Overcoming Spammers in Twitter – A Tale of Five Algorithms[1]

Daniel Gayo-Avello and David J. Brenes

Dept. of Computer Science, University of Oviedo, Calvo Sotelo s/n 33007 Oviedo (SPAIN),
Simplelógica, Fray Ceferino 18 1º Derecha 33001 Oviedo (SPAIN)
dani@uniovi.es,
research@davidjbrenes.info

Abstract. Micro-blogging services such as Twitter can develop into valuable sources of up-to-date information provided the spam problem is overcome. Thus, separating the most relevant users from the spammers is a highly pertinent question for which graph centrality methods can provide an answer. In this paper we examine the vulnerability of five different algorithms to linking malpractice in Twitter and propose a first step towards "desensitizing" them against such abusive behavior.

Keywords: Social networks, Twitter, spamming, graph centrality, prestige.

1 Introduction

Twitter is a service which allows users to publish short text messages (tweets) which are shown to other users following the author of the message. In case the author is not protecting his tweets, they appear in the so-called public timeline and are served as search results in response to user submitted queries. Thus, Twitter can be a source of valuable real-time information and, in fact, several major search engines are including tweets as search results.

Given that tweets are published by individual users, ranking them to find the most relevant information is a crucial matter. In fact, at the moment of this writing, Google seems to be already applying the PageRank method to rank Twitter users to that end [1]. Nevertheless, the behavior of different graph centrality methods and their vulnerabilities when confronted with the Twitter user graph, in general, and Twitter spammers in particular, are still little-known. Hence, the study described in this paper aims to shed some light on this particular issue besides providing some recommendations for future research in the area.

[1] A longer version of this study can be found at http://arxiv.org/abs/1004.0816

2 Literature Review

A social network is any interconnected system whose connections are produce of social interactions among persons or groups. Such networks can be modeled as graphs and, thus, graph theory has become inextricably related to social network analysis with a long history of research. However, for the purpose of this paper it should be enough to briefly sketch the concepts of centrality and prestige.

Both of them are commonly employed as proxy measures for the more subtle ones of importance, authority or relevance. Central actors within a social network are those which are very well connected to other actors and/or relatively close to them. There exist several measures of centrality which can be computed for both undirected and directed graphs. Prestige, in contrast, requires distinguishing inbound from outbound connections. Thus, prestige is only applicable to directed graphs which, in turn, are the most commonly used when analyzing social networks. As with centrality, there are several prestige measures such as indegree (the number of inbound connections), proximity prestige (related to the influence domain of an actor), and rank prestige, where the prestige of a node depends on the respective prestiges of the nodes linking to them. Nonetheless to say, rank prestige is mutually reinforcing and, hence, it requires a series of iterations over the whole network to be computed.

The later is the most commonly used prestige measure and there exist a number of well-known methods to compute one or another "flavor" of such a measure. In the following subsections we will briefly review the popular PageRank and HITS algorithms, in addition to lesser-known techniques such as NodeRanking, TunkRank, and TwitterRank. For the sake of brevity no equations are provided neither for PageRank nor HITS.

2.1 PageRank

PageRank [2] is one of the best known graph centrality methods. It aims to determine a numerical value for each document in the Web indicating the "relevance" of that given document. This value spreads from document to document following the hyperlinks; this way, heavily linked documents tend to have large PageRank values, and those documents receiving few links from highly relevant documents also tend to have large PageRank values. A notable property of the algorithm is that the global amount of PageRank within the graph does not change along the iterations but it is just distributed between the nodes. Thus, if the total amount of PageRank in the Web was arbitrarily fixed at 1 we could see the PageRank for a given document as a proxy for the probability of reaching that given document by following links at random (that's why PageRank is often described as a random surfer model).

2.2 Hyperlink-Induced Topic Search – HITS

HITS [3] is another algorithm to estimate the relevance of a document. The method assumes two different kinds of documents in the Web: authorities and hubs. An

authority is a heavily linked document while a hub is a document collecting links to several authorities. Web pages can exhibit both characteristics and, thus, every document has got two different scores: its authority score and its hub score. HITS was not aimed to be computed across the whole Web graph but, instead, within a query dependent subgraph; however, there is no impediment to apply it to a whole graph.

2.3 NodeRanking

NodeRanking [4] is another installment of the random surfer model. The main differences between NodeRanking and PageRank are two: (1) it is devised for weighted graphs, and (2) the damping parameter is not fixed for the whole graph but is computed for each node and depends on the outbound connections of the node.

