of neighborhood of an instance ι_{j_k} in $\mathcal{I}_k$. Specifically, the neighborhood Nbh_{j_k} of ι_{j_k} is defined as:

$$Nbh_{j_k} = ONbh_{j_k} \cup INbh_{j_k} \tag{8.6}$$

where:

$$\begin{aligned} ONbh_{j_k} &= \{n_{q_k} | (n_{j_k}, n_{q_k}) \in A_I, |TrS_{jq_k}| > 0\} \\ INbh_{j_k} &= \{n_{q_k} | (n_{q_k}, n_{j_k}) \in A_I, |TrS_{qj_k}| > 0\} \end{aligned} \tag{8.7}$$

In other words, Nbh_{j_k} comprises those instances directly connected to ι_{j_k} through an incoming or an outgoing arc, which shared at least one transaction with it.

Finally, we can define the concept of neighborhood of an i-arc $a_{jq_k} = (n_{j_k}, n_{q_k}) \in A_I$. Specifically, the neighborhood Nbh_{jq_k} of the i-arc a_{jq_k} is defined as:

$$Nbh_{jq_k} = ONbh_{jq_k} \cup INbh_{jq_k} \tag{8.8}$$

where:

$$\begin{aligned} ONbh_{jq_k} &= \{(n_{q_k}, n_{r_k}) | (n_{q_k}, n_{r_k}) \in A_I\} \\ INbh_{jq_k} &= \{(n_{l_k}, n_{j_k}) | (n_{l_k}, n_{j_k}) \in A_I\} \end{aligned} \tag{8.9}$$

Hence, $ONbh_{jq_k}$ contains all the arcs of A_I having n_{q_k} as source node, whereas $INbh_{jq_k}$ comprises all the arcs of A_I having n_{j_k} as target node.

8.3.2 Modeling anomalies in a MIoT

In this section, we propose a model allowing for the representation and management of anomalies in MIoTs. The core of our model consists of some possible taxonomies characterizing anomalies in this scenario. Each one will correspond to different analysis viewpoints. Borrowing a terminology typical in data analysis, these taxonomies can be seen as different dimensions of a multi-dimensional model, through which the fact "anomalies in a MIoT" can be investigated. Here, we consider three of these taxonomies, namely: *(i)* presence anomalies vs success anomalies; *(ii)* hard anomalies vs soft anomalies; *(iii)* contact anomalies vs content anomalies. However, we do not exclude that other taxonomies may also be possible in future works.

Continuing with the analogy between our taxonomies and the dimensions of a multi-dimensional model, we have that each combination of the possible values of these dimensions gives rise to a specific type of anomaly to study. Therefore, we have the *Presence-Hard-Contact Anomalies*, the *Success-Hard-Content Anomalies*, and so on. In the following subsections, we briefly illustrate each taxonomy and, then, provide a formalization for some types of combined anomalies. We point out again that the description of our taxonomies is orthogonal to specific anomaly detection techniques. In order to keep the formalization as clear as possible, we will focus on a simple anomaly detection scheme based on frequencies. However, more complex detection schemes may certainly be applied to our taxonomies.

8.3.2.1 Definition of anomaly taxonomies

Presence Anomalies vs Success Anomalies

A *presence anomaly* denotes that there is a strong variation (i.e., *increase* or *decrease*) in the number of transactions carried out from an instance ι_{j_k} to an instance ι_{q_k} in a unit of time. A *success anomaly* shows that, although there is no presence anomaly from ι_{j_k} to ι_{q_k}, there is a strong *decrease* in the number of *successful* transactions from ι_{j_k} to ι_{q_k} in a unit of time.

Hard Anomalies vs Soft Anomalies

A *hard anomaly* indicates that the frequency of successful transactions carried out from an instance ι_{j_k} to an instance ι_{q_k} is higher than (or lower than) a certain threshold. A *soft anomaly* happens when the frequency of the (successful) transactions ranges between the maximum and the minimum thresholds but, for several consecutive instances of time, it is higher (resp., lower) than the mean of these two thresholds and it shows a monotone increasing (resp., decreasing) trend. The rationale

underlying this taxonomy is that hard anomalies are indicators of faults, whereas soft anomalies are indicators of a slow, but constant, degradation. Soft anomalies are extremely precious in applications such as predictive maintenance.

Contact Anomalies and Content Anomalies

A *contact anomaly* from an instance ι_{j_k} to an instance ι_{q_k} considers only the presence or the absence of transactions. By contrast, a *content anomaly* takes the content exchanged in the corresponding transactions into account[2]. Here, we assume that we are capable of identifying possible synonymies or homonymies relating terms. This is a well-known problem in the cooperative information system research field and several thesauruses have been proposed for this purpose. In this chapter, unless otherwise specified, we will refer to Babelnet [498], which is among the most advanced thesauruses. As far as content anomalies are concerned, a reference content set, consisting of some keywords, is necessary for verifying variations with respect to the content of the involved transactions. Two variants of content anomalies can be considered, namely: *(i)* the *strict* content anomalies, where the whole set of the reference keywords must be present in the involved transactions, and *(ii)* the *loose* content anomalies, where at least one of the reference keywords must be present therein.

8.3.2.2 Formalization of anomalies

The combination of the three taxonomies introduced above gives rise to eight possible kinds of anomaly. In the following, we provide the formal definition for representative cases. We recall that, for the sake of clarity, in these definitions we consider frequencies as the basic factor for anomaly detection. However, we point out that, even if frequencies are a well-accepted and widely adopted factor, even more complex factors could easily be incorporated into our taxonomies.

In the next subsections, we present a formalization of a representative selection of the eight anomaly types, providing the method for computing their anomaly degrees. We have not included the formalization for all cases, due to brevity. Yet, their definition would be analogous and straightforward.

The kinds of anomaly that we formalize below include: *(i)* Presence-Hard-Contact anomalies, *(ii)* Success-Hard-Contact anomalies, *(iii)* Presence-Soft-Contact anomalies, and *(iv)* Presence-Hard-Content anomalies. In many of these definitions, the variable "time" plays a key role.

[2] Recall that, given a transaction $Tr_{jq_{k_z}}$, the corresponding content $ct_{jq_{k_z}}$ consists of a set of w keywords.

Presence-Hard-Contact Anomalies

Let t be a time instant and let Δt be a time interval (consisting of one or more time units). The frequency $TrFr_{jq_k}(t, \Delta t)$ of the transactions from ι_{j_k} to ι_{q_k} can be defined as follows:

$$TrF_{jq_k}(t, \Delta t) = \frac{|\{Tr_{jq_{k_z}} \mid Tr_{jq_{k_z}} \in TrS_{jq_k}, st_{jq_{k_z}} \geq t, fh_{jq_{k_z}} \leq (t + \Delta t)\}|}{\Delta t} \tag{8.10}$$

In other words, TrF_{jq_k} is given by the ratio between the number of transactions from ι_{j_k} to ι_{q_k} exchanged in the time interval $[t, t + \Delta t]$ to the length of this time interval (i.e., Δt).

We say that there is a Presence-Hard-Contact anomaly from ι_{j_k} to ι_{q_k} in the time interval $[t, t + \Delta t]$ if:

- TrF_{jq_k} is higher than a certain threshold th_{max}, in which case the anomaly degree is defined as $\alpha_{jq_k}(t, \Delta t) = \frac{TrF_{jq_k}(t,\Delta t) - th_{max}}{th_{max}}$, or
- TrF_{jq_k} is lower than a certain threshold th_{min} and this inequality does not hold in the time instants preceding t. This last condition is necessary to avoid that the lack of transactions from ι_{j_k} to ι_{q_k} is erroneously interpreted as a presence anomaly, as it would the case for instance when two instances have never performed transactions between them in the past. In this case, the anomaly degree is defined as $\alpha_{jq_k}(t, \Delta t) = \frac{th_{min} - TrF_{jq_k}(t,\Delta t)}{th_{min}}$.

If no Presence-Hard-Contact anomaly is detected, $\alpha_{jq_k}(t, \Delta t)$ is set to 0.

Here and in the following, the thresholds th_{max} and th_{min} can either be static or are dynamically computed over the previous observations. For instance, they could be computed considering both the mean and the standard deviation observed for TrF_{jq_k} in a predefined period of time. However, their actual definition depends on the application domain.

Presence-Hard-Contact anomalies focus on anomalies detected in the number of transactions (*presence*) occurring between two *instances* in a MIoT without considering the content they share (*contact*) and focusing on sharp variations of observed values (*hard*).

Their detection could be particularly relevant, for example, to identify faults concerning the ability of a MIoT object to send data. This may happen, for instance, because an object is no longer working.

Here and in the following, thanks to the concept of MIoT, anomalies between pairs of instances can be used to compute anomalies between the corresponding pairs of objects. In particular, given two objects o_j and o_q, let $\mathcal{IS}_{jq}$ be the set of IoTs containing instances of both o_j and o_q connected by an *i-arc*. The anomaly degree $\alpha_{jq}(t, \Delta t)$ between the pair of objects o_j and o_q in a MIoT can be defined as:

$$\alpha_{jq}(t, \Delta t) = \frac{\sum_{\mathcal{I}_k \in \mathcal{IS}_{jq}} \alpha_{jq_k}(t, \Delta t)}{|\mathcal{IS}_{jq}|} \tag{8.11}$$

This way of computing anomalies between pairs of objects in a MIoT, starting from the anomalies of the corresponding pairs of instances, is valid for all kinds of anomalies.

Success-Hard-Contact Anomalies

Similarly to what we have done for Presence-Hard-Contact anomalies, we first define the frequency $TrOkF_{jq_k}(t, t + \Delta t)$ of the transactions from ι_{j_k} to ι_{q_k} that occurred successfully in the time interval $[t, t + \Delta t]$ as:

$$TrOkF_{jq_k}(t, \Delta t) = \frac{|\{Tr_{jq_{k_z}} \mid Tr_{jq_{k_z}} \in TrOkS_{jq_k}, st_{jq_{k_z}} \geq t, fh_{jq_{k_z}} \leq (t + \Delta t)\}|}{\Delta t} \tag{8.12}$$

Now, we can say that, in the time interval $[t, t + \Delta t]$, there is a Success-Hard-Contact anomaly if:

- there is no Presence-Hard-Contact anomaly in the same time interval;
- $TrOkF_{jq_k}$ is lower than a certain threshold th'_{min}.

In this case, the anomaly degree is defined as $\alpha_{jq_k}(t, \Delta t) = \frac{th'_{min} - TrOkF_{jq_k}(t, \Delta t)}{th'_{min}}$. Otherwise, $\alpha_{jq_k}(t, \Delta t) = 0$.

Success-Hard-Contact anomalies are very similar to Presence-Hard-Contact anomalies. However, they focus on the fraction of successful transactions occurring between two instances in a MIoT (*success*); they disregard the content exchanged by transactions (*contact*) and focus on sharp variations of observed values (*hard*).

The detection of this kind of anomaly might be particularly relevant, for example, in recognizing possible difficulties of a MIoT object to deliver requested data. Differently from the previous case, this may happen because there is an issue in the network rather than in the object itself.

Presence-Soft-Contact Anomalies

Let t be a time instant, let Δt be a time interval and let τ be a positive integer representing the number of time units after t into consideration (generally, $\tau \gg \Delta t$), and let $th_{avg} = \frac{th_{min} + th_{max}}{2}$. We can say that, in the time interval $[t, t + \tau]$, there is a Presence-Soft-Contact anomaly if, for each time instant θ such that $t \leq \theta \leq t + \tau$, the following conditions hold:

- $th_{min} \leq TrF_{jq_k}(\theta, \Delta t) \leq th_{max}$, which implies that no Presence-Hard-Contact anomaly exists in the time interval into consideration;

- $TrF_{jq_k}(\theta,\Delta t) > th_{avg}$ (resp., $TrF_{jq_k}(\theta,\Delta t) < th_{avg}$), which denotes that the frequency of the transactions from ι_{j_k} to ι_{q_k} is always higher (resp., smaller) than the average between th_{min} and th_{max};
- $TrF_{jq_k}(\theta+1,\Delta t) \geq TrF_{jq_k}(\theta,\Delta t)$ (resp., $TrF_{jq_k}(\theta+1,\Delta t) \leq TrF_{jq_k}(\theta,\Delta t)$), which implies that the frequency of the transactions from ι_{j_k} to ι_{q_k} is monotonically increasing (resp., decreasing) in the time interval Δt of interest.

If an anomaly is detected, the corresponding anomaly degree $\alpha_{jq_k}(t,\Delta t)$ is set to $\alpha_{jq_k}(t,\Delta t) = \frac{|TrF_{jq_k}(t+\tau,\Delta t)-th_{avg}|}{th_{avg}}$. Otherwise, $\alpha_{jq_k}(t,\Delta t) = 0$.

Presence-Soft-Contact anomalies focus on a smooth (*soft*) decrease in the number of all (*presence*) the transactions exchanged between two instances of a MIoT, without considering the exchanged content (*contact*).

The detection of this kind of anomaly may be useful in identifying a slowly but constantly changing behavior of an object. For instance, it could regard an object that is wearing out, an equipment whose battery has a very low charge level, and so forth.

Presence-Hard-Content Anomalies

Let $\overline{ct}$ be a content consisting of (presumably very few) keywords. We define the set $sTrCtS_{jq_k}(\overline{ct})$ of the transactions from ι_{j_k} to ι_{q_k} *strictly adherent* to $\overline{ct}$, i.e., the set of the transactions from ι_{j_k} to ι_{q_k} that contain *all the keywords* of $\overline{ct}$ as follows:

$$sTrCtS_{jq_k}(\overline{ct}) = \{Tr_{jq_{k_z}} \mid Tr_{jq_{k_z}} \in TrS_{jq_k}, \overline{ct} \subseteq ct_{jq_{k_z}}\} \tag{8.13}$$

As previously pointed out, here we assume that we are capable of identifying possible synonymies or homonymies relating a term of $\overline{ct}$ with a term of $ct_{jq_{k_z}}$. For this purpose, we use Babelnet [498].

Consider, now, a content $\overline{ct}$ consisting of some keywords. We define the set $lTrCtS_{jq_k}(\overline{ct})$ of the transactions from ι_{j_k} to ι_{q_k} that are *loosely adherent* to $\overline{ct}$, i.e., the set of the transactions from ι_{j_k} to ι_{q_k} that contain *at least one keyword* of $\overline{ct}$ as follows:

$$lTrCtS_{jq_k}(\overline{ct}) = \{Tr_{jq_{k_z}} \mid Tr_{jq_{k_z}} \in TrS_{jq_k}, (\overline{ct} \cap ct_{jq_{k_z}}) \neq \emptyset\} \tag{8.14}$$

Let t be a time instant and let Δt be a time interval. By applying the same approach described for Presence-Hard-Contact anomalies, it is possible to define the frequency $sTrCtF_{jq_k}(\overline{ct})$ (resp., $lTrCtF_{jq_k}(\overline{ct})$) of the transactions from ι_{j_k} to ι_{q_k} strictly (resp., loosely) adherent to $\overline{ct}$. Then, it is possible to state that, in the time interval $[t, t+\Delta t]$, there is a strict (resp., loose) Presence-Hard-Content anomaly from ι_{j_k} to ι_{q_k} against $\overline{ct}$ if:

- $sTrCtF_{jq_k}(\overline{ct})$ (resp., $lTrCtF_{jq_k}(\overline{ct})$) is higher than a certain threshold th_{max}, or
- $sTrCtF_{jq_k}(\overline{ct})$ (resp., $lTrCtF_{jq_k}(\overline{ct})$) is lower than a certain threshold th_{min} and this inequality does not hold in the time instants preceding t.

Analogously to what we have done for Presence-Hard-Contact anomalies, if the first condition is verified, the anomaly degree $\alpha_{jq_k}(t, \Delta t)$ can be defined as $\alpha_{jq_k}(t, \Delta t) = \frac{sTrCtF_{jq_k}(\overline{ct}) - th_{max}}{th_{max}}$, for strictly adherent anomalies, and $\alpha_{jq_k}(t, \Delta t) = \frac{lTrCtF_{jq_k}(\overline{ct}) - th_{max}}{th_{max}}$, for loosely adherent ones. Instead, if the second condition is verified, then $\alpha_{jq_k}(t, \Delta t) = \frac{th_{min} - sTrCtF_{jq_k}(\overline{ct})}{th_{min}}$, for strictly adherent anomalies, and $\alpha_{jq_k}(t, \Delta t) = \frac{th_{min} - lTrCtF_{jq_k}(\overline{ct})}{th_{min}}$ for loosely adherent ones. $\alpha_{jq_k}(t, \Delta t) = 0$ in all the other cases.

Presence-Hard-Content anomalies focus on sharp variations (*hard*) in the number of transactions (*presence*) exchanged between two instances in a MIoT, with regard to a certain set of contents (*content*).

The study of content variations paves the way to a wide variety of analyses, ranging from variations in the interests of a user who is adopting the MIoT objects, to variations in the sentiment of a user on a specific topic/service provided through the MIoT objects.

The other kinds of anomaly, whose formalization we have not reported in this chapter because they are very similar to the ones considered above, would provide four further viewpoints of the possible anomalies existing in a MIoT. It would be straightforward to see how these extra anomalies would allow us to model other possible real-world cases, which shows the generic applicability of our approach (three taxonomies and a multi-dimensional perspective).

8.3.3 Investigating the origins and effects of anomalies in a MIoT

After providing a multi-dimensional taxonomy of the possible anomalies present in a MIoT, in this section we aim at investigating their origins and effects. For this purpose, we address two problems that, according to what happens in several other research fields, we dubbed “forward problem” and “inverse problem”, respectively. In the forward problem, given one or more anomalies, we aim at analyzing their effects on a MIoT. In the inverse problem, which is traditionally more complex than the forward one, given the effects of one or more anomalies on the nodes and the arcs of a MIoT, we aim at detecting the origin(s) of them, i.e., the node(s) or the arc(s) from which anomalies have started.

8.3.3.1 Forward Problem

As previously pointed out, this problem aims at understanding the effects that one or more anomalies have on the nodes of a MIoT. In the following, we will investigate the forward problem for one kind of anomaly, namely the Presence-Hard-Contact anomaly. However, all our results can be extended to all the other cases introduced in Section 8.3.2.

First, given a node n_{j_k} of an IoT $\mathcal{I}_k$, along with the anomaly degrees of its outgoing arcs, in the forward problem we want to compute the overall effects of these anomalies over the corresponding IoT, $\mathcal{I}_k$. Specifically, the degree $\delta_{j_k}(t, \Delta t)$ of the anomalies of n_{j_k} in the time instant t and in the time interval Δt depends on the number of nodes belonging to $ONbh_{j_k}$ and, for each of these nodes n_{q_k}, on the degree $\delta_{q_k}(t, \Delta t)$ of the anomalies involving it and on the anomaly degrees measured for the corresponding arcs.

We wish to observe that, by saying that the degree of the anomalies of a node n_{j_k} recursively depends on the degree of the anomalies of the nodes belonging to $ONbh_{j_k}$, we introduce a way of proceeding that is similar to the one underlying the definition of the PageRank [523]. Thus, to compute δ_{j_k}, it is possible to adapt the formula for the computation of the PageRank to our scenario. Specifically:

$$\delta_{j_k}(t, \Delta t) = \gamma + (1 - \gamma) \cdot \frac{\sum_{n_{q_k} \in ONbh_{j_k}} \delta_{q_k}(t, \Delta t) \cdot \alpha_{jq_k}(t, \Delta t)}{\sum_{n_{q_k} \in ONbh_{j_k}} \alpha_{jq_k}(t, \Delta t)} \tag{8.15}$$

This formula says that the degree $\delta_{j_k}(t, \Delta t)$ of the anomalies of n_{j_k} in the time instant t and in the time interval Δt is obtained by summing two components:

- The former component, γ, is the damping factor generally existing in each approach based on PageRank. It ranges in the real interval [0,1] and denotes the minimum absolute anomaly degree that can be assigned to a node of the MIoT.
- The second component, is a weighted sum of the anomaly degree $\delta_{q_k}(t, \Delta t)$ of the nodes n_{q_k} directly connected to n_{j_k} and, therefore, belonging to $ONbh_{j_k}$. The weight of each anomaly degree $\delta_{q_k}(t, \Delta t)$ is given by the value of the parameter α_{jq_k}, which considers the fraction of anomalous transactions performed from n_{j_k} to n_{q_k}.

In this formula, $\delta_{j_k}(t, \Delta t)$ ranges in the real interval [0,1].

The above formula allows us to determine the effects of a faulty node over the corresponding IoT, and consequently on the whole MIoT (as will become clearer next). However, we observe that the current formalization is valid only in the presence of a single faulty node. When multiple nodes simultaneously exhibit some anomalous behavior in one IoT (of the MIoT), our approach fails to distinguish among

the contributions of each anomaly, particularly when the effects are measured in a single node. We wish to point out that this is our very first attempt to investigate MIoT anomalies, proposing a method to evaluate their effects. Our next priority as a follow-up of the present study, will be extending our method accordingly.

Having investigated the effects of an anomaly of an *instance* in an IoT, we can now exploit the features of the MIoT paradigm to analyze the effects of an anomaly of an *object* in a MIoT. In particular, the anomaly degree $\delta_j(t, \Delta t)$ of an object o_j can be computed starting from the anomaly degrees of its instances. Specifically, given the set $\mathcal{IS}_j$ of the IoT containing instances of o_j, $\delta_j(t, \Delta t)$ can be computed as:

$$\delta_j(t, \Delta t) = \frac{\sum_{\mathcal{I}_{j_k} \in \mathcal{IS}_j} \delta_{j_k}(t, \Delta t)}{|\mathcal{IS}_j|} \tag{8.16}$$

We observe that the value of $\delta_j(t, \Delta t)$, if compared with the one of $\delta_{j_k}(t, \Delta t)$, can provide very useful information. In particular, if $\delta_j(t, \Delta t)$ is very similar to $\delta_{j_k}(t, \Delta t)$ for each IoT $\mathcal{I}_{j_k} \in \mathcal{IS}_j$, we can conclude that o_j is really a source of anomaly. Instead, if the standard deviation of $\delta_j(t, \Delta t)$ is high, then we can conclude that o_j is involved in, or affected by, some anomalies in one or more IoTs, but not in some other ones.

8.3.3.2 Inverse Problem

As previously pointed out, the inverse problem is traditionally more complex than the forward one. For this reason, we will focus only on the simplest scenario, i.e., the case in which there is only one anomaly in the MIoT. In the future, we plan to extend our investigation to more complex scenarios. Let $a_{jq_k} = (n_{j_k}, n_{q_k})$ be an i-arc of a MIoT presenting an anomaly whose origin is not known. In the inverse problem we want to detect this origin.

First of all, we must verify if the origin of the anomaly is just a_{jq_k}. For this purpose, we consider the "siblings" of a_{jq_k}, i.e., the other arcs having n_{j_k} as the source node and the other arcs having n_{q_k} as the target node. If none of these present anomalies, then it is possible to conclude that a_{jq_k} is the origin of the observed anomaly and that this last one did not affect other nodes or arcs of the MIoT. In this case, the inverse problem has been solved and the investigation terminates.

However, the situation described above is very particular and, also, quite rare. More typically, anomalies tend to affect multiple nodes and arcs. In that case, given an anomaly found in an arc a_{jq_k}, in order to detect its origin, the first step consists in computing the anomaly degrees of n_{j_k} and n_{q_k} and to choose the maximum between the two. This becomes the current node under investigation.

At this point, an iterative process, aiming at finding the origin of the observed anomaly, is activated. During each step of this process, we apply the PageRank-based

formula for the computation of the anomaly degree of a node, as discussed in Section 8.3.3.1, to all the nodes of the $ONbh$ and the $INbh$ of the current node. After this, we select the node having the maximum anomaly degree. If the degree of this node is higher than the one of the current node, it becomes the new current node and a new iteration starts. Otherwise, our approach concludes that the current node is the origin of the anomaly under consideration.

Clearly, the approach described above is greedy and, therefore, must be intended as a heuristic that could return a local maximum, instead of a global one. However, it is possible to apply to this approach all the techniques for improving the accuracy of a greedy approach already proposed in past literature, spanning from meta heuristics, such as hill climbing [572], to evolutionary optimization algorithms [609].

For instance, if the MIoT is not excessively large, it could be possible to compute the anomaly degree of all its nodes by applying the PageRank-based approach described in Section 8.3.3.1. In this case, the node having the maximum value of anomaly degree would be selected as the anomaly origin. This would correspond to applying an approach returning the optimum solution to the inverse problem, instead of one returning an approximate solution.

On the opposite extreme, if the network is very large, and the anomaly is affecting a vast portion of it, the greedy approach may be prohibitive. In this case, we will need to find an additional way to stop the iterative process, particularly when resources are limited and the process does not stop because, at each iteration, it continues to return a new current node with an anomaly degree higher than the one of the previous iteration. For instance, we could define a maximum number of iterations or a minimum increase of the anomaly degree necessary to activate a further iteration. Furthermore, this required minimum increase could be dynamic and could vary based on the number of steps already performed.

We conclude this section with an important consideration. Since this is our first effort to investigate the inverse problem, we had the necessity to limit our analysis to only one case, i.e., the one in which, in a certain time instant, there is only one anomaly in the MIoT. If at a given time instant, there are more anomalies in the MIoT, the search of the corresponding origins becomes much more complex, because the anomalies could interfere with each other. These interferences could make the search of the anomaly sources extremely complex.

For instance, we argue that, in presence of two anomalies whose source nodes are not known, in case these two nodes were relatively close to each other, the examination of the anomaly degree of their neighbors could be extremely beneficial. In fact, in this scenario, some of these neighbors are influenced only by one anomaly; other ones are influenced only by the other anomaly; a third group of neighbors is influ-

enced by both anomalies; finally, a fourth group is not influenced by any anomalies. By deeply analyzing what happens in these four groups of nodes, it could be possible to derive precious information leading us to identify the sources of the two anomalies. In the future, we plan to conduct specific and accurate investigations about this case, and several other ones possibly characterizing the inverse problem.

8.4 Results

8.4.1 Testbed

To perform this analysis, we considered a reference scenario related to a smart city context. To model it, and to test our approach, we constructed a prototype. Furthermore, we realized a MIoT simulator.

In order to make "concrete" and "plausible" the simulated MIoT, our simulator needs to generate MIoTs having the characteristics specified by the user, whilst being as close as possible to real-world scenarios. In the simulator design, and in the construction of the MIoT used in the experiments, we followed the guidelines outlined in [304, 73, 74], where the authors highlight that one of the main factors used to build links in an IoT is node proximity.

In order to reproduce the creation of transactions among objects, we decided to leverage information about a simulated smart city context. As for a dataset containing real-life paths in a smart city, we selected the one reported in `http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html`. This regards movements of objects, in terms of routes, in the city of Porto from July 1^{st} 2013 to June 30^{th} 2014. Each route contains several Points of Interest, corresponding to the GPS coordinates of each object as it moves in Porto. With this information at hand, our simulator associates an object (thus, creating a node) with one of the routes recorded in the dataset. Furthermore, it creates an arc between two nodes when the distance between the corresponding routes is less than a certain threshold th_d, for a predefined time interval th_t. The value of th_d and th_t can be specified through the constructor interface. Clearly, the higher is this value the more connected the constructed MIoT will be. When we defined the distribution of the transactions among the nodes, we leveraged scientific literature and used the corresponding results to properly tune our simulator. In particular, we adopted the values reported in [301].

The interested reader can find the MIoT created by our simulator for the experiments at the Web address `http://daisy.dii.univpm.it/miot/datasets/anomaly-detection`. It consists of 1,256 nodes and six IoTs having 128, 362, 224, 280, 98 and 164 nodes, respectively. The constructed MIoT is returned in a format that can

be directly processed by the cypher-shell of Neo4J. Some statistics about our dataset are reported in Table 8.2.

Parameter	*Value*
Number of nodes	1,256
Number of relationships	6,860
Mean outdegree	5.44
Mean indegree	5.58

Table 8.2: Parameter values for our simulator

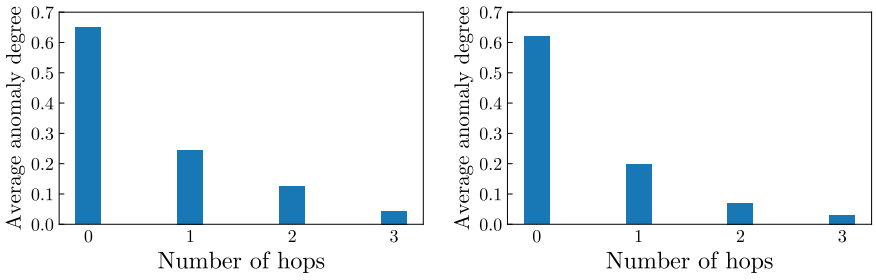
We carried out all the tests presented in this section on a server equipped with an Intel I7 Quad Core 7700 HQ processor and 16 GB of RAM, with the Ubuntu 16.04 operating system. To implement our approach, we adopted Python, as programming language, and Neo4J (Version 3.4.5), as underlying DBMS.

8.4.2 Analysis of the forward problem

Let us preliminarily define the concept of "number of hops" h_{jq_k} between the node n_{j_k} and another node n_{q_k} as the minimum number of arcs of the MIoT that must be traversed in order to reach n_{q_k} from n_{j_k}.

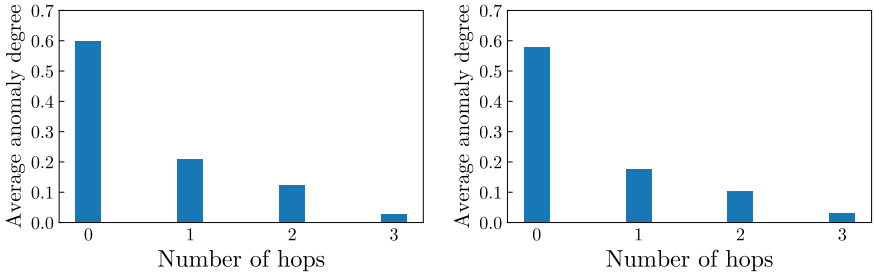
In a first step we analyzed the effects that the anomalous behavior of an object o_j had on the nodes of a MIoT. As pointed out in Sect. 8.3.3.1, given a node n_{j_k} of the IoT $\mathcal{I}_k$, its anomaly degree is represented by the parameter δ_{j_k}. This anomaly may propagate through the MIoT, thus affecting other nodes. To investigate this propagation, given an anomalous instance of an object o_j and the IoT $\mathcal{I}_k$, we measured the anomaly degree δ_{j_k} of n_{j_k} and the average of the anomaly degrees δ_{q_k} of all the nodes n_{q_k}, grouped by the number of hops from n_{j_k} to n_{q_k}. Moreover, we computed the same values but averaged through the IoT belonging to the MIoT. The same test has been run over 100 randomly chosen nodes, and results have been averaged over the runs.

Figure 8.1 shows the results obtained for Presence-Hard-Contact anomalies, while Figure 8.2 presents those regarding Presence-Soft-Contact anomalies. From the analysis of these figures it is possible to observe that the effects of an anomaly on a node spread over the surrounding nodes, even if they rapidly decrease against the number of hops. The corresponding trend follows a power law distribution. If we compare the left and the right distributions of Figures 8.1 and 8.2, we can observe that anomalies propagate more slowly on a MIoT than on a single IoT. However, this difference is negligible. Furthermore, there are no significant differences be-



(a) (b)

Fig. 8.1: Values of δ_{j_k} (corresponding to 0 hops) and average values of the anomaly degrees of all the nodes of $\mathcal{I}_k$ (on the left) and of the MIoT (on the right) being 1, 2 and 3 hops far from n_{j_k} in case of Presence-Hard-Contact anomalies



(a) (b)

Fig. 8.2: Values of δ_{j_k} (corresponding to 0 hops) and average values of the anomaly degrees of all the nodes of $\mathcal{I}_k$ (on the left) and of the MIoT (on the right) being 1, 2 and 3 hops far from n_{j_k} in case of Presence-Soft-Contact anomalies

tween Presence-Hard-Contact anomalies and Presence-Soft-Contact anomalies, except that the latter ones are slightly smaller than the former ones. This trend can be justified by considering that Presence-Soft-Contact anomalies are more difficult to be observed than Presence-Hard-Contact ones, since the former ones are not only required to show values higher (resp., lower) than a given threshold, but should also exhibit a trend that is monotonically increasing (resp., decreasing), within the time interval of interest. As the trends are very similar, in the following tests we focus only on Presence-Hard-Contact anomalies, without loss of generality.

Next, we investigated the effects that the anomaly of an object has on the other objects connected to it. In particular, given an object o_q, whose instances belong to the $ONbh$ of the instances of an anomalous object o_j in at least one IoT of the MIoT, we computed the value and the standard deviation[3] of δ_j and δ_q. We repeated this

[3] Recall that δ_j and δ_q are computed by averaging the anomaly degrees of all the instances of o_j and o_q.

task 100 times with different pairs of objects o_j and o_q. Then, we averaged the values obtained over the runs. The corresponding results are shown in Figure 8.3, under the category ALL. As we can observe, the standard deviation of δ_j is very low. This result can be explained by the fact that all the instances of the anomalous object o_j present anomalies and, consequently, the corresponding anomaly degrees are almost uniform. By contrast, the value of δ_q is lower than the one of δ_j, exhibiting a very high standard deviation. This is explained by observing that the instances of o_q are not in the neighborhoods of the instances of o_j in all the IoTs of the MIoT. In fact, in some of them, they can be 2, 3 or more hops away from the instances of o_j. In some cases, they may even be disconnected from the instances of o_j.

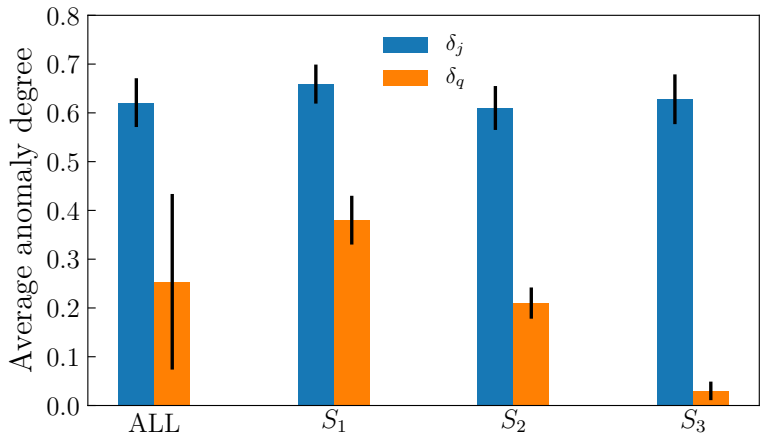


Fig. 8.3: Anomaly degrees and the corresponding standard deviations in different scenarios

As a next step, we repeated the previous experiment, enforcing some extra constraints, which defined three different scenarios. In the first (resp., second, third) one, all the instances of o_q were 1 (resp., 2, more than 2) hop(s) far from the instances of o_j; the third scenario includes also instances of o_q not connected to instances of o_j. The results obtained are shown in Figure 8.3 under the labels S_1, S_2 and S_3, respectively. Looking at the data labelled as ALL, these results are coherent with both the ones of Figure 8.1 and the ones of Figure 8.3. We can see that the effects of a single anomaly are rapidly reduced as soon as we move away from its origin. Furthermore, this experiment confirms what we pointed out in Section 8.3.3.1, i.e., that the anomaly degree δ is a parameter that really helps detecting the object that has caused the anomaly in the first place.

At this point, we investigated the number of nodes in a MIoT that turn out to be anomalous as a consequence of a single anomaly of an object o_j. Again, we repeated this experiment 100 times. Each time, we selected an anomalous object of the MIoT.

The selected objects had different number of instances in the MIoT, ranging from 1 to 6. For each run, we computed the number of anomalous nodes detected in the MIoT. Then, we computed the averages, by grouping the cases based on the number of instances of the anomalous objects and, therefore, based on the number of IoTs of the MIoT involved in the anomaly.

The results obtained are shown in Figure 8.4, which shows how the number of anomalous nodes increases against the number of IoTs in a roughly linear way. This trend can be explained by considering that, even when the number of objects having instances in many IoTs is usually limited with respect to the number of objects having instances in few IoTs, their anomalous behavior affects numerous nodes across several IoT and, consequently, their effect is amplified. On the contrary, anomalies observed on an object having instances in only one or two IoTs are more frequent. Yet, this is counterbalanced by the fact that each of these nodes only exerts a limited and localized impact, which affects only few nodes.

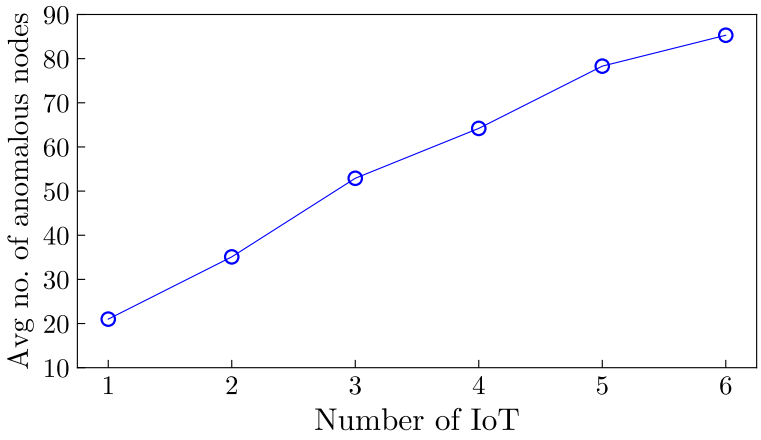


Fig. 8.4: Average number of nodes affected by anomalies against the number of IoT which an anomalous object participates to

Then, we aimed to characterize which of the node properties impacted the spread of anomalies the most. We repeated the previous experiment; but instead of choosing anomalous nodes randomly, we selected them based on their characteristics. A first characteristic that we considered was the outdegree of a node, i.e., the number of its outgoing arcs. In the various runs, we selected nodes with different outdegrees ranging from 10 to 60. For each of these values, we measured the average number of anomalous nodes throughout the MIoT detected by our approach. The results are illustrated in Figure 8.5, which clearly shows that the outdegree of anomalous nodes has a significant impact on the spread of the anomaly over the network. This

result was not surprising, since it is consistent with the results about the information diffusion in social network analysis [647].

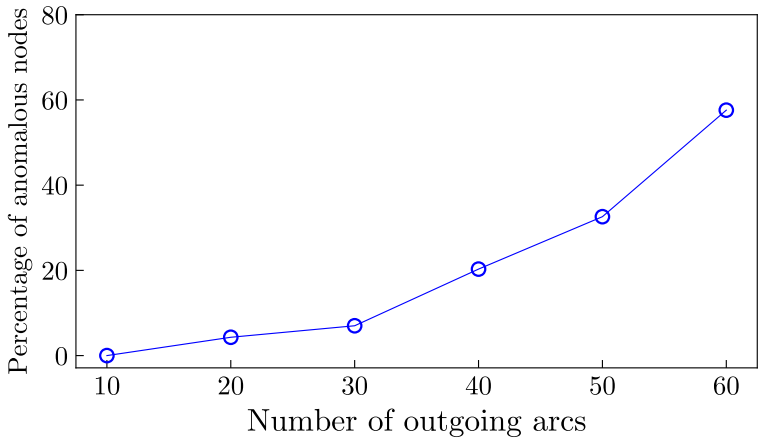


Fig. 8.5: Average percentage of anomalous nodes against their degree centrality

However, we argue that there is another form of centrality in social network analysis, which could be very promising as a node property to impact the spread of anomalies. This measure is closeness centrality. We recall that the closeness centrality of a node is defined as the reciprocal of the sum of the lengths of the shortest paths between the node itself and all the other nodes of the network.

Thus, we repeated the previous experiment; but this time we selected the anomalous nodes based on their closeness centrality. The values of this parameter for the nodes selected ranged from 0.05 to 0.45. The results obtained are shown in Figure 8.6, where we can observe that our intuition was right. Closeness centrality is really a key parameter in the spread of anomalies in a MIoT. It is even more important than degree centrality in this task. In our opinion, this result is extremely interesting because the impact of closeness centrality on anomaly diffusion is substantial, whilst the role of this parameter was a-priori much less obvious than the one of degree centrality.

As a final test on the forward problem, we evaluated the running time necessary to compute the anomaly degree δ_j of an object o_j in a MIoT against the number of its nodes. The results obtained are reported in Figure 8.7, where we can observe a polynomial (specifically, a quadratic) dependency of the running time against the number of nodes of the MIoT. This can be explained by the fact that, during the computation of the recursive formula of δ_{j_k}, the values of α_{jq_k} tend to 0 rapidly while moving away from the node n_{j_k}.

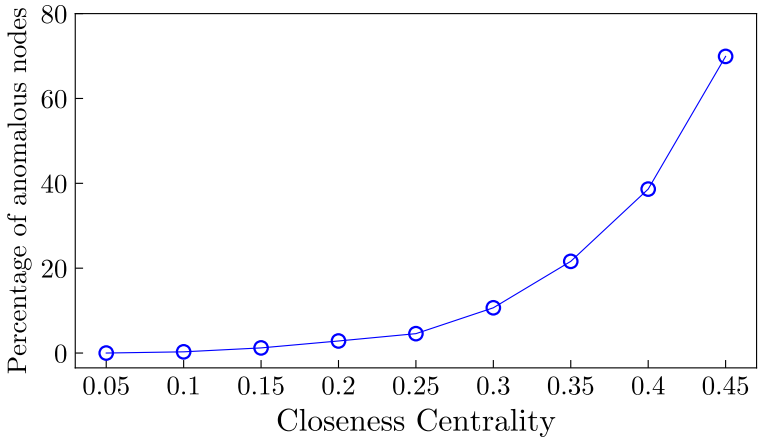


Fig. 8.6: Average percentage of anomalous nodes against their closeness centrality

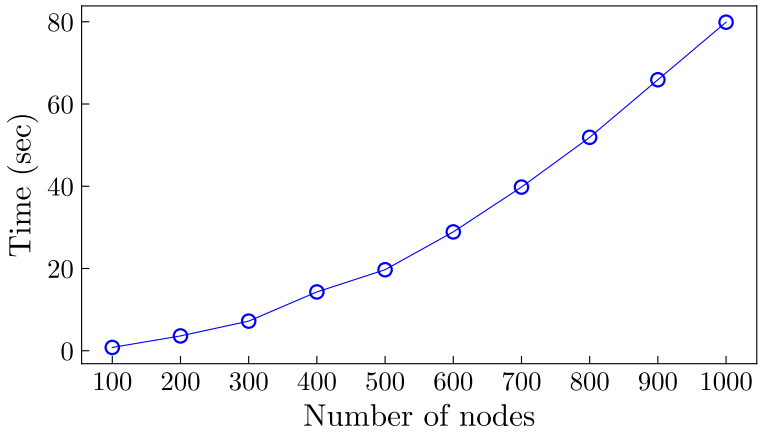


Fig. 8.7: Running time (in seconds) needed to compute δ_j in a MIoT against the number of its nodes

8.4.3 Analysis of the inverse problem

In this section, we present the results of the tests we carried out to validate our approach for solving the inverse problem. We recall that our solution to this problem starts from an i-arc of a MIoT that presents an anomaly whose origin is not known. It applies a greedy algorithm, which aims at detecting the node that originated the anomaly.

During this test, we repeated 100 times the following tasks. We simulated an anomaly on an object and, then, we randomly selected an anomalous i-arc from the whole MIoT. We applied our solution of the inverse problem on this arc and computed the following:

- the number of hits, i.e., the percentage of times our approach detected the anomaly source correctly (we call S_0 this scenario);

- the percentage of times our approach terminated in a node belonging to the $ONbh$ of the anomalous node and, therefore, being 1 hop away from it (we call S_1 this scenario);
- the percentage of times our approach terminated in a node being 2 hops far from the anomalous node (we call S_2 this scenario);
- the percentage of times our approach terminated in a node being more than 2 hops away from the anomalous node (we call S_3 this scenario).

The results obtained are reported in Figure 8.8. They show that our approach is capable of correctly identifying the anomaly source in most cases. In a fraction of cases it stops very near to the anomalous node, i.e., 1 or 2 hops away from it. The slightly higher frequency of the fourth case can be explained by the fact that the starting i-arc of the test is chosen randomly and, therefore, can be very far from the anomalous node. As a consequence, it comprises a relatively high number of cases (3, 4, 5 or more hops away from the anomalous object).

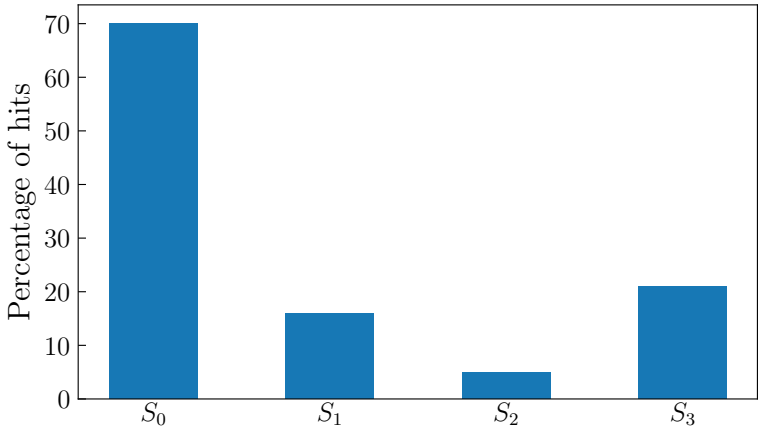


Fig. 8.8: Percentage of times when our approach correctly detects the anomaly source (indicated by the label 0) or terminates in a node being 1, 2 or more than 2 hops far from it

Next, we computed the average running time of our approach. Similarly to what we have done for the forward problem, we evaluated this time against the number of the MIoT nodes. The results obtained are shown in Figure 8.9, where we can observe that the running time increases polynomially against the number of MIoT nodes. This result can be explained by the fact that the greedy algorithm underlying our approach reaches the correct node, or a near one, in few iterations and by the fact that, on average, an anomaly on an i-arc can be observed only when this is not too far away from the node where the anomaly originated.

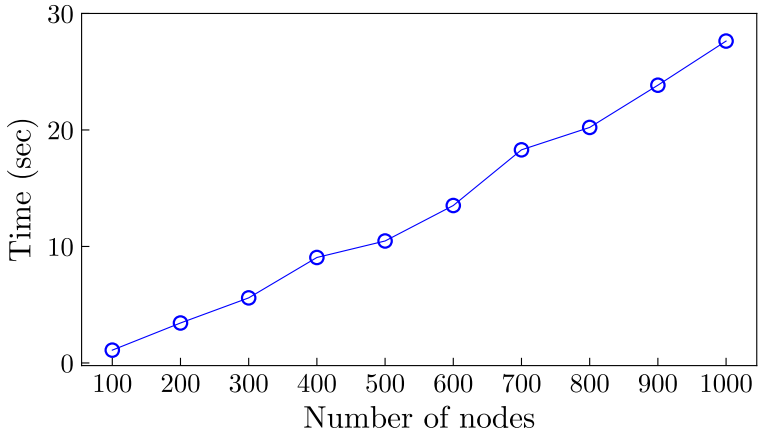


Fig. 8.9: Average running time (in seconds) of our approach for solving the inverse problem

8.5 Use case

All of the devices installed in urban infrastructures, such as smart lighting systems and traffic management ones, contribute to the ecosystem of a so called *smart community*. This last one integrates a series of technological solutions for the definition and implementation of innovative models for the smart management of urban areas. One of the main challenges of the next generation of Information and Communication Technologies (ICT) applied to smart communities is the collection, integration and exploitation of information gathered from heterogeneous data sources, including autonomous smart resources, like SO, sensors, surveillance systems, etc., and human resources, such as posts in social networks. Another key challenge is the application of artificial intelligence tools, such as the ones based on automated reasoning, to advance state-of-the-art in smart community management [162].

The use case we focus on in this section refers to a smart lighting system in a smart city. In particular, we consider a data-centric platform integrated in a smart city environment, in which data coming from sensors and social networks can boost smart lighting, by operating and tuning different smart lighting objects located in the smart city area. The aim of the whole system is to provide citizens with a smart and safe environment.

Data are gathered from three different main sources, namely sensors, social networks and alerts exchanged among citizens on a dedicated social platform. Sensors data are gathered from a set of sensors installed on each smart lamp and handle different measures, such as temperature and humidity, but also several events, such as the presence of a person or the presence of rain. Sensors and smart lamps are organized in a Wireless Sensor Area Network (WSAN). Social networks data include

geo-localized tweets from Twitter and posts from specific Facebook pages and are generated by smart personal devices.

All these data are stored in a data lake, which is directly accessed by a data mining module. This last module includes both sentiment analysis and anomaly detection tasks. The former focuses on the analysis of the data gathered from social posts. A polarity score, i.e., a positiveness/negativeness degree, is assigned to each keyword that can be extracted from a post, and is used to intercept crucial information from the citizens moving around the city. In order to unambiguously single out significant information for the application context, keywords are mapped onto a specific urban taxonomy; this task is also carried out with the support of Babelnet [498]. Furthermore, thanks to the geo-localization of posts, information regarding a specific area of the smart city can be analyzed and assigned to the correct area.

Some data mining tasks are also carried out in order to identify, among other things, situations requiring a variation in the intensity of illumination for some area, for instance because of a variation in the security level perceived by citizens therein. Each smart lamp can communicate with neighboring ones in order to report variations in lighting parameters, as received by the mining module.

Anomaly detection works on both temporal data, gathered from sensors, and polarity scores, extracted by sentiment analysis, in order to detect potential anomalies. It exploits the taxonomies and the techniques presented here (Sections 4 and 5).

In our scenario, the urban area is modeled as a MIoT consisting of a set of IoTs $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m\}$, each one associated with a portion of the area. The set of the objects of $\mathcal{M}$ comprises both the set of sensors, installed in the various smart lamps, and the set of personal devices of people who are moving around them. If an object o_j of the MIoT is active in the k^{th} portion of the urban area, it has an instance ι_{j_k} in the IoT $\mathcal{I}_k$. Clearly, when a person with a smart device o_j moves around different portions of the urban area, each one corresponding to a single IoT, o_j will have different instances, one for each IoT. An object o_j corresponding to a smart lamp sensor in the k^{th} urban area is fixed, and will contain only one instance ι_{j_k} in the corresponding IoT $\mathcal{I}_k$.

A transaction $Tr_{jq_{k_i}}$ between two object instances ι_{j_k} and ι_{q_k} can be generated in different ways. First of all, when citizens move around the various IoTs, they generate posts and alerts with their mobile devices. In this case, the transaction is associated with each post or alert. Sensors send transactions to the platform for sensed data, and smart lamps communicate with each other for parameter adjustments. Each of these events is translated into a transaction $Tr_{jq_{k_z}}$. Even the data mining module may send messages to the various smart lamps, thus generating transactions $Tr_{jq_{k_z}}$ in the MIoT.

Part III

Blockchains

In this part, we apply our complex network-based approach to model blockchains. Indeed, the interactions between wallets in blockchains can be easily modeled thanks to a complex network, which can support a complete representation of the information characterizing this scenario. This part consists of only one chapter, namely Chapter 9.

9

Speculative Bubble Investigation

In this chapter, we present a complex network-based approach to investigate user behavior during a cryptocurrency speculative bubble. Our approach is general and can be applied to any past, present and future cryptocurrency speculative bubble. To verify its potential, we apply it to investigate the Ethereum speculative bubble happened in the years 2017 and 2018. We also describe several knowledge patterns about the behavior of specific categories of users that we obtained from this investigation. Finally, we define how our approach can support the construction of an identikit of the speculators who operated during the Ethereum speculative bubble.

The material present in this chapter is taken from [118].

9.1 Introduction

On October 31st, 2008, a white paper entitled "Bitcoin: A Peer-to-Peer Electronic Cash System" by Satoshi Nakamoto was sent to a cryptography mailing list [573]. Still today, the identity of the author (or, even, authors) is not known, but we can surely recognize the incredible contribution that this paper has had to computer science. Indeed, it introduced Bitcoin, a purely peer-to-peer version of electronic cash without a third-party financial institution and, therefore, the first example of blockchain [704]. It stands as an important step towards a secure, censorship-resistant and trustful system to record transactions, store data, and so forth. A direct consequence of this technology is the concept of cryptocurrency. This is a digital medium of exchange, which leverages encryption techniques (and, therefore, a blockchain) to control the creation of monetary units and to verify the transactions made over the network.

Cryptocurrencies were the subject of a speculative bubble, similar to the tulipans' and stock market ones [670]. Indeed, the popularity of blockchains has been growing continuously from 2008, and the interest on cryptocurrencies followed the same growth. For instance, the price of Bitcoin surged almost 2,800% in four years

and has fallen by 80% in just few weeks, between the end of 2017 and the beginning of 2018, leading to a huge gain for a few people and a big loss for the majority of the investors. These events are interesting to investigate from a data science perspective, because they allow the extraction of knowledge patterns to prevent other similar cases. As a matter of fact, several studies investigate the whole speculative cryptocurrency bubble and its consequences for economy and technology [700, 292].

However, a very limited number of studies take the intrinsic nature of blockchain as a social network into account. Actually, the relationships between blockchain users are extremely relevant in the extraction of unknown patterns and in the disclosure of new viewpoints for analyzing this speculative bubble. For this reason, Social Network Analysis notions [252, 387] can provide a big help to study the relationships in the blockchain network. In this activity, it is reasonable to think of a social network in which each node indicates a user, represented through her/his blockchain address, whereas each arc denotes a transaction between two users. We argue that this social network, and the investigation perspective it makes possible, can be extremely useful to support the extraction of knowledge on the speculative bubble of the years 2017 and 2018. In this chapter, we aim at showing that this conjecture is true. In particular, we focus on the Ethereum blockchain and examine the behavior of its users [158] in these two years, which include the pre-bubble, bubble and post-bubble phases. In carrying out this task, we focus on certain categories of users, namely:

- *The power addresses*, i.e., the most active users on Ethereum, who were responsible for most of the transactions of this network. More specifically, we consider the power addresses for each of the periods of interest (i.e., the pre-bubble, bubble and post-bubble).
- *The Survivors*, i.e., those users who were power addresses in all the three periods of interest.
- *The Missings*, i.e., those users who were power addresses in the pre-bubble period and stopped being power addresses in the bubble and post-bubble periods.
- *The Entrants*, i.e., those users who were not power addresses in the pre-bubble period and became power addresses in the bubble and post-bubble periods.

Then, for each user category, we employ Social Network Analysis based techniques to identify the main characteristics that distinguish the corresponding users from the others. In this activity, the concept of ego network [229] plays an important role.

Afterwards, we check if and when there are backbones linking the users of a certain category. The presence of such backbones can be hypothesized on the basis

of the principle of homophily [468], characterizing many social networks. However, only a set of experimental analyses can indicate whether this hypothesis is true or not. Also in this case, ego networks play a key role to support analytical investigations. They are flanked by k-cores [237], which help in giving a graphical idea of the analytical results.

Finally, the last part of this chapter aims at predicting, given a certain period (i.e., pre-bubble, bubble), who will be the main actors in the next ones (i.e., bubble, post-bubble), based on some parameters. This part ends with an analysis aimed at understanding how the users of the various categories have behaved in the months following the ones considered in our investigation, i.e., from the beginning of 2019 until today.

The outline of this chapter is as follows: in Section 9.2, we present related literature and highlight the novelty of our approach. In Section 9.3, we illustrate the dataset that we used to perform our analysis; in particular, we formally define the user categories of interest and discuss the generalizability of the proposed approach. Finally, in Section 9.4, we evaluate the existence of backbones linking the users of a certain category, define an identikit of bubble speculators, and propose a way to predict the main actors of the next cryptocurrency bubble.

9.2 Related literature

Since the introduction of Bitcoin in 2008 [573], thousands of cryptocurrencies have been created [204], and the interest about them has increased significantly. At the same time, the scientific literature about Blockchain and digital currencies has progressively grown [422, 246, 59, 629, 639]. The spread of this new technology has also created a lively discussion in the economic field on the possibility of speculations around these assets [173, 77, 179, 127].

Indeed, at the end of 2017, the price of Bitcoin (as well as the ones of the other cryptocurrencies, like Ethereum or Litecoin) increased by almost 600% (reaching an all-time high value of $19,475.80) before falling by 80% in few weeks, until January 2018 [700, 433, 112, 250]. This is the biggest bubble in the cryptocurrencies history so far. Researchers have strived to analyze every detail of this particular event to understand the corresponding dynamics in order to prevent other speculations in the future. For instance, in [701], the authors investigate market efficiency and volatility persistence in 12 highly priced and capitalized cryptocurrencies, based on daily data from August 7^{th}, 2015 to November 28^{th}, 2018. They observe a random walk pattern in returns of most cryptocurrencies, including Bitcoin and Ethereum, making the price of these assets unpredictable.

In [205], the authors examine the existence and the time intervals of pricing bubbles in Bitcoin and Ethereum. Specifically, they adopt three measures to best represent the key theoretical components of cryptocurrency pricing structures, namely: *(i)* the mining difficulty, which reflects how difficult it is computing the next block of the blockchain; *(ii)* the hashrate, which represents the speed at which a computer is completing an operation in the blockchain code; *(iii)* the relationship between cryptocurrency returns, volatility and liquidity. This study highlights that there are periods characterized by a clear bubble behavior. The period between 2017 and 2018 could be identified as one of them.

Another interesting research field in digital currencies regards the definition of approaches to predict speculative bubbles [191, 282, 617]. For instance, in [292], the authors introduce an automatic peak detection method that classifies price time series into periods of uninterrupted market growth (i.e., drawups) and periods of uninterrupted market decrease (i.e., drawdowns). In [543], the authors investigate a new approach to predict speculative bubbles involving four cryptocurrencies (Bitcoin, Litecoin, Ethereum, and Monero) based on the behavior of new online social media indicators. For this purpose, they leverage a Hidden Markov Model for detecting epidemic outbreaks in the blockchain setting. In [180], the authors propose another possible way to detect speculative bubbles in cryptocurrencies through an approach based on a social microblogging platform for investors and traders. Specifically, they evaluate the sentiment of users on StockTwits[1] and, then, exploit it as a transition variable in a smooth transition autoregression.

A further approach to investigate the cryptocurrencies market is based on the analysis of the corresponding blockchain. It starts from the consideration that a blockchain represents a public ledger in which all committed transactions are stored in a chain of blocks [724, 704]. This chain can be represented and analyzed like a graph with nodes and edges [361, 618, 175, 356]. This reasoning leads the authors of [431] to examine the transaction network of Bitcoin during the first four years of its existence. The results obtained outline the business distribution by countries and their evolution over time. The authors also show that there is a gambling network that features many small transactions. In [452], the authors present a set of analyses on the user graph obtained by performing a heuristic clustering of the Bitcoin blockchain graph. They figure out a set of interesting properties of the network, including the "rich get richer" property and the existence of central nodes acting as privileged bridges between different parts of the network. Finally, in [618], the authors exploit network analysis techniques to investigate the trading dynamics of ERC20 Blockchain. They model ERC20 as a social network, which nodes represent

[1] https://stocktwits.com/

all trading wallets and which edges stand for the buy-sell trades. This social network is inline with the current network theory expectations and presents strong power law properties.

Our work is in line with the latest ones mentioned above, because it uses Social Network Analysis [337, 599] to investigate a blockchain. However, it presents some novelties with respect to them. Indeed, it introduces several categories of users, based on their behavior in the pre-bubble, bubble and post-bubble periods. Moreover, it leverages ego networks [229] and k-cores [237] to identify the characteristics of the various categories of users. Although ego networks and k-cores are classical tools of Social Network Analysis, to the best of our knowledge, they have never been employed to study the behavior of users during a cryptocurrency bubble. Furthermore, it detects the existence of backbones linking users of certain categories in the pre-bubble, bubble and post-bubble periods, which is a knowledge not found in past literature on the cryptocurrency bubbles. Finally, similarly to other papers mentioned above, it also presents a prediction task. However, it differs from the previous ones for the target of the prediction, which, in this case, concerns the discovery, in a certain period (pre-bubble, bubble), of the most relevant features of the users who will be the main actors in the next period.

9.3 Methods

9.3.1 Dataset description

The dataset we used for our analysis is based on the Ethereum blockchain. As stated on the platform official website[2] "Ethereum is a technology that lets you send cryptocurrency to anyone for a small fee. It also powers applications that everyone can use and no one can take down". Ethereum is a programmable blockchain and represents the technological framework behind the cryptocurrency Ether (ETH).

Our dataset was downloaded from Google BigQuery[3]. It contains all the transactions made on Ethereum from January 1^{st}, 2017 to December 31^{st}, 2018. After some data cleaning operations, a row of the dataset, which represents a transaction, contains four columns, namely:

- `from_address`, the blockchain address starting the transaction;
- `to_address`, the blockchain address receiving the transaction;
- `timestamp`, the transaction timestamp;
- `value`, the amount of Weis[4] transferred during the transactions.

[2] `https://ethereum.org/en/what-is-ethereum/`

[3] `https://www.kaggle.com/bigquery/ethereum-blockchain`

[4] Wei is the smallest denomination of Ether; it corresponds to 10^{-18} Ethers.

The dataset is made of 354,107,563 transactions; the total number of user addresses is 43,537,168. We computed some statistics on it, which are reported in Table 9.1.

Parameter	*Value*
Number of transactions	354,107,563
Total number of `from_addresses`	38,881,752
Total number of `to_addresses`	42,457,991
Cardinality of the intersection between `from_addresses` and `to_addresses`	37,802,576
Number of null `from_addresses`	2,104,863
Number of null `to_addresses`	0

Table 9.1: Some preliminary statistics performed on our dataset

9.3.2 Defining the user categories of interest

In this section, we present some preliminary analyses "depicting" the pre-bubble, bubble and post-bubble periods, as well as the general behavior of users during the two years covered by our dataset and, especially, during the three periods of our interest. At the end of these analyses, we will be able to define the user categories of interest.

A first analysis concerns the distributions of the number of transactions against `from_addresses` and `to_addresses`. They are reported in Figure 9.1. This figure shows that the two distributions follow a power law. We computed some parameters for them; they are reported in Table 9.2.

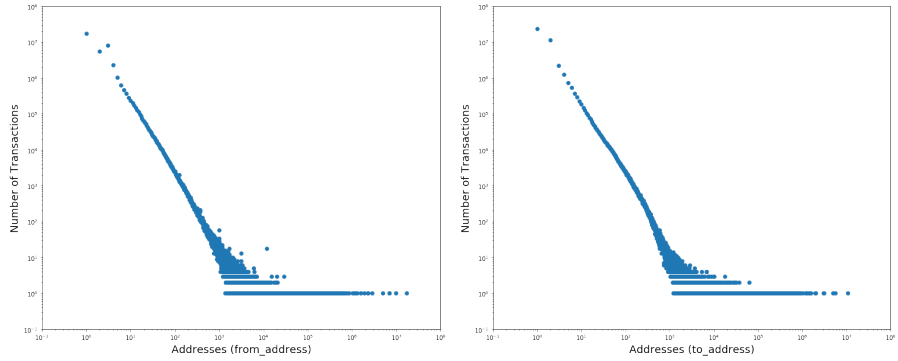


Fig. 9.1: Log-log plots of the distributions of transactions against `from_addresses` (at left) and `to_addresses` (at right)

From the analysis of both Figure 9.1 and Table 9.2 we can observe that the two power law distributions are similar.

Parameter	`from_addresses`	`to_addresses`
Maximum number of transactions	17,509,218	23,404,261
Average number of transactions	5,640.76	5,913.37
α (power law parameter)	1.477	1.565
δ (power law parameter)	0.013	0.074

Table 9.2: Values of the parameters of transaction distributions against addresses

The second analysis that we take into consideration concerns the variation of the number of transactions over time. The purpose of this analysis is the identification of the pre-bubble, bubble and post-bubble periods. This trend is shown in Figure 9.2. From the analysis of this figure we can see that from January 2017 to October 2017 there is a substantially linear growth of the number of transactions. From November 2017 to March 2018 there is first an impressive increase and then an impressive decrease of the same variable. Finally, from April 2018 to December 2018 the number of transactions has an irregular trend, but on average its values are lightly higher than the ones observed before November 2017. Based on these observations, in the following, we assume as pre-bubble period the time interval January - October 2017, as bubble period the time interval November 2017 - March 2018, and as post-bubble period the time interval April - December 2018.

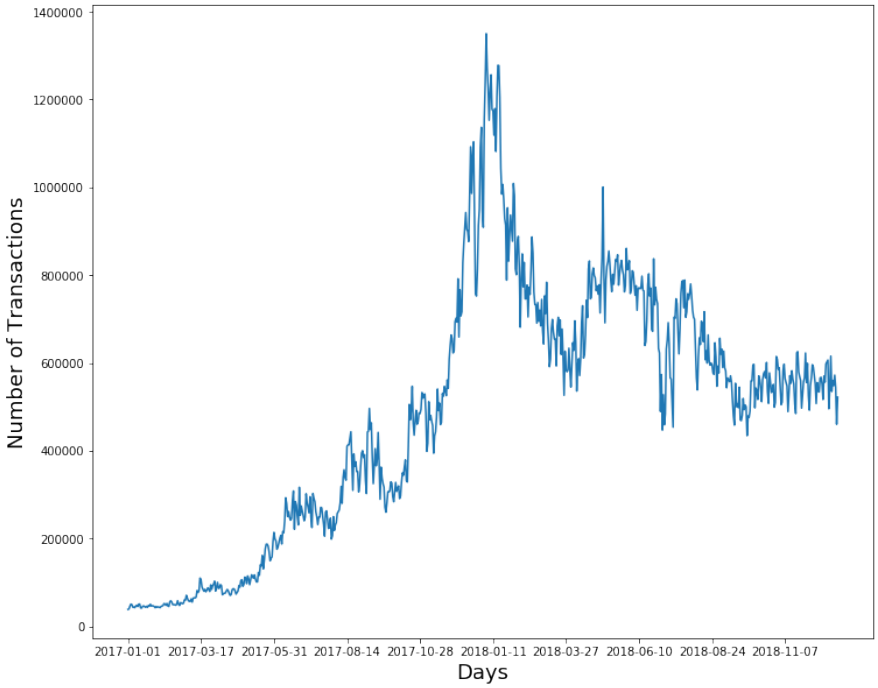


Fig. 9.2: Number of transactions over time

The next analysis focuses on power addresses, i.e., those addresses that have made the most transactions. The analysis of these addresses is extremely relevant for two reasons. First, since the distributions of transactions against addresses follow a power law, the analysis of power addresses covers most of the phenomenon we want to examine. Second, since the number of power addresses is very small, compared to the total number of addresses, it is possible to make very precise and detailed analyses on them, which would be impossible to conduct on all addresses or on a very high fraction of them.

