Defining and detecting k-bridges in a social network: the Yelp case, and more

https://doi.org/10.1016/j.knosys.2020.105721

Defining and detecting k-bridges in a social network: the Yelp case, and more

Enrico Corradini[1], Antonino Nocera[2], Domenico Ursino[1*], and Luca Virgili[1]

[1] DII, Polytechnic University of Marche,
[2] DIII, University of Pavia
[*] Contact Author
e.corradini@pm.univpm.it; antonino.nocera@unipv.it; d.ursino@univpm.it;
l.virgili@pm.univpm.it

Abstract

In this paper, 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 we extract several knowledge patterns about Yelp k-bridges. In particular, we investigate how some basic characteristics of Yelp k-bridges vary against k (i.e., against the number of macro-categories which the businesses reviewed by them belong to). Then, we verify if there exists an influence exerted by k-bridges on their friends and/or on their co-reviewers. In addition, we analyze the relationship between k-bridges and power users. In addition, we investigate the relationship between k-bridges and the main centrality measures in the macro-categories of Yelp. We also propose two further specializations of k-bridges, regarding Reddit and the network of patent inventors, to prove that the knowledge on k-bridges we initially found in Yelp is not limited to this social network. Finally, we present two use cases that can highly benefit from the knowledge on k-bridges detected through our approach.

Keywords: k-bridge; k-bridge detection algorithm; multi-network scenario; influencers; analysis of co-reviewers; Yelp; Reddit; PATSTAT-ICRIOS

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 [24, 49, 35, 8, 9, 60] or linking different networks in a multi-network context [12, 14, 13, 44].

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 paper, 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 [15, 56, 43, 37, 59].

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 [17]. In this paper, 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 [18], 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 in this paper, 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.

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

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

The outline of this paper is as follows: in Section 2, we present related literature. In Section 3, we formalize the concept of k-bridge, present an algorithm for the detection of k-bridges from a social network, and illustrate the specialization of the concept of k-bridge in Yelp, Reddit and the network of patent inventors. In Section 4, we investigate the properties of k-bridges. In Section 5, we provide a deeper analysis of k-bridges as connectors among different communities. The activities described in Sections 4 and 5, and the results obtained, allowed us to consider Yelp as the baseline for further experiments in other social networks. In Section 6, we perform some of the activities in Reddit and the network of patent inventors to verify if what we had found for Yelp was general of specific for this platform. In Section 7, we present two use cases that could benefit from k-bridges and their properties. Finally, in Section 8, we draw our conclusions.

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 [29]. A lot of research has been devoted to understand how users interact in social media and how information diffusion takes place inside them [5, 58, 61, 10].

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 [48, 11, 1].

These studies have proved that there exist different categories of users, each participating to the platform with different levels of activity and heterogeneous contents [7, 38].

Of course, when dealing with user interactions, it is important to consider those that cannot be examined homogeneously [16]. 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 [53, 20, 62, 65]. For instance, in [20], 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 [62] 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 [66]. 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 [53] 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 [53], different studies have been devoted to analyze the role of users in the creation of social communities. In particular, the authors of [30] 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 [25], 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 [24, 49, 35, 3].

Studying nodes bridging communities together has been also a crucial research direction in the context of multi-relation networks [8, 9]. Here, the heterogeneity of the scenario is more evident because of the different nature of the relation considered. In particular, the authors of [8] 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 [9], 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 [12, 14, 13]. Here the concept of community is extended in such a way that a community is mapped to a whole social network. Specifically, in [12], 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 [14]. 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 [13] 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 are 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 in this paper 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 investigations that analyze the dynamics describing how rates are assigned to businesses in Yelp [15, 28, 34, 51, 19, 50]; *(ii) Review Analysis*: It comprises the works focused on the analysis of reviews and of what events drive the users writing them [56, 52, 45, 46, 6, 27]; *(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 [43, 47, 4, 26]; *(iv) Fake review and rate discovery*: It includes the proposals dealing with the detection of fake reviews and rates [37, 42, 39, 33]; *(v) Recommender Systems*: It comprises all the research works devoted to provide Yelp users with recommendations about suitable businesses, other users to interact with, and even text suggestions for new reviews [59, 32, 22, 17, 57].

Despite our work shares some similarities with several other ones described in this section, to the best of our knowledge, this paper represents 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 paper 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 [17] 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 section.

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

In this section, we propose a general model for k-bridges, we specialize it to several social networks and, then, we present an algorithm to extract k-bridges. Specifically: In Section 3.1, we formally define the concept of k-bridges and some other ones related to it; then, we propose a model for representing k-bridges. In Section 3.2, we present an algorithm that, exploiting the anti-monotone property characterizing k-bridges, extracts them from a social network. In Section 3.3, we specialize our concept and model of k-bridges to Yelp that is the main reference social network for this paper. Finally, in Sections 3.4 and 3.5, we propose a further specialization of our concept and model of k-bridge to Reddit and the network of patent inventors.

3.1 Defining and modeling k-bridges

In this section, we present our definition of k-bridges. It can be applied to any social network whose users are organized into partially overlapping communities. In particular, 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.

3.2 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 [2]. 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 3.1 (*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 [36, 64].

3.3 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.

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

Algorithm 1 K-bridges Extraction Algorithm

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

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.

3.4 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.

3.5 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 [23, 21]. 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").
- 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.

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 1: The main notations used throughout this paper

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 reported throughout the remaining part of this paper, in Table 1, we report the main notations introduced.

4 Investigating k-bridge properties

In this section, we analyze on k-bridge properties. We carried out this task focusing on Yelp, which is the main reference network of this paper. However, in the next section, we present some experiments on Reddit and the network of patent inventors devoted to verify if the results on k-bridges found in Yelp are general or specific for this social network. This section is organized as follows: In Subsection 4.1, we illustrate some preliminary investigations performed to better understand the characteristics of Yelp data. In Subsection 4.2, we detect some properties of k-bridges starting from the friendship and co-review networks associated with Yelp. Finally, in Subsection 4.3, we analyze the possible correlations between k-bridges and power users.

4.1 Preliminary investigation of Yelp Data

The data required for the investigation activities described in this paper 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 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.

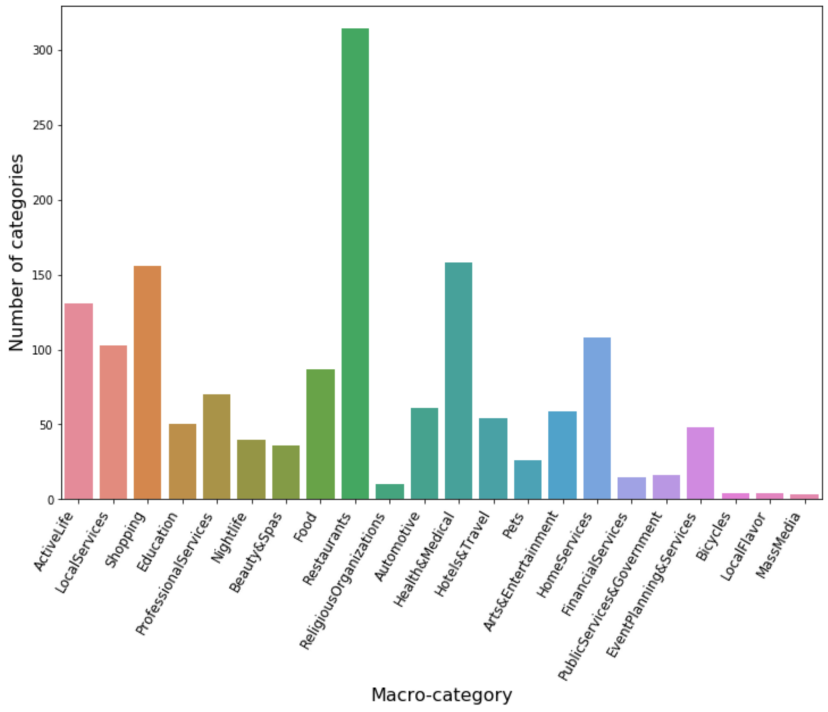

Figure 1: Distribution of categories inside the macro-categories of Yelp

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 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 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 [41]. 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 number of individuals (typically 10-20% of members) perform the majority of the activities (around 80-90% of

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

the overall activities) [63]. Our experiment confirms that this trend also persists in the review tasks in Yelp.