The equations underlying this algorithm are shown below. $P_{jump}(p)$ is the equation driving the damping factor for each node: nodes with few outbound links have greater probability of being damped which should be interpreted as the random surfer getting bored because of the limited set of choices. $P_{choose}(p)$ is the probability of a page p to be chosen by the random surfer which, when ignoring weights in the edges, reduces to the same assumption in PageRank, i.e. a web surfer visiting a given page q would continue to any of the p pages linked from q with equal probability.

$$P_{jump}(p) = \frac{1}{1+|L(p)|}, P_{choose}(p) = \frac{1}{|L(q)|}, (q,p) \in E .$$

2.4 TunkRank

TunkRank [5] is one the first prestige ranking methods tailored to the particular circumstances of social networks. It lies on three assumptions: (1) each user has got a given influence which is a numerical estimator of the number of people who will read his tweets. (2) The attention a user pays to his followees is equally distributed. And (3), if X reads a tweet by Y he will retweet it with a constant probability p.

$$Influence(X) = \sum_{Y \in Followers(X)} \frac{1+p \cdot Influence(Y)}{|Following(Y)|} .$$

2.5 TwitterRank

TwitterRank [6] extends PageRank by taking into account the topical similarity between users in addition to link structure. In fact, TwitterRank is a topic-sensitive method to rank users separately for different topics. Additionally, the transition probability among users relies in both the topical similarity between users, and the number of tweets published not only by the followee, but by all the followees the follower is connected to.

These features make of TwitterRank a highly flexible method. However, we feel that it also makes it difficult to scale to the number of users and tweets that are published on a daily basis. Because of this, and for the sake of better comparison with the rest of graph centrality methods, we slightly modified TwitterRank.

The differences are the following: (1) instead of computing a different TwitterRank

value for each user and topic to be later aggregated, we aimed to compute just one TwitterRank value by user. (2) We also changed the topical similarity measure to compare users: instead of computing Jensen-Shannon Divergence between users' topic distributions we decided to apply the much more usual cosine similarity. And lastly, (3) we simplified the way to compute the damping parameter. The following equations provide a description of our implementation of TwitterRank.

TR(u) is the TwitterRank value for user u; γ is the probability of teleportation – constant for the whole graph, we used the commonly applied value of 0.15; $P(u_j,u_i)$ is the transition probability from user u_j to user u_i; $|\tau_i|$ is the number of tweets published by user u_i, and $|\tau|$ is the total number of tweets published by all the users. Lastly, $sim(u_j,u_i)$ is the cosine similarity between the tweets published by users u_i and u_j.

$$TR(u_i) = (1-\gamma)\sum_{u_j \in followers(u_i)} P(u_j,u_i)\cdot TR(u_j) + \gamma \cdot \frac{|\tau_i|}{|\tau|}$$

$$P(u_j,u_i) = \frac{|\tau_i|}{\sum_{a:u_j \boldsymbol{follows} u_a}|\tau_a|} sim(u_j,u_i)$$

3 Research Motivation

3.1 Research Questions

Social networks are increasingly gaining importance and the contents they provide can be exploited to provide up-to-date information (the so-called real-time Web). Because of the ease of publishing any content, anytime by anyone, it is ever more important to have a way to separate trustworthy sources from the untrustworthy ones.

Given the success of applying graph centrality algorithms to the Web, it seems appealing to do the same with social networks. Thus, the main research questions addressed in this study are the following: 1) How vulnerable to link spamming are common graph centrality algorithms when applied to user graphs from social networks? And 2), is it possible to "desensitize" such algorithms in a way that makes no more necessary to detect spammers but, instead, taking into account their presence and minimize their influence?

In addition to the aforementioned methods this author is proposing a variation of the PageRank method less sensitive to link abusing in social networks. Such a method relies on a de-weighting factor computed from the reciprocal links between users, and is described in the following subsection.

3.2 "Desensitizing" Prestige Ranking Methods against Link Spamming

The indegree is one of the simplest centrality measures. Translated to Twitter terms it is the total number of people following a user: the more followers a user has got the more valuable his tweets should be. Users such as Oprah Winfrey, CNN, or TIME are almost expected to have millions of followers: they are opinion-makers and mass media. One could even find reasons to explain the number of followers for Ashton

Kutcher or Britney Spears: they are celebrities. Which is harder to understand is how can spammers have far more followers than average users [7].

There is, however, a simple explanation for this phenomenon. Twitter has seen the emergence of its own etiquette and following back a new follower is considered "good manners". Once spammers took notice of this behavior, it was relatively easy to get followers: spammers just needed to massively follow other users.