In particular, for each period (pre-bubble, bubble and post-bubble) and for each type of addresses (`from` and `to`), we decided to take the top 1000 addresses as the power ones. For each set thus selected, Table 9.3 shows: *(i)* what percentage of the total number of addresses operating in the reference period the top 1000 addresses correspond to; *(ii)* what percentage of the total number of transactions performed in the reference period the transactions carried out by the top 1000 addresses correspond to. From the analysis of this table, we can deduce that the previous conjectures on the opportunity to carry out the power address analyses were correct.

Set	*Percentage of addresses*	*Percentage of transactions*
Pre-bubble, top 1000 `from_addresses`	0.01549%	89.81%
Bubble, top 1000 `from_addresses`	0.00599%	78.48%
Post-bubble, top 1000 `from_addresses`	0.00534%	77.87%
Pre-bubble, top 1000 `to_addresses`	0.01325%	86.02%
Bubble, top 1000 `to_addresses`	0.00495%	82.29%
Post-bubble, top 1000 `to_addresses`	0.00548%	86.34%

Table 9.3: Percentage of the addresses and transactions covered by each set of power addresses

A first analysis of power addresses concerned the possible overlap between `from_addresses` and `to_addresses`. For this purpose, for each period, we computed the intersection between the top 1000 `from_addresses` and the top 1000 `to_addresses`. The result obtained is reported in Table 9.4. This table shows that only a small fraction of power addresses is simultaneously present in the top 1000 `from_addresses` and in the top 1000 `to_addresses`. Another information emerging from Table 9.4 is that this fraction significantly decreases in the transition from pre-bubble to bubble and from bubble to post-bubble periods.

A further analysis on power addresses led us to compute the possible intersections of the top 1000 addresses during the pre-bubble, bubble and post-bubble periods. The results obtained are reported in Table 9.5. Here, T^F_{Pre} (resp., T^F_B, T^F_{Post}) is the set of the top 1000 `from_addresses` during the pre-bubble (resp., bubble,

Pre-bubble	*Bubble*	*Post-Bubble*
173	115	81

Table 9.4: Number of power addresses simultaneously belonging to the set of the top 1000 `from_addresses` and to the set of the top 1000 `to_addresses` in the three periods of interest

post-bubble) period. Analogously, T^T_{Pre}, T^T_B and T^T_{Post} are the corresponding sets for `to_addresses`. From the analysis of this table we can see that:

Set	*Cardinality*
$\lvert T^F_{Pre} \cap T^F_B \rvert$	267
$\lvert T^F_B \cap T^F_{Post} \rvert$	268
$\lvert T^F_{Pre} \cap T^F_{Post} \rvert$	107
$\lvert T^T Pre \cap T^T_B \rvert$	288
$\lvert T^T_B \cap T^T_{Post} \rvert$	309
$\lvert T^T_{Pre} \cap T^T_{Post} \rvert$	114
$\lvert T^F_{Pre} \cap T^F_B \cap T^F_{Post} \rvert$	102
$\lvert T^T_{Pre} \cap T^T_B \cap T^T_{Post} \rvert$	112

Table 9.5: Cardinalities of the possible intersections of the top 1000 addresses during the pre-bubble, bubble and post-bubble periods

- The trends of `from_addresses` and `to_addresses` are very similar.
- The bubble has changed the power address scenario considerably. In fact, while the cardinality of the sets $\lvert T^F_{Pre} \cap T^F_B \rvert$, $\lvert T^F_B \cap T^F_{Post} \rvert$, $\lvert T^T Pre \cap T^T_B \rvert$ and $\lvert T^T_B \cap T^T_{Post} \rvert$ is quite large, the one of the sets $\lvert T^F_{Pre} \cap T^F_{Post} \rvert$ and $\lvert T^T_{Pre} \cap T^T_{Post} \rvert$ is much smaller. This tells us that, during the bubble period, most of the power addresses present in the pre-bubble period disappeared and new power addresses appeared; these last continued to exist during the post-bubble period. Finally, we observe that there are some power addresses, which we call "Survivors", that are present in the pre-bubble, bubble and post-bubble periods.

Based on the intersections introduced in Table 9.5, we can define three categories of addresses whose analysis appears extremely interesting for the extraction of knowledge on the bubble of Ethereum (and, presumably, of other cryptocurrencies). These categories are:

- *the Survivors*, which are the power addresses present in the pre-bubble, bubble and post-bubble periods;
- *the Missings*, which are the power addresses present in the pre-bubble period, but absent in the bubble and post-bubble ones;

- *the Entrants*, which are the power addresses absent in the pre-bubble period, but present in the bubble and post-bubble ones.

In the following, we aim at extracting knowledge patterns about these categories of addresses (and, ultimately, of users).

The next analysis aims at identifying how many power addresses are present in each category. We conducted this analysis for `from_addresses`, `to_addresses` and the intersection of these two sets. The results obtained are shown in Table 9.6.

Addresses	*Survivors*	*Entrants*	*Missings*
`from_addresses`	102	166	728
`to_addresses`	112	197	710
Intersection of `from_addresses` and `to_addresses`	21	17	114

Table 9.6: Number of power addresses belonging to the Survivors, Entrants and Missings categories

To fully understand the knowledge that can be extracted from this table, we must recall that: *(i)* the maximum number of power addresses for each category is equal to 1000; *(ii)* the Survivors, the Entrants and the Missings are obtained carrying out intersection operations. According to this reasoning, we can observe that the Survivors are very few; this result was expected because this category of addresses is obtained performing the intersection of three sets. The Entrants are also few while the Missings are many. This confirms that the bubble completely revolutionized the power address scenario in Ethereum, making the previous "main actors" (i.e., power addresses) disappear while introducing new ones.

Observe that, for all categories, the intersections between `from_addresses` and `to_addresses` are very small. This is totally in line with Table 9.4, where we have seen that only a few addresses are `from_addresses` and `to_addresses` simultaneously.

9.3.3 Detecting the main features of the user categories of interest

Given a period (pre-bubble, bubble and post-bubble) and the set of the corresponding power addresses, we build a support social network. More specifically, let

$$\mathcal{N}_{Pre} = \langle NS_{Pre}, AS_{Pre} \rangle \qquad \mathcal{N}_{B} = \langle NS_{B}, AS_{B} \rangle \qquad \mathcal{N}_{Post} = \langle NS_{Post}, AS_{Post} \rangle$$

be the social networks associated with the pre-bubble, bubble and post-bubble periods.

NS_{Pre} (resp., NS_B, NS_{Post}) represents the set of the nodes of $\mathcal{N}_{Pre}$ (resp., $\mathcal{N}_B$, $\mathcal{N}_{Post}$). In this set, there is a node n_i for each power address. A label is associated

with n_i; it allows us to specify if the corresponding address belongs to one of the categories of interest (Survivors, Entrants, Missings) or to none of them. Since there is a biunivocal correspondence between power addresses and nodes, in the following we will use these two terms interchangeably.

AS_{Pre} (resp., AS_B, AS_{Post}) denotes the set of the arcs of $\mathcal{N}_{Pre}$ (resp., $\mathcal{N}_B$, $\mathcal{N}_{Post}$). There is an arc $(n_i, n_j, TS_{ij}) \in AS_{Pre}$ (resp., AS_B, AS_{Post}) if there was at least one transaction from n_i to n_j. TS_{ij} represents the set of transactions from n_i to n_j made during the pre-bubble (resp., bubble, post-bubble) period. It consists of a set of pairs (t_{ij_k}, τ_{ij_k}), where t_{ij_k} represents the k^{th} transaction and τ_{ij_k} indicates the corresponding timestamp.

Having defined the support social networks, we can start our analyses on the address categories of interest. Below, we use the following notations:

- $\mathcal{S}^F$ (resp., $\mathcal{S}^T$), to indicate the Survivors `from_addresses` (resp., `to_addresses`).
- $\mathcal{E}^F$ (resp., $\mathcal{E}^T$), to denote the Entrants `from_addresses` (resp., `to_addresses`).
- $\mathcal{M}^F$ (resp., $\mathcal{M}^T$), to represent the Missings `from_addresses` (resp., `to_addresses`).

In order to conduct our analyses on the address categories, we have considered the adoption of ego networks extremely useful. We recall that the ego network of a node n_i (called, precisely, "ego") consists of n_i, the nodes (called "alters") to which n_i is directly connected, the arcs connecting the ego to the alters and the arcs connecting the alters to each other. An ego network provides a clear indication of the relationships the corresponding ego is involved in, the nodes it interacts with, and the relationships existing between these last ones. In our analysis, which aims at detecting the features of each address category, ego network can play an important role because, due to the principle of homophily characterizing social networks [468], the features of a node are strongly influenced by the nodes belonging to its neighborhood.

As a first task, we computed the average number of nodes, the average number of arcs and the average density of the ego networks of the nodes belonging to each address category of interest. First, we examined the pre-bubble period. The results obtained are reported in Table 9.7.

From the analysis of this table we can see that the ego networks of the Survivors nodes have an average number of nodes and arcs significantly higher than the ego networks of the nodes belonging to the other two categories. If such a result was expected for the Entrants (because the corresponding nodes were not power addresses during the pre-bubble period), it is instead surprising for the Missings. In fact, the latter, like the Survivors, were power addresses during the pre-bubble period. This allows us to conclude that having a very large ego-network during the pre-bubble

Parameter	$\mathcal{S}^F$	$\mathcal{S}^T$	$\mathcal{M}^F$	$\mathcal{M}^T$	$\mathcal{E}^F$	$\mathcal{E}^T$
Average number of nodes	36,177.84	27,335.21	1,710.52	2,864.44	537.69	886.02
Average number of arcs	115,290.27	68,051.82	4,561.86	7,342.89	795.53	1,718.39
Average density	0.1120	0.0639	0.3852	0.2423	0.2125	0.1568

Table 9.7: Average number of nodes, average number of arcs and average density of the ego networks of the nodes belonging to the address categories of interest - Pre-bubble period

period increases the possibility of remaining power addresses during the bubble and post-bubble periods. As far as density is concerned, there are no particular observations to make taking into account that the low density of Survivor's ego networks can be explained simply by the large number of nodes characterizing them.

After this, we examined the bubble period. The results obtained are reported in Table 9.8.

Parameter	$\mathcal{S}^F$	$\mathcal{S}^T$	$\mathcal{M}^F$	$\mathcal{M}^T$	$\mathcal{E}^F$	$\mathcal{E}^T$
Average number of nodes	82,832.51	59,339.83	366.58	798.29	17,180.69	18,945.69
Average number of arcs	325,179.44	172,713.37	587.84	2563.00	59,733.11	67,956.61
Average density	0.074	0.019	0.401	0.282	0.211	0.031

Table 9.8: Average number of nodes, average number of arcs and average density of the ego networks of the nodes belonging to the address categories of interest - Bubble period

From the analysis of this table we can observe that both the Survivors and the Entrants have much larger ego networks than the Missings. Actually, this result was expected since, in the bubble period, the nodes belonging to the Survivors and the Entrants are power addresses. Instead, it is unexpected that the Survivors have much larger ego networks than the Entrants. In fact, the addresses of both categories are power addresses during the bubble period. However, it seems that the Survivors tend to include the strongest power addresses. Note also that the one of the Survivors' ego networks during the bubble period is about twice the size of the Survivors' ego networks during the pre-bubble period. Also, the Survivors' ego networks have by far the largest size during the bubble period. This allows us to conclude that it is exactly the activity of the Survivors that could have caused the bubble; this activity has led to the exit of the Missings from the power addresses and to the arrival of the Entrants among them. However, these last ones enter into the power addresses "on tiptoe"; in fact, they are not the ones who dictate the line and cause the bubble; this task is carried out by the Survivors.

Finally, we considered the post-bubble period. The results obtained are reported in Table 9.9.

Parameter	$\mathcal{S}^F$	$\mathcal{S}^T$	$\mathcal{M}^F$	$\mathcal{M}^T$	$\mathcal{E}^F$	$\mathcal{E}^T$
Average number of nodes	47,237.20	46,661.02	162.10	572.93	19,686.75	22,373.64
Average number of arcs	174,537.78	148,359.25	425.70	1,360.52	93,099.84	70,518.77
Average density	0.1045	0.039	0.411	0.233	0.178	0.0157

Table 9.9: Average number of nodes, average number of arcs and average density of the ego networks of the nodes belonging to the address categories of interest - Post-bubble period

The analysis of this table confirms the trends we observed in Table 9.8 for the bubble period. This is not surprising because also during the post-bubble period both the Survivors and the Entrants are power addresses. Note that, during this period, the size of the Survivors' ego networks is much smaller than the one of the Survivors' ego networks during the bubble period, although it is slightly larger than the size of the Survivors' ego networks during the pre-bubble period. This trend perfectly reflects the one of the number of transactions reported in Figure 9.2. This is a further confirmation that the trend shown by Ethereum in the years 2017 and 2018, which led to a bubble, was mainly caused by the Survivors. We note that the size of the Entrants' ego network during the post-bubble period shows a slight growth compared to the bubble period. This is an indication that, during the post-bubble period, the Entrants consolidate their presence among the power addresses, even though they are not dictating the line yet: this is still a responsibility of the Survivors.

The analysis of Tables 9.7 - 9.9, along with the previous reasoning, indicates that having very large ego networks seems to be an intrinsic feature of the Survivors, regardless of the pre-bubble, bubble or post-bubble period.

9.3.4 Generalizability of the proposed analyses

In Section 9.3.1, we saw that our dataset was derived from Ethereum. Furthermore, we saw that each record in it corresponds to a transaction and stores only four fields related to it, namely: *(i)* the blockchain address starting it; *(ii)* the blockchain address receiving it; *(iii)* its timestamp; *(iv)* the amount of money transferred during it. These four fields are very general and available for any cryptocurrency blockchain. Therefore, although our analysis was performed on Ethereum, our approach can be extended to any cryptocurrency blockchain. To facilitate this extension, in the following we abstract the analyses described here into a well-structured algorithm of

which they represent single steps. The pseudo-code of this algorithm is shown in Algorithms 3 and 4.

Input

- $\mathcal{B}$: the cryptocurrency blockchain of interest
- I: the time interval to investigate

Output

- PA^F_{Pre}, PA^F_B, PA^F_{Post}, PA^T_{Pre}, PA^T_B, PA^T_{Post}: power addresses of the dataset
- SPA_{Pre}, SPA_B, SPA_{Post}, PA^F_{Pre-B}, PA^F_{B-Post}, PA^T_{Pre-B}, PA^T_{B-Post}: power addresses of the dataset
- $\mathcal{S}^F$, $\mathcal{S}^T$: the Survivors; $\mathcal{M}^F$, $\mathcal{M}^T$: the Missings; $\mathcal{E}^F$, $\mathcal{E}^T$: the Entrants
- $EgoKPSet$: a set of knowledge patterns derived from the ego network analyses
- $BackboneKPSet$: a set of knowledge patterns on the possible presence of backbones
- $BSurvivorsSet$: a set of potential Survivors
- $PBSurvivorsSet$: a set of potential Survivors
- $PBEntrantsSet$: a set of potential Entrants

Require:

- D: a dataset of transactions;
- I_{Pre}, I_B, I_{Post}: time intervals;
- $\mathcal{N}_{Pre}$, $\mathcal{N}_B$, $\mathcal{N}_{Post}$: social networks;
- $ENSet^{\mathcal{S},F}_{Pre}$, $ENSet^{\mathcal{S},T}_{Pre}$, $ENSet^{\mathcal{M},F}_{Pre}$, $ENSet^{\mathcal{M},T}_{Pre}$, $ENSet^{\mathcal{E},F}_{Pre}$, $ENSet^{\mathcal{E},T}_{Pre}$: a set of ego networks;
- $ENSet^{\mathcal{S},F}_{B}$, $ENSet^{\mathcal{S},T}_{B}$, $ENSet^{\mathcal{M},F}_{B}$, $ENSet^{\mathcal{M},T}_{B}$, $ENSet^{\mathcal{E},F}_{B}$, $ENSet^{\mathcal{E},T}_{B}$: a set of ego networks;
- $ENSet^{\mathcal{S},F}_{Post}$, $ENSet^{\mathcal{S},T}_{Post}$, $ENSet^{\mathcal{M},F}_{Post}$, $ENSet^{\mathcal{M},T}_{Post}$, $ENSet^{\mathcal{E},F}_{Post}$, $ENSet^{\mathcal{E},T}_{Post}$: a set of ego networks;
- T^F_{Pre}, T^T_{Pre}, T^F_B, T^T_B, T^F_{Post}, T^T_{Post}: top power addresses of the dataset;

D = Extract_Dataset($\mathcal{B}, I$)
$\langle I_{Pre}, I_B, I_{Post}\rangle$ = Determine_Intervals(D)
$\langle PA^F_{Pre}, PA^F_B, PA^F_{Post}\rangle$ = Detect_From_Power_Addresses($I_{Pre}, I_B, I_{Post}, D$)
$\langle PA^T_{Pre}, PA^T_B, PA^T_{Post}\rangle$ = Detect_To_Power_Addresses($I_{Pre}, I_B, I_{Post}, D$)
$\langle SPA_{Pre}, SPA_B, SPA_{Post}\rangle$ = Detect_Super_Power_Addresses($PA^F_{Pre}, PA^F_B, PA^F_{Post}, PA^T_{Pre}, PA^T_B, PA^T_{Post}$)
$\langle SPA^F_{Pre-B}, SPA^F_{B-Post}\rangle$ = Detect_Multi_Interval_From_Power_Addresses($PA^F_{Pre}, PA^F_B, PA^F_{Post}$)
$\langle SPA^T_{Pre-B}, SPA^T_{B-Post}\rangle$ = Detect_Multi_Interval_To_Power_Addresses($PA^T_{Pre}, PA^T_B, PA^T_{Post}$)
$\langle \mathcal{S}^F, \mathcal{S}^T\rangle$ = Detect_Survivors($PA^F_{Pre}, PA^F_B, PA^F_{Post}, PA^T_{Pre}, PA^T_B, PA^T_{Post}$)
$\langle \mathcal{M}^F, \mathcal{M}^T\rangle$ = Detect_Missings($PA^F_{Pre}, PA^F_B, PA^F_{Post}, PA^T_{Pre}, PA^T_B, PA^T_{Post}$)
$\langle \mathcal{E}^F, \mathcal{E}^T\rangle$ = Detect_Entrants($PA^F_{Pre}, PA^F_B, PA^F_{Post}, PA^T_{Pre}, PA^T_B, PA^T_{Post}$)
$\langle \mathcal{N}_{Pre}, \mathcal{N}_B, \mathcal{N}_{Post}\rangle$ = Construct_Social_Networks($I_{Pre}, I_B, I_{Post}, D$)
$\langle ENSet^{\mathcal{S},F}_{Pre}, ENSet^{\mathcal{S},T}_{Pre}, \rangle$ = Construct_Survivors_Ego_Networks_Pre($I_{Pre}, \mathcal{N}_{Pre}, \mathcal{S}^F, \mathcal{S}^T$)
$\langle ENSet^{\mathcal{S},F}_{B}, ENSet^{\mathcal{S},T}_{B}, \rangle$ = Construct_Survivors_Ego_Networks_Bubble($I_B, \mathcal{N}_B, \mathcal{S}^F, \mathcal{S}^T$)
$\langle ENSet^{\mathcal{S},F}_{Post}, ENSet^{\mathcal{S},T}_{Post}, \rangle$ = Construct_Survivors_Ego_Networks_Post($I_{Post}, \mathcal{N}_{Post}, \mathcal{S}^F, \mathcal{S}^T$)

Algorithm 3: Investigating user behavior during a cryptocurrency speculative bubble (first part)

Our algorithm receives the cryptocurrency blockchain $\mathcal{B}$ of interest and the time interval I during which there was a speculative bubble involving $\mathcal{B}$.

It first calls the function *Extract_Dataset* that returns the dataset D of the transactions of $\mathcal{B}$ during I. Next, it calls the function *Determine_Intervals* to partition I into three sub-intervals I_{Pre}, I_B and I_{Post}, relating to the pre-bubble, bubble and post-bubble periods, respectively. After that, it calls the functions

Require:

$\langle ENSet^{\mathcal{M},F}_{Pre}, ENSet^{\mathcal{M},T}_{Pre}, \rangle$ = Construct_Missings_Ego_Networks_Pre($I_{Pre}, \mathcal{N}_{Pre}, \mathcal{M}^F, \mathcal{M}^T$)
$\langle ENSet^{\mathcal{M},F}_{B}, ENSet^{\mathcal{M},T}_{B}, \rangle$ = Construct_Missings_Ego_Networks_Bubble($I_B, \mathcal{N}_B, \mathcal{M}^F, \mathcal{M}^T$)
$\langle ENSet^{\mathcal{M},F}_{Post}, ENSet^{\mathcal{M},T}_{Post}, \rangle$ = Construct_Missings_Ego_Networks_Post($I_{Post}, \mathcal{N}_{Post}, \mathcal{M}^F, \mathcal{M}^T$)
$\langle ENSet^{\mathcal{E},F}_{Pre}, ENSet^{\mathcal{E},T}_{Pre}, \rangle$ = Construct_Entrants_Ego_Networks_Pre($I_{Pre}, \mathcal{N}_{Pre}, \mathcal{E}^F, \mathcal{E}^T$)
$\langle ENSet^{\mathcal{E},F}_{B}, ENSet^{\mathcal{E},T}_{B}, \rangle$ = Construct_Entrants_Ego_Networks_Bubble($I_B, \mathcal{N}_B, \mathcal{E}^F, \mathcal{E}^T$)
$\langle ENSet^{\mathcal{E},F}_{Post}, ENSet^{\mathcal{E},T}_{Post}, \rangle$ = Construct_Entrants_Ego_Networks_Post($I_{Post}, \mathcal{N}_{Post}, \mathcal{E}^F, \mathcal{E}^T$)
$EgoKPSet$ = Analyze_Ego_Pre($ENSet^{\mathcal{S},F}_{Pre}, ENSet^{\mathcal{S},T}_{Pre}, ENSet^{\mathcal{M},F}_{Pre}, ENSet^{\mathcal{M},T}_{Pre}, ENSet^{\mathcal{E},F}_{Pre}, ENSet^{\mathcal{E},T}_{Pre}$)
$EgoKPSet = EgoKPset \cup$ Analyze_Ego_Bubble($ENSet^{\mathcal{S},F}_{B}, ENSet^{\mathcal{S},T}_{B}, ENSet^{\mathcal{M},F}_{B}, ENSet^{\mathcal{M},T}_{B}, ENSet^{\mathcal{E},F}_{B}, ENSet^{\mathcal{E},T}_{B}$)
$EgoKPSet = EgoKPset \cup$ Analyze_Ego_Post($ENSet^{\mathcal{S},F}_{Post}, ENSet^{\mathcal{S},T}_{Post}, ENSet^{\mathcal{M},F}_{Post}, ENSet^{\mathcal{M},T}_{Post}, ENSet^{\mathcal{E},F}_{Post}, ENSet^{\mathcal{E},T}_{Post}$)
$BackboneKPSet$ = Detect_Backbones_Survivor_Pre($ENSet^{\mathcal{S},F}_{Pre}, ENSet^{\mathcal{S},T}_{Pre}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Survivor_Bubble($ENSet^{\mathcal{S},F}_{B}, ENSet^{\mathcal{S},T}_{B}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Survivor_Post($ENSet^{\mathcal{S},F}_{Post}, ENSet^{\mathcal{S},T}_{Post}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Missing_Pre($ENSet^{\mathcal{M},F}_{Pre}, ENSet^{\mathcal{M},T}_{Pre}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Missing_Bubble($ENSet^{\mathcal{M},F}_{B}, ENSet^{\mathcal{M},T}_{B}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Missing_Post($ENSet^{\mathcal{M},F}_{Post}, ENSet^{\mathcal{M},T}_{Post}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Entrants_Pre($ENSet^{\mathcal{E},F}_{Pre}, ENSet^{\mathcal{E},T}_{Pre}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Entrants_Bubble($ENSet^{\mathcal{E},F}_{B}, ENSet^{\mathcal{E},T}_{B}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$BackboneKPSet = BackboneKPSet \cup$ Detect_Backbones_Entrants_Post($ENSet^{\mathcal{E},F}_{Post}, ENSet^{\mathcal{E},T}_{Post}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T$)
$\langle T^F_{Pre}, T^T_{Pre}, T^F_B, T^T_B, T^F_{Post}, T^T_{Post}, \rangle$ = Detect_Top_Power_Addresses($I_{Pre}, I_B, I_{Post}, D$)
$BSurvivorsSet$ = Predict_Bubble_Survivors($T^F_{Pre}, T^T_{Pre}, T^F_B, T^T_B, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T, I_{Pre}, I_B, D$)
$PBSurvivorsSet$ = Predict_Post_Survivors($T^F_B, T^T_B, T^F_{Post}, T^T_{Post}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T, I_B, I_{Post}, D$)
$PBEntrantsSet$ = Predict_Post_Entrants($T^F_B, T^T_B, T^F_{Post}, T^T_{Post}, \mathcal{S}^F, \mathcal{S}^T, \mathcal{M}^F, \mathcal{M}^T, \mathcal{E}^F, \mathcal{E}^T, I_B, I_{Post}, D$)

return all outputs

Algorithm 4: Investigating user behavior during a cryptocurrency speculative bubble (second part)

Detect_From_Power_Addresses, *Detect_To_Power_Addresses* and *Detect_Super_Power_Addresses* to determine the power addresses with the largest number of incoming arcs, outgoing arcs and both. Finally, it calls the functions *Detect_Multi_Interval_From_Power_Addresses* and *Detect_Multi_Interval_To_Power_Addresses* to determine the addresses that remain From_Power_Addresses and To_Power_Addresses when passing from the pre-bubble period to the bubble one and from the bubble period to the post-bubble one.

At this point, our algorithm has all the data it needs to activate *Detect_Survivors*, *Detect_Missings* and *Detect_Entrants*, which aim at determining the Survivors $\mathcal{S}^F$ and $\mathcal{S}^T$, the Missings $\mathcal{M}^F$ and $\mathcal{M}^T$ and the Entrants $\mathcal{E}^F$ and $\mathcal{E}^T$. Next, it calls the function *Construct_Social_Network* that returns the social networks $\mathcal{N}_{Pre}$, $\mathcal{N}_B$ and $\mathcal{N}_{Post}$ relative to the pre-bubble, bubble and post-bubble period. After that, it calls the functions *Construct_Survivors_Ego_Network_Pre*, *Construct_Survivors_Ego_Network_Bubble* and *Construct_Survivors_Ego_Network_Post* to construct the ego networks of the Survivors of the social networks $\mathcal{N}_{Pre}$, $\mathcal{N}_B$ and $\mathcal{N}_{Post}$. Similarly, it proceeds to call the suitable functions for constructing the ego networks of the Missings and the Entrants for the same social networks mentioned above.

The ego networks thus constructed represent the basis for the next analyses aimed at extracting a set *EgoKPSet* of knowledge patterns on the characteristics of the Survivors, the Missings and the Entrants in the pre-bubble, bubble and post-bubble periods. Our algorithm performs this extraction by calling the functions *Analyze_Ego_Pre*, *Analyze_Ego_Bubble* and *Analyze_Ego_Post*. The next analysis performed by it concerns the possible existence of backbones linking Survivors, Missings or Entrants in the pre-bubble, bubble and post-bubble periods. To this end, it calls some functions having the objective of extracting the set BackboneKPSet of knowledge patterns concerning the possible existence of backbones among the various kinds of address of interest.

Once the backbone analysis is finished, our algorithm proceeds with the last analysis which, unlike the previous ones, is predictive. In fact, it aims at predicting, during a certain period, the nodes that will become protagonists in the next period. To this end, it calls the functions *Predict_Bubble_Survivors*, *Predict_Post_Survivors* and *Predict_Post_Entrants*. The first examines nodes during the pre-bubble period and predicts which of them constitute the set $BSurvivorsSet$ of potential Survivors during the bubble period. The second and the third examine the nodes during the bubble period and predict which of them will form the set $PBSurvivorsSet$ and $PBEntrantsSet$ of potential Survivors and Entrants during the post-bubble period.

The algorithm terminates returning in output all the information extracted through the calls of the functions mentioned above.

A more abstract and simplified graphical representation of it is shown in Figure 9.3.

9.4 Results

In this section, we provide some considerations regarding the proposed analyses, the results obtained and their applicability for future cryptocurrency speculative bubbles. In particular, we aim at answering the following questions:

- Are there backbones linking users of a certain category? Can we apply the concept of ego networks and k-cores to detect them?
- The graphical evaluation of the existence of a backbone should have been based on the concept of clique. However, due to computational complexity issues, our experiments were performed on k-cores, which represent a relaxation of the clique concept. Could the results obtained have been affected by the adoption of k-cores instead of cliques?

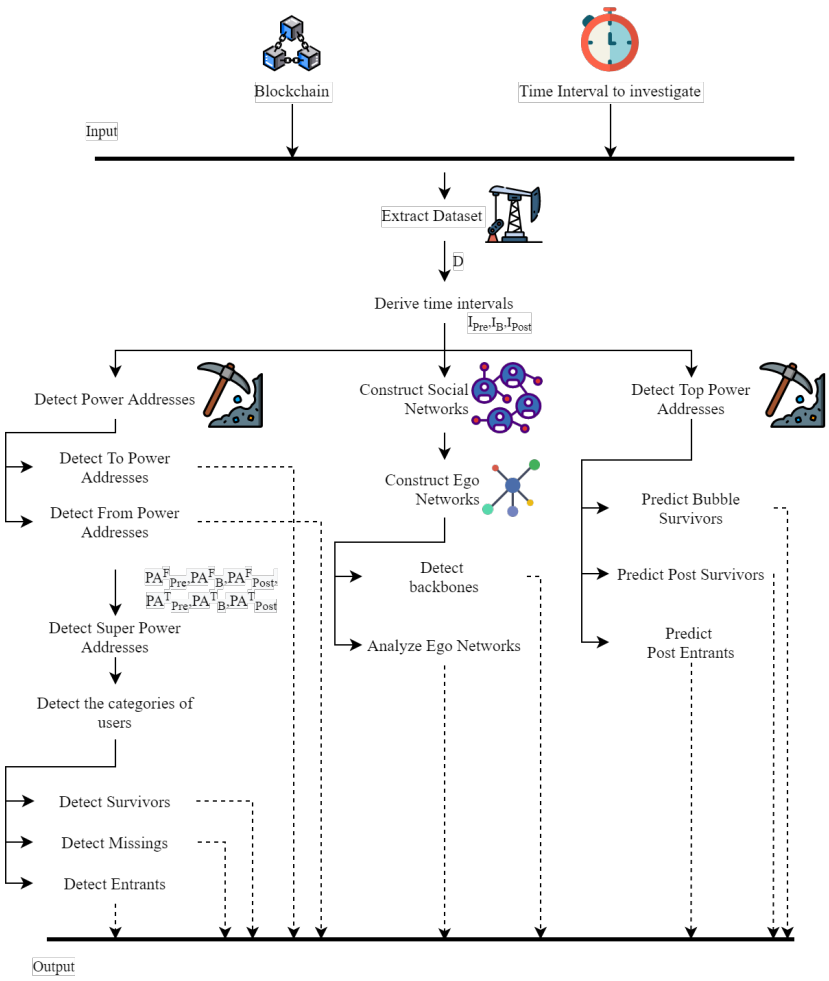


Fig. 9.3: A graphical abstract representation of our algorithm

- Do the described outcomes allow us to infer that there was a group of speculators who managed the Ethereum bubble in the years 2017-2018? If so, what can be said about their profile?
- Can we predict the characteristics of the main future users for the next periods?

In the following, we devote a subsection to each of the four issues mentioned above.

9.4.1 Evaluating the existence of backbones linking users of a certain category

The ego networks introduced previously represent a considerable tool to also estimate the possible existence of backbones linking addresses of the same category. In fact, a way to do this consists in verifying, given an address category, the fraction of the corresponding ego networks having, among the alters, at least k addresses belonging to it. Clearly, the higher the value of k and the fraction of the ego networks

satisfying this property, the stronger the hypothesis that a backbone exists among the addresses of the category into examination.

To better clarify this idea, let us consider Table 9.10 that refers to the Survivors' ego networks during the pre-bubble period. In the left part of this table, we examine the set $\mathcal{S}^F$ of the Survivors `from_addresses`. The fifth row of this table tells us that 19.6% of the ego networks of the nodes of $\mathcal{S}^F$ contains at least 5 nodes of $\mathcal{S}^F$ among the alters. This percentage decreases to 0.9% if we consider the presence of at least 5 nodes of $\mathcal{E}^F$ and increases to 33.3% if we take into account the presence of at least 5 nodes of $\mathcal{M}^F$.

	Ego networks of $\mathcal{S}^F$			*Ego networks of* $\mathcal{S}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k = 1$	0.755	0.088	0.676	0.580	0.223	0.696
$k = 2$	0.512	0.058	0.529	0.339	0.071	0.509
$k = 3$	0.392	0.049	0.402	0.169	0.0	0.348
$k = 4$	0.294	0.019	0.353	0.098	0.0	0.304
$k = 5$	0.196	0.009	0.333	0.080	0.0	0.277
$k = 6$	0.147	0.0	0.284	0.062	0.0	0.268
$k = 7$	0.118	0.0	0.265	0.053	0.0	0.241
$k = 8$	0.078	0.0	0.235	0.036	0.0	0.196
$k = 9$	0.078	0.0	0.216	0.027	0.0	0.196

Table 9.10: Analysis of the presence of backbones linking the Survivors during the pre-bubble period

Once we have clarified the kind of information we want to look for, let us consider Table 9.10, which concerns the Survivors' ego networks during the pre-bubble period. From the analysis of this table we can see that many of the ego-networks of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$) have, among their alters, several nodes belonging to $\mathcal{S}^F$ (resp., $\mathcal{S}^T$), along with several nodes belonging to $\mathcal{M}^F$ (resp., $\mathcal{M}^T$). Instead, the number of ego networks of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$) having one or more nodes of $\mathcal{E}^F$ (resp., $\mathcal{E}^T$) among the alters is very small. This allows us to assume that there is a backbone linking the nodes of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$). The presence of many nodes of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$) among the alters of the ego networks of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$) is not surprising because, during the pre-bubble period, the nodes of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$) were power addresses. Finally, we observe that the presence of Survivors and Missings nodes among the alters of the ego networks of Survivors nodes is more marked for `from_addresses` than for `to_addresses`, as we can see comparing the first three and the last three columns of Table 9.10.

Consider, now, Table 9.11 that refers to the Missings' ego networks during the pre-bubble period. The structure and the semantics of this table are analogous to the ones of Table 9.10. From the analysis of this table, we can observe that many ego

networks of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$) have one or two nodes of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$) or of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$) among their alters. However, compared to the case of the Survivors, reported in Table 9.10, this phenomenon is much smaller both as fraction of ego-networks and as value of k. Therefore, we can conclude that there is a backbone also among the nodes of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$), although this is less strong than the one observed for the nodes of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$). The presence of many nodes of $\mathcal{S}^F$ (resp., $\mathcal{S}^T$) among the alters of the ego networks of $\mathcal{M}^F$ (resp., $\mathcal{M}^T$) is justified by the fact that both these categories of nodes were power addresses during the pre-bubble period. The difference between `from_addresses` and `to_addresses` in the Missings' ego networks is much smaller than the one observed in the Survivors' ego networks.

	Ego networks of $\mathcal{M}^F$			*Ego networks of* $\mathcal{M}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k = 1$	0.466	0.010	0.497	0.390	0.024	0.406
$k = 2$	0.277	0.0	0.214	0.162	0.0	0.225
$k = 3$	0.165	0.0	0.115	0.093	0.0	0.138
$k = 4$	0.098	0.0	0.070	0.056	0.0	0.089
$k = 5$	0.059	0.0	0.049	0.039	0.0	0.068
$k = 6$	0.040	0.0	0.033	0.031	0.0	0.052
$k = 7$	0.018	0.0	0.029	0.025	0.0	0.037
$k = 8$	0.004	0.0	0.027	0.021	0.0	0.032
$k = 9$	0.004	0.0	0.027	0.018	0.0	0.028

Table 9.11: Analysis of the presence of backbones linking the Missings during the pre-bubble period

Now, we conduct the same analysis for the Entrants' ego networks. The results obtained are shown in Table 9.12. The structure and the semantics of this table are similar to the ones of Tables 9.10 and 9.11. From the analysis of Table 9.12 we can conclude that there is no backbone linking the Entrants during the pre-bubble period. This result is justified considering that, during this period, the Entrants were not power addresses. The presence of some nodes of the Survivors or of the Missings in the alters of the Entrants is simply due to the fact that the Survivors and the Missings were power addresses during the pre-bubble period.

To also give a graphical idea of the results on the presence of backbones obtained above, we consider a social network $\mathcal{N}^F_{Pre}$, obtained from $\mathcal{N}_{Pre}$ considering only the power `from_addresses`.

In order to extract a subnet of $\mathcal{N}^F_{Pre}$ containing nodes strongly connected to each other, we should consider the cliques of $\mathcal{N}^F_{Pre}$. However, since the computation of cliques is an NP-hard problem, we decided to use a relaxation of the concept of clique and focused on k-core. We recall that a k-core of a network $\mathcal{N}$ is a connected

	Ego networks of $\mathcal{E}^F$			*Ego networks of* $\mathcal{E}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k=1$	0.326	0.140	0.163	0.194	0.0	0.222
$k=2$	0.140	0.0	0.023	0.083	0.0	0.056
$k=3$	0.070	0.0	0.0	0.056	0.0	0.056
$k=4$	0.0	0.0	0.0	0.056	0.0	0.056
$k=5$	0.0	0.0	0.0	0.056	0.0	0.028
$k=6$	0.0	0.0	0.0	0.056	0.0	0.028
$k=7$	0.0	0.0	0.0	0.056	0.0	0.028
$k=8$	0.0	0.0	0.0	0.056	0.0	0.028
$k=9$	0.0	0.0	0.0	0.056	0.0	0.028

Table 9.12: Analysis of the presence of backbones linking the Entrants during the pre-bubble period

maximal induced subnetwork of $\mathcal{N}$ in which all nodes have degree at least k. A k-core can be used as an indicator of the presence of backbones. In fact, if some nodes, say $n_1, n_2, \ldots, n_q$, belong to a k-core, then each of them will be connected to at least k of the other ones.

Consider the 5-core of $\mathcal{N}^F_{Pre}$ shown in Figure 9.4. In it, we indicate in yellow the Survivors nodes, in red the Missings nodes and in blue all the other ones. The 5-core consists of 175 nodes. As we can see from the figure, there is a strong backbone connecting 32 Survivors nodes and another weaker backbone connecting 13 Missings nodes. Consider, now, the 7-core of $\mathcal{N}^F_{Pre}$ shown in Figure 9.5. It contains even more strongly connected nodes than the 5-core. The total number of its nodes is 86. Again, there is a strong backbone connecting 19 Survivors nodes and a weaker backbone connecting 5 Missings nodes. Both these figures provide a graphical idea of the analytical results found previously.

The next analysis concerns the Survivors', the Missings' and the Entrants' ego networks during the bubble period. The results obtained by carrying out the same tasks seen for the pre-bubble period are reported in Tables 9.13, 9.14 and 9.15.

From the analysis of these tables we can detect the following knowledge patterns:

- There is a very strong backbone linking the Survivors, as can be seen by examining Table 9.13.
- In the same table, we can observe that there are some Entrants and Missings nodes among the alters of the Survivors' ego networks. This can be explained taking into account that the Entrants are power addresses during the bubble period, while the Missings, although not anymore, were power addresses in the period immediately before.
- Table 9.14 shows that there is no longer a backbone linking the Missings.

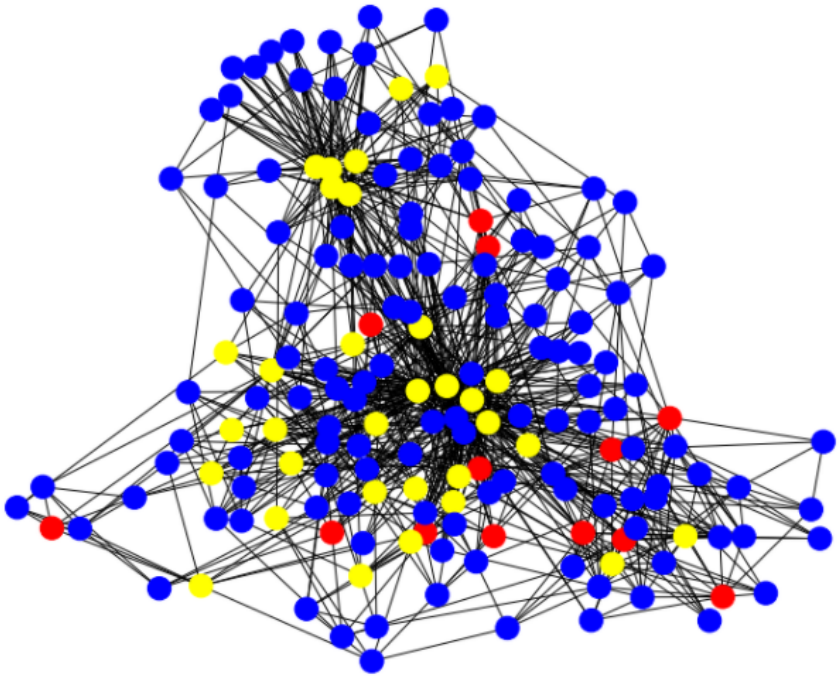


Fig. 9.4: A 5-core of $\mathcal{N}^F_{Pre}$

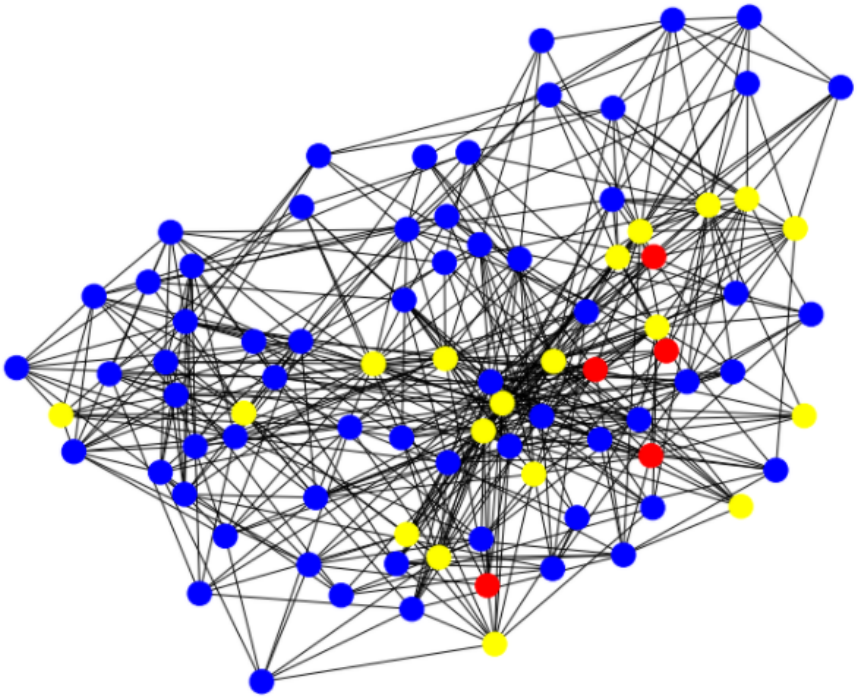


Fig. 9.5: A 7-core of $\mathcal{N}^F_{Pre}$

- Table 9.15 reveals that a backbone linking the Entrants starts to exist, even if it is not very strong yet.

	Ego networks of $\mathcal{S}^F$			*Ego networks of* $\mathcal{S}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k=1$	0.824	0.451	0.461	0.750	0.688	0.714
$k=2$	0.598	0.245	0.333	0.554	0.509	0.491
$k=3$	0.431	0.167	0.284	0.312	0.357	0.339
$k=4$	0.373	0.127	0.265	0.143	0.223	0.232
$k=5$	0.304	0.078	0.225	0.098	0.152	0.161
$k=6$	0.265	0.069	0.216	0.071	0.062	0.134
$k=7$	0.196	0.029	0.147	0.036	0.054	0.098
$k=8$	0.147	0.020	0.137	0.027	0.045	0.089
$k=9$	0.108	0.020	0.118	0.027	0.036	0.089

Table 9.13: Analysis of the presence of backbones linking the Survivors during the bubble period

	Ego networks of $\mathcal{M}^F$			*Ego networks of* $\mathcal{M}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k=1$	0.338	0.125	0.138	0.283	0.166	0.217
$k=2$	0.163	0.054	0.023	0.095	0.034	0.049
$k=3$	0.111	0.035	0.006	0.042	0.014	0.026
$k=4$	0.065	0.021	0.004	0.026	0.010	0.014
$k=5$	0.044	0.015	0.002	0.020	0.008	0.012
$k=6$	0.021	0.013	0.0	0.020	0.006	0.010
$k=7$	0.019	0.010	0.0	0.016	0.004	0.008
$k=8$	0.010	0.004	0.0	0.010	0.004	0.006
$k=9$	0.006	0.002	0.0	0.010	0.004	0.006

Table 9.14: Analysis of the presence of backbones linking the Missings during the bubble period

	Ego networks of $\mathcal{E}^F$			*Ego networks of* $\mathcal{E}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k=1$	0.337	0.572	0.217	0.335	0.477	0.335
$k=2$	0.175	0.295	0.127	0.152	0.284	0.152
$k=3$	0.096	0.169	0.084	0.081	0.142	0.081
$k=4$	0.066	0.096	0.054	0.061	0.076	0.051
$k=5$	0.048	0.066	0.042	0.061	0.046	0.030
$k=6$	0.036	0.030	0.036	0.056	0.030	0.025
$k=7$	0.024	0.024	0.036	0.046	0.030	0.020
$k=8$	0.024	0.0	0.036	0.041	0.025	0.015
$k=9$	0.024	0.0	0.036	0.036	0.025	0.015

Table 9.15: Analysis of the presence of backbones linking the Entrants during the bubble period

To also give a graphical idea of these results, we consider the $\mathcal{N}_B^F$ network. It is defined similarly to N_{Pre}^F, but taking the bubble period into account. We also consider the corresponding 5-core and 7-core, shown in Figures 9.6 and 9.7, respectively. In them, we represent the Survivors nodes in yellow, the Entrants nodes in

green and all the other nodes in blue. The 5-core consists of 149 nodes. Here, there is a very strong backbone involving 47 Survivors nodes and a weaker one involving 17 Entrants nodes. The 7-core consists of 67 nodes. Also in this case there is a very strong backbone connecting 30 Survivors nodes and a weaker backbone connecting 13 Entrants nodes.

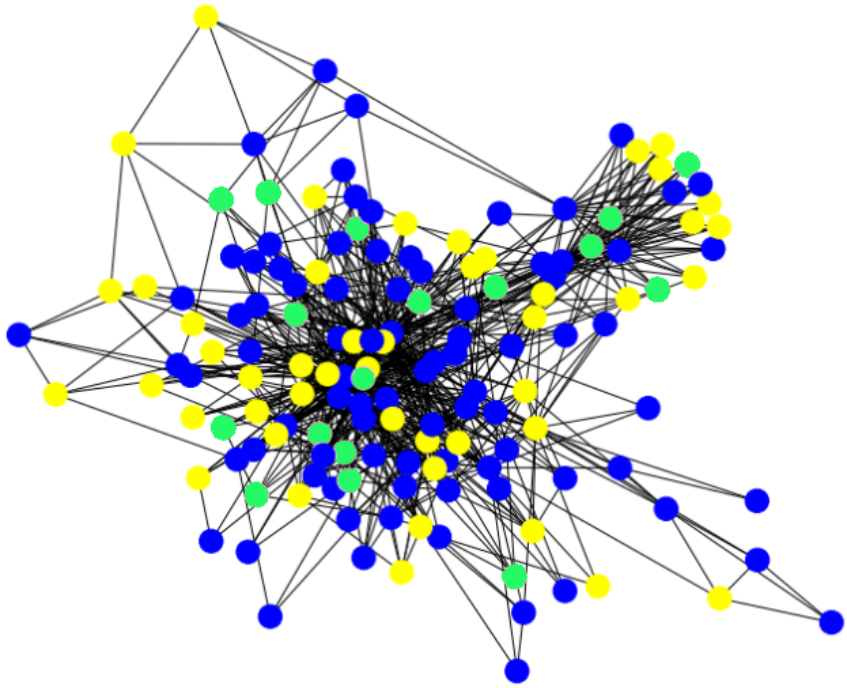


Fig. 9.6: A 5-core of $\mathcal{N}_B^F$

The last analysis concerns the Survivors', the Missings' and the Entrants' ego networks during the post-bubble period. The results obtained are reported in Tables 9.16, 9.17 and 9.18.

	Ego networks of $\mathcal{S}^F$			*Ego networks of* $\mathcal{S}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k = 1$	0.716	0.490	0.353	0.741	0.768	0.518
$k = 2$	0.510	0.265	0.206	0.607	0.598	0.330
$k = 3$	0.363	0.167	0.167	0.384	0.446	0.188
$k = 4$	0.265	0.147	0.108	0.223	0.366	0.143
$k = 5$	0.216	0.137	0.088	0.116	0.268	0.089
$k = 6$	0.186	0.098	0.078	0.080	0.223	0.089
$k = 7$	0.108	0.069	0.059	0.062	0.134	0.080
$k = 8$	0.088	0.059	0.049	0.045	0.098	0.062
$k = 9$	0.059	0.039	0.039	0.045	0.062	0.045

Table 9.16: Analysis of the presence of backbones linking the Survivors during the post-bubble period

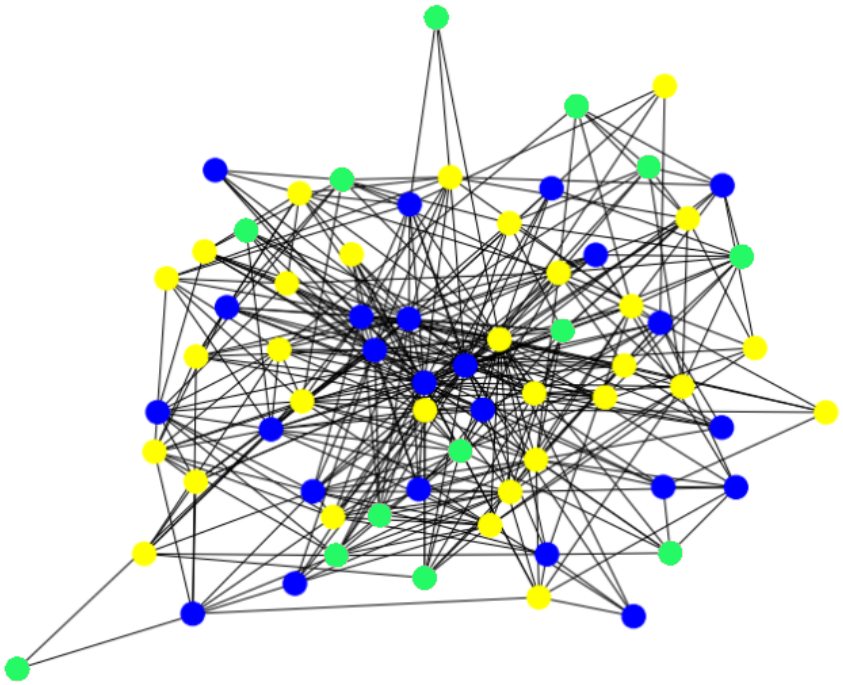


Fig. 9.7: A 7-core of $\mathcal{N}_B^F$

	Ego networks of $\mathcal{M}^F$			*Ego networks of* $\mathcal{M}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k = 1$	0.263	0.193	0.119	0.274	0.167	0.070
$k = 2$	0.122	0.126	0.015	0.067	0.040	0.027
$k = 3$	0.056	0.081	0.007	0.032	0.019	0.013
$k = 4$	0.033	0.059	0.007	0.027	0.011	0.008
$k = 5$	0.026	0.052	0.004	0.019	0.011	0.008
$k = 6$	0.015	0.041	0.004	0.016	0.008	0.005
$k = 7$	0.011	0.033	0.004	0.013	0.005	0.003
$k = 8$	0.011	0.022	0.0	0.011	0.005	0.0
$k = 9$	0.007	0.011	0.0	0.008	0.005	0.0

Table 9.17: Analysis of the presence of backbones linking the Missings during the post-bubble period

From the analysis of these tables we can deduce the following knowledge patterns:

- There is a strong backbone linking the Survivors, as can be seen in Table 9.16. Comparing Tables 9.13 and 9.16 we can see that this backbone, while continuing to remain strong, undergoes a weakening, compared to the pre-bubble period. This is physiological because, during the post-bubble period, the number of transactions made decreased considerably with respect to the ones of the bubble period.

	Ego networks of $\mathcal{E}^F$			*Ego networks of* $\mathcal{E}^T$		
	Nodes of $\mathcal{S}^F$	*Nodes of* $\mathcal{E}^F$	*Nodes of* $\mathcal{M}^F$	*Nodes of* $\mathcal{S}^T$	*Nodes of* $\mathcal{E}^T$	*Nodes of* $\mathcal{M}^T$
$k = 1$	0.331	0.651	0.211	0.431	0.675	0.376
$k = 2$	0.187	0.380	0.133	0.223	0.457	0.096
$k = 3$	0.133	0.193	0.084	0.091	0.310	0.036
$k = 4$	0.090	0.108	0.048	0.076	0.198	0.020
$k = 5$	0.054	0.078	0.048	0.071	0.122	0.015
$k = 6$	0.036	0.066	0.048	0.061	0.086	0.015
$k = 7$	0.036	0.042	0.048	0.061	0.056	0.015
$k = 8$	0.030	0.018	0.048	0.056	0.051	0.015
$k = 9$	0.024	0.018	0.042	0.056	0.046	0.010

Table 9.18: Analysis of the presence of backbones linking the Entrants during the post-bubble period

- We continue to observe the presence of some Entrants and Missings nodes among the alters of the Survivors' ego networks. The reasons for this fact are the same as those seen for the bubble period.
- The backbone linking the Missings, which had already started to disappear during the bubble period, has completely dissolved, as evidenced by the further decrease of the values in the fourth and seventh columns of Table 9.17, compared to the corresponding ones of Table 9.14.
- The backbone linking the Entrants, which was already visible during the bubble period, is further consolidated during the post-bubble period, as can be seen by examining Table 9.18.

Also in this case we can use k-cores to give a graphical idea of the results obtained. For this purpose, we consider the network $\mathcal{N}^F_{Post}$, obtained similarly to $\mathcal{N}^F_{Pre}$ and $\mathcal{N}^F_B$. We also consider the corresponding 5-core and 7-core, shown in Figures 9.8 and 9.9, respectively. The meaning of the colors of the nodes in this figure is the same as the one seen for Figures 9.6 and 9.7. In this case, the 5-core consists of 202 nodes. Here, there is a strong backbone linking 42 Survivors nodes. Furthermore, there is a backbone linking 31 Entrants nodes. Note that, compared to the bubble period, the backbone linking the Entrants nodes has strengthened. A similar reasoning also applies to the 7-core. It consists of 111 nodes. In it, we can observe a strong backbone linking 24 Survivors nodes and a backbone linking 16 Entrants nodes. Also this last backbone appears strengthened compared to the corresponding one relative to the 7-core during the bubble period shown in Figure 9.7. All these graphical results are totally in line with the analytical ones relative to the post-bubble period presented above.

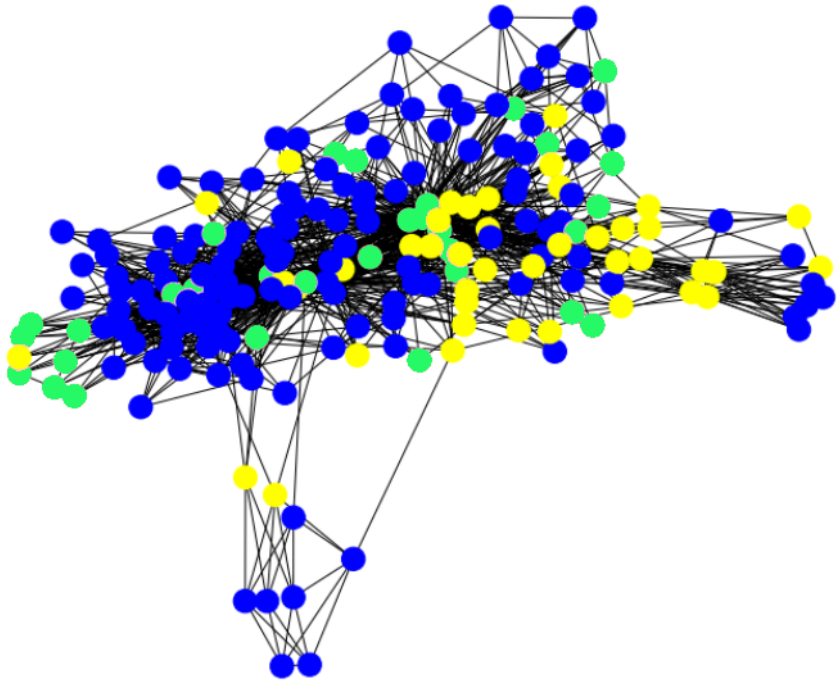


Fig. 9.8: A 5-core of $\mathcal{N}^F_{Post}$

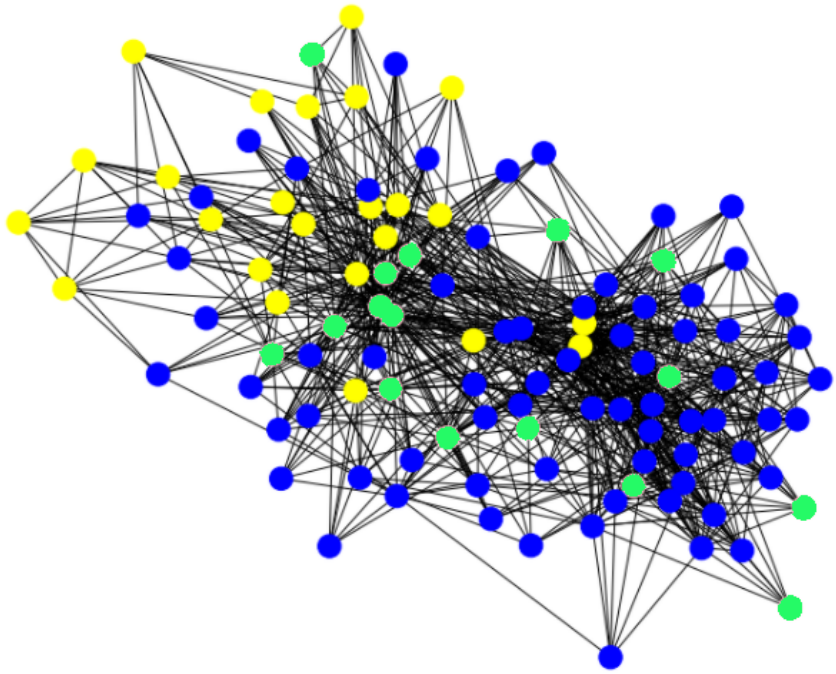


Fig. 9.9: A 7-core of $\mathcal{N}^F_{Post}$

9.4.2 Graphical backbone evaluations through k-trusses

In Section 9.4.1, we have said that, in order to verify the possible existence of backbones among Survivors, Missings or Entrants, the concept of cliques could be used.

We have also said that the computation of cliques was a NP-hard problem and, for this reason, we chose to replace cliques with k-cores. In fact, the k-core concept is a relaxation of the clique concept and, unlike cliques, the computation of k-cores can be done in polynomial time. However, it is worth checking that the results obtained with k-core are not unduly influenced by the properties of this structure. One way to carry out this verification is to repeat the experiments performed with k-cores using another data structure that can be considered a relaxation of the clique concept and can be computed in polynomial time. To this end, we focused on the concept of k-truss [200]. A k-truss is a non-trivial, one component subgraph such that each edge is reinforced by at least $k-2$ pairs of edges making a triangle with that edge. Observe that each clique of order k is contained in a k-truss, whereas a k-truss does not necessarily contain a clique of order k. Furthermore, each k-truss is a subgraph of a (k-1)-core. All these properties support the idea that a k-truss is a concept that lies somewhere between the clique concept, which is too restrictive, and the k-core concept, which is too lax. Furthermore, similarly to k-cores and unlike cliques, the computation of k-trusses requires polynomial time.

At this point, similarly to what we did for the k-core, we computed the 5-truss of $\mathcal{N}^F_{Pre}$ and we saw that: *(i)* it consists of 152 nodes; *(ii)* there is a strong backbone connecting 27 Survivors; *(iii)* there is a weaker backbone connecting 7 Missings. Next, we computed the 7-truss of $\mathcal{N}^F_{Pre}$ and we obtained that: *(i)* it consists of 74 nodes; *(ii)* there is a strong backbone connecting 16 Survivors; *(iii)* there is no significant backbone among Missings.

Proceeding with our analysis, we computed the 5-truss of $\mathcal{N}^F_B$; analyzing it, we saw that: *(i)* it consists of 134 nodes; *(ii)* there is a very strong backbone involving 41 Survivors; *(iii)* there is an additional backbone involving 15 Entrants. The analysis of the 7-truss of $\mathcal{N}^F_B$ allows us to say that: *(i)* it consists of 61 nodes; *(ii)* there is a very strong backbone involving 26 Survivors; *(iii)* there is a weaker backbone involving 10 Entrants.

Our analysis on k-trussed ends with the computation of the 5-truss and 7-truss of $\mathcal{N}^F_{Post}$. Regarding the 5-truss we obtained that: *(i)* it consists of 194 nodes; *(ii)* there is a strong backbone connecting 36 Survivors; *(iii)* there is an additional backbone connecting 26 Entrants. Regarding the 7-truss of $\mathcal{N}^F_{Post}$ we saw that: *(i)* it consists of 96 nodes; *(ii)* there is a strong backbone connecting 22 Survivors; *(iii)* there is an additional backbone connecting 12 Entrants.

Comparing the results obtained through the k-truss analysis with those regarding the k-core analysis shown in Section 9.4.1, we can observe that they are similar. In fact, the k-truss analysis confirms everything was found through the k-core analysis. The only exception regards the fact that the k-core analysis detects a backbone

(albeit a very weak one) between the Missings in the 7-core associated with $\mathcal{N}^{F}_{Pre}$. Such a backbone is not detected in the corresponding 7-truss. However, this minimal difference can be explained considering that the detected backbone of the 7-core is anyway very weak as well as taking into account that the concept of k-truss is more "severe" than the one of k-core.

At the end of this analysis, we can conclude that the strong similarity of the results obtained using k-cores and k-trusses allows us to say that these are intrinsic in the data and are not unduly caused by the properties of the k-cores.

9.4.3 Defining the identikit of bubble speculators

In the previous section, we extracted some knowledge patterns involving various kinds of addresses present in a cryptocurrency blockchain. In this section, we want to verify whether the suitable integration of these knowledge patterns allows us to build an identikit of speculators.

In performing this task we start with the information about the ego network obtained in Section 9.3.3. It tells us that: *(i)* in the pre-bubble period, the Survivors have much larger ego networks than the other nodes; *(ii)* in the bubble and post-bubble periods, the Survivors have larger ego networks than the other nodes; *(iii)* in the bubble period, the Survivors' ego networks are much larger than even the Entrants' ego networks; this difference fades in the post-bubble period. Recall that having a large ego network means having the possibility to influence a large number of nodes.

Now, we consider the information on backbones extracted in Section 9.4.1. It tells us that: *(i)* in the pre-bubble period, there is a strong backbone among the Survivors and a weaker backbone among the Missings; *(ii)* in the bubble period, there is a very strong backbone among the Survivors and a weaker backbone among the Entrants; this last is stronger than the corresponding one of the bubble period. Recall that the presence of a backbone among a set of nodes is an indicator that they tend to act in a coordinated way with each other.

We continue our investigation by considering the characteristics of the future main actors extracted in Section 9.4.4. In that section, we saw that the address that best survive a bubble must be sought among those that, in the pre-bubble and bubble periods, made the most transactions and had the most contacts. But, from what we saw in Section 9.3.3, the addresses with such characteristics are first those of the Survivors and then those of the Entrants.

Finally, an analysis of the nodes active in the period corresponding to the Ethereum bubble of the years 2017-2018 that are still active today also leads us to the same re-

sults, namely that most of the Survivors and a good portion of the Entrants present in the 2017-2018 Ethereum bubble are still active today.

All these considerations lead us to conclude that indeed in the Ethereum speculative bubble of 2017-2018, a group of speculators existed. Regarding the profile of the users belonging to this group, we can conclude that most of them were Survivors and were already present in the pre-bubble period. They are flanked in the bubble period by a group of speculators that formed the Entrants set. Initially, these were not the leaders of the phenomenon; at first, the leadership was of the Survivors alone. However, as time passed, the Entrants gradually consolidated and reached the level of leadership that previously characterized the Survivors alone.

9.4.4 Predicting the characteristics of the main future actors

All the previous analyses are mainly descriptive and diagnostic. In this section, instead, we want to go one step further proposing a predictive analysis with the aim of understanding, during a period (specifically, pre-bubble, bubble), what are the features of the addresses that will probably play a leading role during the next period (specifically, bubble, post-bubble). The importance of this analysis (in itself already evident) is reinforced by the results obtained in the previous section, telling us that these main actors are often connected by backbones. Consequently, identifying (and possibly acting on) some of them gives the possibility to identify (and act on) most of the others connected through the backbones.