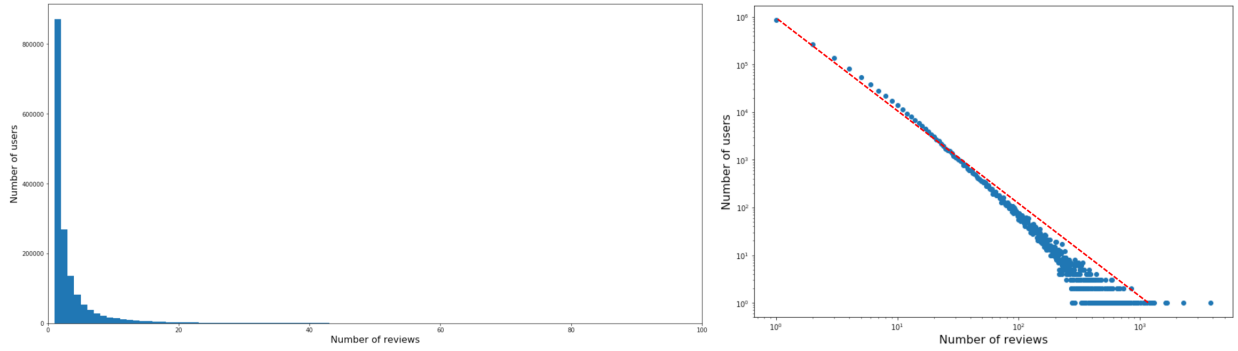

Figure 2: Distribution of user reviews in Yelp - Linear scale (on the left) and Logarithmic scale (on the right)

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. 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 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 confirm 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.

4.2 K-bridges at a first glance

After the preliminary analysis on Yelp data, described in the previous section, we started our investigation on k-bridges.

Firstly, we verified the possible existence of a backbone among the *bridges*. In other words, we

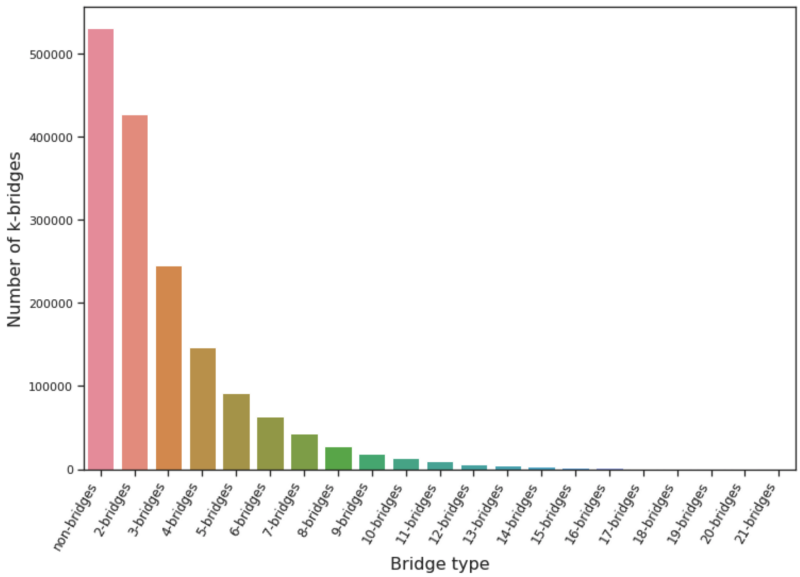

Figure 3: Distribution of the k-bridges against k in Yelp

wanted to verify if a bridge tends to relate more to other bridges or not. This property can be seen as the application of the homophily principle [40] to our context. Furthermore, we wanted to understand if there exists an influence exerted by bridges on their friends and/or on their co-reviewers.

We conducted this study on both the friendship network $\mathcal{U}^f$ and on the co-review network $\mathcal{U}^{cr}$. In the next two subsections we describe it in more detail.

4.2.1 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. 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: 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 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 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). 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 [63], 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 5. Differently from what emerges from Table 3, where there is a little difference between the *fraction of bridges* in the neighborhoods of bridges and non-bridges, in Table 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 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 4 and 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.

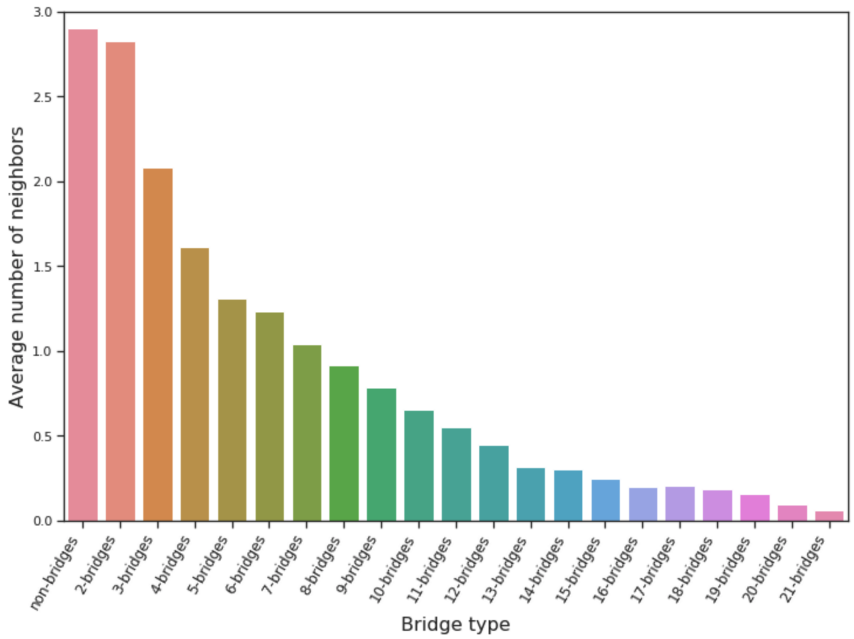

Figure 4: Distribution of the neighbors of *bridges* in $\mathcal{U}^f$

4.2.2 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 in this network. Preliminarily, we removed those nodes corresponding to users who reviewed businesses not belonging to any macro-category of Yelp (see Section 4.1). 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 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.

