Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity

Investigating Reddit to detect subreddit and author stereotypes and to evaluate author assortativity

Abstract

In recent years, Reddit has attracted the interest of many researchers due to its popularity all over the world. In this paper, we aim at providing a contribution in the knowledge of this social network by investigating three of its aspects, interesting from the scientific viewpoint, and, at the same time, by analyzing a large number of applications. In particular, we first propose a definition and an analysis of several stereotypes of both subreddits and authors. This analysis is coupled with the definition of three possible orthogonal taxonomies that help us to classify stereotypes in an appropriate way. Then, we investigate the possible existence of author assortativity in this social medium; specifically, we focus on co-posters, i.e. authors who submitted posts on the same subreddit.

Keywords: Reddit; Author Stereotypes; Community Stereotypes; Assortativity; Social Network Analysis; Subreddit Lifecycle

1 Introduction

Reddit[1] is a heterogeneous crowd-sourced news aggregator and online social platform, originally self-declared as "the front page of Internet". It was founded in 2005 and, in few years, has become an ecosystem of 430M+ average monthly active users[2]. At the time of writing, it ranks 19^{th} and 5^{th} in the Alexa's top 500 global and US websites, respectively[3]. Reddit is built on the concept of *subreddit*, which is an interest-based community where users can post and comment contents. A subreddit is identified by a name, and is referred to using the /r/ prefix within Reddit, such as /r/science and /r/cats. Currently, there are more than 1.9M subreddits[4]. They are mainly topical, although more general cases exist.

In Reddit, users can submit contents in the form of texts, images and links to external resources. Submitted contents (also simply called posts) can be read by other users and discussed via comments. Users can subscribe to multiple subreddits in order to receive the latest posted contents on their front pages. An important feature of Reddit is *voting*, which represents the mechanism affecting the visibility and the ranking of both posts and comments. In fact, users are allowed to *upvote* or *downvote*

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

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

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

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

posts of other users, so that each submission has a *score*. This is a metric based on the difference between the number of upvotes and the number of downvotes, and it significantly affects the order through which posts and comments are shown to users. However, the exact numbers of upvotes and downvotes are not shown publicly.

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

In this paper, we aim at providing a contribution in the knowledge of Reddit by investigating subreddit and author stereotypes and by evaluating author assortativity in this social platform. For this purpose, we built a dataset with all the posts published from January 1^{st}, 2019 to September 1^{st}, 2019, which we used for our analyses. We started with some preliminary investigations on Reddit data. They focused on three aspects, namely posts submitted to subreddits, comments under these posts and, finally, users who created a subreddit, posted or commented. The aim of this preliminary descriptive analysis was not to discover new specific knowledge about Reddit. Instead, it allowed us to better understand the dataset, and to check if some theoretical trends, which should have characterized these aspects on Reddit, were verified on it. Furthermore, the results found, which were partially expected, represented the starting point of the next knowledge detection activities, which are the core of our paper. They were also useful to explain the knowledge patterns extracted.

After this preliminary analysis, we discuss our investigation on how to stereotype subreddits. For this purpose, we first investigated the lifecycle of a subreddit, depicting its typical characteristics. Then, starting from this, we identified several subreddit stereotypes and, finally, we defined and applied three orthogonal taxonomies in order to characterize them. After the analysis of subreddit stereotypes, we proceeded similarly for Reddit authors. In particular, we extracted several author stereotypes and, then, we classified them according to some orthogonal taxonomies that we defined for this purpose.

The last part of this paper is devoted to verify the possible existence of a degree assortativity in Reddit. We recall that assortativity in a social network expresses the inclination of a node to associate with other nodes that are somewhat similar. Assortativity has been largely investigated by social media analysts [**?**, **?**]. We aimed at performing this analysis for Reddit authors and degree assortativity to verify if authors very active in Reddit tend to form a backbone or not.

The findings on stereotypes and degree assortativity explained in this paper have several applications. Just to cite a few of them, we mention: *(i)* the definition of some guidelines to follow in order to make a subreddit successful; *(ii)* the definition and realization of different categories of recommender systems for Reddit; *(iii)* the definition of an algorithm that finds subreddits to merge or, at least, to integrate; *(iv)* the detection of possible targets for an advertising campaign; *(v)* the definition of an algorithm that builds blacklists of users based on author stereotypes.

The outline of this paper is as follows. In Section 2, we describe related literature. In Section

3, we illustrate the dataset that we used for our investigations. In Section 4, we present several preliminary analyses concerning posts, comments and users in Reddit. In Section 5, we illustrate the activities performed to detect subreddit stereotypes and to determine their features. In Section 6, we describe the same tasks done to detect author stereotypes. In Section 7, we analyze author assortativity in Reddit. In Section 8, we describe some possible applications of the knowledge we extracted in the previous sections. Finally, in Section 9, we draw our conclusions and have a look at future developments concerning our research.

2 Related work

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

The swift increase of scientific literature related to Reddit produced a discrete number of papers with several goals and methodologies. An overall survey is introduced in [**?**], discussing various studies spanning in time from 2005 to 2018. An interesting longitudinal analysis on the evolution of Reddit is presented in [**?**].

Due to the heterogeneity of Reddit data, different structures and points of view can be adopted to analyze and study phenomena concerning this social medium. In order to understand all of the literature revolving around the Reddit ecosystem, we first look in some detail at works considering the underlying network structure. Then, we provide a brief bird's eye view considering all the other ones.

An interesting and in-depth analyzed aspect is the "multi-community interaction". In [**?**], the authors examine multi-community engagement using longitudinal posting behavior on Reddit and DBLP. They find out that users continually post in new communities, while those who eventually leave a community are intended to do so from the very early beginning of their history. A study regarding inter-community aspects in Reddit is presented in [**?**]. Here, the authors focus on anti-social behaviors in the form of inter-community conflicts, studying subreddits where a user shows social or anti-social behaviors. The studies of [**?**] and [**?**] focus on specific behavioral aspects of authors, namely multi-community engagement and anti-social behaviors.

Another work regarding community interactions and conflicts is presented in [**?**], where the authors study inter-community interactions across 36,000 communities. In particular, they examine cases where users of one community, driven by a negative sentiment, comment in another community. They highlight how such conflicts emerge from a very small number of communities. They also discuss strategies for predicting conflicts and mitigating their negative impacts.

In [**?**], the authors focus on studying loyal communities, and find that they tend to be less assortative as long as their interaction level increases. Assortativity is studied on monthly interaction networks, where users are connected if they submit a comment in the same comment chain with a

gap of at most two comments. The authors also carry out a comparison with a null model and find that the difference between loyal communities and their random counterparts disappears. This result implies that users in loyal communities tend to interact with dissimilar users as a consequence of the community's activity.

User posting behavior is explored in [**?**], where the authors show how the "answer-person" role is present in Reddit, and define an automated method based on user interactions for identifying this role, avoiding expensive content analysis. In [**?**], the authors investigate both the behavioral context of user posting and the polarization of user responses. Furthermore, the authors of [**?**] present a broad exploration of posts, with a particular interest to comments. Here, they aim at fulfilling three different tasks. The first is analyzing a comment thread by looking at its topical structure and evolution; the second consists of using comment threads to enhance web search; the third aims at distilling useful features to predict the final score of a comment.

The authors of [**?**] investigate the success and group dynamics of online communities, focusing on Reddit ones. In detail, they identify four success measures desirable for most communities, spanning from the growth of the numbers of members to the volume of activities within the community, and capturing different kinds of success. They also consider the prediction of the final success of a new community.

In [**?**], the authors discuss the rise of new trends in complex networks by looking at vertices that "shine", i.e. high-degree vertices, also called network stars. They study the evolution of some complex networks, with Reddit among them. They analyze the temporal dynamics of the networks by looking at how different features, such as density and average clustering coefficient, change over time.

A relatively large set of different approaches and methodologies for characterizing several properties and aspects of Reddit can be found in literature. For instance, in [**?**], a mixed-methods approach studies a particular subreddit representing an online User Experience community (/r/userexperience), whose members socialize and learn together. Here, the authors identify five distinct social roles, such as the knowledge broker (i.e., a member that introduces knowledge to the community by sharing links) and the translator (i.e., a member that offers her academic knowledge into the community). Similarly, the authors of [**?**] present a study regarding highly related communities; they introduce a taxonomy considering two kinds of user, i.e. explorers and non explorers. Both in [**?**] and in [**?**], the introduced taxonomy is particularly specific and addresses only those users belonging to a particular subset of communities. Instead, the stereotypes we are proposing in this paper are general and can be applied to all subreddits.

In [**?**], the authors use text classification and computational critical discourse analysis to distinguish and interpret ideological differences between subreddits. In [**?**], the authors present a study regarding a quantitative, language-based typology of communities' identity, revealing how various social phenomena manifest across communities. The introduced taxonomy is based on two aspects of community identity, i.e. distinctiveness and dynamicity. User migration is studied in [**?**], where Reddit is examined during a period of community unrest, resulting in the identification of motivations for this kind of behavior. Political and ideological aspects emerging in Reddit are discussed in [**?**, **?**, **?**, **?**]. Finally, in [**?**], the authors present a mixed-method study of 100,000 subreddits and their rules, whose aim is to characterize effective mechanisms for community governance.

3 Dataset description

We start depicting the overall structure of Reddit in Figure 1. In the left part of this figure, each rounded box represents a subreddit. The central part shows a list of posts in the example subreddit /r/subreddit, where each color identifies a different type of posts (text, image or link to external resource). Finally, the right part illustrates the structure of a post, including its title and its comments, which are presented as a tree having the post as root.

Figure 1: A graphical overview of Reddit structure

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

In order to carry out our experiments, we used a server equipped with 16 Intel Xeon E5520 CPUs and 96 GB of RAM with the Ubuntu 18.04.3 operating system. We adopted Python 3.6 as programming language, its library Pandas to perform ETL operations on data, and its library NetworkX to perform operations on networks.

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

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

4 Preliminary investigations on Reddit data

In this section, we describe some preliminary investigations that we performed on Reddit. As pointed out in the Introduction, these are not the core of our paper, but they confirmed us the suitability of our dataset. Furthermore, some knowledge extracted here was extremely useful in the analyses described in the next sections. We group the following analyses in three subsets, which regard posts, comments, and authors, respectively. We describe each subset in a separate subsection.

4.1 Investigation on posts

In our first investigation on this topic, we determined the distribution of subreddits against posts. In Figure 2 we report the results obtained. This figure shows that the distribution follows a power law. This implies that most of the subreddits have very few posts, whereas very few subreddits have lots of posts. We computed the coefficients α and δ of the power law and we found that $\alpha = 1.651$ and $\delta = 0.014$. We also detected that the maximum number of posts in a subreddit is 2,370,456.