In Table 9.19, we show the number of transactions, the number of contacts and the average value of transactions for the following addresses:

- T^F_{Pre}: the power `from_addresses` in the pre-bubble period.
- $\mathcal{S}^F$: the Survivors `from_addresses`. By definition, each element of $\mathcal{S}^F$ must also be an element of T^F_{Pre} and an element of T^F_B, i.e., the power `from_addresses` in the bubble period.
- $\mathcal{M}^F$: the Missings `from_addresses`. By definition, each element of $\mathcal{M}^F$ must also be an element of T^F_{Pre}, while it cannot belong to T^F_B.
- $\mathcal{E}^F_{Pre}$: the `from_addresses` that appeared in the bubble period but were already present (albeit not as power addresses) in the pre-bubble period. By definition, each element of $\mathcal{E}^F_{Pre}$ must also be an element of T^F_B, while it cannot belong to T^F_{Pre}.

From the analysis of this table we can see that the addresses of $\mathcal{S}^F$ have a significantly higher number of transactions and contacts than the corresponding ones not only of $\mathcal{M}^F$ and $\mathcal{E}^F_{Pre}$ but also of T^F_{Pre}. Instead, the average value of transactions is smaller for $\mathcal{S}^F$, $\mathcal{M}^F$ and T^F_{Pre} than for $\mathcal{E}^F_{Pre}$.

	T^F_{Pre}	$\mathcal{S}^F$	$\mathcal{M}^F$	$\mathcal{E}^F_{Pre}$
Average Number of Transactions	30,346.55	175,729.30	11,064.18	473.83
Average Number of Contacts	4,817.39	27,088.52	1,259.26	242.98
Average Value of Transactions (Eth)	8.65	8.18	7.32	106.53

Table 9.19: Average number of transactions, average number of contacts and average values of transactions for T^F_{Pre}, $\mathcal{S}^F$, $\mathcal{M}^F$ and $\mathcal{E}^F_{Pre}$

This result is even more evident considering Figure 9.10 (resp., 9.11). Here, we show the distribution of the addresses of $\mathcal{S}^F$ and $\mathcal{M}^F$ against the number of transactions (resp., contacts) of T^F_{Pre}. The abscissae axis is divided into deciles. In the figure, we indicate the decile with the highest values with D_{10} and the one with the lowest value with D_1. Figure 9.10 shows that most of the addresses of $\mathcal{S}^F$ belong to the highest deciles of T^F_{Pre}. This does not happen for the addresses of $\mathcal{M}^F$ that show a rather uniform distribution among the deciles of T^F_{Pre}, except for the lowest decile where they are almost absent. Figure 9.11 shows a similar trend except for the lowest decile, which comprises a lot of addresses for both $\mathcal{S}^F$ and $\mathcal{M}^F$.

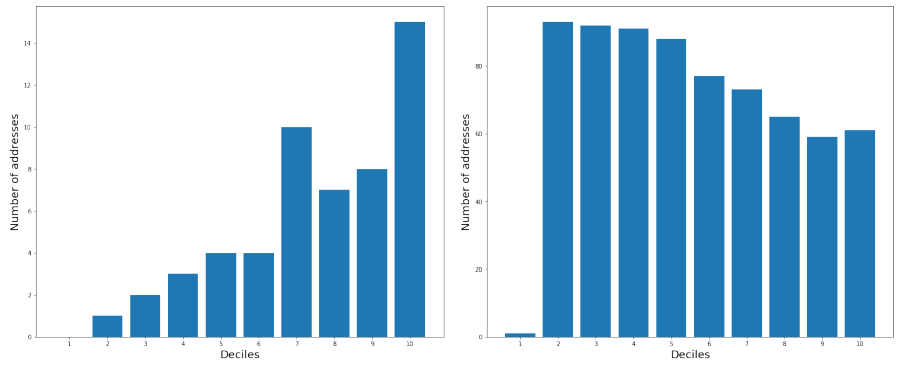


Fig. 9.10: Distribution of the addresses of $\mathcal{S}^F$ (at left) and $\mathcal{M}^F$ (at right) against the number of transactions of T^F_{Pre}

Both Table 9.19 and Figures 9.10 and 9.11 give us the same important following indication: "The addresses that will survive a bubble are to be searched among the ones that, in the pre-bubble period, have carried out the highest numbers of transactions and have the highest numbers of contacts". This indication is very strong for the number of transactions while it is a bit weaker for the number of contacts. In fact, as for this last parameter, we can see that the lowest decile contains a certain number not only of Missings nodes but also of Survivors ones.

Instead, Table 9.19 does not seem to give any indication on how searching, in the pre-bubble period, the future Entrants that will be among the main actors in the bubble and post-bubble periods.

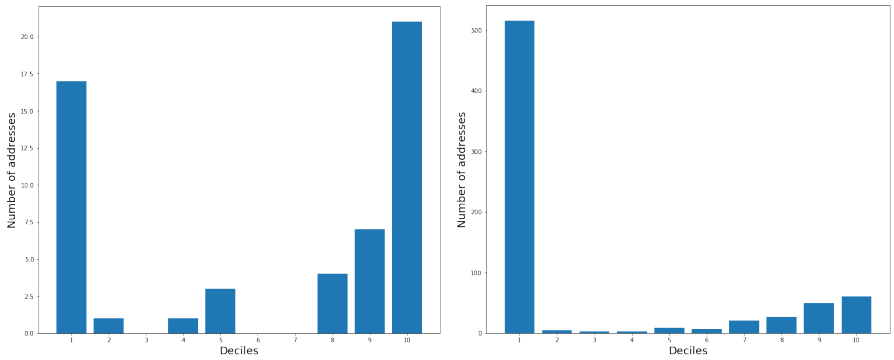


Fig. 9.11: Distribution of the addresses of $\mathcal{S}^F$ (at left) and $\mathcal{M}^F$ (at right) against the number of contacts of T^F_{Pre}

All previous analyses performed for `from_addresses` in the pre-bubble period can be repeated for `to_addresses` in the same period. In Table 9.20, we report the average number of transactions, the average number of contacts and the average value of transactions for T^T_{Pre}, $\mathcal{S}^T$, $\mathcal{M}^T$ and $\mathcal{E}^T_{Pre}$ (the latter defined similarly to $\mathcal{E}^F_{Pre}$, but for `to_addresses` instead of `from_addresses`). Furthermore, in Figure 9.12 (resp., 9.13), we show the distribution of the addresses of $\mathcal{S}^T$ and $\mathcal{M}^T$ against the number of transactions (resp., contacts) of T^T_{Pre}. Both the table and the two figures confirm, for `to_addresses`, the same results that we found previously for `from_addresses`.

	T^T_{Pre}	$\mathcal{S}^T$	$\mathcal{M}^T$	$\mathcal{E}^T_{Pre}$
Average Number of Transactions	28,035.76	138,663.66	10,121.69	599.78
Average Number of Contacts	5,329.76	23,007.33	2,165.56	294.28
Average Value of Transactions (Eth)	9.05	6.79	14.17	4.86

Table 9.20: Average number of transactions, average number of contacts and average value of transactions for T^T_{Pre}, $\mathcal{S}^T$, $\mathcal{M}^T$ and $\mathcal{E}^T_{Pre}$

So far we have examined pre-bubble data to identify some characteristics allowing us to predict who will be the main actors of the bubble period. Now, we want to do the same activity but examining bubble data to look for features allowing us to predict who will be the protagonists of the post-bubble period. In this analysis, we consider the following addresses:

- T^F_B: the top 1000 `from_addresses` in the bubble period;
- $\mathcal{S}^F$: the Survivors `from_addresses`;
- $\mathcal{E}^F$: the Entrants `from_addresses`.

In Table 9.21, we show the average number of transactions, the average number of contacts and the average value of transactions for T^F_B, $\mathcal{S}^F$ and $\mathcal{E}^F$.

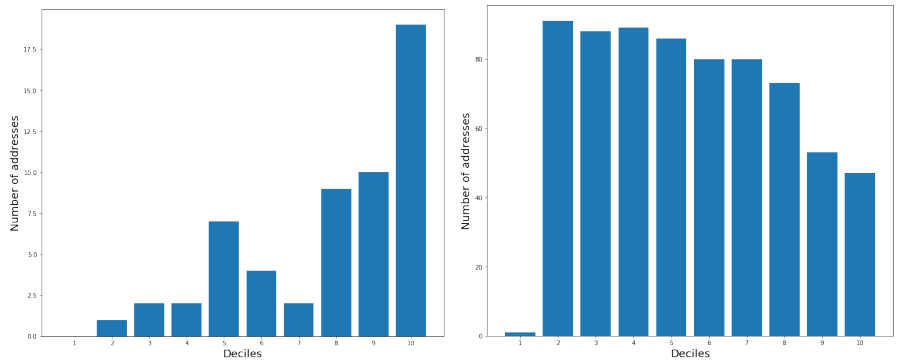


Fig. 9.12: Distribution of the addresses of $\mathcal{S}^T$ (at left) and $\mathcal{M}^T$ (at right) against the number of transactions of T^T_{Pre}

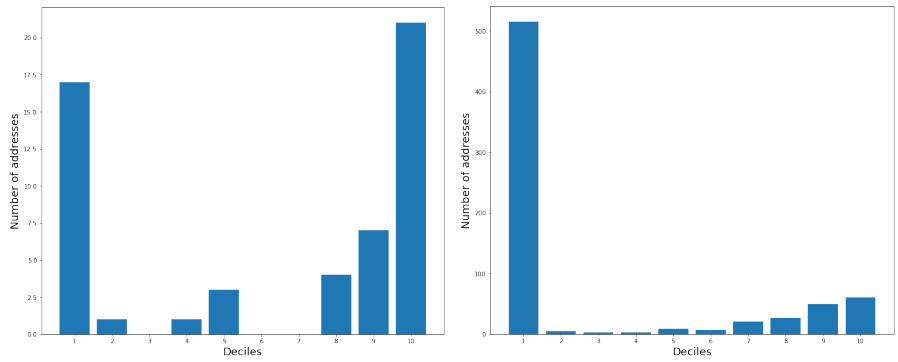


Fig. 9.13: Distribution of the addresses of $\mathcal{S}^T$ (at left) and $\mathcal{M}^T$ (at right) against the number of contacts of T^T_{Pre}

	T^F_B	$\mathcal{S}^F$	$\mathcal{E}^F$
Average Number of Transactions	45,418.29	266,183.77	46,010.31
Average Number of Contacts	10,100.95	55,029.89	12,851.75
Average Value of Transactions (Eth)	2.43	2.49	3.73

Table 9.21: Average number of transactions, average number of contacts and average value of transactions for T^F_B, $\mathcal{S}^F$ and $\mathcal{E}^F$

From the analysis of Table 9.21 we can see that, once again, it is easy to identify the Survivors of the post-bubble period. In fact, they generally have a significantly higher number of transactions and contacts than the other power `from_addresses`. Instead, the Entrants are not easily distinguishable, because they have only slightly more transactions and contacts than the other power `from_addresses`. This represents a confirmation of what we had deduced from the analysis of Tables 9.13 - 9.18 and Figures 9.6 - 9.9, where we derived that the set of the Entrants is formed during the bubble period but it consolidates only during the post-bubble period.

This result is confirmed and substantially reinforced by Figures 9.14 and 9.15. In them, we can see that the Survivors are in the highest deciles, and this was expected considering the results of Table 9.21. However, a similar trend, although less marked, is also found for the Entrants. This represents a further important result because it allows us to define, at least partially, which nodes will be the Entrants in the post-bubble period. Similarly to what happened in the pre-bubble period, the distribution against the number of transactions is better than the one against the number of contacts in discriminating the Survivors and the Entrants against the other nodes during the post-bubble period. Indeed, in the case of the number of contacts, there is a certain number of addresses in the lowest decile, which, in fact, represents an outlier.

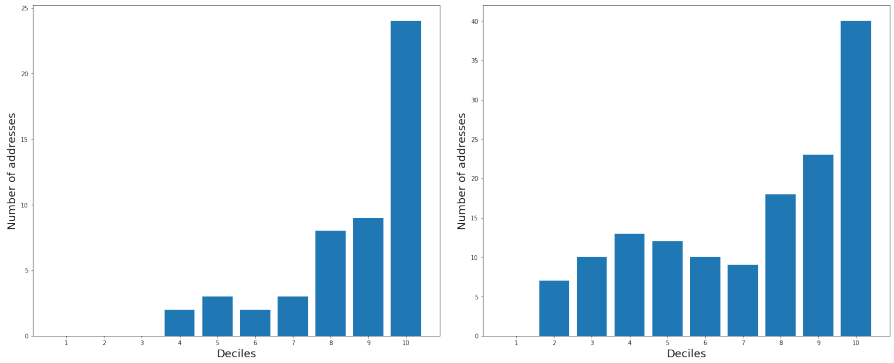


Fig. 9.14: Distribution of the addresses of $\mathcal{S}^F$ (at left) and $\mathcal{E}^F$ (at right) against the number of transactions of T_B^F

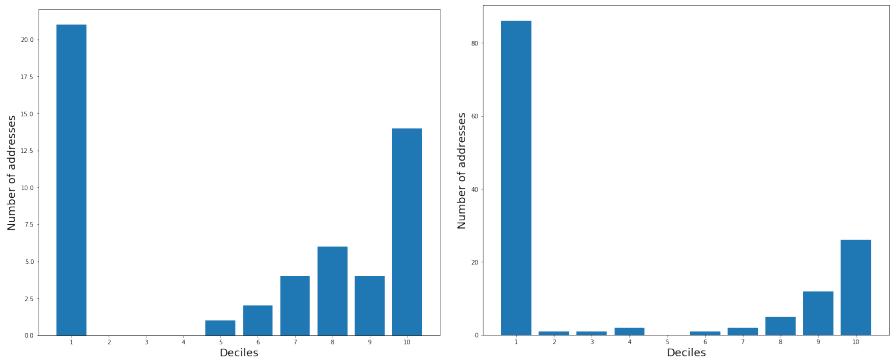


Fig. 9.15: Distribution of the addresses of $\mathcal{S}^F$ (at left) and $\mathcal{E}^F$ (at right) against the number of contacts of T_B^F

Both Table 9.21 and Figures 9.14 and 9.15 give us the same important following indication: "The addresses that will survive a speculative bubble are to be searched

among those that, in the bubble period, have carried out the highest numbers of transactions and have the highest numbers of contacts. If they also had this property in the pre-bubble period they belong to the Survivors, otherwise they belong to the Entrants.".

All previous analyses performed for `from_addresses` in the bubble period can be repeated for `to_addresses` in the same period. In Table 9.22, we report the average number of transactions, the average number of contacts and the average value of transactions for T_B^T, $\mathcal{S}^T$ and $\mathcal{E}^T$. Furthermore, in Figure 9.16 (resp., 9.17), we show the distribution of the addresses of $\mathcal{S}^T$ and $\mathcal{E}^T$ against the number of transactions (resp., contacts) of T_B^T.

	T_B^T	$\mathcal{S}^T$	$\mathcal{E}^T$
Average Number of Transactions	49,912.89	219,068.94	58,823.91
Average Number of Contacts	11,963.66	45,949.34	14,134.10
Average Value of Transactions (Eth)	1.90	1.98	1.71

Table 9.22: Average number of transactions, average number of contacts and average value of transactions for T_B^T, $\mathcal{S}^T$ and $\mathcal{E}^T$

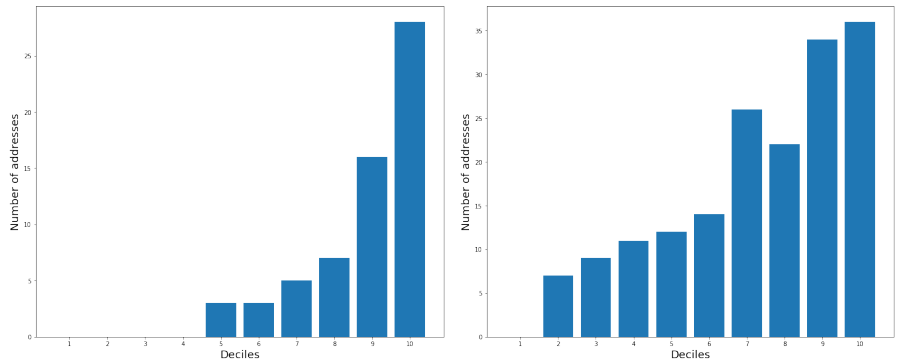


Fig. 9.16: Distribution of the addresses of $\mathcal{S}^T$ (at left) and $\mathcal{E}^T$ (at right) against the number of transactions of T_B^T

Table 9.22 and Figure 9.16 confirm, for `to_addresses`, the same results we found previously for `from_addresses`. Figures 9.17, if compared with Figure 9.15, shows that, as for the number of contacts of the Survivors, the outlier represented by the lowest decile is strongly reduced. Instead, this outlier remains for the Entrants. However, for this last category of addresses, we can observe that, similarly to what happens for the Survivors, and differently from what happened in Figure 9.15, most of the addresses are in the highest deciles, even if, once again, this phenomenon is less marked than the corresponding one observed for the Survivors.

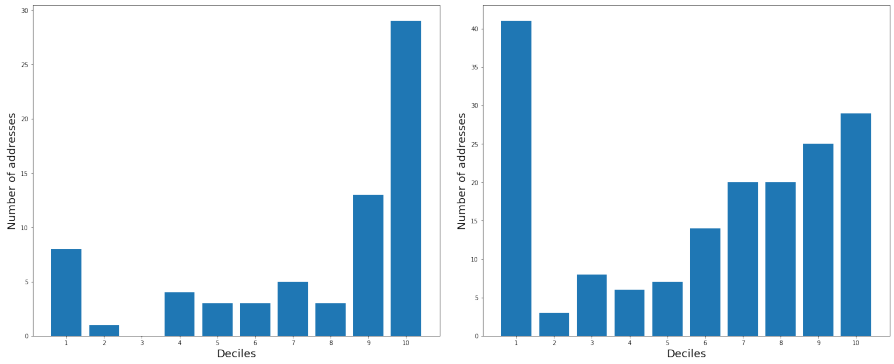


Fig. 9.17: Distribution of the addresses of $\mathcal{S}^T$ (at left) and $\mathcal{E}^T$ (at right) against the number of contacts of T_B^T

As a last analysis, we investigated how the power addresses of the post-bubble period behaved during the months following the time interval considered for our dataset, i.e., from January 2019 until today. For this purpose, we considered three subsets of the power addresses, i.e., the Survivors, the Entrants and the other nodes (hereafter, the Others), and we examined the date of the last transaction for them. The distribution of the Survivors (resp., the Entrants, the Others) against this date is shown in Figure 9.18 (resp., 9.19, 9.20) for `from_addresses`, and in Figure 9.21 (resp., 9.22, 9.23) for `to_addresses`. From the analysis of these figures we can observe that:

- As for `from_addresses`, we can see that most of the Survivors are still active. Many Entrants are also active but, unlike the Survivors, there is a fraction of them that ceased to operate in the second half of 2019. The date of the end of activity of the Others is, instead, more uniformly distributed. This is a further confirmation that the Survivors represent the vast part of the guiding users in Ethereum.
- As far as `to_addresses` are concerned, we can see that most of the Survivors and the Entrants are still active. The date of the end of activity of the Others is distributed in a more balanced way, even if there is a large amount of addresses still active also in this case. Therefore, as for `to_addresses`, we can deduce that the Survivors include most of the guiding users in Ethereum. However, differently from what happens for `from_addresses`, they have been flanked as leaders by the Entrants.

9.4.5 Adoption of our approach in the next speculative bubble

The main objective of this Chapter was to study the cryptocurrency speculative bubble during the years 2017-2018 to understand the behavior of some particularly in-

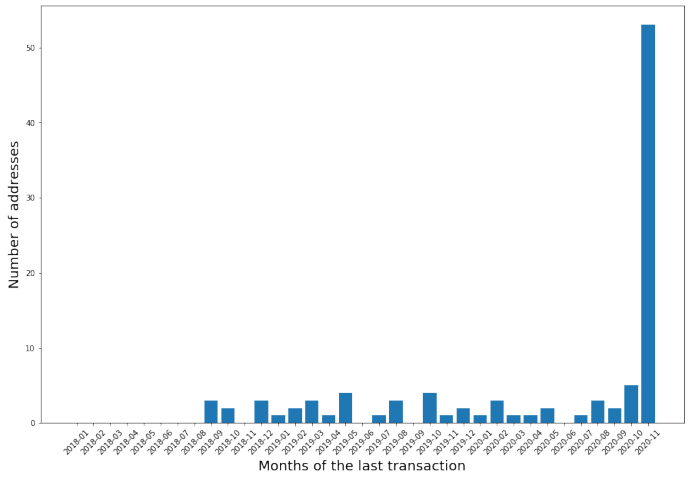


Fig. 9.18: Distribution of the Survivors (`from_addresses`) against the date of the last transaction

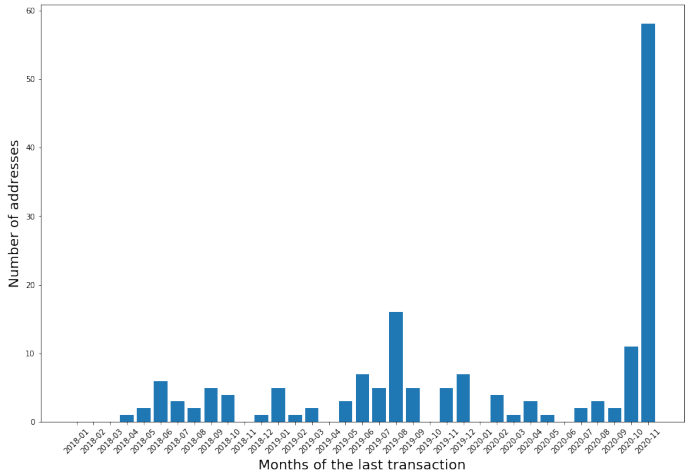


Fig. 9.19: Distribution of the Entrants (`from_addresses`) against the date of the last transaction

teresting categories of users and to try to identify a profile of possible speculators. However, the knowledge pattern extracted in this way do not represent only an abstract knowledge related to a past event, but can become an extremely valuable tool for the future.

In fact, the cryptocurrency context is considered a highly speculative environment by many graduates of the Nobel Memorial Prize in Economic Sciences, central bankers and investors. Speculations on cryptocurrencies have also been observed recently. For example, on March 8^{th}, 2020 the price of Bitcoin was 8,901 USD. On March 12^{th}, 2020, it was 6,206 USD, with a decrease of about 30%. In October 2020 this price was already doubled again and was about 13,000 USD. On January 3^{rd}, 2021 the price of Bitcoin was 34,792 USD; the next day it decreased by 17%. On

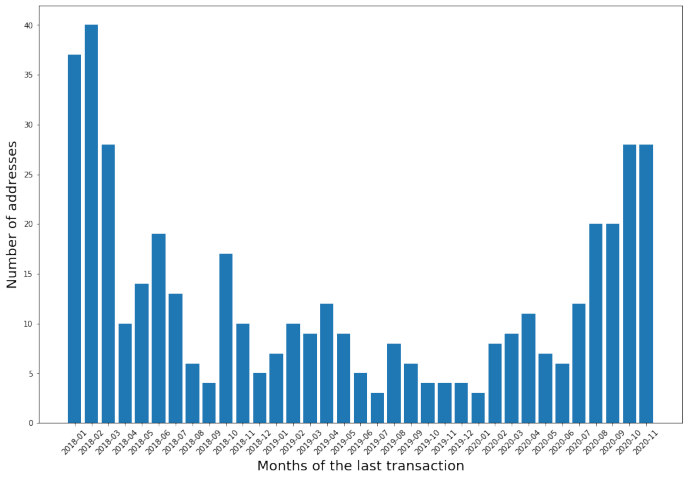


Fig. 9.20: Distribution of the Others (`from_addresses`) against the date of the last transaction

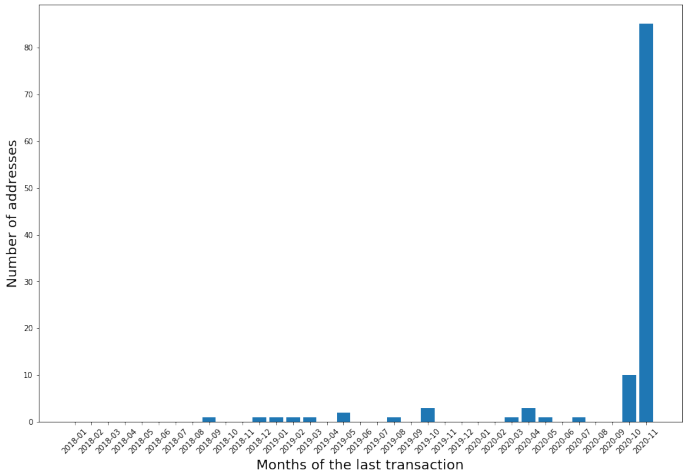


Fig. 9.21: Distribution of the Survivors (`to_addresses`) against the date of the last transaction

January 8^{th}, 2021 its value exceeded 40,000 USD and on February 16^{th}, 2021 it exceeded 50,000 USD. In March 2021 its value was 58,734 USD, while on May 9^{th}, 2021 it reached its highest value in history being 58,788 USD. On May 18^{th}, 2021 (which corresponds to the time of writing of this section) it had fallen again to 43,144 USD losing 26.61% of its value in 9 days.

Similar trends apply to other cryptocurrencies. For example, the value of Ether was about 750 USD in December 2020, about 1,350 USD in January 2021, about 1,800 USD in March 2021 and about 2,700 USD in April 2021. On May 12^{th}, 2021 this value was equal to 4,132.76 USD and represents the highest value reached by this currency so far. On May 15^{th}, 2021 its value was still 4,100.03 USD. On May

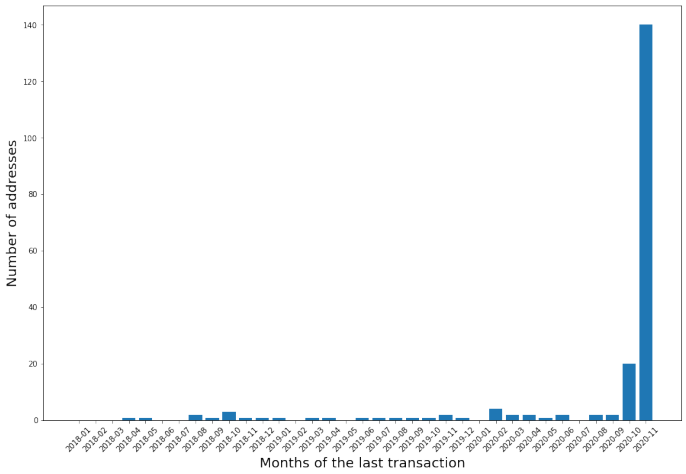


Fig. 9.22: Distribution of the Entrants (`to_addresses`) against the date of the last transaction

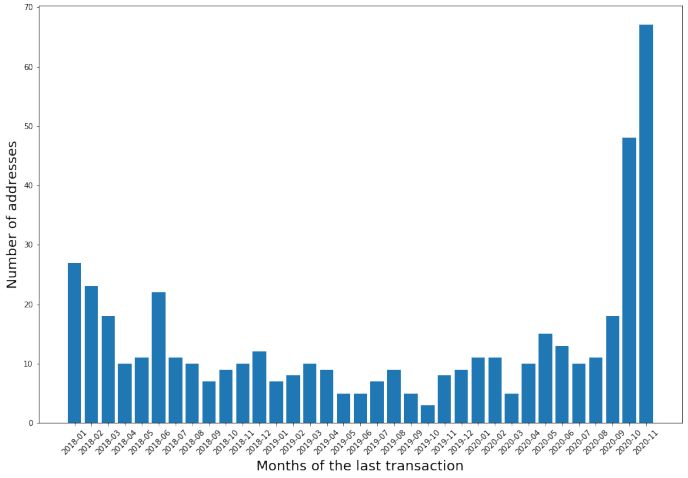


Fig. 9.23: Distribution of the Others (`to_addresses`) against the date of the last transaction

16^{th}, 2021 (which corresponds to the time of writing of this section) the value of the Ether was 3,231.94 USD with a collapse of 21.81% in 6 days.

The above examples highlight how prone the cryptocurrency world is to speculation. In addition, the trends of the last month lead us to believe that we are in the midst of a speculative bubble similar to the one of 2017-2018. If this is the case, the proposed approach would allow us to extract many knowledge patterns about the behaviors of the various players operating in this market and could even support analysts in understanding who are the speculators behind these bubbles. Therefore, we believe that the proposed approach has not only a value for the past but it provides useful predictive tools for the present and for the future.

Part IV

Further Areas

In this part, we apply our complex network-based model to further areas. Indeed, this model can deal with many scenarios in which the interactions between actors play a key role. Specifically, we show how our model can be applied to three further scenarios, namely: (i) *Innovation Management*, (ii) *Neurological Disorders, and* (iii) *Extraction of semantic relationships among concepts. This part is organized as follows: in Chapter 10, we illustrate the application of our model to the investigation of the patent citations and their influence in the innovation management. In Chapter 11, we describe its application to manage the ElectroEncephaloGram (i.e., EEG) signals of a patient in order to investigate neurological disorders, such as Alzheimer's Disease and Mild Cognitive Impairment. Finally, in Chapter 12, we apply our model for data lake management, which brings together network-based and semantics-driven representation of metadata.*

10

Innovation Management

The impressive development of innovations in all the R&D fields is making the adoption of big data centered-techniques compulsory for their analysis. Here, network analysis-based approaches are extremely promising. Centrality is one of the most investigated issues in network analysis and, in the past, several centrality measures have been proposed. However, none of them is tailored to the specificity of the patent citations scenario. In this chapter, we propose a well-tailored centrality measure for evaluating patents and their citations and experimentally prove that it is well-suited to capture the peculiarities of this domain. We also present three possible applications of our measure: the computation of the scope of a patent, the computation of the lifecycle of a patent, and the detection of the so-called power patents.

The material present in this chapter is taken from [236].

10.1 Introduction

Patents have been largely investigated in the past scientific literature [8, 442, 662, 245, 622, 403]. In fact, their analysis can supply a large amount of information concerning both the state of art and the protagonists of a certain Research & Development (R&D) field [684, 267, 305, 329, 340, 473, 613, 435]. This also because the submission of a patent is usually the first public claim of a new invention or innovation. Patent analysis allows decision makers to investigate the experiences of other (possible competitor) institutions and/or countries, in such a way as to know the past and the current R&D activities in the fields of interest, to delineate their evolution and to foresee their future developments. Furthermore, patent analysis allows the construction of a detailed picture of the R&D cooperations among different institutions and/or countries and can be an indicator of geo-political evolutions happening all over the world [202, 601, 122].

Most of the past approaches for patent analysis were based on classical statistics. However, the impressive development of innovations in all the R&D fields is leading

to a huge increase of patent data. Therefore, it is reasonable to foresee that, in the next future, Big Data centered techniques will be compulsory to fully exploit the potential of patent data. In this last scenario, the adoption of approaches based on network analysis is extremely promising [676, 197, 198, 415, 709, 149]. As a matter of facts, network analysis allows a full comprehension and a complete management of those phenomena where relationships among objects to investigate play the key role and, at the same time, the corresponding variables are strictly related to each other. This is exactly the future scenario characterizing patent and innovation management, and, at the same time, it is the "worst-case scenario" for classic statistic-based approaches, which present several limitations when operating therein [647].

As a confirmation of the adequacy of network analysis for patent investigation, in the past literature, several approaches to facing this issue can be found [149, 349, 247, 341, 695].

Centrality is one of the most investigated issues in network analysis, which aims at measuring the importance of a node in a network.

Several centrality measures have been proposed in the literature [181, 575, 281, 314, 280, 621, 133], but they are not tailored to this scenario and could return only approximate results. This because patents have a very relevant peculiarity that is not found elsewhere (for instance, in scientific papers [262]), in that, if a patent p_i cites a patent p_j, then p_i loses a part of its value.

If we report this reasoning to the network analysis context, we have that, for a node, having incoming arcs is extremely positive; by contrast, having outgoing arcs is negative. Past centrality measures certainly distinguish between these two kinds of arc; for instance, degree centrality distinguishes between indegree and outdegree [319]. However, they do not combine centrality values originated from the incoming arcs with those derived from the outgoing ones. We are missing a centrality measure that first assigns a positive ranking to incoming arcs and a negative ranking to outgoing ones and, then, combines these rankings to obtain a unique value.

In this chapter, we propose a well-tailored centrality measure for evaluating patents and their citations.

For this purpose, we preliminarily introduce a suitable support directed network, whose nodes represent patents. An arc from a node v_i to a node v_j indicates that the patent represented by v_i cited the patent represented by v_j.

After this, we introduce our centrality measures, namely "Naive Patent Degree" and "Refined Patent Degree", and we show that they are well tailored to capture the specificities of the patent scenario. To investigate the adequacy of our centrality measures, we carried out several experiments. The corresponding patent data derives from PATSTAT-ICRIOS database [199]. It stores patent data, from 1978 to the cur-

rent year, coming from about 90 patent offices worldwide, including, of course, the most important and largest ones, such as European Patent Office (EPO) and United States Patent and Trademark Office (USPTO).

Finally, we present three possible applications of our measures, namely: *(i)* the computation of the "scope" of a patent, whose purpose is the evaluation of the width and the strength of the influence of a patent on a given R&D field; *(ii)* the computation of the lifecycle of a patent; *(iii)* the detection of the so-called "power patents", i.e., the most relevant patents, and the investigation of the importance, for a patent, to be cited by a power patent.

The plan of this chapter is as follows: in Section 10.2, we present related literature. In Section 10.3, we define the support model and the theoretical definition of our new centrality measures. Then, in Section 10.4, we describe the patent database that we used for our experiments, the evaluation of our centrality measures, and three possible applications of them in the patent scenario.

10.2 Related Literature

Centrality has always been one of the core topics of network analysis and has been largely investigated in the literature. It allows people to quantify the importance of nodes in their network and to understand the structural properties of this last one. As a matter of facts, already [545] developed a self-consistent methodology for determining citation-based influence measures for scientific journals, subfields and fields. Specifically, these authors formulate an eigenvalue problem leading to a size-independent influence weight for each journal or aggregate. Then, they define two other measures, namely the influence per publication and the total influence. Finally, they present some hierarchical influence diagrams and numerical data to display inter-relationships for journals on physics. In the same years, [281] examined and explained the role of centrality metrics in network analysis.

As illustrated in detail in [438, 210], the influence of a node mainly depends on its position in the corresponding network, as well as on the structural properties of this last one. Centrality metrics aim at assigning a rank to each network node, summarizing its importance in the network. As previously pointed out, this rank is strictly related to the needs of the application scenario, which the network refers to. Since these needs can be heterogeneous, several different metrics have been proposed in the past network analysis literature.

The study of the neighborhood of a node is adopted as the starting point of some of the most important centrality metrics. In this context, degree centrality is one of the most famous metrics; it aims at measuring the visibility of a node within its net-

work. Degree centrality presents several strengths but also some weaknesses. This is the reason why, in the literature, researchers proposed some approaches that try to overcome the problems of this metric. An example is ClusterRank, proposed in [181]; it also considers clustering coefficient in the score computation. In [237], the authors, starting from the observation that the position of a node is more important than its degree for measuring its relevance, apply k-core decomposition. It iteratively breaks down the network according to the residual degree of its nodes. K-core decomposition is considered as one of the most valid approaches to understanding the influence of a node and its role in information diffusion. Another well known centrality measure is h-index [330], which returns the influence of a user in a social network.

Another family of centrality approaches is based on the number of paths, which a node is involved in. In this path-based centrality, the higher the number of paths where a certain node is present the higher the node's importance. Closeness centrality [575], eccentricity centrality [314] and betweenness centrality [280] belong to this family of approaches. From a general point of view, a node with a high closeness centrality can have access to a high number of communications; therefore, it can perform a high control on information flow. Instead, a node with a high betweenness centrality, in most cases, operates as a bridge between two communities; therefore, it can have a strong control on information exchange. Other techniques belonging to this family of centrality metrics are Kats centrality [370], subgraph centrality [253], and information index [621].

As pointed out in [683], in most cases, centrality does not depend only on the number of neighbors of a node on the paths it is involved in. In some cases, not only the number of neighbors, but also their relevance is important to assess the relevance of a node in its network. Starting from this consideration, authors have defined a third family of centrality measures. Eigenvector centrality [117], PageRank [133] and HITs [384] are the most known metrics of this family.

Even if centrality is one of the most important topics in network analysis, it was rarely adopted for investigating the relevance of a patent based on citations. Actually, the idea of analyzing patents based on their citations was proposed by Seidel in 1949 [594]. From that time, a large variety of tools for performing this analysis has been proposed in the literature. Network analysis is one of the most adopted tools because it allows the creation of suitable networks representing patent citations.

Bibliometrics is certainly an optimum starting point for patent investigation, as it shares many common aspects with patent analysis. Clearly, besides many similarities, paper and patent citations also present several significant differences, as evidenced in [472].

If we focus on patent citations, several variegate approaches to investigating patents based on them have been proposed in the past. For instance, the authors of [695] consider both direct and indirect citations, as well as patent couplings co-citations. An approach to investigating patent outliers is described in [566], whereas the small world phenomenon in the context of patent citation networks is analyzed in [349]. The definition of the lifecycle of a given technology starting from patent citation networks is proposed in [345], whereas the technological focus of patents is studied in [344].

In several cases, the typical problems of network analysis are investigated in the context of patent citation networks. For instance, the approach to analyzing network connectivity proposed in [346] is extended to patent citation networks in [84, 268, 657]. Specifically, [84] shows how the analysis of network connectivity can be extended to the patent scenario for detecting reliable knowledge on technological evolutions. [268] exploits network connectivity to reconstruct the most relevant technological trajectories of data communication standards. [657] performs a similar investigation but for fuel cells technology.

Finally, the application of the standard centrality metrics to patent citation networks has been proposed in very few cases. For instance, the authors of [149] propose an approach to determining the relevance of companies in the industry they operate on, based on the application of classic centrality metrics on the citation networks of the patents published by them. An analogous effort can be found in [177], but for Intelligent Transportation System companies. The authors of [412] apply degree centrality, betweenness centrality and closeness centrality on patent citation networks to investigate several mechanisms underlying technological innovations. Finally, in [251, 463], the authors carefully examine the usage of PageRank in patent citation networks, and evidence its strengths and weaknesses.

However, to the best of our knowledge, none of the approaches proposing the application of centrality measures to patent citation networks considers the main peculiarity of this scenario, i.e., that, if a patent p_i cites a patent p_j, then the value of p_i decreases. By contrast, this important feature represents the core of our approach.

10.3 Methods

10.3.1 Definition of a support model

In this section, we introduce the data model representing data about patents and used by our approach. Before illustrating it, we must introduce two sets allowing us to formalize data at our disposal. These are: *(i)* the set *Pat* of all the patents stored

in a dataset, and *(ii)* the set Pat_k of the patents filed by at least one inventor of the country k.

We are now able to present our data model. It consists of a network $N = \langle V, A \rangle$. V denotes the set of the nodes (or vertices) of N. A node $v_i \in V$ corresponds exactly to a patent $p_i \in Pat$. Since there is a biunivocal correspondence between a node of V and the corresponding patent of Pat, in the following, in some cases, we adopt the symbol v_i to represent both of them and we adopt the terms "patent" and "node" interchangeably. Each node $v_i \in V$ has an associated label l_i, denoting the set of the countries of the inventors of p_i. A is the set of the arcs of N. There exists an arc $a_{ij} = (v_i, v_j) \in A$ if p_i cites p_j. Clearly, N is a directed network.

Starting from N, we can define some sets representing the neighborhoods of a node in V. In particular, given a node $v_i \in V$, we can define the following neighborhoods:

- $Cited_i$, i.e., the set of the patents cited by p_i:

  $$Cited_i = \{v_j | (v_i, v_j) \in A, v_j \neq v_i\}$$

  In other words, $Cited_i$ is the set of the nodes (and, therefore, the set of the patents) v_j such that there exists an arc from v_i to v_j (which implies that v_j was cited by v_i) in the set A of the arcs of N.
- $Citing_i$, i.e., the set of the patents citing p_i:

  $$Citing_i = \{v_j | (v_j, v_i) \in A, v_j \neq v_i\}$$

  In other words, $Citing_i$ is the set of the nodes (and, therefore, the set of the patents) v_j such that there is an arc from v_j to v_i (which implies that v_j cited v_i) in the set A of the arcs of N.
- V_k, i.e., the set of the nodes associated with the patents of Pat_k:

  $$V_k = \{v_i | v_i \in V, k \in l_i\}$$

  or, analogously:

  $$V_k = \{v_i | v_i \in V, p_i \in Pat_k\}$$

  In other words, V_k is the set of the nodes of N having the country k among the ones forming its label l. This is equivalent to say that V_k is the set of the patents having at least one inventor of the country k.

10.3.2 Definition of a new centrality measure

Patent citations have a very important specificity because, if a patent p_i cites a patent p_j, the value of p_i decreases. As a consequence, differently from many other contexts, such as scientific papers, making a citation is not painless for the citing patent.

If we report this reasoning to our model, it implies that having incoming arcs is extremely positive for a node (and this is in line with the classic centrality metrics of network analysis). By contrast, having outgoing arcs is penalizing for a node (and this fact is not captured by classic centrality measures).

Since our support network is a directed one, it is necessary to define both the indegree and the outdegree of a node. The former indicates the number of its incoming arcs (i.e., the number of citations received by the corresponding patent), whereas the latter denotes the number of its outgoing arcs (i.e., the number of citations performed by the corresponding patent).

We propose two centrality measures, which we call:

- Naive Patent Degree (NPD);
- Refined Patent Degree (RPD).

We start by analyzing Naive Patent Degree. Given a node $v_i \in V$, the corresponding Naive Patent Degree NPD_i is defined as:

$$NPD_i = |Citing_i| - |Cited_i|$$

Clearly, this definition is immediate and captures the specificity mentioned above. However, we tried to find a more rigorous centrality metric, capable of capturing the synergies characterizing the patent scenario. Refined Patent Degree is the result of this effort. Its definition is based on the following considerations:

- C_1: given a patent p_i, the higher its capability of being cited by patents making very few citations, the higher its importance.
- C_2: given a patent p_i, the higher its capability of being cited by important patents, the higher, in turn, its importance. Observe that, in principle, Condition C_2 is very complex because it implies that the RPD of a node n_i depends on the RPD of a node n_j. This implies that, for the computation of this metric, complex systems characterized by hundreds, or even thousands, of equations and variables should be solved, at least in the most complex cases. As a consequence, the computation of RPD appears difficult to handle without a heuristic. A reasonable one could consider the NPD of n_j, instead of the RPD of this node, in the computation of the RPD of n_i.
- C_3: the weight of a citation of a patent p_j, which a patent p_i must make, is inversely proportional to the number of citations received by p_j. In other words, if p_j is a very important patent, which received a very high number of citations, the fact that p_i must cite p_j does not considerably decrease the innovativity of p_i. By contrast, if p_i must cite a little cited patent p_j, it is possible to conclude that it is strongly influenced by p_j, and this significantly undermines its innovativity.

Taking all these conditions into account, RPD_i can be defined as:

$$RPD_i = \sum_{j=1}^{|Citing_i|} \omega_j - \sum_{q=1}^{|Cited_i|} \frac{1}{1+|Citing_q|}$$

where:

$$\omega_j = \alpha\left(\frac{1}{1+|Cited_j|}\right) + (1-\alpha)\left(\frac{NPD_j}{NPD_{max}}\right)$$

Here, $|Citing_i|$ (resp., $|Cited_i|$) is the cardinality of the set $Citing_i$ (resp., $Cited_i$). ω_j is a weighted mean of two terms. The former expresses Condition C_1, whereas the latter represents Condition C_2. The weight α allows the tuning of the mutual relevance of these two terms. In our case, we chose to assign the same importance to them; as a consequence, we set α equal to 0.5. Finally, the second term of the formula for RPD_i allows the formalization of Condition C_3.

As it will be clear in the next subsection, RPD does not overturn NPD. It simply refines this last metric, thanks to the three conditions, which it is based on. Specifically, it can produce acceptable distributions also for those countries having a low number of patents associated with them. This is exactly the scenario where NPD shows its main weaknesses.

10.4 Results

Our new patent centrality measures can have several applications. In this Section, we firstly present the reference dataset extracted from the PATSTAT-ICRIOS database. Then, we evaluate our centrality measures on it, and finally we describe three possible applications, namely: *(i)* the computation of the "scope" of a patent; *(ii)* the definition of the lifecycle of a patent; *(iii)* the detection of "power patents".

10.4.1 Patent Database

Data regarding patents adopted in our analyses has been taken from PATSTAT-ICRIOS database [199]. This is a large database about patents handled by ICRIOS Center at Bocconi University.

PATSTAT (i.e., EPO worldwide PATent STATistical database) is a database storing raw data about patents. It was constructed by EPO in cooperation with the World Intellectual Property Organization (WIPO), OECD and Eurostat. It is currently managed by EPO. It stores data about all patents, from 1978 to the current year, coming from about 90 patent offices worldwide, comprising the most relevant ones, such as EPO and USPTO.

As pointed out above, data is registered in PATSTAT in a raw format. To facilitate its analysis, ICRIOS processed it and produced a cleaned and harmonized database, i.e., PATSTAT-ICRIOS. This includes all bibliographic variables concerning each patent application. In particular, it stores application number and date, publication number and date, priority, title and abstract, application status, designed states for protection, main and secondary International Patent Classification (IPC) codes, name and address of both the applicant and the inventor, references (i.e., citations) to prior-art patent and non-patent literature, the corresponding Nomenclature of Units for Territorial Statistics (NUTS3) and, finally, File Index concordance tables, allowing the conversion of IPC codes into more aggregated and manageable technological classes.

To perform our investigation in the most correct and effective way, we carried out a pre-processing activity on the data of our interest. For this purpose, we used the framework R [3]. Our pre-processing activity consisted of the following tasks:

- *Data Extraction*. During this task, we first identified all the tables of PATSTAT-ICRIOS necessary for our analyses. To increase the effectiveness of the next tasks, we removed all the unnecessary and redundant attributes from these tables. This led to a strong reduction of the size of the data to process.
- *Data Normalization*. During this task, we removed some inhomogeneities regarding the data types of some fields (i.e., strings and dates).
- *Data Aggregation*. During this task, we performed a data integration activity aiming at storing all data about a concept in a unique collection.
- *Data Loading*. During this task, we loaded available data (represented in the CSV format) into a MongoDB [2] final database, which we used for our next activities.

At the end of these four tasks, the size of the dataset to analyze was reduced from 12.5 GB to 2.5 GB.

10.4.2 Centrality measures evaluation

We started the evaluation of our metrics by computing the distribution of NPD for many world countries. Obtained results show that, for most countries, the distribution of NPD follows a power law. However, this power law is very singular and completely different from the ones generally characterizing degree distribution in network analysis.

In order to give an idea of the peculiarities of the distribution of NPD, in Figure 10.1, we show its values for Italy. From the analysis of this figure, we can see that, actually, there are two power law distributions almost mirrored with respect to the zero value of NPD.

Another interesting phenomenon, which can be observed in this figure, regards the two tails of the power law distributions. In fact, the right tail is much longer than the left one. This means that the number of citations received by Italian patents is much higher than the number of citations made by them. Furthermore, if we consider the shape of the tails, we can observe that the right tail is much steeper than the left one. This means that the distribution of citations received by Italian patents follows a more pronounced power law than the distribution of citations made by them. Finally, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.55.

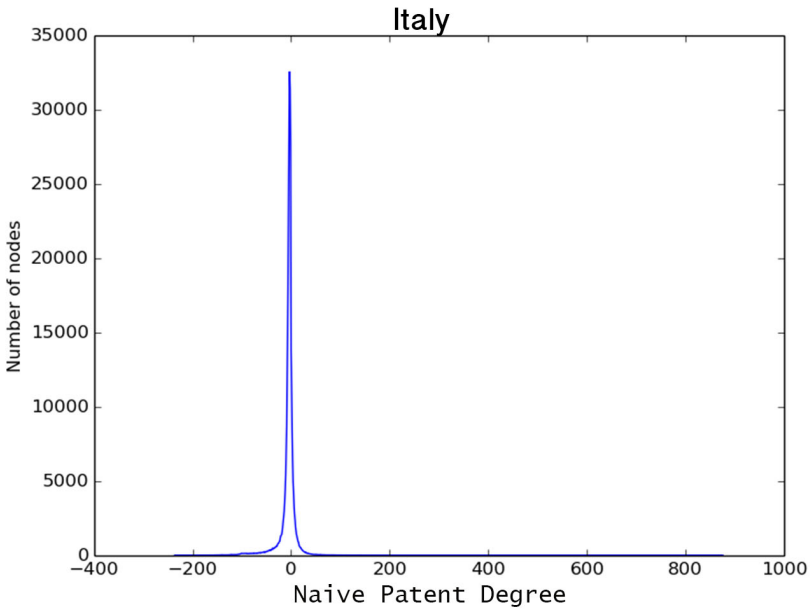


Fig. 10.1: Distribution of the values of NPD for Italy

As previously pointed out, the same trend (with the same specificities) can be observed for most countries.

For some countries, the distribution of NPD is similar to the one of Italy, even if much more disturbed than it. An example of this trend is shown in Figure 10.2, where we report the case of Estonia. A first result emerging from the comparison of this figure with Figure 10.1 is that the number of patents of Estonia is much lower than the one of Italy. Furthermore, we can note that, in this case, the trend of NPD values differs from the optimal one. This fact is more evident in the left power law distribution. Here, it is possible to observe some peaks that evidence the presence of a considerable number of Estonian patents that make many citations, especially if we compare their number with the total number of Estonian patents. As a further result, we observe that the length of the right and the left tails are comparable.

However, also in this case, the right tail is steeper than the left one. All the previous observations are valid for all the countries with such a kind of trend for NPD. In this case, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 1.05.

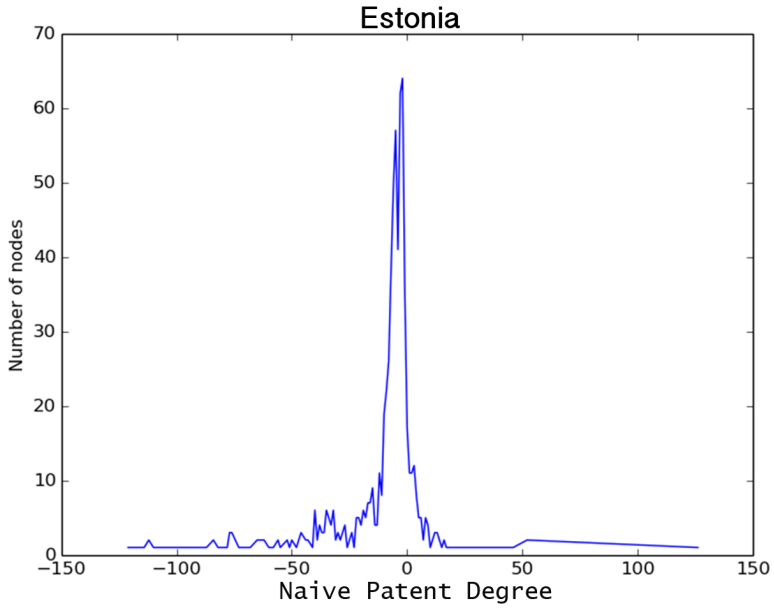


Fig. 10.2: Distribution of the values of NPD for Estonia

For some countries, the distribution of NPD does not follow a power law. As an example of this situation consider Figure 10.3, where we report the distribution of NPD for Tunisia. In this figure, we can also observe that the left tail is longer than the right one and that the number of Tunisian patents is very low. Even if this case is not very significant from a statistic point of view, we can again observe that the right "tail" is "steeper" than the left one. Furthermore, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 2.64. This also happens for the other countries with an analogous distribution of NPD.

The comparison of the results obtained for the three kinds of country mentioned above suggests that the most innovative and rich countries present a power law distribution for NPD. Furthermore, since these countries drive the innovation and the technological progress of the other ones, their patents receive many more citations than the ones they must make.

Those countries, like Estonia, showing a disturbed power law for NPD do not have a patent patrimony allowing them to be innovation leaders currently. However,

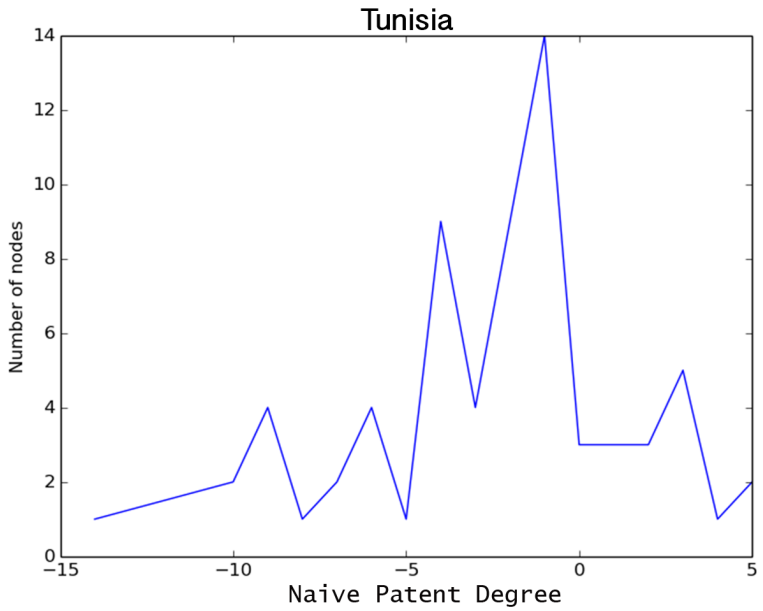


Fig. 10.3: Distribution of the values of NPD for Tunisia

they are accumulating a certain number of patents allowing them to become innovation leaders in the near future.

Finally, those countries, like Tunisia, having an irregular distribution of NPD are characterized by a very low number of patents. They have not reached an adequate research and innovation level yet. Their very limited number of patents does not allow a detailed analysis about their situation.

After having evaluated NPD, we proceed to investigate RPD. We start with the most innovative countries. In Figure 10.4, we report the distribution of the values of RPD for Italy on the left, and a zoomed representation of the same distribution around the zero value of RPD on the right. If we compare the distribution of RPD with the one of NPD, reported in Figure 10.1, we can observe that RPD confirms (or, even better, magnifies) all the results returned by NPD. The only exception regards the steepness of the two tails. In fact, differently from NPD, in this case, the left tail is steeper than the right one. Finally, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.14.

In Figure 10.5, we report the distribution of the values of RPD for Estonia, as a representative of the countries with an intermediate number of patents. If we compare this distribution with the corresponding one of NPD for the same country, we can observe that RPD removes many of the disturbances observed in NPD. Therefore, the corresponding distribution is much "cleaner". Differently from what happens in Figure 10.4, and analogously to the trend shown in Figure 10.2, we have that,

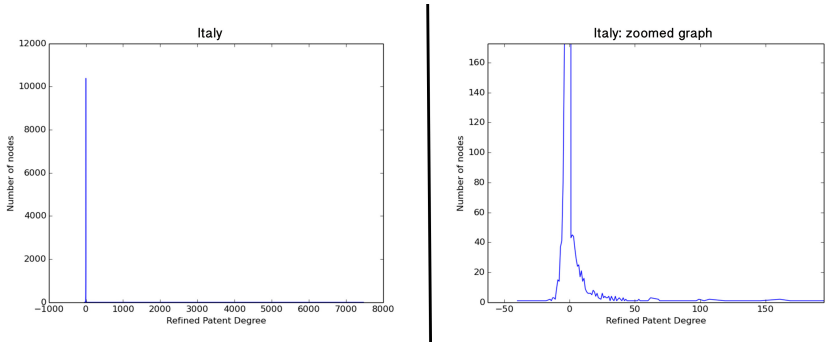


Fig. 10.4: Distribution of the values of RPD for Italy

in this case, the right tail is steeper than the left one. In this case, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.20.

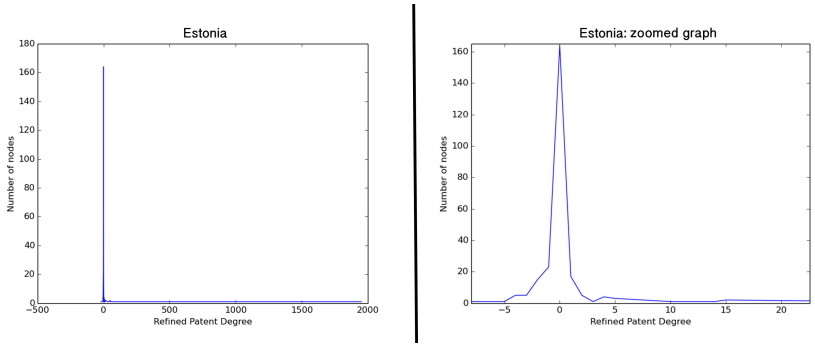


Fig. 10.5: Distribution of the values of RPD for Estonia

An analogous reasoning can be drawn for those countries having a low number of patents. If we compare the distribution of RPD for Tunisia, shown in Figure 10.6, with the corresponding one of NPD, shown in Figure 10.3, we can see that the RPD's capability of cleaning the distortions of NPD is even magnified for countries with a small number of patents. In this case, the steepness of the left tail is slightly higher than the one of the right tail, even if the differences are not remarkable. Furthermore, the ratio between the area formed by the curve of NPD and the axis of the abscissae to the left and the right of NPD=0 is equal to 0.33.

In conclusion, both NPD and RPD appear well suited as centrality measures for patents. However, RPD is capable of removing some distortions that have been shown by NPD when this last is adopted for evaluating countries with a small number of patents.

To make our analysis about NPD and RPD more exhaustive, we computed the "similarity rate" of the results returned by NPD and RPD. For this purpose, given a country k, we computed the set Top_k^{NPD} (resp., Top_k^{RPD}) of the top 5% of the patents

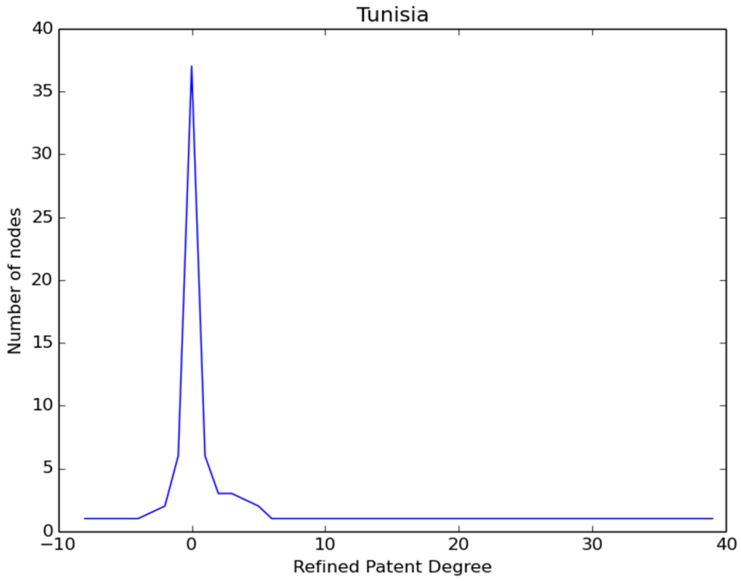


Fig. 10.6: Distribution of the values of RPD for Tunisia

of Pat_k with the highest values of NPD (resp., RPD). Then, we computed the parameter:

$$rTop_k = \frac{|Top_k^{NPD} \cap Top_k^{RPD}|}{|Top_k^{NPD}|}$$

The possible values of $rTop_k$ range between 0 and 1, where 0 denotes that NPD and RPD return completely different results, whereas 1 indicates that they have exactly the same behavior.

We computed the value of $rTop_k$ for the world countries and, in Table 10.1, we report some of them. From the analysis of this table, we can observe that the value of $rTop_k$ is generally much higher than 0.5. Its average value for all world countries is 0.65. This result, along with the previous ones specified above, allows us to conclude that RPD does not overturn NPD. Actually, the former refines the latter thanks to the three conditions, which it is based on. RPD can return acceptable and clean distributions also for those countries having a low number of patents, in which case NPD is excessively sensitive to disturbances.

10.4.3 Computation of the scope of a patent

We use the term "scope" to indicate the width and the strength of the influence of a patent $p_i \in Pat$ on the other patents, that is the width and the strength of the influence of a node $v_i \in V$ on the other nodes of N. We argue that the scope of v_i is strictly connected to the number and the centrality of the nodes citing it, either directly or

Country	$rTop_k$
Algeria	1.00
Austria	0.86
Brazil	0.62
Bulgaria	0.68
China	0.56
South Korea	0.62
Denmark	0.59
Estonia	0.77
Finland	0.52
France	0.57
Germany	0.65
Japan	0.73
Greece	0.50
India	0.61
Italy	0.59
Luxembourg	1.00
Poland	0.63
United Kingdom	0.59
Romania	0.67
Russia	0.59
Spain	0.48
South Africa	0.57
Taiwan	0.60
Tunisia	0.67

Table 10.1: Similarity Rate of NPD and RPD for some countries

indirectly. As a consequence, in the scope definition, the main roles are played by the centrality measure, which we have already seen, and by the neighborhood of a node, which we introduce now.

With regard to this last concept, we point out that there could exist several levels of neighborhood of a node v_i. For this reason, it is possible to introduce the neighborhood of level t of a node $v_i \in V$. This is defined as follows:

$$nbh_i^t = \begin{cases} Citing_i & if\ t = 0 \\ \{v_j | (v_j, v_l) \in A, v_l \in nbh_i^{t-1}\} & if\ t > 0 \end{cases}$$

We are now able to define the Naive Scope NS_i^t and the Refined Scope RS_i^t of a node $v_i \in V$ w.r.t. the nodes of its t^{th} neighborhood nbh_i^t as follows:

$$NS_i^t = \sum_{j \in nbh_i^t} NPD_j \qquad RS_i^t = \sum_{j \in nbh_i^t} RPD_j$$

Once the scope of a node has been defined, it is possible to perform an investigation at the country level to analyze the average trend of the scope of the nodes of a country k. In particular, the Average Naive Scope ANS_k^t and the Average Refined Scope ARS_k^t of the patents of a country k with respect to their t^{th}-level neighbors can be defined as:

$$ANS_k^t = \frac{\sum_{v_i \in V_k} NS_i^t}{|V_k|} \qquad ARS_k^t = \frac{\sum_{v_i \in V_k} RS_i^t}{|V_k|}$$

We computed the trends of ANS_k^t and ARS_k^t for most world countries. As an example, in Figures 10.7 - 10.9, we show the trend of ANS_k^t (in blue) and ARS_k^t (in red) for three countries, namely China, Luxembourg and Poland. Analogous trends have been found for the other countries. From the analysis of Figures 10.7 - 10.9, we can observe that, for all cases, the average scope decreases when the neighborhood level increases. This general result was expected. However, the really interesting analysis concerns *how fast* this decrease is. As for this issue, we generally observe a steep decrease so that, after the third-level neighborhoods, patent scopes are almost null. If we compare the trends of ANS_k^t and ARS_k^t in these figures, we can observe that they are similar, even if the trends of ARS_k^t are always steeper than the ones of ANS_k^t. This is in line with the results of the comparison of NPD and RPD presented in Section 10.4.2, where we have seen that RPD refines and magnifies the trends characterizing NPD.

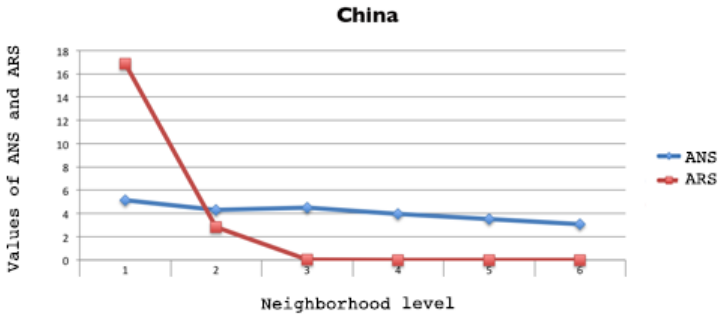


Fig. 10.7: Trend of ANS_k^t and ARS_k^t against the neighborhood level t for China

10.4.4 Computation of the lifecycle of a patent

This activity aims at verifying if, by computing, year by year, the NPD and the RPD of patents published all over the world, it is possible to determine one or more characteristic patterns. In the affirmative case, each characteristic pattern would represent a lifecycle template for the patents following it. Defining lifecycle templates for specific categories of patents is extremely useful because, given a new patent p_i belonging to a category for which there exists a lifecycle template, it is possible to foresee the NPD and the RPD of p_i over time, and, ultimately, the number and the relevance of the citations received by it.

In order to show how lifecycle templates could be defined, in the following, we associate categories with years and introduce a category per year. However, we could

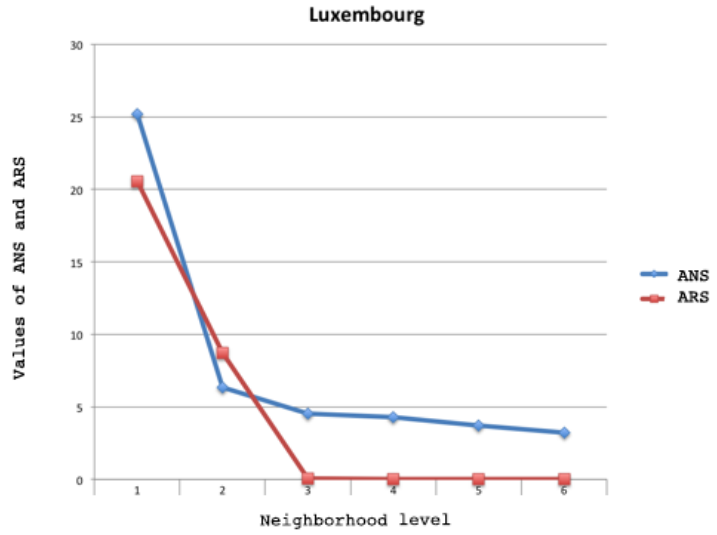


Fig. 10.8: Trend of ANS_k^t and ARS_k^t against the neighborhood level t for Luxembourg

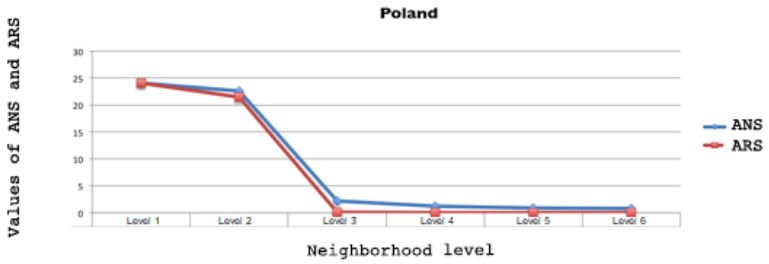


Fig. 10.9: Trend of ANS_k^t and ARS_k^t against the neighborhood level t for Poland

adopt the same technique with a completely different taxonomy, for instance by associating a category per IPC class (in such a way as to define a patent lifecycle template for each IPC class), a category per country, and so forth.

