Proceedings of the 2nd International Workshop on Social Influence Analysis (SocInf 2016)
July 9th, 2016 - New York, USA

Pinpointing Influence in Pinterest

Panagiotis Liakos[1], Katia Papakonstantinopoulou[1], Michael Sioutis[2], Konstantinos Tsakalozos[3], and Alex Delis[1]

[1] University of Athens, GR15703, Athens, Greece
{p.liakos,katia,ad}@di.uoa.gr
[2] Université d'Artois, CRIL UMR 8188, Lens, France
sioutis@cril.fr
[3] Canonical Group Ltd., London, UK
konstantinos.tsakalozos@canonical.com

Abstract. The success of most applications that run on top of a social network infrastructure is due to the social ties among their users; the users can get informed about the activity of their friends and acquaintances, and, hence, new ideas, habits, and products have the opportunity to gain popularity. Therefore, understanding the influence dynamics on social networks provides us with insights that are useful in designing efficient social network applications. In this work we focus on Pinterest, a social network that is often used to promote commercial products, and investigate the influence mechanisms in it. We examine the user indegree and PageRank as potential estimators of the number of repins and likes the user may receive. We observe that, although both measures are weakly associated with user influence in Pinterest, PageRank is much more powerful than indegree in revealing how much influential a user is.

Keywords: Social Influence; PageRank; Pinterest.

1 Introduction

Over the past two decades the rise of social networking sites has shaped new forms of interaction. Social media such as Facebook, Twitter, and Pinterest are drawing millions of individuals, who now depend on them to keep up with friends, follow breaking news, share their interests, and discover products or events. The popularity of online social networks and the diversity of the activities of their users make them ideal candidates for studying influence patterns. In particular, we want to find out whether certain individuals in a social network have the power to affect their social contacts and are in the position to convince them to buy a product or adopt a political idea.

The diffusion of influence has received significant attention from the fields of sociology, advertising, and political science. Early studies argue for the existence

Copyright © 2016 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.

of a small minority of *opinion leaders*, that are able to persuade the majority of the society to mimic their behavior [6]. More recent studies, however, do not outright accept this hypothesis. In [4] the authors consider the probability of a customer buying some product as a function of the influence of other customers, as well as her intrinsic desire for the product. In [16] it is suggested that although an *influential* minority may exist, it is seldom responsible for spreading ideas; it is a critical mass of easily influenced individuals who trigger *chain-reactions* of influence. Identifying *influential* individuals allows for cost-effective viral marketing techniques to increase brand awareness or even sway the public opinion.

The massive activity of online social network users enables us to collect rich large-scale data and assess the presence and potential of *influential* users. A related recent effort [2] investigates influence in the Twitter social network. Topological measures, such as the number of one's followers (indegree), are reported to fail in exposing user influence. This is attributed to the users' tendency to follow others for courtesy and is often referred to as the *million follower fallacy.*[4] However, user activity varies across different social networking sites. Thus, it is important to examine whether indegree reveals user influence in other social media. Moreover, considering more refined measures of importance implied by the network topology may lead to more effective identification of influential users.

We focus on Pinterest, an image-based online social network in which users can post (or *pin*) content they find interesting and browse the content of others in their feed, where they can re-post (or *repin*), comment, or endorse (*like*) other pins. Consequently, users focus on the *curation* and *discovery* of existing content and many businesses opt to join Pinterest in an effort to promote their products. We perform an in-depth empirical analysis to measure the extent to which the indegree of Pinterest users is associated with the number of repins and likes they receive. In addition to this, we propose the use of PageRank [8] to better estimate one's potential to influence others. Indegree considers the expressed opinions of all users as equally important. In contrast, PageRank distinguishes *important* users based on the network topology. Our intuition is that the more authoritative a user is, the more influential she is expected to be. Therefore, the use of PageRank to identify influential users sounds promising. In summary, we make the following contributions:

- We examine the association between indegree and user influence in a very popular social network. We find that, similarly to Twitter [2], the indegree of Pinterest users reveals very little regarding the number of repins and likes a user may receive.
- We propose the use of PageRank for identifying influential individuals and investigate its effectiveness. Our findings suggest that, even though the association between PageRank and user influence appears limited, PageRank is much more powerful than indegree in revealing the influence of a user.