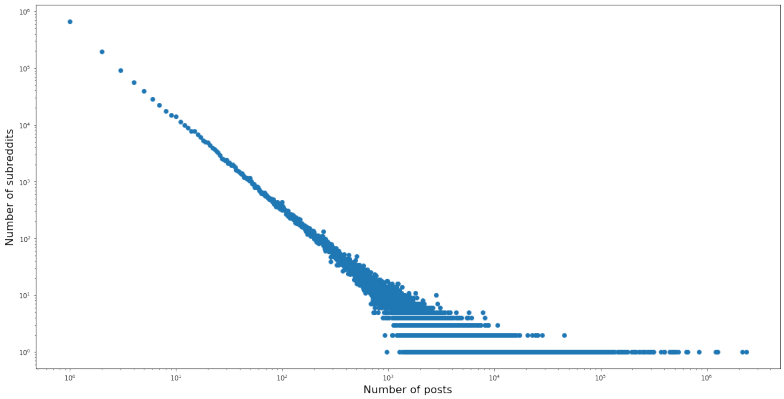

Figure 2: Distribution of subreddits against posts

Then, we determined the distribution of authors against posts. The results obtained are reported in Figure 3. This figure highlights that also this distribution follows a power law. Almost all the authors submitted very few posts, whereas only very few authors submitted lots of posts. We computed the values of the coefficients α and δ. Specifically, we obtained $\alpha = 1.431$ and $\delta = 0.016$. The maximum number of posts submitted by an author is 25,331.

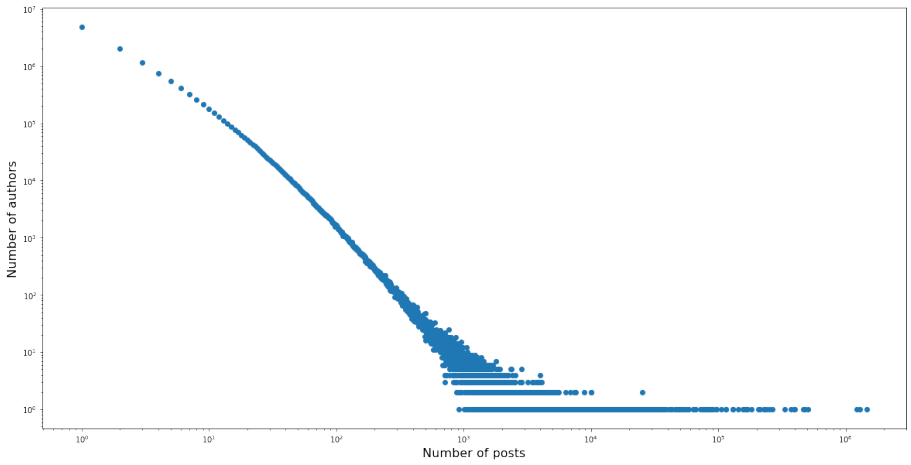

Figure 3: Distribution of authors against posts

Afterwards, we computed the distribution of posts against scores. The results obtained are reported in Figure 4, whereas, in Figure 5, we show a zoom of it focusing on very low values of score. Both

figures clearly show that the distribution follows a power law. In this case, we found that $\alpha = 1.600$ and $\delta = 0.005$. We also determined that the maximum score received by at least one post is 212,631, whereas the maximum number of posts with the same score is 51,721,824. Interestingly, these posts have associated a score equal to 1. Instead, the number of posts with a score equal to 0 or to 2 is much lesser. This trend can be explained by considering that a post submitted on Reddit starts with a score of 1. As a consequence, when no other author upvotes or downvotes it, the final score of the post is 1.

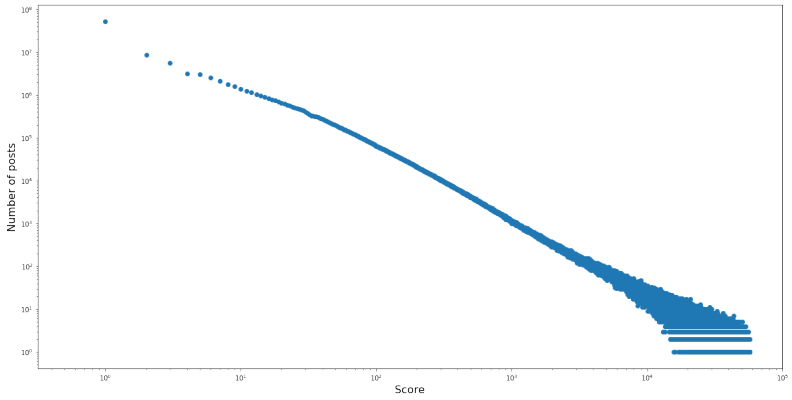

Figure 4: Distribution of posts against scores

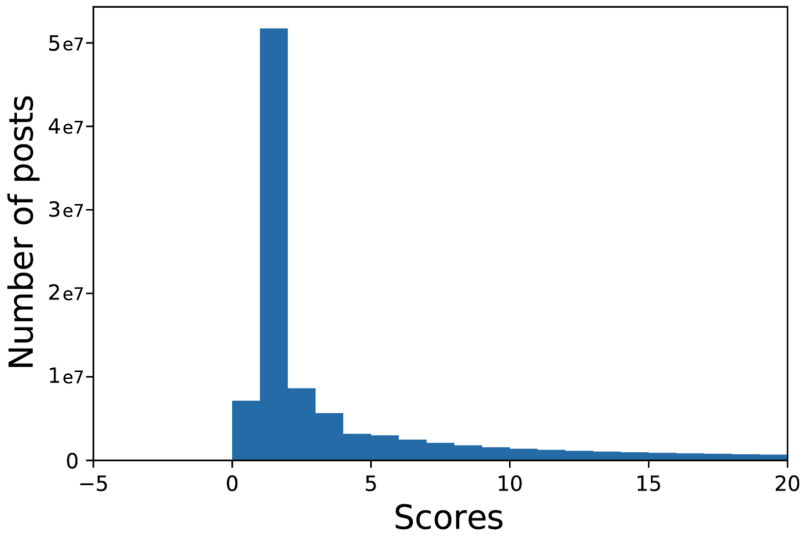

Figure 5: Zoom of the distribution of posts against scores focused on low values of scores

We also observe that no post has a negative score. This fact is probably due to `pushshift.io` that removed the posts with a negative score. So, the posts with a score equal to 0 become particularly important, because they are the only ones at our disposal that were judged negatively by at least one Reddit user. For this reason, we decided to investigate them deeply and, in the following, we call them

"negative" posts.

The number of authors who submitted at least one negative post is 2,907,549. Instead, the number of negative posts is 7,142,699.

We computed the distribution of authors against negative posts. The result is reported in Figure 6. From the analysis of this figure we can see that it follows a power law with $\alpha = 2.274$ and $\delta = 0.030$. The maximum number of negative posts submitted by a single author is 10,415. The number of authors with more than 100 negative posts is 1,884.

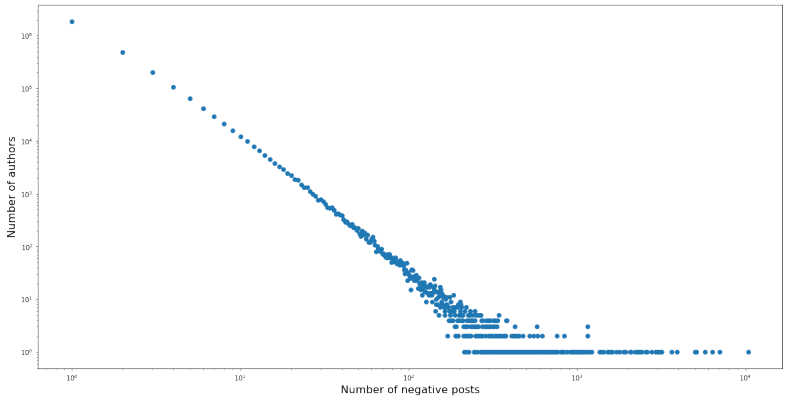

Figure 6: Distribution of authors against negative posts

At this point, we found particularly interesting to determine the distribution of authors against positive posts. The result is reported in Figure 7. Again, we have a power law distribution with $\alpha = 2.074$ and $\delta = 0.014$. This figure shows that the number of authors who submitted at least one positive post is 49,565,132. The number of positive posts is 115,425,931, whereas the maximum number of positive posts submitted by a single author is 1,471,177.

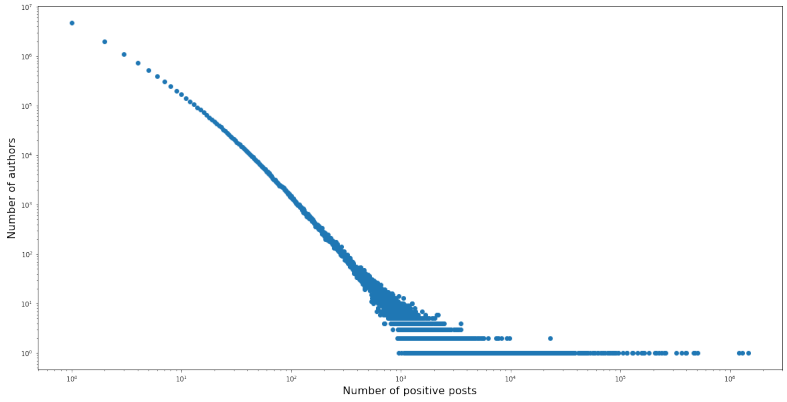

Figure 7: Distribution of authors against positive posts

Comparing the two distributions shown in Figures 6 and 7, we found that the number of positive posts is about 16 times the number of negative ones.

4.2 Investigation on comments

Firstly, we determined the distribution of subreddits against comments. The results are reported in Figure 8. Even in this case, we observe a power law distribution with $\alpha = 1.730$ and $\delta = 0.015$. We also found that the maximum number of comments for a single subreddit is 48,010,026.

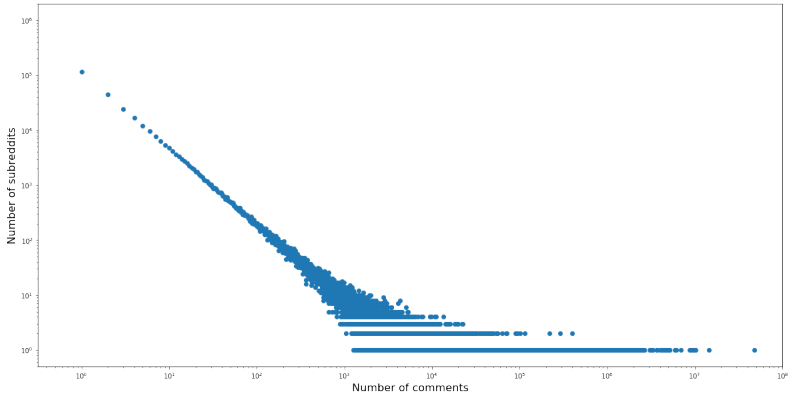

Figure 8: Distribution of subreddits against comments

Then, we determined the distribution of the average number of comments against the scores of the posts they refer to. The results obtained are reported in Figure 9. From the analysis of this figure, we can observe that we have a Gaussian distribution whose mean is at a score near to 50,000. The distribution, even if roughly Gaussian, presents several outliers. For instance, for a score equal to 79,470, we have a unique post with a number of comments equal to 71,225.