Hence, the number of followers is not to be trusted and, indeed, it has been suggested that the follower-followee ratio is what really matters. In fact, such ratio can be interpreted as the user's "value" regarding the introduction of new original information from the "outside world" into the Twitter global ecosystem. Users publishing valuable tweets get followers in spite of being "impolite" (i.e. they do not follow back). That way they have huge number of followers but small numbers of followees and, thus, their ratios tend to be large. On the other hand, users who tend to discuss relatively personal matters with their close group of acquaintances do not get large audiences and, in turn, their ratios tend to be small.

How should we tackle with the spam problem, then? We think the answer lies on reciprocal connections. It seems obvious that those users with huge numbers of followers simply cannot follow-back everybody (not if they want to actually read what their followees are writing). Spammers, however, do not read tweets and, thus, they have no constrain in the number of people to follow. In other words, reciprocal links should be under suspicion and, hence, we define the follower-followee ratio with discounted reciprocity.

$$ratio\_discounted = \frac{followers - reciprocal}{followees - reciprocal}$$

However, putting under suspicion all reciprocal links seems a bit obnoxious; that's why we suggest employing either the follower-followee ratio or the discounted version depending on the possible outcome: if a user would "benefit" of using the original ratio then we use the discounted one, and vice versa. Because of that the complete name for our proposed ratio is in fact *followers to followee ratio with paradoxical discounted reciprocity*:

$$paradoxical\_discounted(p) = \begin{cases} \dfrac{followers(p)}{followees(p)} & if\ followers(p) > followees(p) \\ \dfrac{followers(p) - reciprocal(p)}{followees(p) - reciprocal(p)} & otherwise \end{cases}$$

It must be noticed that this ratio is not aimed to be directly applied to users in the graph but, instead, as a weight within an algorithm such as PageRank:

$$PR(p_i) = \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{|L(p_j)|} \cdot \frac{paradoxical\_discounted(p_j)}{max\_paradoxical\_discounted}$$

Nevertheless, as we will show below we also employed this ratio to obtain a "pruned" version of the user graph. That is, we removed all those users (and their connections) with a zero ratio to then apply standard PageRank to the "pruned" graph.

4 Research Design

The main goal of this study was to compare the performance of different graph centrality algorithms when applied to social networks. To do that, a dataset, and an objective criterion against which to compare the performance of the different methods were needed. The dataset was crawled by the author from Twitter. Then, a subset of "abusive" users was obtained from that dataset. The basic idea is to compare the different algorithms by analyzing the ranking spammers reach with each of them.

4.1 Dataset Description

We relied on the Twitter search API to create the dataset for this study. To that end, we employed a query composed of frequent English stop words. That query was submitted once every minute from January 26, 2009 to August 31, 2009 collecting about 28 million tweets.

Then we obtained followers and followees for each of the 4.98 million users appearing in that collection. For the final graph we did not take into consideration links from/to users not appearing in the sample and we also dropped isolated users. Besides, a substantial amount of user accounts were suspended at the moment of the second crawl. Thus, the finally collected Twitter graph consisted of 1.8 million users and 134 million connections.

4.2 Data Preparation: A Subset of Abusive Users within Twitter

As we have already said, Twitter spammers have both more followees and followers than legitimate users. According to [7], they triple the number of both kinds of connections; these researchers argue that *"spammers invest a lot of time following other users (and hoping other users follow them back)"*.

Thus, we decided to focus on Twitter spammers and a method to detect them was needed. We implemented a version of the method described in [7] achieving similar performance: 87.32% precision versus the 91% reported by them. This spam detection system detected 9,369 spammers in the dataset. By examining a representative sample we found that about 24% of them were already suspended by Twitter.

We obtained a list of distinctive terms from spammers biographies and terms such as `business`, `money`, `internet marketing`, `social media` and `SEO` were at the top of the list. Those terms are not only popular among spammers but among other Twitter users also. As Yardi *et al.* [7] said of them: *"[They] tread a fine line between reasonable and spam-like behavior in their efforts to gather followers and attention"*. We denoted those users as "aggressive marketers", and we decided to expand the target group from pure spammers to also include them. This way we found another 22,290 users which cannot be labeled as spammers but exhibit some common behaviors with them (see Table 1).

4.3 Evaluation Method

We did not assume any prior "correct" ranking for the users, instead we consider a ranking algorithm should be judged by its ability to avoid abusive users achieving undeserved rankings. Hence, the evaluation process was quite straightforward. All of the different methods were applied to the Twitter graph to obtain a user ranking. Then, we compared the positions reached by spammers and marketers in relation to average users. The lower the rankings abusive users reach, the better the method is.

Table 1. Features characterizing the behavior of spammers, marketers, and average users.

