Co-posting Author Assortativity in Reddit

Francesco Cauteruccio[1], Enrico Corradini[2], Giorgio Terracina[1], Domenico Ursino[2], and Luca Virgili[2]

[1] DEMACS, University of Calabria
[2] DII, Polytechnic University of Marche

Abstract. In the context of social networks, a renowned paper of Newman introduced the notion of "assortativity", also known as "assortative mixing". Strictly akin to the concept of homophily, it shows how much a node tends to associate with other nodes somewhat similar to it. Degree centrality is the most used similarity metrics for evaluating assortativity between nodes, but several more could be dealt with. Assortativity was deeply investigated in many past researches, given different social platforms. However, Reddit was not one of the social networks taken into account, even if it is a really popular social medium. In this paper, we want to find out the possible presence of a form of assortativity in Reddit; in particular, we focus our analysis on co-posters, i.e. authors posting contents on the same subreddit.

Keywords: Reddit; Co-posters; Assortativity; Social Network Analysis; Degree Centrality

1 Introduction

Assortativity and degree assortativity were introduced in a renowned paper of Newman [17]. Here, the author defines a measure of assortativity for networks showing that real social networks are often assortative, whereas technological and biological networks tend to be disassortative. He also models an assortative network and exploits it for analytic and numeric studies. At the end of this analysis, he finds that assortative networks tend to percolate more easily than disassortative ones and that they are more robust to node removal. Another important study concerning social network assortativity was proposed in [18]. In this paper, the authors confirm the results of [17] and analyze the relation between clustering and assortativity in communities inside a social network. Recently, a detailed overview of assortative mixing in complex network was presented in [19]. Here, the authors investigate assortativity, and in particular degree assortativity, in different kinds of complex network. The concept of assortativity in social networks is a specific case of homophily. It comes from the famous homophily principle “similarity breeds connection” [13] that can be applied for network ties of every type. The result is that people’s personal networks are homogeneous w.r.t. many sociodemographic, behavioral, and intrapersonal characteristics.

After the famous paper of Newman, a lot of researchers started to investigate assortativity in social networks. However, in spite of this, there are several platforms (many of them famous) where assortativity has not been yet investigated. One of them is Reddit[3]. This is a heterogeneous crowd-sourced news aggregator and online social network, 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[4]. In Reddit, users can post their contents as texts, images or 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 latests content 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.

This paper aims at fulfilling the gap mentioned above and presents some analyses we performed in order to evaluate assortativity in Reddit. For this purpose, we first built a dataset with all the posts published in Reddit from January 1^{st}, 2019 to September 1^{st}, 2019. Then, we performed several analyses on it. Starting from this dataset, we built a suitable social network representing co-posting activities in Reddit. Then, we carried out several investigations on this network and we compared the results obtained from them with the ones returned by operating on a corresponding null model. At the end of this task, we found that Reddit is assortative with respect to degree centrality, as far as the co-posting relationship is concerned and we defined a hypothesis that explains this result.

The outline of this paper is as follows: In Section 2, we describe related literature. In Section 3, we illustrate the dataset we used for our analyses. In Section 4, we perform our investigation on assortativity in Reddit. Finally, in Section 5, we draw our conclusions and have a look at future developments of our research.

2 Related work

As previously pointed out, [17] and [18] can be considered as the founding fathers of the notion of assortativity. After this papers, in [7], the authors modeled biological, technological and online social networks starting from microscopical mechanisms of growth. Exploiting this model to perform statistical evaluations, they found that the statistical properties of biological, technological and online social networks are in good agreement with those of the real-world social networks of scientists co-authoring papers in condensed matter physics. Here, assortativity plays a key role. Indeed, the authors show that online social networks are generally assortative, whereas the majority of technological and bi-

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

[4] `https://www.redditinc.com/`

ological networks appear to be disassortative with respect to degree centrality. The investigation of [7] was expanded in [10]. Here, the authors proposed an analysis on assortativity/disassortativity for different kinds of network. Specifically, they considered the same network categories highlighted in [17, 18, 7]. They both confirmed the disassortativity of biological and technological networks and the assortativity of real social networks, analogously to what was shown in [7]. Differently from the wide-spread belief and the results of [7], they found that not all of online social networks are assortative. Almost all the results of [10] were confirmed in [9].