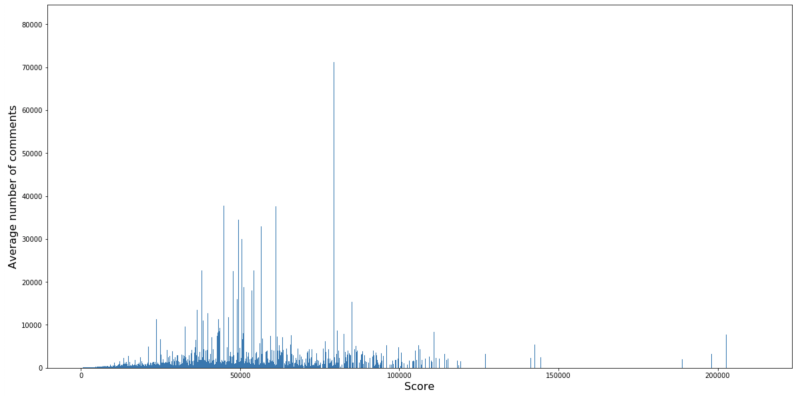

Figure 9: Distribution of the average number of comments against the scores of the posts they refer to

Next, we determined the distribution of posts against comments. The results obtained are reported in Figure 10. Again, we observe that this distribution follows a power law with $\alpha = 1.455$ and $\delta = 0.011$. The number of posts with only one comment is 16,531,169, whereas the maximum number of comments for a single post is 100,072.

Finally, we considered the 150 posts with the highest number of comments and the subreddits

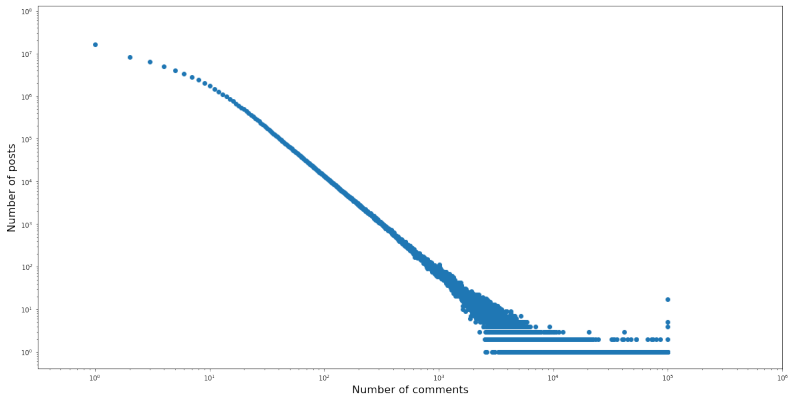

Figure 10: Distribution of posts against comments

they were submitted to. We obtained only 31 subreddits. Then we computed the average number of comments for *all* the posts submitted in each of these subreddits. The results obtained are reported in Figure 11. From the analysis of this figure, we can observe that the distribution is very irregular. It decreases quickly for the first three subreddits, very slowly for the next 13 subreddits, quickly for the next 9 subreddits and, finally, it suddenly drops and becomes almost zero.

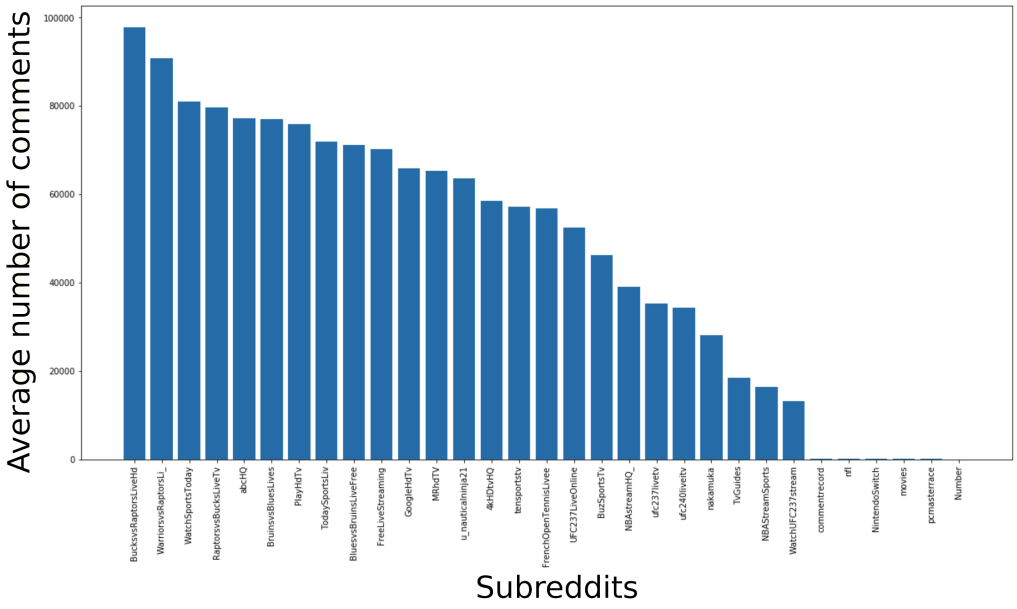

Figure 11: Distribution of the average number of comments submitted to the subreddits receiving the 150 most commented posts

4.3 Investigation on authors

First, we determined the distribution of authors against subreddits. The results are reported in Figure 12. From the analysis of this figure, we observe that it follows a power law with $\alpha = 1.702$

and $\delta = 0.081$. We have also found that the number of authors who posted on only one subreddit is 67,315, whereas the maximum number of subreddits where a single author submitted at least one post is 3,456.

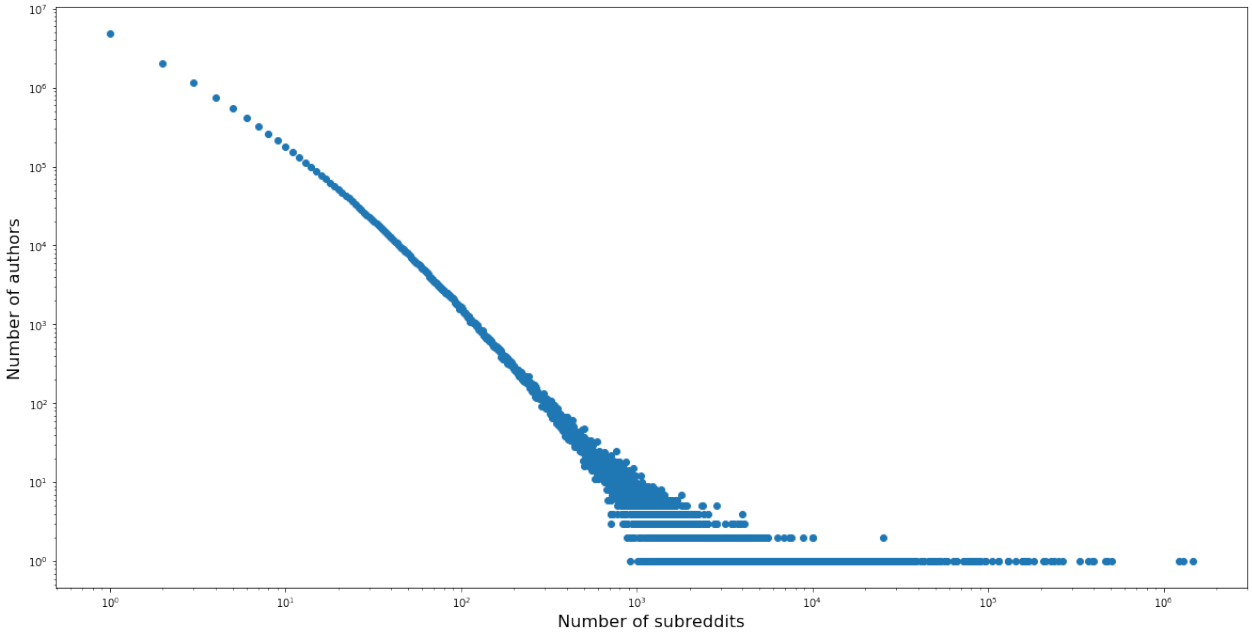

Figure 12: Distribution of authors against subreddits

Afterwards, we selected the 150 posts with the highest number of comments and the corresponding authors. Interestingly, we had only 26 authors for all the 150 posts. These can be considered as the most commented authors in Reddit and, maybe, they are influencers. Then, we computed the average number of comments for *all* the posts each author submitted. The results obtained are reported in Figure 13. From the analysis of this figure we can observe that the decrease of the distribution is roughly stepwise.

5 Stereotyping subreddits

In order to determine some possible stereotypes of subreddits, we start investigating the subreddit lifespan. As a first step, we considered the subreddits created in January 2019 and then verified the month when they performed their last activity (and, therefore, presumably died). The results obtained are reported in Figure 14. Here, an activity level of 1 implies that the subreddit died in the same month it was born, an activity level of 2 suggests that it died one month after it was born, and so on. An activity level of 8 indicates that it is still alive (we recall that our dataset comprises data from January 1^{st}, 2019 to September 1^{st}, 2019). We proceeded in the same way for the subreddits created in February, March, and so forth. For instance, in Figure 15, we report the trends of the subreddits created in February 2019 and in March 2019.