The rest of this paper is organized as follows. In Section 2 we review the related work. Section 3 provides an insight into Pinterest's fundamental elements

[4] `http://bit.ly/1WaTAeK`

and outlines our approach. In Section 4 we present an empirical analysis of Pinterest activity to study the extent of influence in the network. Finally, Section 5 concludes this work.

2 Related work

The following fundamental algorithmic problem for social network processes dealing with influence was posed by Domingos and Richardson in [15,4]: *if we can try to convince a subset of individuals to adopt a new product or innovation, and the goal is to trigger a large cascade of further adoptions, which set of individuals should we target?* This problem is studied in [7] in the context of the most popular models in social network analysis. The authors show that the aforementioned problem is essentially the optimization problem of selecting the most influential nodes, which is $\mathcal{NP}$-hard, and provide the first provable approximation guarantees for efficient algorithms. The results of this study gave rise to a number of works that try to identify influential users by employing heuristics that are based on real network data.

In [2], Cha et al. study user influence on Twitter; based on a Twitter crawl, they compare three measures of influence, namely, indegree, retweets, and mentions, and observe that users who have high indegree, although popular, are not necessarily influential in terms of causing retweets or mentions. Their findings suggest that topological measures alone, probably do not reveal much about how influential a user is. However, they leave room for examining more refined measures of influence, which do not consider all links in a network to be transferring the same authority. Weng et al. [17] focus on the problem of identifying influential users of micro-blogging services, with Twitter being their service of choice for studying such influence. In a dataset that they prepared for their study, they observe that 72.4% of the users in Twitter follow more than 80% of their followers, and 80.5% of the users have 80% of users they are following follow them back. Their study reveals that the presence of such reciprocity can be explained by the phenomenon of *homophily* [11], i.e., the tendency of individuals to associate and bond with similar others. Based on this finding, they propose TwitterRank, an extension of the PageRank algorithm, to measure the influence of users in Twitter. Identifying patterns of influence also serves the purpose of evaluating sociological models. In another line of research, in [10] the authors study real user activity in the Digg social network, based on sociological studies and the game theory framework, and identify a model that closely describes the observed influence. In particular, they model user activity as an opinion formation game in which each user is influenced by her social contacts and find that the Nash equilibria of the game are nice illustrations of how users real behave.

Gilbert et al. [5] focus on identifying certain properties of Pinterest. In particular, they try to answer the questions of what drives activity on Pinterest, what role gender plays in the site's social connections, and, finally, what distinguishes Pinterest from other networks, such as Twitter. With respect to those questions, they conclude that being female implies more repins but fewer followers, and

that four verbs set Pinterest apart from Twitter, namely, *use*, *look*, *want*, and *need*. Similarly to [5], properties of Pinterest are also highlighted in [3]. Chang et al. study a fundamental issue for social curation sites where people collect, organize, and share pictures of items, with a focus on Pinterest, as it is the most prominent example of such a site. In particular, they study the issue of what patterns of activity attract attention. They organize their study around two key factors, namely, the extent to which users specialize in particular topics, and homophily among users. Further, they also consider the existence of differences between female and male users. Their study reveals that women and men differ in the types of content they collect and the degree to which they specialize. These findings suggest strategies both for users (e.g., to attract an audience) and for maintainers (e.g., to explore content recommendation methods) of social curation sites. Mittal et al. [12] characterize Pinterest on the basis of large scale crawls of 3.3 million user profiles, and 58.8 million pins. In particular, they explore various attributes of users, pins, boards, pin sources, and user locations in detail, and perform topical analysis of user generated textual content. This characterization revealed the most prominent topics among users and pins, such as design, fashion, photography, food, travel, music, and art. Moreover the top image sources and geographical distribution of users on Pinterest were obtained.