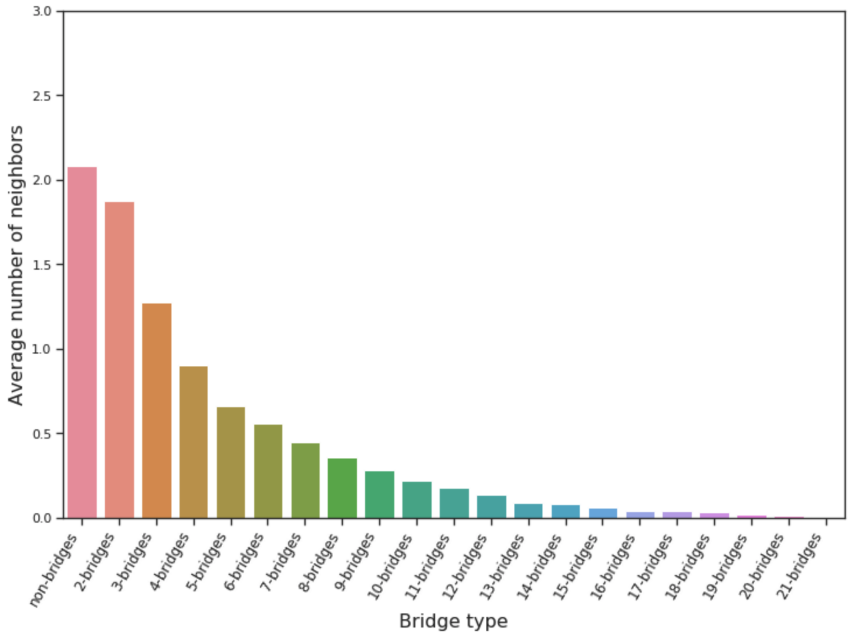

Figure 5: Distribution of the neighbors of *non-bridges* in $\mathcal{U}^f$

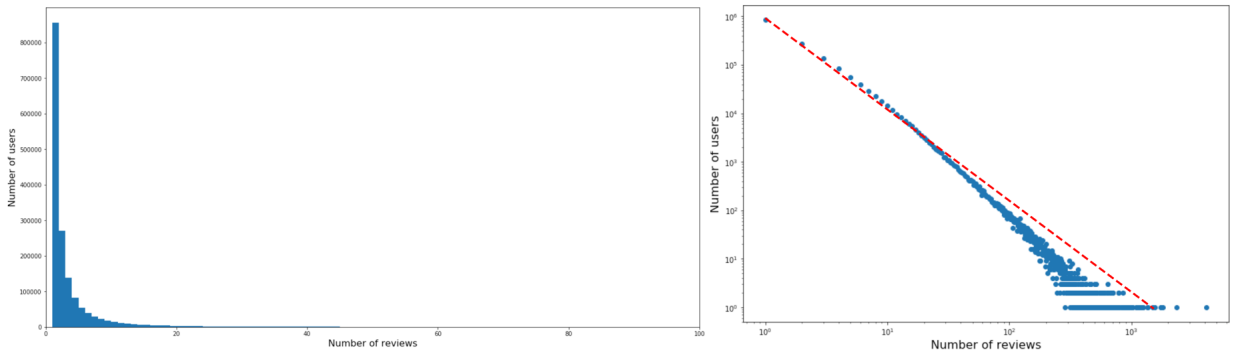

Figure 6: Distribution of reviews for users in $\mathcal{U}^{cr}$ - Linear scale (on the left) and Logarithmic scale (on the right)

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 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 $\mathcal{U}^{cr}$.

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 7 and 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

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

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.

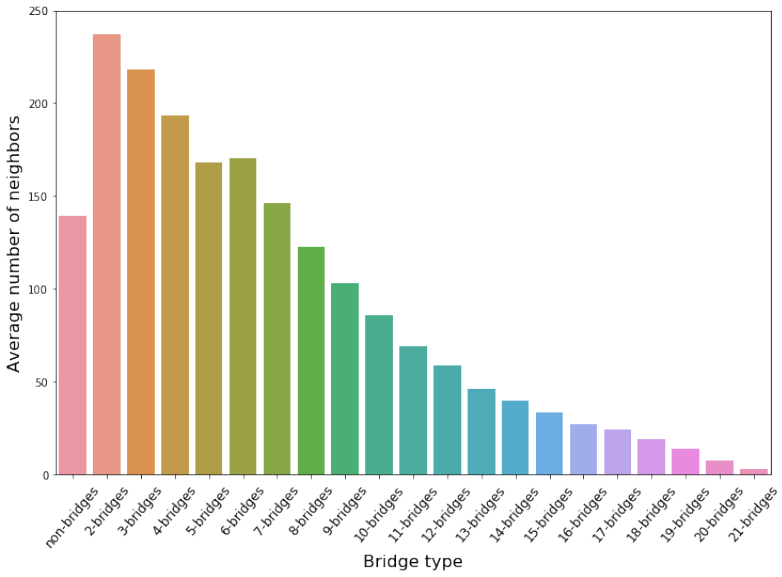

Figure 7: Distribution of the neighbors of *bridges* in $\mathcal{U}^{cr}$

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 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 identical 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 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:

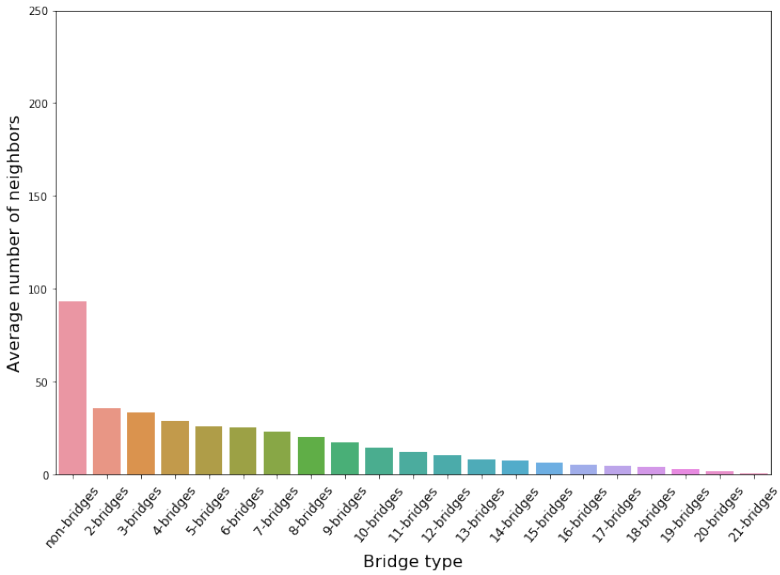

Figure 8: Distribution of the neighbors of *non-bridges* in $\mathcal{U}^{cr}$

- 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 $\mathcal{U}^{cr}$.

4.3 Analysis of the possible correlation between k-bridges and power users in $\mathcal{U}^{cr}$

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 9. As we can see from this figure, all distributions follow power laws; their corresponding coefficients α and δ are reported in Table 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 centrality 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.