After this, we focused on those subreddits died in the same month they were born. We analyzed

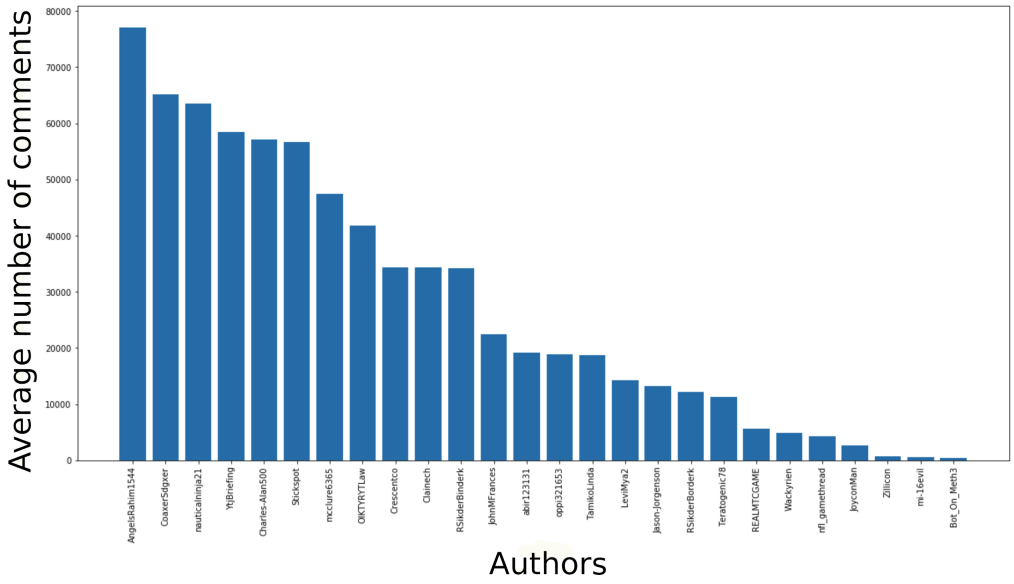

Figure 13: Distribution of the average number of comments received against the authors submitting the 150 most commented posts

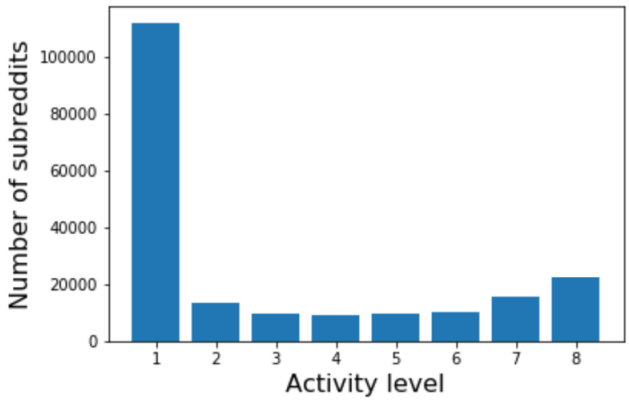

Figure 14: Lifespan of the subreddits created in January 2019

their corresponding lifespan and we observed that almost all of them died in the same day they were born. For instance, in Figure 16, we report the trends of the subreddits born and died in February 2019 and in March 2019.

Then, we decided to deeply investigate those subreddits died in the same day they were born. We computed their distribution against the number of their posts. Figure 17 shows what happens for January 2019; the same trend can be observed for the other months of this year. Clearly, this distribution follows a power law, a trend that can be observed also for similar subreddits born in the other months. From its analysis we observe that most of the subreddits, which died in the same day they were born, have only one post. At this point, we computed the distribution of these subreddits against the number of comments. In Figure 18, we show the subreddits of January 2019, even if the

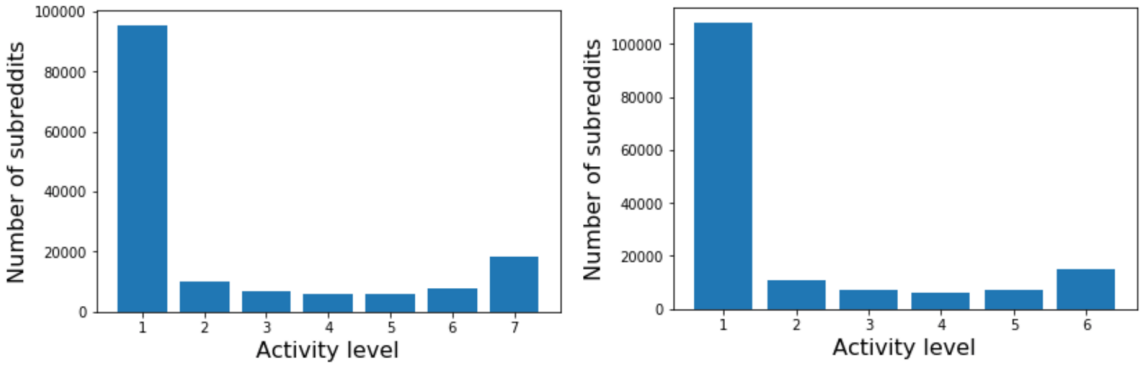

Figure 15: Lifespan of the subreddits created in February 2019 (at left) and March 2019 (at right)

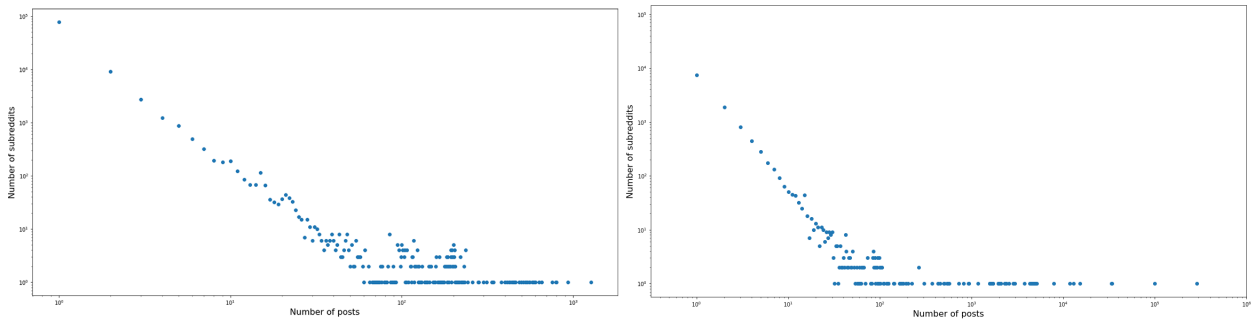

Figure 16: Lifespan of the subreddits born and died in February 2019 (at left) and March 2019 (at right)

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

Next, we examined a second class of subreddits, similar to the previous one. In fact, we selected all those subreddits that died one day after they were born. Again, we first computed their distribution against the number of posts. In Figure 19, we show what happens for the subreddits of January 2019; again, the same trend was found for all the other months. This distribution follows a power law, which was expected. The unexpected thing was that the minimum number of posts was 2 and not 1. Even more unexpectedly, this trend is also confirmed for the subreddits with the same features born in the other months. After that, we computed the distribution of these subreddits against the number of comments. In Figure 20, we show it for the subreddits of January 2019; the same trend can be observed for all the other months. From the analysis of this figure, we note that this distribution follows a power law. Furthermore, most of these subreddits have no comments.

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

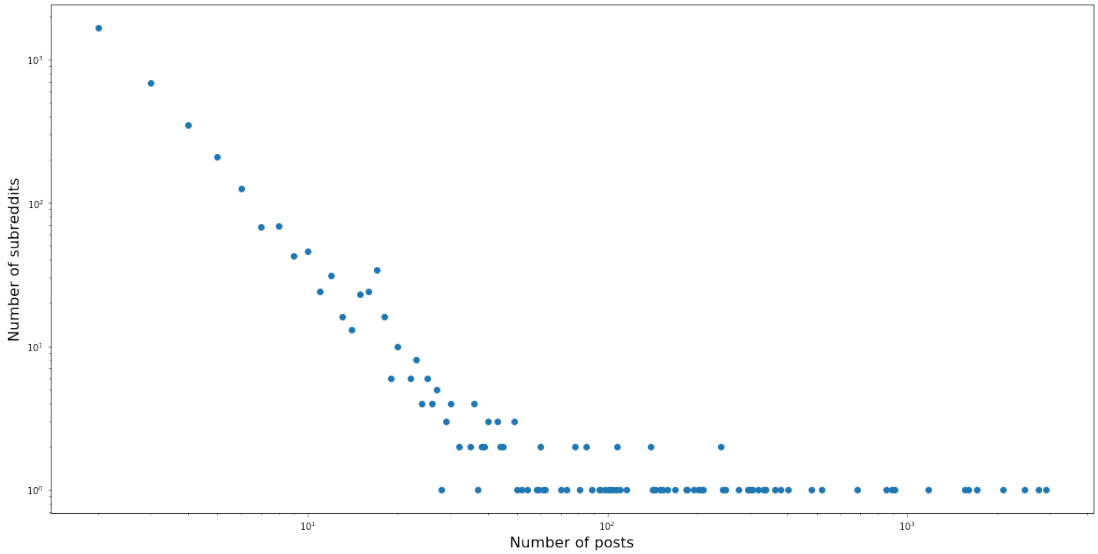

Figure 17: Distribution of the subreddits of January 2019 died in the same day they were born against the number of their posts

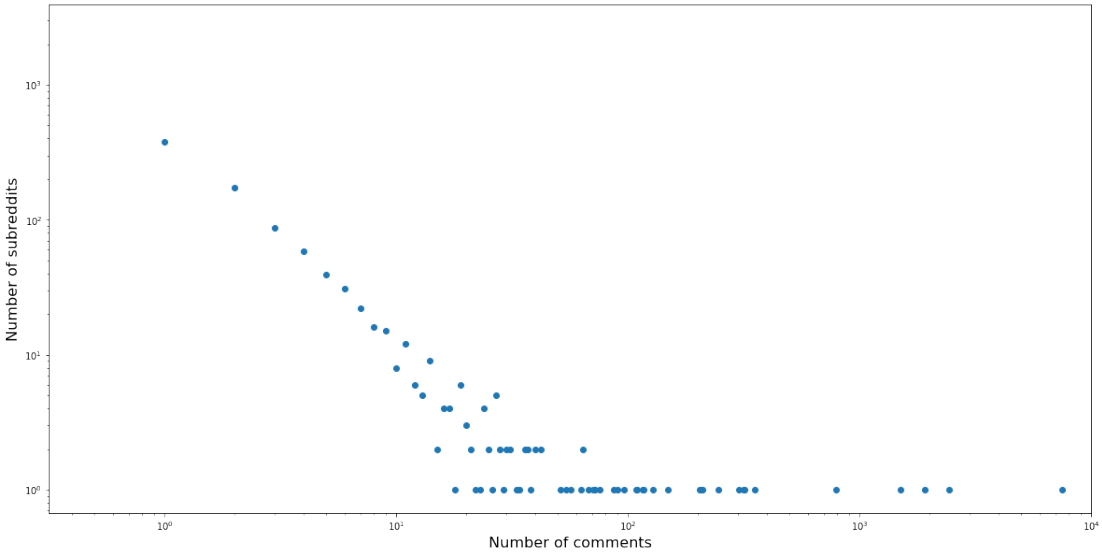

Figure 18: Distribution of the subreddits of January 2019 died in the same day they were born against the number of their comments

deepening what we have found previously, we have determined the following stereotypes characterizing the subreddits "dead in crib" (i.e., those subreddits who died at most one day after they were born):

- *User Profile*: it is associated with a user profile.
- *Unsuccessful Subreddit*: it initially stimulated several interactions. However, after few hours, these interactions finished and it quickly died.