Ottoni et al. [13] make a first attempt towards a more complete understanding of user behavior across multiple online social networks, such as Twitter and Pinterest. They collect a sample of over $30,000$ users that have accounts on both Twitter and Pinterest, by crawling their profile information and activity on a daily basis for a period of almost three months. Then, they develop a novel methodology for comparing activity across these two sites, that builds on the Labeled Latent Dirichlet Allocation model (L-LDA) [14], a supervised topic model for credit attribution in multi-labeled corpora. The authors find that the global patterns of use across the two sites differ significantly, and that users tend to post items to Pinterest before posting them on Twitter. These findings can assist in the understanding of user behavior on individual sites, as well as the dynamics of sharing across the social web.

3 Tracing Influence in Pinterest

In this section we discuss the Pinterest social network; its immense popularity, the actions allowed to its users, and its most prominent categories indicate that Pinterest is extremely interesting business-wise. Moreover, we detail the measures of influence that we will employ in our empirical analysis.

3.1 Understanding Pinterest

Pinterest is an image sharing social network that was founded in 2010 and was the fastest site to surpass the 10,000,000 monthly active users milestone.[5] During

[5] `http://techcrunch.com/2012/02/07/pinterest-monthly-uniques/`

the year 2015, Pinterest was also reported to have broken the mark of 100,000,000 monthly active users.[6] Notably, the vast majority of these users are female,[7] which makes Pinterest particularly interesting to study.

Pinterest users are able to create a profile along with one or more *boards* where they may *pin* images or other media content, for example videos. They are also able to follow other users or specific boards in order to receive personalized updates in their feed. Pins can be *liked* or *repinned* (shared) by other users, as it is also the case with the content of other social networks such as Facebook or Twitter.

Pinterest boards are organized into a broad range of categories, typical ones being *Food & Drink*, *Weddings*, and *Home Decor*. These categories are indicative of the users' habit of creating digital shopping lists of products they are interested in buying. This tendency has attracted significant commercial attention, as many businesses invest in creating compelling boards to increase their revenue.[8]

3.2 Defining influence in Pinterest

A user's behavior in a social network is often affected by others. However, the task of singling out those individuals who actually cause social influence in an online social network is not trivial. In this paper, we focus on three fundamental actions performed by Pinterest users, namely, *follow*, *repin*, and *like*. We consider that the number of followers a user has, as well as the number of repins and likes she receives, are indicative of her potential to affect others in the network and can be used as a measure of influence.

The use of indegree as a quality measure may be flawed due to its local nature. To alleviate this issue we additionally consider the use of PageRank values. In information networks, with the World Wide Web being the most popular example, the authority of each node is estimated using the PageRank algorithm [8], a quality metric introduced by the creators of the Google search engine that is based on the network's link structure. According to the definition of PageRank, a node's authority is distributed uniformly to the nodes it has a link to. As a consequence, a node is important if many important nodes link to it. The use of PageRank, however, is not restricted to search engine optimization. PageRank is used in recommendation systems, in ranking tweets in Twitter, and even in suggesting friends in online social networks.

The measures of influence that can be drawn from the Pinterest network and are examined in this work are the following ones:

- **Indegree influence**: the number of followers of a user directly indicates the *size of the audience* of that user.
- **PageRank influence**: the PageRank of a user indicates the *strength of her influence* on her followers.

[6] `http://mobile.nytimes.com/blogs/bits/2015/09/17/pinterest-crosses-user-milestone-of-100-million/`
[7] `http://www.omnicoreagency.com/pinterest-statistics/`
[8] `https://business.pinterest.com/en/success-stories`

– **Like influence**: the number of likes containing one's name indicates the ability of that user to generate *popular* content.
– **Repin influence**: the number of repins containing one's name indicates the ability of that user to generate content with *pass-along value*.

Notice that the first two measures depend solely on the *topology* of the social network, while the other two take into account the *user activity* on the network as well.