To construct a lifecycle template for each year, we must preliminarily introduce the measures NPD_i^y and RPD_i^y. These two measures are analogous to NPD_i and RPD_i, except that they consider only the patents published in the year y.

To carry out our analysis, for each year from 1985 to 2013, we considered all the patents published in that year and, for each of them, we computed the values of NPD and RPD from that year until 2013. For instance, in Figure 10.10 (resp., 10.11, 10.12 and 10.13), we show the trends of RPD for the patents published in the year 1985 (resp., 1990, 1995 and 2000). By analyzing the obtained results we have seen that, independently of the publication year of patents, there exists a unique pattern representing the patent lifecycle.

We aimed at expressing this lifecycle template mathematically and we observed that it can be represented by a sixth-degree polynomial function of the form:

$$y = ax^6 + bx^5 + cx^4 + dx^3 + ex^2 + fx + e$$

To give a visual intuition of this fact, in Figures 10.10 – 10.13, we traced, along with the real values of patent lifecycle, the sixth-degree polynomial function that best approximates it. It is possible to observe that the deviations between the real values and the ones of the polynomial function are very small.

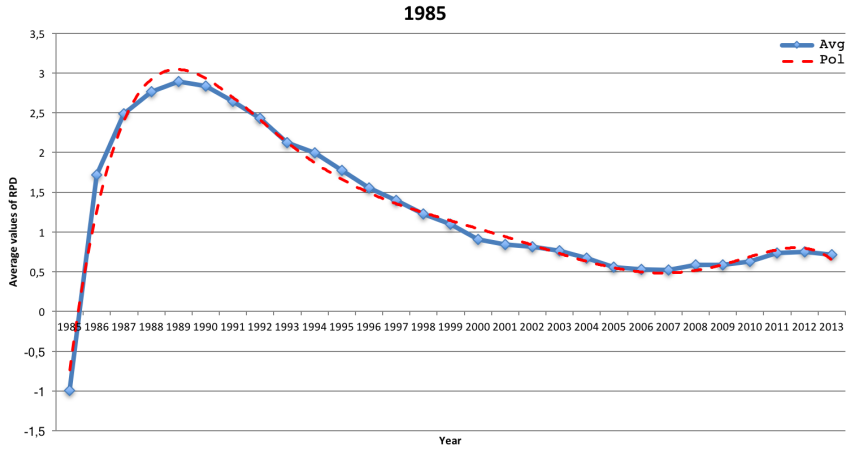


Fig. 10.10: Average values of RPD over time for the patents published in 1985

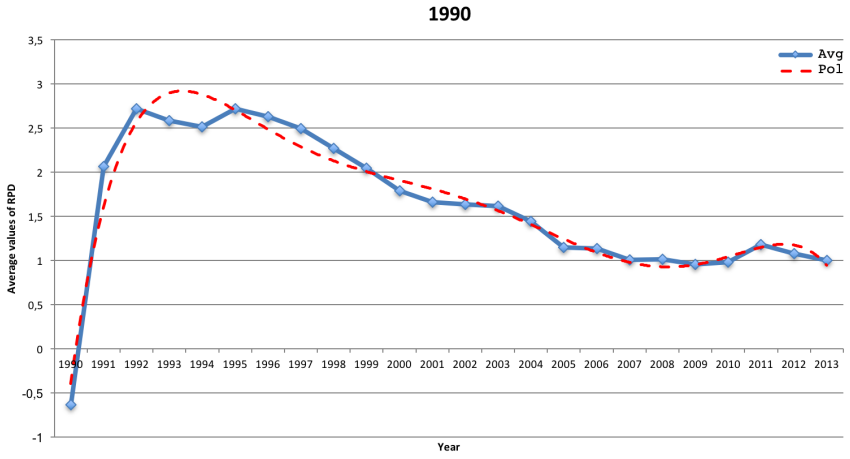


Fig. 10.11: Average values of RPD over time for the patents published in 1990

By analyzing each figure, we can observe that RPD is negative in the publication year of patents. This is due to the fact that all the citations performed by a given patent p_i are concentrated in its publication year, whereas, in that year, no patents, or a little number of them, cite p_i. After the first year from the publication of p_i, the corresponding RPD starts to increase. This increase reaches a maximum after about 5 years from publication. Then, a stall phase can be observed until to about

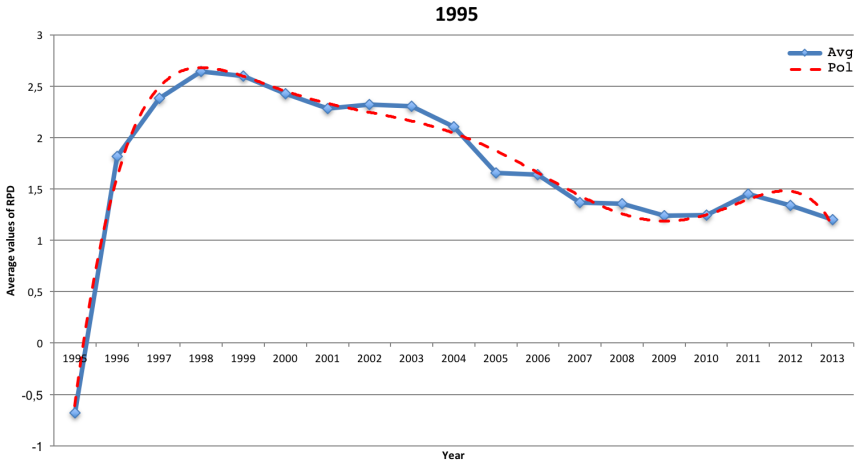


Fig. 10.12: Average values of RPD over time for the patents published in 1995

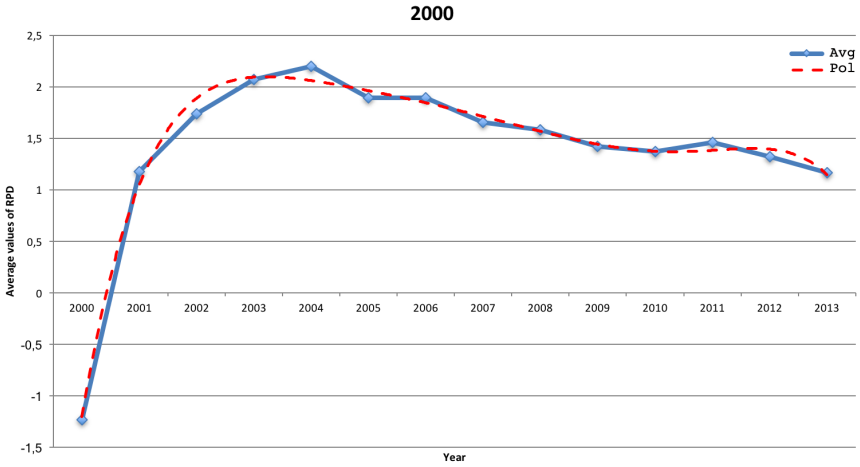


Fig. 10.13: Average values of RPD over time for the patents published in 2000

the eighth year; this phase is followed by a phase of decline, which becomes stronger and stronger until the RPD of p_i reaches an almost null value. This decline can be easily explained by considering that, for most patents, after about ten years from their publication, new technologies and/or more innovative patents appear, which make them obsolete.

In Table 10.2, we report the values of the coefficients of the sixth-degree polynomial function that represents the lifecycle templates regarding patents published in the years 1985-2000, obtained by applying the least square method. The coefficients of the lifecycles regarding patents published after 2000 are not reported because these lifecycles are too recent and, consequently, they are not complete yet.

Very similar trends and conclusions can be derived for NPD.

Years	*a*	*b*	*c*	*d*	*e*	*f*	*g*
1985	-1E-06	1E-04	-0,0039	0,0778	-0,8166	3,9637	-3,9546
1986	-1E-06	0,0001	-0,0046	0,0890	-0,8942	4,1551	-3,9498
1987	-2E-06	0,0002	-0,0056	0,1033	-0,9902	4,4030	-4,0791
1988	-2E-06	0,0002	-0,0066	0,1171	-1,0779	4,6154	-4,0942
1989	-3E-06	0,0002	-0,0078	0,1312	-1,1494	4,7282	-4,1012
1990	-3E-06	0,0003	-0,0084	0,1350	-1,1406	4,5921	-3,9704
1991	-4E-06	0,0004	-0,0113	0,1668	-1,3066	4,9941	-4,4076
1992	-5E-06	0,0005	-0,0118	0,1768	-1,4087	5,2030	-4,7034
1993	-8E-06	0,0006	-0,0154	0,2149	-1,5778	5,6661	-4,9479
1994	-1E-05	0,0008	-0,0198	0,2619	-1,8236	6,2006	-5,1879
1995	-1E-05	0,0009	-0,0225	0,2841	-1,8956	6,2383	-5,2146
1996	-2E-05	0,0011	-0,0260	0,3124	-1,9979	6,3822	-5,4142
1997	-2E-05	0,0014	-0,0305	0,3474	-2,1273	6,5878	-5,6143
1998	-3E-05	0,0014	-0,0306	0,3380	-2,0453	6,3775	-5,5960
1999	-3E-05	0,0016	-0,0341	0,3659	-2,1663	6,6400	-5,8417
2000	-4E-05	0,0020	-0,0393	0,4066	-2,3270	6,9163	-6,1626

Table 10.2: Values of the coefficients of the sixth-degree polynomial function that best approximates the lifecycles of patents published from 1985 to 2000

10.4.5 Definition of power patents and investigation of their importance

The definition of patent-tailored centrality measures like ours allows the identification of the most relevant patents. As a matter of fact, since both NPD and RPD follow a power law, it is reasonable to assume that there exist some *power patents*, i.e., a very small number of patents that have been cited very much. In order to investigate this aspect, in the following, we will consider RPD, even if analogous reasonings can be made for NPD.

Clearly, in principle, the fraction of power patents could differ for each country because it depends on the trend of the corresponding distribution of the RPD values. However, thanks to the features of RPD illustrated in Section 10.3.2, if we choose to select as power patents those ones whose values of RPD lie at the right of the elbow of the RPD distribution function, we obtain that, for most countries, it is sufficient to take as power patents the top 5% of patents having the highest values of RPD. To give an idea of this reasoning, in Figures 10.14, 10.15 and 10.16, we show three examples concerning the RPD value distribution of India, France and Japan. In all the three cases, it is evident that taking as power patents the top 5% of patents is sufficient. Analogous trends can be found for almost all the other world countries. In the following, we indicate with $\overline{Pat_k}$ the power patents of the country k.

After having defined a way to detect the power patents of each country, we aimed at investigating if, for a patent p_j, being cited by a power patent p_i can bring benefits, i.e., citations performed by patents that, having cited p_i, must also cite p_j.

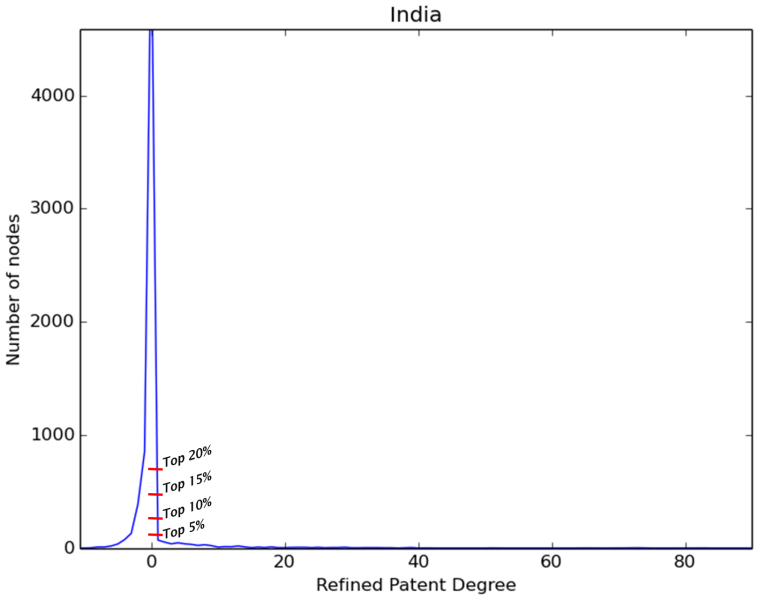


Fig. 10.14: Distribution of the values of RPD for India, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

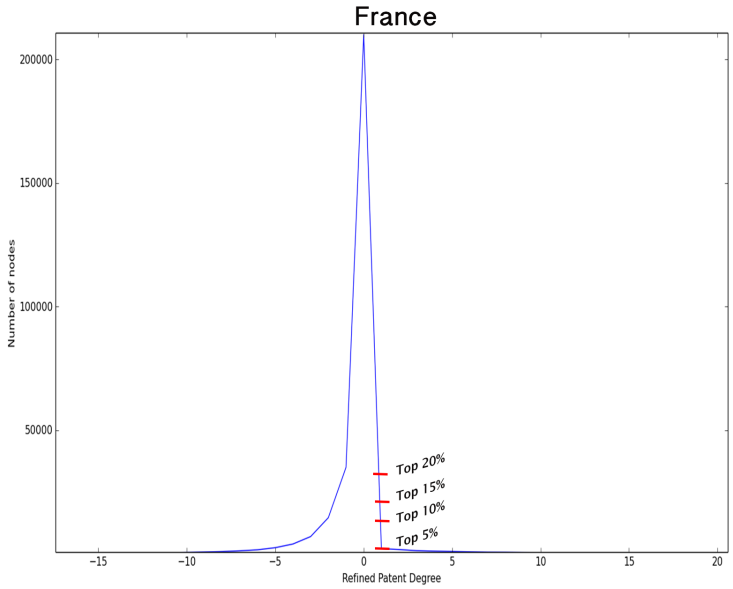


Fig. 10.15: Distribution of the values of RPD for France, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

To answer this question, we must preliminarily introduce some parameters. In particular, let $p_i \in Pat_k$ be a patent of the country k:

- The set of potential beneficiaries PB_i of p_i is defined as:

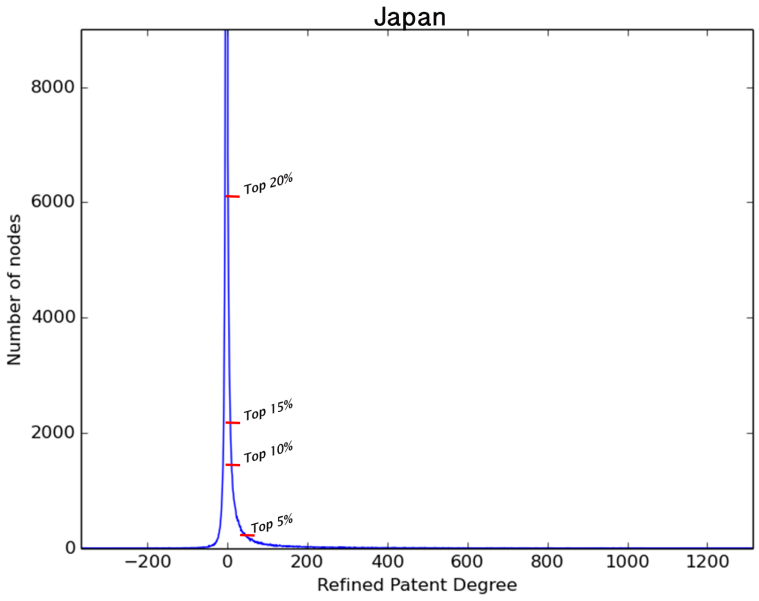


Fig. 10.16: Distribution of the values of RPD for Japan, along with the levels corresponding to the top 5%, 10%, 15% and 20% of patents with the highest values

$$PB_i = \{p_j | p_j \in Cited_i, p_i \in Cited_r, p_j \in Cited_r\}$$

- The fraction of potential beneficiaries of p_i is defined as:

$$F_i^{PB} = \frac{|PB_i|}{|Cited_i|}$$

- The average fraction of the potential beneficiaries of the patents of a country k is defined as:

$$AvgF_k^{PB} = \frac{\sum_{p_i \in Pat_k} F_i^{PB}}{|Pat_k|}$$

- The average fraction of the potential beneficiaries of the power patents of a country k is defined as:

$$\overline{AvgF_k^{PB}} = \frac{\sum_{p_i \in \overline{Pat_k}} F_i^{PB}}{|\overline{Pat_k}|}$$

We are now able to define the benefit capability bc_k of the power patents of a country k. Specifically:

$$bc_k = \frac{\overline{AvgF_k^{PB}}}{AvgF_k^{PB}}$$

The value of bc_k ranges between 0 and $+\infty$. If $bc_k \leq 1$, the power patents of k do not provide benefits to the patents cited by them. By contrast, if $bc_k > 1$, they are beneficial for the patents cited by them, and the higher bc_k the greater these benefits.

In Table 10.3, we report the value of bc for several countries. From the analysis of this table, we can see that bc is generally much higher than 1. This clearly evidences that, for a patent, obtaining a citation from a power patent is highly beneficial.

Country	*bc*
Austria	10.73
Brazil	0.47
China	13.30
South Korea	17.23
Denmark	6.58
Finland	7.93
France	10.37
Germany	9.72
Japan	5.19
Greece	1.47
India	21.63
Italy	10.11
Poland	6.32
United Kingdom	4.98
Romania	12.46
Russia	21.23
Spain	12.36
South Africa	6.24
Taiwan	17.73

Table 10.3: Values of *bc* for several countries

11

Neurological Disorders

In this chapter, we propose a complex network analysis-based approach to help experts in their investigations of patients with Mild Cognitive Impairment (i.e., MCI) and Alzheimer's Disease (i.e., AD) and investigate the evolution of these neurological diseases over time. The inputs of our approach are the ElectroEncephaloGrams (i.e., EEGs) of the patients, performed at a certain time and, again, three months later. Given an EEG, our approach constructs a complex network with nodes that represent the electrodes and edges that denote connections between electrodes. Then, we apply several network-based techniques for the investigation of subjects with MCI and AD and the analysis of their evolution over time. Our main results are: (i) a connection coefficient that distinguishes patients with MCI from patients with AD; (ii) a conversion coefficient that verifies if a subject with MCI is converting to AD; (iii) some network motifs, i.e., network patterns very frequent in one kind of patient and absent, or very rare, in the other.

The material present in this chapter is taken from [293, 436].

11.1 Introduction

In recent years, the incidence of Alzheimer's Disease (hereafter, AD) is growing because the population is aging in most countries. For this reason, the efforts to design approaches capable of determining the onset of this disease in advance are intensifying [326, 561]. Even if this issue is challenging, it is extremely complex, as also evidenced in past literature. As a matter of fact, it was shown that: *(i)* AD shares many clinical features with other forms of dementia, and *(ii)* the molecular pathomechanism of AD becomes active several years before neurons start dying and cognitive deficits appear. For a definitive diagnosis of AD, the biopsy of brain tissues is necessary.

Another important issue that makes the diagnosis on these patients difficult concerns the fact that they, just by the very nature of their disease, do not easily undergo

examinations, like Magnetic Resonance Imaging that force them to stay still for a long time.

On the other hand, a non-invasive and well tolerated examination, which can be done on patients with neurological disorders, is ElectroEncephaloGram (hereafter, EEG) [579, 368]. In the scientific literature, several signal theory-based approaches employing EEG to investigate patients with AD and Mild Cognitive Impairment (hereafter, MCI) have been proposed [220, 222, 401, 489, 488, 355, 336]. The EEGs of patients with AD present some peculiarities, namely slowing, reduced complexity and perturbations in synchrony. However, it was shown that these effects can be observed with different intensities in different patients. For this reason, none of them alone allows a reliable diagnosis of AD at an early stage so far.

In this setting, one approach that is obtaining interesting results is the complex network-based one. It relies on the concept that the EEG data can be easily modeled as a network, with nodes that represent electrodes and edges that denote connections between electrodes [571, 225, 294]. Modeling the EEG in this way helps the experts to study the subjects with MCI and AD over time and also to observe the change on the level of interactions between different brain areas, which could be relevant for monitoring the evolution of the diseases.

In this Chapter, we aim at presenting a complex network analysis-based approach, whose inputs are the EEGs of the patients to analyze, performed at time t_0 and, then again three months later, at time t_1. Given an EEG of a patient, our approach constructs a network with nodes that represent the electrodes and edges that denote connections between electrodes. Each edge has associated a weight representing a measure of the connection level between the brain areas covered by the corresponding electrodes.

Once the network associated with an EEG has been constructed, it is possible to employ the enormous wealth of knowledge already existing in network analysis to face the issues of our interest. In particular, since it is well known that, in AD progression and in MCI progression towards AD [542], a key role is played by the loss of connectivity among the different cortical areas, it appears reasonable to start our analysis from the knowledge on connectivity gained in network analysis in the past. Here, one of the most important tools for this purpose is the concept of clique. We recall that a clique of dimension k in a network represents a completely connected subnetwork formed by k nodes.

Our approach applies this concept to construct a suitable data structure, which we call *clique network, and an indicator of the connectivity level of the brain areas, called* connection coefficient, allowing us to distinguish patients with MCI from patients with AD. A further indicator called *conversion coefficient*, which associates the

quantification of connection loss with the probability that such a loss corresponds to the MCI converting to AD, has proven particularly useful in helping experts to understand if a patient with MCI is going towards AD. In our opinion, connection and conversion coefficients represent a first relevant contribution. Indeed, the literature lacks longitudinal studies on MCI/AD, due to the difficulty in keeping such patients and their caregivers loyal to a periodical follow-up program. We believe that the present research can be a starting point for motivating other people to engage longitudinal studies on MCI and AD. In order to face this issue of a limited-size database, we performed a further experimental campaign on virtual patients with MCI or AD, suitably constructed from the real ones (see Section 11.4.1).

In addition, our approach aims at verifying if *network motifs* exist, i.e., specific sub-networks, or network patterns, which are very frequent in one kind of patient and absent, or very rare, in the other. Also for this issue we have obtained interesting results, since we have found some motifs characterizing patients with MCI from patients with AD. Interestingly, our concept of motif has a further, much more important, feature. Indeed, it could provide a characterization of the behavior of brain areas in presence of a disorder (or when a patient converts from a disorder to another). For instance, motifs could denote what brain areas are more connected and/or more active in presence of MCI and in absence of AD or, dually speaking, what brain areas are most affected or damaged when a patient with MCI converts to AD. As for this topic, the results obtained by our approach are very similar to the ones obtained by the approach described in [386], acquired by applying a completely different methodology.

Besides these two major contributions, we present some minor ones. For instance, our analysis confirms the previous results, obtained in past literature through completely different approaches [671, 278, 221, 106], about the capability of helping experts to understand if a patient with MCI converts to AD, which characterizes the tracings of some of the four sub-bands (i.e., α, β, δ and θ) of an EEG. In particular, according to past results obtained in the literature, we have shown that the sub-bands δ and θ play a key role in this context. Furthermore, we introduce the connection coefficient. This parameter is strictly dependent on both the number and the dimension of the cliques that can be found in the network. Since cliques represent completely connected subnetworks, connection coefficient is well suited as an indicator of the connection degree of a network. Actually, as we will show below, connection coefficient shows a much better performance than clustering coefficient, which is the parameter classically adopted in Social Network Analysis to measure the connectivity degree of a network. Finally, it is worth noting that our approach might be extended to other neurological disorders, related to an impairment of cor-

tical connectivity (Parkinson's disease [624] [331], schizophrenia [196, 29], epilepsy [389, 659], ADHD [20] and autism [21]).

This chapter is organized as follows: in Section 11.2, we present the related literature. Then, in Section 11.3, we illustrate the proposed approach in detail, and introduce the definition of the connection and conversion coefficients. In Section 11.4, we describe the results of the experimental campaign we conducted to determine the adequacy of our approach and discuss them.

11.2 Related Literature

Several approaches investigate the slowing of EEGs in patients with AD (see, for instance, [214, 658]). In particular, some of these last also investigate the effect of AD in the tracings of EEGs in the sub-bands α, β, δ and θ. The changes in spectral power are determined by means of Fourier Transform [214, 658] or sparsified time-frequency maps [658]. Other approaches analyze the reduced complexity of EEG signals in patients with AD (see, for instance, [336, 105]). In this context, to quantify this reduction, the authors apply several measures, namely approximate entropy [336], auto mutual information [336], sample entropy [336], multiscale entropy [336], Lempel-Ziv complexity [336], and fractal dimension [105, 534]. Finally, further approaches investigate the decrease of synchrony in patients with MCI and AD w.r.t. age-matched control subjects (see, for instance, [214, 222]). To quantify this decrease, many measures have been proposed, e.g., Pearson correlation coefficient [222], coherence [222, 584], Granges causality [222], information-theoretic [222], state space-based synchrony measures [214, 222], phase synchrony indices [214, 222] and stochastic event synchrony [222].

Few studies evidence an increase of EEG synchrony in patients, recorded during working memory task [692]. This inverse effect is often interpreted as the result of a compensatory mechanism in the brain. Several works (e.g., [19, 53]) examine the changes of brain activity in patients with MCI using MagnetoEncephaloGram (MEG), instead of EEG.

Network analysis [27, 231, 116, 300] has been frequently applied in the investigation of modern brain mapping techniques. Indeed, it provides several neurobiologically meaningful and easily computable measures [325, 315] to reliably quantify the main characteristics of brain networks. Furthermore, it is extremely useful to detect possible connectivity abnormalities characterizing neurological and psychiatric disorders [549, 667]. Typical network analysis parameters and structures adopted for this purpose are functional segregation [677, 504], functional integration [14], paths

in functional networks [335], anatomical motifs [480, 619, 518]. Network analysis was also adopted to quantify the resilience of brain to insults [49].

Many approaches (e.g., [225, 705, 654, 353]) focus on the usage of network analysis to investigate MCI or AD through the EEGs of the corresponding patients (an overview of these studies can be found in [482]). The parameter generally adopted to measure the connection level of brain areas is clustering coefficient, even if other basic network analysis parameters, such as characteristic path length, global efficiency, connectivity degree and connectivity density, have been proven able to partially evidence the loss of connectivity characterizing the progression of AD [487]. In some cases, these measures are applied not only to the overall EEG but also to one or more sub-bands (for instance, [620] considers the β sub-band). In [386], the authors investigate the spatial distribution of EEG phase synchrony in patients with AD. For this purpose, they analyze the surface topography of the Multivariate Phase Synchronization of multichannel EEG. They investigate these features for both the overall EEG and its sub-bands.

11.3 Methods

11.3.1 Input and Support Data Structures

The input of our approach consists of a set $EEGSet$ of EEGs at our disposal. It has the following structure:

$$EEGSet = \{CtrlSet, MCISet_0, ADSet_0, MCISet_1, ADSet_1\}$$

where: *(i)* $CtrlSet$ is the set of the EEGs of the control subjects; *(ii)* $MCISet_0$ (resp., $MCISet_1$) is the set of the EEGs of the patients with MCI at t_0 (resp., t_1); *(iii)* $ADSet_0$ (resp., $ADSet_1$) is the set of the EEGs of the patients with AD at t_0 (resp., t_1).

Let eeg be an EEG[1] of $EEGSet$. Starting from eeg, it is possible to define a network:

$$\mathcal{N} = \langle V, E \rangle$$

Here, V is the set of nodes of $\mathcal{N}$. Each node $v_i \in V$ corresponds to an electrode of the EEG. In our EEGs, electrodes were applied by following the 10-20 system and $|V| = 19$.

E is the set of the edges of $\mathcal{N}$. Each edge e_{ij} connects the nodes v_i and v_j and can be represented as:

[1] At this moment, we do not make any assumptions about the subject whom eeg refers to. She/he could be a control subject, a patient with MCI or a patient with AD.

$$e_{ij} = (v_i, v_j, w_{ij})$$

Here, w_{ij} is a measure of "distance" between v_i and v_j. It is an indicator of the disconnection level of v_i and v_j. Even if our approach is orthogonal to the measure adopted for estimating synchrony, in our experiments we chose to employ PDI (*Permutation Disalignment Index*), which proved to be well suited in quantifying the overall coupling strength between EEG signals associated with MCI progression towards AD [459]. In particular, PDI was compared with Coherence and Dissimilarity Index, a nonlinear and symbolic measure that proved to be promising in the pairwise analysis of EEG data. PDI was shown to outperform both Coherence and Dissimilarity Index [459]. It can help whenever a multivariate, amplitude invariant, robust to noise, nonlinear coupling strength analysis is necessary. All the above mentioned features are useful in EEG processing because EEG is multivariate, influenced by the distance from the reference electrode, affected by noise and nonlinear behavior. For all these reasons, in our experiments, w_{ij} was set to the average PDI between v_i and v_j.

In order to make our model more "user-friendly" and "expressive" and, at the same time, more capable of discriminating strong and weak connections between the different brain areas, we decided to construct a new network, namely $\mathcal{N}_\pi$, obtained from $\mathcal{N}$ by removing the edges with an "excessive" weight (see below) and by coloring the others on the basis of their weight. As a matter of fact, edges with an "excessive" weight represent connections between portions of the brain having a low connection degree. In particular, blue edges denote strong connections (i.e., small weights), red edges represent intermediate ones and, finally, green edges indicate weak connections. In the following, we formalize this reasoning:

$$\mathcal{N}_\pi = \langle V, E_\pi \rangle$$

Here, the nodes of $\mathcal{N}_\pi$ are the same as the ones of $\mathcal{N}$. To define E_π, we employ the distribution of the weights of the edges of $\mathcal{N}$. Specifically, let max_E (resp., min_E) be the maximum (resp., minimum) weight of an edge of E. Starting from them, it is possible to define a parameter $step_E = \frac{max_E - min_E}{10}$, which represents the length of a "step" of the interval between min_E and max_E. We can define $d^k(E)$, $0 \leq k \leq 9$, as the number of the edges of E with weights that belong to the interval between $min_E + k \cdot step_E$ and $min_E + (k+1) \cdot step_E$. All these intervals are closed on the left and open on the right, except for the last one that is closed both on the left and on the right. We are now able to formalize E_π. Specifically, it consists of all the edges of E belonging to $d^k(E)$, where $k \leq th_{max}$.

Now, we can "color" the edges composing E_π. Specifically, $E_\pi = E_\pi^b \cup E_\pi^r \cup E_\pi^g$. Here:

- $E^b_\pi = \left\{e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{min} \leq k \leq th_{br}} d^k(E)\right\}$;
- $E^r_\pi = \left\{e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{br} < k \leq th_{rg}} d^k(E)\right\}$;
- $E^g_\pi = \left\{e_{ij} \in E \mid e_{ij} \in \bigcup_{th_{rg} < k \leq th_{max}} d^k(E)\right\}$.

In this definition, we determined the bounds of E^b_π, E^r_π and E^g_π experimentally. In particular, we set the values of th_{min}, th_{br}, th_{rg} and th_{max} to 0, 1, 4 and 6, respectively. From this definition, it is clear that discarded edges are those belonging to the eighth, ninth and tenth intervals of the range $[min_E, max_E]$.

To give an idea of the expressiveness of colored networks, in Figure 11.1 we report the distribution of the edge weights and the colored network of a control subject (resp., a patient with MCI, a patient with AD). The disposal of nodes in the network reflects the 10-20 system, even if they are rotated 90 degrees clockwise. It is straightforward to observe that the control subject presents a weight distribution more biased on the left than the patient with MCI, who, in turn, presents a weight distribution more biased on the left than the patient with AD. A direct consequence of this fact is that the colored network of the patient with AD presents lesser and weaker edges than the colored network of the patient with MCI that, in turn, presents lesser and weaker edges than the colored network of the control subject.

In order to quantify this phenomenon, in Table 11.1 we report the values of some measures characterizing the three colored networks shown in the three figures above. Specifically, the considered measures are: *(i)* the total number of colored edges; *(ii)* the total number of blue (resp., red, green) edges[2]; *(iii)* the percentage of colored edges against the total number of original edges; *(iv)* the percentage of blue (resp., red, green) edges against the total number of original edges. The quantitative results reported in Table 11.1, fully confirm the qualitative analysis mentioned above.

Parameter	*Control Subject*	*Patient with MCI*	*Patient with AD*
Total number of colored edges	170	141	69
Total number of blue edges	105	35	2
Total number of red edges	59	75	40
Total number of green edges	6	31	27
Percentage of colored edges	99.4%	82.5%	40.3%
Percentage of blue edges	61.4%	20.5%	1.2%
Percentage of red edges	34.5%	43.8%	23.4%
Percentage of green edges	3.5%	18.1%	15.8%

Table 11.1: Quantitative results representing the networks of Figure 11.1

[2] Recall that blue edges are the strongest ones, red edges have an intermediate weight, whereas green edges are the weakest ones.

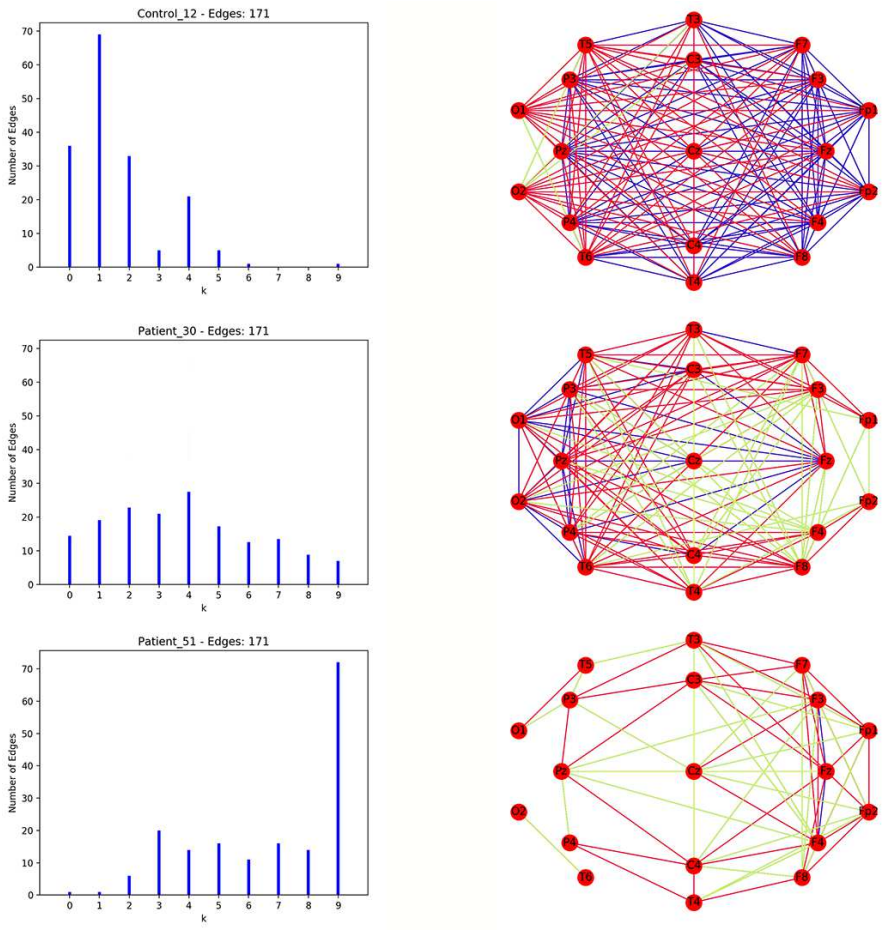


Fig. 11.1: Distributions of the edge weights and colored networks for the possible kinds of subjects into consideration

As pointed out in the Introduction, the concept of clique[3] can play a key role in the investigation of those neurological diseases, like MCI and AD, where it is extremely important to analyze the connection level between brain areas. For this reason, in our approach, we introduce a further support data structure, called clique network.

In particular, let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi = \langle V, E_\pi \rangle$ be the corresponding colored network and let $\mathcal{C}$ be the set of the cliques of $\mathcal{N}_\pi$. The clique network $\mathcal{CN}$, corresponding to $\mathcal{N}_\pi$ and $\mathcal{C}$, is defined as:

$$\mathcal{CN} = \langle CV, CE \rangle$$

Here:

- CV denotes the set of the nodes of $\mathcal{CN}$. There is a node $v_i \in CV$ for each node of $\mathcal{N}_\pi$. A weight w_i is associated with v_i. It represents the number of cliques, which

[3] Recall that a clique of dimension k in a network represents a completely connected subnetwork formed by k nodes.

v_i is involved in. Formally speaking, let v_i be a node of $\mathcal{N}_\pi$ and let $\mathcal{C}_i$ be the set of the cliques of $\mathcal{N}_\pi$ which v_i is involved in (clearly $\mathcal{C}_i \subseteq \mathcal{C}$). Then CV is defined as:

$$CV = \{(v_i, w_i) | v_i \in V, w_i = |\mathcal{C}_i|\}$$

- CE represents the set of the edges of $\mathcal{CN}$. There is an edge $(v_i, v_j, w_{ij}) \in CE$ if the edge (v_i, v_j) is present in at least one clique of $\mathcal{C}$. w_{ij} denotes the number of cliques of $\mathcal{C}$, which (v_i, v_j) is involved in.

The edges of $\mathcal{CN}$ can be "colored" in an analogous way to the edges of $\mathcal{N}_\pi$. Also in this case, blue edges are the strongest ones, red edges have an intermediate strength and green edges are the weakest ones. Formally speaking:

$$CE = CE^b \cup CE^r \cup CE^g$$

Here:

- $CE^b = \left\{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ w_{ij} > th_{rb}\right\}$;
- $CE^r = \left\{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE,\ (w_{ij} > th_{gr}) \wedge (w_{ij} \leq th_{rb})\right\}$;
- $CE^g = \left\{(v_i, v_j, w_{ij}) \mid (v_i, v_j, w_{ij}) \in CE, w_{ij} \leq th_{gr}\right\}$.

Analogously to what we have seen for $\mathcal{N}_\pi$, we experimentally determined the values of th_{rb} and th_{gr}. In particular, we found that the best values for them are $th_{gr} = 4$ and $th_{rb} = 6$. We point out that clique network is very expressive from a visual point of view. Indeed, the color of an edge is an indicator of the strength of the connection between the corresponding brain areas, whereas the dimension of a node is an indicator of the connection degree of the corresponding brain area, and, ultimately, an indicator of its activity level.

In Figure 11.2, we report the clique networks corresponding to the EEGs of three patients at the time instants t_0 and t_1. Here, the dimension of a node is directly proportional to the associated weight. In this figure and in the following, we use the notation Patient X (MCI-MCI) - where X is a number - to denote a patient suffering from MCI at both t_0 and t_1. Analogously, Patient X (MCI-AD) indicates a patient with MCI at t_0 and AD at t_1. Finally, Patient X (AD-AD) represents a patient with AD at both t_0 and t_1.

Analogously to what we have done for colored networks, also in this case, in Table 11.2, we provide some quantitative measures characterizing the clique networks of Figure 11.2. Specifically, in this case, the considered measures are: *(i)* the total number of colored edges; *(ii)* the total number of blue (resp., red, green) edges; *(iii)* the percentage of colored edges against the total number of theoretically possible edges; *(iv)* the number of nodes with weights from 1 to 10. Even in this case, the quantitative values reported in this table fully confirm the qualitative analysis mentioned above.

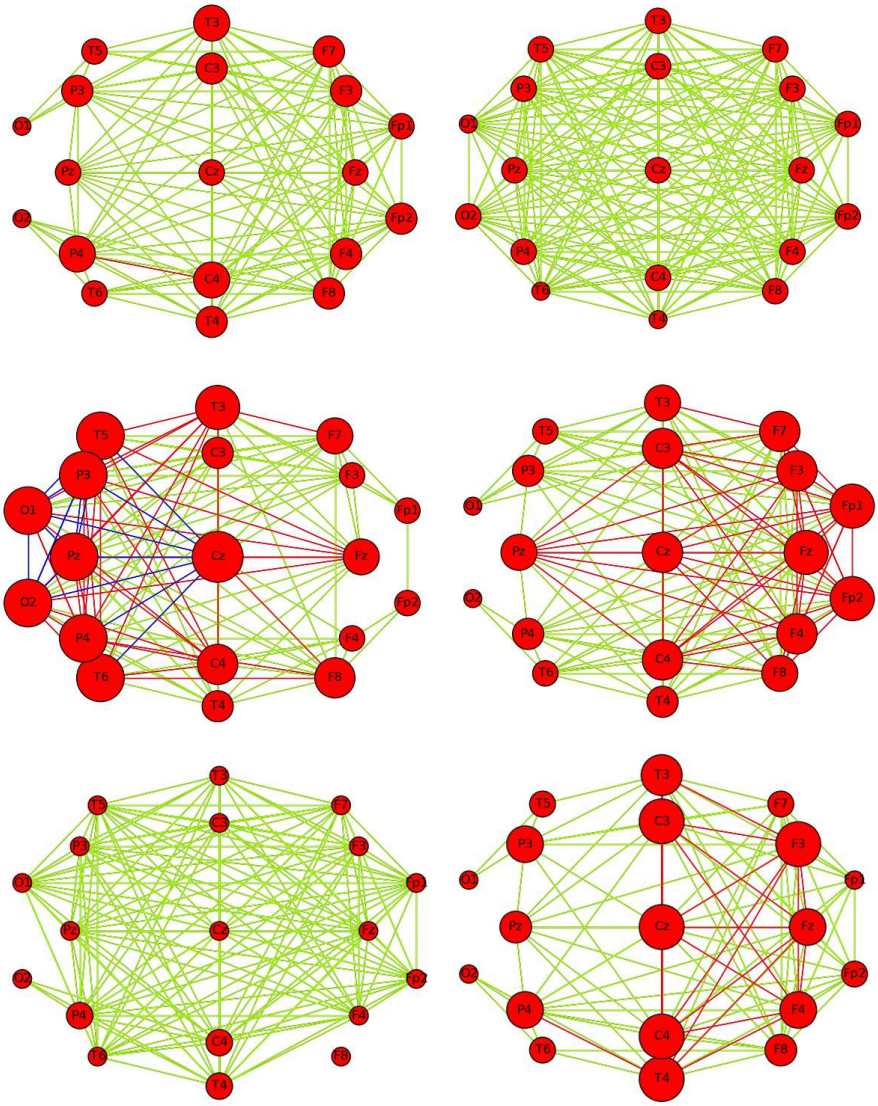


Fig. 11.2: The clique networks of Subjects 12 (Control Subject), 30 (MCI-MCI) and 51 (MCI-AD) at t_0 (on the left) and t_1 (on the right)

11.3.2 Connection Coefficient

As pointed out in the Introduction, one of the main features to investigate in neurodegenerative patients is the connection level of the brain areas. Previously, we introduced the concept of clique, which is one of the most powerful tools in network analysis for investigating the connection level of a network. Starting from cliques, it is possible to define a quantitative coefficient, which we call *connection coefficient*, capable of measuring the connectivity level of a network associated with an EEG.

This coefficient should take the following considerations into account:

- Both the dimension and the number of cliques are important as connectivity indicators.

Parameter	*Control 12* at t_0 *(Control)*	*Control 12* at t_1 *(Control)*	*Patient 30* at t_0 *(MCI)*	*Patient 30* at t_1 *(MCI)*	*Patient 51* at t_0 *(MCI)*	*Patient 51* at t_1 *(AD)*
Total number of colored edges	129	171	123	122	148	107
Total number of blue edges	0	0	23	0	0	0
Total number of red edges	1	0	40	48	0	21
Total number of green edges	128	171	60	74	148	86
Percentage of colored edges	75.4%	100%	71.9%	71.3%	86.5%	62.6%
Percentage of blue edges	0%	0%	13.6%	0%	0%	0%
Percentage of red edges	0.6%	0%	23.4%	28%	0%	12.3%
Percentage of green edges	74.8%	100%	35.1%	43.3%	86.5%	50.3%
Number of nodes whose weight is 0	0	0	0	0	1	0
Number of nodes whose weight is 1	2	19	0	2	15	3
Number of nodes whose weight is 2	6	0	4	2	3	4
Number of nodes whose weight is 3	8	0	1	3	0	3
Number of nodes whose weight is 4	3	0	3	3	0	4
Number of nodes whose weight is 5	0	0	2	6	0	5
Number of nodes whose weight is 6	0	0	1	3	0	0
Number of nodes whose weight is 7	0	0	6	0	0	0
Number of nodes whose weight is 8	0	0	2	0	0	0
Number of nodes whose weight is 9	0	0	0	0	0	0
Number of nodes whose weight is 10	0	0	0	0	0	0

Table 11.2: Quantitative results representing the networks of Figure 11.2

- The concept of clique is intrinsically exponential. In other words, a clique of dimension $n+1$ is exponentially more complex than a clique of dimension n.
- It is necessary to avoid the possible presence of outliers and noise. As a consequence, it is inappropriate to consider only the cliques with the maximum dimension. By contrast, it is more balanced to consider, in addition to them, the cliques with the sub-maximum and sub-sub-maximum dimension. On the other hand, it is unnecessary and time consuming to consider the other cliques because their contribution decreases exponentially against their dimension.

Starting from these considerations, we now define our connection coefficient. Let $\mathcal{N}_\pi = \langle V, E_\pi \rangle$ be the colored network associated with an EEG of $EEGSet$. Let $\mathcal{C}$ be the set of the cliques of $\mathcal{N}_\pi$ and let $dim(\cdot)$ be a function returning the dimension of a set of cliques, all of the same dimension, received in input. Then, it is possible to define: *(i)* the subset $\mathcal{C}_{M_1} \subseteq \mathcal{C}$ of the cliques with the maximum dimension; *(ii)* the subset $\mathcal{C}_{M_2} \subset \mathcal{C}$ of the cliques with the sub-maximum dimension; *(iii)* the subset $\mathcal{C}_{M_3} \subset \mathcal{C}$ of the cliques with the sub-sub-maximum dimension.

Finally, let $|\mathcal{C}_{M_1}|$, $|\mathcal{C}_{M_2}|$ and $|\mathcal{C}_{M_3}|$ be the cardinalities (i.e., the number of cliques) of $\mathcal{C}_{M_1}$, $\mathcal{C}_{M_2}$ and $\mathcal{C}_{M_3}$, respectively. Then, the connection coefficient $cc_{\mathcal{N}_\pi}$, associated with $\mathcal{N}_\pi$, is defined as:

$$cc_{\mathcal{N}_\pi} = \sum_{i=1}^{3} \left(|\mathcal{C}_{M_i}| \cdot 2^{dim(\mathcal{C}_{M_i})} \right)$$

This definition considers all the above observations in the most suitable way.

11.3.3 Sub-band Analysis

In the previous sections, we have always considered the complete EEG tracing. However, in the literature, it is well known that an EEG tracing can be separated in several sub-bands (e.g., α, β, δ and θ) whose analysis can provide significant information in several neurological disorders. For instance, in the past, it was shown that the sub-bands δ and θ can help in investigating the conversion from MCI to AD [221, 106]. For this reason, we decided to extend all the previous analysis from the overall tracing to the ones of the sub-bands α, β, δ and θ. In this section, we illustrate this extension and the most important results we have obtained from it.

Preliminarily, we must introduce further support data structures and parameters. Specifically, let *eeg* be a generic EEG of $EEGSet$. Starting from *eeg*, it is possible to define four further tracings, namely eeg^{α}, eeg^{β}, eeg^{δ} and eeg^{θ}, referred to the sub-bands α, β, δ and θ.

In Section 11.3.1, we have defined the network $\mathcal{N} = \langle V, E \rangle$ corresponding to *eeg*. In an analogous way, it is possible to define the networks:

$$\mathcal{N}^{\alpha} = \langle V, E^{\alpha} \rangle \qquad \mathcal{N}^{\beta} = \langle V, E^{\beta} \rangle \qquad \mathcal{N}^{\delta} = \langle V, E^{\delta} \rangle \qquad \mathcal{N}^{\theta} = \langle V, E^{\theta} \rangle$$

Here, V is the set of nodes, which coincides with the nodes of $\mathcal{N}$. E^{α} (resp., E^{β}, E^{δ}, E^{θ}) represents the set of the edges of $\mathcal{N}^{\alpha}$ (resp., $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$, $\mathcal{N}^{\theta}$). Each edge of E^{α} (resp., E^{β}, E^{δ}, E^{θ}), connecting the nodes v_i and v_j, has the form (v_i, v_j, w_{ij}), where w_{ij} is a measure of the "distance" between v_i and v_j in $\mathcal{N}^{\alpha}$ (resp., $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$, $\mathcal{N}^{\theta}$). As seen in Section 11.3.1, this "distance" is an indicator of the disconnection level of v_i and v_j, and each measure representing this feature could be adopted in our model. Analogously to the overall tracing, in the experiments associated with this research, we adopted the Permutation Disalignment Index [459]. As a consequence, for the edge $(v_i, v_j, w_{ij}) \in E^{\alpha}$ (resp., E^{β}, E^{δ}, E^{θ}), w_{ij} is equal to the average value of PDI in eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}).

Beside $\mathcal{N}^{\alpha}$, $\mathcal{N}^{\beta}$, $\mathcal{N}^{\delta}$ and $\mathcal{N}^{\theta}$, it is possible to define:

- the colored networks $\mathcal{N}_{\pi}^{\alpha} = \langle V, E_{\pi}^{\alpha} \rangle$ (resp., $\mathcal{N}_{\pi}^{\beta}$, $\mathcal{N}_{\pi}^{\delta}$, $\mathcal{N}_{\pi}^{\theta}$), corresponding to eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}), by extending to this tracing what we have already done in Section 11.3.1 for the overall tracing;
- the connection coefficient $cc_{\mathcal{N}_{\pi}^{\alpha}}$ (resp., $cc_{\mathcal{N}_{\pi}^{\beta}}$, $cc_{\mathcal{N}_{\pi}^{\delta}}$, $cc_{\mathcal{N}_{\pi}^{\theta}}$), corresponding to eeg^{α} (resp., eeg^{β}, eeg^{δ}, eeg^{θ}), by extending to this tracing what we have already done in Section 11.3.2 for the overall tracing.

11.3.4 Conversion Coefficient

We have introduced the connection coefficient and we have shown that it is well suited for determining the connection degree of a network and, in our case, of the

brain. In this task, this parameter presents a better performance than clustering coefficient that is the parameter adopted in classical Network Analysis for this purpose. It also proved to be adequate to verify the conversion from MCI to AD. Finally, its adoption in sub-bands δ and θ proved to be well suited to predict the same conversion.

All these results, in the whole, suggest that, in order to quantitatively predict the conversion from MCI to AD, it is reasonable to define a new coefficient (which we call *conversion coefficient*) capable of detecting the conversion of a patient from MCI to AD more exactly, by taking the connection coefficient relative to all these three tracings into account.

The conversion coefficient can be defined as follows: let eeg be an EEG of $EEGSet$, let $\mathcal{N}_\pi$ (resp., $\mathcal{N}_\pi^\delta$, $\mathcal{N}_\pi^\theta$) be the corresponding colored network associated with the overall tracing (resp., the sub-bands δ and θ) of eeg, let $cc^0_{\mathcal{N}_\pi}$, $cc^0_{\mathcal{N}_\pi^\delta}$, $cc^0_{\mathcal{N}_\pi^\theta}$ (resp., $cc^1_{\mathcal{N}_\pi}$, $cc^1_{\mathcal{N}_\pi^\delta}$, $cc^1_{\mathcal{N}_\pi^\theta}$) be the corresponding connection coefficients at t_0 (resp., t_1). The conversion coefficient $conv_{eeg}$, corresponding to eeg, is defined as:

$$conv_{eeg} = \frac{1}{3} \cdot \left(\frac{cc^1_{\mathcal{N}_\pi} - cc^0_{\mathcal{N}_\pi}}{cc^0_{\mathcal{N}_\pi}} + \frac{cc^1_{\mathcal{N}_\pi^\delta} - cc^0_{\mathcal{N}_\pi^\delta}}{cc^0_{\mathcal{N}_\pi^\delta}} + \frac{cc^1_{\mathcal{N}_\pi^\theta} - cc^0_{\mathcal{N}_\pi^\theta}}{cc^0_{\mathcal{N}_\pi^\theta}} \right)$$

In other words, the conversion coefficient $conv_{eeg}$ of an electroencephalogram eeg considers the variations of the connection coefficients $cc_{\mathcal{N}_\pi}$, $cc_{\mathcal{N}_\pi^\delta}$ and $cc_{\mathcal{N}_\pi^\theta}$ associated with the overall tracing and with the tracings corresponding to the sub-bands δ and θ. All these contributions are taken with the same weight.

11.3.5 Network Motifs

In this section, we aim at investigating the possible presence of motifs characterizing patients with MCI from patients with AD, and vice versa.

As a matter of fact, motifs have been already investigated and used in past approaches adopting network analysis (see, for instance, [480, 619, 518]). In this scenario, they are considered as [480]:

> "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks".

In our approach, we use motifs in a completely different way. Indeed, we do not examine a unique complex network to find patterns frequently repeated therein. By contrast, we search for patterns appearing frequently in the networks corresponding to the tracing segments of patients with MCI (resp., AD) and being absent in the ones of patients with AD (resp., MCI), thus characterizing the patients with MCI (resp., AD) from the ones with AD (resp., MCI).

First, we must formalize our concept of motif. Specifically, let $EEGSet$ be a set of EEGs, let $MCISet$ (resp., $ADSet$) be the subset of $EEGSet$ corresponding to patients with MCI (resp., AD). Let $NSet_M$ (resp., $NSet_A$) be the set of colored networks corresponding to the EEGs of $MCISet$ (resp., $ADSet$). Let $\mathcal{C}_M$ (resp., $\mathcal{C}_A$) be the set of the cliques of $NSet_M$ (resp., $NSet_A$) and let $\mathcal{T}_M$ (resp., $\mathcal{T}_A$) be the set of totally connected triads of $\mathcal{C}_M$ (resp., $\mathcal{C}_A$)[4]. Finally, let t be a generic triad. We call $nocc_M$ (resp., $nocc_A$) the number of occurrences of t in $NSet_M$ (resp., $NSet_A$).

After having defined all support data structures and parameters, we are able to describe our motif extraction approach. It consists of two main steps, the former devoted to the extraction of basic motifs and the latter conceived for the construction of derived ones. In this section, we illustrate the extraction of basic motifs. Preliminarily, it is necessary to specify what a basic motif is in our context. Specifically:

> Let t be a totally connected triad of $NSet_M$. If: (1) t is *absent* in the networks of $NSet_A$ and is *frequent* in the networks of $NSet_M$, or (2) t is *very rare* in the networks of $NSet_A$ and *very frequent* in the networks of $NSet_M$, then t is a motif characterizing patients with MCI from patients with AD.

To really extract basic motifs, it is necessary to specify a quantitative definition of this rule. To carry out this task, it is preliminarily necessary to associate numeric values with the concepts of *very rare*, *frequent* and *very frequent*. For this purpose, we can define the following thresholds:

$$th_{VR} = \alpha_{VR} \cdot |NSet_A| \qquad th_F = \alpha_F \cdot |NSet_M|$$
$$th_{VF} = \alpha_{VF} \cdot |NSet_M|$$

We experimentally set the values of α_{VR}, α_F and α_{VF} to 0.10, 0.25 and 0.40, respectively. We chose these values as the default ones of our approach. In fact, they proved to be the most "equilibrate" (i.e., neither extremely permissive nor extremely restrictive) ones.

Therefore, let $t \in \mathcal{T}_M$ be a totally connected triad of $NSet_M$ and let $nocc_M$ (resp., $nocc_A$) be the number of occurrences of t in $NSet_M$ (resp., $NSet_A$). If:

- (1) $(nocc_A = 0) \wedge (nocc_M \geq th_F)$, or
- (2) $(nocc_A \leq th_{VR}) \wedge (nocc_M \geq th_{VF})$

then t is a basic motif characterizing patients with MCI from patients with AD.

[4] We recall that a triad is a subnetwork consisting of three nodes. The totally connected triad is considered the most stable structure in network analysis. A totally connected triad can be considered as a clique of dimension 3.

In a dual fashion, it is possible to define the basic motifs characterizing patients with AD from patients with MCI. Also in this case, we experimentally set the values of α_{VR}, α_F and α_{VF} to 0.10, 0.25 and 0.40, respectively.

In the following, we indicate by $\mathcal{M}_M$ (resp., $\mathcal{M}_A$) the set of motifs extracted starting from the triads of $NSet_M$ (resp., $NSet_A$).

Observe that a motif is not only an indicator of the tracing segments of the EEGs of patients with MCI (or with AD). As a matter of fact, it is much more. Indeed, it allows us to characterize the behavior of the brain areas of patients with MCI (resp., AD) from patients with AD (resp., MCI). For instance, it denotes what brain areas are most connected (and, therefore, most active) in patients with MCI before converting to AD (resp., in patients that converted from MCI to AD).

Once basic motifs have been extracted, and a first version of $\mathcal{M}_M$ and $\mathcal{M}_A$ has been obtained, it is possible to construct derived (and, possibly, much more complex and significant) motifs starting from them.

Our approach constructs new derived motifs starting from the already known ones. For this purpose, it uses nodes common to two or more known motifs as "junction points". Formally speaking, let $m_i = \langle V_i, E_i \rangle$ and $m_j = \langle V_j, E_j \rangle$ be two motifs of $\mathcal{M}_M$ such that $V_i \cap V_j \neq \emptyset$. Then, it is possible to construct a candidate motif by computing the union of the nodes and the edges of m_i and m_j:

$$m_{ij} = \langle V_i \cup V_j, E_i \cup E_j \rangle$$

Once m_{ij} has been constructed, analogously to what we have seen for basic motifs, it is necessary to evaluate $nocc_M$ and $nocc_A$[5]. If, for these parameters, condition (1) or condition (2) for the extraction of basic motifs hold, then m_{ij} can be added to $\mathcal{M}_M$, i.e., $\mathcal{M}_M = \mathcal{M}_M \cup \{m_{ij}\}$.

The addition of a new motif in $\mathcal{M}_M$ could make possible the construction of new candidate motifs. As a consequence, the enrichment process of $\mathcal{M}_M$ is iterative and terminates when, during an iteration, no new motif is added to $\mathcal{M}_M$. In an analogous fashion, the derived motifs of $\mathcal{M}_A$ can be extracted.

11.4 Results

11.4.1 Testbed

We enrolled seven patients with AD and eight patients with MCI monitored at the IRCCS Centro Neurolesi Bonino Pulejo of Messina (Italy), within a three-month

[5] Clearly, for derived motifs, $nocc_M$ and $nocc_A$ refer to the number of occurrences of motifs, instead of triads.

follow-up program. The main characteristics of these patients are reported in Table 11.3.

Patient ID	*Age*	*Gender*	*Diagnosis at* t_0	*Diagnosis at* t_1
pt_03	68	M	MCI	AD
pt_23	84	F	MCI	MCI
pt_30	69	M	MCI	MCI
pt_41	78	M	MCI	MCI
pt_51	71	F	MCI	AD
pt_57	83	M	MCI	MCI
pt_71	79	F	MCI	AD
pt_72	65	F	MCI	MCI
pt_31	74	M	AD	AD
pt_54	83	F	AD	AD
pt_64	74	F	AD	AD
pt_65	76	M	AD	AD
pt_76	79	F	AD	AD
pt_86	83	F	AD	AD
pt_87	78	F	AD	AD

Table 11.3: Main characteristics of the patients enrolled for our experiments

Every subject signed an informed consent form, in agreement with a clinical protocol approved by the Ethical Committee. We also enrolled eighteen control subjects. The diagnostic procedure followed the guidelines of the Diagnostic and Statistical Manual of Mental Disorders (fifth edition, DSM-5) [67] and consisted of a full cognitive and clinical assessment, carried out by a multidisciplinary team of neurologists, psychologists, psychiatrists and EEG experts. Each patient was evaluated at baseline (time t_0) and then again three months later (time t_1). The patients were evaluated neuroradiologically, in order to rule out other clinical conditions, like brain lesions, which might have caused cognitive impairment. Current medical treatment (particularly cholinesterase inhibitors - ChEis, Memantine, anti-depressants, anti-psychotics and anti-epileptic drugs) was also taken into account in AD patients. MCI subjects were not under medical treatment. Furthermore, we also had 18 EEGs of control subjects.

The EEGs were recorded according to the 10-20 International System (19 channels), with 1024 *Hz* sampling rate. A 50 *Hz* notch filter was used, with linked earlobe (A1-A2) reference. The EEG recordings were performed in a comfortable resting state. The patients kept their eyes closed but remained awake. The EEG was band-pass filtered at 0.5-32 *Hz* through the Matlab toolbox *EEGLab* (`https://sccn.ucsd.edu/eeglab/`) [230]. EEG preprocessing was fully carried out in Matlab (The MathWorks, Inc., Natick, MA, USA). After filtering, the artifactual segments in the EEG recordings were manually detected by the EEG experts and the

artifactual epochs were discarded. The average time length of the recordings, after artifact cancellation, is 5.44 *mins*. After that, the four major EEG rhythms, i.e., α, β, δ and θ were extracted from the EEG signals. In this way, a n-channels EEG recording was eventually split into 4 n-channels sub-band EEG recordings: EEG^{α}, EEG^{β}, EEG^{δ}, EEG^{θ}. Each sub-band of the EEG was then downsampled to 256 Hz. Every recording of the sub-bands was partitioned into 5 s non-overlapping windows, and analyzed window by window.

On the basis of the diagnosis at times t_0 and t_1, the patients into examination were partitioned in three groups, namely: *(i)* patients with MCI at t_0 that were still diagnosed MCI at t_1; *(ii)* patients with AD at t_0 that remained with AD at t_1; *(iii)* patients with MCI at t_0 that converted to AD at t_1.

As pointed out in the Introduction, we have striven to (at least partially) face the issue of the narrowness of the set of available patients. For this purpose, we realized a simulator aimed to construct virtual control subjects and virtual patients with MCI or AD. The simulator behaves as follows:

- It receives a set $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$) of matrices. Each element of this set represents the adjacency matrix of the complex network associated with the EEG of a control subject (resp., a patient with MCI, a patient with AD). The set of real control subjects (resp., patients with MCI, patients with AD) from which we constructed $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$) consisted of the 50% of the control subjects (resp., patients with MCI, patients with AD) at our disposal, selected at random. In fact, as we will see below, the other 50% of control subjects (resp., patients with MCI, patients with AD) were necessary for testing our simulator.
- It constructs a new adjacency matrix $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$) whose generic element $\overline{A^{CS}}[i,j]$ (resp., $\overline{A^{MCI}}[i,j]$, $\overline{A^{AD}}[i,j]$) represents the mean of the (i,j) elements of the matrices of $ASet^{CS}$ (resp., $ASet^{MCI}$, $ASet^{AD}$).
- It computes the standard deviation σ^{CS} (resp., σ^{MCI}, σ^{AD}) of the elements of $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$).
- It constructs the set $\widehat{ASet^{CS}}$ (resp., $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$) of the adjacency matrices representing the complex networks associated with the EEGs of virtual control subjects (resp., patients with MCI, patients with AD). In particular, the generic element $\widehat{A}[i,j]$ of a matrix of $\widehat{ASet^{CS}}$ (resp., $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$) is obtained by perturbing the corresponding element $\overline{A^{CS}}$ (resp., $\overline{A^{MCI}}$, $\overline{A^{AD}}$) of a random value comprising between $-\frac{1}{2}\sigma^{CS}$ (resp., $-\frac{1}{2}\sigma^{MCI}$, $-\frac{1}{2}\sigma^{AD}$) and $\frac{1}{2}\sigma^{CS}$ (resp., $\frac{1}{2}\sigma^{MCI}$, $\frac{1}{2}\sigma^{AD}$).

After having obtained the three sets $\widehat{ASet^{CS}}$, $\widehat{ASet^{MCI}}$, $\widehat{ASet^{AD}}$, it was necessary to couple the corresponding matrices appropriately in such a way as to represent

virtual control subjects (having an element of $\widehat{ASet^{CS}}$ at t_0 and another one of the same set at t_1), virtual patients with MCI at both t_0 and t_1 (therefore, having an element of $\widehat{ASet^{MCI}}$ at t_0 and another one of the same set at t_1), and patients with MCI at t_0 and with AD at t_1, and, finally, patients with AD at both t_0 and t_1. Each of the four sets constructed above consisted of 27 elements. After this, by following the holdout technique, for each of the four groups mentioned above, we chose 18 elements to train our approach and 9 elements to test it. After having verified the adequacy of our approach on virtual people, we tested it on the 50% of the real people not used for constructing the virtual models, in such a way as to verify its suitability on real patients. We applied this technique first to evaluate the connection coefficient on the overall EEG tracing, then to test the same coefficient on the four EEG sub-bands and, finally, to evaluate the conversion coefficient.

Before discussing the "adequacy" of our approach, a discussion about the enrollment of patients in neurological tests is in order. Nowadays it is still very difficult to keep MCI and AD subjects and their caregivers actively involved in the follow-up programs. On the other side, these programs are strictly necessary to develop biomarkers for the objective quantification of the degeneration degree of cortical electrical connectivity caused by dementia. Many subjects do not fulfil the timing of the periodic assessments. This is often due to the difficulties caused by the disease itself. This means that many recruited subjects must be later excluded from the analysis because their EEGs were not recorded following the predetermined scheduling, which implies that their inclusion would not allow the construction of a dataset with homogeneous characteristics. As a result, there are only a few longitudinal studies in which the EEG of the subjects has been recorded and evaluated twice over time. To the best of our knowledge, the largest sample ever analyzed (143 MCI subjects) was constructed within a multicentric study described in [144]. There, the authors introduced a methodology, named Implicit Function As Squashing Time (IFAST), based on artificial neural networks. IFAST succeeded to predict the conversion from amnestic MCI to AD with an 85.98% accuracy in a 1-year follow-up study. Later, this methodology was improved; however, it has been so far tested only on a classification study concerning cross-sectional MCI vs AD.