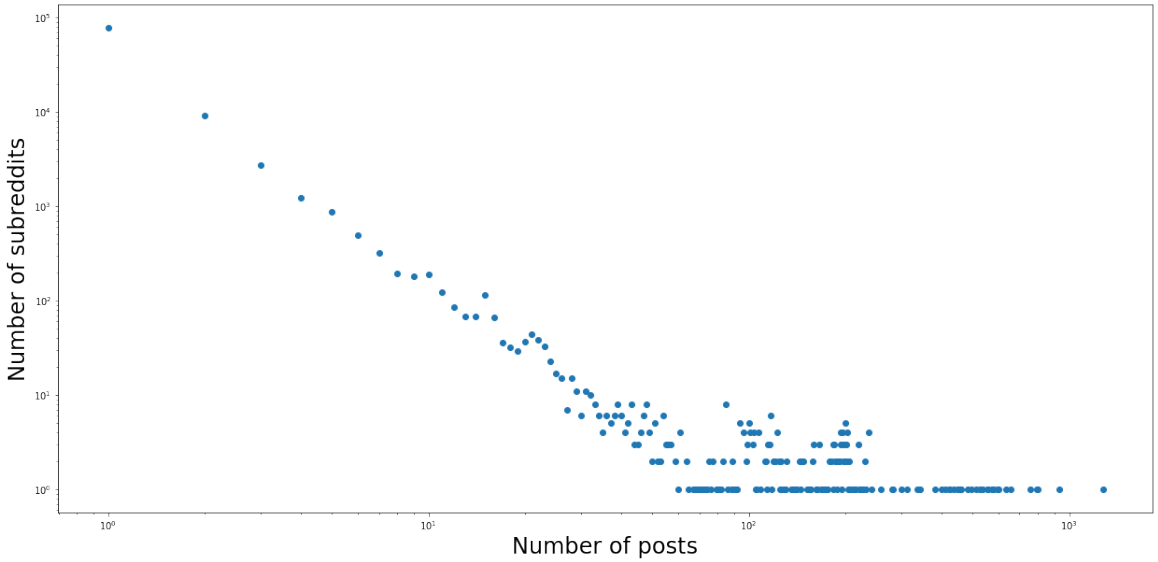

Figure 19: Distribution of the subreddits of January 2019 died one day after they were born against the number of their posts

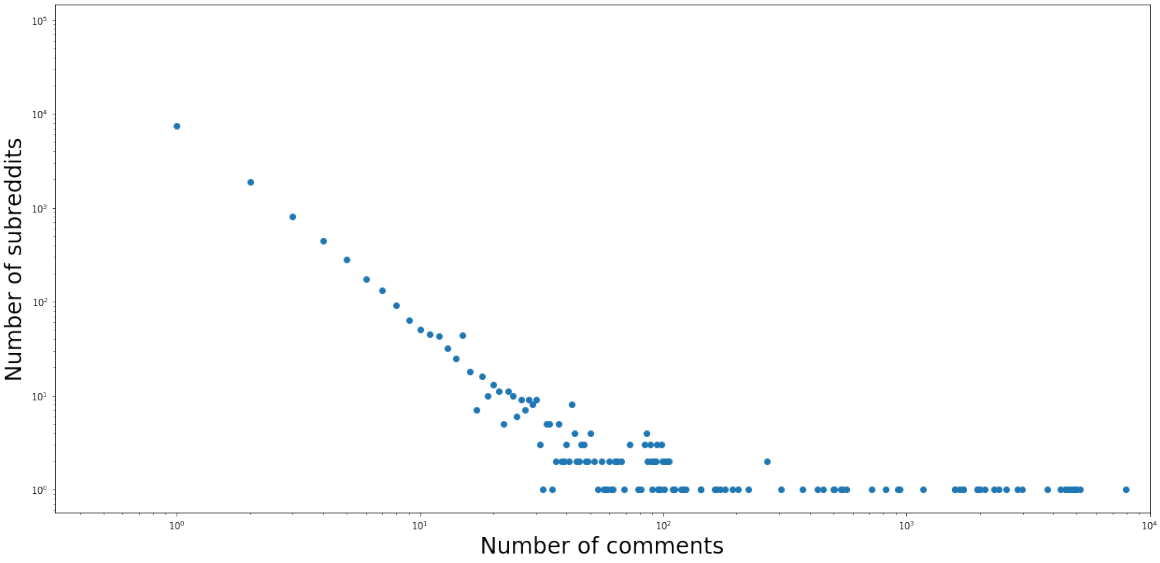

Figure 20: Distribution of the subreddits of January 2019 died one day after they were born against the number of their comments

- *Comment Grabber*: it had at least one post capable of stimulating a debate, even if minimal.
- *Private Community*: it requires an invitation to be accessed. It is often associated with a specific event of interest for a specific community.
- *Banned Subreddit*: it was banned probably because it was associated with a spammer.

- *Bot*: it can be recognized because its posts are always similar and consist of links and comments with links.

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

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

Taking these three taxonomies into consideration, the previous stereotypes can be classified as shown in Tables 1 and 2.

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

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

Table 1: Classification of stereotypes concerning the subreddits "dead in crib" - Few posts case

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

Table 2: Classification of stereotypes concerning the subreddits "dead in crib" - Many posts case

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

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

Based on the three taxonomies defined above, the previous stereotypes can be classified as shown in Tables 3 and 4.

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

Table 3: Classification of stereotypes concerning the subreddits "survivors" - Few posts case

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

Table 4: Classification of stereotypes concerning the subreddits "survivors" - Many posts case

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

- *User Profile*, *Niche Subreddit*, *Bot*, *Cringe / NSFW Subreddit*, *Private Community*, *Banned Subreddit*: these are the same ones we have seen for the previous categories.
- *Unsuccessful Boomer*: it was successful for a while, but died after a period of decline.
- *Unsuccessful Zombie*: it was born without infamy nor praise, managed to survive for a while in a gray way and, finally, died.

Based on the three taxonomies that we defined above, the previous stereotypes can be classified as shown in Tables 5 and 6.

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

Table 5: Classification of stereotypes concerning the subreddits "undelivered promises" - Few posts case

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

Table 6: Classification of stereotypes concerning the subreddits "undelivered promises" - Many posts case

6 Stereotyping authors

In order to determine the possible author stereotypes, we proceeded in a way analogous to what we have done for defining subreddit stereotypes. In fact, also for authors, we found three macro-categories of stereotypes, namely "very positive", "neutral" and "very negative" authors. To better understand the reasoning underlying these categories, we recall that, in Section 4.1, we have found that the number of positive posts is about 16 times the number of negative ones in Reddit. As a consequence, it is possible to use this result as a baseline for a preliminary author classification. Specifically, we considered an author as "very positive" if the number of positive posts submitted by her is at least $2 \cdot 16 = 32$ times the number of negative ones, which means at least twice the typical number of positive posts submitted for each negative one by a user. Instead, we considered an author as "neutral" if the number of positive posts submitted by her is between 1 and 16 times the number of negative ones. Finally, we considered an author as "very negative" if the number of negative posts submitted by her is at least 16 times the number of positive ones. Clearly, this classification is not exhaustive and it is also empirical because it derives from our observation on the behaviors of users in Reddit. However, we feel that it is useful to provide a first definition of three macro-categories of author stereotypes possibly interesting for application scenarios.

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

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

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

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

Based on the two taxonomies that we defined above, the previous stereotypes can be classified as shown in Table 7.

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

Table 7: Classification of the stereotypes concerning "very positive" authors

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

- *Unsuccessful Author*: this stereotype is the same as we have seen for "very positive" authors.
- *Spammer*: she is an author submitting a lot of spam posts evaluated negatively by other users.
- *Hatred Sower*: she is a user whose goal is attacking minority groups with hate posts or comments.

- *Instigator*: she is an author using every opportunity to make herself known. For her, it is not important how she is judged, but the fact that one speaks of her.

Based on the two taxonomies defined above, the previous stereotypes can be classified as shown in Table 8.

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

Table 8: Classification of the stereotypes concerning "very negative" authors

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

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

Based on the two taxonomies defined above for authors, the previous stereotypes can be classified as shown in Table 9.

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

Table 9: Classification of the stereotypes concerning "neutral" authors

7 Analyzing author assortativity

The concept of "assortativity" or "assortative mixing" in a social network was introduced in a famous paper of Newman [**?**]. It is strictly related to the concept of homophily [**?**] and indicates a network node's predilection to relate to other nodes that are somewhat similar. Several possible similarities could be considered in assortativity, but the most investigated one is node degree. In the past, assortativity has been largely analyzed in several social media [**?**]. In this section, we aim at checking if a form of degree assortativity exists in Reddit; in particular, we focus on co-posters, i.e. authors submitting posts on the same subreddit.

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

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

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

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

First of all, we computed the degree centrality of the nodes of $\mathcal{P}$. In Figure 21, we report the corresponding distribution. This figure shows that degree centrality follows a power law, even if disturbed. This result is in line with the theory regarding this kind of centrality [**?**]. The maximum value of degree centrality is 1,820,412, while the minimum value is 0.

We sorted the corresponding authors in a descending order, based on their degree centrality, to verify the possible presence of assortativity in Reddit. Then, we divided the sorted list into intervals of authors. In particular, we considered equi-width intervals $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_{40}\}$, each consisting of 312,500 authors[5]. As a consequence, the interval $\mathcal{I}_k$, $1 \leq k \leq 39$, contained the authors of the sorted list comprised in the interval $(312,500 \cdot (k-1), 312,500 \cdot k]$, open at left and closed at right. The interval $\mathcal{I}_{40}$ contained the authors comprised in the interval $(12,187,500\ ,\ 12,464,188]$.

First of all, we considered the first interval $\mathcal{I}_1$ and, for each interval $\mathcal{I}_k$, $1 \leq k \leq 40$, we determined how many authors of $\mathcal{I}_1$ are connected to at least one author of $\mathcal{I}_k$. The results obtained are reported in Figure 22. Then, we computed the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_1$. The results obtained are reported in Figure 23. From the analysis of Figures 22 and 23, it is clear that a strict correlation (i.e., a sort of backbone) exists among the authors with the highest degree centrality.