Online social networks simulating real life activities show an opposite behavior. The authors of [23] analyzed the assortativity and other network parameter on both standard social graphs and interaction graphs. They showed that the latter present a higher assortativity than the former. The authors of [6] present a study on degree assortativity for co-author networks. In [8], an interesting investigation on the relationship between assortativity and centrality is presented. Here, the authors study the relation between the degree-degree correlation coefficient and the BC-BC (i.e., Betweenness Centrality-Betweenness Centrality) one. In [11], a detailed study on the relation between Shannon entropy and degree assortativity was presented. Here, the authors defined a general class of degree-degree correlated networks and obtained the corresponding Shannon entropy starting from some suitable parameters. They found that the maximum entropy does not typically correspond to neutral networks but to either assortative or disassortative ones.

In [4], the authors investigated the assortativity of psychological states in real world social networks and online social networks. Specifically, they wanted to check the tendency of online social networks to be assortative, as it happens for real world social networks. The authors of [12] further analyzed assortativity on Twitter. They crawled this network and obtained 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. They took into consideration several network parameters, like degree distribution, diameter, reciprocity of user friendship declaration, homophily and assortativity. The authors of [3] proposed an interesting application of degree assortativity. They exploited this measure, along with several other ones, to classify YouTube users in spammers, promoters, and legitimates.

The concept of assortativity was also expanded along several directions. For instance, the authors of [22] extended and evaluated this concept on a weighted social network representing research collaborations. Another interesting extension is the concept of type assortativity that defines a way to measure if and how a social graph belonging to a single type exhibits homophily. In [2], the authors used paths, walks and random walks to define the concept of high order assortativity and showed that classical assortativity can be considered as a particular case of the new proposed notion. They also presented several examples and applications to airline networks and Enron e-mail networks.

Assortativity was also considered in several other analyses, such as node classification and network robustness measurements. The authors of [16] used assortativity to improve the prediction of node attributes, based on the fact that this measure provides information about each node, given its neighbors. This approach is particularly useful in those situations where data is inaccurate or missing. In [20], the authors measured the robustness of network community by means of a metric called “community assortativity”, based on the classical notion of assortativity. Finally, in [5], the authors expanded the concept of assortativity from social networks to social internetworking systems, i.e. systems where two or more social networks interact with each other through common users called bridges.

3 Dataset description

Figure 1 reproduces how Reddit is structured. Each rounded rectangle on the left represents a subreddit. The central part depicts an example list of posts from the subreddit /r/subreddit. Here, each post is associated with a type (texts, images or link to external resources) determined by a color. In the right, we can see the structure of a post, made up of a title (the root) and comments.

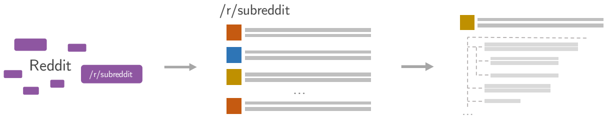

Fig. 1. A graphical overview of Reddit structure

Data used in our investigation activity was downloaded from `pushshift.io`, which is a website well-known as a Reddit data source. We obtained all the posts found on Reddit from January 1^{st}, 2019 to September 1^{st}, 2019. All contents posted in a month were added to the dataset at the end of the next month. We had a total of 150,795,895 posts available for our analyses. Each post had the following set of attributes provided by `pushshift.io`: `id`, `subreddit`, `title`, `author`, `created_utc`, `score`, `num_comments` and `over_18`.

The server used for our experiment was equipped with 16 Intel Xeon E5520 CPUs and 96 GB of RAM with the Ubuntu 18.04.3 operating system. Python 3.6 was the programming language used for the analyses, along with its library Pandas, for ETL operations on data, and its library NetworkX, for operations on networks. Performing ETL operations, we found that some authors who left Reddit wrote posts being in our dataset. So, we decided to delete them. After this activity, we had 122,568,630 posts in total. Beginning from this cleaned data, we figured out that the number of authors who wrote these posts was equal to 12,464,188. The number of subreddits they posted was 1,356,069.

4 Analyzing author assortativity

This section represents the core of our paper. In fact, it aims at verifying if a form of assortativity exists in Reddit. To do so, we focused on co-posters, i.e. authors who post on the same subreddit.

Co-posting network $\mathcal{P}$ is the support network we defined to perform our analyses. 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 posted at least once. 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) posted at least once 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.

We have that the number of nodes of $\mathcal{P}$, which is 12,464,188, is exactly the same as the number of authors of our testbed. On the other hand, the arcs of $\mathcal{P}$ are about 925 billions. We computed that the density of the network is 0.00596, while the average clustering coefficient is 0.43753.