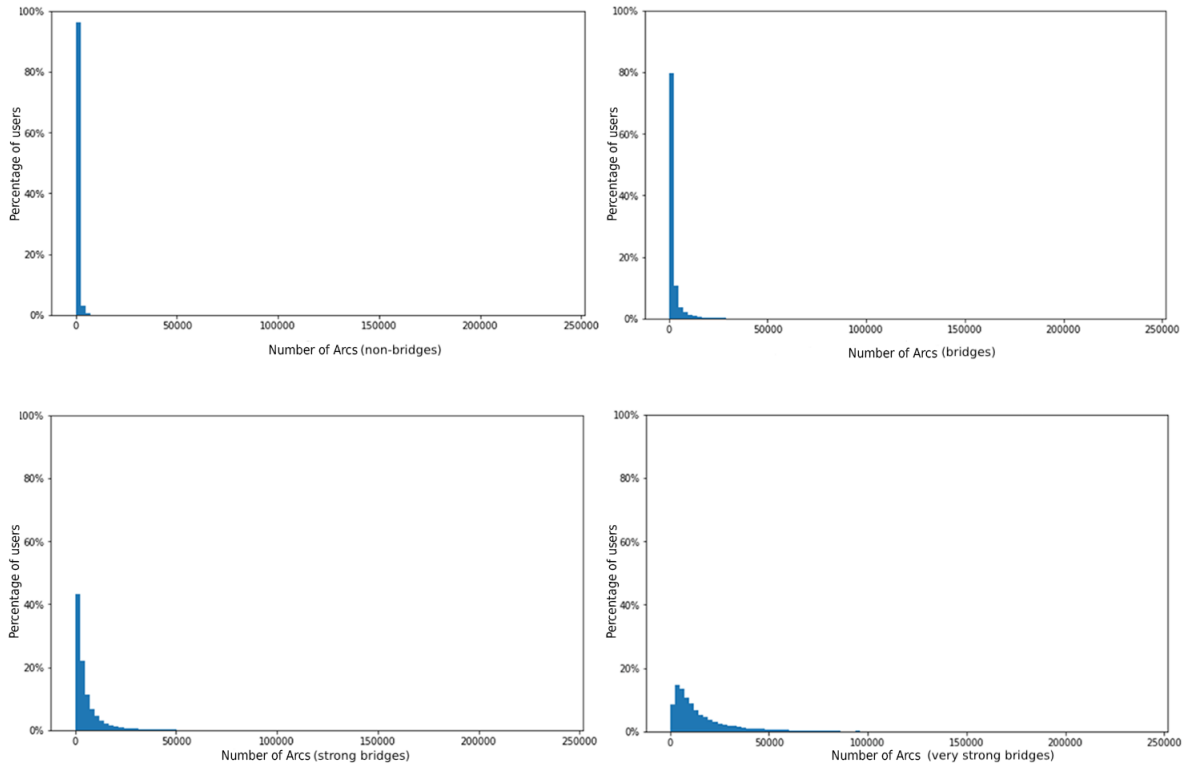

Figure 9: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges

	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 8: Coefficients α and δ for the power law distributions of Figure 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 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 distribution of their degree. We

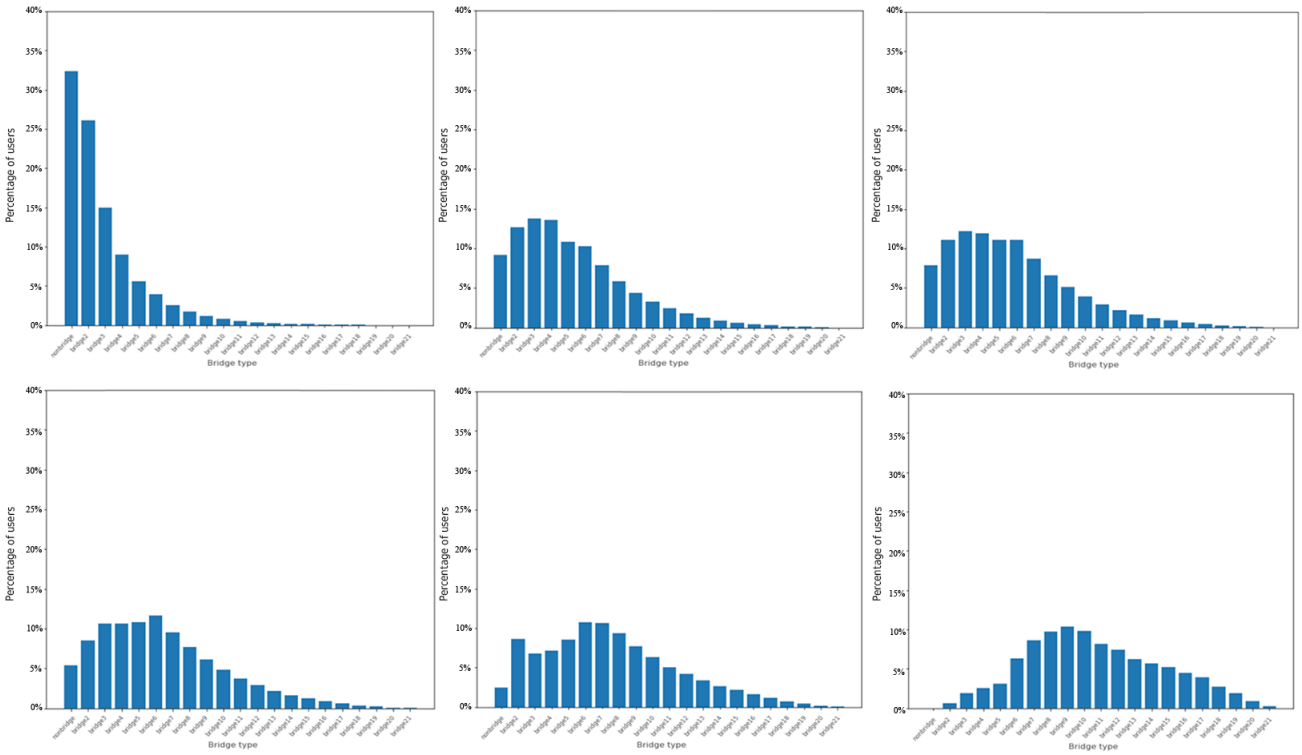

Figure 10: Distributions of (power) users against the strength of bridges

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 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.

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

5 K-bridges and macro-categories: a deeper analysis

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 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.