In order to prove the statistical significance of our results, we generated a null model to compare our findings with the ones obtained in an unbiasedly random scenario. Specifically, we built our null model shuffling the arcs of $\mathcal{P}$ (that, in our case, represent co-posting relationships) among the nodes of this network. In this way, we left unchanged all the original features of $\mathcal{P}$ with the exception of

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

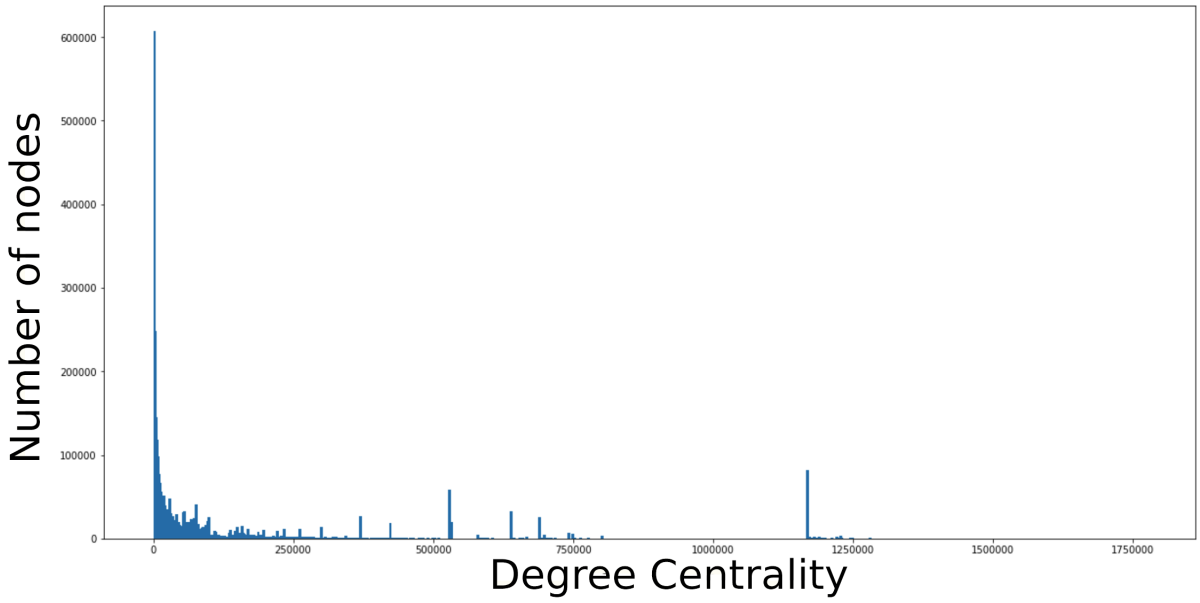

Figure 21: Distribution of degree centrality for the nodes of $\mathcal{P}$

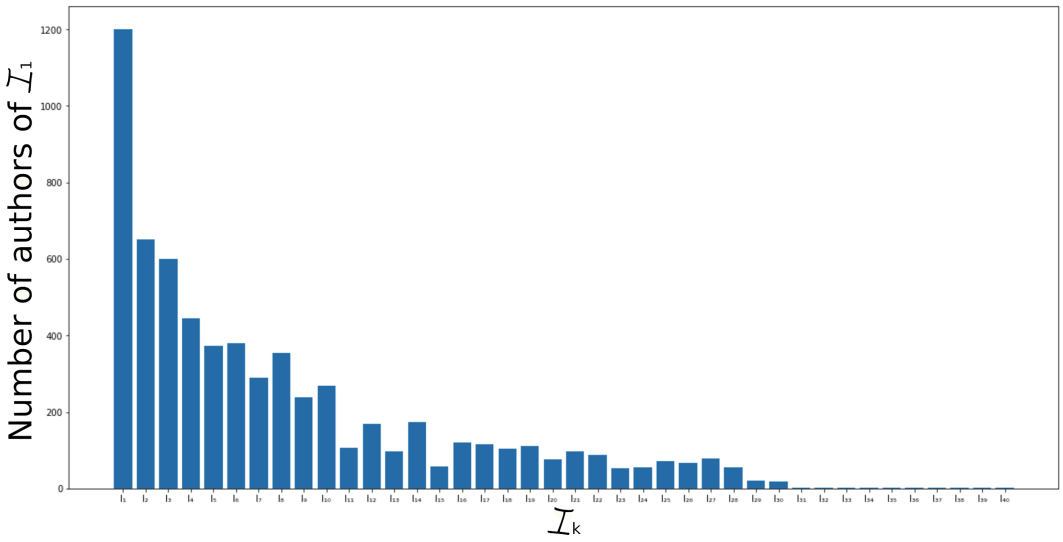

Figure 22: Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$

the distribution of co-posting tasks, which became unbiasedly random in the null model. After that, we repeated the previous analyses on the null model. The results obtained are reported in Figures 24 and 25. Comparing these figures with Figures 22 and 23, we can see that the distributions represented therein are similar, in a way that many of the intervals with the highest values in Figures 22 and 23 continue to reach the highest values in Figures 24 and 25. However, in this last case, the values are much smaller. Therefore, we can conclude that the behavior observed in Figures 22 and 23 (and the consequent possible degree assortativity revealed by them) is not random but it is intrinsic to Reddit.

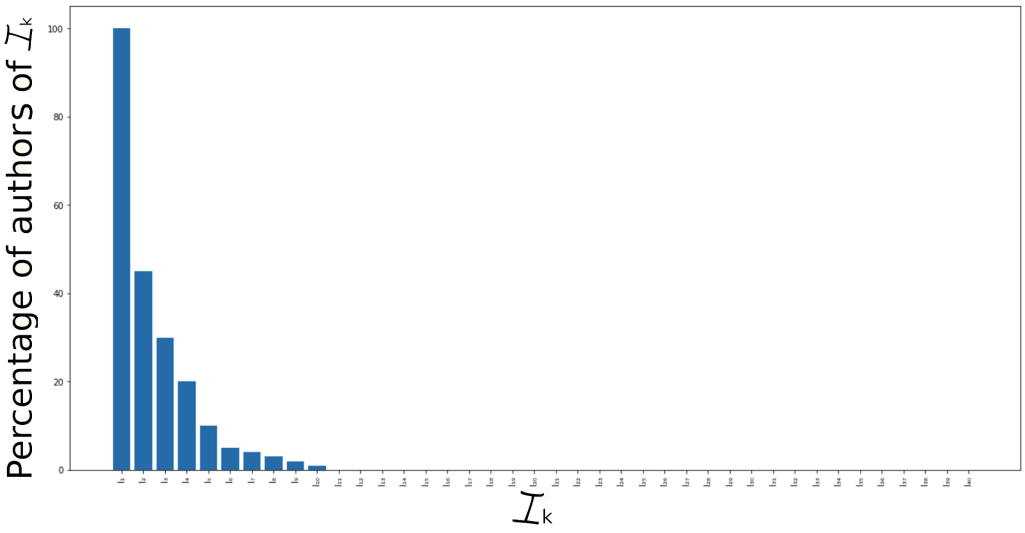

Figure 23: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$

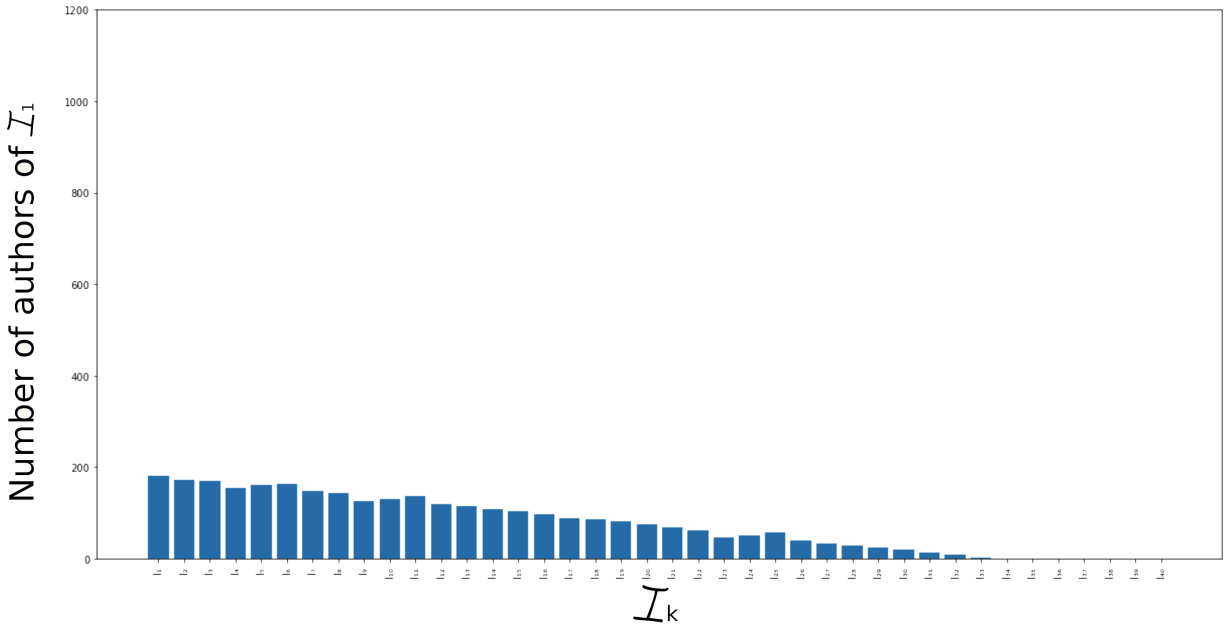

Figure 24: Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ in the null model

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

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

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

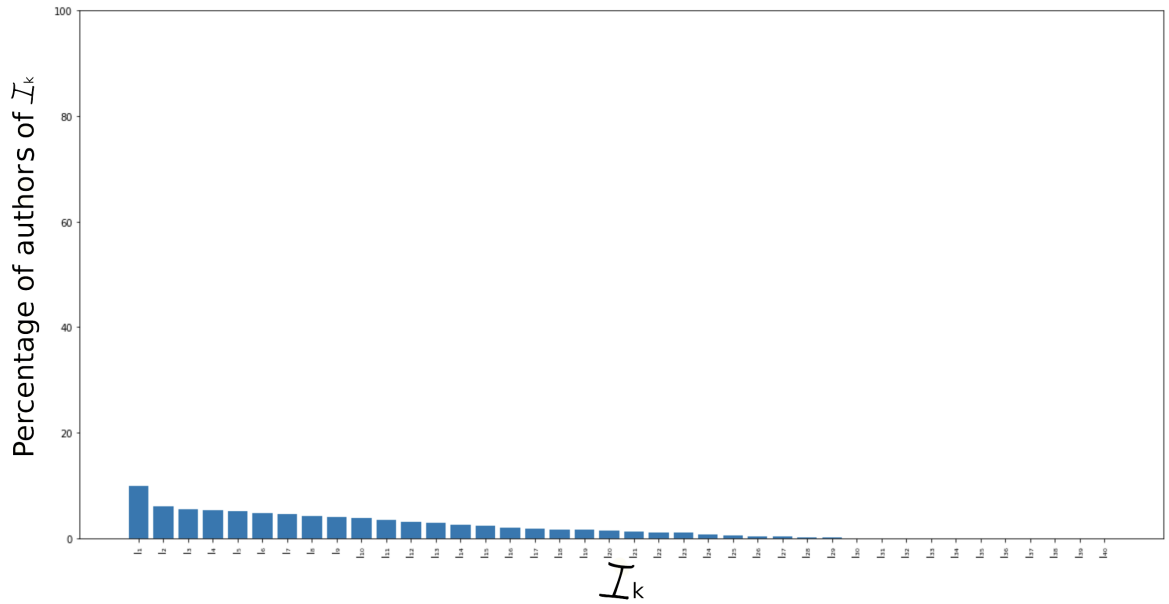

Figure 25: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$ in the null model

Figure 26 reports the number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$, whereas Figure 27 shows the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{20}$. From the analysis of these figures, it emerges a strict correlation between the authors with an intermediate degree centrality.