The first task done was evaluating the degree centrality of the nodes of $\mathcal{P}$. In Figure 2, we show the corresponding distribution.

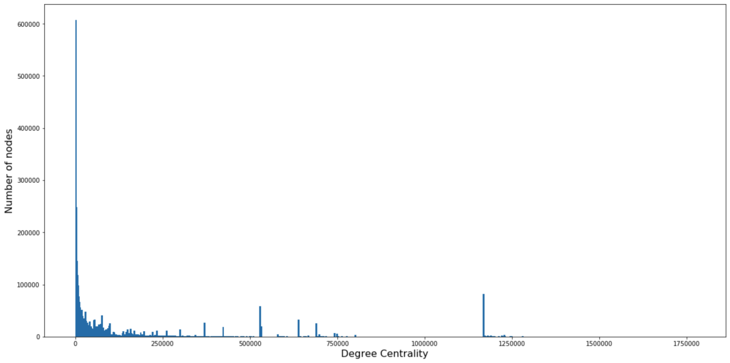

Fig. 2. Distribution of the degree centrality for the nodes of $\mathcal{P}$

As we can see from this figure, degree centrality follows a power law; this result is aligned with the theory underlying this form of centrality [21]. The maximum value of degree centrality is 1,820,412, while the minimum one is 0.

In order to check a possible existence of assortativity in Reddit, we sorted the authors according to their degree centrality, in a descending order. We then partitioned the resulting list into intervals. Specifically, we took intervals with equal width[5] $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_{40}\}$, each made up of 312,500 authors. As a consequence, $\mathcal{I}_k$, $1 \leq k \leq 39$, contained all the authors comprised in the interval $(312,500 \cdot (k-1), 312,500 \cdot k]$, open at left and closed at right of the

[5] Eventually, the last interval had a width a bit lower than the other ones.

sorted list. The interval $\mathcal{I}_{40}$ contained all the authors comprised in the interval $(12,187,500\ ,\ 12,464,188]$.

First of all, we considered the first interval, i.e. $\mathcal{I}_1$. For each interval $\mathcal{I}_k$, $1 \leq k \leq 40$, we determined how many authors of $\mathcal{I}_1$ are connected through an arc to at least one author of $\mathcal{I}_k$. The results obtained are reported in Figure 3. Then, we determined the percentage of the authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_1$. The results obtained are reported in Figure 4.

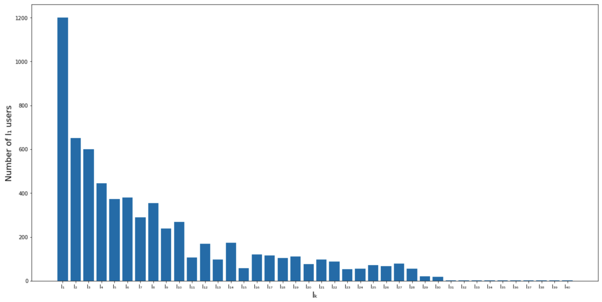

Fig. 3. Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$

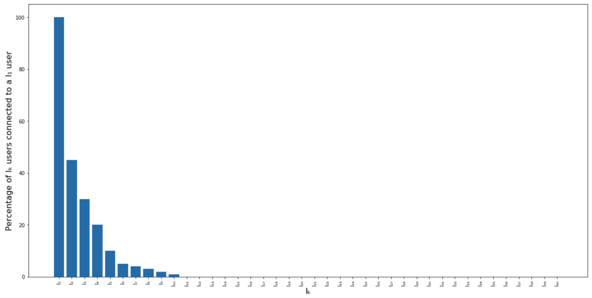

Fig. 4. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$

The analysis of Figures 3 and 4 clearly shows a strict correlation, i.e. a sort of backbone, between the authors with the highest degree centrality.

We compared our findings with the ones obtained through a null model, in order to verify the statistical significance of our results in an unbiasedly random scenario. In particular, we shuffled all the arcs between the nodes of $\mathcal{P}$ (that, in our case, represent co-postings), in order to build the null model. In this way, we left unchanged all the features of $\mathcal{P}$, excluding the distribution of co-posting relationships, which was unbiasedly random in the null model. Next, we repeated all the previous analyses on the null model. Figures 5 and 6 show the obtained results. The comparison between these two last figures and Figures 3 and 4 highlights the similarity of the distributions represented therein. Many of the intervals that obtained the highest values in Figures 3 and 4 continue to reach the highest values in Figures 5 and 6. However, in the null model, the values are much smaller. So, we can conclude that the behaviors observed are not random, but intrinsic to Reddit.