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

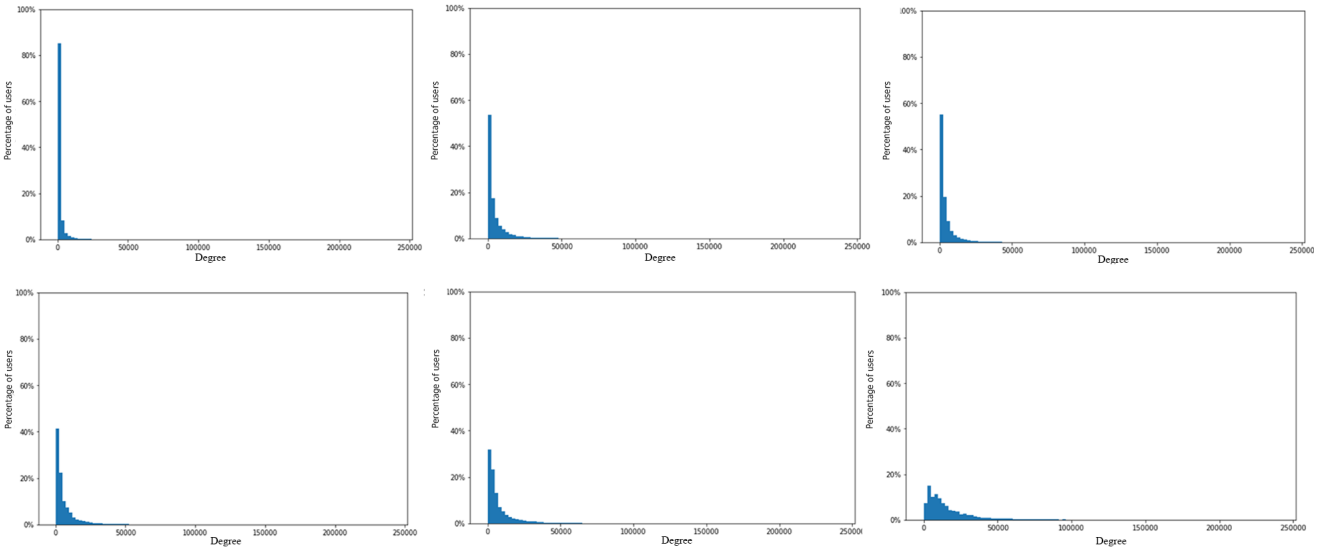

Figure 11: Distributions of k-bridges against their degree

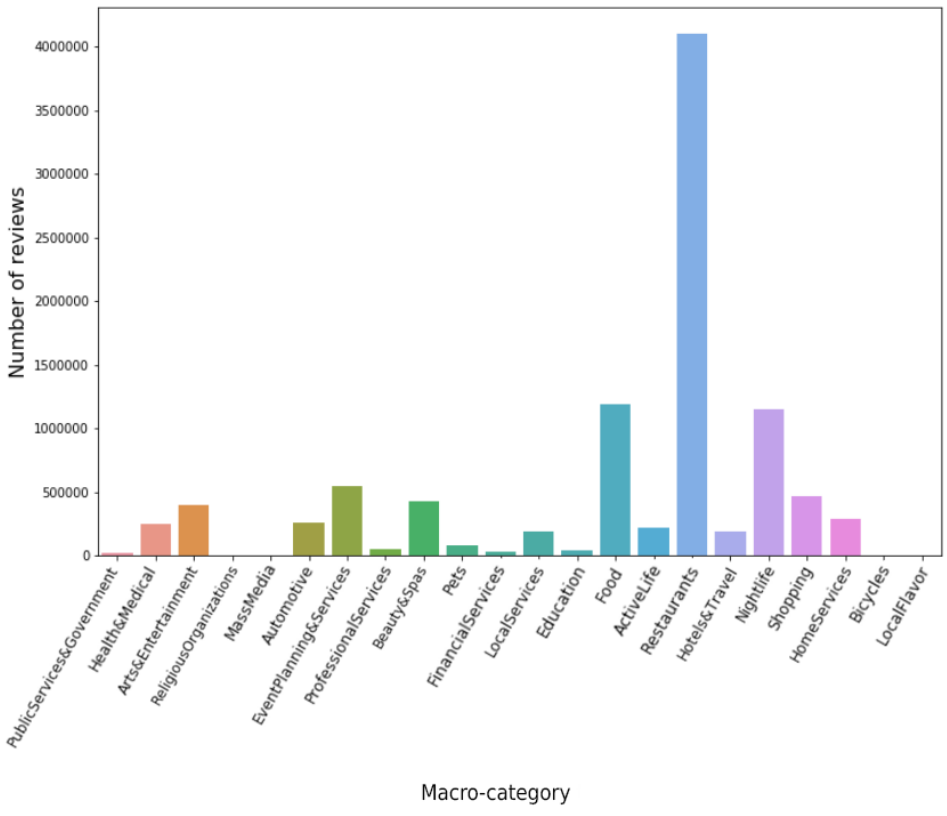

Figure 12: Distribution of the reviews of Yelp users against the Yelp macro-categories

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 13 - 16.

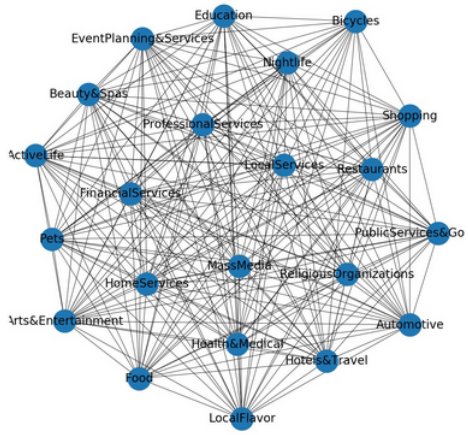

Figure 13: The network $\mathcal{M}^{1\%}$

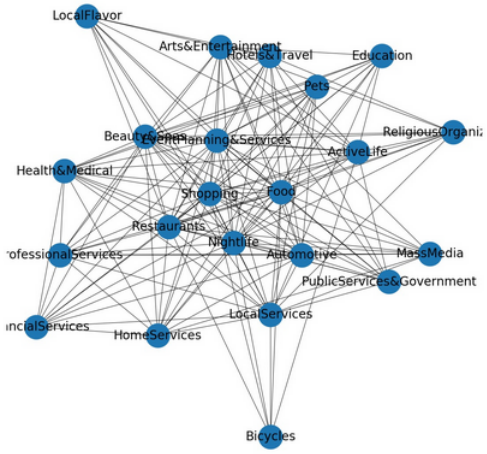

Figure 14: The network $\mathcal{M}^{5\%}$

The corresponding density and average clustering coefficient are reported in Table 9. Figures 17 and 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

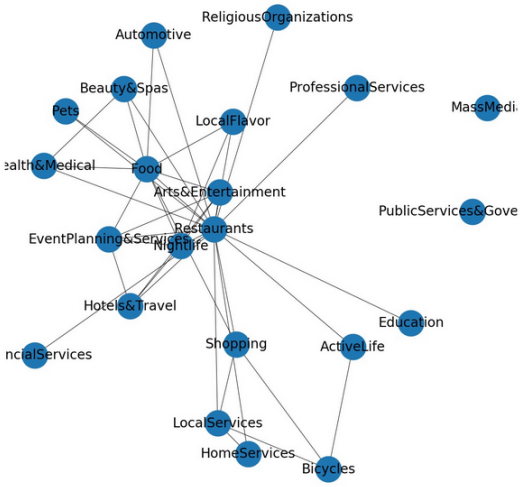

Figure 15: The network $\mathcal{M}^{10\%}$

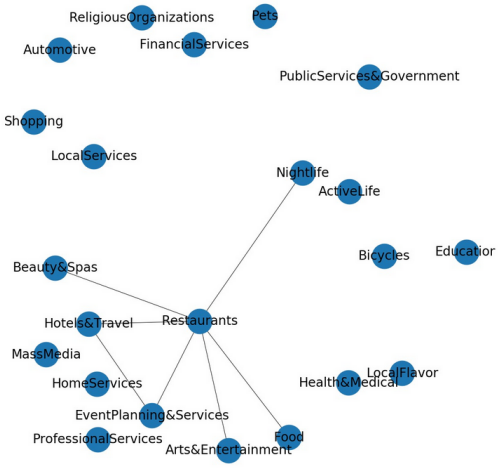

Figure 16: The network $\mathcal{M}^{15\%}$

by at least 1% of the users. However, if we are more demanding on the fraction of users 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.