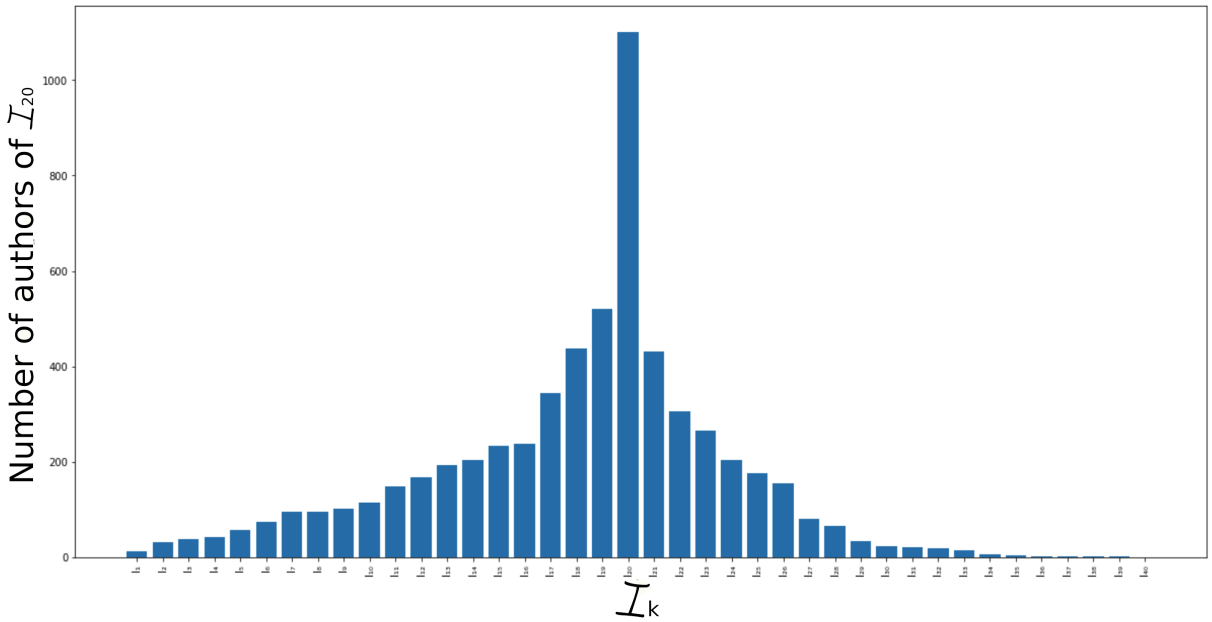

Figure 26: Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$

Also in this case, we compared these findings with the ones obtained in the null model. These last

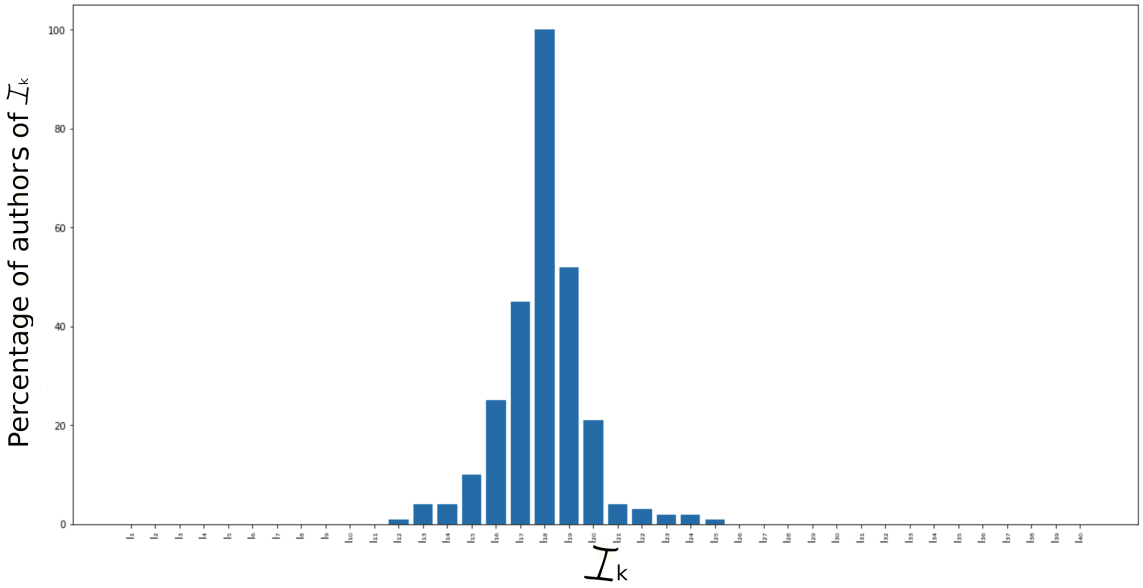

Figure 27: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$

ones are reported in Figures 28 and 29. Looking at these results and the ones represented in Figures 26 and 27, we can conclude that, again, the behavior observed in these last figures is not random but it is a property of Reddit.

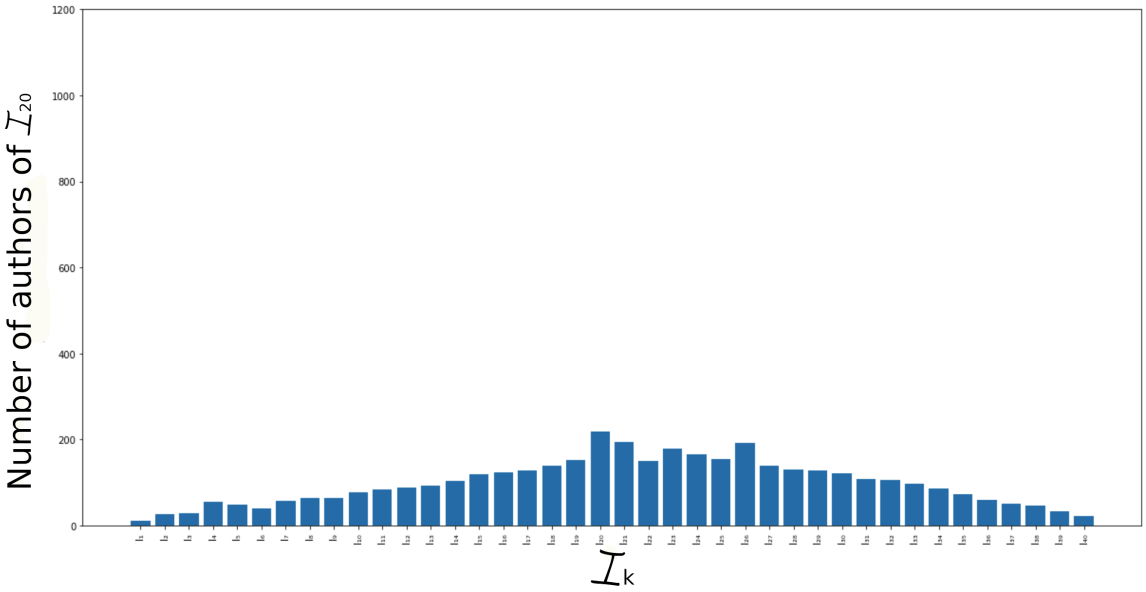

Figure 28: Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ in the null model

Finally, Figure 30 reports the number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$, whereas Figure 31 shows the percentage of authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{39}$. Again, there is a strict correlation between authors with a low degree centrality. Also for this last

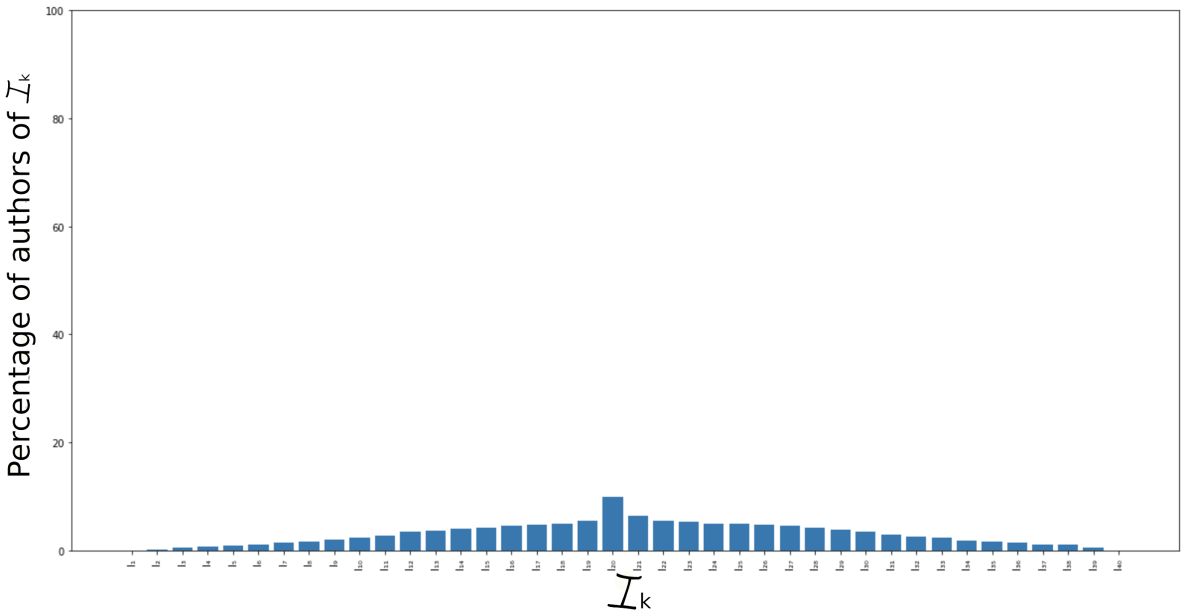

Figure 29: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$ in the null model

case, we compared the results obtained with the ones returned using the null model. We report these last ones in Figures 32 and 33. The comparison of these figures with Figures 30 and 31 confirms that the behavior observed in them is a property intrinsic to Reddit.