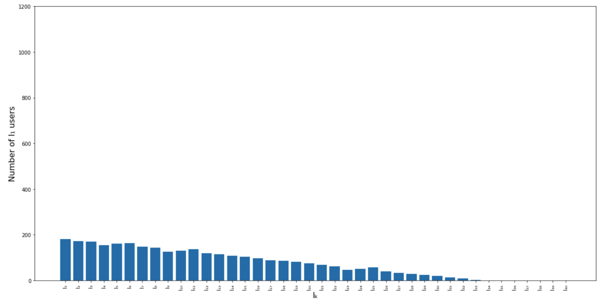

Fig. 5. Number of authors of $\mathcal{I}_1$ connected to at least one author of $\mathcal{I}_k$ in the null model

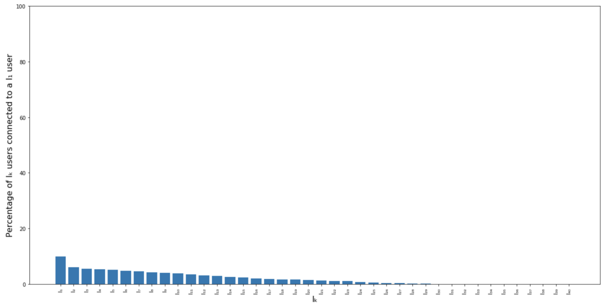

Fig. 6. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_1$ in the null model

However, this is not enough to prove the existence of a degree assortativity for co-posters in Reddit. Indeed, we must check if this trend is also verified for authors with an intermediate degree centrality and for ones with a low degree centrality.

For this reason, we have to redo the previous tasks done for $\mathcal{I}_1$ for all intervals. Due to space constraints, we consider only the intervals $\mathcal{I}_{20}$, as the representative of the intermediate degree centrality author intervals, and the interval $\mathcal{I}_{39}$, as the representative of the low degree centrality author intervals[6].

Figure 7 shows the number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$. Figure 8 shows the percentage of the authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{20}$. These figures clearly highlight the existence of a correlation between the authors with an intermediate degree centrality.

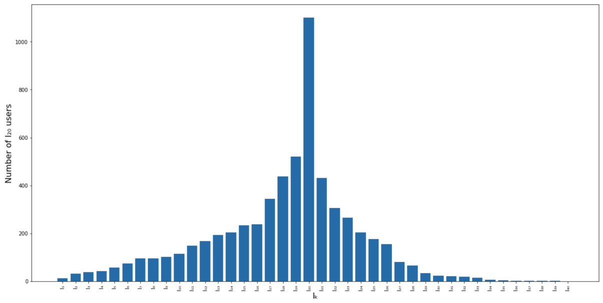

Fig. 7. Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$

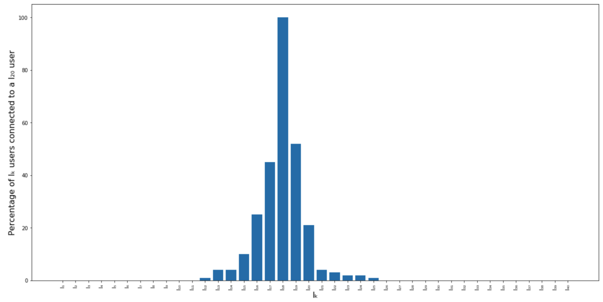

Fig. 8. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$

Also here, we compared the results with the null model. Figures 9 and 10 present the results obtained. Comparing them with Figures 7 and 8, we note that, again, the behaviors observed are not random, but they are a feature of Reddit.

Finally, Figure 11 reports the number of the authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$, while Figure 12 shows the percentage of the authors of $\mathcal{I}_k$ connected with at least one author of $\mathcal{I}_{39}$. Here too, a strict correlation exists

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

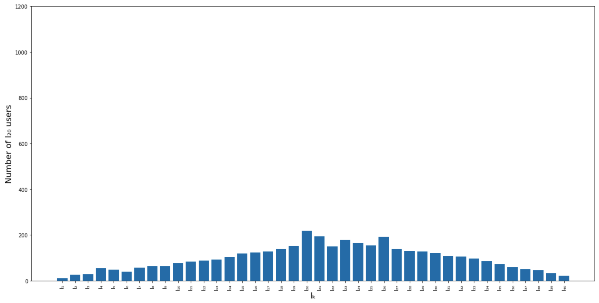

Fig. 9. Number of authors of $\mathcal{I}_{20}$ connected to at least one author of $\mathcal{I}_k$ in the null model