4 Empirical Analysis

This section details the analysis we performed on the Pinterest network and our empirical observations concerning the influence patterns exhibited. We first introduce the dataset we used for our analysis and the experimental setup needed for the PageRank algorithm. Then, we proceed with answering the following questions:

– Do the indegree of Pinterest users and the repins and likes they receive follow long-tailed distributions?
– Is there significant overlap among the top users based on indegree or PageRank and the top users based on the number of repins or likes they received?
– How strong is the pairwise rank correlation between indegree, repins and likes in Pinterest?
– Is the rank correlation improved when we consider the use of PageRank instead of indegree?

4.1 Experimental Setting

Our empirical analysis is based on the Pinterest dataset described in [19,18]. The activity of the users was collected from January 3rd 2013 to January 21st 2013, while the social graph was crawled during April 2013. We focused on the subset of this dataset that contains only links that are common to both Pinterest and Facebook. Our experimental setup comprises a social graph of 36,198,633 users and 983,520,986 social ties. Regarding the activity of these users, we analyzed a total of 18,957,340 repins, and 9,066,973 likes on pins. These pins were created by a total of 1,253,189 users.

Calculating the PageRank values of large-scale graphs calls for the use of Pregel-like graph processing systems like Apache Giraph.[9] Using Juju[10], we deployed an Apache Hadoop 2.7.1 cluster on a Dell PowerEdge R630 server with an Intel®Xeon® E5-2630 v3, 2.40 GHz processor, and 256 GB of RAM. Then, we built Apache Giraph in the client node of our cluster and submitted a Giraph job that calculated the PageRank values of the users in our social graph by performing 80 iterations. This allowed us to fully utilize our infrastructure and obtain the desired results in less than 3 hours.

[9] `http://giraph.apache.org/`
[10] `https://jujucharms.com/big-data`

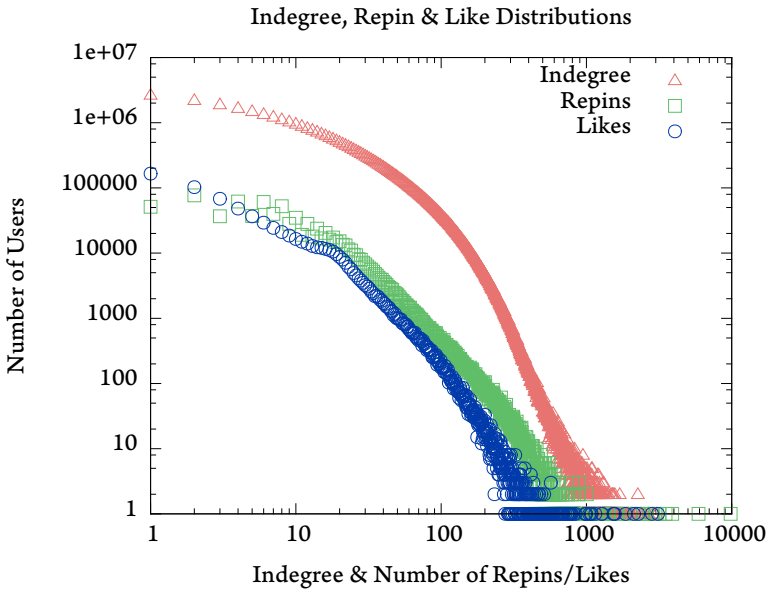


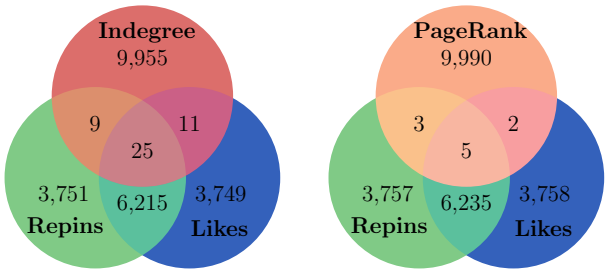
Fig. 1. Distribution of indegree and received repins/likes. Most of the Pinterest activity is centered around a small number of users.