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

Table 9: Values of the density and the average clustering coefficient for the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

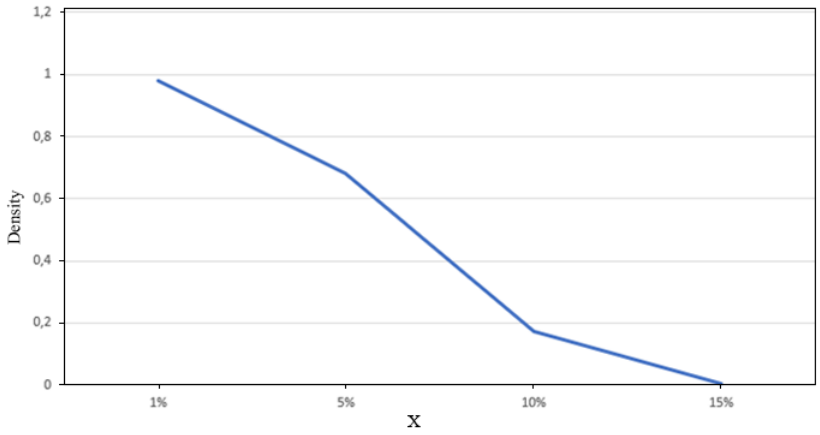

Figure 17: Variation of the density of the macro-category networks $\mathcal{M}^{X\%}$ against the increase of X

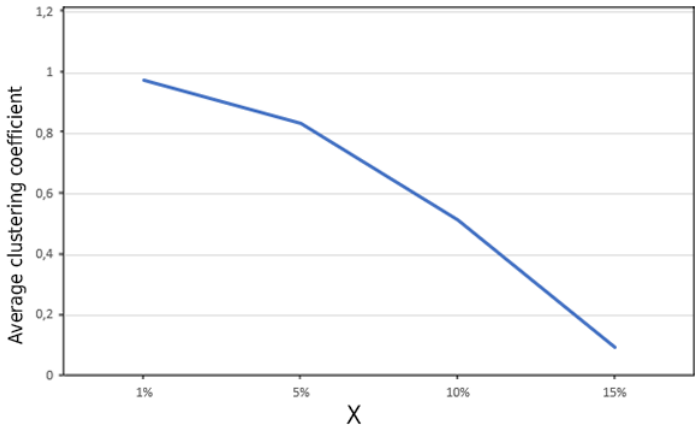

Figure 18: Variation of the average clustering coefficient of the macro-category networks $\mathcal{M}^{X\%}$ against the increase of X

Table 10 shows the maximum and sub-maximum values of the degree centrality for the networks of Figures 13 - 16, along with the macro-categories 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 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 10.

Table 11 shows the maximum and sub-maximum values of the closeness centrality for the networks of Figures 13 - 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

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
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 10: Maximum and sub-maximum values of degree centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

centrality are generally different [55], 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 11: Maximum and sub-maximum values of closeness centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

Table 12 shows the maximum and sub-maximum values of the betweenness centrality for the networks of Figures 13 - 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 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 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 12: Maximum and sub-maximum values of betweenness centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

Table 13 shows the maximum and sub-maximum values of the eigenvector centrality for the networks of Figures 13 - 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".

The analysis of the distributions and the ones of all the different forms of centrality show that

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
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 13: Maximum and sub-maximum values of eigenvector centrality and the corresponding macro-categories in the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$

"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 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.

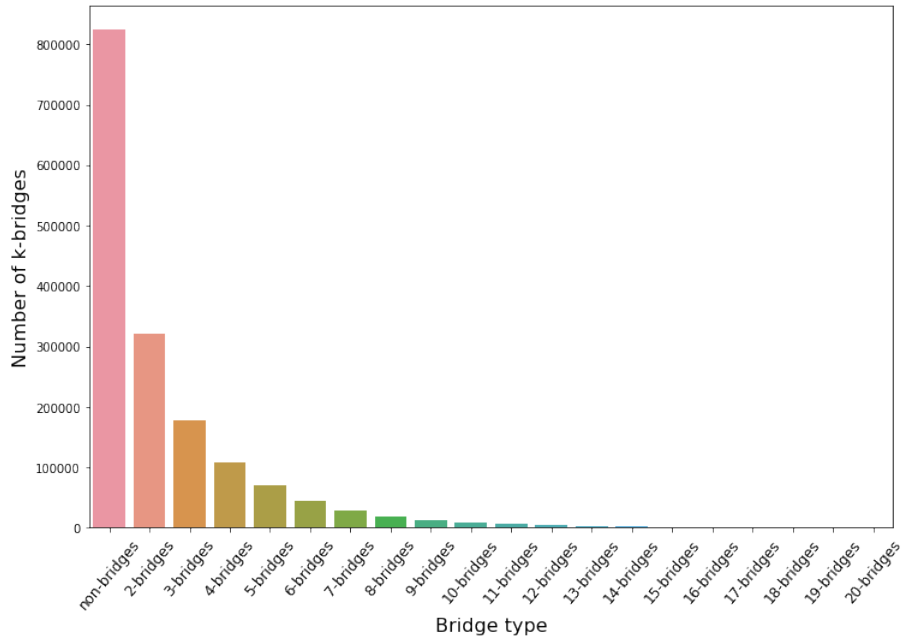

Figure 19: Distribution of the k-bridges against k in Yelp after the removal of "Restaurants"

Then, we computed again the networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$. They are shown in Figure 20. From the analysis of this figure, we can observe that the connection level of these networks slightly decrease 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 14.

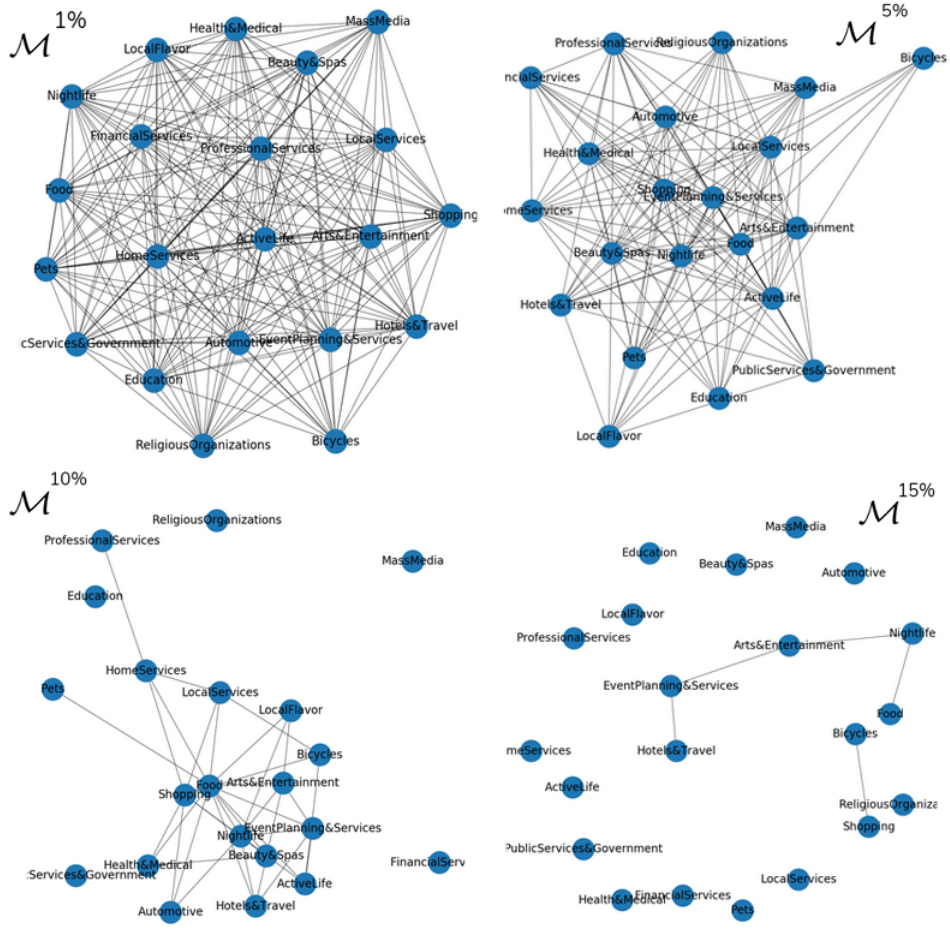

Figure 20: The networks $\mathcal{M}^{1\%}$ - $\mathcal{M}^{15\%}$ after the removal of "Restaurants"

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

Table 14: Values of the density and the average clustering coefficient for 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 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.