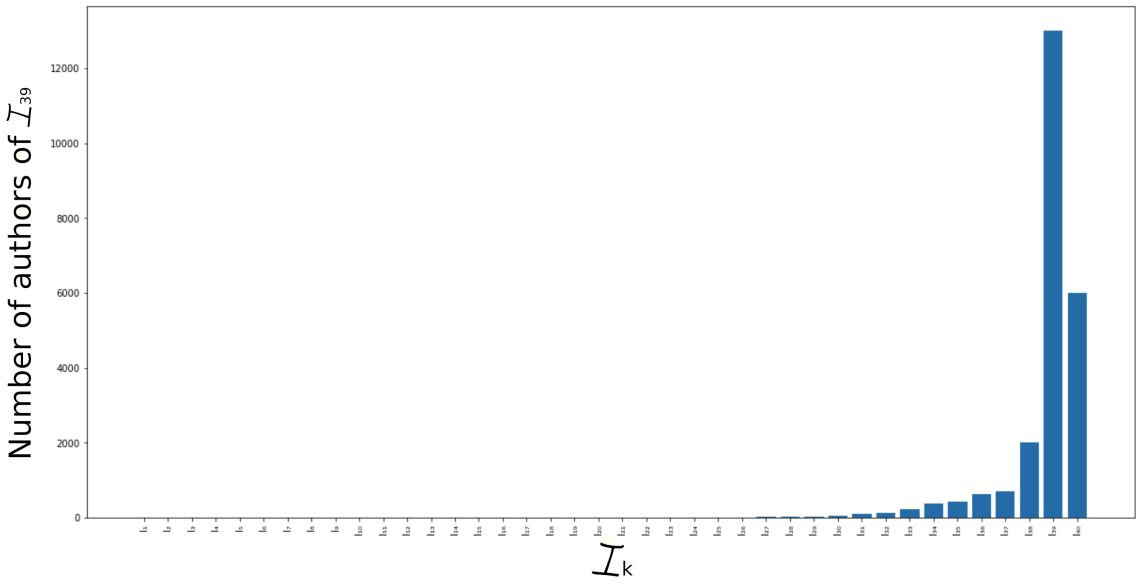

Figure 30: Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$

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

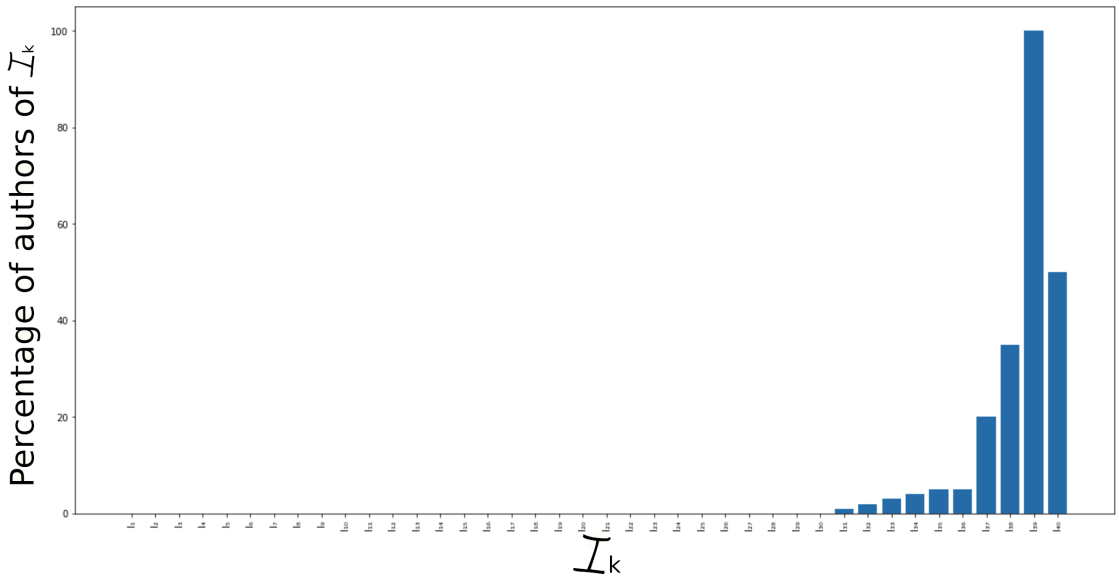

Figure 31: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$

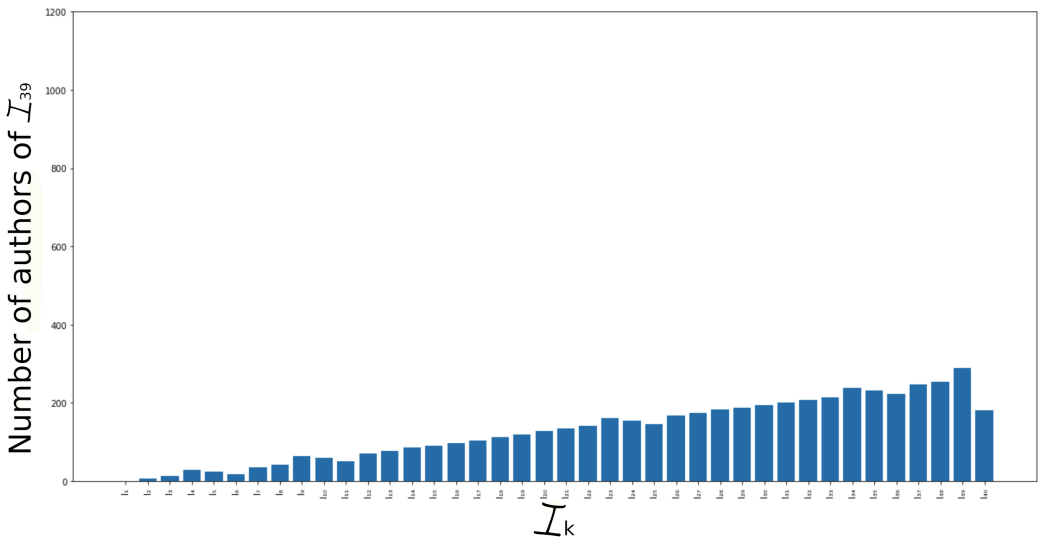

Figure 32: Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ in the null model

This important result can be explained considering the concept of karma and the posting rules in Reddit. Indeed, in this platform, each user has associated a karma, which is a score taking her past "reputation" into account. Generally, users with high karma are very active and, often, submit a lot of appreciated posts. As a consequence, it is presumable that they have a high degree centrality. In other words, a direct correlation between karma and degree centrality can be recognized for authors. Now, the posting rules of Reddit state that each subreddit has associated a minimum threshold of karma [**?**, **?**, **?**] so that only the authors with a karma higher than this threshold can submit a post on it. This threshold is dynamic and changes over time. Clearly, when it is low, all the authors can submit

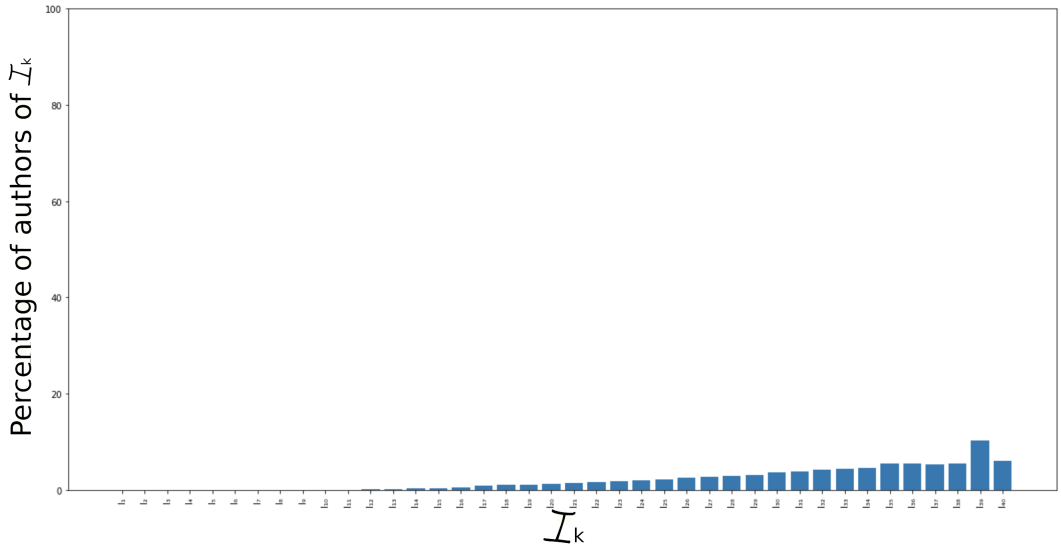

Figure 33: Percentage of authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$ in the null model

their posts on the subreddit. When it grows, the authors with a low karma (and, presumably, with a low degree centrality) cannot submit posts on it. Finally, when it becomes high, only the authors with a high karma (and, presumably, a high degree centrality) can submit posts on it. This way of proceeding tends to segment users into groups having homogeneous degree centralities.

8 Possible applications of stereotypes

This section presents some possible applications of the stereotypes previously investigated. It consists of two subsections. The first explains how subreddit stereotypes could be employed to make a subreddit successful. The second highlights how particular types of author stereotypes prove to be useful to improve the content quality of subreddits.

8.1 Subreddit stereotypes

In Section 5, we defined several subreddit stereotypes belonging to three macro-categories, namely "dead in crib", "survivors" and "undelivered promises". A first application of this research can be the definition of some guidelines to follow in order to make a subreddit successful. Indeed, knowing how a subreddit became successful (resp., unsuccessful) can lead to the characterization of "positive" (resp., "negative") actions that can influence the "lifespan" of a new subreddit. For instance, consider the subreddit /r/meme. It was activated during 2008 and, at the time of writing, has about 806,000 users. Certainly, it represents an example of a successful subreddit. Here, the authors post high quality and engaging contents. This kind of behavior could be registered as a "best practice" in the guidelines. On the other hand, a subreddit containing only few contents from few authors is an example of an unsuccessful subreddit. This failure could be caused by a lack of engaging contents posted in it. Clearly, what said above provides just an idea of what these guidelines could contain.

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

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

8.2 Author stereotypes

In Section 6, we defined some possible author stereotypes. Some of them are strictly related to the homonymous or corresponding subreddit stereotypes. Other ones, instead, are intrinsic to human behavior and, in particular, to the concept of author. For example, consider "Fame Seekers" and "Content Creators". These users could represent the target of a proposal of an advertising campaign aiming at promoting them. Take, for instance, a painter or a digital artist, who has been classified as "Fame Seeker". An advertising company can easily persuade her to give it an engagement to promote her image.

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

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

9 Conclusion

In this paper, we have presented an investigation on Reddit, whose aim was analyzing three aspects of this social platform that are interesting for both the theory and the practice.

First, we have examined related literature and we have described the dataset used for our investigation. Then, we have illustrated some preliminary analyses that allowed us to gather some (partially expected) information, useful to correctly carry out the following activities and interpret the corresponding results.

The first knowledge detected in our investigation is subreddit stereotypes. We have explained the way of proceeding that we followed to determine them, we have defined three macro-categories and, for each of them, a certain number of stereotypes. Finally, we have proposed three orthogonal taxonomies and we have classified the detected stereotypes according to them. We have proceeded in the same way performing the second main task of our investigation, namely the definition and the classification of author stereotypes.

Afterwards, we have focused on a more theoretical issue. In fact, analogously to what has been carried out for other social platforms, we have verified if Reddit is assortative, and in which way. We have found that a degree assortativity exists in Reddit and that it involves co-posters. Finally, we have presented several applications that could benefit from subreddit and author stereotypes.

In the future, we plan to develop our research on Reddit along several directions. First of all, we would like to carry out a deep investigation on NSFW subreddits. In fact, in spite they are very numerous, few analyses on them have been performed in the past literature. Furthermore, in Section 8.1, we have seen that the merge, or at least the integration, of related subreddits could be extremely beneficial. Therefore, we plan to define an approach that finds possible subreddits to merge or to integrate and, then, suggests the tasks necessary to carry out this activity. Last, but not the least, we would like to define an approach to find duplicate accounts, i.e. two or more Reddit accounts belonging to the same person. We would like to understand the main motivations leading a user to adopt multiple accounts and verify if she has different behaviors in different accounts.