4.2 Distribution of indegree and received repins/likes

In Figure 1 we illustrate the indegree distribution of the social graph of Pinterest as well as the distribution of influence based on repins and likes. We use a logarithmic scale to highlight the fact that the users' indegree, as well as the users' number of received repins or likes, varies by several orders of magnitude. In particular, we observe that there are a few users with more than 1,000 followers, repins or likes, whereas the majority of users have no more than one follower and have received no more than one repin or like. Therefore, most of the activity on the network is centered around a small minority of users. Moreover, we observe that users tend to receive slightly more repins than likes.

4.3 Overlap of top-ranked users

We continue our analysis of the Pinterest dataset by examining the overlap of the top-ranked users for each measure. In particular, the Venn diagram of Figure 2(a) depicts the overlap of the top-10,000 users according to their indegree, the number of received repins, and the number of received likes. We observe that the overlap of indegree with both of the other measures is marginal. In contrast, there is significant overlap between the users whose pins received many repins and those whose pins received many likes.



(a) (b)

Fig. 2. *Venn* diagrams of the top-10,000 users across indegree/repins/likes (a) and PageRank/repins/likes (b). The overlap of indegree and PageRank with repins and likes is marginal. In contrast, repins and likes are strongly correlated.

Moreover, we present the corresponding Venn diagram for the top-10,000 users according to their PageRank in Figure 2(b). We see that the overlap of PageRank with repins and likes is even more insignificant in this case.

These observations hint that there is very weak correlation of the indegree or PageRank of users with the frequency they receive repins and likes. In addition to this, we showed that users who receive many repins tend to receive many likes as well.

4.4 Comparing Influence Measures

For each of the $36,198,633$ users of our dataset we calculated the value of indegree and PageRank and the number of repins and likes, and adopted the methodology of [2] to quantify the association between them. In particular, we characterized our comparison on the relative order of users' ranks as a measure of difference. We assigned the rank of 1 to the most influential user, and increased the rank as we proceeded to less influential users. Identical values were each assigned fractional ranks equal to the average of their positions in the ascending order of the values [1]. Then, we used the *Spearman's rank correlation coefficient* [9], a non-parametric measure that is used to assess the degree of association between two variables, to examine whether two ranked variables covary. This measure is especially useful as it does not make any assumptions regarding the distribution of the data.

The Spearman's rank correlation coefficient is defined by the following formula:[11]

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = rg(X_i) - rg(Y_i)$ is the difference between the two ranks of user i, and n is the total number of users.

[11] In our case all ranks are distinct integers, hence, we present the simplified formula.

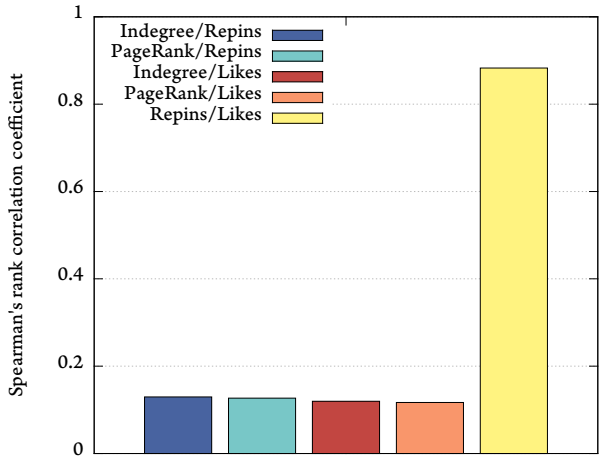


Fig. 3. Spearman's rank correlation coefficients for all the users of our dataset.

Figure 3 depicts the correlation between all pairs of measures examined in this work, for all the users of our dataset. We observe that both the indegree and the PageRank measures exhibit very weak correlation with the repin and like measures. A weak correlation between topological measures and user influence has already been reported in [2]. However, our results suggest that the association is much weaker on Pinterest than on Twitter.