	$\mathcal{M}^{1\%}$	$\mathcal{M}^{5\%}$	$\mathcal{M}^{10\%}$	$\mathcal{M}^{15\%}$
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 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"

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.

6 Verifying k-bridge properties in Reddit and in the network of patent inventors

This section aims at verifying if k-bridge properties we had found in Yelp are general or specific to this social network. Due to space constraints, we limit our analysis to only some of the properties found above. We verify their validity first in Reddit (Subsection 6.1) and, then, in the network of patent inventors (Subsection 6.2).

6.1 Verifying k-bridge properties in Reddit

In this section, we show the results that we obtained by performing on Reddit some of the experiments previously carried out on Yelp. 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 3.4, 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 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.

As a second experiment, we considered the co-posting network $\mathcal{U}^{cp}$, defined in Section 3.4. 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

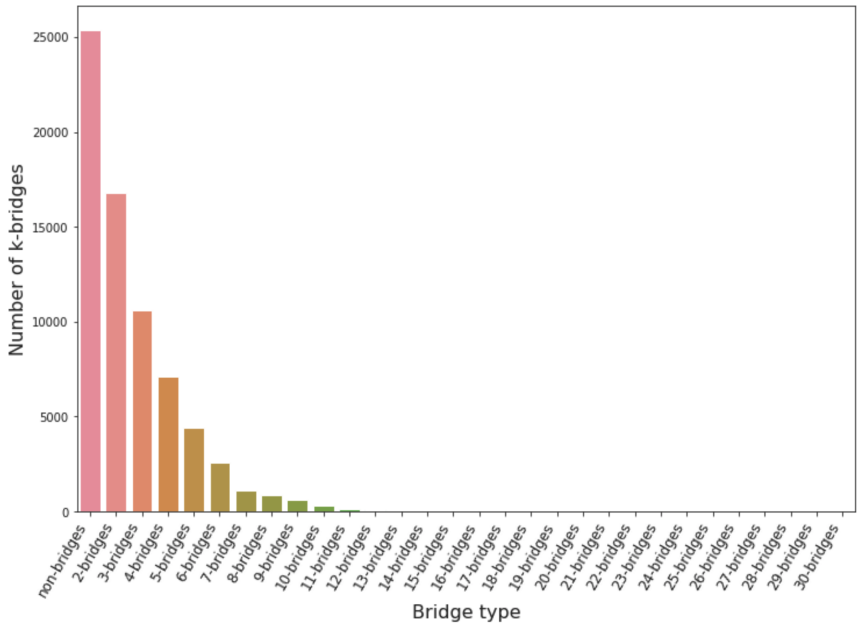

Figure 21: Distribution of the k-bridges against k in Reddit

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 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 6.

	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 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 4.2.1 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 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 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.

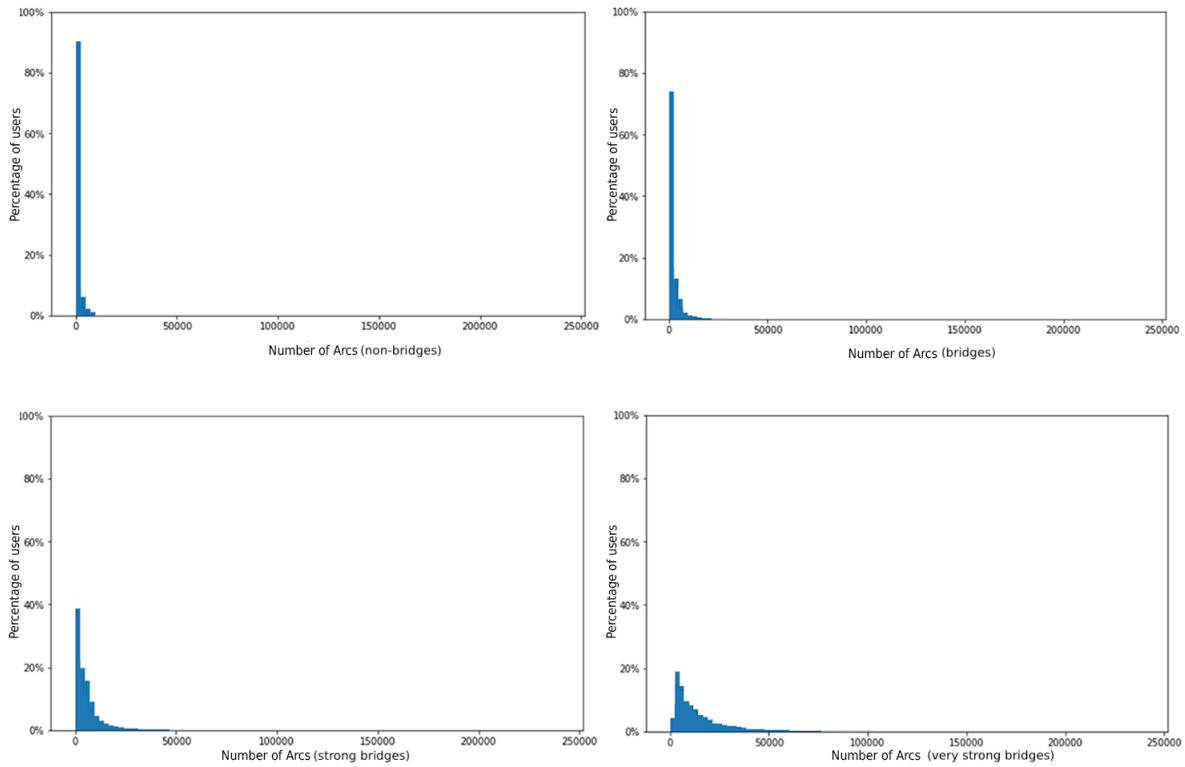

Figure 22: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges in Reddit

6.2 Verifying k-bridge properties in the network of patent inventors

In this section, we report the results that we obtained by performing on the network of patent inventors some of the experiments previously carried out on Yelp. 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.

According to our model, as described in Section 3.5, 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 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 3.5. 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.

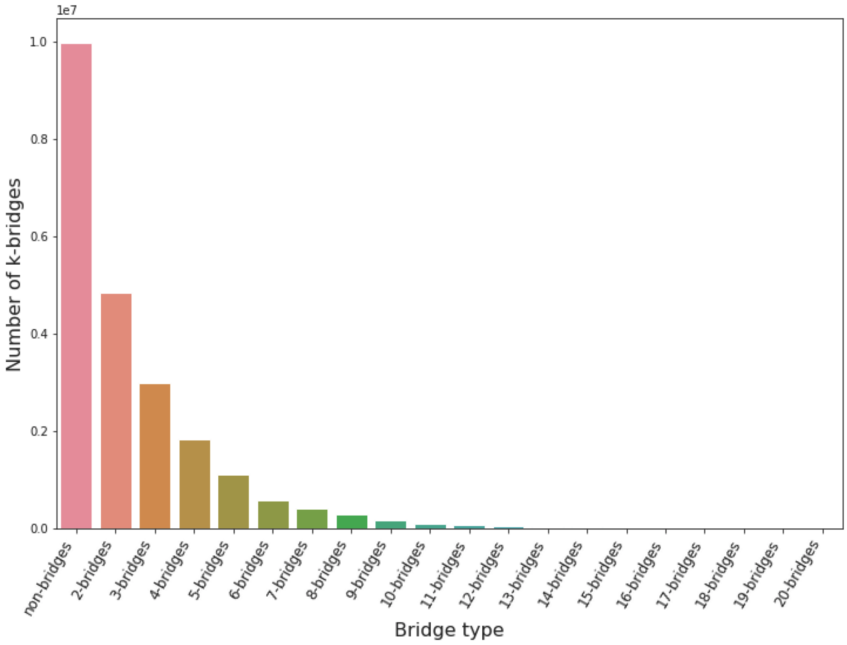

Figure 23: Distribution of the k-bridges against k in the network of patent inventors

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 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.