Some other follow-up studies were carried out, but the EEG was recorded and assessed only at baseline (i.e., at t_0) and was later interpreted on the basis of the new diagnosis formulated at time t_1. In particular, [322] examined 35 amnestic MCI subjects whose EEGs were recorded at time t_0. Then, they retrospectively classified these EEGs according to the diagnosis reformulated at time t_1. The features were extracted through a Phase Lag Index (PLI)-based connectivity analysis. [490] analyzed the correlation between higher alpha3/alpha2 frequency power, cortical decay and

perfusion rate with conversion to AD in a group of 76 subjects diagnosed as MCI patients at time t_0 and, then, re-evaluated at time t_1. [297] recruited 205 nondemented amyloid positive subjects (142 of them were MCI), and computed peak frequencies and relative power in the four major sub-bands (δ, θ, α, β). Then, they retrospectively evaluated the relationship between normalized EEG measures and the probability of conversion to AD. The study proposed by [548] included 86 MCI subjects. These authors introduced a Neurophysiological Biomarker Toolbox, based on β band features, to predict the conversion from MCI to AD.

All the aforementioned studies consisted in a retrospective cross-sectional classification between groups of subjects. They do not perform the longitudinal quantification of changes in the EEGs of the same subject, which is the only way to find possible correlations between changes in the characteristics of EEG signals and/or physiological changes caused by the progression of the disease.

After this premise, we can proceed to quantitatively measure the "adequacy" of our approach, we adopted the parameters generally used in the literature for this purpose (see [318] for all details). In particular, let *pos* be the number of positives in a clinical analysis (in our case, the number of patients converting from MCI to AD in real life), let *t_pos* be the number of true positives (in our case, the number of patients converting from MCI to AD in real life and correctly detected by the connection coefficient), let *f_pos* be the number of false positives, let *neg* be the number of negatives and, finally, let *t_neg* be the number of true negatives. Starting from these parameters, it is possible to define:

- *sensitivity*, or *true positive rate*, as the proportion of positives correctly identified by the approach to evaluate: $sensitivity = \frac{t\_pos}{pos}$;
- *specificity*, or *true negative rate*, as the proportion of negatives correctly identified by the approach to evaluate: $specificity = \frac{t\_neg}{neg}$;
- *precision*, as the proportion of subjects labeled as positives by the approach to evaluate and being really positives: $precision = \frac{t\_pos}{t\_pos+f\_pos}$.

Clearly, in this medical context, sensitivity is much more important than specificity and precision.

As a final remark, we performed a comparative evaluation of our connection and conversion coefficients against clustering coefficient, which is much simpler and is the classical parameter adopted in network analysis to evaluate the connection level of a network.

11.4.2 Training of the approach

First, we decided to perform a preliminary, yet rough, verification of the capability of our EEG generator to produce plausible results. For this purpose, we computed the average minimum weight, the average maximum weight and the average mean weight for the following sets: *(i)* 50% of real EEGs (control subjects, patients with MCI, patients with AD) used to "train" the EEG generator; *(ii)* virtual EEGs produced through our generator and used to train our approach; *(iii)* virtual EEGs produced through our generator and used to test our approach; *(iv)* 50% of real EEGs used to test our approach. Obtained results are reported in Table 11.4.

Set of persons	*Avg Min Weight*	*Avg Mean Weight*	*Avg Max Weight*
Real control subjects for generator training	1.2852	1.8534	3.0923
Real control subjects for approach testing	1.2114	1.8355	3.0954
Virtual control subjects for approach training	1.1887	1.8543	2.9367
Virtual control subjects for approach testing	1.1511	1.8446	2.8912
Real patients with MCI for generator training	1.3612	2.0812	3.0224
Real patients with MCI for approach testing	1.2729	1.8854	2.7689
Virtual patients with MCI for approach training	1.2723	1.8838	2.4678
Virtual patients with MCI for approach testing	1.2863	1.8856	2.4643
Real patients with AD for generator training	1.2867	2.0243	2.9498
Real patients with AD for approach testing	1.3412	2.0976	3.0657
Virtual patients with AD for approach training	1.2643	2.0385	2.9564
Virtual patients with AD for approach testing	1.2712	2.0501	2.9504

Table 11.4: Average minimum weight, average mean weight and average maximum weight for the sets of interest

From the analysis of this table we can observe that they appear plausible, similar to the corresponding real ones and, at the same time, present a reasonable heterogeneity. For instance, the maximum variation of the average minimum (resp., mean, maximum) weight is 7.90% (resp., 9.62%, 18.24%).

After this verification, we trained our approach for making it able to detect the conversion from MCI to AD. With regard to this task, we found that a decrease of the connection coefficient higher than 80% is a potentially good indicator of the conversion phenomenon. We found the identical threshold value also for the conversion coefficient.

11.4.3 Testing of the approach

The first test that we performed regarded the connection coefficient's capability of detecting the conversion of a patient from MCI to AD.

First we operated on virtual EEGs. As previously specified, we considered 27 virtual patients with MCI at both t_0 and t_1, 27 virtual patients with AD at both t_0 and

t_1 and 27 virtual patients with MCI at t_0 that converted to AD at t_1. Obtained results are shown in the first row of Table 11.5. Then, we considered real people and operated exactly as in the previous test. Obtained results are reported in the second row of Table 11.5. The analysis of this table shows that the connection coefficient appears a good parameter for predicting the conversion from MCI to AD. Sensitivity, specificity and precision obtained by this coefficient are very high, even if improvable, both for virtual patients and for real ones. Interestingly, the values obtained for real patients are higher than the ones returned for virtual patients.

Set	Sensitivity	Specificity	Precision
Virtual patients	0.94	0.91	0.72
Real patients	1.00	0.91	0.75

Table 11.5: Sensitivity, specificity and precision of the connection coefficient associated with overall EEGs

The second test was analogous to the first one, but it regarded the sub-bands of EEGs, instead of the overall tracing. The corresponding results are reported in Tables 11.6 and 11.7. These tables show that δ and θ sub-bands are very adequate for investigating the conversion of a patient from MCI to AD. This result is in line with the ones obtained by [221, 106]. Also for these sub-bands, real patients behave better than virtual ones. α and β sub-bands, instead, do not present particularly satisfying results. For all these reasons, we decided to not consider these two sub-bands in the computation of the conversion coefficient.

Set	Sensitivity	Specificity	Precision
Virtual patients (α sub-band)	0.75	0.94	0.71
Virtual patients (β sub-band)	0.85	0.80	0.72
Virtual patients (δ sub-band)	0.94	0.95	0.69
Virtual patients (θ sub-band)	0.92	0.97	0.54

Table 11.6: Sensitivity, specificity and precision of the connection coefficient associated with the sub-bands of EEGs (virtual patients)

The next test regarded the conversion coefficient's capability of detecting the conversion of a patient from MCI to AD. For this purpose, we operated in an analogous way to what we have seen for the connection coefficient, i.e., first we considered virtual EEGs and, then, real ones. Obtained results are reported in Table 11.8.

As shown in this table, the values of sensitivity, specificity and precision returned by conversion coefficient are extremely high for virtual patients and maximum for real ones. Again, real patients behave better than virtual ones.

Set	Sensitivity	Specificity	Precision
Real patients (α sub-band)	0.67	0.91	0.67
Real patients (β sub-band)	0.80	0.80	0.67
Real patients (δ sub-band)	1.00	1.00	0.75
Real patients (θ sub-band)	1.00	1.00	0.60

Table 11.7: Sensitivity, specificity and precision of the connection coefficient associated with the sub-bands of EEGs (real patients)

Set	Sensitivity	Specificity	Precision
Virtual patients	0.95	0.94	0.92
Real patients	1.00	1.00	1.00

Table 11.8: Sensitivity, specificity and precision of the conversion coefficient

As for the comparison between the connection and the clustering coefficients in distinguishing control subjects from patients with MCI and patients with AD, from the analysis of Table 11.9 we observe that:

- The average connection coefficient of virtual (resp., real) patients with MCI decreases of 14.45% (resp., 11.32%) w.r.t. the corresponding value of virtual (resp., real) control subjects. Instead, the average clustering coefficient of virtual (resp., real) patients with MCI decreases of 2.46% (resp., 2.39%) w.r.t. the corresponding value of virtual (resp., real) control subjects.
- The average connection coefficient of virtual (resp., real) patients with AD decreases of 75.77% (resp., 69.63%) w.r.t. the corresponding value of virtual (resp., real) patients with MCI. Instead, the average clustering coefficient of virtual (resp., real) patients with AD decreases of 15.16% (resp., 12.81%) w.r.t. the corresponding value of virtual (resp., real) patients with MCI.

These values clearly evidence that the connection coefficient is much better than the clustering coefficient in distinguishing control subjects, patients with MCI and patients with AD. As a consequence, even if the computation of this coefficient is more expensive than the one of the clustering coefficient, this is balanced by its much better capability of distinguishing the states of a person.

11.4.4 Comparison between Connection and Clustering coefficients

As previously pointed out, in Social Network Analysis, the most commonly used parameter for evaluating the connection level of a network is clustering coefficient. This coefficient is simpler to compute than the connection and the conversion coefficients. As a consequence, the adoption of these last ones makes sense only if they provide more accurate results. To verify if this happens, we performed some tests.

The first one aimed at computing the average connection coefficient and the average clustering coefficient for virtual and real control subjects, patients with MCI and patients with AD. The obtained results are reported in Table 11.9.

Set	Average Connection Coefficient	Average Clustering Coefficient
Virtual control subjects	232523	0.9675
Virtual patients with MCI	198785	0.9422
Virtual patients with AD	48223	0.7889
Real control subjects	226169	0.9592
Real patients with MCI	200548	0.9363
Real patients with AD	60904	0.8164

Table 11.9: Average connection coefficient and average clustering coefficient for all the sets of virtual and real people of interest

The second test aimed at comparing the capability of the conversion and the clustering coefficients in determining the conversion of a patient from MCI to AD. In Table 11.8, we report sensitivity, specificity and precision of the conversion coefficient in carrying out this task.

We performed the same analysis for clustering coefficient. In this case, we experimentally set to 80% the percentage of the decrease of the clustering coefficient necessary for saying that a patient converted from MCI to AD. The corresponding sensitivity, specificity and precision are reported in Table 11.10.

Set	Sensitivity	Specificity	Precision
Virtual patients	0.77	0.71	0.68
Real patients	0.82	0.84	0.75

Table 11.10: Sensitivity, specificity and precision of the clustering coefficient

The analysis of Tables 11.8 and 11.10 allows us to point out that conversion coefficient returned much better results than clustering coefficient. In fact, for virtual (resp., real) patients, sensitivity, specificity and precision increase of 26.31%, 32.86% and 34.78% (resp., 21.95%, 19.05% and 33.33%) if the conversion coefficient is adopted in place of the clustering coefficient.

These two tests allow us to conclude that, even if our coefficients are more complex than the clustering coefficient, they can provide much better results and, therefore, are worthwhile to be adopted.

11.4.5 Network Motifs

The basic motifs belonging to $\mathcal{M}_M$ derived by our approach are reported in Table 11.11.

Condition (1)	Condition (2)
$[Fp1, Fp2, O2]$; $[Fp1, F3, O2]$; $[Fp1, Fz, O2]$	$[Fz, C3, O2]$
$[Fp1, F4, O2]$; $[Fp1, F8, O2]$; $[Fp1, C3, O2]$	$[Fz, Cz, O2]$
$[Fp1, Cz, O2]$; $[Fp1, C4, O2]$; $[Fp1, T4, O2]$	$[Fz, C4, O2]$
$[Fp1, Pz, O2]$; $[Fp1, P4, O2]$; $[Fp1, T6, O2]$	$[Fz, T4, O2]$
$[Fp2, F3, O2]$; $[Fp2, Fz, O2]$; $[Fp2, F4, O2]$	$[Fz, Pz, O2]$
$[Fp2, F8, O2]$; $[Fp2, C3, O2]$; $[Fp2, Cz, O2]$	$[Fz, P4, O2]$
$[Fp2, C4, O2]$; $[Fp2, T4, O2]$; $[Fp2, Pz, O2]$	$[Fz, T6, O2]$
$[Fp2, P4, O2]$; $[Fp2, T6, O2]$; $[F7, F3, O2]$	$[C3, Cz, O2]$
$[F7, Fz, O2]$; $[F7, Cz, O2]$; $[F7, C4, O2]$	$[C3, C4, O2]$
$[F7, P4, O2]$; $[F3, Fz, O2]$; $[F3, F4, O2]$	$[C3, Pz, O2]$
$[F3, F8, O2]$; $[F3, T3, O2]$; $[F3, C3, O2]$	$[C3, P4, O2]$
$[F3, Cz, O2]$; $[F3, C4, O2]$; $[F3, T4, O2]$	$[C3, T6, O2]$
$[F3, T5, O2]$; $[F3, P3, O2]$; $[F3, Pz, O2]$	
$[F3, P4, O2]$; $[F3, T6, O2]$; $[F3, O1, O2]$	
$[Fz, F4, O2]$; $[Fz, F8, O2]$; $[Fz, T3, O2]$	
$[F4, C3, O2]$; $[F8, C3, O2]$; $[F8, P3, O2]$	
$[T3, C4, O2]$	

Table 11.11: The basic motifs belonging to $\mathcal{M}_M$ derived by applying condition (1) and condition (2)

On the top of Figure 11.3, we represent two basic motifs belonging to $\mathcal{M}_M$, obtained by applying our approach to the EEGs of the patients at our disposal.

With the current values of α_{VR}, α_F and α_{VF}, we did not extract any motif belonging to $\mathcal{M}_A$. This is in line with the results shown in Sections 11.4.4, where we have seen that the networks corresponding to patients with AD are much less connected than the ones corresponding to patients with MCI. However, if the human expert wants to be more "permissive", she/he can decrease the values of α_F and α_{VF} and can increase the value of α_{VR} w.r.t. the default ones specified above. In this case, she/he could find basic motifs also in $\mathcal{M}_A$.

On the bottom of Figure 11.3, we show the most significant derived motifs extracted by our approach. In order to provide a quantitative evaluation of derived motifs (which implies characterizing the tracing segments of patients with MCI from patients with AD), in Table 11.12, we report some quantitative measures characterizing them. Specifically, the considered measures are: *(i)* the number of edges linking two nodes of the right part of the brain (r-r edges); *(ii)* the number of edges linking a node of the left part and a node of the right part of the brain (l-r edges); *(iii)* the number of edges linking two nodes of the left part of the brain (l-l edges); *(iv)* the

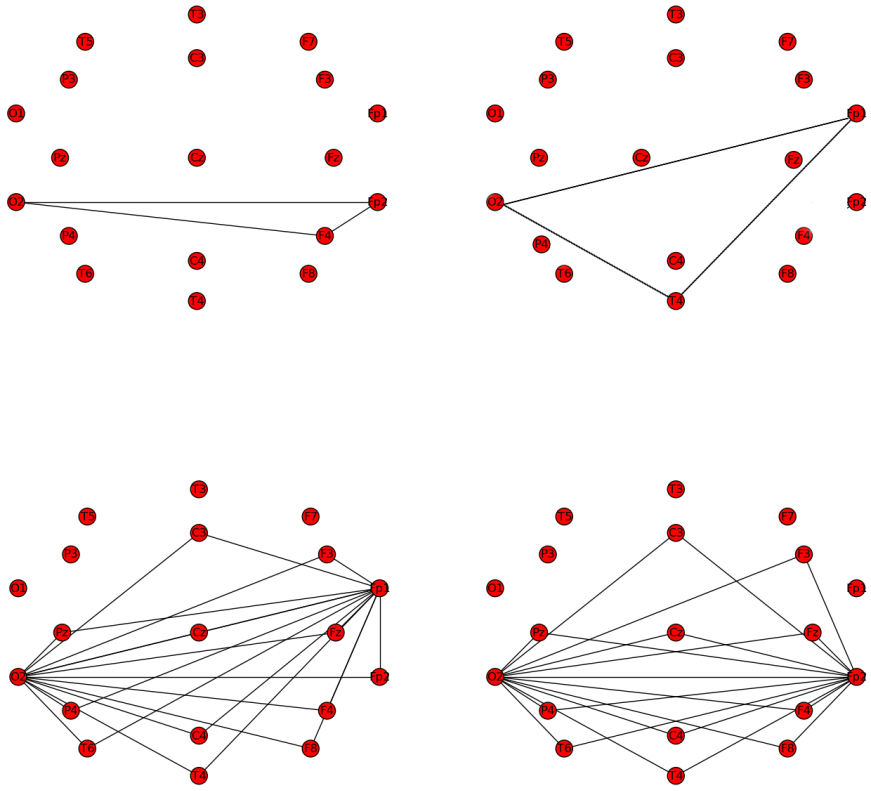


Fig. 11.3: Two of the most significant basic motifs (on the top) and two of the most significant derived motifs (on the bottom) characterizing the tracing segments of patients with MCI from patients with AD

number of edges linking a node of the central part and a node of the right part of the brain (c1-r edges); *(v)* the number of edges linking a node of the central part and a node of the left part of the brain (c1-l edges); *(vi)* the number of edges linking two nodes of the central part of the brain (c1-c1 edges); *(vii)* the number of edges linking two nodes of the frontoparietal part of the brain (f-f edges); *(viii)* the number of edges linking a node of the frontoparietal part and a node of the occipital part of the brain (f-o edges); *(ix)* the number of edges linking two nodes of the occipital part of the brain (o-o edges); *(x)* the number of edges linking a node of the central part and a node of the frontoparietal part of the brain (c2-f edges); *(xi)* the number of edges linking a node of the central part and a node of the occipital part of the brain (c2-o edges); *(xii)* the number of edges linking two nodes of the central part of the brain (c2-c2 edges).

Let us now examine in detail the two derived motifs shown in Figure 11.3. The former is centered on the electrodes $O2$ and $Fp1$, whereas the latter is centered on the electrodes $O2$ and $Fp2$. The analysis of these motifs provides important information about what happens in the brain areas when a patient converts from MCI to AD. In fact, in both cases, the node $O2$ is central. This indicates that the corresponding

brain area is very active in patients with MCI and little active (or inactive) in patients with AD. Furthermore, in both cases, it emerges a very intense activity in the right part of the brain in patients with MCI, which reduces or disappears in patients with AD. This could lead to conclude that the conversion from MCI to AD creates deeper damages in the right part of the brain (especially, in the area corresponding to the electrode $O2$) than in the left one.

As a further confirmation of these results, consider the quantitative values reported in Table 11.12. They show that most of the edges connect two nodes of the right part of the brain and that often one node is situated in the frontopolar area and the other resides in the occipital area.

Parameter	*First Derived Motif*	*Second Derived Motif*
r-r edges	7	13
l-r edges	8	4
l-l edges	2	0
c1-r edges	3	6
c1-l edges	3	0
c1-c1 edges	0	0
f-f edges	5	4
o-f edges	6	5
o-o edges	3	3
c2-f edges	4	4
c2-o edges	4	4
c2-c2 edges	0	0

Table 11.12: Quantitative results representing the derived motifs of Figure 11.3

11.4.6 Comparison with other existing approaches

In this section, we compare our approach with the one illustrated in [459]. In our opinion, this comparison is extremely interesting to highlight the potential of our approach because: *(i)* both our approach and the one of [459] use the same metric (i.e., Permutation Disalignment Index) for evaluating the connection degree of brain areas; *(ii)* the authors of [459] showed that their approach is well-suited for evaluating the conversion from MCI to AD, and they support their claim by means of comparisons between their approaches and some related ones proposed in the past.

In [459], the authors used boxplots to verify whether a subject with MCI at t_0 converts to AD at t_1 or not. We applied both the approach of [459] and our own to the EEGs of four patients. Two of them suffered from MCI at both t_0 and t_1, whereas two other ones converted from MCI at t_0 to AD at t_1. Clearly, the number of patients we are considering is very small. However, we point out that we do not aim at precisely quantifying how much the performance of our approach is better (or worse) than

the one of the approach of [459]. Actually, we simply want to provide the reader with an idea of the way of proceeding of our approach (which implies the need to graphically show the colored networks and the boxplots associated with the EEGs of the patients we are examining) and, possibly, to give a rough comparative estimation of its performance.

In Figure 11.4, we report the boxplots of the four patients into examination. In Table 11.13, we present the values of some parameters helping us to quantify the results shown therein. Analogously, in Figure 11.5, we present the colored networks of the same four patients. In Table 11.14, we show the values of the corresponding conversion coefficient.

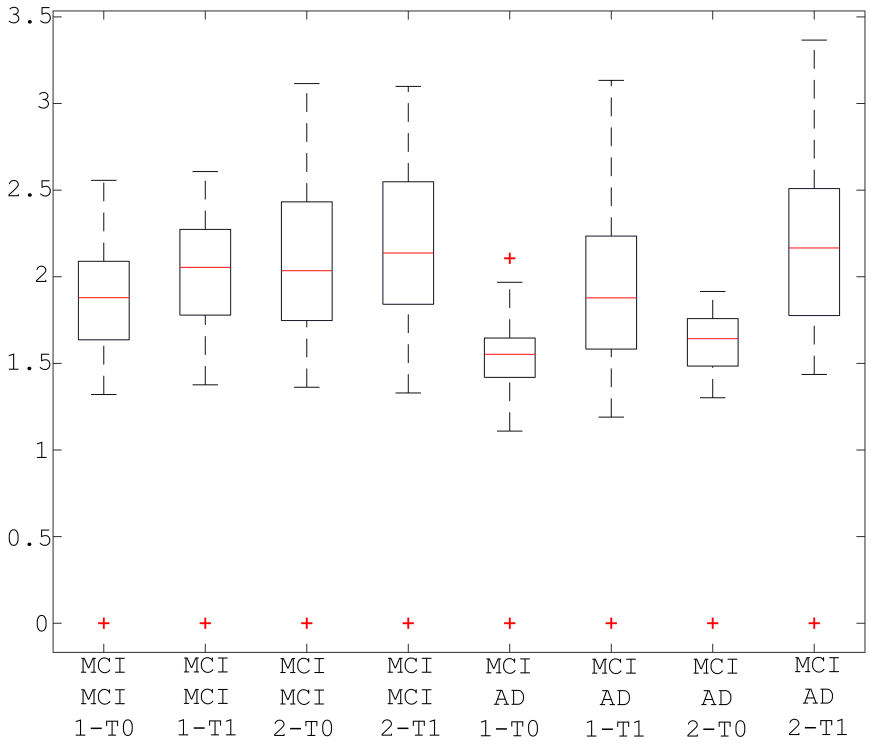


Fig. 11.4: Results of the application of the approach of [459] to the four subjects into consideration

	Variation of medians from $T0$ to $T1$	Variation of 25^{th} percentile from $T0$ to $T1$	Variation of 75^{th} percentile from $T0$ to $T1$
I subject MCI-MCI	9.04%	8.59%	9.13%
II subject MCI-MCI	4.93%	5.75%	4.53%
I subject MCI-AD	20.65%	11.27%	35.97%
II subject MCI-AD	31.70%	19.59%	43.43%

Table 11.13: Quantitative results representing the results shown in Figure 11.4

	Conversion coefficient $conv_{eeg}$
I subject MCI-MCI	-25.00%
II subject MCI-MCI	-4.96%
I subject MCI-AD	-89.06%
II subject MCI-AD	-99.41%

Table 11.14: Values of the conversion coefficient $conv_{eeg}$ for the four patients into examination

From the analysis of Figures 11.4 and 11.5 and from the comparison of Tables 11.13 and 11.14, we can observe that our approach appears more adequate than the one of [459] in distinguishing patients converting from MCI to AD from the ones who do not convert. Indeed:

- When passing from t_0 to t_1 boxplot positions certainly vary more for patients converting to AD than for patients who do not convert. However, this variation is not very clear and marked (see Figure 11.4). Vice versa, when passing from t_0 to t_1, the number and the color of network edges do not present a great variation for patients who do not convert to AD, whereas both these indicators strongly vary for patients converting to AD (see Figure 11.5).
- The variation of medians (resp., 25^{th} percentile and 75^{th} percentile) is about 6.5% (resp., 7%, 6.5%) for patients who do not convert to AD, whereas it is about 26% (resp., 15%, 44%) for patients converting to AD (see Table 11.13).

Instead, if we consider our conversion coefficient, we can observe that its value is about 12% for patients not converting to AD, whereas it is about 9% for patients converting to *AD* (see Table 11.14).

All these evaluations allow us to claim that our approach is really more adequate than the one of [459] to help an expert to visually and quantitatively evaluate the longitudinal history of a patient suffering from MCI and/or AD.

11.4.7 Findings and limitations

Clearly, the results presented in all the previous subsections will require much more efforts and investigations in the future, especially by experts in neurological diseases, in order to completely "capture" their meaningfulness. Nevertheless, they are an interesting "food for thought" that our approach is providing to researchers in this sector.

At the end of this research we can generalize the found results and draw the following hypothesis about the conversion from MCI to AD:

- Conversion coefficient is a well suited indicator of the transition of a patient from MCI to AD. In particular, a decrease of this coefficient of more than 80% in three

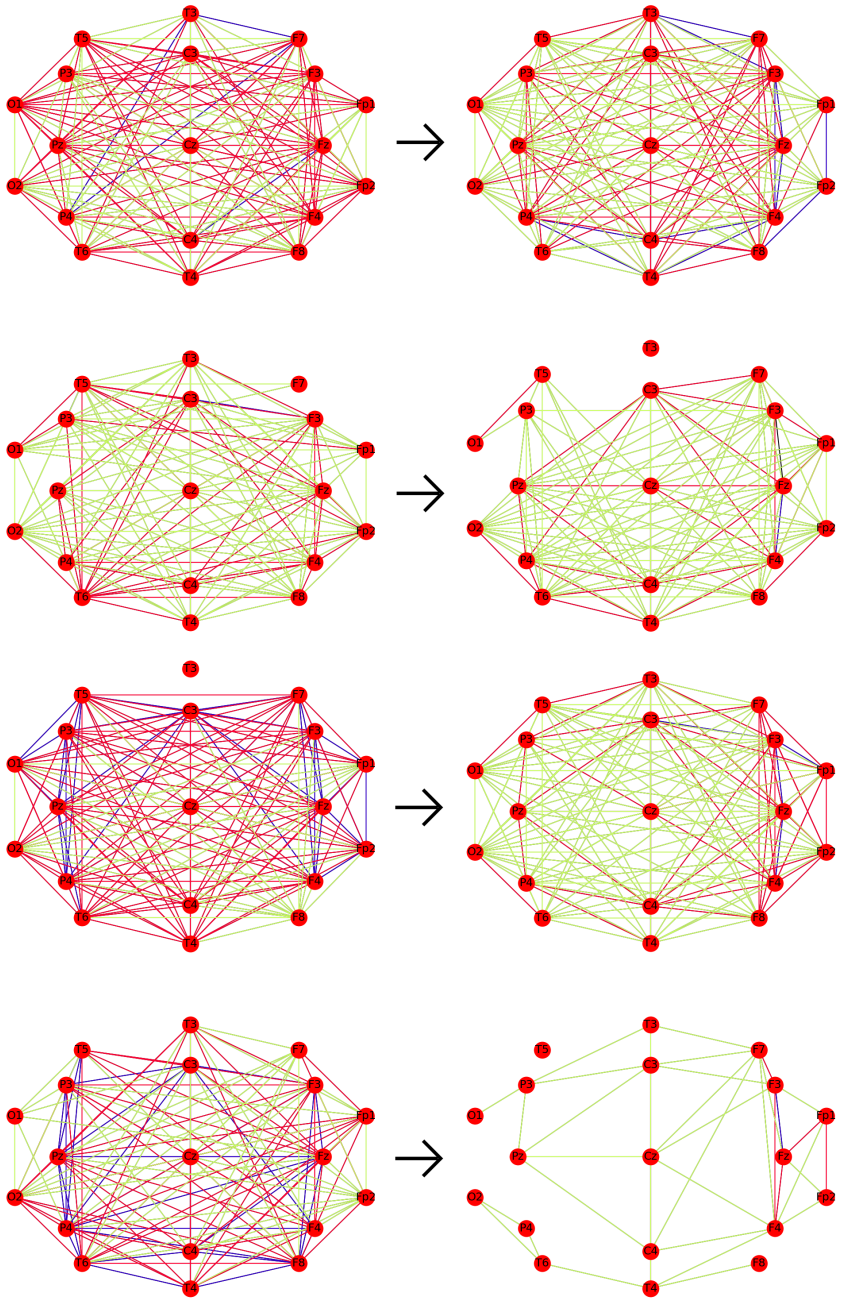


Fig. 11.5: The networks $\mathcal{N}_{0_\pi}$ and $\mathcal{N}_{1_\pi}$ for the two patients not converting to AD (above) and for the two other ones converting to AD (below)

months is a clear indicator that the corresponding patient is converting from MCI to AD.

- The activity of the brain area underlying the electrode $O2$ and of the right part of the brain is a potential indicator of a possible transition of a patient from MCI to AD. In particular, a marked reduction of the activity of these two brain parts

is a possible indicator that the corresponding patients is converting from MCI to AD.

12

Extraction of Semantic Relationships among Concepts

The term "interschema properties" is used to indicate semantic relationships (e.g., synonymies, homonymies, hyponymies, subschema similarities) among concepts. The knowledge of interschema properties plays a key role for allowing decision making using data sources of heterogeneous formats. In the past, a wide amount and variety of approaches to derive interschema properties from structured and semi-structured data have been proposed. However, currently, it is esteemed that more than 80% of data sources are unstructured and the number of sources generally involved in a data-driven task is much higher than in the past. As a consequence, we have the necessity of defining new approaches to address the interschema property derivation issue. In this chapter, we propose an approach capable of uniformly extracting interschema properties from a huge number of structured, semi-structured and unstructured sources.

The material present in this chapter is taken from [168].

12.1 Introduction

In the last few years, the number and the size of available data sources have dramatically increased, and most of them (i.e., roughly 80%) are unstructured [203, 185]. These facts are rapidly changing the scientific and technological approach of the information system research field [101, 392, 390, 11, 497, 154]. As a consequence, issues successfully addressed in the past must be re-considered and re-investigated. One of them is certainly the derivation of interschema properties (i.e., *intensional* relationships between concepts represented in different data sources [526], like synonymies, homonymies, hyponymies, overlappings, subschema similarities). This topic has been widely studied in the past [558, 102]; however, the proposed approaches generally considered structured or, at most, semi-structured sources. Furthermore, the number of involved sources, for which most of past approaches were targeted to, was very small, if compared with a typical current source interaction and cooperation scenario.

Interschema property derivation is not just one of the many topics to re-investigate in information systems cooperation field. Actually, it represents the bottom line of other issues: for instance, the knowledge of interschema properties is necessary for source integration, the construction of data warehouses and data lakes, data analytics, and so forth.

Here, we propose a novel approach to uniformly perform the extraction of interschema properties from structured, semi-structured and unstructured sources. Our approach has been specifically conceived having in mind two peculiarities, namely: *(i)* the capability of handling unstructured sources; *(ii)* the lightweightness, making it capable of managing a huge number of data sources.

As for the capability of handling unstructured sources, our approach is provided with a preliminary step capable of "structuring" unstructured sources, i.e., of (at least partially) deriving a structure for them. This is possible because it assumes that each unstructured source (e.g., a video, an audio, an image, a text) has associated a list of keywords describing it. The "structuring" process is based exactly on these keywords. This is a main contribution of our approach, which, generally speaking, allows the unstructured sources to be uniformly handled as the structured and the semi-structured ones. With regard to this aspect, some clarifications of what we mean with the terms "structured" and "semi-structured" sources are in order. In particular, we use these terms as they are generally adopted in databases and information systems research field. Here, a structured source consists of some concepts, each having a precise set of attributes and relationships with other concepts of the source (e.g., a relational database). A semi-structured source has similar characteristics, but the set of attributes and relationships characterizing a given concept is handled in a more flexible fashion (e.g., XML document).

Unstructured sources are videos, audios, images or texts, where we do not generally have a conceptual representation of their concepts, along with the corresponding properties and relationships. However, they are generally provided with a set of keywords, denoting their main features. The purpose of our approach for "structuring" unstructured sources is exactly the derivation of the relationships existing among the concepts represented by the keywords associated with unstructured sources. If we are capable of performing this task, unstructured sources can be handled similarly to structured and semi-structured ones. Furthermore, their analysis and management could benefit from the wide amount of results found in the past for structured and semi-structured sources. Finally, the integration, cooperation and simultaneous querying of structured, semi-structured and unstructured sources are possible.

Our approach also differs from other ones previously presented in related research fields and that could be in principle extended to address the problem we are considering in this paper. Think, for instance, of ontologies. We could link each available keyword to an ontology and use this last one as the "infrastructure" through which establishing the relationships among the keywords, once these last have been linked to it. This approach is certainly valid, but it needs a support ontology. As a consequence, it can be employed only in those application fields for which an ontology exists and only if all the involved information sources can be mapped onto a unique ontology. If only some of the involved unstructured sources can be referred to an ontology and/or some of them can be mapped onto another ontology and/or, finally, some of them cannot be referred to any ontology, this way of proceeding cannot be adopted. From this point of view, our approach is more general because it can be applied in all cases, independently of the presence of none, one or more ontologies, which the unstructured sources can be referred to. It only needs a thesaurus. If there exists a specific thesaurus for the scenario which the unstructured source into examination belongs to, then it uses this thesaurus. Otherwise, it can still work with a general-purpose thesaurus, like BabelNet [498]. Clearly, if the unstructured sources are specific of a certain field, the availability of a specific thesaurus can help to obtain a better accuracy. However, if this kind of thesaurus is not available, a general-purpose one is sufficient to proceed even if, in this case, accuracy could be lower.

As for the lightweightness of our approach, we observe that, in a big data scenario a new proposed approach must take scalability into account [426, 423]. As a matter of fact, the sources interacting in every task are always very numerous and large (think, for instance, of a data lake constructed to support data analytics in an organization) and the time allowed for each transaction is very limited (think, for instance, of streaming applications). As a consequence, even approaches considered very scalable in the past (such as DIKE [528], MOMIS [92], and Cupid [451]) are not adequate anymore. The tests performed to evaluate our approach and described in Section 12.4 confirm that it is really capable of satisfying the lightweightness requirement without sacrificing, if not to a very small extent, result accuracy.

This chapter is organized as follows: in Section 12.2, we examine related literature. In Section 12.3.1, we introduce a source representation model that we exploit in our tasks. In Section 12.3.2, we show our approach for the construction of a "structured representation" of unstructured data sources. In Section 12.3.3, we present our interschema property derivation approach. Finally, in Section 12.4, we present some experiments that we performed to test our approach.

12.2 Related Literature

Schema matching is one of the most investigated topics in past database research. The first schema matching approaches proposed by researchers were manual and operated only on structured databases. Subsequently, researchers proposed semi-automatic or automatic schema matching approaches capable of handling both structured and semi-structured data sources. With the advent of big data, unstructured sources are becoming more and more frequent and important.

A process preliminary to schema matching is data profiling. It is necessary when the metadata available for schema matching are not sufficient. Data profiling aims at discovering metadata starting from a data source [10, 11, 497, 103]. Generally speaking, data profiling activities encompass ad-hoc approaches, such as selecting random subsets of the data or formulating aggregation queries, the systematic inference of structural information and statistics of a dataset using dedicated tools, and the discovery of inclusion and functional dependencies [131, 132, 154, 155, 592].

After having introduced this preliminary task, we can move to the analysis of schema matching approaches available in the literature. These lasts were thought to consider several kinds of heterogeneity; the most relevant of them are lexicographic, structural and semantic ones. The first deals with names and terms; the second considers type formats, data representation models and structural relationships among concepts; the third regards the meaning of involved data. In the following, an overview of several approaches to perform schema matching from the beginning to the present day.

In [125], an approach to transform structured documents by leveraging schema graph matching is proposed. In particular, an XML schema to map each structured document is defined; for this purpose, some XSLT scripts are automatically generated. In [451], Cupid, a system for deriving interschema properties among heterogeneous sources, is proposed. Cupid leverages two different matchings, namely the *structure* and the *linguistic* ones. In [92], MOMIS, a system supporting querying and information source integration in a semi-automatic fashion, is presented. MOMIS implements a clustering procedure for the extraction of interschema properties. DIKE and XIKE [528, 227, 527], as well as the approaches described in [159, 241], also belong to this generation. An overview of this generation of schema matching approaches can be found in [558, 102].

More recent approaches, which significantly differ from the classical ones, are based on probabilistic methods, applied to networks of schemas [348]. They allow the definition of network-level integrity constraints for matching, as well as the anal-

ysis of query/click logs [248, 495], specifying the class of desired user-based schema matching.

In [42], an XML-based schema matching approach conceived to operate on large-scale schemas is presented. This approach leverages Prufer sequences. It performs a two-step activity; during the former step it parses XML schemas in schema trees; during the latter one, it exploits Label Prufer Sequences (LPS) to capture schema tree semantic information. In [506], SMART, a Schema Matching Analyzer and Reconciliation Tool, designed for the detection and the subsequent reconciliation of matching inconsistencies, is proposed. SMART is semi-automatic because it requires the intervention of an expert for the validation of results. In [462], the authors propose an approach to determine the semantic similarity of terms using the knowledge present in the search history logs from Google. For this purpose, they exploit four techniques that evaluate: *(i)* frequent co-occurrences of terms in search patterns; *(ii)* relationships between search patterns; *(iii)* outlier coincidence on search patterns; *(iv)* forecasting comparisons. In [47], a framework for the management of a data lake through the corresponding metadata is proposed. This framework leverages schema matching techniques to identify similarities between the attributes of different datasets. These techniques consider both schemas (specifically, attribute types and dependencies) and instances (specifically, attribute values) [102]. The framework integrates different schema matching approaches proposed in the last years, like graph matching, usage-based matching, document content similarity detection and document link similarity detection. [471] proposes an instance-based approach to find 1-1 schema matches. It combines the semantics provided by Google and regular expressions. It does not work well in a scenario where sources are very heterogeneous and data are stored in their raw way. Another instance-based approach is presented in [352]. It faces the heterogeneity of the different schemas by leveraging an ad-hoc mapping language.

Most schema matching approaches based on similarities often filter out unnecessary matchings and information [536] in such a way as to operate easier and faster.

As we have seen in this overview, schema matching has been widely investigated in the past for very heterogeneous scenarios, and very different approaches have been adopted to reach disparate goals. Among all these approaches, ours is characterized by the following features: *(i)* it has been specifically conceived to handle also unstructured sources; *(ii)* it has been designed to be scalable and, therefore, it is lightweight; *(iii)* it is automatic; *(iv)* in spite of these two last features, it presents a good accuracy, as we will see in Section 12.4.

On the other hand, the representation mechanisms of unstructured sources (basically texts) are mainly based on two strategies, namely analysis of contents and

analysis of references [635]. The former infers a representation of a document from the corresponding content, whereas the latter focuses on relationships among documents. Clearly, our interest is mainly on the former strategy, because its objective is similar to the one of our approach.

The most basic approach to represent texts leverages Bags of Words (BOW) [78, 589]. In this case, machine learning techniques are used to identify the set of words that mostly characterizes a text [391, 418]. Some more sophisticated strategies are based on the extraction of sentences [261]. In this case, a text is mapped onto semantic spaces, such as WordNet or Wikipedia. Another strategy is Explicit Semantic Analysis (ESA) [284], which mixes BOW and document references. In ESA, the relatedness between documents is computed by extracting similarities between the concepts identified within them, thanks to the cross-references expressed therein.

An important model in the BOW context is word2vec [474, 475]. This model is based on neural networks. It constructs a vector space and associates each word of the text into examination with a vector in this space in such a way that words sharing common contexts have close corresponding vectors in the vector space. The word2vec model was extended to the doc2vec one [405], which exploits similarities and contextual information of each word to reduce the dimensionality of the vector space. Other approaches reach the same objective (i.e., dimensionality reduction) by means of Latent Semantic Analysis [379], which exploits matrix decomposition methods.

Word-based methods are currently flanked by concept-based ones. As an example, [577, 576] introduce the idea of Bag of Concepts, in place of Bag of Words. In this approach, concepts are generated by disregarding semantic similarities between words. Semantic similarities have been considered only recently [381].

Another relevant set of approaches use ontologies or, in general, external sources of semantics, to generate conceptual representations of documents by matching document terms with ontology concepts (see, for instance, [111, 357, 665, 40]). The performance of these approaches is strongly related to the quality of the adopted external sources. As a consequence, in these approaches, very specific domains can strongly benefit from the availability of dedicated ontologies.

The approaches examined above generally consider only texts; they do not operate with other forms of unstructured sources, such as videos. Furthermore, they terminate with the derivation of keywords or key concepts representing a source. In fact, none of them tries to go a step over, i.e., to define a certain "structure" for an unstructured source, which is one of the objectives of our approach.

An attempt to define a "structure" for an unstructured source can be found in [458]. This approach generates a rowset with n attributes, i.e., a tabular represen-

tation from unstructured data. A single rowset is a set of tuples and is equivalent to a relation in relational databases; logical associations may exist between rowsets, but these are not explicitly defined. The schema of a rowset may be defined on read. Transformation functions, possibly based on fuzzy logic, are used to properly read the complex unstructured data and map them on the rowset schema. These functions are also exploited to address the data variety issue, by means of an interface for the dataset extraction, which is unified and valid for all the sources. Different transformation functions can be used to map different unstructured data onto the same schema. The content of a rowset depends on the membership function associated with a fuzzy logic and on the possible constraints regarding it. However, data extraction is only one of the steps defined in [458], which develops a general data processing system based on an Extract, Process, and Store (EPS) paradigm.

From the above description, it appears evident that the approach of [458] shares several features with ours; in particular, the purpose of structuring unstructured data is common to both of them. However, the two approaches also present several differences. Indeed, for the structuring task, the approach of [458] strongly depends on user defined transformation functions and on rowset schemas (which are not automatically inferred from the sources). Now, the definition of both the functions and the schema may be difficult for complex sources. Furthermore, mapping more sources on the same schema requires a manual integration step, which, again, may be a difficult task when the number of involved sources is high. On the other hand, querying obtained data sources is particularly effective with the use of fuzzy techniques and the declarative U-SQL query language characterizing the approach of [458]. On the contrary, in our proposal, to perform the structuring of unstructured sources, we leverage network analysis, as well as lexical and string similarities, for automatically deriving a general and uniform schema of different unstructured sources. In fact, as we will see in the following, unstructured sources are "structured" by first representing them as a network, starting from a set of keywords associated with them; then, this structure is enriched by adding arcs that link nodes having lexical or string similarities even if they belong to different sources. As a consequence, it is possible to state that the approach presented in [458] is more effective and flexible in querying data lake contents, but it requires a more complex design phase, with a heavy human intervention, difficult to sustain in presence of numerous data sources. On the contrary, our approach simplifies the structuring phase, because it does not need a preliminary structure to be used as a model, and it does not require a human intervention. However, its querying capabilities are limited to the summarization of unstructured sources provided by the keywords representing

them. Therefore, in a certain sense, our approach and the one of [458] can be considered orthogonal.

12.3 Methods

12.3.1 A network-based model for uniformly representing structured, semi-structured and unstructured sources

In this section, we present a network-based model for uniformly representing data sources of different formats. In order to understand the peculiarities of our model, we assume to have a set DS of m data sources of interest possibly characterized by different data formats.

$$DS = \{D_1, D_2, \cdots, D_m\}$$

Each data source D_k has associated a rich set $\mathcal{M}_k$ of metadata. We indicate with $\mathcal{M}_{DS}$ the repository of the metadata of all the data sources of DS:

$$\mathcal{M}_{DS} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$$

Given the source D_k, in order to represent the information content stored in $\mathcal{M}_k$, our model starts from a notation typical of XML, JSON and many other semi-structured data models. According to this notation, Obj_k denotes the set of all the objects stored in $\mathcal{M}_k$. Obj_k consists of the union of three subsets:

$$Obj_k = Att_k \cup Smp_k \cup Cmp_k$$

where:

- Att_k denotes the set of the attributes of $\mathcal{M}_k$;
- Smp_k indicates the set of the simple elements of $\mathcal{M}_k$;
- Cmp_k represents the set of the complex elements of $\mathcal{M}_k$.

Here, the meaning of the terms "attribute", "simple element" and "complex element" is the one typical of semi-structured data models.

$\mathcal{M}_k$ can be also represented as a graph:

$$\mathcal{M}_k = \langle N_k, A_k \rangle$$

N_k is the set of the nodes of $\mathcal{M}_k$. There is a node n_{k_j} in N_k for each object o_{k_j} of Obj_k. According to the structure of Obj_k, N_k consists of the union of three subsets:

$$N_k = N_k^{Att} \cup N_k^{Smp} \cup N_k^{Cmp}$$

where N_k^{Att} (resp., N_k^{Smp}, N_k^{Cmp}) denotes the set of the nodes corresponding to Att_k (resp., Smp_k, Cmp_k). There is a biunivocal correspondence between a node of N_k and an object of Obj_k. Therefore, in the following, we will use these two terms interchangeably. Each node has associated a name that identifies it in the schema which the corresponding element or attribute belongs to.

Let x be a complex element of $\mathcal{M}_k$. We denote by Obj_x the set of the objects directly contained in x and by N_x^{Obj} the set of the corresponding nodes. Finally, let x be a simple element of $\mathcal{M}_k$. We indicate by Att_x the set of the attributes directly contained in x and by N_x^{Att} the set of the corresponding nodes.

A_k denotes the set of the arcs of $\mathcal{M}_k$. It consists of three subsets:

$$A_k = A'_k \cup A''_k \cup A'''_k$$

where:

- $A'_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k^{Cmp}, n_y \in N_{n_x}^{Obj}\}$; in other words, there is an arc in A'_k from a complex element of $\mathcal{M}_k$ to each object directly contained in it. L_{xy} represents the label of A'_k.
- $A''_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k^{Smp}, n_y \in N_{n_x}^{Att}\}$; in other words, there is an arc in A''_k from a simple element of $\mathcal{M}_k$ to each attribute directly contained in it. L_{xy} represents the label of A''_k.
- $A'''_k = \{(n_x, n_y, L_{xy}) | n_x \in N_k, n_y \in N_k, D_k$ is unstructured, $\sigma(n_x, n_y)$=true$\}$. Here, $\sigma(n_x, n_y)$ is a function that receives two nodes and returns true if there exists a similarity between n_x and n_y. For instance, $\sigma(n_x, n_y)$ could return true if the concepts represented by n_x and n_y are semantically similar or if the names identifying n_x and n_y in the corresponding schema present a high string similarity. L_{xy} represents the label of A'''_k.

As for the label L_{xy} associated with each arc, in the current version of this model, it is NULL for the arcs of A'_k and A''_k. However, we do not exclude that, in the future, enrichments of our model might lead us to use this field for storing some knowledge. Instead, L_{xy} has an important meaning for the arcs of A'''_k. In fact, as will be clear in Section 12.3.3, it is used to denote the strength of the correlation between n_x and n_y.

From an abstract point of view, there is a "fil rouge" linking the three subsets of A_k, which leads to the concept of homophily in Social Network Analysis. Indeed, A'_k, A''_k and A'''_k are the three possible ways to represent the links between a concept and its "direct homophiles", i.e., the other concepts that can contribute to define (at least partially) its meaning.

12.3.2 Structuring an unstructured source

Our network-based model for uniformly representing and handling data sources with disparate formats is perfectly fitted for semi-structured sources. Indeed, it is sufficient:

- deriving the metadata of the source (if not yet provided) by applying one of the several techniques and tools defined for this purpose w.r.t. the various kinds of format;
- defining a complex element to represent the source as a whole;
- introducing a complex element, a simple element and an attribute for each complex element, simple element and attribute present in the metaschema of the source;
- defining an arc of A'_k from the source to the root of the document;
- introducing an arc of A'_k or A''_k for each relationship existing between the objects composing the source metadata.

Clearly, our model is sufficiently powerful to represent structured data. Indeed, it is sufficient:

- deriving the E/R schema of the source (if not yet provided) by performing a classical database reverse engineering activity;
- defining a complex element to represent the source as a whole;
- introducing a complex element for each entity of the E/R schema and an attribute for each attribute of the schema;
- defining an arc of A'_k from the complex element corresponding to the source to each complex element associated with an entity of the E/R schema;
- introducing an arc of A''_k from an entity to each of its attributes;
- defining an arc of A'_k for each one-to-many relationship of the E/R schema; this arc is from the entity participating to the relationship with a maximum cardinality equal to 1 to the entity participating with a maximum cardinality equal to N;
- representing a many-to-many relationship without attributes as a pair of one-to-many relationships and, then, modeling them accordingly;
- representing a many-to-many relationship R with attributes that connects two entities E_1 and E_2 as an entity having the same attributes as R and linked to E_1 and E_2 by means of two one-to-many relationships; the new entity and the new relationships are then suitably modelled by applying the rules defined in the previous cases.

The highest modeling difficulty regards unstructured data because it is worth avoiding a flat representation consisting of a simple element for each keyword provided to denote the source content. As a matter of fact, this flat representation would make the reconciliation, and the next integration, of an unstructured source with the other semi-structured and structured sources of DS very difficult. This is a very challenging issue to address. In the following, we propose our approach to "structure" unstructured sources. It is in itself a major issue in the current information systems scenario and, at the same time, plays a key role to provide our interschema property derivation approach with the capability of operating on sources with disparate formats.

Our approach assumes that each unstructured source into consideration (e.g., a video, an audio, an image, a text) is provided with a list of keywords describing it[1]. They will play a key role, as will be clarified in the following. We observe that this assumption is not particularly strong or out of place. As a matter of fact, in the reality, most video, image or audio providers associate a list of keywords (sometimes, in the form of tags) with the contents they deliver. As for text, representing keywords can be also easily derived through suitable techniques, like TF-IDF [460].

Our approach consists of four phases, namely: *(1)* creation of nodes; *(2)* management of lexical similarities; *(3)* management of string similarities; *(4)* management of (temporary) duplicated arcs. We describe these phases below.

- **Phase 1: Creation of nodes.** During this phase, our approach creates a complex node representing the source as a whole and a simple node for each keyword[2]. Furthermore, it adds an arc of A'_k from the node associated with the source to any node corresponding to a keyword. Initially, there is no arc between two keywords. To determine the arcs to add, the next phases are necessary.
- **Phase 2: Management of lexical similarities.** During this phase, our approach handles lexical similarities. For this purpose, it leverages a suitable thesaurus. Taking the current trends into account, this thesaurus should be a multimedia one; for this purpose, in our experiments, we have adopted BabelNet [498]. In particular, our approach adds an arc of A''_k from the node n_{k_1}, corresponding to the keyword k_1, to the node n_{k_2}, corresponding to the keyword k_2, and vice versa,

[1] Here, we assume that the list is ordered and the order is the one in which the keywords appear in the list.

[2] Here and in the following, to make the presentation smoother, we use the term "complex node" to indicate a node belonging to N_k^{Cmp} and the term "simple node" to denote a node of N_k^{Smp}. Furthermore, we use the term "source" (resp., "keyword") to denote both the source (resp., a keyword) and the corresponding node associated with it.

if k_1 and k_2 have at least one common lemma[3] in the thesaurus. Furthermore, it transforms the nodes n_{k_1} and n_{k_2} from simple to complex. The new arcs have a label corresponding to the number of common lemmas for k_1 and k_2 in the thesaurus.

- **Phase 3: Management of string similarities.** During this phase, our approach derives string similarities and states that there exists a similarity between two keywords k_1 and k_2 if the string similarity degree $kd(k_1,k_2)$, computed by applying a suitable string similarity metric on k_1 and k_2, is "sufficiently high" (see below). In this case, it adds an arc of A_k''' from n_{k_1} to n_{k_2}, and vice versa. Both the arcs have $kd(k_1,k_2)$ as their label. We have chosen N-Grams [388] as string similarity metric because we have experimentally seen that it provides the best results in our context. In particular, we have selected bi-grams as the best trade-off between accuracy and costs. In fact, mono-grams would require a lower cost but they would also return a lower accuracy than bi-grams. By contrast, tri-grams would guarantee a very high accuracy but at the expense of the computational cost, which would be excessive. Again, if n_{k_1} and n_{k_2} are simple nodes, our approach transforms them into complex ones.

  Now, we illustrate in detail what "sufficiently high" means and how our approach operates. Let $KeySim$ be the set of the string similarities for each pair of keywords of the source into consideration. Each record in $KeySim$ has the form $\langle k_i,k_j,kd(k_i,k_j)\rangle$. Our approach first computes the maximum keyword similarity degree kd_{max} present in $KeySim$. Then, it examines each keyword similarity registered therein. Let $\langle k_1,k_2,kd(k_1,k_2)\rangle$ be one of these similarities. If $((kd(k_1,k_2) \geq th_k \cdot kd_{max})$ and $(kd(k_1,k_2) \geq th_{kmin}))$, which implies that the keyword similarity degree between k_1 and k_2 is among the highest ones in $KeySim$ and that, in any case, it is higher than or equal to a minimum threshold, then it concludes that there exists a similarity between n_{k_1} and n_{k_2}. We have experimentally set $th_k = 0.70$ and $th_{kmin} = 0.50$.

  Observe that the choice to consider string similarities, in particular the one to adopt N-Grams as the technique for detecting string similarities, makes our approach robust against misspellings possibly present in the keywords. In fact, as shown in [320], N-Grams is well suited to handle also this kind of error.
- **Phase 4: Management of (temporary) duplicated arcs.** This phase is devoted to handle the possible simultaneous presence of both lexical and string similarities for the same pair of keywords. Indeed, it may occur that, for a pair of nodes n_{k_1}

[3] Here, we use the term "lemma" according to the meaning it has in BabelNet [498]. Given a term, its lemmas are other objects (terms, emoticons, etc.) that contribute to specify its meaning.

and n_{k_2}, there are two arcs from n_{k_1} to n_{k_2} belonging to A_k''' and generated by both lexical and string similarities, and two arcs from n_{k_2} to n_{k_1}. In this case, the two arcs from n_{k_1} to n_{k_2} corresponding to these two forms of similarities, must be merged in only one arc, which has associated a label denoting both the number of common lemmas between k_1 and k_2 in BabelNet and the value of $kd(k_1, k_2)$. The same happens for the two arcs from n_{k_2} to n_{k_1}.

From this description, it emerges that, at the end of the four phases, given two nodes n_{k_1} and n_{k_2}, four cases may exist, namely:

1. There is no arc from n_{k_1} to n_{k_2}.
2. A pair of arcs derived from a lexical similarity links them. In this case, the two arcs actually coincide (also in their labels); therefore, one of them can be removed. Note that the choice of the arc to be removed has deep implications in the definition of the topology of the corresponding network. Indeed, one of the two nodes involved (i.e., the source node of the maintained arc) will be certainly a complex node, whereas the other one may be a simple node (if no other arc starts from it) or a complex node (if at least another arc, different from the removed one, starts from it). In turn, the topology of the network has implications in the nature and the quality of the interschema properties that can be extracted, as will be clear in Section 12.3.3. Therefore, it is appropriate that the choice of the arc to be removed is not random and that a clear rule guiding it is defined. The rule that we chose for our approach is the following: given a pair of arcs between two nodes n_{k_1}, corresponding to the keyword k_1, and n_{k_2}, corresponding to the keyword k_2, with k_1 preceding k_2 in the list of keywords associated with the source D_k, the arc from n_{k_1} to n_{k_2} is maintained and the one from n_{k_2} to n_{k_1} is removed.
3. A pair of arcs derived from a string similarity links them. As in the previous case, the two arcs coincide and one of them is removed. The policy adopted to determine the arc to remove is the same as the one followed in the previous case.
4. A pair of arcs derived from Phase 4 links them. As in the previous case, the two arcs coincide and one of them is removed.

Actually, arc labels introduced above are not necessary in our approach for the extraction of semantic relationships described in Section 12.3.3. However, we have decided to maintain them in our model because we aim at providing an approach to "structure" unstructured sources that is general and that may be adopted in several future applications, some of which could benefit from this information.

Moreover, we point out that, in the prototype implementing our approach, in order to increase its efficiency, we directly added only one arc, namely (n_{k_1}, n_{k_2}),

during Phases 2, 3 and 4, instead of adding two arcs and of removing one of them at the end of the four phases.

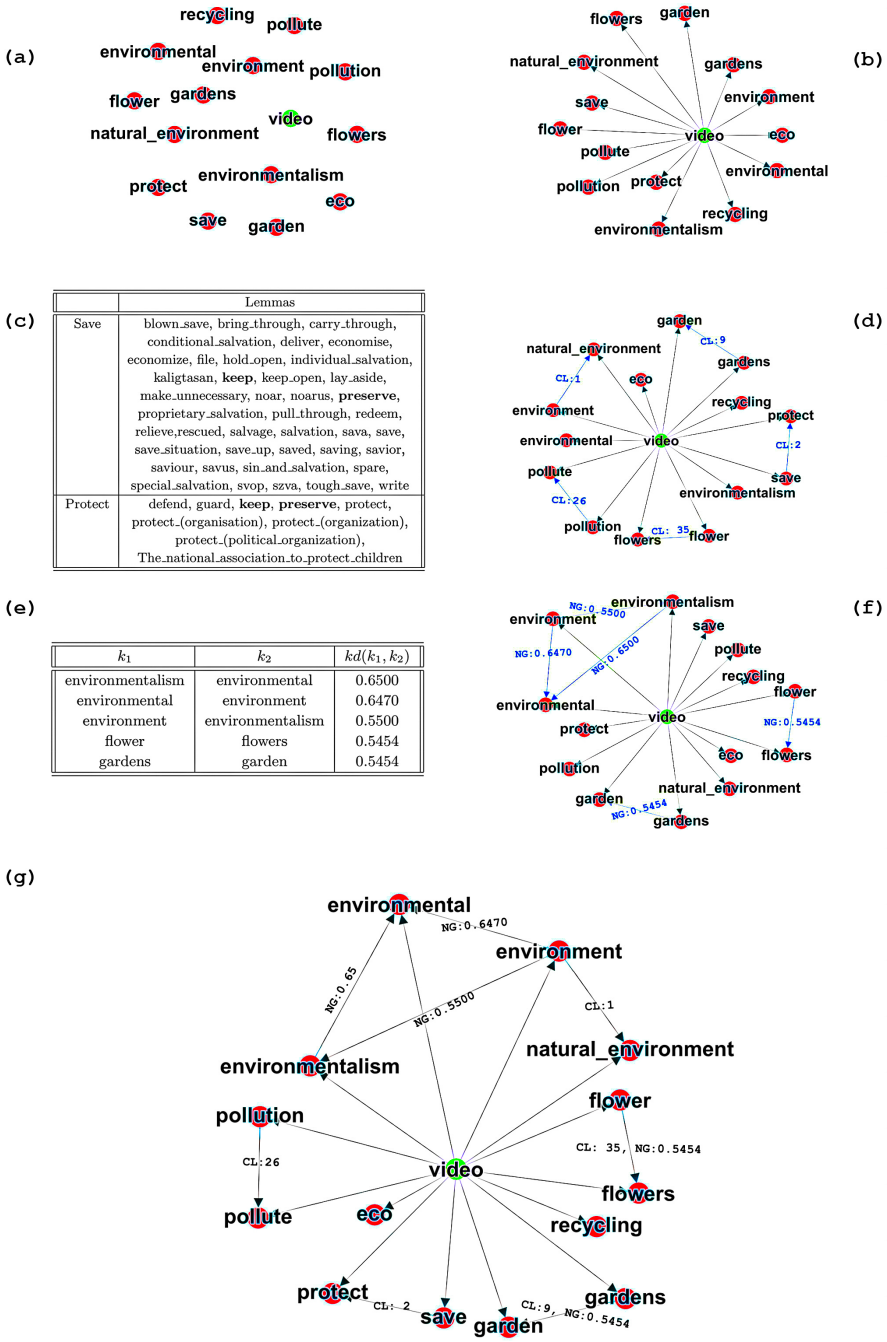
Finally, we want to show a possible example of how our approach is able to construct a "structured" representation of an unstructured source. In particular, the unstructured source into consideration is a video, which talks about environment and pollution. As we said before, for each unstructured source, our approach begins from a list of keywords representing its content. In order to keep our description simple and clear, in this example, we assume that our video has a limited number of keywords, namely the ones shown in Figure 12.1.

Our approach starts with Phase 1. As we can see in Figure 12.1(a), during this phase, it constructs a graph having a node for each keyword. A further node is added to represent the video as a whole; nodes representing keywords are colored in red, whereas the other one is colored in green. Following our strategy, in Figure 12.1(b), we added an arc from the node representing the whole video to each node associated with a keyword.

Now, Phase 2 starts. During this phase, our approach uses a thesaurus. In our example, we leveraged BabelNet. In particular, let k_1 and k_2 be two keywords of Figure 12.1(a) having at least one common lemma in BabelNet. An arc is added from the node n_{k_1}, associated with k_1, to the node n_{k_2}, associated with k_2, and vice versa. In Figure 12.1(c), we show two keywords ("Save" and "Protect") and the corresponding lemmas in BabelNet. Common lemmas (i.e., "keep" and "preserve") are in bold. Since "Save" and "Protect" have at least one common lemma, an arc is added between the corresponding nodes in Figure 12.1(d)[4]. This arc is highlighted in blue. Each arc has a label representing the number of common lemmas between the corresponding keywords in BabelNet.

After having examined lexical similarities, Phase 2 terminates and our approach proceeds with Phase 3, which leverages string similarities. In particular, let k_1 and k_2 be two keywords of Figure 12.1(a) having a string similarity degree higher than or equal to $th_k \cdot kd_{max}$ and, at the same time, higher than or equal to th_{kmin}. An arc is added from the node n_{k_1}, corresponding to k_1, to the node n_{k_2}, corresponding to k_2. In Figure 12.1(e), we report the pairs of keywords that satisfy this feature. In Figure 12.1(f), we added an arc for each pair of keywords of Figure 12.1(e). Here, to better highlight them, we have omitted the arcs constructed during Phase 2. Again, these arcs are highlighted in blue. Each arc has a label representing the string similarity degree (computed by means of N-Grams) between the corresponding keywords.

[4] Here, we have directly added only one arc between "Save" and "Protect", instead of adding two arcs and removing one of them later, after the four phases.



	Lemmas
Save	blown_save, bring_through, carry_through, conditional_salvation, deliver, economise, economize, file, hold_open, individual_salvation, kaligtasan, **keep**, keep_open, lay_aside, make_unnecessary, noar, noarus, **preserve**, proprietary_salvation, pull_through, redeem, relieve,rescued, salvage, salvation, sava, save, save_situation, save_up, saved, saving, savior, saviour, savus, sin_and_salvation, spare, special_salvation, svop, szva, tough_save, write
Protect	defend, guard, **keep**, **preserve**, protect, protect_(organisation), protect_(organization), protect_(political_organization), The_national_association_to_protect_children

k_1	k_2	$kd(k_1, k_2)$
environmentalism	environmental	0.6500
environmental	environment	0.6470
environment	environmentalism	0.5500
flower	flowers	0.5454
gardens	garden	0.5454

Fig. 12.1: Graphical representation of our approach to derive a "structure" for an unstructured source

Finally, in Figure 12.1(g), Phase 4 of our approach combines the arcs derived in Phases 2 and 3. In particular, it may happen that, for a pair of keywords (see, for instance, the keywords "garden" and "gardens"), two arcs have been generated, one in Figure 12.1(d) and one in Figure 12.1(f). In this case, in Figure 12.1(g), the two arcs are substituted by only one arc, representing both of them. The label of this arc reports the label of both the original ones.

12.3.3 Extracting interschema properties from different sources

We are now ready to illustrate our strategy for uniformly extracting interschema properties from structured, semi-structured and unstructured sources. Here, we assume that the content of the sources of interest is represented by means of the model described in Section 12.3.1, and that our approach to "structure" unstructured sources, described in Section 12.3.2, has been already applied on all unstructured sources.

Before delving into a detailed description of our approach, a discussion about the role played by source metadata, and about the consequences of this role, is in order. Indeed, as previously pointed out, our approach assumes that some metadata are available for each structured, semi-structured and unstructured source. This assumption is important because both our approach for structuring unstructured sources and our approach for extracting interschema properties use these metadata. It is, then, of outmost importance to analyze the possible issues (and the corresponding solutions) in obtaining good quality metadata, when they are not directly provided with the sources, and the impact that they have on the results returned by our approach.

Metadata generation received much attention in the literature. According to [38], metadata relative to a data source are currently generated by crawlers, by professional metadata creators, or, finally, by source creators. Generating metadata by means of automatic crawlers has great advantages, such as low cost and high efficiency; however, in some cases, the quality of generated metadata could be poor. In this context, it could be extremely useful the support of several mechanisms for controlling the quality of metadata, as well as the aid of metadata professionals, such as cataloguers and indexers; these are people who have had a formal training and are efficient in using metadata. Generally, they produce high-quality metadata. However, it has been observed that, in some cases, even metadata generated by professionals or by source authors may have poor quality and might hamper, rather than aid, the usage of the corresponding sources. This happens because most authors have little previous knowledge on metadata creation [38].

As pointed out in [531], the widespread adoption of several mechanisms for controlling the quality of metadata witnesses a strong awareness of the importance of having high-quality metadata at disposal. However, despite the relevance and the impact of metadata quality are universally recognized in the literature, there is no agreement yet on what metadata quality actually means. This implies, among the other things, the impossibility of defining systematic approaches to its automatic measurement and enhancement [637]. Metadata quality assurance should be verified simultaneously to metadata creation [532]. Indeed, a poor quality of metadata negatively affects the performance of systems using them and the overall user satisfaction. Quality assurance procedures are generally complemented by manual quality review and, if necessary, by the assistance of the technical staff during the process of metadata creation. Other mechanisms, such as metadata creation guidelines (sometimes embedded into the metadata creation system) and metadata generation tools, are on the rise.

The great relevance given to the metadata quality improvement is observed in the study presented in [364]. Here, the authors introduce a quality measure and analyze the metadata quality in the European context over the years. They observe that the metadata quality improves not only in new collections but also in the same collection over the years.

As pointed out in [531], in the metadata generation process, accuracy and consistency are prioritized over completeness, whereas the semantics of metadata elements is perceived to be less important. In principle, this might be an issue for our approach, since it strongly relies on semantics. The authors of [531] also point out that semantic overlaps and ambiguities are by far the two most critical factors. Actually, as our approach exploits thesauruses, string, and semantic similarities to relate keywords, these negative factors are significantly mitigated.