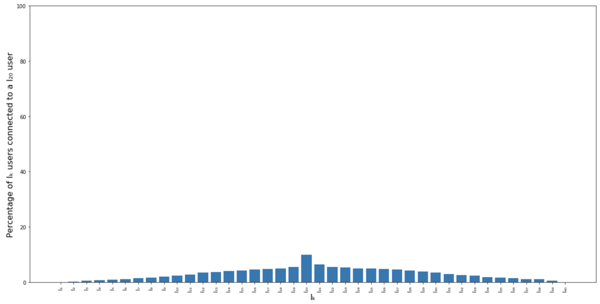

Fig. 10. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{20}$ in the null model

between the authors with a low degree centrality. We compared these results with the ones obtained through the null model reported in Figures 13 and 14. Again, this comparison confirms that the behaviors observed is a property intrinsic to Reddit. The existence of a backbone among the authors with a high (resp., intermediate, low) degree centrality helps us to conclude that, actually, Reddit is assortative with respect to degree centrality, as far as the co-posting relationship is concerned.

This important finding can be explained through the concept of karma and the posting rules existing in Reddit. Indeed, each user has associated a karma, i.e. a score taking her past "reputation" into account. Users with high karma are generally very active and often submit high quality contents, appreciated by others. So, they likely have a high degree centrality. In other words, we can recognize a direct correlation between karma and degree centrality for authors. Reddit's posting rules state that each subreddit has associated a minimum threshold of karma that authors must have to post on it [14, 15, 1]. This threshold is dynamic and changes over time. When it is low, all users can post on that subreddit. When it becomes moderate, users with low karma (and maybe low degree centrality) cannot post on it. When it becomes high, only users with high karma (and maybe high degree centrality) can post on it. In this way we can segment users into groups with homogeneous degree centrality.

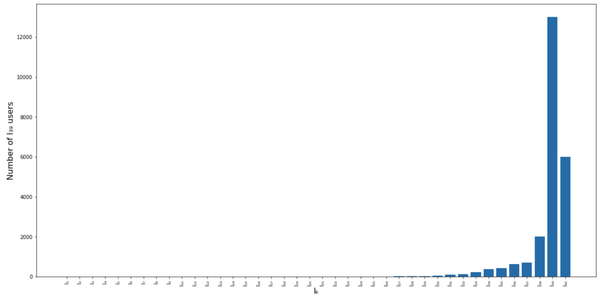

Fig. 11. Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$

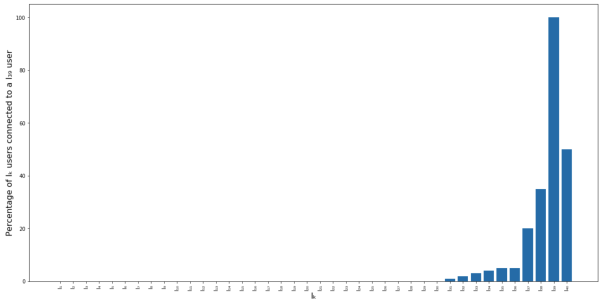

Fig. 12. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$

5 Conclusion

In this paper, we have presented several investigations that we performed to evaluate assortativity in Reddit. First, we have built a dataset comprising all posts found in Reddit from January 1^{st}, 2019 to September 1^{st}, 2019. Then, we have constructed a co-posting network that represented the reference structure on which performing our analyses. Afterwards, we have carried out several investigations on both the co-posting network and a corresponding null model. Finally, we have compared the results obtained and we have found that Reddit is assortative with respect to degree centrality, as far as the co-posting relationship is concerned.

In the future, we plan to extend this work in several directions. For example, we plan to evaluate the possible existence of other forms of assortativity or

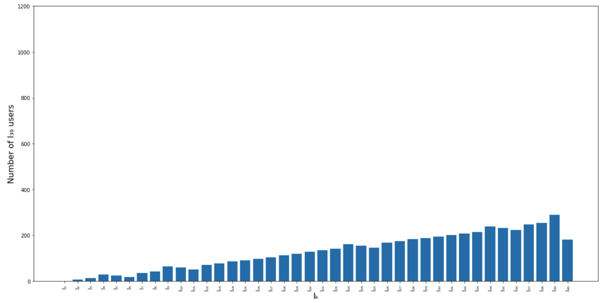

Fig. 13. Number of authors of $\mathcal{I}_{39}$ connected to at least one author of $\mathcal{I}_k$ in the null model