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

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 4.2.1 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.

We computed the distribution of the number of arcs for non-bridges, bridges, strong and very strong bridges. The results are reported in Figure 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 9 for Yelp and in Figure 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

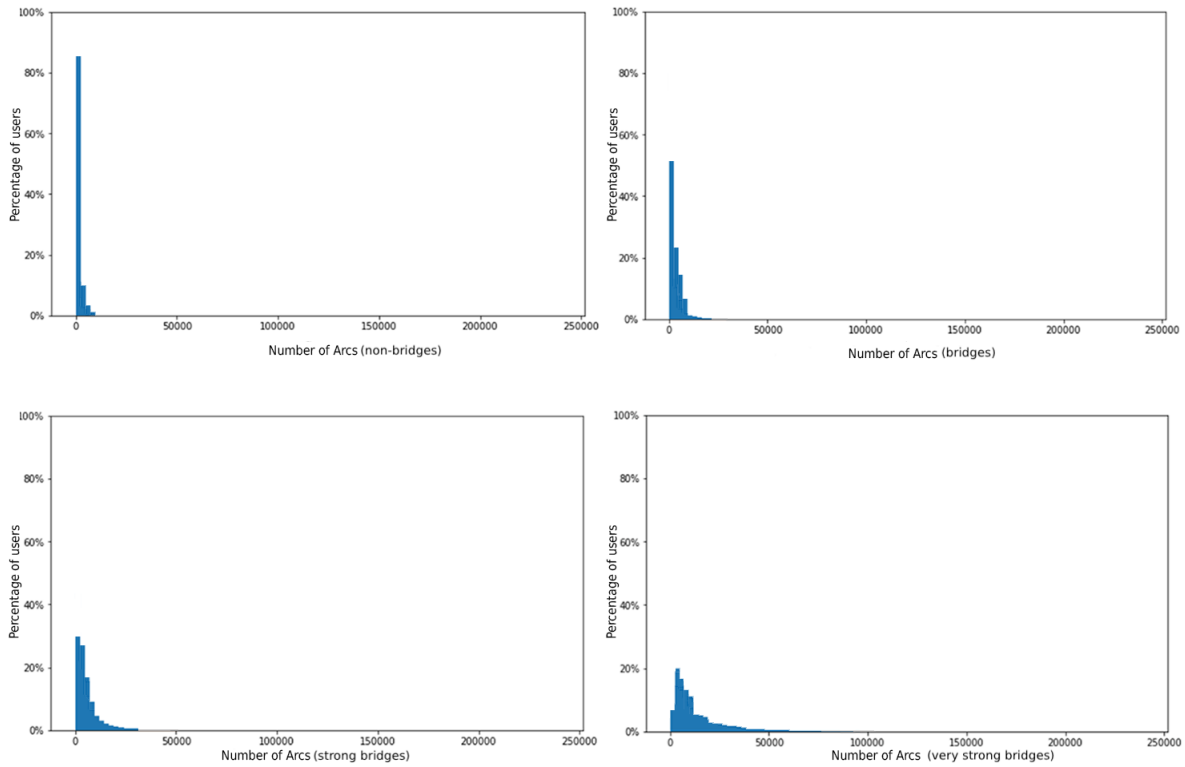

Figure 24: Distributions of the number of arcs for non-bridges, bridges, strong and very strong bridges in the network of patent inventors

knowledge of k-bridges.

7 Two possible k-bridge applications

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.

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 [54]. 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

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

ad-hoc marketing and advertising campaigns is extremely important. In the next subsections, we describe in details two case studies related to this concept, which massively exploit k-bridges to conduct marketing campaigns and support business decisions in Yelp.

7.1 Case A: 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 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 1, we report a possible simple implementation of such a metric (clearly, future research efforts could be made to define a more sophisticated metric):

$$\mu_i = \frac{nr_i \cdot nf_i}{nd_i} \tag{1}$$

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.

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

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 $\overline{B_k}$

7.2 Case B: 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 [31], 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.

8 Conclusion

In this paper, we have introduced the concept of k-bridge and we have found that it enjoys the anti-monotone property. Starting from this result, we have proposed an algorithm for detecting k-bridges from a social network. With Yelp as the main reference platform, we have discovered several features characterizing k-bridges and we have detected several knowledge patterns about them. Afterwards, by performing on Reddit and the network of patent inventors some of the experiments we had already carried out on Yelp, we have seen that the properties and the knowledge patterns characterizing k-bridges, that we have found through Yelp, are general and not limited to this social network. Finally, we have presented two use cases that could benefit from the presence of k-bridges; the former regards the application of k-bridges to find the best targets for a marketing campaign. The latter concerns the role of k-bridges to find new products/services to propose.

[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.

In the future, we plan to extend this research in several directions. First of all, we would like to investigate other properties of k-bridges, for instance their capability of being influencers in a social context. Negative influencers are particularly interesting for us because this user stereotype is less studied in the literature even if its impact in real life is enormous. Then, we would like to extend the analysis of k-bridges from Yelp to other platforms similar to it, for instance TripAdvisor, to understand the analogies and the difference with Yelp. Afterwards, we would like to extend, realize and test the approaches described in the two use cases described in this paper. Finally, we plan to realize a research campaign that performs a profile-based analysis of users for most of the challenges described in this paper in order to extract a deep knowledge about k-bridges and their behaviors in the social platforms they belong to.

Acknowledgments

This work was partially funded by the Department of Information Engineering at the Polytechnic University of Marche under the project "A network-based approach to uniformly extract knowledge and support decision making in heterogeneous application contexts" (RSAB 2018).

References

[1] L. Adamic and E. Adar. Friends and Neighbors on the Web. *Social Networks*, 25(3):211–230, 2003. Elsevier.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the International Conference on Very Large Data Bases (VLDB'94)*, volume 1215, pages 487–499, Santiago, Chile, 1994.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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. IGI Global, 2019.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] D.W. Franks, J. Noble, P. Kaufmann, and S. Stagl. Extremism propagation in social networks with hubs. *Adaptive Behavior*, 16(4):264–274, 2008.

[25] M.S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. JSTOR.

[26] 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*, 2020. Forthcoming. Elsevier.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] J. Leskovec, L.A. Adamic, and B.A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, 2007. ACM.

[36] 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, Artificial Intelligence, Networking and Parallel/Distributed Computing (ACIS'13)*, pages 236–241, 2012. IEEE.

[37] M. Luca and G. Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016.

[38] 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.

[39] J. Malbon. Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36(2):139–157, 2013. Springer.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. Sebastopol, CA, USA, 2011. O'Reilly Media, Inc.

[56] T. Tucker. Online word of mouth: characteristics of Yelp.com reviews. *Elon Journal of Undergraduate Research in Communications*, 2(1):37–42, 2011.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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*, 2020. Forthcoming.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.