In [2], a preprocessing step is applied to the dataset that removes all users with a limited number of tweets since the creation of their account. We investigated the impact of an equivalent preprocessing step to our dataset by examining only those users who have created at least 15 pins, i.e., a total of $20,011,513$ users. This resulted to a weaker correlation of indegree with both repins (0.1068) and likes (0.0997). Therefore, we attribute the deviation of our results from the findings of [2] to the different patterns of use across Pinterest and Twitter.

Another important finding illustrated in Figure 3, is the extremely strong correlation between repins and likes. This is indicative of the different focus of Pinterest compared with Twitter; in Twitter the action of a retweet focuses on the content of a tweet and the action of a mention on the user, whereas in Pinterest both the action of a repin and the action of a like focus on the *content* of a pin. As such, the correlation between the two corresponding measures in Twitter is reported to be moderate (0.58) [2], whereas the two measures in Pinterest are very strongly associated.

We have already discussed that most users have very few followers, as clearly shown in Figure 1. Consequently, most users are tied when ranked according to their indegree. This may lead to a fabricated result regarding the correlation of measures, as users with low indegree most likely also have limited repins and

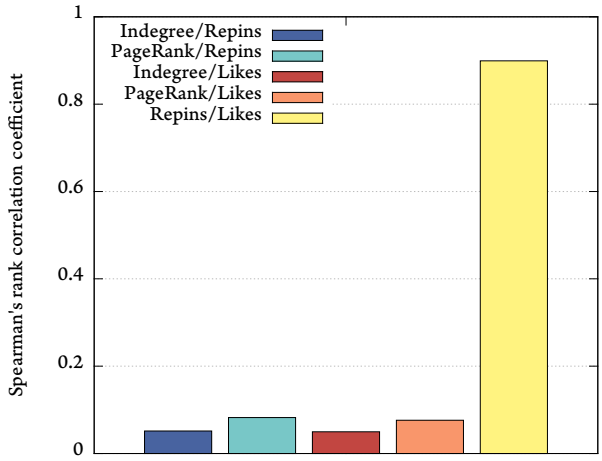


Fig. 4. Spearman's rank correlation coefficients for the top 10th percentile of users according to their indegree.

likes. To alleviate this issue, we considered users ranked in the top 10th and 1st percentiles, respectively, according to their indegree, as suggested in [2].

We observe in Figures 4 and 5 that the correlation of indegree with repins or likes is indeed even weaker for the top 10th and 1st percentiles of users respectively. However, what stands out is that the correlation of PageRank with repins and likes is stronger than that of indegree for these two cases. In particular, in Figure 4 we observe that for the top 10th percentile of users according to their indegree, the association of PageRank with user influence is about twice as strong as that of indegree. The results for the 1st percentile, depicted in Figure 5, are even more impressive, as the statistical dependence with PageRank instead of indegree is more than twice as strong.

This verifies our intuition that using PageRank instead of indegree allows for capturing the importance of users more accurately. Given that the focus of both repins and likes is on the content of the pin and not on the user who posted it, we expect that PageRank will perform even better against actions that are *centered toward individuals*, e.g., Twitter's mentions.

5 Conclusion and Future Work

The study of influence patterns on social networks is essential for the design of a successful advertising strategy. The recent unprecedented growth in the number of online social networking services allows us to empirically validate relevant theories and uncover opportunities for viral marketing techniques. We performed an analysis of a large amount of activity on the Pinterest network. The case of Pinterest is particularly interesting as it is embraced by an ever-increasing

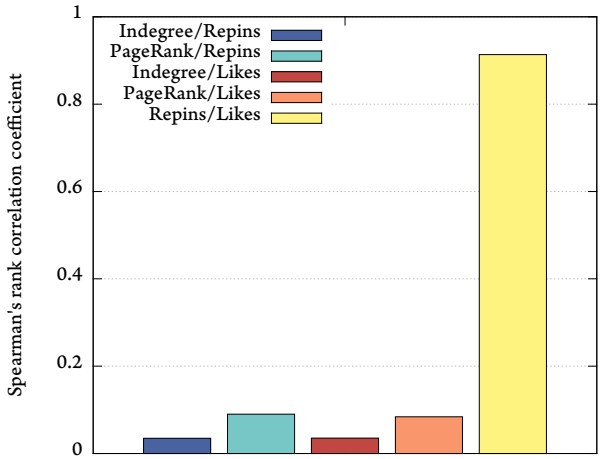


Fig. 5. Spearman's rank correlation coefficients for the top 1st percentile of users according to their indegree.

number of businesses eager to promote their products to a wide audience through captivating pin-boards.