After this important discussion about the metadata of the involved sources, we can start our discussion about the derivation of interschema properties. We recall that, in the current big data scenario, any interschema property extraction strategy must be lightweight. For this reason, in our effort to define a new approach for this task, we avoided highly complex choices, such as the fixpoint computation characterizing DIKE [528, 527] and XIKE [227], or the clustering-based computation characterizing MOMIS [93], or, again, the wide range of parameter computation characterizing Cupid [451]. These choices, as well as most of the other ones present in the past approaches proposed for reconciling and integrating structured and semistructured data sources (e.g., the construction of a data warehouse) [558, 102], would certainly return very accurate results. However, their speed is incompatible with the one required in many current applications, which must allow the derivation of se-

mantic relationships "on-the-fly" from a very high number of data sources, most of which are unstructured, i.e., in a format not considered by classic approaches. As a consequence, our strategy must necessarily privilege quickness over accuracy even if, clearly, accuracy must be high. In Section 12.4, we will see if, and how, this issue has been addressed.

Our strategy consists of two phases; the former computes the semantic similarity degree of each pair of objects stored in the metadata of the involved sources. The latter derives semantic relationships between the same objects starting from the results returned by the former.

12.3.4 Semantic similarity degree computation

Our approach to semantic similarity degree computation consists of three steps, namely:

- basic similarity computation;
- standard similarity computation;
- refined similarity computation.

In the next subsections, we illustrate these three steps in detail.

12.3.4.1 Basic similarity computation

Basic similarities consider only lexicon (determined with the support of suitable thesauruses, such as BabelNet [498] and WordNet [477], and string similarity metrics, such as N-Grams [388]), and object types.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The basic similarity degree $bs(x_1, x_2)$ between x_1 and x_2 can be computed as:

$$bs(x_1, x_2) = \omega \cdot \sigma_L(x_1, x_2) + (1 - \omega) \cdot \sigma_T(x_1, x_2)$$

In other words, the basic similarity degree between x_1 and x_2 can be computed as a weighted mean of two components. The former, σ_L, returns their lexical similarity, whereas the latter, σ_T, specifies the similarity of their types. ω is a weight belonging to the real interval $[0,1]$ and used to tune the importance of σ_L w.r.t. σ_T. We have experimentally set ω to 0.90.

σ_L can be directly detected from a thesaurus. In our experiments, we used WordNet in the first beat, because it provides the similarity degree between the two objects, and BabelNet, when WordNet did not provide any result. Since this last thesaurus does not return the similarity degree of two objects that it considers similar, we coupled BabelNet with a suitable string similarity metric (in particular, N-Grams). This last is applied to the objects and the corresponding lemmas returned

by BabelNet; obtained results are, then, combined to compute the lacking similarity degree. Furthermore, in very specific application contexts, specialized thesauruses could be used.

σ_T is defined as follows:

$$\sigma_T = \begin{cases} 1 & \textit{if } (x_1 \in Cmp_1 \textit{ and } x_2 \in Cmp_2) \textit{ or } (x_1 \in Smp_1 \textit{ and } x_2 \in Smp_2) \textit{ or} \\ & (x_1 \in Att_1 \textit{ and } x_2 \in Att_2) \\ 0.5 & \textit{if } (x_1 \in Cmp_1 \textit{ and } x_2 \in Smp_2) \textit{ or } (x_1 \in Smp_1 \textit{ and } x_2 \in Cmp_2) \textit{ or} \\ & (x_1 \in Smp_1 \textit{ and } x_2 \in Att_2) \textit{ or } (x_1 \in Att_1 \textit{ and } x_2 \in Smp_2) \\ 0 & \textit{otherwise} \end{cases}$$

12.3.4.2 Standard similarity computation

Standard similarities take both basic similarities and the neighbors of the involved objects into account.

Let D_k be a source of the set DS of the sources of interest, let $\mathcal{M}_k = \langle N_k, A_k \rangle$ be the corresponding set of metadata, let Obj_k be the set of the objects of $\mathcal{M}_k$. The set $nbh(x)$ of the neighbors of an object $x \in Obj_k$ is defined as:

$$nbh(x) = \left\{ y | y \in Obj_k, (n_x, n_y) \in A_k \right\}$$

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The standard similarity degree $ss(x_1, x_2)$ between x_1 and x_2 can be computed as follows:

- If both $nbh(x_1) = \emptyset$ and $nbh(x_2) = \emptyset$, then $ss(x_1, x_2) = bs(x_1, x_2)$ [5].
- If either $nbh(x_1) = \emptyset$ and $nbh(x_2) \neq \emptyset$ or $nbh(x_2) = \emptyset$ and $nbh(x_1) \neq \emptyset$, then $ss(x_1, x_2) = f_p \cdot bs(x_1, x_2)$. Here, f_p is a factor, whose possible values belong to the real interval $[0, 1]$, which "penalizes" the value obtained for basic similarities. Indeed, these are the only similarities that we can compute and, therefore, we must base our standard similarity computation on them. However, we must consider that the sets of neighbors of x_1 and x_2 have different features, because one of them is empty and the other one is not empty, and this fact must be taken into account. We have experimentally set $f_p = 0.85$.
- In all the other cases, i.e., if $x_1 \in (Smp_1 \cup Cmp_1)$ and $x_2 \in (Smp_2 \cup Cmp_2)$, then $ss(x_1, x_2)$ can be computed as follows:
  1. $nbh(x_1)$ and $nbh(x_2)$ are determined.

[5] For instance, this happens when both x_1 and x_2 are attributes; indeed, the nodes corresponding to attributes do not have outgoing arcs.

2. A bipartite graph, whose nodes are the ones of $nbh(x_1)$ and $nbh(x_2)$, is constructed.
3. For each pair (p,q), such that $p \in nbh(x_1)$ and $q \in nbh(x_2)$, an arc is added in the bipartite graph; the weight of this arc is set to $bs(p,q)$.
4. The maximum weight matching is computed on this bipartite graph. Let A_M be the set of the returned arcs. Then:

$$ss(x_1,x_2) = \begin{cases} \frac{2 \cdot \sum_{(p,q) \in A_M} bs(p,q)}{|nbh(x_1)|+|nbh(x_2)|} & \textit{if neither } D_1 \textit{ nor } D_2 \textit{ are unstructured} \\ \\ \frac{2 \cdot \sum_{(p,q) \in A_M} bs(p,q)}{2 \cdot min(|nbh(x_1)|,|nbh(x_2)|)} & \textit{otherwise} \end{cases}$$

In this formula, if neither D_1 nor D_2 are unstructured, $ss(x_1,x_2)$ returns the value of an objective function that takes into account how many nodes of $nbh(x_1)$ and $nbh(x_2)$ are linked by basic similarity relationships and how strong these relationships are. Furthermore, the objective function penalizes the presence of dangling nodes, i.e., nodes of $nbh(x_1)$ or $nbh(x_2)$ that do not participate to the maximum weight matching.

If D_1 and/or D_2 are unstructured, then it is necessary to consider that, even if our approach performed a "structuring" task, its final structure is limited, if compared with the rich structure characterizing the other kinds of source. As a consequence, the sets of neighbors of the nodes belonging to unstructured sources are generally much smaller than the ones characterizing the other kinds of source. Therefore, in this case, using the same objective function adopted when neither D_1 nor D_2 are unstructured would not take this important feature into account, and the overall result would be biased. To address this issue, if D_1 and/or D_2 are unstructured, in the denominator of $ss(x_1,x_2)$ we consider the minimum size between $|nbh(x_1)|$ and $|nbh(x_2)|$, clearly multiplied by 2 to indicate the maximum number of nodes that could be linked by a similarity relationship in this situation.

12.3.4.3 Refined similarity computation

Refined similarities are based on standard similarities (for simple and complex objects), basic similarities (for attributes) and object neighbors.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The refined similarity degree $rs(x_1,x_2)$ between x_1 and x_2 can be computed as follows:

- If $nbh(x_1) = \emptyset$ and/or $nbh(x_2) = \emptyset$, then $rs(x_1,x_2) = ss(x_1,x_2)$.

- Otherwise, if $x_1 \in (Smp_1 \cup Cmp_1)$ and $x_2 \in (Smp_2 \cup Cmp_2)$, then $rs(x_1, x_2)$ is obtained by applying the same four steps described for $ss(x_1, x_2)$ with the only difference that, in Step 3, the weight of the arc (p, q), such that $p \in nbh(x_1)$ and $q \in nbh(x_2)$, is set to $ss(p, q)$, and no more to $bs(p, q)$. In other words, while standard similarity computation leverages basic similarities, refined similarity computation is based on standard similarities.

Clearly, from a theoretical point of view, it would be possible to perform other refinement steps. In this case, at the i^{th} refinement step, the similarities would be computed starting from the ones obtained at the $(i-1)^{th}$ step, by setting these last ones as the weights of the arcs of the bipartite graph. However, the advantages in accuracy that these further refinement steps could produce do not justify the computational costs introduced by them (see Section 12.4), especially in an agile and lightweight context, such as the one characterizing the big data scenario.

12.3.5 Semantic relationship detection

The derivation of semantic relationships among the objects of the sources of DS represents the second phase of our strategy. It takes the refined semantic similarities among the objects of DS as input. The semantic relationships that it can return are the following:

- *Synonymies*: A synonymy between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have a high similarity degree, the same type (i.e., both of them are complex objects or simple objects or attributes) and (possibly) different names.
- *Type Conflicts*: A type conflict between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have a high similarity degree but different types.
- *Overlappings*: An overlapping exists between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ if they have (possibly) different names, the same type and an intermediate similarity degree, in such a way that they can be considered neither synonymous nor distinct.
- *Homonymies*: A homonymy between two objects $x_1 \in Obj_1$ and $x_2 \in Obj_2$ exists if they have the same name and the same type but a low similarity degree.

Let D_1 and D_2 be two sources, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the corresponding sets of metadata, let $x_1 \in Obj_1$ and $x_2 \in Obj_2$ be two objects belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Finally, let $RefSim_{12}$ be the set of refined similarities involving the objects of Obj_1 and Obj_2.

First, our approach computes the maximum refined similarity degree rs_{max} present in $RefSim_{12}$. Then, it examines each similarity $\langle x_1, x_2, rs(x_1, x_2)\rangle$ registered

in $RefSim_{12}$ and verifies if a semantic relationship exists between the corresponding objects as follows:

- If $\left(rs(x_1, x_2) \geq th_{Syn} \cdot rs_{max}\right)$ and $(rs(x_1, x_2) \geq th_{min})$, which implies that the refined similarity degree between x_1 and x_2 is among the highest ones in $RefSim_{12}$ and, in any case, higher than or equal to a minimum threshold, then:
  - if x_1 and x_2 have the same type, it is possible to conclude that a synonymy exists between them;
  - if x_1 and x_2 have different types, it is possible to conclude that a type conflict exists between them.
- If $\left(rs(x_1, x_2) < th_{Syn} \cdot rs_{max}\right)$ and $(rs(x_1, x_2) \geq th_{Ov} \cdot rs_{max})$ and $(rs(x_1, x_2) \geq th_{min})$, which implies that the refined similarity degree between x_1 and x_2 is higher than or equal to a minimum threshold, it is not among the highest ones in $RefSim_{12}$, but it is significant, then:
  - if x_1 and x_2 have the same type, it is possible to conclude that an overlapping exists between them.
- If $(rs(x_1, x_2) < th_{Hom} \cdot rs_{max})$ and $(rs(x_1, x_2) < th_{max})$, which implies that the refined similarity degree between x_1 and x_2 is among the lowest ones in $RefSim_{12}$ and, in any case, lower than a maximum threshold, then:
  - if x_1 and x_2 have the same name and the same type, it is possible to conclude that a homonymy exists between them.

Here, th_{Syn}, th_{min}, th_{Ov}, th_{Hom} and th_{max} have been experimentally set to 0.85, 0.50, 0.65, 0.25 and 0.15, respectively.

As pointed out in the Introduction, the knowledge of interschema properties is very relevant for several applications, for instance source integration, source querying, data warehouse and/or data lake construction, data analytics, and so forth. As an example, as far as source integration is concerned:

- If a synonymy exists between $x_1 \in Obj_1$ and $x_2 \in Obj_2$, then x_1 and x_2 must be merged in a unique object, when the integrated schema is constructed.
- If a homonymy exists between x_1 and x_2, then it is necessary to change the name of x_1 and/or x_2, when the integrated schema is constructed.
- If an overlapping exists between x_1 an x_2, then it is necessary to restructure the corresponding portion of network. Specifically, a node x_{12}, representing the "common part" of x_1 and x_2, is added to the network. Furthermore, each pair of arcs (x_1, x_T) and (x_2, x_T), starting from x_1 and x_2 and having the same target x_T, is substituted by a unique arc (x_{12}, x_T). Finally, an arc from x_1 to x_{12} and another arc from x_2 to x_{12} are added to the network.

- If a type conflict exists between x_1 and x_2, then it is necessary to change the type of x_1 and/or x_2 in such a way as to transform the type conflict into a synonymy. Then, it is necessary to handle this last relationship by applying the corresponding integration rule seen above.

The way of proceeding described above can be extended to the detection of hyponymies. In particular, the extension already proposed in [525] for structured and semi-structured data can be probably adapted to this scenario. We plan to investigate this issue in the future. Finally, an analogous way of proceeding can be performed when querying or other activities must be carried out on a set of sources of interest.

Example

In this section, we provide an example of the behavior of our approach to the extraction of semantic relationships. To fully illustrate its potentialities, we derive these relationships between objects belonging to an unstructured source and a semi-structured one.

The unstructured source is a video. The corresponding keywords are reported in Table 12.1. Its "structured" representation, in our network-based model, obtained after the application of the approach described in Section 12.3.2, is reported in Figure 12.2. The semi-structured source is a JSON file whose structure is shown in Figure 12.3. Its representation in our network-based model is reported in Figure 12.4.

Keywords
video,reuse,flower,easy,tips,plastic,simple,environment,pollution,garbage, wave,recycle, reduce,pollute,help,natural_environment,educational,green, environment_awareness,bike,life, environmentalism, planet,earth,climate,clime, save,nature,environmental,gardens,power,recycling, garden, protect,flowers,eco,fine_particle,o_3, atmospheric_condition,ocean,metropolis,weather,spot,waving,aurora

Table 12.1: Keywords of the unstructured source of our interest

By applying the first phase of our approach we obtained the refined semantic similarity degrees between all the possible pairs of nodes (n_U, n_S), such that n_U belongs to the unstructured source and n_S belongs to the semi-structured one. To give an idea of these similarity degrees, in Figure 12.5, we report their distribution in a semi-logarithmic scale. From the analysis of this figure, we can observe that a very few numbers of pairs have a significant similarity degree, which could make them eligible to be selected for synonymies, type conflicts and overlappings. At a first

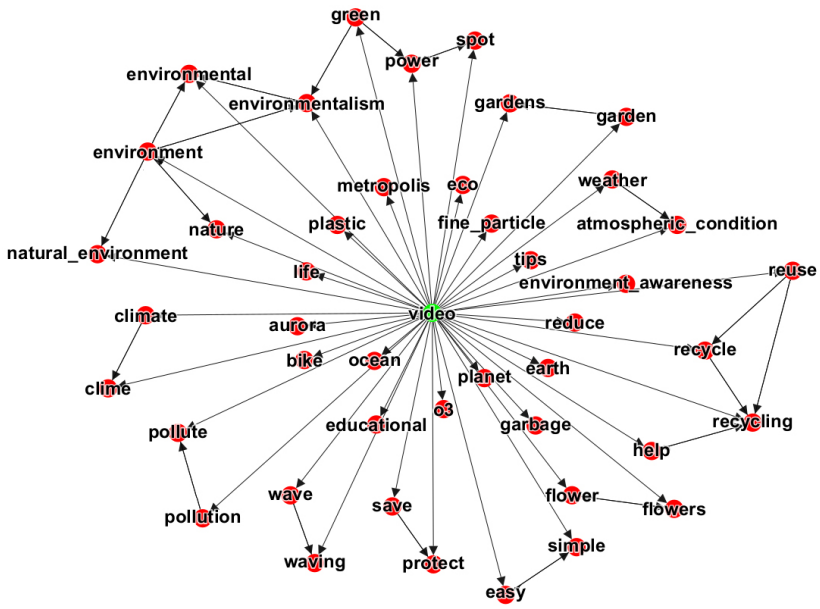


Fig. 12.2: Representation of the unstructured source of our interest through our network-based model

glance, this trend appeared correct and intuitive, even if this conclusion had to be confirmed or rejected by a much deeper analysis (see below).

By applying the second phase of our approach, we obtained the synonymies, the type conflicts and the overlappings reported in Tables 12.2 - 12.4. Instead, as for this pair of sources, we found no homonymies.

Semi-Structured Source Node	*Unstructured Source Node*
climate	*climate*
climate	*clime*

Table 12.2: Derived synonymies between objects of the two sources of interest

Semi-Structured Source Node	*Unstructured Source Node*
*pm*10	*fine_particle*
ozone	o_3

Table 12.3: Derived type conflicts between objects of the two sources of interest

We asked a human expert to validate these results. At the end of this task, he reported the following considerations:

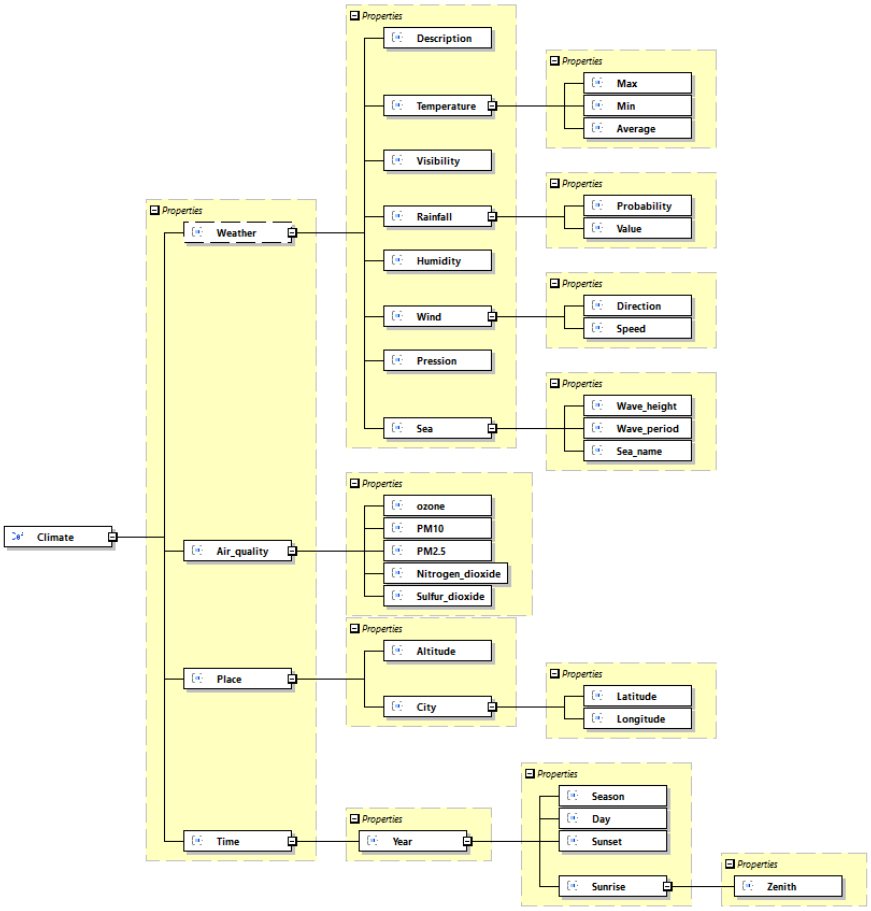


Fig. 12.3: Structure of the JSON file associated with the semi-structured source of our interest

Semi-Structured Source Node	*Unstructured Source Node*
sea	*ocean*
city	*metropolis*
sunrise	*aurora*
place	*spot*
wind	*tips*
sulfur_dioxide	*garbage*
weather	*clime*

Table 12.4: Derived overlappings between objects of the two sources of interest

- The synonymies provided by our approach are correct. No further synonymy can be manually found in the two considered sources.
- The type conflicts provided by our approach are correct. No further type conflict can be manually found in the two sources.

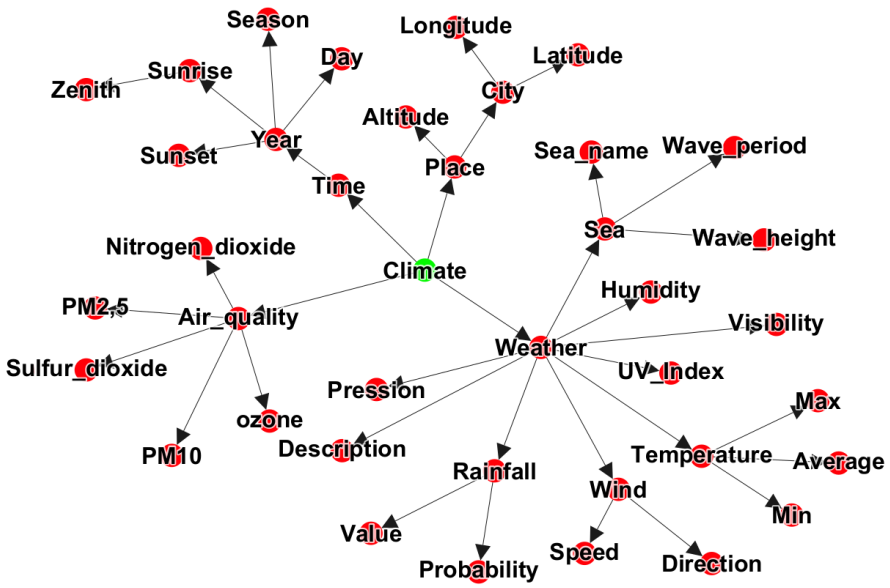


Fig. 12.4: Representation, in our network-based model, of the semi-structured source of our interest

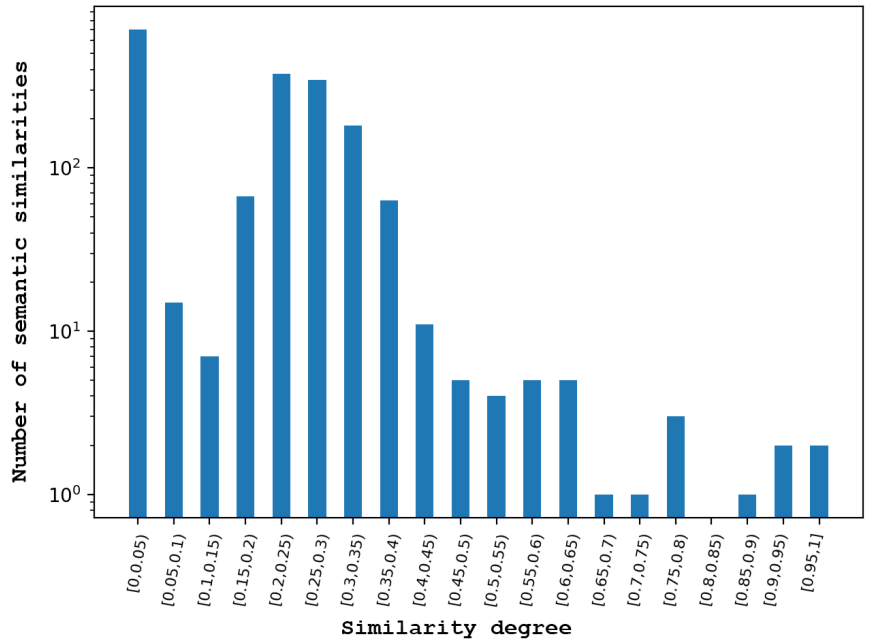


Fig. 12.5: Distribution, in a semi-logarithmic scale, of the values of the semantic similarity degrees of the objects belonging to the two sources of interest

- The overlappings provided by our approach are correct, except for the one linking “wind” and “tips”, which actually represents two different concepts. A very interesting overlapping found by our approach is the one between “sulfur_dioxide” and “garbage”, in that, even if they represent two seemingly dif-

ferent concepts, both of them denote harmful substances. Some further overlappings could be manually found in the two sources into consideration (for instance, the one between "climate" and "environment"), even if they are semantically weak, and considering them as overlappings or as distinct concepts is subjective.

12.4 Results

Our test campaign had four main purposes, namely: *(i)* evaluating the performance of our interschema property derivation approach when applied to the scenario for which it was thought, *(ii)* evaluating the pros and the cons of this approach w.r.t. analogous ones thought for structured and semi-structured sources, *(iii)* evaluating its scalability, *(iv)* evaluating the role of our approach for structuring unstructured sources, and *(v) evaluating the effectiveness and efficiency of our approach*. We describe these five experiments in the next subsections.

12.4.1 Overall performances of our approach

To perform our experiments, we constructed a set *DS* of data sources consisting of 2 structured sources, 4 semi-structured ones (2 of which were XML sources and 2 were JSON ones), and 4 unstructured ones (2 of which were books and 2 were videos). All these sources stored data about environment and pollution. To describe unstructured sources, we considered a list of keywords for each of them. These keywords were derived from Google Books, for books, and from YouTube, for videos. The interested reader can find the schemas, in case of structured and semi-structured sources, and the keywords, in case of unstructured sources, at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`. The password to type is "`za.12&;lq74:#`". A summary of the size of these sources is reported in Table 12.5.

Data Source	*Size (order)*
Structured Sources	Gigabytes
Semi-structured Sources	Gigabytes (2 sources), Hundreds of Gigabytes (2 further sources)
Unstructured (books)	Megabytes
Unstructured (videos)	Gigabytes

Table 12.5: Size of the sources involved in the tests

It could appear that taking only 10 sources is excessively limited. However, we made this choice because we wanted to fully analyze the behavior and the performance of our approach and, as it will be clear below, this requires the human inter-

vention for verifying obtained results. This intervention would have become much more difficult with a higher number of sources to examine. At the same time, our test set is fully scalable. As a consequence, an interested reader, starting from the data sources provided at the address `http://daisy.dii.univpm.it/dl/datasets/dl1`, can construct a data set with a much higher number of sources, if necessary.

For our experiments, we used a server equipped with an Intel I7 Dual Core 5500U processor and 16 GB of RAM with the Ubuntu 16.04.3 operating system. Clearly, the capabilities of this server were limited. However, they were adequate for the (small) data set DS we have chosen to use in our tests.

As the first task of our experiment, we represented the metadata of all the sources by means of the data model described in Section 12.3.1. Then, we applied the approach described in Section 12.3.2 to (at least partially) "structure" the unstructured sources of our test data set. Finally, we extracted semantic relationships existing between all the possible pairs of objects belonging to our test sources. After this, we asked the human expert to examine all the possible pairs of our test sources and to indicate us the semantic relationships that, in his opinion, existed among the corresponding objects.

At this point, we were able to evaluate the correctness and the completeness of our approach by measuring the classical parameters adopted in the literature for this purpose, i.e., Precision, Recall, F-Measure and Overall [652].

Precision is a measure of correctness. It is defined as:

$$Precision = \frac{|TP|}{|TP|+|FP|}$$

where TP are the true positives (i.e., semantic relationships detected by our approach and confirmed by the human expert), whereas FP are the false positives (i.e., semantic relationships proposed by our approach but not confirmed by our expert).

Recall is a measure of completeness. It is defined as:

$$Recall = \frac{|TP|}{|TP|+|FN|}$$

where FN are the false negatives (i.e., semantic relationships detected by the human expert that our system was unable to find).

F-Measure is the harmonic mean of Precision and Recall. It is defined as:

$$F\text{-}Measure = 2 \cdot \frac{Pecision \cdot Recall}{Precision + Recall}$$

Overall measures the post-match effort needed for adding false negatives and removing false positives from the set of matchings returned by the system to evaluate. It is defined as:

$$Overall = Recall \cdot (2 - \frac{1}{Precision})$$

Property	*Precision*	*Recall*	*F-Measure*	*Overall*
Synonymies	0.82	0.87	0.84	0.68
Overlappings	0.77	0.69	0.73	0.48
Type Conflicts	0.78	0.73	0.75	0.52
Homonymies	0.95	0.92	0.93	0.87

Table 12.6: Precision, Recall, F-Measure and Overall of our approach

Precision, Recall and F-Measure fall within the interval $[0,1]$, whereas Overall ranges between $-\infty$ and 1; the higher Precision, Recall, F-Measure and Overall, the better the performance of the evaluated approach.

In Table 12.6, we report obtained results. From the analysis of this table, we can observe that, although our approach has been designed with the intent of privileging quickness and lightweightness over accuracy, for the reasons explained in the Introduction, its performance, in terms of correctness and completeness, is extremely satisfying.

We also point out that the values reported in Table 12.6 are those obtained by applying the threshold values reported in Section 12.3.3. These are the ones guaranteeing the best tradeoff between Precision and Recall and, consequently, the best values of F-Measure and Overall.

Interestingly, if, in a given application context, a user must privilege correctness (resp., completeness) over completeness (resp., correctness), it is sufficient to increase (resp., decrease) the values of th_{min} and to decrease (resp., increase) the values of th_{Ov} and th_{max}.

12.4.2 Evaluation of the pros and the cons of our approach

In order to provide a quantitative evaluation of the pros and the cons of our interschema property extraction approach w.r.t. the past ones thought for structured and semi-structured sources[6] [558, 102], we compared our approach with XIKE [227]. Indeed, in [227], XIKE was already compared with several other systems having the same purposes (namely, Autoplex, COMA, Cupid, LSD, GLUE, SemInt, Similarity Flooding) and it was shown that it obtained comparable or better results.

First, we evaluated Precision, Recall, F-Measure and Overall of our approach and XIKE. Clearly, since this last system (as well as all the other ones mentioned above) did not handle unstructured data sources, we had to limit ourselves to consider only

[6] Actually, to the best of our knowledge, no approach to uniformly extract interschema properties from structured, semi-structured and unstructured sources have been proposed in the past.

Application context	*Number of Schemas*	*Max depth*	*Average Number of nodes*	*Average Number of complex elements*
Biomedical Data	11	8	26	8
Project Management	3	4	40	7
Property Register	2	4	70	14
Industrial Companies	5	4	28	8
Universities	5	5	17	4
Airlines	2	4	13	4
Biological Data	5	8	327	60
Scientific Publications	2	6	18	9

Table 12.7: Characteristics of the sources adopted for evaluating our approach

structured or semi-structured sources. Furthermore, as performed in [227], we limited our attention to synonymies and homonymies.

In a first experiment, we considered the same sources adopted in [227] for evaluating the performance of XIKE. In particular, we considered sources relative to Biomedical Data, Project Management, Property Register, Industrial Companies, Universities, Airlines, Biological Data and Scientific Publications. According to what reported in [227], Biomedical Schemas have been derived from several sites; among them we cite `http://www.biomediator.org`[7]. Project Management, Property Register and Industrial Companies Schemas have been derived from Italian Central Governmental Office (ICGO) sources and are shown at the address `http://www.mat.unical.it/terracina/tests.html`. Universities Schemas have been downloaded from the address `http://anhai.cs.uiuc.edu/archive/domains/courses.html`[8]. Airlines Schemas have been found in [535]; Biological Schemas have been downloaded from the addresses `http://smi-web.stanford.edu/projects/helix/pubs/ismb02/schemas/`[9], `http://www.cs.toronto.edu/db/clio/data/GeneX/\_RDB-s.xsd`[10] and `http://www.genome.ad.jp/kegg/soap/v3.0/KEGG.wsdl`. Finally, Scientific Publications Schemas have been supplied by the authors of [411].

[7] Currently, this web address is no more available. However, the interested reader can find the corresponding source at the address `https://web.archive.org/web/20100412034606/http://www.biomediator.org/`

[8] Currently, this web address is no more available. However, the interested reader can find the corresponding source at the address `https://web.archive.org/web/20061212142107/http://anhai.cs.uiuc.edu/archive/domains/courses.html`

[9] Currently, this web address is no more available. However, the interested reader can find the corresponding source at the address `https://web.archive.org/web/20050314041246/http://smi-web.stanford.edu/projects/helix/pubs/ismb02/schemas/`

[10] Currently, this web address is no more available. However, the interested reader can find the corresponding source at the address `https://web.archive.org/web/20060718122245/http://www.cs.toronto.edu/db/clio/data/GeneX_RDB-s.xsd`

We considered 35 sources whose characteristics are reported in Table 12.7. The minimum, the maximum and the average number of concepts of our sources were 12, 829 and 79, respectively.

A summary of the size of tested sources is shown in Table 12.8.

Data Source	*Size (order)*
Biomedical Data	Between Gigabytes and Hundreds of Gigabytes
ICGO Databases	Between Hundreds of Gigabytes and Terabytes
Universities Data	Megabytes
Airlines Data	Gigabytes
Biological Data	Terabytes and more
Scientific Publication Data	Hundreds of Gigabytes

Table 12.8: Size of the sources involved in the tests

The number of synonymies (resp., homonymies) really present in these sources was 498 (resp, 66). The number of synonymies (resp., homonymies) returned by XIKE was 541 (resp, 76). Finally, the number of synonymies (resp., homonymies) returned by our system was 593 (resp., 84). By comparing real synonymies and homonymies with the ones returned by XIKE and our approach we computed Precision, Recall, F-Measure and Overall for these two systems. They are reported in Table 12.9.

From the analysis of this table we can observe that Precision, Recall, F-Measure and Overall are better in XIKE, even if those obtained by our approach are satisfying. This was expected because our approach has been designed to be lightweight and, therefore, it introduces some approximations. For instance, while XIKE considers the neighbors of many levels in the computation of the similarity degree of two objects, our approach considers only the neighbors of levels 1 and 2.

Until now, our experimental campaign highlighted the cons of our approach. To evidence and quantify the pros, we measured its response time and the one of XIKE when the number of involved concepts represented in the corresponding metadata to examine increases. Obtained results are reported in Figure 12.6.

From the analysis of this figure, it clearly emerges that, as for this aspect, our approach is much better than XIKE. Indeed, the difference in the computation time between it and XIKE is of various orders of magnitude and is such to make XIKE, and the other systems mentioned above, unsuitable to handle the number and the size of the data sources characterizing the current big data scenario.

With reference to this claim, we observe that, in this experiment, the response time is measured against the number of concepts in the source metaschema. As such, already a set of sources with 1500 concepts can be considered “large”. Indeed, it

System	*Precision*	*Recall*	*F-Measure*	*Overall*
XIKE (Synonymies)	0.92	0.90	0.91	0.82
XIKE (Homonymies)	0.87	0.95	0.91	0.81
Our approach (Synonymies)	0.84	0.87	0.85	0.70
Our approach (Homonymies)	0.79	0.92	0.85	0.68

Table 12.9: Precision, Recall, F-Measure and Overall of XIKE and our approach

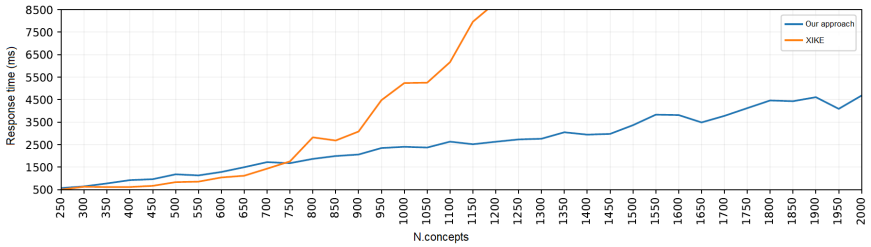


Fig. 12.6: Computation time of XIKE and our approach against the number of concepts to process

would correspond, for instance, to a set of E/R schemas consisting of about 1500 entities or a set of XML Schemas defining about 1500 different element types.

Furthermore, in this analysis, we must not forget that XIKE and the approaches mentioned above are not capable of handling unstructured data, which represents the second (and, for many verses, most important) peculiarity of our approach.

12.4.3 A deeper investigation on the scalability of our approach

The previous experiment represents a first confirmation of the quickness and the scalability of our approach. In this section, we aim at finding a further confirmation of this trend by considering a much more numerous and articulated set of sources and by comparing the accuracy and the response time of our approach, of XIKE [227] and DIKE [526]. This last is one of the approaches of its generation showing the highest accuracy, as witnessed by the comparison tests described in [558].

Clearly, if we want to compare these three approaches, the only way of proceeding is to consider structured sources because they are the only ones handled by DIKE. In particular, we considered the database schemas of Italian Central Government Offices (hereafter, ICGO). They include about 300 databases that are heterogeneous both in the data model and languages (e.g., hierarchical, network, relational), as well as in their structure and complexity, ranging from simple databases, with schemas including few objects, to very complex databases [528]. Information about the size of these data sources is provided in Table 12.8.

System	*Precision*	*Recall*	*F-Measure*	*Overall*
DIKE (Synonymies)	0.98	0.93	0.95	0.91
DIKE (Homonymies)	0.96	0.95	0.95	0.91
XIKE $u = 5$ (Synonymies)	0.96	0.91	0.93	0.87
XIKE $u = 5$ (Homonymies)	0.93	0.93	0.93	0.86
XIKE $u = 2$ (Synonymies)	0.84	0.86	0.85	0.70
XIKE $u = 2$ (Homonymies)	0.85	0.86	0.85	0.71
Our approach (Synonymies)	0.83	0.81	0.82	0.64
Our approach (Homonymies)	0.81	0.83	0.82	0.64

Table 12.10: Precision, Recall, F-Measure and Overall of DIKE, XIKE ($u = 5$, $u = 2$) and our approach

Observe that our approach, XIKE and DIKE are all based on graphs and on the computation of similarities of the neighbors of the involved objects. However, DIKE was thought for relatively small structured databases. As a consequence, when it computes the similarity of two objects belonging to different sources, it considers the similarity of their direct neighbors, the one of the neighbors of their direct neighbors, and so forth, until it terminates a fixpoint computation. In the worst case, the number of iterations of the fixpoint computation could be equal to the number of concepts of one of the involved sources. Clearly, performing such a high number of iterations allows DIKE to return very accurate results, but the required computation time is generally very high not only from the worst case computational complexity viewpoint, but also from the real computation time point of view. In XIKE, the possible number and dimension of data sources is higher than DIKE and they can be both structured and semi-structured. As a consequence, there is the need to limit the number of iterations of the fixpoint computation. For this reason, the concept of severity level is introduced. In particular, a user can choose a severity level u between 1 and n and the fixpoint computation is not completed but terminates after u iterations. The higher u the more accurate and slower XIKE. Our approach privileges lightweightness over accuracy for the reasons explained above. As a consequence, in this case, we limited the fixpoint computation to only 2 iterations. This could cause a reduction of accuracy but it is the only way to extend the approach of DIKE and XIKE also to a big data scenario.

Analogously to what happened in the previous section, in order to verify the theoretical conjectures explained above, we applied our approach, DIKE and XIKE (with $u = 5$ and, then, with $u = 2$) to ICGO databases. The obtained results are reported in Table 12.10.

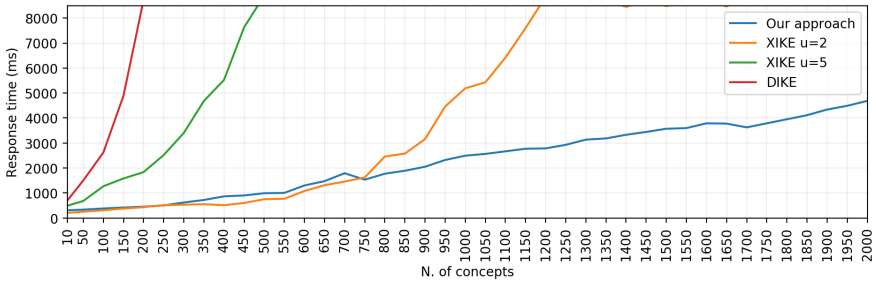


Fig. 12.7: Computation time of DIKE, XIKE ($u = 5$ and $u = 2$) and our approach against the number of concepts to process

The results of this table confirm our conjectures. DIKE provides a higher Precision, Recall, F-Measure and Overall than XIKE which, in turn, provides better results than our approach. Finally, XIKE, with a severity level equal to 5, provides better results than XIKE with a severity level equal to 2. The former tends to be comparable with the ones of DIKE; the latter tend to be comparable with the ones of our approach. This is in line with the fact that, when u tends to 5 the fixpoint computation tends to be complete; instead, when $u = 2$, it is substituted by only three iterations.

In any case, we would like to remark that, analogously to what happened in the previous experiment, the results obtained by our approach are still acceptable.

After having verified our conjectures about accuracy, we analyzed the ones regarding computation time. In particular, the average computation time of DIKE, XIKE (with $u = 5$ and $u = 2$) and our approach is reported in Figure 12.7.

From the analysis of this figure, it is easy to observe that the lower performance in terms of accuracy of our approach is largely balanced by an increased performance in terms of computation time. In a big data context, this aspect is mandatory. As a matter of fact, Figure 12.7 shows that DIKE and XIKE (especially when the severity level is high), even if very accurate, could not be applied in a big data scenario.

12.4.4 Evaluation of the role of our approach for structuring unstructured sources

In this section, we test the accuracy of our approach for structuring unstructured sources by comparing it with an alternative approach. For this purpose, we extended to unstructured data the clustering-based family of approaches defined for structured and semi-structured sources (see, for instance [43, 551]).

We performed this extension as follows: we considered the keywords associated with an unstructured source and used WordNet to derive a semantic distance coefficient for each pair of keywords. Then, we applied a clustering algorithm (specifically, Expectation Maximization [318]) to group keywords into homogeneous clusters. In

Property	*Precision*	*Recall*	*F-Measure*	*Overall*
Synonymies	0.76	0.82	0.79	0.56
Overlappings	0.69	0.65	0.67	0.36
Type Conflicts	0.72	0.64	0.68	0.39
Homonymies	0.91	0.88	0.89	0.79

Table 12.11: Precision, Recall, F-Measure and Overall of our approach when a clustering-based technique for structuring unstructured sources is applied

this way, we obtained a possible structure for unstructured sources. This structure is in line with what was done in the past for the clustering-based family of approaches, when they were applied on structured and semi-structured sources. This way of proceeding gave us the possibility to still apply the interschema property extraction approach defined in Section 12.3.3. In this case, we assumed that, given a keyword, the corresponding neighborhood consisted of the other keywords of its clusters.

We performed the same experiment described in Section 12.4.1 on the same sources. The only difference was the substitution of our approach for structuring unstructured sources with the clustering-based approach outlined above. The obtained results are shown in Table 12.11. Clearly, the differences between the performance reported in Tables 12.6 and 12.11 were due exclusively to the merits or demerits of our approach for structuring unstructured sources. From the analysis of this table we can observe that our approach presents a better performance than the corresponding clustering-based one described above. The differences are evident even if not extremely marked. For instance, we can observe a gain in Precision (resp., Recall, F-Measure, Overall) ranging from 4% (resp., 4%, 4%, 9%) to 10% (resp., 12%, 10%, 25%).

The results of this experiment, coupled with the theoretical analysis performed in the Introduction and mentioned above, allow us to conclude that our approach for structuring unstructured data is really capable of satisfying the requirements for which it was defined.

12.4.5 Effectiveness vs Efficiency

In any context characterized by a huge amount of data, such as those of interest to most current computer applications, efficiency plays a fundamental role. In fact, in these contexts, effectiveness (defined in terms of accuracy, precision, recall, etc.) is certainly an aspect to be taken into account, but it is not the only one and, in some cases, it may not be the main one. Indeed, if a high level of effectiveness can be achieved only at the price of adopting methods computationally incapable of han-

dling huge data, then it is necessary to resort to approaches that, while preserving an acceptable level of effectiveness, are able to guarantee a computation time compatible with the huge amount of data to process. From what we have seen in the previous subsections, our approach falls exactly in this case. In fact, it may be extremely useful in all those cases in which it is necessary to obtain interschema properties, extracted from huge amounts of data, to be used in other applications, such as querying, integration, data lake and data warehouse construction, knowledge extraction, etc. In all these cases, although our approach is not paramount as far as effectiveness is concerned, it continues to return acceptable results and is able to complete its tasks. By contrast, the approaches of the previous generations examined above, which can give better results in terms of effectiveness, are not able to complete their tasks in a reasonable amount of time.

In the scenario described above, our approach presents another interesting feature as it is able to extract interschema properties from unstructured data. In this feature, it differs from the ones presented in the past. Therefore, it is extremely interesting to investigate the effectiveness/efficiency of our approach with regard to this kind of data source. In fact, all the experiments proposed above have shown that our approach is the only one, among those analyzed, able to operate with the sizes characterizing the current data sources. On the other side, a great number of these sources are unstructured. Therefore, analyzing the efficiency and effectiveness of our approach when it works with huge unstructured sources is compulsory.

In this analysis, there are two important points to consider. The first concerns the fact that our approach assumes that the keywords representing each unstructured source are already known. If these keywords were unknown, it would be necessary to extract them. In this case, if the extraction task requires an excessive effort, for instance of some orders of magnitude higher than the subsequent extraction of interschema properties, our approach would become inefficient, and therefore not usable, in all those cases in which the keywords of the unstructured sources are not known a priori. The second point concerns the performance of our approach in terms of effectiveness, compared to a naive approach that considers only the basic similarities between keywords (see Section 12.3.4.1). Indeed, this last approach would presumably be more efficient than ours.

To address both these points we conducted the following experiment. We selected four popular approaches to text/information extraction, namely RAKE (Rapid Automatic Keyword Extraction) [570], LDA (Latent Dirichlet Allocation) [113], YAKE! (Yet Another Keyword Extractor) [150] and TopicRank [124], and applied them to the unstructured data sources used in the experiments in Section 12.4.1. Each of these approaches returned its own set of keywords for each source. Let $\mathcal{K}^R$ (resp.,

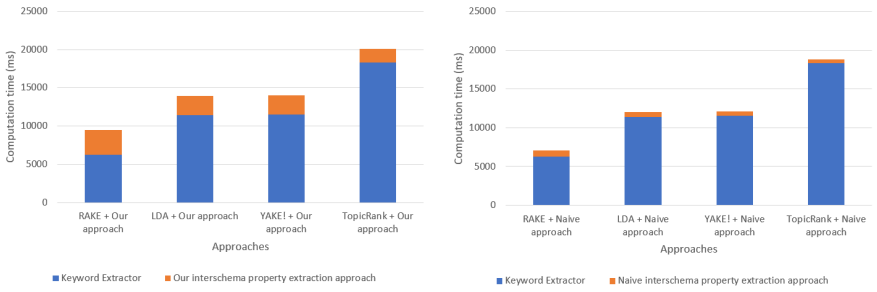


Fig. 12.8: Computation time of RAKE, LDA, YAKE! and TopicRank coupled with our interschema property extraction approach and a naive one considering only basic similarities

$\mathcal{K}^L$, $\mathcal{K}^Y$ and $\mathcal{K}^T$) be the set of the sets of keywords returned by RAKE (resp., LDA, YAKE! and TopicRank) when applied to the unstructured sources considered in our tests. We applied our interschema property extraction approach, as well as the naive one based only on basic similarities, on the sets of the keywords of $\mathcal{K}^R$ (resp., $\mathcal{K}^L$, $\mathcal{K}^Y$ and $\mathcal{K}^T$). The computation times characterizing the eight overall approaches under consideration are shown in Figure 12.8, while the approaches' average Precision, Recall, F-Measure and Overall are shown in Table 12.12.

In our opinion, the results reported in Figure 12.8 and Table 12.12 are very important and encouraging. In fact, they tell us that, in case of unstructured sources without associated keywords, the keyword computation requires a longer time, but of a comparable order of magnitude, than the interschema property extraction task. Therefore, the possible preliminary detection of the keywords does not change the conclusions emerged from the analysis of Figures 12.6 and 12.7, i.e., that our approach is the only one that can be adopted in presence of huge data sources. At the same time, the adoption of our approach, which, as far as the examination of neighborhoods is concerned, is a compromise between DIKE and XIKE (which consider all possible neighborhoods) and the naive approach (which considers only the immediate neighborhoods), guarantees an effectiveness certainly lesser than the one of DIKE and XIKE, but much greater than the one of the naive approach.

Therefore, our approach appears to be the best compromise between the ones of the past generation, having a very high effectiveness but an unacceptable efficiency, and a naive one, having a slightly higher efficiency but a much lower effectiveness than our approach.

Property	*Precision*	*Recall*	*F-Measure*	*Overall*
Synonymies (RAKE + our approach)	0.80	0.83	0.82	0.62
Overlappings (RAKE + our approach)	0.74	0.65	0.69	0.42
Type Conflicts (RAKE + our approach)	0.75	0.71	0.73	0.47
Homonymies (RAKE + our approach)	0.92	0.89	0.91	0.81
Synonymies (RAKE + naive approach)	0.68	0.70	0.69	0.37
Overlappings (RAKE + naive approach)	0.63	0.62	0.63	0.26
Type Conflicts (RAKE + naive approach)	0.63	0.59	0.61	0.24
Homonymies (RAKE + naive approach)	0.81	0.77	0.79	0.59
Synonymies (LDA + our approach)	0.81	0.88	0.84	0.67
Overlappings (LDA + our approach)	0.78	0.68	0.73	0.49
Type Conflicts (LDA + our approach)	0.77	0.74	0.76	0.52
Homonymies (LDA + our approach)	0.96	0.90	0.93	0.86
Synonymies (LDA + naive approach)	0.68	0.75	0.71	0.40
Overlappings (LDA + naive approach)	0.65	0.57	0.61	0.26
Type Conflicts (LDA + naive approach)	0.66	0.63	0.65	0.31
Homonymies (LDA + naive approach)	0.84	0.77	0.80	0.62
Synonymies (YAKE! + our approach)	0.83	0.85	0.84	0.68
Overlappings (YAKE! + our approach)	0.76	0.70	0.73	0.48
Type Conflicts (YAKE! + our approach)	0.80	0.71	0.75	0.53
Homonymies (YAKE! + our approach)	0.92	0.90	0.91	0.82
Synonymies (YAKE! + naive approach)	0.70	0.74	0.72	0.42
Overlappings (YAKE! + naive approach)	0.64	0.57	0.60	0.25
Type Conflicts (YAKE! + naive approach)	0.67	0.58	0.62	0.29
Homonymies (YAKE! + naive approach)	0.78	0.80	0.79	0.57
Synonymies (TopicRank + our approach)	0.84	0.89	0.86	0.72
Overlappings (TopicRank + our approach)	0.79	0.70	0.74	0.51
Type Conflicts (TopicRank + our approach)	0.79	0.74	0.76	0.54
Homonymies (TopicRank + our approach)	0.95	0.94	0.95	0.89
Synonymies (TopicRank + naive approach)	0.71	0.76	0.73	0.45
Overlappings (TopicRank + naive approach)	0.67	0.59	0.63	0.30
Type Conflicts (TopicRank + naive approach)	0.68	0.60	0.64	0.32
Homonymies (TopicRank + naive approach)	0.85	0.81	0.83	0.67

Table 12.12: Precision, Recall, F-Measure and Overall of RAKE, LDA, YAKE! and TopicRank coupled with our interschema property extraction approach and a naive one considering only basic similarities

Part V

Closing Remarks

This part of the thesis is devoted to point out some final remarks and describe the possible extensions of the approaches presented previously. In particular, in Chapter 14 we have a look at some future research, whereas in Chapter 13 we draw our conclusions.

13

Conclusions

In this thesis, we have proposed a complex network-based model capable of supporting the representation and management of heterogeneous scenarios. Specifically, we have proved the validity of our model through its application in the following six domains: *(i)* Social Networks, *(ii)* Internet of Things, *(iii)* Blockchain, *(iv)* Innovation Management, *(v)* Neurological Disorders, and *(vi)* Extraction of Semantic Relationships among Concepts in Data Lakes. In all these domains, we have shown that our approach is able to uniformly extract knowledge and support decision making.

As for Social Networks, we focused on Reddit and Yelp. In both cases, we employed the proposed model to represent the users in those social media and the relationships between them (e.g. friendship, writing a post for the same forum, reviewing the same business). Then, we investigated its properties and derived interesting knowledge from it. As for Reddit, we proposed an approach to detect user and subreddit stereotypes and verified the assortativity property for the co-posting network. Furthermore, we investigated the differences between Safe For Work (SFW) and Not Safe For Work (NSFW) posts, and verified the assortativity property of NSFW co-posting network. As for Yelp, we introduced the definition of k-bridge user and investigated the corresponding properties. Then, we defined three stereotypes of Yelp users, and we studied their characteristics and the profile of negative influencers.

As for the Internet of Things (IoT), we started from the Multiple Internet of Things (MIoT) paradigm, which models the IoT as a set of networks that can communicate with each other. In this way, it provides a complex network-based representation of the IoT. Thanks to it, we defined some extensions of the MIoT paradigm, along with approaches to address some open issues in the IoT scenario. Our contributions regard the following topics: *(i)* analysis and optimization of the communication between devices; *(ii)* evaluation of the reliability of these interactions; *(iii)* safeguard of privacy and security; *(iv)* anomaly detection.

As for Blockchain, we applied the complex network-based approach to investigate the user behavior during the cryptocurrency speculative bubble at the end of

2017 and the beginning of the 2018. We focused on Ethereum and described several knowledge patterns on the behavior of specific categories of users.

As for Innovation Management, we proposed a well-tailored centrality measure for evaluating patents and their citations. We also presented three possible applications of our measure.

As for Neurological Disorders, we modeled the ElectroEncephaloGram signals through a complex network in order to help experts investigating two neurological disorders, namely Mild Cognitive Impairment and Alzheimer's Disease.

As for the Extraction of Semantic Relationships among Concepts in Data Lakes, we have defined new models and paradigms for metadata representation and management in a data lake scenario. The proposed approach is able to define a structure for unstructured data and extract thematic views from heterogeneous data sources.

14

Future Works

14.1 Premise

This thesis should not be considered as an ending point. On the contrary, it represents a first evidence that complex networks and suitable approaches built on top of them, are capable of uniformly addressing different issues in heterogeneous scenarios. As an evidence of this claim, in the next sections, we present some examples of future works for the three main areas previously presented.

14.2 Social Networks

As stated before, Reddit is one of the few social networks that handles Not Safe For Work (NSFW) content in an explicit and well-structured way. The past literature has a very limited number of studies on this topic. In order to fill this gap, we could think to employ a complex network to model NSFW posts and comments on Reddit, and then extract their content. It would be interesting to study the most frequent text patterns present in NSFW posts and their utility (e.g. expressed by likes from the community, the sentiment of the pattern, etc.). These text patterns could describe the language adopted by users and how these last ones tend to interact with NSFW posts. Probably, text patterns could also unveil user communities with a similar way to deal with these particular posts.

In a similar perspective, we could think of extracting frequent and utility patterns from a set of comments and, then, build a content semantic network. The nodes of this network could represent comment lemmas, while the arcs could denote either the co-occurrence of two lemmas in a sentence or the semantic similarity between two lemmas. Once we have two content semantic networks extracted by two different sets of comments, we can compute the graph distance between them (e.g., through NetSimile [97]). Depending on the set of comments (they could come from two subreddits, forums, users, etc.), this approach could have several applications. For instance, starting from the graph distances between content semantic networks,

we can think of building a content-based or collaborative filtering recommender system, define new user communities, create new thematic forums (e.g., subreddits in Reddit) from the existing ones, and so on and so forth.

14.3 Internet of Things

In the Internet of Things (i.e., IoT) setting, one interesting future work concerns the definition of a framework to enable protection and automation. Indeed, we recall that the IoT is characterized by a large number of smart objects with several constraints/features, such as: *(i)* limited storage and computing capability; *(ii)* great dynamism, due to the high number of nodes that join or leave the IoT at any time; *(iii)* criticality and sensitiveness of used services and applications. In this scenario, the protection of objects and the possibility to guarantee them a certain autonomy represent two interesting issues to address. These two aspects are related to each other and the solution to face both of them should consider the distributed nature of IoT. Indeed, some approaches present in the literature propose frameworks with a centralized authority or empowered devices [608, 184], which does not respect the fully distributed nature of IoT.

In a distributed solution, each smart object should be able to build a complete representation of the other objects' behavior in the IoT. It should also be able to link a sequence of actions (defining a behavior) to each object. This requires the definition of an authentication mechanism to map each action (e.g., a transaction) to the object making it.

One possible authentication mechanism involves the use of the blockchain technology in the IoT as a shared and reliable environment among all objects [240, 544, 602]. However, the application of blockchain without a central or empowered actor in the IoT poses a lot of research challenges that must be considered. One issue regards the high computational power required for deploying a blockchain-based solution in an IoT context. Another issue, from the blockchain perspective, concerns the handling of the big volume of transactions generated by smart objects, which harms the scalability and environmental impact (in terms of energy consumption). A partial solution concerns the definition of lightweight blockchains for the IoT. Typically, these approaches work on the reduction of the information necessary to mine and validate transactions published in the ledger by proposing alternative consensus algorithms [239]. However, even the simple monitoring of the public ledger can be a heavy and expensive task for smart objects with minimal computation capability in presence of a very high volume of transactions.

For all these reasons, it could be interesting to develop a two-tier blockchain framework to increase the protection and autonomy of smart objects in the IoT. Following the intuition proposed in Chapter 7, we could consider smart objects as organized in communities. Hence, the first, local, blockchain tier could manage the trust measures of each smart object inside the community it belongs to, and adopt a solution leveraging both a lightweight blockchain and a validity window to control transaction volume. By splitting the overall set of objects into communities, we can control the size of the local blockchain in order to avoid excessive loads for smart objects. The second, global, blockchain tier could record aggregated data related to the individual communities (it could contain any information, such as trust and reputation between devices or also between communities). Once the data is aggregated and saved in the global tier, we can think of emptying out the local one, in order to save space, computational power and the device battery.

14.4 Blockchains

Anyone can participate to a blockchain network, and anyone can provide a specific service to other users. For example, if we consider a blockchain that manages smart contracts, some actors (called miners) maintain the blockchain network, while others (called exchanges) allow users to trade different cryptocurrencies. Some actors deal with auctions, others offer games or services, and so on and forth. Therefore, different actors can be identified in this scenario.

In some cases, there are online systems that provide a label to identify the class of the services provided by these actors in a blockchain network. One of the most known of these services is Etherscan[1], which is designed for Ethereum. Through it, the developer of a smart contract can publish the corresponding code and request verification. Etherscan performs such a task and provides a categorization of the corresponding user. However, it alone is not able to classify and give information about all addresses present in the Ethereum blockchain, but only of those submitted for verification.

Some researchers have studied the user classification in this domain [429, 729, 643], but they have not considered the social factor of the blockchain. Hence, it could be interesting to represent this scenario as a complex network and, then, extract both patterns and features useful for the classification of user behaviors.

Specifically, in order to define user behaviors in a certain time interval, first we could build a complex network representing the users involved in the blockchain and their transactions. Then, starting from this complex network, we compute a set

[1] https://etherscan.io/

of features for each user. These could be the number of incoming and outgoing arcs of the node corresponding to the user, the number of incoming and outgoing transactions, the amount of incoming and outgoing cryptocurrencies, the clustering coefficient, some centrality measures, and so forth. Of course, these features can change over time, and joining them together creates a multivariate time series which fully represents the user behaviors in a time interval.

Once we model the scenario in this way, we are actually dealing with a classification problem in which each element to classify and each available class are represented by multivariate time series. The solution of this problem is surely not straightforward and investigating it is certainly a challenging issue.

References

1. Thingful: A Search Engine for the Internet of Things. *https://thingful.net/*, 2017.
2. Mongodb. https://www.mongodb.org/, 2019.
3. The R Project for Statistical Computing. https://www.r-project.org/, 2019.
4. Concise Oxford Dictionary. *https://en.oxforddictionaries.com/*, 2020.
5. IPSO Alliance. *https://www.ipso-alliance.org/*, 2020.
6. Six stereotypes you follow on Instagram. https://www.kaindefoecommunications.com/new-england-social-media-marketing/6-stereotypes-you-follow-on-instagram/, 2020.
7. The Stereotypes of Facebook. https://www.ericsson.com/en/blog/2011/9/facebook-stereotypes-which-type-are-you, 2020.
8. A. Abbas, L. Zhang, and S.U. Khan. A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13, 2014. Elsevier.
9. M. Abe. Mix-networks on permutation networks. In *Proc. of the International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT'99)*, pages 258–273, Singapore, 1999. Springer.
10. Z. Abedjan, L. Golab, and F. Naumann. Profiling relational data: a survey. *The VLDB Journal*, 24(4):557–581, 2015. Springer.
11. Z. Abedjan, L. Golab, and F. Naumann. Data profiling: A tutorial. In *Proc. of the International Conference on Management of Data (MOD'17)*, pages 1747–1751, New York, NY, USA, 2017. ACM.
12. M. Abomhara and G.M. Koien. Security and privacy in the Internet of Things: Current status and open issues. In *Proc. of the International Conference on Privacy and Security in mMbile Systems (PRISMS'14)*, pages 1–8, Aalborg, Denmark, 2014. IEEE.
13. M. Abulaish, A. Kamal, and M.J. Zaki. A Survey of Figurative Language and Its Computational Detection in Online Social Networks. *ACM Transaction on the Web*, 14(1):3:1–3:52, 2020. ACM.
14. S. Achard and E. Bullmore. Efficiency and cost of economical brain functional networks. *PLoS Computational Biology*, 3(2):e17, 2007. Public Library of Science.
15. L. Adamic and E. Adar. Friends and Neighbors on the Web. *Social Networks*, 25(3):211–230, 2003. Elsevier.

16. A.K. Agarwal, A.P. Pelullo, and R.M. Merchant. "ld": the Word Most Correlated to Negative Online Hospital Reviews. *Journal of General Internal Medicine*, pages 1–2, 2019. Springer.
17. R. Agarwal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the International VLDB Conference (VLDB'94)*, pages 487–499, Santiago de Chile, Chile, 1994. Morgan Kaufmann.
18. R. Aggarwal, R. Gopal, A. Gupta, and H. Singh. Putting Money Where the Mouths Are: The Relation Between Venture Financing and Electronic Word-of-Mouth. *Information Systems Research*, 23(3):976–992, 2012. INFORMS.
19. M. Ahmadlou, A. Adeli, R. Bajo, and H. Adeli. Complexity of functional connectivity networks in mild cognitive impairment subjects during a working memory task. *Clinical Neurophysiology*, 125(4):694–702, 2014. Elsevier.
20. M. Ahmadlou and H. Adeli. Wavelet-synchronization methodology: a new approach for EEG-based diagnosis of ADHD. *Clinical EEG and Neuroscience*, 41(1):1–10, 2010. SAGE Publications.
21. M. Ahmadlou, H. Adeli, and A. Adeli. Improved visibility graph fractality with application for the diagnosis of autism spectrum disorder. *Physica A: Statistical Mechanics and its Applications*, 391(20):4720–4726, 2012. Elsevier.
22. M. Ahmed. Collective anomaly detection techniques for network traffic analysis. *Annals of Data Science*, 5(4):497–512, 2018. Springer.
23. M. Ahmed and A.N. Mahmood. Novel approach for network traffic pattern analysis using clustering-based collective anomaly detection. *Annals of Data Science*, 2(1):111–130, 2015. Springer.
24. M. Ahmed, A.N. Mahmood, and J. Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016. Elsevier.
25. M. Ahmed and A.S.S.M. Barkat Ullah. Infrequent pattern mining in smart healthcare environment using data summarization. *The Journal of Supercomputing*, 74(10):5041–5059, 2018. Springer.
26. Y.Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the International Conference on World Wide Web (WWW'07)*, pages 835–844, Banff, Alberta, Canada, 2007. ACM.
27. R. K. Ahuja. *Network Flows: Theory, Algorithms, and Applications*. Boston, MA, USA, 2017. Pearson Education.
28. N.Z. Aitzhan and D. Svetinovic. Security and privacy in decentralized energy trading through multi-signatures blockchain and anonymous messaging streams. *IEEE Transactions on Dependable and Secure Computing*, 15(5):840–852, 2018. IEEE.
29. S.A. Akar, S. Kara, F. Latifoğlu, and V. Bilgiç. Analysis of the Complexity Measures in the EEG of Schizophrenia Patients. *International Journal of Neural Systems*, 26(02):1650008, 2016. World Scientific.
30. L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Proc. of the Pacific-Asia Conference on Advances in Knowledge Discovery and*

Data Mining, (PAKDD'10) Part II, pages 410–421, Hyderabad, India, 2010. Lecture Notes in Computer Science, Springer.

31. L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015. Springer.
32. F. Al-Turjman. *Cognitive Sensors and IoT: Architecture, Deployment, and Data Delivery*. Boca Raton, Florida, USA, 2017. CRC Press.
33. F. Al-Turjman. Information-centric sensor networks for cognitive IoT: an overview. *Annals of Telecommunications*, 72(1-2):3–18, 2017. Springer.
34. F. Al-Turjman and S. Alturjman. Context-sensitive access in industrial internet of things (IIoT) healthcare applications. *IEEE Transactions on Industrial Informatics*, 14(6):2736–2744, 2018. IEEE.
35. F. Al-Turjman and I. Baali. Machine learning for wearable iot-based applications: A survey. *Transactions on Emerging Telecommunications Technologies*, page e3635, 2019. Wiley Online Library.
36. F. Al-Turjman, H. Zahmatkesh, and R. Shahroze. An overview of security and privacy in smart cities' IoT communications. *Transactions on Emerging Telecommunications Technologies*, page e3677, 2019. Wiley Online Library.
37. A. Al-Zoubi, J. Alqatawna, H. Faris, and M. A Hassonah. Spam profiles detection on social networks using computational intelligence methods: The effect of the lingual context. *Journal of Information Science*, 47(1):58–81, 2019. SAGE.
38. B.M. Albassuny. Automatic Metadata Generation Applications: a Survey Study. *International Journal of Metadata, Semantics and Ontologies*, 3(4):260–282, 2008. Inderscience Publishers.
39. A. Alchihabi, A. Dervis, E. Ever, and F. Al-Turjman. A generic framework for optimizing performance metrics by tuning parameters of clustering protocols in WSNs. *Wireless Networks*, 25(3):1031–1046, 2019. Springer.
40. Z. Aleksovski, M.C.A. Klein, W.T. Kate, and F. van Harmelen. Matching Unstructured Vocabularies Using a Background Ontology. In *Proc. of the International Conference on Knowledge Engineering and Knowledge Management (EKAW'06)*, pages 182–197, Prague, Czech Republic, 2006. Lecture Notes in Computer Science. Springer.
41. A. Alexandrov. Characteristics of single-item measures in Likert scale format. *The Electronic Journal of Business Research Methods*, 8(1):1–12, 2010.
42. A. Algergawy, E. Schallehn, and G. Saake. Improving XML schema matching performance using Pr. *Data & Knowledge Engineering*, 68(8):728–747, 2009. Elsevier.
43. S. P. Algur and P. Bhat. Web Video Object Mining: Expectation Maximization and Density Based Clustering of Web Video Metadata Objects. *International Journal of Information Engineering and Electronic Business*, 8(1):69, 2016. Modern Education and Computer Science Press.
44. S.A. Aljawarneh and R. Vangipuram. Garuda: Gaussian dissimilarity measure for feature representation and anomaly detection in internet of things. *The Journal of Supercomputing*, (11227):1–38, 2018. Springer US.