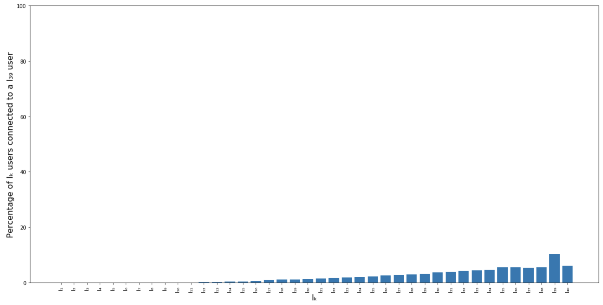

Fig. 14. Percentage of the authors of $\mathcal{I}_k$ connected to at least one author of $\mathcal{I}_{39}$ in the null model

disassortativity in Reddit. They could involve, for instance, centrality measures, other than degree centrality, or user activities, other than posting. In addition, we plan to investigate other issues analyzed in other social platforms and not yet investigated in Reddit.

References

1. K.E. Anderson. Ask me anything: what is Reddit? 2015. Emerald.
2. A. Arcagni, R. Grassi, S. Stefani, and A. Torriero. Higher order assortativity in complex networks. *European Journal of Operational Research*, 262(2):708–719, 2017. Elsevier.
3. F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proc. of the International Conference on Research and Development in Information Retrieval (SIGIR '09)*, pages 620–627, Boston, MA, USA, 2009. ACM.
4. J. Bollen, B. Gonçalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, 2011. MIT Press.
5. F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Internetworking assortativity in Facebook. In *Proc. of the International Conference on Social Computing and*

its Applications (SCA 2013), pages 335–341, Karlsruhe, Germany, 2013. IEEE Computer Society.

6. M. Catanzaro, G. Caldarelli, and L. Pietronero. Assortative model for social networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 70(3):037101–037104, 2004. The American Physical Society.
7. M. Catanzaro, G. Caldarelli, and L. Pietronero. Social network growth with assortative mixing. *Physica A: Statistical Mechanics and its Applications*, 338(1):119–124, 2004. Elsevier.
8. K.I. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Physical Review E*, 67(1):017101, 2003. APS.
9. H. B. Hu and X. F. Wang. Evolution of a large online social network. *Physics Letters A*, 373(12):1105–1110, 2009. Elsevier.
10. H.B. Hu and X.F. Wang. Disassortative mixing in online social networks. *EPL (Europhysics Letters)*, 86(1):18003, 2009. IOP Publishing.
11. S. Johnson, J.J. Torres, J. Marro, and M.A. Munoz. Entropic origin of disassortativity in complex networks. *Physical Review Letters*, 104(10):108702, 2010. APS.
12. H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. of the International Conference on World Wide Web (WWW'10)*, pages 591–600, Raleigh, NC, USA, 2010. ACM.
13. 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.
14. J. Meese. It belongs to the Internet: Animal images, attribution norms and the politics of amateur media production. *M/C Journal*, 17(2):1–3, 2014. M/C.
15. D. Morrison and C. Hayes. Here, have an upvote: Communication behaviour and karma on Reddit. *Informatik*, pages 2258–2268, 2013. Gesellschaft für Informatik eV.
16. D. Mulders, C. de Bodt, J. Bjelland, A. Pentland, M. Verleysen, and Y.-A. de Montjoye. Inference of node attributes from social network assortativity. *Neural Computing and Applications*, pages 1–21, 2019. Springer Nature Switzerland AG.
17. M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002. APS.
18. M.E.J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003. APS.
19. R. Noldus and P. Van Mieghem. Assortativity in complex networks. *Journal of Complex Networks*, 3(4):507–542, 2015. Oxford University Press.
20. D. Shizuka and D.R. Farine. Measuring the robustness of network community structure using assortativity. *Animal Behaviour*, 112:237–246, 2016. Elsevier.
21. M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web.* 2011. O'Reilly Media, Inc.
22. M. Vaanunu and C. Avin. Homophily and nationality assortativity among the most cited researchers' social network. In *Proc. of 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 584–586, Barcelona, Spain, 2018. IEEE Computer Society.
23. C. Wilson, B. Boe, A. Sala, K.P.N Puttaswamy, and B.Y. Zhao. User interactions in social networks and their implications. In *Proc. of the ACM European Conference on Computer systems (EuroSys'09)*, pages 205–218, Nuremberg, Germany, 2009. ACM.