We examined the association of the indegree of Pinterest users with the number of times their pins are repinned or liked. Our results reveal that there is very little correlation between the ranking of users based on their indegree and their ranking based on the number of either repins or likes they receive. We attribute this result to the *million follower fallacy* that is evident in this network, and the fact that the focus of repins and likes is more on the *content* of pins rather than on the user who posted them. Furthermore, we proposed the use of PageRank instead of indegree for the identification of *influential* users. As PageRank is a measure that captures how authoritative a user is in a network, we expect that a ranking of users based on their PageRank value will provide us with a more accurate view of their potential to influence their social contacts. Indeed, even though correlation with the ranking of users based on repins or likes received is still limited, PageRank performed much better than indegree.

We will further investigate the influence patterns of the Pinterest network with regard to the variance exhibited across different topics. In particular, we will consider the categories of pins when ranking users based on the repins and likes received. The overlap of influential users across different Pinterest categories will indicate how many are able to spread information over a variety of topics. Moreover, we will examine whether PageRank proves to be even more successful when quantifying its association with an action targeted towards the user instead of the content (both repins and likes are related to the content of a pin). We expect that for online social networks that focus more on original content, like Twitter or Instagram, and with actions targeted towards the user,

such as *mentions*, PageRank's performance will be significantly superior to that of the indegree.

References

1. W. Buck. Tests of significance for point-biserial rank correlation coefficients in the presence of ties. *Biometrical Journal*, 22(2):153–158, 1980.
2. M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*, 2010.
3. S. Chang, V. Kumar, E. Gilbert, and L. G. Terveen. Specialization, Homophily, and Gender in a Social Curation Site: Findings from Pinterest. In *CSCW*, 2014.
4. P. M. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, 2001.
5. E. Gilbert, S. Bakhshi, S. Chang, and L. G. Terveen. "I Need to Try This!": A Statistical Overview of Pinterest. In *CHI*, 2013.
6. E. Katz and P. F. Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications.* Transaction Publishers, 1955.
7. D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.
8. P. Lawrence, B. Sergey, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.
9. E. L. Lehmann and H. J. D'Abrera. *Nonparametrics: statistical methods based on ranks.* Springer-Verlag New York, 2006.
10. P. Liakos and K. Papakonstantinopoulou. On the Impact of Social Cost in Opinion Dynamics. In *ICWSM*, 2016.
11. M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
12. S. Mittal, N. Gupta, P. Dewan, and P. Kumaraguru. Pinned it! A Large Scale Study of the Pinterest Network. In *IKDD*, 2014.
13. R. Ottoni, D. B. L. Casas, J. P. Pesce, W. M. Jr., C. Wilson, A. Mislove, and V. Almeida. Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks. In *ICWSM*, 2014.
14. D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *EMNLP*, 2009.
15. M. Richardson and P. Domingos. Mining Knowledge-sharing Sites for Viral Marketing. In *KDD*, 2002.
16. D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.
17. J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: Finding Topic-Sensitive Influential Twitterers. In *WSDM*, 2010.
18. C. Zhong, M. Salehi, S. Shah, M. Cobzarenco, N. Sastry, and M. Cha. Social Bootstrapping: How Pinterest and Last.fm Social Communities Benefit by Borrowing Links from Facebook. In *WWW*, 2014.
19. C. Zhong, S. Shah, K. Sundaravadivelan, and N. Sastry. Sharing the Loves: Understanding the How and Why of Online Content Curation. In *ICWSM*, 2013.