45. M. Aloqaily, S. Otoum, I. Al Ridhawi, and Y. Jararweh. An intrusion detection system for connected vehicles in smart cities. *Ad Hoc Networks*, 90:101842, 2019. Elsevier.
46. A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng. Fog computing for the Internet of Things: Security and privacy issues. *IEEE Internet Computing*, 21(2):34–42, 2017. IEEE.
47. A. Alserafi, A. Abello, O. Romero, and T. Calders. Towards information profiling: data lake content metadata management. In *Proc. of the International Conference on Data Mining Workshops (ICDMW'16)*, pages 178–185, Barcelona, Spain, 2016. IEEE.
48. M.D. Alshehri and F.K. Hussain. A fuzzy security protocol for trust management in the internet of things (Fuzzy-IoT). *Computing*, 101(7):791–818, 2019. Springer.
49. J. Alstott, M. Breakspear, P. Hagmann, L. Cammoun, and O. Sporns. Modeling the impact of lesions in the human brain. *PLoS Computational Biology*, 5(6):e1000408, 2009. Public Library of Science.
50. M. Amadeo, C. Campolo, A. Iera, and A. Molinaro. Named data networking for IoT: An architectural perspective. In *Proc. of the European Conference on Networks and Communications (EuCNC'2014)*, pages 1–5, Bologna, Italy, 2014. IEEE.
51. M. Amadeo, C. Campolo, J. Quevedo, D. Corujo, A. Molinaro, A. Iera, R. Aguiar, and A. Vasilakos. Information-centric networking for the internet of things: challenges and opportunities. *IEEE Network*, 30(2):92–100, 2016. IEEE.
52. F. Amato, V. Moscato, A. Picariello, and F. Piccialli. SOS: A multimedia recommender System for Online Social networks. *Future Generation Computer Systems*, 93:914–923, 2019. Elsevier.
53. J. Amezquita-Sanchez, A. Adeli, and H. Adeli. A new methodology for automated diagnosis of mild cognitive impairment (MCI) using magnetoencephalography (MEG). *Behavioural brain research*, 305:174–180, 2016. Elsevier.
54. B. Amiri, L. Hossain, J. W. Crawford, and R.T. Wigand. Community detection in complex networks: Multi–objective enhanced firefly algorithm. *Knowledge-Based Systems*, 46:1–11, 2013. Elsevier.
55. J. An, H. Kwak, O. Possega, and A. Jungherr. Political Discussions in Homogeneous and Cross-Cutting Communication Spaces. In *Proc. of the International Conference on Web and Social Media (ICWSM 2019)*, pages 68–79, Munich, Germany, 2019. AAAI.
56. K. Anand, J. Kumar, and K. Anand. Anomaly detection in online social network: A survey. In *Proc. of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT '17)*, pages 456–459, Coimbatore, India, 2017. IEEE.
57. K.E. Anderson. Ask me anything: what is Reddit? 2015. Emerald.
58. S. Angelidis and M. Lapata. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31, 2018. MIT Press.
59. N. Antonakakis, I. Chatziantoniou, and D. Gabauer. Cryptocurrency market contagion: Market uncertainty, market complexity, and dynamic portfolios. *Journal of International Financial Markets, Institutions and Money*, 61:37–51, 2019. Elsevier.
60. M. Antunes, D. Gomes, and R. L Aguiar. Towards IoT data classification through semantic features. *Future Generation Computer Systems*, 86:792–798, 2018. Elsevier.

61. A.P. Plageras and K.E. Psannis and C. Stergiou and H. Wang and B.B. Gupta. Efficient iot-based sensor big data collection–processing and analysis in smart buildings. *Future Generation Computer Systems*, 82:349–357, 2018. Elsevier.
62. G. Araniti, A. Orsino, L. Militano, L. Wang, and A. Iera. Context-aware information diffusion for alerting messages in 5G mobile social networks. *IEEE Internet of Things Journal*, 4(2):427–436, 2017. IEEE.
63. G. Araniti, A. Orsino, L. Militano, L. Wang, and A. Iera. Context-Aware Information Diffusion for Alerting Messages in 5G Mobile Social Networks. *IEEE Internet of Things Journal*, 4(2):427–436, 2017.
64. J.R. Arthur, D. Etzioni, and A.J. Schwartz. Characterizing extremely negative reviews of total joint arthroplasty practices and surgeons on yelp.com. *Arthroplasty Today*, 2019. Elsevier.
65. D. Artz and Y. Gil. A survey of trust in computer science and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.
66. C. Aslay, L.V.S. Lakshmanan, W. Lu, and X. Xiao. Influence maximization in online social networks. In *Proc. of the ACM International Conference on Web Search and Data Mining (WSDM'18)*, pages 775–776, Marina del Rey, CA, USA, 2018. ACM.
67. American Psychiatric Association, editor. *Diagnostic and statistical manual of mental disorders (5th ed.)*. 2013.
68. L. Atzori, A. Iera, and G. Morabito. The Internet of Things: A survey. *Computer networks*, 54(15):2787–2805, 2010. Elsevier.
69. L. Atzori, A. Iera, and G. Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010. Elsevier.
70. L. Atzori, A. Iera, and G. Morabito. SIoT: Giving a social structure to the Internet of Things. *IEEE Communications Letters*, 15(11):1193–1195, 2011. IEEE.
71. L. Atzori, A. Iera, and G. Morabito. From "smart objects" to "social objects": The next evolutionary step of the Internet of Things. *IEEE Communications Magazine*, 52(1):97–105, 2014. IEEE.
72. L. Atzori, A. Iera, and G. Morabito. From smart objects to social objects: The next evolutionary step of the internet of things. *IEEE Communications Magazine*, 52(1):97–105, 2014. IEEE.
73. L. Atzori, A. Iera, and G. Morabito. Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. *Ad Hoc Networks*, 56:122–140, 2017. Elsevier.
74. L. Atzori, A. Iera, G. Morabito, and M. Nitti. The Social Internet of Things (SIoT)– when social networks meet the Internet of Things: Concept, architecture and network characterization. *Computer networks*, 56(16):3594–3608, 2012. Elsevier.
75. S. Babar, P. Mahalle, A. Stango, N. Prasad, and R. Prasad. Proposed security model and threat taxonomy for the Internet of Things (IoT). In *Proc. of the International Conference on Network Security and Applications (CNSA'10)*, pages 420–429, Chennai, India, 2010. Springer.

76. D.A. Bader, S. Kintali, K. Madduri, and M. Mihail. Approximating betweenness centrality. In *Proc. of the International Workshop on Algorithms and Models for the Web-Graph (WAW'07)*, volume 4863, pages 124–137, San Diego, CA, USA, 2007. Springer.
77. C. Baek and M. Elbeck. Bitcoins as an investment or speculative vehicle? A first look. *Applied Economics Letters*, 22(1):30–34, 2015. Taylor & Francis.
78. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. 1999. Addison Wesley Longman.
79. P. Bajpai, A. K Sood, and R. J Enbody. The art of mapping IoT devices in networks. *Network Security*, 2018(4):8–15, 2018. Elsevier.
80. U.A.B.U.A. Bakar, H. Ghayvat, S.F. Hasanm, and S.C. Mukhopadhyay. *Activity and Anomaly Detection in Smart Home: A Survey*, pages 191–220. Cham, 2016. Springer International Publishing.
81. G. Baldassarre, P. Lo Giudice, L. Musarella, and D. Ursino. A paradigm for the cooperation of objects belonging to different IoTs. In *Proc. of the International Database Engineering & Applications Symposium (IDEAS 2018)*, pages 157–164, Villa San Giovanni, Italy, 2018. ACM.
82. G. Baldassarre, P. Lo Giudice, L. Musarella, and D. Ursino. The MIoT paradigm: main features and an "ad-hoc" crawler. *Future Generation Computer Systems*, 92:29–42, 2019. Elsevier.
83. F. Bao, R. Chen, and J. Guo. Scalable, adaptive and survivable trust management for community of interest based internet of things systems. In *Proc. of the International Symposium on Autonomous Decentralized Systems (ISADS'13)*, pages 1–7, Mexico City, Mexico, 2013. IEEE.
84. D. Barbera-Tomas, F. Jimenez-Saez, and I. Castello-Molina. Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks. *Research Policy*, 40(3):473–486, 2011. Elsevier.
85. M. Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):163–168, 2004. Springer.
86. R. Barthwal, S. Misra, and M.S. Obaidat. Finding overlapping communities in a complex network of social linkages and Internet of things. *The Journal of Supercomputing*, 66(3):1749–1772, 2013. Springer.
87. S. Basuroy, S. Chatterjee, and S.A. Ravid. How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, 67(4):103–117, 2003. SAGE Publications.
88. K. Bauman and A. Tuzhilin. Discovering contextual information from user reviews for recommendation purposes. In *Proc. of the International Workshop on New Trends in Content-Based Recommender Systems (CBRecSys @ RecSys 2014)*, pages 2–9, Foster City, CA, USA, 2014.
89. J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The pushshift Reddit dataset. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM'20)*, volume 14, pages 830–839, Atlanta, GA, USA, 2020. AAAI Press.

90. M. Behniafar, A.R. Nowroozi, and H.R. Shahriari. A survey of anomaly detection approaches in internet of things. *The ISC International Journal of Information Security*, 10(2):79–92, 2018. Iranian Society of Cryptology.
91. F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proc. of the ACM SIGCOMM Conference on Internet measurement*, pages 49–62, Chicago, IL, USA, 2009. ACM.
92. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.
93. S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. Semantic integration and query of heterogeneous information sources. *Data & Knowledge Engineering*, 36(3):215–249, 2001.
94. J. Berger, A.T. Sorensen, and S.J. Rasmussen. Positive effects of negative publicity: When negative reviews increase sales. *Marketing science*, 29(5):815–827, 2010. INFORMS.
95. M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. Foundations of Multidimensional Network Analysis. In *Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011)*, pages 485–489, Kaohsiung, Taiwan, 2011. IEEE.
96. M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web*, 16(5-6):567–593, 2013. Springer.
97. M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos. Netsimile: A scalable approach to size-independent network similarity. *arXiv preprint arXiv:1209.2684*, 2012.
98. M. Berlingerio, F. Pinelli, and F. Calabrese. Abacus: frequent pattern mining-based community discovery in multidimensional networks. *Data Mining and Knowledge Discovery*, 27(3):294–320, 2013. Springer.
99. J.B. Bernabe, J.L. Hernández, M.V. Moreno, and A.F.S. Gomez. Privacy-preserving security framework for a social-aware Internet of Things. In *Proc. of the International Conference on Ubiquitous Computing and Ambient Intelligence (UCAMI'14)*, pages 408–415, Belfast, Northern Ireland, UK, 2014. Springer.
100. J. Bernabé-Moreno, A. Tejeda-Lorente, C. Porcel, H. Fujita, and E. Herrera-Viedma. Quantifying the emotional impact of events on locations with social media. *Knowledge-Based Systems*, 146:44–57, 2018. Elsevier.
101. J. Bernabé-Moreno, A. Tejeda-Lorente, C. Porcel-Gallego, and E. Herrera-Viedma. Leveraging Localized Social Media Insights for Industry Early Warning Systems. *International Journal of Information Technology & Decision Making*, 17(01):357–385, 2018. World Scientific.
102. P.A. Bernstein, J. Madhavan, and E. Rahm. Generic Schema Matching, Ten Years Later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011.
103. L. Berti-Equille. Reinforcement learning for data preparation with active reward learning. In *Proc. of the International Conference on Internet Science (INSCI'19)*, pages 121–132, 2019. Springer.
104. D. Bertram. Likert scales. *Retrieved November*, 2:2013, 2007.

105. C. Besthorn, H. Sattel, C. Geiger-Kabisch, R. Zerfass, and H. Forstl. Parameters of EEG dimensional complexity in Alzheimer's disease. *Electroencephalography and clinical neurophysiology*, 95(2):84–89, 1995. Elsevier.

106. C. Besthorn, R. Zerfass, C. Geiger-Kabisch, H. Sattel, S. Daniel, U. Schreiter-Gasser, and H. Forstl. Discrimination of Alzheimer's disease and normal aging by EEG data. *Electroencephalography and clinical neurophysiology*, 103(2):241–248, 1997. Elsevier.

107. P.K. Bhanodia, A. Khamparia, B. Pandey, and S. Prajapat. Online social network analysis. In *Hidden Link Prediction in Stochastic Social Networks*, pages 50–63, 2019. IGI Global.

108. A.Q. Bhatti, M. Umer, S. H. Adil, M. Ebrahim, D. Nawaz, and F. Ahmed. Explicit Content Detection System: An Approach towards a Safe and Ethical Environment. *Applied Computational Intelligence and Soft Computing*, page 1463546, 2018. Hindawi.

109. A.K. Bhowmick, S. Suman, and B. Mitra. Effect of information propagation on business popularity: A case study on yelp. In *Proc. of the International Conference on Mobile Data Management (MDM'17)*, pages 11–20, Daejeon, South Korea, 2017. IEEE.

110. P.V. Bindu, P. Santhi Thilagam, and D. Ahuja. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior*, 73:568–582, 2017. Elsevier.

111. L. Bing, S. Jiang, W. Lam, Y. Zhang, and S. Jameel. Adaptive Concept Resolution for document representation and its applications in text mining. *Knowledge Based Systems*, 74:1–13, 2015.

112. B.M. Blau. Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance*, 41:493–499, 2017. Elsevier.

113. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. Microtone Publishing.

114. V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. IOP Publishing.

115. A. Blum, K. Ligett, and A. Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013. ACM.

116. Bela Bollobas. *Modern Graph Theory (Graduate Texts in Mathematics)*. Salmon Tower Building, New York City, United States, 2013. Springer.

117. P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972. Taylor & Francis.

118. G. Bonifazi, E. Corradini, D. Ursino, and L. Virgili. A Social Network Analysis based approach to investigate user behavior during a cryptocurrency speculative bubble. *Journal of Information Science*, Forthcoming. SAGE.

119. L. Bontemps, V.L. Cao, J. McDermott, and N. Le-Khac. Collective anomaly detection based on long short-term memory recurrent neural networks. In *Proc. of the International Conference on Future Data and Security Engineering (FDSE'16)*, pages 141–152, Can Tho City, Vietnam, 2016.

120. S. Borgatti and M. Everett. A graph-theoretic perspective on centrality. *Social networks*, 28(4):466–484, 2006. Elsevier.

121. S.P. Borgatti. Centrality and Network flow. *Social Networks*, 27(1):55–71, 2005. Elsevier B.V.

122. L. Bornmann and H. Daniel. What do citation counts measure? A review of studies on citing behavior. *Journal of documentation*, 64(1):45–80, 2008. Emerald Group Publishing Limited.

123. M. Bouabdellah, N. Kaabouch, F. El Bouanani, and H. Ben-Azza. Network layer attacks and countermeasures in cognitive radio networks: A survey. *Journal of Information Security and Applications*, 38:40–49, 2018. Elsevier.

124. A. Bougouin, F. Boudin, and B. Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In *Proc. of the International Joint Conference on Natural Language Processing (IJCNLP'13)*, pages 543–551, Nagoya, Japan, 2013. Asian Federation of Natural Language Processing.

125. A. Boukottaya and C. Vanoirbeek. Schema matching for transforming structured documents. In *Proc. of the ACM Symposium on Document Engineering (DocEng'05)*, pages 101–110, Bristol, United Kingdom, 2005. ACM.

126. P. Bourelos, G. Kousiouris, O. Voutyras, and T.A. Varvarigou. Heating schedule management approach through decentralized knowledge diffusion in the context of social internet of things. In *Proc. of the Panhellenic Conference on Informatics (PCI 2015)*, pages 103–108, Athens, Greece, 2015. ACM.

127. E. Bouri, C.K.M. Lau, B. Lucey, and D. Roubaud. Trading volume and the predictability of return and volatility in the cryptocurrency market. *Finance Research Letters*, 29:340–346, 2019. Elsevier.

128. U. Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001. Taylor & Francis.

129. U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008. Elsevier.

130. S. Breschi and F. Lissoni. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4):439–468, 2009. Oxford University Press.

131. B. Breve, L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese. Visualizing Dependencies during Incremental Discovery Processes. In *Proc. of the International Conference on Extending Database Technology (EDBT'20)*, 2020. CEUR-WS.

132. B. Breve, L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese. Dependency visualization in data stream profiling. *Big Data Research*, 25:100240, 2021. Elsevier.

133. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117, 1998.

134. F. Buccafurri, V.D. Foti, G. Lax, A. Nocera, and D. Ursino. Bridge Analysis in a Social Internetworking Scenario. *Information Sciences*, 224:1–18, 2013. Elsevier.

135. F. Buccafurri, G. Lax, A.Nocera, and D. Ursino. A system for extracting structural information from Social Network accounts. *Software Practice & Experience*, 45(9):1251–1275, 2015. John Wiley & Sons.

136. F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera. Comparing Twitter and Facebook user behavior: Privacy and other aspects. *Computers in Human Behavior*, 52:87–95, 2015. Elsevier.

137. F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera. Interest Assortativity in Twitter. In *Proc. of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016)*, pages 239–246, Rome, Italy, 2016. "SCITEPRESS – Science and Technology Publications, Lda".

138. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Supporting Information Spread in a Social Internetworking Scenario. *Post-Proceedings of the International Workshop on New Frontiers in Mining Complex Knowledge Patterns at ECML/PKDD 2012 (NFMCP 2012)*, 200–214. Lecture Notes in Artificial Intelligence, Springer.

139. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. SISO: a conceptual framework for the construction of "stereotypical maps" in a Social Internetworking Scenario. In *Proc. of the International Workshop on New Frontiers in Mining Complex Knowledge Patterns at ECML/PKDD 2012 (NFMCP 2012)*, Bristol, UK, 2012.

140. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Internetworking assortativity in Facebook. In *Proc. of the International Conference on Social Computing and its Applications (SCA 2013)*, pages 335–341, Karlsruhe, Germany, 2013. IEEE Computer Society.

141. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences*, 256:126–137, 2014. Elsevier.

142. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Discovering Missing Me Edges across Social Networks. *Information Sciences*, 319:18–37, 2015. Elsevier.

143. C. Buntain and J. Golbeck. Identifying Social Roles in Reddit Using Network Structure. In *Proc. of the International Conference on World Wide Web (WWW 2014)*, page 615–620, Seoul, Korea, 2014. ACM.

144. M. Buscema, E. Grossi, M. Capriotti, C. Babiloni, and P. Rossini. The IFAST model allows the prediction of conversion to Alzheimer disease in patients with mild cognitive impairment with high degree of accuracy. *Current Alzheimer Research*, 7(2):173–187, 2010. Bentham Science Publishers.

145. J.W. Byers, M. Mitzenmacher, and G. Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proc. of the ACM Conference on Electronic Commerce (EC'12)*, pages 248–265, Valencia, Spain, 2012. ACM.

146. C. Stergiou and K.E. Psannis and B.B. Gupta and Y. Ishibashi. Security, privacy & efficiency of sustainable cloud computing for big data & iot. *Sustainable Computing: Informatics and Systems*, 19:174–184, 2018. Elsevier.

147. D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. In *Proc. of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'05)*, pages 445–452, Porto, Portugal, 2005. Springer.

148. D. Camacho, A. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria. The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, 63:88–120, 2020. Elsevier.

149. A. Cammarano, F. Michelino, E. Lamberti, and M. Caputo. From social network analysis to business network analysis: Roles and features of companies involved in joint patenting activities. In *Proc. of the International Business Information Management Association Conference - Innovation Vision 2020: From Regional Development Sustainability to Global Economic Growth(IBIMA 2015)*, pages 955–964, Madrid, Spain, 2015.

150. R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020. Elsevier.

151. U. Can and B. Alatas. A new direction in social network analysis: Online social network analysis problems and applications. *Physica A: Statistical Mechanics and its Applications*, 535:122372, 2019. Elsevier.

152. Q. Cao, W. Duan, and Q. Gan. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521, 2011. Elsevier.

153. M. Carpenter and M. Garner. NSFW: An Empirical Study of Scandalous Trademarks. *Cardozo Arts & Ent. LJ*, 33:321, 2015. HeinOnline.

154. L. Caruccio and S. Cirillo. Incremental discovery of imprecise functional dependencies. *Journal of Data and Information Quality*, 12(4):1–25, 2020. ACM.

155. L. Caruccio, V. Deufemia, F. Naumann, and G. Polese. Discovering relaxed functional dependencies based on multi-attribute dominance. *IEEE Transactions on Knowledge and Data Engineering*, 33(9):3212–3228, 2020. IEEE.

156. R. Casadei, G. Fortino, D. Pianini, W. Russo, C. Savaglio, and M. Viroli. A development approach for collective opportunistic edge-of-things services. *Information Sciences*, 498:154–169, 2019. Elsevier.

157. R. Casadei, G. Fortino, D. Pianini, W. Russo, C. Savaglio, and M. Viroli. Modelling and simulation of Opportunistic IoT Services with Aggregate Computing. *Future Generation Computer Systems*, 91:252–262, 2019.

158. N. Cassavia, E. Masciari, C. Pulice, and D. Saccà. Discovering User Behavioral Features to Enhance Information Search on Big Data. *ACM Transactions on Interactive Intelligent Systems*, 7(2), 2017. ACM.

159. S. Castano, V. De Antonellis, and S. De Capitani di Vimercati. Global viewing of heterogeneous data sources. *IEEE Transactions on Data and Knowledge Engineering*, 13(2):277–297, 2001.

160. M. Cataldi, L. Di Caro, and C. Schifanella. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. In *Proc. of the International Workshop on Multimedia Data Mining (MDMKDD 2010)*, pages 4–13, Washington, DC, USA, 2010. ACM.

161. F. Cauteruccio, L. Cinelli, E. Corradini, G. Terracina, D. Ursino, L. Virgili, G. Fortino, A. Liotta, and C. Savaglio. A framework for anomaly detection and classification in Multiple IoT scenarios. *Future Generation Computer Systems*, 114:322–335, 2021. Elsevier.

162. F. Cauteruccio, L. Cinelli, G. Fortino, C. Savaglio, and G. Terracina. Using sentiment analysis and automated reasoning to boost smart lighting systems. In *Proc. of the 12th*

International Conference in Internet and Distributed Computing Systems (IDCS 2019), volume 11874 of *LNCS*, pages 69–78, Naples, Italy, 2019. Springer.

163. F. Cauteruccio, L. Cinelli, G. Fortino, C. Savaglio, G. Terracina, D. Ursino, and L. Virgili. An Approach to Compute the Scope of a Social Object in a Multi-IoT Scenario. *Pervasive and Mobile Computing*, 67:101223, 2020. Elsevier.
164. F. Cauteruccio, L. Cinelli, G. Terracina, D. Ursino, and L. Virgili. Investigating the scope of a thing in a multiple Internet of Things scenario. In *Atti del Ventisettesimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati (SEBD'19)*, Castiglione della Pescaia (GR), Italy, 2019.
165. F. Cauteruccio, E. Corradini, G. Terracina, D. Ursino, and L. Virgili. Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *Journal of Information Science*. Forthcoming.
166. F. Cauteruccio, E. Corradini, G. Terracina, D. Ursino, and L. Virgili. Co-posting Author Assortativity in Reddit. In *Atti del Ventottesimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati (SEBD'20)*, pages 222–233, Villasimius (CA), Italy, 2020. CEUR Workshop Proceedings.
167. F. Cauteruccio, G. Fortino, A. Guerrieri, A. Liotta, D.C. Mocanu, C. Perra, G. Terracina, and M.T. Vega. Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance. *Information Fusion*, 52:13–30, 2019. Elsevier.
168. F. Cauteruccio, P. Lo Giudice, L. Musarella, G. Terracina, D. Ursino, and L. Virgili. A lightweight approach to extract interschema properties from structured, semi-structured and unstructured sources in a big data scenario. *International Journal of Information Technology & Decision Making*, 19(3):849–889, 2020. World Scientific.
169. F. Cauteruccio, G. Terracina, and D. Ursino. Generalizing identity-based string comparison metrics: Framework and Techniques. *Knowledge-Based Systems*, 187:104820, 2020. Elsevier.
170. F. Cauteruccio, G. Terracina, D. Ursino, and L. Virgili. Redefining Betweenness Centrality in a Multiple IoT Scenario. In *Proc. of the International Workshop on Artificial Intelligence & Internet of Things (AI&IOT'19)*, pages 16–27, Rende (CS), Italy, 2019.
171. J. Caverlee, L. Liu, and S. Webb. The socialtrust framework for trusted social information management: Architecture and algorithms. *Information Sciences*, 180(1):95–112, 2010.
172. S. Cha, T. Tsai, W. Peng, T. Huang, and T. Hsu. Privacy-aware and blockchain connected gateways for users to access legacy IoT devices. In *Proc. of the Global Conference on Consumer Electronics (GCCE'17)*, pages 1–3, Nagoya, Japan, 2017. IEEE.
173. P. Chaim and M.P. Laurini. Is Bitcoin a bubble? *Physica A: Statistical Mechanics and its Applications*, 517:222–232, 2019. Elsevier.
174. D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10(4), 2008. ACM.
175. W. Chan and A. Olmsted. Ethereum transaction graph analysis. In *Proc. of the International Conference for Internet Technology and Secured Transactions (ICITST'17)*, pages 498–500, Cambridge, MA, USA, 2017. IEEE.

176. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computer Surveys*, 41(3):15:1–15:58, 2009. ACM.

177. Y. Chang, W.G. Yang, M. Yang, K. Lai, C.Y. Lin, and H.Y. Chang. Locate the technological position by technology redundancy and centralities: Patent citation network perspective. In *Proc. of the International Conference on Management of Engineering and Technology (PICMET 2016)*, pages 1550–1559, Honolulu, HI, USA, 2016. IEEE.

178. Y.C. Chang, C.H. Ku, and C.H. Chen. Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. *International Journal of Information Management*, 48:263–279, 2019. Elsevier.

179. E.T. Cheah and J. Fry. Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters*, 130:32–36, 2015. Elsevier.

180. C.Y.H. Chen and C.M. Hafner. Sentiment-induced bubbles in the cryptocurrency market. *Journal of Risk and Financial Management*, 12(2):53, 2019. Multidisciplinary Digital Publishing Institute.

181. D. Chen, H. Gao, L. Lu, and T. Zhou. Identifying influential nodes in large-scale directed networks: the role of clustering. *PloS one*, 8(10):e77455, 2013. Public Library of Science.

182. D. Chen, L. Lü, M. Shang, Y. Zhang, and T. Zhou. Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4):1777–1787, 2012. Elsevier.

183. F. Chen and D.B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD'14)*, pages 1166–1175, New York, NY, USA, 2014. ACM.

184. G. Chen, B.D. Ward, C. Xie, W. Li, Z. Wu, J. Jones, M. Franczak, P. Antuono, and S. Li. Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology*, 259(1):213–221, 2011. Radiological Society of North America, Inc.

185. J. Chen, N. Zhong, and J. Feng. Developing a Provenance Warehouse for the Systematic Brain Informatics Study. *International Journal of Information Technology & Decision Making*, 16(06):1581–1609, 2017. World Scientific.

186. R. Chen, J. Guo, and F. Bao. Trust management for SOA-based IoT and its application to service composition. *IEEE Transactions on Services Computing*, 9(3):482–495, 2016. IEEE.

187. X. Chen, Z. Qin, Y. Zhang, and T. Xu. Learning to rank features for recommendation over multiple categories. In *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*, pages 305–314, New York, NY, USA, 2016. ACM.

188. X. Chen, Y. Yuan, and M. Ali Orgun. Using bayesian networks with hidden variables for identifying trustworthy users in social networks. *Journal of Information Science*, 46(5):600–615, 2019. SAGE.

189. Z. Chen, W. Hendrix, and N.F. Samatova. Community-based anomaly detection in evolutionary networks. *Journal of Intelligent Information Systems*, 39(1):59–85, 2012. Springer.

190. J. Cheng, J. M. Kleinberg, J. Leskovec, D. Liben-Nowell, B. State, K. Subbian, and L. A. Adamic. Do Diffusion Protocols Govern Cascade Growth? In *Proc. of the International Conference on Web and Social Media (ICWSM 2018)*, pages 32–41, Stanford, CA, USA, 2018. AAAI Press.

191. A. Cheung, E. Roca, and J. Su. Crypto-currency bubbles: an application of the Phillips–Shi–Yu (2013) methodology on Mt. Gox bitcoin prices. *Applied Economics*, 47(23):2348–2358, 2015. Taylor & Francis.

192. C.M.K. Cheung and M.K.O Lee. What Drives Consumers to Spread Electronic Word of Mouth in Online Consumer-Opinion Platforms. *Decision Support Systems*, 53(1):218–225, 2012. Elsevier.

193. C.M.K. Cheung and D.R. Thadani. The impact of Electronic Word-of-Mouth Communication: A Literature Analysis and Integrative Model. *Decision Support Systems*, 54(1):461—470, 2012. Elsevier.

194. H.S. Choi and W.S. Rhee. Social based Trust Management System for Resource Sharing Service. In *Proc. of International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI'18)*, pages 148–152, Phuket, Thailand, 2018. ACM.

195. N. Chouhan, H.K. Saini, and S. Jain. Internet of things: Illuminating and study of protection and justifying potential countermeasures. In *Soft Computing and Signal Processing*, pages 21–27. 2019. Springer.

196. D. Chyzhyk, M. Graña, D. Öngür, and A.K. Shinn. Discrimination of schizophrenia auditory hallucinators by machine learning of resting-state functional MRI. *International Journal of Neural Systems*, 25(03):1550007, 2015. World Scientific.

197. A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review Part E*, 70(6):066111, 2004.

198. A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. SIAM.

199. M. Coffano and G. Tarasconi. CRIOS - Patstat Database: Sources, Contents and Access Rules. *Center for Research on Innovation, Organization and Strategy, CRIOS Working Paper*, 2014.

200. J. Cohen. Trusses: Cohesive subgraphs for social network analysis. *National security agency technical report*, 16(3.1), 2008.

201. T. Connie, M. Al-Shabi, and M. Goh. Smart content recognition from images using a mixture of convolutional neural networks. In *IT Convergence and Security 2017*, pages 11–18. 2018. Springer.

202. L. Cooke and H. Hall. Facets of DREaM: A social network analysis exploring network development in the UK LIS research community. *Journal of Documentation*, 69(6):786–806, 2013. Emerald Group Publishing Limited.

203. A. Corbellini, C. Mateos, A. Zunino, D. Godoy, and S.N. Schiaffino. Persisting big-data: The NoSQL landscape. *Information Systems*, 63:1–23, 2017. Elsevier.

204. S. Corbet, B. Lucey, A. Urquhart, and L. Yarovaya. Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62:182–199, 2019. Elsevier.

205. S. Corbet, B. Lucey, and L. Yarovaya. Datestamping the Bitcoin and Ethereum bubbles. *Finance Research Letters*, 26:81–88, 2018. Elsevier.

206. E. Corradini, A. Nocera, D. Ursino, and L. Virgili. Defining and detecting k-bridges in a social network: the Yelp case, and more. *Knowledge-Based Systems*, 187:104820, 2020. Elsevier.

207. E. Corradini, A. Nocera, D. Ursino, and L. Virgili. Investigating negative reviews and detecting negative influencers in Yelp through a multi-dimensional social network based model. *International Journal of Information Management*, 60:102377, 2021. Elsevier.

208. E. Corradini, A. Nocera, D. Ursino, and L. Virgili. Investigating the phenomenon of NSFW posts in Reddit. *Information Sciences*, 566:140–164, 2021. Elsevier.

209. D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi. The many shades of anonymity: Characterizing anonymous social media content. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM 2015)*, pages 71–80, Oxford, UK, 2015. AAAI.

210. L. Costa, F. Rodrigues, G. Travieso, and P. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007. Taylor & Francis.

211. Y. Cui. An Evaluation of Yelp Dataset. *arXiv preprint arXiv:1512.06915*, 2015.

212. T. Cunha, D. Jurgens, C. Tan, and D. Romero. Are All Successful Communities Alike? Characterizing and Predicting the Success of Online Communities. In *Proc. of the World Wide Web Conference (WWW 2019)*, pages 318–328, San Francisco, CA, USA, 2019. ACM.

213. E. Curry, W. Derguech, S. Hasan, C. Kouroupetroglou, and U. ul Hassan. A Real-time Linked Dataspace for the Internet of Things: Enabling "Pay-As-You-Go" Data Management in Smart Environments. *Future Generation Computer Systems*, 90:405–422, 2019. Elsevier.

214. B. Czigler, D. Csikós, Z. Hidasi, Z. Gaál, E. Csibri, E. Kiss, P. Salacz, and M. Molnár. Quantitative EEG in early Alzheimer's disease patients - power spectrum and complexity features. *International Journal of Psychophysiology*, 68(1):75–80, 2008. Elsevier.

215. W. Dai, G.Z Jin, J. Lee, and M. Luca. Optimal aggregation of consumer ratings: an application to yelp.com. *NBER Working Paper Series*, page 18567, 2012.

216. K. Darwish, P. Stefanov, M.J. Aupetit, and P. Nakov. Unsupervised User Stance Detection on Twitter. In *Proc. of the International Conference on Web and Social Media (ICWSM 2020)*, pages 141–152, Atlanta, GA, USA, 2020. AAAI Press.

217. K. Das, S. Samanta, and M. Pal. Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, 8(1):13, 2018.

218. S. Datta and E. Adar. Extracting Inter-Community Conflicts in Reddit. In *Proc. of the International Conference on Web and Social Media (ICWSM 2019)*, pages 146–157, Munich, Germany, 2019. AAAI.

219. S.K. Datta. Towards securing discovery services in Internet of Things. In *Proc. of the International Conference on Consumer Electronics (ICCE'16)*, pages 506–507, Las Vegas, NV, USA, 2016. IEEE.

220. J. Dauwels, F. Vialatte, and A. Cichocki. Diagnosis of Alzheimer's disease from EEG signals: where are we standing? *Current Alzheimer Research*, 7(6):487–505, 2010. Bentham Science Publishers.

221. J. Dauwels, F. Vialatte, and A. Cichocki. On the early diagnosis of Alzheimer's disease from EEG signals: A mini-review. In *Advances in Cognitive Neurodynamics (II)*, pages 709–716. 2011. Springer.

222. J. Dauwels, F. Vialatte, T. Musha, and A. Cichocki. A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. *NeuroImage*, 49(1):668–693, 2010. Elsevier.

223. D. Davis, R. Lichtenwalter, and N.V. Chawla. Multi-relational link prediction in heterogeneous information networks. In *Proc. of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2011)*, pages 281–288, Kaohsiung, Taiwan, 2011. IEEE.

224. P.V.A. de Freitas, G.N.P. Santos, A.J.G. Busson, A.L.V. Guedes, and S. Colcher. A baseline for NSFW video detection in e-learning environments. In *Proc. of the Brazillian Symposium on Multimedia and the Web (WebMedia 2019)*, pages 357–360, Rio de Janeiro, Brazil, 2019. ACM.

225. W. de Haan, Y.A. Pijnenburg, R.L. Strijers, Y. van der Made, W. van der Flier, P. Scheltens, and C.J. Stam. Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory. *BMC neuroscience*, 10(1):1, 2009. BioMed Central.

226. P. De Meo, A. Nocera, D. Rosaci, and D. Ursino. Recommendation of reliable users, social networks and high-quality resources in a Social Internetworking System. *AI Communications*, 24(1):31–50, 2011. IOS Press.

227. P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. Integration of XML Schemas at various "severity" levels. *Information Systems*, 31(6):397–434, 2006.

228. P. De Meo, G. Quattrone, and D. Ursino. A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI)*, 20(1):41–86, 2010. Springer.

229. R. DeJordy and D. Halgin. Introduction to ego network analysis. *Boston MA: Boston College and the Winston Center for Leadership & Ethics*, 2008.

230. A. Delorme and S. Makeig. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including ındependent component ʌnalysis. *Journal of Neuroscience Methods*, 134:9–21, 2004.

231. N. Deo. *Graph Theory with Applications to Engineering and Computer Science*. Mineola, New York, United States, 2016. Dover Publications.

232. D. Diamantini, D. Potena, and E. Storti. A virtual mart for knowledge discovery in databases. *Information Systems Frontiers*, 15(3):447–463, 2013. Springer.

233. S. Distefano, G. Merlino, and A. Puliafito. Enabling the cloud of things. In *Proc. of the International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS'2012)*, pages 858–863, Taichung, Taiwan, 2012. IEEE.

234. J. Domingo-Ferrer. On the connection between t-closeness and differential privacy for data releases. In *Proc. of the International Conference on Security and Cryptography (SECRYPT'13)*, pages 1–4, Reykjavík, Iceland, 2013. IEEE.

235. J. Domingo-Ferrer and J. Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151–158, 2015. Elsevier.

236. C. Donato, P. Lo Giudice, R. Marretta, D. Ursino, and L. Virgili. A well-tailored centrality measure for evaluating patents and their citations. *Journal of Documentation*, 75(4):750–772, 2019. Emerald.

237. S. Dorogovtsev, A. Goltsev, and J. Mendes. K-core organization of complex networks. *Physical Review Letters*, 96(4):040601, 2006. APS.

238. A. Dorri, S. Kanhere, R. Jurdak, and P. Gauravaram. Blockchain for IoT security and privacy: The case study of a smart home. In *Proc. of the International Conference on Pervasive Computing and Communications Workshops (PerCom'17 Workshops)*, pages 618–623, Kona, HI, USA, 2017. IEEE.

239. A. Dorri, S. Kanhere, R. Jurdak, and P. Gauravaram. LSB: A Lightweight Scalable Blockchain for IoT security and anonymity. *Journal of Parallel and Distributed Computing*, 134:180–197, 2019. Elsevier.

240. A. Dorri, S.S. Kanhere, and R. Jurdakand P. Gauravaram. Blockchain for IoT security and privacy: The case study of a smart home. In *Proc. of the IEEE international Conference on Pervasive Computing and Communications Workshops (PerCom'17 workshops)*, pages 618–623, Kona, HI, USA, 2017. IEEE.

241. R. dos Santos Mello, S. Castano, and C.A. Heuser. A method for the unification of XML schemata. *Information & Software Technology*, 44(4):241–249, 2002. Elsevier.

242. M. Du, R. Christensen, W. Zhang, and F. Li. Pcard: Personalized restaurants recommendation from card payment transaction records. In *Proc. of the World Wide Web Conference (WWW 2019)*, pages 2687–2693, San Francisco, CA, USA, 2019. ACM.

243. C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011. ACM.

244. C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. Now Publishers.

245. O. Ejermo and C. Karlsson. Interregional inventor networks as studied by patent coinventorships. *Research Policy*, 35(3):412–430, 2006. Elsevier.

246. A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli. Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science*, 4(11):170623, 2017. The Royal Society Publishing.

247. P. Ellis, G. Hepburn, and C. Oppenhein. Studies on patent citation networks. *Journal of documentation*, 34(1):12–20, 1978. MCB UP Ltd.

248. H. Elmeleegy, M. Ouzzani, and A.K. Elmagarmid. Usage-Based Schema Matching. In *Proc. of the International Conference on Data Engineering (ICDE'08)*, pages 20–29, Cancún, México, 2008. IEEE.

249. K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008. BMJ Group BMA House, Tavistock Square, London, WC1H 9JR.

250. Y. Eom. Premium and speculative trading in bitcoin. *Finance Research Letters*, page 101505, 2020. Elsevier.

251. P. Erdi and P. Bruck. Patent Citation Network Analysis: Ranking: from web pages to patents. In *Proc. of the International Conference on Artificial Neural Networks, (ICANN 2016)*, page 544, Barcelona, Spain, 2016. Springer Verlag.

252. A. Esfandyari, M. Zignani, S. Gaito, and G.P. Rossi. User identification across online social networks in practice: Pitfalls and solutions. *Journal of Information Science*, 44(3):377–391, 2018. SAGE Publications.

253. E. Estrada and J. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005. APS.

254. M.G. Everett and S.P. Borgatti. The centrality of groups and classes. *The Journal of mathematical sociology*, 23(3):181–201, 1999. Taylor & Francis.

255. M.G. Everett and S.P. Borgatti. Ego network betweenness. *Social networks*, 27(1):31–38, 2005. Elsevier.

256. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, and A. Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3):267–279, 2014. IEEE.

257. S. Fakhraei, J.R. Foulds, M.V.S. Shashanka, and L. Getoor. Collective Spammer Detection in Evolving Multi-Relational Social Networks. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 1769–1778, Sydney, Australia, 2015. ACM.

258. U. M. Farooq, M. Waseem, A. Khairi, and S. Mazhar. A critical analysis on the security concerns of Internet of Things (IoT). *International Journal of Computer Applications*, 111(7), 2015. Foundation of Computer Science.

259. I. Farris, R. Girau, L. Militano, M. Nitti, L. Atzori, A. Iera, and G. Morabito. Social virtual objects in the edge cloud. *IEEE Cloud Computing*, 2(6):20–28, 2015. IEEE.

260. A. Felfernig, S. Polat-Erdeniz, C. Uran, S. Reiterer, M. Atas, T.N.T. Tran, P. Azzoni, C. Kiraly, and Dolui K. An overview of recommender systems in the Internet of Things. *Journal of Intelligent Information Systems*, pages 1–25, 2018.

261. F. Feng and W.B. Croft. Probabilistic techniques for phrase extraction. *Information Processing & Management*, 37(2):199–220, 2001.

262. M. Ferrara, D. Fosso, D. Lanatà, R. Mavilia, and D. Ursino. A Social Network Analysis based approach to extracting knowledge patterns about innovation geography from patent databases. *International Journal of Data Mining, Modelling and Management*, 10(1):23–71, 2018. Inderscience.

263. B. Ferwerda and M. Schedl. Personality-Based User Modeling for Music Recommender Systems. In *Proc. of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016)*, pages 254–257, Riva del Garda, Italy, 2016. Springer International Publishing.

264. C. Fiesler, J. Jiang, J. McCann, K. Frye, and J. Brubaker. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proc. of International Conference on Web and Social Media (ICWSM 2018)*, pages 72–81, Stanford, CA, USA, 2018. AAAI.

265. M. Fire and C. Guestrin. The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time. *Information Processing & Management*, 57(2):102041, 2020. Elsevier.

266. J. Fogel and S. Zachariah. Intentions to use the yelp review website and purchase behavior after reading reviews. *Journal of Theoretical and Applied Electronic Commerce Research*, 12(1):53–67, 2017. Universidad de Talca.

267. R. Fontana, A. Nuvolari, H. Shimizu, and A. Vezzulli. Reassessing patent propensity: evidence from a data-set of R&D awards, 1977-2004. *Research Policy*, 42(10):1780–1792, 2013. Elsevier.

268. R. Fontana, A. Nuvolari, and B. Verspagen. Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of Innovation and New Technology*, 18(4):311–336, 2009. Taylor & Francis.

269. A. Forestiero. Multi-agent recommendation system in Internet of Things. In *Proc. of the IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, (CCGRID 2017)*, pages 772–775, Madrid, Spain, 2017. IEEE Computer Society / ACM.

270. C. Forman, A. Ghose, and B. Wiesenfeld. Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Information Systems Research*, 19(3):291—-313, 2008. INFORMS.

271. G. Fortino, A. Rovella, W. Russo, and C. Savaglio. On the Classification of Cyberphysical Smart Objects in the Internet of Things. In *Proc. of the Internation Workshop on Networks of Cooperating Objects for Smart Cities (UBICITEC'14)*, pages 86–94, Berlin, Germany, 2014.

272. G. Fortino, W. Russo, C. Savaglio, W. Shen, and M. Zhou. Agent-oriented cooperative smart objects: From IoT system design to implementation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(11):1939–1956, 2017. IEEE.

273. G. Fortino, W. Russo, C. Savaglio, W. Shen, and M. Zhou. Agent-Oriented Cooperative Smart Objects: From IoT System Design to Implementation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(11):1939–1956, 2018.

274. G. Fortino, W. Russo, C. Savaglio, M. Viroli, and M. Zhou. Modeling Opportunistic IoT Services in Open IoT Ecosystems. In *Proc. of the Workshop "From Objects to Agents" (WOA'17)*, pages 90–95, Scilla (RC), Italy, 2017.

275. G. Fortino, C. Savaglio, C. E. Palau, J. S. de Puga, M. Ganzha, M. Paprzycki, M. Montesinos, A. Liotta, and M. Llop. Towards multi-layer interoperability of heterogeneous IoT platforms: The INTER-IoT approach. pages 199–232, 2018. Springer.

276. G. Fortino, C. Savaglio, C.E. Palau, J. Suarez de Puga, M. Ganzha, M. Paprzycki, M. Montesinos, A. Liotta, and M. Llop. Towards multi-layer interoperability of heterogeneous iot platforms: The inter-iot approach. In *Integration, interconnection, and interoperability of IoT systems*, pages 199–232. 2018. Springer.

277. G. Fortino and P. Trunfio, editors. *Internet of Things Based on Smart Objects, Technology, Middleware and Applications*. Springer, 2014.

278. F.J. Fraga, T.H. Falk, P.A. Kanda, and R. Anghinah. Characterizing Alzheimer's disease severity via resting-awake EEG amplitude modulation analysis. *PloS one*, 8(8):e72240, 2013. Public Library of Science.

279. D.W. Franks, J. Noble, P. Kaufmann, and S. Stagl. Extremism propagation in social networks with hubs. *Adaptive Behavior*, 16(4):264–274, 2008. SAGE Publications Sage UK: London, England.

280. L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. JSTOR.

281. L. C. Freeman. Centrality in Social Networks Conceptual and Clarification. *Social Networks*, 1(3):215–239, 1979. Elsevier.

282. J. Fry and E.T. Cheah. Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis*, 47:343–352, 2016. Elsevier.

283. J. Furukawa and K. Sako. Mix-net system, 2010. Google Patents.

284. E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1606–1611, Hyderabad, India, 2007.

285. D. Gambetta. Can we trust trust? In *Trust: Making and breaking cooperative relations*, chapter 13, pages 213–237. 2000.

286. M. Ganzha, M. Paprzycki, W. Pawłowski, P. Szmeja, and K. Wasielewska. Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective. *Journal of Network and Computer Applications*, 81:111–124, 2017. Elsevier.

287. P. Garcia-Teodoro, J.E. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, 28(1-2):18–28, 2009. Elsevier.

288. S. Garg, K. Kaur, S. Batra, G. Kaddoum, N. Kumar, and A. Boukerche. A multi-stage anomaly detection scheme for augmenting the security in iot-enabled applications. *Future Generation Computer Systems*, 104:105–118, 2020. Elsevier.

289. S. Garruzzo, S. Modafferi, D. Rosaci, and D. Ursino. X-Compass: an XML agent for supporting user navigation on the Web. In *Proc. of the International Conference on Flexible Query Answering Systems (FQAS 2002)*, pages 197–211, Copenhagen, Denmark, 2002. Lecture Notes in Artificial Intelligence, Springer-Verlag.

290. R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In *Proc. of the Workshop on Algorithm Engineering & Expermiments (ALENEX'08)*, pages 90–100, San Francisco, CA, USA, 2008.

291. D. Georgakopoulos and P.P. Jayaraman. Internet of things: from internet scale sensing to smart services. *Computing*, 98(10):1041–1058, 2016. Springer.

292. J.C. Gerlach, G. Demos, and D. Sornette. Dissection of Bitcoin's multiscale bubble history from January 2012 to February 2018. *Royal Society Open Science*, 6(7):180643, 2019. The Royal Society.

293. P. Lo Giudice, N. Mammone, F.C. Morabito, R.G. Pizzimenti, D. Ursino, and L. Virgili. Leveraging Network Analysis to support experts in their analyses of subjects with MCI and AD. *Medical & Biological Engineering & Computing*, 57(9):1961–1983, 2019. Springer.

294. P. Lo Giudice, D. Ursino, N. Mammone, F.C. Morabito, U. Aguglia, V. Cianci, E. Ferlazzo, and S. Gasparini. A network analysis based approach to characterizing Periodic Sharp Wave Complexes in electroencephalograms of patients with sporadic CJD. *International Journal of Medical Informatics*, 121:19–29, 2019. Elsevier.

295. P. Gogoi, D.K. Bhattacharyya, B. Borah, and J. K. Kalita. A survey of outlier detection methods in network anomaly identification. *The Computer Journal*, 54(4):570–588, 2011. Elsevier.

296. D. Goldschlag, M. Reed, and P. Syverson. Onion routing for anonymous and private internet connections. *Communications of the ACM*, 42(2):39–40, 1999. Association for Computing Machinery, Inc.

297. A.A. Gouw, A.M. Alsema, B.M. Tijms, A. Borta, P. Scheltens, C.J. Stam, and W.M. van der Flier. EEG spectral analysis as a putative early prognostic biomarker in nondemented, amyloid positive subjects. *Neurobiology of aging*, 57:133–142, 2017. Elsevier.

298. M.S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. JSTOR.

299. A. Grewal and J. Lin. The evolution of content analysis for personalized recommendations at Twitter. In *Proc. of the International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*, pages 1355–1356, Ann Arbor, MI, USA, 2018. ACM.

300. J. L. Gross and J. Yellen. *Graph Theory and Its Applications*. New York, United States, 2005. Chapman and Hall/CRC.

301. Peerless Research Group. Sensors in Distribution: On the Cusp of New Performance Efficiencies. *https://www.logisticsmgmt.com/wp_content/honeywell_wp_sensors_022316b.pdf*, 2015. Honeywell.

302. T. Grover and G. Mark. Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. In *Proc. of the International Conference on Web and Social Media (ICWSM 2019)*, pages 193–204, Munich, Germany, 2019. AAAI.

303. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013. Elsevier.

304. I.D. Guedalia, J. Guedalia, R.P. Chandhok, and S. Glickfield. Methods to discover, configure, and leverage relationships in Internet of Things (IoT) networks, feb 20 2018. Google Patents.

305. D. Guellec and B. van Pottelsberghe de la Potterie. The internationalisation of technology analysed with patent data. *Research Policy*, 30(8):1253–1266, 2001. Elsevier.

306. J. Guerreiro and P. Rita. How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *Journal of Hospitality and Tourism Management*, 43:269–272, 2020. Elsevier.

307. L. Gui, Y. Zhou, R. Xu, Y. He, and Q. Lu. Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*, 124:34–45, 2017. Elsevier.

308. A. Guimaraes, O. Balalau, E. Terolli, and G. Weikum. Analyzing the Traits and Anomalies of Political Discussions on Reddit. In *Proc. of the International Conference on Web and Social Media (ICWSM 2019)*, pages 205–213, Munich, Germany, 2019. AAAI.

309. D. Guinard, M. Fischer, and V. Trifa. Sharing using social networks in a composable web of things. In *Proc. of the International Conference on Pervasive Computing and Communications (PERCOM 2010)*, pages 702–707, Mannheim, Germany, 2010. IEEE.

310. D. Guinard, V. Trifa, F. Mattern, and E. Wilde. From the internet of things to the web of things: Resource-oriented architecture and best practices. *Architecting the Internet of Things*, pages 97–129, 2011. Springer.

311. A. Gulati and M. Eirinaki. With a Little Help from My Friends (and Their Friends): Influence Neighborhoods for Social Recommendations. In *Proc. of the World Wide Web Conference (WWW'19)*, pages 2778–2784, San Francisco, CA, USA, 2019. ACM.

312. R. K. Gunupudi, M. Nimmala, N. Gugulothu, and S. R. Gali. Clapp: A self constructing feature clustering approach for anomaly detection. *Future Generation Computer Systems*, 74:417–429, 2017. Elsevier.

313. B. Guo, Z. Yu, X. Zhou, and D. Zhang. Opportunistic IoT: Exploring the social side of the Internet of Things. In *Proc. of the International Conference on Computer Supported Cooperative Work in Design (CSCWD'12)*, pages 925–929, Wuhan, China, 2012. IEEE.

314. P. Hage and F. Harary. Eccentricity and centrality in networks. *Social networks*, 17(1):57–63, 1995. Elsevier.

315. P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C.J. Honey, J.V. Wedeen, and O. Sporns. Mapping the structural core of human cerebral cortex. *PLoS Computational Biology*, 6(7):e159, 2008. Public Library of Science.

316. M.A.M Hail, M. Amadeo, A. Molinaro, and S. Fischer. On the performance of caching and forwarding in information-centric networking for the IoT. In *Proc. of the International Conference on Wired/Wireless Internet Communication (WWIC'15)*, pages 313–326, Malaga, Spain, 2015. Springer.

317. W. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Loyalty in Online Communities. In *Proc. of the International Conference on Web and Social Media (ICWSM 2017)*, pages 540–543, Montreal, Canada, 2017. AAAI.

318. J. Han and M. Kamber. *Data Mining: Concepts and Techniques - Second Edition*. 2006. Morgan Kaufmann notes.

319. R. Hanneman and M. Riddle. *Introduction to social network methods*. `http://faculty.ucr.edu/~hanneman/nettext/` , 2005. University of California, Riverside.

320. S.M. Harding, W.B. Croft, and C. Weir. Probabilistic retrieval of ocr degraded text using n-grams. In *Proc. of the International Conference on Theory and Practice of Digital Libraries (ECDL'97)*, pages 345–359, Pisa, Italy, 1997. Springer.

321. M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *Proc. of the Advances in Neural Information Processing Systems (NIPS'12)*, pages 2339–2347, Stateline, NV, USA, 2012.

322. F. Hatz, M. Hardmeier, N. Benz, M. Ehrensperger, U. Gschwandtner, S. Rüegg, C. Schindler, A.U. Monsch, and P. Fuhr. Microstate connectivity alterations in patients with early Alzheimer's disease. *Alzheimer's research & therapy*, 7(1):78, 2015. BioMed Central.

323. D.M. Hawkins. *Identification of outliers / D.M. Hawkins*. New York, 1980. Chapman and Hall London ; New York.

324. Q. He, X. Wang, F. Mao, J. Lv, Y. Cai, M. Huang, and Q. Xu. CAOM: A community-based approach to tackle opinion maximization for social networks. *Information Sciences*, 513:252–269, 2020. Elsevier.

325. Y. He, Z.J. Chen, and A.C. Evans. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cerebral Cortex*, 17(10):2407–2419, 2007. Oxford University Press.

326. L. Hebert, J. Weuve, P. Scherr, and D. Evans. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783, 2013. AAN Enterprises.

327. J. Hessel, C. Tan, and L. Lee. Science, AskScience, and BadScience: On the Coexistence of Highly Related Communities. In *Proc. of the International Conference on Web and Social Media (ICWSM 2016)*, pages 171–180, Cologne, Germany, 2016. AAAI.

328. A. Hicks, S. Comp, J. Horovitz, M. Hovarter, M. Miki, and J.L Bevan. Why people use Yelp. com: An exploration of uses and gratifications. *Computers in Human Behavior*, 28(6):2274–2279, 2012. Elsevier.

329. P. Hingley and S. Bas. Numbers and sizes of applicants at the European Patent Office. *World Patent Information*, 31(4):285–298, 2009. Elsevier.

330. J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005. National Academy of Sciences.

331. T.J. Hirschauer, H. Adeli, and J.A. Buford. Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. *Journal of Medical Systems*, 39(11):179, 2015. Springer.

332. A.O. Hirschman. The paternity of an index. *The American Economic Review*, 54(5):761–762, 1964. JSTOR.

333. Y.C. Ho, J. Wu, and Y. Tan. Disconfirmation Effect on Online Rating Behavior: A Structural Model. *Information Systems Research*, 28(3):626—642, 2008. INFORMS.

334. L. Holmquist, F. Mattern, B. Schiele, P. Alahuhta, M. Beigl, and H. Gellersen. Smart-its friends: A technique for users to easily establish connections between smart artefacts. In *Proc. of the International Conference on Ubiquitous Computing (Ubicomp'2001)*, pages 116–122, Atlanta, GA, USA, 2001. Springer.

335. C.J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J.P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural con-

nectivity. *Proc. of the National Academy of Sciences*, 106(6):2035–2040, 2009. National Academy of Sciences.

336. R. Hornero, D. Abásolo, J. Escudero, and C. Gómez. Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with Alzheimer's disease. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1887):317–336, 2009. The Royal Society.
337. M. Hosseini-Pozveh, K. Zamanifar, and A.R. Naghsh-Nilchi. A community-based approach to identify the most influential nodes in social networks. *Journal of Information Science*, 43(2):204–220, 2017. SAGE Publications.
338. A. Høst-Madsen and J. Zhang. Coding of graphs with application to graph anomaly detection. *CoRR*, abs/1804.02469, 2018.
339. R. Hsu, J. Lee, T.Q. Quek, and J. Chen. GRAAD: Group Anonymous and Accountable D2D Communication in Mobile Networks. *IEEE Transactions on Information Forensics and Security*, 13(2):449–464, 2018. IEEE.
340. C.C. Hsueh and C.C. Wang. The Use of Social Network Analysis in Knowledge Diffusion Research from Patent Data. In *Proc. of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009)*, pages 393–398, Athens, Greece, 2009. IEEE Computer Society.
341. J. Hu and Y. Zhang. Structure and patterns of cross-national Big Data research collaborations. *Journal of Documentation*, 73(6):1119–1136, 2017. Emerald Publishing Limited.
342. L. Hu, A. Sun, and Y. Liu. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proc. of the International ACM SIGIR Conference on Research & development in information retrieval (SIGIR'14)*, pages 345–354, Gold Coast, Queensland, Australia, 2014. ACM.
343. Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta Structure: Computing Relevance in Large Heterogeneous Information Networks. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1595–1604, New York, NY, USA, 2016. ACM.
344. J. Huenteler, J. Ossenbrink, T. Schmidt, and V. Hoffmann. How a product's design hierarchy shapes the evolution of technological knowledge – Evidence from patent-citation networks in wind power. *Research Policy*, 45(6):1195–1217, 2016. Elsevier.
345. J. Huenteler, T. Schmidt, J. Ossenbrink, and V. Hoffmann. Technology life-cycles in the energy sector – Technological characteristics and the role of deployment for innovation. *Technological Forecasting and Social Change*, 104:102–121, 2016. Elsevier.
346. N. Hummon and P. Dereian. Connectivity in a citation network: The development of DNA theory. *Social networks*, 11(1):39–63, 1989. Elsevier.
347. K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Söderström. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD'18)*, pages 387–395, London, UK, 2018. ACM.
348. N.Q.V. Hung, N.T. Tam, V.T. Chau, T.K. Wijaya, Z. Miklós, K. Aberer, A. Gal, and M. Weidlich. SMART: A tool for analyzing and reconciling schema matching networks. In *Proc.*

of the International Conference on Data Engineering (ICDE'15), pages 1488–1491, Seoul, South Korea, 2015. IEEE.

349. S. Hung and A. Wang. Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, 82(1):121–134, 2010. Springer.
350. C.J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. of the International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, pages 216–225, Ann Arbor, MI, USA, 2014.
351. L. Jain and R. Katarya. Discover opinion leader in online social network using firefly algorithm. *Expert Systems with Applications*, 122:1–15, 2019. Elsevier.
352. S. Jain and S. Tanwani. Schema matching technique for heterogeneous web database. In *Proc. of the International Conference on Reliability (ICRITO'15)*, pages 1–6, Noida, India, 2015. IEEE.
353. M. Jalili. Graph theoretical analysis of Alzheimer's disease: Discrimination of AD patients from healthy subjects. *Information Sciences*, 384:145–156, 2017. Elsevier.
354. F. Jamour, S. Skiadopoulos, and P. Kalnis. Parallel Algorithm for Incremental Betweenness Centrality on Large Graphs. *IEEE Transactions on Parallel and Distributed Systems*, 2017. IEEE.
355. J. Jeong. EEG dynamics in patients with Alzheimer's disease. *Clinical neurophysiology*, 115(7):1490–1505, 2004. Elsevier.
356. L. Jiang and X. Zhang. BCOSN: A blockchain-based decentralized online social network. *IEEE Transactions on Computational Social Systems*, 6(6):1454–1466, 2019. IEEE.
357. S. Jiang, L. Bing, B. Sun, Y. Zhang, and W. Lam. Ontology enhancement and concept granularity learning: keeping yourself current and adaptive. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD'11)*, pages 1244–1252, San Diego, CA, USA, 2011. ACM.
358. F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNAKDD '13)*, Chicago, Illinois, 2013. ACM.
359. A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644, 2007. Elsevier.
360. V. Jyothsna and V.V. Rama Prasad. Article: A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7):26–35, 2011. IJCA Journal.
361. H. Kalodner, M. Möser, K. Lee, S. Goldfeder, M. Plattner, A. Chator, and A. Narayanan. Blocksci: Design and applications of a blockchain analysis platform. In *Proc. of the International Security Symposium (USENIX'20)*, pages 2721–2738, 2020. USENIX Association.
362. A. Kamra, E. Terzi, and E. Bertino. Detecting anomalous access patterns in relational databases. *The International Journal on Very Large Data Bases*, 17(5):1063–1077, 2008. Springer.

363. Y.S. Kang, J. Min, J. Kim, and H. Lee. Roles of alternative and self-oriented perspectives in the context of the continued use of social network sites. *International Journal of Information Management*, 33(3):496–511, 2013. Elsevier.

364. S. Kapidakis. Rating Quality in Metadata Harvesting. In *Proc. of the International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*, pages 65:1–65:8, New York, NY, USA, 2015. ACM.

365. K. K. Kapoor, K. Tamilmani, N. P Rana, P. Patil, Y. K Dwivedi, and S. Nerur. Advances in social media research: past, present and future. *Information Systems Frontiers*, 20(3):531–558, 2018. Springer.

366. S. Karnouskos. Smart houses in the smart grid and the search for value-added services in the cloud of things era. In *Proc. of the International Conference on Industrial Technology (ICIT'2013)*, pages 2016–2021, Cape Town, Western Cape, South Africa, 2013. IEEE.

367. V. Karyotis, K. Tsitseklis, K. Sotiropoulos, and S. Papavassiliou. Big Data Clustering via Community Detection and Hyperbolic Network Embedding in IoT Applications. *Sensors*, 18(4):1205, 2018. Multidisciplinary Digital Publishing Institute.

368. N. Kasabov and E. Capecci. Spiking neural network methodology for modelling, classification and understanding of EEG spatio-temporal data measuring cognitive processes. *Information Sciences*, 294:565–575, 2015. Elsevier.

369. W. Kasper and M. Vela. Sentiment analysis for hotel reviews. In *Proc. of the International Computational Linguistics-Applications Conference*, volume 231527, pages 45–52, Jachranka, Poland, 2011.

370. L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. Springer.

371. A.L. Kavanaugh, D.D. Reese, J.M. Carroll, and M.B. Rosson. Weak ties in networked communities. *The Information Society*, 21(2):119–131, 2005. Taylor & Francis.

372. K. Kaviya, C. Roshini, V. Vaidhehi, and J.D. Sweetlin. Sentiment analysis for restaurant rating. In *Proc. of the International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM'17)*, pages 140–145, Chennai, India, 2017. IEEE.

373. C. Ke-Jia, Z. Pei, Y. Zinong, and L. Yun. iBridge: Inferring bridge links that diffuse information across communities. *Knowledge-Based Systems*, 192, 2020. Elsevier.

374. D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the International ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2003)*, pages 137–146, Washington, DC, USA, 2003. ACM.

375. M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938. Oxford University Press.

376. W.O. Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A.*, 115(772):700–721, 1927. The Royal Society London.

377. D.J. Ketchen and C.L. Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996. Wiley Online Library.

378. H. Kim and E. Gelenbe. Anomaly detection in gene expression via stochastic models of gene regulatory networks. *BMC Genomics*, 10(3):S26, 2009. BMC Genomics.

379. H. Kim, P. Howland, and H. Park. Dimension Reduction in Text Classification with Support Vector Machines. *Journal of Machine Learning Research*, 6:37–53, 2005.

380. H. Kim, R. Çetin Atalay, and E. Gelenbe. G-Network Modelling Based Abnormal Pathway Detection in Gene Regulatory Networks. In *Proc. of the International Symposium on Computer and Information Sciences (ISCIS'11)*, pages 257–263, London, UK, 2011.

381. H.K. Kim, H. Kim, and S. Cho. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing*, 266:336–352, 2017.

382. J. Kim, J. Bae, and M. Hastak. Emergency information diffusion on online social media during storm Cindy in US. *International Journal of Information Management*, 40:153–165, 2018. Elsevier.

383. J. Kim and M. Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.

384. J.K. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. ACM.

385. J. Knoll and J. Matthes. The Effectiveness of Celebrity Endorsements: A Meta-Analysis. *Journal of the Academy of Marketing Science*, 45(1):55–75, 2017. Springer.

386. M. Knyazeva, M. Jalili, A. Brioschi, I. Bourquin, E. Fornari, M. Hasler, R. Meuli, P. Maeder, and J. Ghika. Topography of EEG multivariate phase synchronization in early Alzheimer's disease. *Neurobiology of aging*, 31(7):1132–1144, 2010. Elsevier.

387. Y. Ko, D. Chae, and S. Kim. Influence maximisation in social networks: A target-oriented estimation. *Journal of Information Science*, 44(5):671–682, 2018. SAGE Publications.

388. G. Kondrak. N-gram similarity and distance. In *String Processing and Information Retrieval*, pages 115–126, 2005. Springer.

389. M. Koppert, S. Kalitzin, D. Velis, F. Lopes Da Silva, and M.A. Viergever. Preventive and Abortive Strategies for Stimulation Based Control of Epilepsy: A Computational Model Study. *International Journal of Neural Systems*, 26(08):1650028, 2016. World Scientific.

390. G. Kou, X. Chao, Y. Peng, F.E. Alsaadi, and E. Herrera-Viedma. Machine learning methods for systemic risk analysis in financial sectors. *Technological and Economic Development of Economy*, 25(5):716–742, 2019.

391. G. Kou, Y. Lu, Y. Peng, and Y. Shi. Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, 11(01):197–225, 2012. World Scientific.

392. G. Kou, Y. Peng, and G. Wang. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275:1–12, 2014. Elsevier.

393. Y. Kou, C.M. Gray, A.L. Toombs, and R.S. Adams. Understanding Social Roles in an Online Community of Volatile Practice: A Study of User Experience Practitioners on Reddit. *ACM Transactions on Social Computing*, 1(4):17:1–17:22, 2018. ACM.

394. E.L. Koua, A.M. MacEachren, and M. Kraak. Evaluating the usability of visualization methods in an exploratory geovisualization environment. *International Journal of Geographical Information Science*, 20(4):425–448, 2006. Taylor & Francis.

395. P. Kouvaris, E. Pirogova, H. Sanadhya, A. Asuncion, and A. Rajagopal. Text enhanced recommendation system model based on yelp reviews. *SMU Data Science Review*, 1(3):8, 2018.

396. A.M. Kowshalya and M.L. Valarmathi. Community Detection in the Social Internet of Things Based on Movement, Preference and Social Similarity. *Studies in Informatics and Control*, 25(4):499–506, 2016. National Institute for R&D in Informatics.

397. M. Kranz, L. Roalter, and F. Michahelles. Things that Twitter: social networks and the Internet of Things. In *Proc. of the International Workshop on Pervasive Computing (Pervasive 2010)*, pages 1–10, Helsinki, Finland, 2010.

398. N. Kumar and I. Benbasat. Research note: the influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research*, 17(4):425–439, 2006. INFORMS.

399. S. Kumar, J. Cheng, and J. Leskovec. Antisocial Behavior on the Web: Characterization and Detection. In *Proc. of the International Conference on World Wide Web Companion*, page 947–950, Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

400. S. Kumar, W.L. Hamilton, J. Leskovec, and D. Jurafsky. Community Interaction and Conflict on the Web. In *Proc. of the World Wide Web Conference (WWW 2018)*, pages 933–943, Lyon, France, 2018. ACM.

401. D. Labate, F. La Foresta, G. Morabito, I. Palamara, and F.C. Morabito. Entropic measures of EEG complexity in Alzheimer's disease through a multivariate multiscale approach. *IEEE Sensors Journal*, 13(9):3284–3292, 2013. IEEE.

402. A. Landherr, B. Friedl, and J. Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2(6):371–385, 2010. Springer.

403. J. Lanjouw and M. Schankerman. Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495):441–465, 2004. Wiley Online Library.

404. J. LaViolette and B. Hogan. Using Platform Signals for Distinguishing Discourses: The Case of Men's Rights and Men's Liberation on Reddit. In *Proc. of the International Conference on Web and Social Media (ICWSM 2019)*, pages 323–334, Munich, Germany, 2019. AAAI.

405. Q.V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proc. of the International Conference on Machine Learning (ICML'14)*, pages 1188–1196, Beijing, China, 2014. JMLR.org.

406. R. Lea and M. Blackstock. Smart Cities: an IoT-centric Approach. In *Proc. of the International Workshop on Web Intelligence and Smart Sensing (IWWISS '14)*, pages 12:1–12:2, Saint Etienne, France, 2014. ACM.

407. J. Lee, S. Yu, K. Park, Y. Park, and Y. Park. Secure three-factor authentication protocol for multi-gateway iot environments. *Sensors*, 19(10):2358, 2019. Multidisciplinary Digital Publishing Institute.

408. K. Lee, J. Ham, S. Yang, and C. Koo. Can You Identify Fake or Authentic Reviews? An fsQCA Approach. In *Information and Communication Technologies in Tourism 2018*, pages 214–227, Jonkoping, Sweden, 2018. Springer.

409. K. Lee, D. Kim, D. Ha, U. Rajput, and H. Oh. On security and privacy issues of fog computing supported Internet of Things environment. In *Proc. of the International Conference on the Network of the Future (NOF'15)*, pages 1–3, Montreal, Quebec, Canada, 2015. IEEE.

410. M. Lee, J. Lee, J.Y. Park, R.H. Choi, and C. Chung. Qube: a quick algorithm for updating betweenness centrality. In *Proc. of the International Conference on World Wide Web (WWW'12)*, pages 351–360, Lyon, France, 2012. ACM.

411. M.L. Lee, L.H. Yang, W. Hsu, and X. Yang. XClust: clustering XML schemas for effective integration. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM 2002)*, pages 292–299, McLean, Virginia, USA, 2002. ACM Press.

412. P. Lee, H. Su, and F. Wu. Quantitative mapping of patented technology – The case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, 77(3):466–478, 2010. Elsevier.

413. D. Leggio, G. Marra, and D. Ursino. Defining and investigating the scope of users and hashtags in Twitter. In *Proc. of the International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2014)*, pages 674–681, Amantea (CS), Italy, 2014. Lecture Notes in Computer Science. Springer.

414. X. Lei and X. Qian. Rating prediction via exploring service reputation. In *Proc. of the International Workshop on Multimedia Signal Processing (MMSP'15)*, pages 1–6, Xiamen, China, 2015. IEEE.

415. E.A. Leicht, P. Holme, and M. E. J. Newman. Vertex similarity in networks. *Physical Review Part E*, 73(2):026120, 2006.

416. J. Leskovec, L.A. Adamic, and B.A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, 2007. ACM.

417. J. Li, H. Wang, and S.U. Khan. A semantics-based approach to large-scale mobile social networking. *Mobile Networks and Applications*, 17(2):192–205, 2012. Springer.

418. M.X. Li, C.H. Tan, K.K. Wei, and K.L. Wang. Sequentiality of Product Review Information Provision: An Information Foraging Perspective. *MIS Q.*, 41(3):867–892, 2017. Management Information Systems Research Center.

419. N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. of the International Conference on Data Engineering (ICDE'07)*, pages 106–115, Istanbul, Turkey, 2007. IEEE.

420. N. Li, Li. Zeng, Q. He, and Z. Shi. Parallel implementation of apriori algorithm based on mapreduce. In *Proc. of the International Conference on Software Engineering, Arti-*

ficial Intelligence, Networking and Parallel/Distributed Computing (ACIS'13), pages 236–241, 2012. IEEE.

421. W. Li, S. Tug, W. Meng, and Y. Wang. Designing collaborative blockchained signature-based intrusion detection in iot environments. *Future Generation Computer Systems*, 96:481–489, 2019. Elsevier.
422. X. Li, P. Jiang, T. Chen, X. Luo, and Q. Wen. A survey on the security of blockchain systems. *Future Generation Computer Systems*, 107:841–853, 2020. Elsevier.
423. X. Li, Y. Tian, F. Smarandache, and R. Alex. An extension collaborative innovation model in the context of big data. *International Journal of Information Technology & Decision Making*, 14(01):69–91, 2015. World Scientific.
424. Y. Li, Z. Su, J. Yang, and C. Gao. Exploiting similarities of user friendship networks across social networks for user identification. *Information Sciences*, 506:78–98, 2020. Elsevier.
425. Y. Lim and B. Van Der Heide. Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on Yelp. *Journal of Computer-Mediated Communication*, 20(1):67–82, 2014. Oxford University Press.
426. C. Lin, G. Li, Z. Shan, and Y. Shi. Thinking and Modeling for Big Data from the Perspective of the I Ching. *International Journal of Information Technology & Decision Making*, 16(06):1451–1463, 2017. World Scientific.
427. J. Lin, E.J. Keogh, A.W. Fu, and H. Van Herle. Approximations to magic: Finding unusual medical time series. In *Proc. of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS 2005), 23-24 June 2005, Dublin, Ireland*, pages 329–334, 2005. IEEE Computer Society.
428. X. Lin and X. Wang. Examining gender differences in people's information-sharing decisions on social networking sites. *International Journal of Information Management*, 50:45–56, 2020. Elsevier.
429. Y.J. Lin, P.W. Wu, C.H. Hsu, I.P. Tu, and S.W. Liao. An evaluation of bitcoin address classification based on transaction history summarization. In *Proc. of the IEEE International Conference on Blockchain and Cryptocurrency (ICBC'19)*, pages 302–310, Seoul, South Korea, 2019. IEEE.
430. W. Lippmann. *Public Opinion*. 1922. Macmillan.
431. M. Lischke and B. Fabian. Analyzing the bitcoin network: The first four years. *Future Internet*, 8(1):7, 2016. Multidisciplinary Digital Publishing Institute.
432. J. Liu, Y. Xiao, and C.P. Chen. Authentication and access control in the Internet of Things. In *Proc. of the International Conference on Distributed Computing Systems Workshops*, pages 588–592, Macau, China, 2012. IEEE.
433. R. Liu, S. Wan, Z. Zhang, and X. Zhao. Is the introduction of futures responsible for the crash of Bitcoin? *Finance Research Letters*, 34:101259, 2020. Elsevier.
434. P. Lo Giudice, A. Nocera, D. Ursino, and L. Virgili. Building Topic-Driven Virtual IoTs in a Multiple IoTs Scenario. *Sensors*, 19(13):2956, 2019. MDPI.
435. P. Lo Giudice, P. Russo, and D. Ursino. A new Social Network Analysis-based approach to extracting knowledge patterns about research activities and hubs in a set of countries.

International Journal of Business Innovation and Research, 17(2):147–186, 2018. Inderscience.

436. P. Lo Giudice, D. Ursino, and L. Virgili. A "big data oriented" and "complex network based" model supporting the uniform investigation of heterogeneous personalized medicine data. In *Proc. of the International Conference on Bioinformatics and Biomedicine (BIBM'18)*, pages 2094–2101, Madrid, Spain, 2018. IEEE.
437. D.M. Low, R. K. Huang, P. Mishra, G. Jain, and J.B. Gosnell. Group formation using anonymous broadcast information, 2013. Google Patents.
438. L. Lu, D. Chen, X. Ren, Q. Zhang, Y. Zhang, and T. Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016. Elsevier.
439. L. Lü, D. Chen, X.L. Ren, Q.M. Zhang, Y.C. Zhang, and T. Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016. Elsevier.
440. R. Lu, K. Heung, A.H. Lashkari, and A. Ghorbani. A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT. *IEEE Access*, 5:3302–3312, 2017. IEEE.
441. R. Lu, X. Liang, X. Li, X. Lin, and X. Shen. Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications. *IEEE Transactions on Parallel and Distributed Systems*, 23(9):1621–1631, 2012. IEEE.
442. L.M. Lubango. The effect of co-inventors' reputation and network ties on the diffusion of scientific and technical knowledge from academia to industry in South Africa. *World Patent Information*, 43:5–11, 2015. Elsevier.
443. M. Luca. Reviews, reputation, and revenue: The case of Yelp.com. *Harvard Business School Working Paper*, 12-016, 2016.
444. M. Luca and G. Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016. INFORMS.
445. X. Luo. Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science*, 28(1):148–165, 2009. INFORMS.
446. M. López, A. Peinado, and A. Ortiz. An extensive validation of a sir epidemic model to study the propagation of jamming attacks against iot wireless networks. *Computer Networks*, 165:106945, 2019.
447. J. Ma and Y. Luo. The classification of rumour standpoints in online social network based on combinatorial classifiers. *Journal of Information Science*, 46(2):191–204, 2020. SAGE.
448. Z. Ma, A. Sun, and G. Cong. Will this #Hashtag be Popular Tomorrow? In *Proc. of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 1173 – 1174, Portland, OR, USA, 2012. ACM.
449. Z. Ma, A. Sun, and G. Cong. On Predicting the Popularity of Newly Emerging Hashtags in Twitter. *Journal of the Association for Information Science and Technology*, 64(7):1399–1410, 2013.
450. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proc. of the International Conference on Data Engineering (ICDE'06)*, pages 24–24, New York, NY, USA, 2006. IEEE.

451. J. Madhavan, P.A. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proc. of the International Conference on Very Large Data Bases (VLDB 2001)*, pages 49–58, Rome, Italy, 2001. Morgan Kaufmann.

452. D.D.F. Maesa, A. Marino, and L. Ricci. Uncovering the bitcoin blockchain: an analysis of the full users graph. In *Proc. of the International Conference on Data Science and Advanced Analytics (DSAA'16)*, pages 537–546, Montreal, Quebec, Canada, 2016. IEEE.

453. C. Magnien and F. Tarissan. Time evolution of the importance of nodes in dynamic networks. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015)*, pages 1200–1207, Paris, France, 2015. IEEE.

454. W. Maharani, Adiwijaya, and A.A. Gozali. Degree centrality and eigenvector centrality in twitter. In *Proc. of the International Conference on Telecommunication Systems Services and Applications (TSSA'14)*, pages 1–5, Bali, Indonesia, 2014. IEEE.

455. M. Maia, J. Almeida, and V. Almeida. Identifying user behavior in online social networks. In *Proc. of the International Workshop on Social Network Systems*, pages 1–6, Glasgow, Scotland, UK, 2008. ACM.

456. J. Malbon. Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36(2):139–157, 2013. Springer.

457. F. Malliaros, M. Rossi, and M. Vazirgiannis. Locating influential nodes in complex networks. *Scientific reports*, 6:19307, 2016. Nature.

458. B. Malysiak-Mrozek, M. Stabla, and D. Mrozek. Soft and Declarative Fishing of Information in Big Data Lake. *IEEE Transactions on Fuzzy Systems*, 26(5):2732–2747, 2018. IEEE.

459. N. Mammone, L. Bonanno, S. De Salvo, S. Marino, P. Bramanti, A. Bramanti, and F.C. Morabito. Permutation disalignment index as an indirect, EEG-based, measure of brain connectivity in MCI and AD patients. *International journal of neural systems*, 27(05):1750020, 2017. World Scientific.

460. C.D. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. 2008. Cambridge University Press Cambridge.

461. G. Marra, F. Ricca, G. Terracina, and D. Ursino. Information Diffusion in a Multi-Social-Network Scenario: A framework and an ASP-based analysis. *Knowledge and Information Systems*, 48(3):619–648, 2016. Springer.

462. J. Martinez-Gil and J.F. Aldana-Montes. Semantic similarity measurement using historical google search patterns. *Information Systems Frontiers*, 15(3):399–410, 2013. Springer.

463. S. Maslov and S. Redner. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, 28(44):11103–11105, 2008. Society for Neuroscience.

464. J.N. Matias. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proc. of the International Conference on Human Factors in Computing Systems (ACM CHI 2016)*, pages 1138–1151, San Jose, CA, USA, 2016. ACM.

465. J. O. Maximo, E.J. Cadena, and R.K. Kana. The implications of brain connectivity in the neuropsychology of autism. *Neuropsychology review*, 24(1):16–31, 2014. Springer.

466. S. Mazhari, S.M. Fakhrahmad, and H. Sadeghbeygi. A user-profile-based friendship recommendation solution in social networks. *Journal of Information Science*, 41(3):284–295, 2015. SAGE.

467. D.H. McKnight and N.L. Chervany. The meanings of trust. *Technical Report MISRC (Management Information Systems Research Center) - Working Paper Series 96-04*, 1996. University of Minnesota.

468. M. McPherson, L. Smith-Lovin, and J.M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. JSTOR.

469. A.N. Medvedev, R. Lambiotte, and J.C. Delvenne. The Anatomy of Reddit: An Overview of Academic Research. In *Dynamics On and Of Complex Networks III*, pages 183–204, Cham, 2019. Springer International Publishing.

470. J. Meese. "It belongs to the Internet": Animal images, attribution norms and the politics of amateur media production. *M/C Journal*, 17(2):1–3, 2014. M/C.

471. O. Mehdi, H. Ibrahim, and L. Affendey. An approach for instance based schema matching with Google similarity and regular expression. *International Arab Journal of Information Technology*, 14(5):755–763, 2017.

472. M. Meyer. What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1):93–123, 2000. Springer.

473. E. Miguélez and R. Moreno. Research networks and inventors' mobility as drivers of innovation: evidence from Europe. *Regional Studies*, 47(10):1668–1685, 2013. Taylor & Francis.

474. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

475. T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of the International Conference on Advances in Neural Information Processing Systems (NIPS'13)*, pages 3111–3119, Lake Tahoe, NV, USA, 2013.

476. L. Militano, M. Nitti, L. Atzori, and A. Iera. Enhancing the navigability in a social network of smart objects: A Shapley-value based approach. *Computer Networks*, 103:1–14, 2016. Elsevier.

477. A.G. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

478. B.A. Miller, N. Arcolano, and N. T. Bliss. Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data. In *Proc. of the 2013 IEEE International Conference on Intelligence and Security Informatics, Seattle, WA, USA, June 4-7, 2013*, pages 179–184, 2013. IEEE.

479. Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter Spammer Detection Using Data Stream Clustering. *Information Sciences*, 260:64–73, 2014.

480. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. American Association for the Advancement of Science.

481. D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac. Internet of things: Vision, applications and research challenges. *Ad Hoc Networks*, 10(7):1497–1516, 2012. Elsevier.

482. F. Miraglia, F. Vecchio, and P. Rossini. Searching for signs of aging and dementia in EEG through network analysis. *Behavioural Brain Research*, 317:292–300, 2017. Elsevier.

483. G. Miritello, E. Moro, and R. Lara. Dynamical strength of social ties in information spreading. *Phys. Rev. E*, 83:045102, 2011. American Physical Society.

484. A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the ACM SIGCOMM International Conference on Internet Measurement (IMC'07)*, pages 29–42, San Diego, CA, USA, 2007. ACM.

485. S. Misra, R. Barthwal, and M. S. Obaidat. Community detection in an integrated Internet of Things and social network architecture. In *Proc. of IEEE Global Communications Conference (GLOBECOM 2012)*, pages 1647–1652, Anaheim, CA, USA, 2012. IEEE.

486. S. Misra, R. Barthwal, and M.S. Obaidat. Community detection in an integrated Internet of Things and social network architecture. In *Proc. of the Global Communications Conference (GLOBECOM'12)*, pages 1647–1652, Anaheim, CA, USA, 2012. IEEE.

487. F.C. Morabito, M. Campolo, D. Labate, G. Morabito, L. Bonanno, A. Bramanti, S. De Salvo, A. Marra, and P. Bramanti. A longitudinal EEG study of Alzheimer's disease progression based on a complex network approach. *International journal of neural systems*, 25(02):1550005, 2015. World Scientific.

488. F.C. Morabito, D. Labate, A. Bramanti, F. La Foresta, G. Morabito, I. Palamara, and H. Szu. Enhanced compressibility of EEG signal in Alzheimer's disease patients. *IEEE Sensors Journal*, 13(9):3255–3262, 2013. IEEE.

489. F.C. Morabito, D. Labate, G. Morabito, I. Palamara, and H. Szu. Monitoring and diagnosis of Alzheimer's disease using noninvasive compressive sensing EEG. In *SPIE Defense, Security, and Sensing*, pages 87500Y–87500Y. 2013. International Society for Optics and Photonics.

490. D.V. Moretti. Association of EEG, MRI, and regional blood flow biomarkers is predictive of prodromal Alzheimer's disease. *Neuropsychiatric disease and treatment*, 11:2779, 2015. Dove Press.

491. D. Morrison and C. Hayes. Here, have an upvote: Communication behaviour and karma on Reddit. *Informatik*, pages 2258–2268, 2013. Gesellschaft für Informatik eV.

492. A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. What yelp fake review filter might be doing? In *Proc. of the International AAAI Conference on Weblogs and Social Media (ICDSM'13)*, Boston, MA, USA, 2013.

493. M. Nakayama and Y. Wan. The cultural impact on social commerce: A sentiment analysis on yelp ethnic restaurant reviews. *Information & Management*, 56(2):271–279, 2019. Elsevier.

494. H. Nam, Y.V. Joshi, and P.K. Kannan. Harvesting brand information from social tags. *Journal of Marketing*, 81(4):88–108, 2017. SAGE.

495. A. Nandi and P.A. Bernstein. HAMSTER: Using Search Clicklogs for Schema and Taxonomy Matching. *Proceedings of the VLDB Endowment*, 2(1):181–192, 2009.

496. B. K. Narayanan and M. Nirmala. Adult content filtering: Restricting minor audience from accessing inappropriate Internet content. *Education and Information Technologies*, 23(6):2719–2735, 2018. Springer.

497. F. Naumann. Data profiling revisited. *ACM SIGMOD Record*, 42(4):40–49, 2014. ACM.

498. R. Navigli and S.P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. Elsevier.

499. B. Negash, T. Westerlund, and H. Tenhunen. Towards an interoperable Internet of Things through a web of virtual things at the Fog layer. *Future Generation Computer Systems*, 91:96–107, 2019. Elsevier.

500. N. Nesa, T. Ghosh, and I. Banerjee. Non-parametric sequence-based learning approach for outlier detection in iot. *Future Generation Computer Systems*, 82:412–421, 2018. Elsevier.

501. E. Newell, D. Jurgens, H.M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proc. of the International Conference on Web and Social Media (ICWSM 2016)*, pages 279–288, Cologne, Germany, 2016. AAAI.

502. M.E.J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001. APS.

503. M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002. APS.

504. M.E.J. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003. JSTOR.

505. M.E.J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005. Elsevier.

506. Q.V.H. Nguyen, T.T. Nguyen, V.T. Chau, T.K. Wijaya, Z. Miklos, K. Aberer, A. Gal, and M. Weidlich. SMART: A tool for analyzing and reconciling schema matching networks. In *Proc. of the International Conference on Data Engineering (ICDE'15)*, pages 1488–1491, Seoul, Korea, 2015. IEEE.

507. T.V. Nguyen, N.T. Tran, and S. Le Thanh. An anomaly-based network intrusion detection system using deep learning. In *Proc. of the 2017 International Conference on System Science and Engineering (ICSSE)*, pages 210–214, Ho Chi Minh City, Vietnam, 2017. IEEE.

508. S. Nicolazzo, A. Nocera, D. Ursino, and L. Virgili. A Privacy-Preserving Approach to Prevent Feature Disclosure in an IoT Scenario. *Future Generation Computer Systems*, 105:502–512, 2020. Elsevier.

509. H. Ning and Z. Wang. Future internet of things architecture: like mankind neural system or social organization framework? *IEEE Communications Letters*, 15(4):461–463, 2011. IEEE.

510. Z. Ning, X. Wang, X. Kong, and W. Hou. A social-aware group formation framework for information diffusion in narrowband internet of things. *IEEE Internet of Things Journal*, 5(3):1527–1538, 2018.

511. M. Nitti, L. Atzori, and I.P. Cvijikj. Network navigability in the social internet of things. In *Proc. of the International Conference on Internet of Things (WF-IoT'14)*, pages 405–410, Seoul, South Korea, 2014. IEEE.

512. M. Nitti, R. Girau, and L. Atzori. Trustworthiness management in the social internet of things. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1253–1266, 2014. IEEE.

513. M. Nitti, R. Girau, L. Atzori, A. Iera, and G. Morabito. A subjective model for trustworthiness evaluation in the social internet of things. In *Proc. of the International Conference on Personal Indoor and Mobile Radio Communications (PIMRC'12)*, pages 18–23, Sydney, Australia, 2012. IEEE.

514. A. Nocera and D. Ursino. PHIS: a system for scouting potential hubs and for favoring their "growth" in a Social Internetworking Scenario. *Knowledge-Based Systems*, 36:288–299, 2012. Elsevier.

515. P. Nokhiz and F. Li. Understanding rating behavior based on moral foundations: The case of Yelp reviews. In *Proc. of the International Conference on Big Data (Big Data 2017)*, pages 3938–3945, Boston, MA, USA, 2017. IEEE.

516. S. Oh, Y. Kim, and S. Cho. An interoperable access control framework for diverse iot platforms based on oauth and role. *Sensors*, 19(8):1884, 2019. Multidisciplinary Digital Publishing Institute.

517. Y. Okada, K. Masui, and Y. Kadobayashi. Proposal of Social Internetworking. In *Proc. of the International Human.Society@Internet Conference (HSI 2005)*, pages 114–124, Asakusa, Tokyo, Japan, 2005. Lecture Notes in Computer Science, Springer.

518. J.P. Onnela, J. Saramäki, J. Kertész, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005. APS.

519. A. Oram. *Managing the Data Lake*. Sebastopol, CA, USA, 2015. O'Reilly.

520. A. Ortiz, D. Hussein, S. Park, S. Han, and N. Crespi. The cluster between internet of things and social networks: Review and research challenges. *IEEE Internet of Things Journal*, 1(3):206–215, 2014. IEEE.

521. S. Otoum, B. Kantarci, and H. T Mouftah. Hierarchical trust-based black-hole detection in wsn-based smart grid monitoring. In *Proc. of the International Conference on Communications (ICC'17)*, pages 1–6, Paris, France, 2017. IEEE.

522. S. Otoum, B. Kantarci, and H.T Mouftah. On the feasibility of deep learning in sensor network intrusion detection. *IEEE Networking Letters*, 1(2):68–71, 2019. IEEE.

523. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Proc. of the Seventh International World-Wide Web Conference (WWW 1998)*, pages 161–172, Brisbane, Australia, 1998. Elsevier.

524. H. Haddad Pajouh, R. Javidan, R. Khayami, D. Ali, and K.R. Choo. A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in iot backbone networks. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2019. IEEE.

525. L. Palopoli, D. Rosaci, G. Terracina, and D. Ursino. A graph-based approach for extracting terminological properties from information sources with heterogeneous formats. *Knowledge and Information Systems*, 8(4):462–497, 2005.

526. L. Palopoli, D. Saccà, G. Terracina, and D. Ursino. Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):271–294, 2003.

527. L. Palopoli, G. Terracina, and D. Ursino. DIKE: a system supporting the semi-automatic construction of Cooperative Information Systems from heterogeneous databases. *Software Practice & Experience*, 33(9):847–884, 2003.

528. L. Palopoli, G. Terracina, and D. Ursino. Experiences using DIKE, a system for supporting cooperative information system and data warehouse design. *Information Systems*, 28(7):835–865, 2003.

529. A. Parikh, C. Behnke, M. Vorvoreanu, B. Almanza, and D. Nelson. Motives for reading and articulating user-generated restaurant reviews on yelp. com. *Journal of Hospitality and Tourism Technology*, 5(2):160–176, 2014. Emerald Group Publishing Limited.

530. A.A. Parikh, C. Behnke, B. Almanza, D. Nelson, and M. Vorvoreanu. Comparative content analysis of professional, semi-professional, and user-generated restaurant reviews. *Journal of Foodservice Business Research*, 20(5):497–511, 2017. Taylor & Francis.

531. J.-R. Park and Y. Tosaka. Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly*, 48(8):696–715, 2010. Routledge.

532. J.R. Park. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, 47(3-4):213–228, 2009. Routledge.

533. M. Park, H. Oh, and K. Lee. Security risk measurement for information leakage in iot-based smart homes from a situational awareness perspective. *Sensors*, 19(9):2148, 2019. Multidisciplinary Digital Publishing Institute.

534. E. Parvinnia, M. Sabeti, M. Jahromi, and R. Boostani. Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *Journal of King Saud University-Computer and Information Sciences*, 26(1):1–6, 2014. Elsevier.

535. K. Passi, L. Lane, S.K. Madria, B.C. Sakamuri, M.K. Mohania, and S.S. Bhowmick. A model for XML Schema integration. In *Proc. of the International Conference on E-Commerce and Web Technologies (EC-Web 2002)*, pages 193–202, Aix-en-Provence, France, 2002. Lecture Notes in Computer Science, Springer.

536. M. Patella and P. Ciaccia. Approximate similarity search: A multi-faceted problem. *Journal of Discrete Algorithms*, 7(1):36–48, 2009. Elsevier.

537. G. Peeters and J. Czapinski. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European review of social psychology*, 1(1):33–60, 1990. Taylor & Francis.

538. M. Pennacchiotti and A. Popescu. Democrats, republicans and starbucks afficionados: user classification in Twitter. In *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 430–438, San Diego, CA, USA, 2011. ACM.

539. C. Perera, Y. Qin, J. Estrella, S. Reiff-Marganiec, and A. Vasilakos. Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys (CSUR)*, 50(3):32, 2017. ACM.

540. C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context aware computing for the internet of things: A survey. *IEEE communications surveys & tutorials*, 16(1):414–454, 2013. IEEE.

541. C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context aware computing for the Internet of Things: A survey. *IEEE Communications Surveys & Tutorials*, 16(1):414–454, 2014. IEEE.

542. R. Petersen. Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine*, 256(3):183–194, 2004. Wiley Online Library.

543. R.C. Phillips and D. Gorse. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *Proc. of the International Symposium Series on Computational Intelligence (SSCI'17)*, pages 1–7, Honolulu, HI, USA, 2017. IEEE.

544. R. Di Pietro, X. Salleras, M. Signorini, and E. Waisbard. A blockchain-based Trust System for the Internet of Things. In *Proc. of the ACM International Symposium on Access Control Models and Technologies (SACMAT'18)*, pages 77–83, Indianapolis, IN, USA, 2018. ACM.

545. G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information processing & management*, 12(5):297–312, 1976. Elsevier.

546. A.P. Plageras, C. Stergiou, G. Kokkonis, K.E. Psannis, Y. Ishibashi, B.G. Kim, and B.B. Gupta. Efficient large-scale medical data (ehealth big data) analytics in internet of things. In *Proc. of the Conference on Business informatics (CBI'17)*, volume 2, pages 21–27, Thessaloniki, Greece, 2017. IEEE.

547. M. Plantié and M. Crampes. Survey on social community detection. In *Social media retrieval*, pages 65–85. 2013. Springer.

548. S.S. Poil, W. De Haan, W.M. van der Flier, H.D. Mansvelder, P. Scheltens, and K. Linkenkaer-Hansen. Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. *Frontiers in aging neuroscience*, 5:58, 2013. Frontiers.

549. S.C. Ponten, L. Douw, F. Bartolomei, J.C. Reijneveld, and C.J. Stam. Indications for network regularization during absence seizures: weighted and unweighted graph theoretical analyses. *Experimental Neurology*, 217(1):197–204, 2009. Elsevier.

550. M. Potamias. The warm-start bias of Yelp ratings. *arXiv preprint arXiv:1202.5713*, 2012.

551. D.V. Prasad, S. Madhusudanan, and S. Jaganathan. uCLUST - A new algorithm for clustering unstructured data. *ARPN Journal of Engineering and Applied Sciences*, 10(5):2108–2117, 2015.

552. R. Puzis, Y. Elovici, and S. Dolev. Fast algorithm for successive computation of group betweenness centrality. *Physical Review E*, 76(5):056709, 2007. APS.

553. Z. Qasem, M. Jansen, T. Hecking, and H.U. Hoppe. On the Detection of Influential Actors in Social Media. In *Proc. of the International Conference on Signal-Image Technology & Internet-Based Systems, (SITIS 2015)*, pages 421–427, Bangkok, Thailand, 2015. IEEE Computer Society.

554. Y. Qin, Q. Sheng, N. Falkner, S. Dustdar, H. Wang, and A. Vasilakos. When things matter: A survey on data-centric internet of things. *Journal of Network and Computer Applications*, 64:137–153, 2016. Elsevier.

555. J. Qiu, Y. Li, and Z. Lin. Does Social Commerce Work in Yelp? An Empirical Analysis of Impacts of Social Relationship on the Purchase Decision-making. In *Proc. of the Pacific Asia Conference on Information Systems (PACIS'18)*, page 16, Yokohama, Japan, 2018.

556. J. Qiu, Y. Li, and Z. Lin. Detecting Social Commerce: An Empirical Analysis on Yelp. *Journal of Electronic Commerce Research*, 21(3):168–179, 2020. Journal of Electronic Commerce Research.

557. J. Quevedo, M. Antunes, D. Corujo, D. Gomes, and R.L. Aguiar. On the application of contextual iot service discovery in information centric networks. *Computer Communications*, 89:117–127, 2016. Elsevier.

558. E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

559. Y. Rahulamathavan, R.C. Phan, M. Rajarajan, S. Misra, and A. Kondoz. Privacy-preserving blockchain based IoT ecosystem using attribute-based encryption. In *Proc. of the International Conference on Advanced Networks and Telecommunications Systems (ANTS'17))*, pages 1–6, Bhubaneswar, India, 2017. IEEE.

560. A.R.G. Ramirez, I. González-Carrasco, G.H. Jasper, A.L. Lopez, J.L. Lopez-Cuadrado, and A. García-Crespo. Towards human smart cities: internet of things for sensory impaired individuals. *Computing*, 99(1):107–126, 2017. Springer.

561. J. Ramirez, J. Gorriz, D. Salas-Gonzalez, A. Romero, M. Lopez, I. Alvarez, and M. Gomez-Rio. Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. *Information Sciences*, 237:59–72, 2013. Elsevier.

562. G. Ramponi, M. Brambilla, S. Ceri, F. Daniel, and M. Di Giovanni. Content-based characterization of online social communities. *Information Processing & Management*, page 102133, 2019. Elsevier.

563. A. Rehman, S. U. Rehman, and H. Raheem. Sinkhole attacks in wireless sensor networks: A survey. *Wireless Personal Communications*, pages 1–23, 2018. Springer.

564. P. Resnick and R. Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce*, pages 127–157. 2002. Emerald Group Publishing Limited.

565. M. Riondato and E. Kornaropoulos. Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475, 2016. Springer.

566. A. Rodriguez, A. Tosyali, B. Kim, J. Choi, J. Lee, B. Coh, and M. Jeong. Patent Clustering and Outlier Ranking Methodologies for Attributed Patent Citation Networks for Technology Opportunity Discovery. *IEEE Transactions on Engineering Management*, 63(4):426–437, 2016. IEEE.

567. R. Roman, P. Najera, and J. Lopez. Securing the Internet of Things. *Computer*, 1(9):51–58, 2011. IEEE.

568. R. Roman, J. Zhou, and J. Lopez. On the features and challenges of security and privacy in distributed Internet of Things. *Computer Networks*, 57(10):2266–2279, 2013. Elsevier.

569. D.M. Romero, W. Galuba, S. Asur, and B.A. Huberman. Influence and passivity in social media. In *Proc. of the International Conference on World Wide Web (WWW'11)*, pages 113–114, Hyderabad, India, 2011. ACM.

570. S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010. Wiley, New York.

571. M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010. Elsevier.

572. S.J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. 2016. Malaysia; Pearson Education Limited,.

573. S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *The Cryptography Mailing List*, 2008.

574. J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

575. G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966. Springer.

576. M. Sahlgren and R. Cöster. Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In *Proc. of the International Conference on Computational Linguistics (COLING'04)*, page 487, Geneva, Switzerland, 2004.

577. M. Sahlgren and J. Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341, 2005.

578. Y.B. Saied, A. Olivereau, D. Zeghlache, and M. Laurent. Trust management system design for the Internet of Things: A context-aware and multi-service approach. *Computers & Security*, 39:351–365, 2013. Elsevier.

579. V. Sakkalis. Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in biology and medicine*, 41(12):1110–1117, 2011. Elsevier.

580. Y. Saleem, N. Crespi, M. H. Rehmani, R. Copeland, D. Hussein, and E. Bertin. Exploitation of social IoT for recommendation services. In *Proc. of the IEEE World Forum on Internet of Things (WF-IoT 2016)*, pages 359–364, Reston, VA, USA, 2016. IEEE Computer Society.

581. Y. Saleem, N. Crespi, M.H. Rehmani, R. Copeland, D. Hussein, and E. Bertin. Exploitation of social IoT for recommendation services. In *Proc. of the World Forum on Internet of Things (WF-IoT'16)*, pages 359–364, Reston, VA, USA, 2016. IEEE.

582. A. Salinca. Business reviews classification using sentiment analysis. In *Proc. of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC'15)*, pages 247–250, Timisoara, Romania, 2015. IEEE.

583. S. Salvador, P. Chan, and J. Brodie. Learning states and rules for time series anomaly detection. In *Proc. of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*, pages 306–311, 2004. AAAI Press.

584. Z. Sankari, H. Adeli, and A. Adeli. Intrahemispheric, interhemispheric, and distal EEG coherence in Alzheimer?s disease. *Clinical Neurophysiology*, 122(5):897–906, 2011. Elsevier.

585. J. Santos, T. Wauters, B. Volckaert, and F. De Turck. Resource provisioning in fog computing: From theory to practice. *Sensors*, 19(10):2238, 2019. Multidisciplinary Digital Publishing Institute.

586. D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang. Anomaly detection in online social networks. *Social Networks*, 39:62–70, 2014. Elsevier.

587. C. Savaglio, G. Fortino, and M. Zhou. Towards interoperable, cognitive and autonomic IoT systems: An agent-based approach. In *Proc. of the World Forum on Internet of Things (WF-IoT'16)*, pages 58–63, Reston, VA, USA, 2016. IEEE.

588. A. Saxena, R. Gera, I. Bermudez, D. Cleven, E.T. Kiser, and T. Newlin. Twitter Response to Munich July 2016 Attack: Network Analysis of Influence. *Frontiers in Big Data*, 2:17, 2019. Frontiers.

589. S.F. Sayeedunnissa, A.R. Hussain, and M.A. Hameed. Supervised Opinion Mining of Social Network Data Using a Bag-of-Words Approach on the Cloud. In *Proc. of the International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA'12)*, pages 299–309, Gwalior, India, 2012.

590. S.E. Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007. Elsevier.

591. S. Schiaffino and A. Amandi. Intelligent user profiling. In *Artificial Intelligence An International Perspective*, pages 193–216. 2009. Springer.

592. P. Schirmer, T. Papenbrock, I. Koumarelas, and F. Naumann. Efficient discovery of matching dependencies. *ACM Transactions on Database Systems*, 45(3):1–33, 2020. ACM.

593. D. Schuff and S. Mudambi. What makes a helpful online review? A study of customer reviews on Amazon.com. *Social Science Electronic Publishing*, 34(1):185–200, 2012. Elsevier.

594. A. Seidel. Citation system for patent office. *Journal of the Patent Office Society*, 31(554):26–31, 1949.

595. V. Setyani, Y.Q. Zhu, A.N. Hidayanto, P.I. Sandhyaduhita, and B. Hsiao. Exploring the psychological mechanisms from personalized advertisements to urge to buy impulsively on social media. *International Journal of Information Management*, 48:96–107, 2019. Elsevier.

596. M. Shao, J. Li, F. Chen, H. Huang, S. Zhang, and X. Chen. An efficient approach to event detection and forecasting in dynamic multivariate social media networks. In *Proc. of the 26th International Conference on World Wide Web*, pages 1631–1639, Perth, Australia, 2017. ACM.

597. V. Sharma, I. You, D.N. Jayakody, and M. Atiquzzaman. Cooperative trust relaying and privacy preservation via edge-crowdsourcing in social Internet of Things. *Future Generation Computer Systems*, 92:759–776, 2019. Elsevier.

598. V. Sharma, I. You, and R. Kumar. Isma: Intelligent sensing model for anomalies detection in cross platform osns with a case study on iot. *IEEE Access*, 5:3284–3301, 2017. IEEE.

599. A. Sheikhahmadi and M. Nematbakhsh. Identification of multi-spreader users in social networks for viral marketing. *Journal of Information Science*, 43(3):412–423, 2017. SAGE Publications.

600. S. Shekhar, C. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA, August 26-29, 2001*, pages 371–376, 2001. ACM.

601. L. Shen, B. Xiong, and J. Hu. Research status, hotspots and trends for information behavior in China using bibliometric and co-word analysis. *Journal of Documentation*, 73(4):618–633, 2017. Emerald Publishing Limited.

602. M. Shen, X. Tang, L. Zhu, X. Du, and M. Guizani. Privacy-Preserving Support Vector Machine Training over Blockchain-Based Encrypted IoT Data in Smart Cities. *IEEE Internet of Things Journal*, Forthcoming. IEEE.

603. Q. Shen and R. Carolyn. The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In *Proc. of the International Workshop on Abusive Language Online (ALW 2019)*, pages 58–69, Florence, Italy, 2019. Association for Computational Linguistics.

604. W. Shen, Y.J. Hu, and J.R. Ulmer. Competing for Attention: An Empirical Study of Online Reviewers' Strategic Behavior. *MIS Q.*, 39(3):683–696, 2015. Management Information Systems Research Center.

605. E. Shi, H. Chan, E. Rieffel, R. Chow, and D. Song. Privacy-preserving aggregation of time-series data. In *Proc. of the Annual Network & Distributed System Security Symposium (NDSS'11)*, San Diego, CA, USA, 2011. Internet Society.

606. X. Shi, B.L. Tseng, and L.A. Adamic. Looking at the blogosphere topology through different lenses. In *Proc. of the International Conference on Weblogs and Social Media (ICWSM'07)*, Boulder, CO, USA, 2007.

607. N. Shrivastava, A. Majumder, and R. Rastogi. Mining (social) network graphs to detect random link attacks. In *Proc. of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, Mexico*, pages 486–495, 2008. IEEE Computer Society.

608. S. Sicari, A. Rizzardi, L.A. Grieco, and A. Coen-Porisini. Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks*, 76:146–164, 2015. Elsevier.

609. D. Simon. *Evolutionary optimization algorithms*. 2013. John Wiley & Sons.

610. Ö. Şimşek and D. Jensen. Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35):12758–12762, 2008. National Academy of Sciences.

611. P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier. Evolution of Reddit: From the Front Page of the Internet to a Self-Referential Community? In *Proc. of the International Conference on World Wide Web (WWW 2014)*, page 517–522, Seoul, Korea, 2014. ACM.

612. J. Singh. Collaborative networks as determinants of knowledge diffusion patterns. *Management Science*, 51(5):756–770, 2005. INFORMS.

613. J. Singh. Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Research Policy*, 37(1):77–96, 2008. Elsevier.

614. R. Singh, J. Woo, N. Khan, J. Kim, H.J. Lee, H.A. Rahman, J. Park, J. Suh, M. Eom, and N. Gudigantala. Applications of machine learning models on yelp data. *Asia Pacific Journal of Information Systems*, 29(1):117–143, 2019.

615. J.V. Sobral, J.J. Rodrigues, R.A. Rabêlo, J. Al-Muhtadi, and V. Korotaev. Routing protocols for low power and lossy networks in internet of things applications. *Sensors*, 19(9):2144, 2019. Multidisciplinary Digital Publishing Institute.

616. A. Soliman, J. Hafer, and F. Lemmerich. A Characterization of Political Communities on Reddit. In *Proc. of the ACM Conference on Hypertext and Social Media (HT'19)*, page 259–263, Hof, Germany, 2019. ACM.

617. V.N. Soloviev and A. Belinskiy. Complex systems theory and crashes of cryptocurrency market. In *Proc. of the International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications (ICTERI'18)*, pages 276–297, Kyiv, Ukraine, 2018. Springer.

618. S. Somin, G. Gordon, and Y. Altshuler. Network analysis of ERC20 tokens trading on ethereum blockchain. In *Proc. of the International Conference on Complex Systems (ICCS'18)*, pages 439–450, Cambridge, MA, USA, 2018. Springer.

619. O. Sporns and Rolf R. Kötter. Motifs in brain networks. *PLoS Computational Biology*, 2(11):e369, 2004. Public Library of Science.

620. C.J. Stam, B. Jones, G. Nolte, M. Breakspear, and P. Scheltens. Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral cortex*, 17(1):92–99, 2007. Oxford Univ Press.

621. K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, 1989. Elsevier.

622. C. Sternitzke, A. Bartkowski, and R. Schramm. Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, 30(2):115–131, 2008. Elsevier.

623. I. Stojmenovic and S. Olariu. Data-centric protocols for wireless sensor networks. *Handbook of sensor networks: algorithms and architectures*, pages 417–456, 2005. Wiley.

624. F. Su, J. Wang, B. Deng, X.L. Wei, Y.Y. Chen, C. Liu, and H.Y. Li. Adaptive control of Parkinson's state based on a nonlinear computational model with unknown parameters. *International Journal of Neural Systems*, 25(01):1450030, 2015. World Scientific.

625. R.P. Subbanarasimha, S. Srinivasa, and S. Mandyam. Invisible Stories That Drive Online Social Cognition. *IEEE Transactions on Computational Social Systems*, pages 1–14, 2020. IEEE.

626. G. Suciu, S. Halunga, A. Vulpe, and V. Suciu. Generic platform for IoT and cloud computing interoperability study. In *Proc. of the International Symposium on Signals, Circuits and Systems (ISSCS'13)*, pages 1–4, Iasi, Romania, 2013. IEEE.

627. S. Sudrich, J. De Melo Borges, and M. Beigl. Anomaly detection in evolving heterogeneous graphs. In *Proc. of the International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and So-*

cial Computing (CPSCom) and IEEE Smart Data (SmartData), pages 1147–1149, Exeter, UK, 2017. IEEE Computer Society.

628. B. Sun and V.T.Y. Ng. Identifying influential users by their postings in social networks. In *Proc. of the International Workshop on Modeling Social Media (MSM 2012)*, pages 1–8, Milwaukee, WI, USA, 2012. ACM.
629. H. Sun, N. Ruan, and H. Liu. Ethereum Analysis via Node Clustering. In *Proc. of the International Conference on Network and System Security (NSS'19)*, pages 114–129, Sapporo, Japan, 2019. Springer.
630. Y. Sun and J.D.G. Paule. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards*, 2(1):5, 2017. SpringerOpen.
631. Y. Sun, J. Zhang, Y. Xiong, and G. Zhu. Data security and privacy in cloud computing. *International Journal of Distributed Sensor Networks*, 10(7):190903, 2014. SAGE Publications Sage.
632. S. Sussman, R. Garcia, T. B. Cruz, L. Baezconde-Garbanati, M. A. Pentz, and J. B Unger. Consumers' perceptions of vape shops in southern california: an analysis of online yelp reviews. *Tobacco induced diseases*, 12(1):22, 2014. BioMed Central.
633. L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.
634. M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010. National Acad Sciences.
635. J. Szymanski. Comparative Analysis of Text Representation Methods Using Classification. *Cybernetics and Systems*, 45(2):180–199, 2014.
636. C. Tan and L. Lee. All Who Wander: On the Prevalence and Characteristics of Multi-Community Engagement. In *Proc. of the International Conference on World Wide Web (WWW 2015)*, page 1056–1066, Florence, Italy, 2015. ACM.
637. A. Tani, L. Candela, and D. Castelli. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, 49(6):1194–1205, 2013.
638. K. Tei and L. Gurgen. ClouT: Cloud of things for empowering the citizen clout in smart cities. In *Proc. of the World Forum on Internet of Things (WF-IoT'2014)*, pages 369–370, Seoul, South Korea, 2014. IEEE.
639. M. Thelwall. Can social news websites pay for content and curation? The SteemIt cryptocurrency model. *Journal of Information Science*, 44(6):736–751, 2018. SAGE Publications.
640. M.T.P.M.B. Tiago and J.M.C. Veríssimo. Digital marketing and social media: Why bother? *Business horizons*, 57(6):703–708, 2014. Elsevier.
641. K. Tiidenberg. Boundaries and conflict in a NSFW community on tumblr: The meanings and uses of selfies. *New Media & Society*, 18(8):1563–1578, 2016. Sage Publications.
642. P.L. Ting, S.L. Chen, H. Chen, and W.C. Fang. Using big data and text analytics to understand how customer experiences posted on yelp. com impact the hospitality industry. *Contemporary Management Research*, 13(2), 2017. Academy of Taiwan Information Systems Research.

643. K. Toyoda, T. Ohtsuki, and P.T. Mathiopoulos. Multi-class bitcoin-enabled service identification based on transaction history summarization. In *Proc. of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 1153–1160, Halifax, NS, Canada, 2018. IEEE.

644. N.B. Truong, T.W. Um, and G.M. Lee. A reputation and knowledge based trust service platform for trustworthy social internet of things. In *Proc. of the International Conference on Innovations in Clouds, Internet and Networks (ICIN '16)*, Paris, France, 2016.

645. N.B. Truong, T.W. Um, B. Zhou, and G.M. Lee. From personal experience to global reputation for trust evaluation in the social internet of things. In *Proc. of the International IEEE Global Communications Conference (GLOBECOM'17)*, pages 1–7, Singapore, 2017. IEEE.

646. C. Tsai, C. Lai, and A. Vasilakos. Future Internet of Things: open issues and challenges. *Wireless Networks*, 20(8):2201–2217, 2014. Springer.

647. M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. Sebastopol, CA, USA, 2011. O'Reilly Media, Inc.

648. T. Tucker. Online word of mouth: characteristics of Yelp.com reviews. *Elon Journal of Undergraduate Research in Communications*, 2(1):37–42, 2011.

649. D. Ursino and L. Virgili. An approach to evaluate trust and reputation of things in a Multi-IoTs scenario. *Computing*, 102:2257–2298, 2020. Springer.

650. D. Ursino and L. Virgili. Humanizing IoT: defining the profile and the reliability of a thing in a Multi-IoT scenario. *Towards Social Internet of Things: Enabling Technologies, Architectures and Applications. Studies in Computational Intelligence*, 846:51–76, 2020. Springer Nature.

651. V. Ďurčeková, L. Schwartz, V. Hottmar, and B. Adamec. Detection of Attacks Causing Network Service Denial. *Advances in Military Technology*, 13(1), 2018.

652. C.J. Van Rijsbergen. *Information Retrieval*. 1979. Butterworth.

653. J.M. Vanerio and P Casas. Ensemble-learning approaches for network security and anomaly detection. In *Proc. of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks, Big-DAMA@SIGCOMM 2017*, pages 1–6, Los Angeles, CA, USA, 2017. ACM.

654. F. Vecchio, F. Miraglia, C. Marra, D. Quaranta, M. Vita, P. Bramanti, and P. Rossini. Human brain networks in cognitive decline: a graph theoretical analysis of cortical connectivity from EEG data. *Journal of Alzheimer's Disease*, 41(1):113–127, 2014. IOS Press.

655. A. M. Vegni, V. Loscri, and A. Benslimane. SOLVER: A Framework for the Integration of Online Social Networks with Vehicular Social Networks. *IEEE Network*, 34(1):204–213, 2020. IEEE.

656. O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, I. S. Jubert, M. Mazura, M. Harrison, M. Eisenhauer, and P. Doody. Internet of things strategic research roadmap. *Internet of things-global technological and societal trends*, 1(2011):9–52, 2011. River Publishers.

657. B. Verspagen. Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(01):93–115, 2007. World Scientific.

658. F. Vialatte, A. Cichocki, G. Dreyfus, T. Musha, S.L. Shishkin, and R. Gervais. Early detection of Alzheimer's disease by blind source separation, time frequency representation, and bump modeling of EEG signals. In *Proc. of the International Conference on Artificial Neural Networks (ICANN'05)*, pages 683–692, Warsaw, Poland, 2005. Lecture Notes in Computer Science, Springer.

659. J.R. Villar, P. Vergara, M. Menéndez, E. de la Cal, V.M. González, and J. Sedano. Generalized Models for the Classification of Abnormal Movements in Daily Life and its Applicability to Epilepsy Convulsion Recognition. *International Journal of Neural Systems*, 26(06):1650037, 2016. World Scientific.

660. C. Villavicencio, S. Schiaffino, J.A. Diaz-Pace, and A. Monteserin. Group recommender systems: A multi-agent solution. *Knowledge-Based Systems*, 164:436–458, 2019. Elsevier.

661. B. Viswanath, A. Mislove, M. Cha, and K.P. Gummadi. On the evolution of user interaction in Facebook. In *Proc. of the ACM Workshop on Online Social Networks (WOSN'09)*, pages 37–42, Barcelona, Spain, 2009. ACM.

662. C.S. Wagner and L. Leydesdorff. Network structure, self-organization, and the growth of the internationalcollaboration in science. *Research Policy*, 34(10):1608–1618, 2005. Elsevier.

663. J. Wan, J. Liu, Z. Shao, A. Vasilakos, M. Imran, and K. Zhou. Mobile crowd sensing for traffic prediction in internet of vehicles. *Sensors*, 16(1):88, 2016. Multidisciplinary Digital Publishing Institute.

664. J. Wan, S. Tang, Z. Shu, D. Li, S. Wang, M. Imran, and A. Vasilakos. Software-defined industrial internet of things in the context of industry 4.0. *IEEE Sensors Journal*, 16(20):7373–7380, 2016. IEEE.

665. F. Wang, Z. Wang, Z. Li, and J.R. Wen. Concept-based Short Text Classification and Ranking. In *Proc. of the International Conference on Information and Knowledge Management (CIKM'14)*, pages 1069–1078, Shangai, China, 2014. ACM.

666. L. Wang. Using the relationship of shared neighbors to find hierarchical overlapping communities for effective connectivity in IoT. In *Proc. of the International Conference on Pervasive Computing and Applications (ICPCA'11)*, pages 400–406, Port Elizabeth, South Africa, 2011. IEEE.

667. L. Wang, C. Zhu, Y. He, Y. Zang, Q. Cao, H. Zhang, Q. Zhong, and Y. Wang. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Human Brain Mapping*, 30(2):638–649, 2009. Wiley Online Library.

668. L.H. Wang, R.C. Bucelli, E. Patrick, D. Rajderkar, E. Alvarez III, M.M. Lim, G. DeBruin, V. Sharma, S. Dahiya, R.E. Schmidt an T.S. Benzinger, B.A. Ward, and B.M. Ances. Role of magnetic resonance imaging, cerebrospinal fluid, and electroencephalogram in diagnosis of sporadic Creutzfeldt-Jakob disease. *Journal of Neurology*, 260(2):498–506, 2013. Springer.

669. N. Wang, H. Wang, Y. Jia, and Y. Yin. Explainable recommendation via multi-task learning in opinionated text data. In *Proc. of the International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*, pages 165–174, Ann Arbor, MI, USA, 2018. ACM.

670. P. Wang and Y. Wen. Speculative bubbles and financial crises. *American Economic Journal: Macroeconomics*, 4(3):184–221, 2012.

671. R. Wang, J. Wang, H. Yu, X. Wei, C. Yang, and B. Deng. Power spectral density and coherence analysis of Alzheimer's EEG. *Cognitive neurodynamics*, 9(3):291–304, 2015. Springer.

672. S. Wang, Y. Du, and Y. Deng. A new measure of identifying influential nodes: Efficiency centrality. *Communications in Nonlinear Science and Numerical Simulation*, 47:151 – 163, 2017. Elsevier.

673. T. Wang, G. Zhang, A. Liu, M. Z. A. Bhuiyan, and Q. Jin. A secure iot service architecture with an efficient balance dynamics based on cloud and edge computing. *IEEE Internet of Things Journal*, 6(3):4831–4843, 2019. IEEE.

674. Y. Wang and J. Vassileva. Toward trust and reputation based web service selection: A survey. *International Transactions on Systems Science and Applications*, 3(2):118–132, 2007.

675. Z. Wang. A privacy-preserving and accountable authentication protocol for IoT end-devices with weaker identity. *Future Generation Computer Systems*, 82:342–348, 2018. Elsevier.

676. S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge, UK, 1994. Cambridge University Press.

677. D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. Nature Publishing Group.

678. J. Weng, E. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proc. of the ACM International Conference on Web Search and Data Mining (WSDM 2010)*, pages 261–270, New York, NY, USA, 2010. ACM.

679. T. Weninger. An exploration of submissions and discussions in social news: mining collective intelligence of Reddit. *Social Network Analysis and Mining*, 4:173–192, 2014. Springer.

680. D.R. White and S.P. Borgatti. Betweenness centrality measures for directed graphs. *Social Networks*, 16(4):335–346, 1994. Elsevier.

681. D. Wikström. A universally composable mix-net. In *Proc. of the Theory of Cryptography Conference (TCC'04)*, pages 317–335, Cambridge, MA, USA, 2004. Springer.

682. F. Wilcoxon. Individual Comparisons by Ranking Methods. In *Breakthroughs in statistics*, pages 196–202. 1992. Springer.

683. G. Wittenbaum, A. Hubbell, and C. Zuckerman. Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of personality and social psychology*, 77(5):967, 1999. American Psychological Association.

684. World Intellectual Property Organization. *Patent-Based Technology Analysis Report - Alternative Energy Technology*. Geneva, Switzerland, 2009.

685. D. Wu, A. Taly, A. Shankar, and D. Boneh. Privacy discovery and authentication for the Internet of Things. In *Proc. of the European Symposium on Research in Computer Security (ESORICS 2016)*, pages 301–319, Heraklion, Crete, Greece, 2016. Springer.

686. P. Wu, Z. Lu, Q. Zhou, Z. Lei, X. Li, M. Qiu, and P.C.K. Hung. Bigdata logs analysis based on seq2seq networks for cognitive Internet of Things. *Future Generation Computer Systems*, 90:477–488, 2019. Elsevier.

687. Y. Wu, H. Huang, N. Wu, Y. Wang, M.Z.A. Bhuiyan, and T. Wang. An incentive-based protection and recovery strategy for secure big data in social networks. *Information Sciences*, 508:79–91, 2020. Elsevier.

688. K. Xu, Y. Qu, and K. Yang. A tutorial on the internet of things: From a heterogeneous network integration perspective. *IEEE Network*, 30(2):102–108, 2016. IEEE.

689. Y. Xu, H. Xu, D. Zhang, and Y. Zhang. Finding overlapping community from social networks based on community forest model. *Knowledge-Based Systems*, 109:238–255, 2016. Elsevier.

690. Q. Xuan, X. Shu, Z. Ruan, J. Wang, C. Fu, and G. Chen. A self-learning information diffusion model for smart social networks. *IEEE Transactions on Network Science and Engineering*, 7(3):1466–1480, 2019. IEEE.

691. O. Yagan, D. Qian, J. Zhang, and D. Cochran. Conjoining speeds up information diffusion in overlaying social-physical networks. *IEEE Journal on Selected Areas in Communications*, 31(6):1038–1048, 2013. IEEE.

692. J.Z. Yan. Abnormal cortical functional connections in Alzheimer's disease: analysis of inter-and intra-hemispheric EEG coherence. *Journal of Zhejiang University Science B*, 6(4):259–264, 2005. Springer.

693. S.R. Yan, X.L. Zheng, Y. Wang, W.W. Song, and W.Y. Zhang. A graph-based comprehensive reputation model: Exploiting the social context of opinions to enhance trust in social commerce. *Information Sciences*, 318:51–72, 2015. Elsevier.

694. Z. Yan, P. Zhang, and A.V. Vasilakos. A survey on trust management for Internet of Things. *Journal of Network and Computer Applications*, 42:120–134, 2014. Elsevier.

695. G. Yang, G. Li, C. Li, Y. Zhao, J. Zhang, T. Liu, D. Chen, and M. Huang. Using the comprehensive patent citation network (CPC) to evaluate patent value. *Scientometrics*, 105(3):1319–1346, 2015. Springer.

696. W. Yang, Y. Wang, Z. Lai, Y. Wan, and Z. Cheng. Security Vulnerabilities and Countermeasures in the RPL-based Internet of Things. In *Proc. of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC'18)*, pages 49–495, ZhengZhou, China, 2018. IEEE.

697. Y. Yang, N. Chawla, Y. Sun, and J. Hani. Predicting links in multi-relational and heterogeneous networks. In *Proc. of the International Conference on Data Mining (ICDM'12)*, pages 755–764, Bruxelles, Belgium, 2012. IEEE.

698. Z. Yang, A.X. Cui, and T. Zhou. Impact of heterogeneous human activities on epidemic spreading. *Physica A: Statistical Mechanics and its Applications*, 390(23-24):4543–4548, 2011. Elsevier.

699. L. Yao, Q. Z. Sheng, A.H.H. Ngu, and X. Li. Things of interest recommendation by leveraging heterogeneous relations in the internet of things. *ACM Transaction on Internet Technology*, 16(2):9:1–9:25, 2016.

700. O.S. Yaya, A.E. Ogbonna, and O.E. Olubusoye. How persistent and dynamic inter-dependent are pricing of Bitcoin to other cryptocurrencies before and after 2017/18 crash? *Physica A: Statistical Mechanics and its Applications*, 531:121732, 2019. Elsevier.

701. O.S. Yaya, E.A. Ogbonna, and R. Mudida. Market Efficiency and Volatility Persistence of Cryptocurrency during Pre-and Post-Crash Periods of Bitcoin: Evidence based on Fractional Integration. *International Journal of Finance and Economics*, 2020. John Wiley & Sons.

702. Y. Ye and C.C. Chiang. A parallel apriori algorithm for frequent itemsets mining. In *Proc. of the International Conference on Software Engineering Research, Management and Applications (SERA'06)*, pages 87–94, 2006. IEEE.

703. D. Yin, S. Mitra, and H. Zhang. When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Information Systems Research*, 27(1):131–144, 2016. INFORMS.

704. J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander. Where is current research on blockchain technology? A systematic review. *PloS one*, 11(10):e0163477, 2016. PloS ONE.

705. M. Yu, A. Gouw, A. Hillebrand, B. Tijms, C. Stam, E. van Straaten, and Y. Pijnenburg. Different functional connectivity and network topology in behavioral variant of frontotemporal dementia and Alzheimer's disease: an EEG study. *Neurobiology of aging*, 42:150–162, 2016. Elsevier.

706. W. Yu, J. Li, M. Z. A. Bhuiyan, R. Zhang, and J. Huai. Ring: Real-time emerging anomaly monitoring system over text streams. *IEEE Transactions on Big Data*, 5(4):506–519, 2019. IEEE.

707. Y. Yu, H. Yan, H. Guan, and H. Zhou. Deephttp: Semantics-structure model with attention for anomalous http traffic detection and pattern mining. *CoRR*, abs/1810.12751, 2018. IEEE.

708. M. Zamani, J. Saia, M. Movahedi, and J. Khoury. Towards provably-secure scalable anonymous broadcast. In *Proc. of the International Workshop on Free and Open Communications on the Internet (FOCI'13)*, Washington, D.C., USA, 2013.

709. H. Zardi, L.B. Romdhane, and Z. Guessoum. Efficiently mining community structures in weighted social networks. *International Journal of Data Mining, Modelling and Management*, 8(1):32–61, 2016. Inderscience Publishers (IEL).

710. B.B. Zarpelão, R.S. Miani, C.T. Kawakani, and S.C. de Alvarenga. A survey of intrusion detection in internet of things. *Journal of Network and Computer Applications*, 84:25–37, 2017. Elsevier.

711. C. Zhang, L. Zhu, C. Xu, K. Sharif, X. Du, and M. Guizani. LPTD: Achieving lightweight and privacy-preserving truth discovery in CIoT. *Future Generation Computer Systems*, 90:175–184, 2019. Elsevier.

712. D. Zhang, J. Yin, X. Zhu, and C. Zhang. User Profile Preserving Social Network Embedding. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'17)*, pages 3378–3384, Melbourne, Australia, 2017. ijcai.org.

713. J. Zhang, W. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Community Identity and User Engagement in a Multi-Community Landscape. In *Proc. of the International Conference on Web and Social Media (ICWSM 2017)*, pages 377–386, Montreal, Canada, 2017. AAAI.

714. K.Z. Zhang, S.J. Zhao, C.M. Cheung, and M.K. Lee. Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems*, 67:78–89, 2014. Elsevier.

715. M. Zhang, L. Guo, M. Hu, and W. Liu. Influence of customer engagement with company social networks on stickiness: Mediating effect of customer value creation. *International Journal of Information Management*, 37(3):229–240, 2017. Elsevier.

716. Y. Zhang, D. Raychadhuri, L. Grieco, E. Baccelli, J. Burke, R. Ravindran, G. Wang, A. Lindgren, B. Ahlgren, and O. Schelen. Requirements and Challenges for IoT over ICN. *https://tools.ietf.org/html/draft-zhang-icnrg-icniot-requirements-00*, 2015. IETF Internet-Draft.

717. Y. Zhang, D. Raychadhuri, R. Ravindran, and G. Wang. ICN based Architecture for IoT. *https://tools.ietf.org/html/draft-zhang-iot-icn-challenges-02*, 2013. IRTF contribution.

718. Y. Zhang, S. Shi, S. Guo, X. Chen, and Z. Piao. Audience management, online turbulence and lurking in social networking services: A transactional process of stress perspective. *International Journal of Information Management*, 56:102233, 2021. Elsevier.

719. Z. Zhang, Q. Li, D. Zeng, and H. Gao. User community discovery from multi-relational networks. *Decision Support Systems*, 54(2):870–879, 2013. Elsevier.

720. F. Zhao, Z. Sun, and H. Jin. Topic-centric and semantic-aware retrieval system for internet of things. *Information Fusion*, 23:33–42, 2015. Elsevier.

721. Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. Herrera Viedma. An incremental method to detect communities in dynamic evolving social networks. *Knowledge-Based Systems*, 163:404–415, 2019. Elsevier.

722. D. Zhelonkin and N. Karpov. Training Effective Model for Real-Time Detection of NSFW Photos and Drawings. In *Proc. of the International Conference on Analysis of Images, Social Networks and Texts (AIST 2019)*, pages 301–312, Kazan, Russia, 2019. Springer.

723. B. Zheng, O. Liu, J. Li, Y. Lin, C. Chang, B. Li, T. Chen, and H. Peng. Towards a distributed local-search approach for partitioning large-scale social networks. *Information Sciences*, 508:200–213, 2020. Elsevier.

724. Z. Zheng, S. Xie, H.N. Dai, X. Chen, and H. Wang. Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, 14(4):352–375, 2018. Inderscience.

725. M. Zhou, X. Cai, Q. Liu, and W. Fan. Examining continuance use on social network and micro-blogging sites: Different roles of self-image and peer influence. *International Journal of Information Management*, 47:215–232, 2019. Elsevier.

726. Z. Zhou, C. Gao, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez. Social Big-Data-Based Content Dissemination in Internet of Vehicles. *IEEE Transactions on Industrial Informatics*, 14(2):768–777, 2018.

727. Z. Zhu, J. Su, and L. Kong. Measuring influence in online social network based on the user-content bipartite graph. *Computers in Human Behavior*, 52:184–189, 2015.

728. D. Zissis and D. Lekkas. Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3):583–592, 2012. Elsevier.

729. F. Zola, M. Eguimendia, J.L. Bruse, and R.O. Urrutia. Cascading Machine Learning to Attack Bitcoin Anonymity. In *Proc. of the International Conference on Blockchain (ICBC'19)*, pages 10–17, Atlanta, GA, USA, 2019. IEEE.