	Spammers	**Aggressive marketers**	**Average Users**
Avg. in-degree	**3203.28**	**1338.83**	82.36
Avg. out-degree	**3156.09**	**1245.35**	82.36
Avg. # of tweets over the whole period and Std. Dev.	**41.25 (80.99)**	**12.93 (34.07)**	5.60 (19.45)
% of tweets including URLs	**90.42%**	**32.86%**	18.21%
Avg. # of URLs per tweet including URLs	1.018	1.015	1.014
% of tweets including hashtags	11.54%	8.83%	7.98%
Avg. # of hashtags per tweet including hashtags	1.41	1.42	1.50
% of retweets over total tweets	2.97%	**6.50%**	2.87%
% of "conversations" over total (excluding retweets)	6.86%	**21.48%**	**19.26%**
Avg. # of users referred in conversational tweets (excluding retweets)	1.17	1.13	1.09

Table 2. Amount of the global prestige grabbed by abusive users and top rankings reached by 90% and 50% of the users from each class under different ranking algorithms.

Ranking method	**Spammers**			**Aggressive marketers**		
	% of global prestige	Ranking of 90% of spammers	Ranking of 50% of spammers	% of global prestige	Ranking of 90% of marketers	Ranking of 50% of marketers
PageRank	1.4%	Top-60%	Top-10%	3.3%	Top-80%	Top-20%
HITS	5.2%	Top-40%	Top-10%	11%	Top-60%	Top-20%
NodeRanking	1.62%	Top-60%	Top-10%	3.86%	Top-70%	Top-20%
TunkRank	**0.74%**	**Top-70%**	**Top-20%**	**1.94%**	**Top-90%**	**Top-40%**
TwitterRank	**0.0003%**	Top-30%	Top-10%	**0.00025%**	Top-80%	**Top-40%**
Discounted PageRank	**0.22%**	N/A: 40% spammers tie for the last position	**Top-20%**	**1.05%**	N/A: 55% of the aggressive marketers tie for the last position	
Pruned PageRank	1.84%	Top-50%	Top-10%	4.27%	Top-70%	Top-20%

5 Results

About 50% of the spammers detected in the collection of tweets do not appear in the graph. Those present account for 0.25% of the users. Regarding the aggressive marketers, 98% of them appear in the graph. The acute difference from spammer to marketer presence in the graph gives an idea of the work devoted by Twitter to get rid of spammers. Hence, the whole set of spammers and marketers represent a mere 1.5% of the users. The results obtained by the different ranking methods when confronted to these abusive users are summarized in Table 2, and Figures 1 and 2.

In addition to check the ability of the different algorithms to penalize abusive users it would be interesting to also check the level of agreement between the induced rankings and their implications. Figure 3 shows the agreement between the different methods and PageRank. Such agreements were computed according to the normalized version of Kendall distance with a zero penalty parameter [8][9].

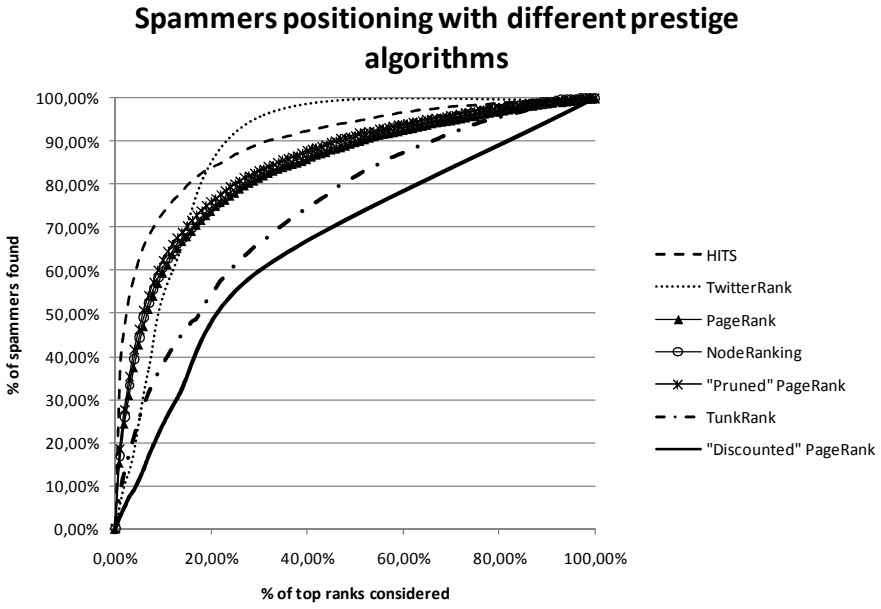

Fig. 1. Percent of spammers found for different slices of the users ranking.

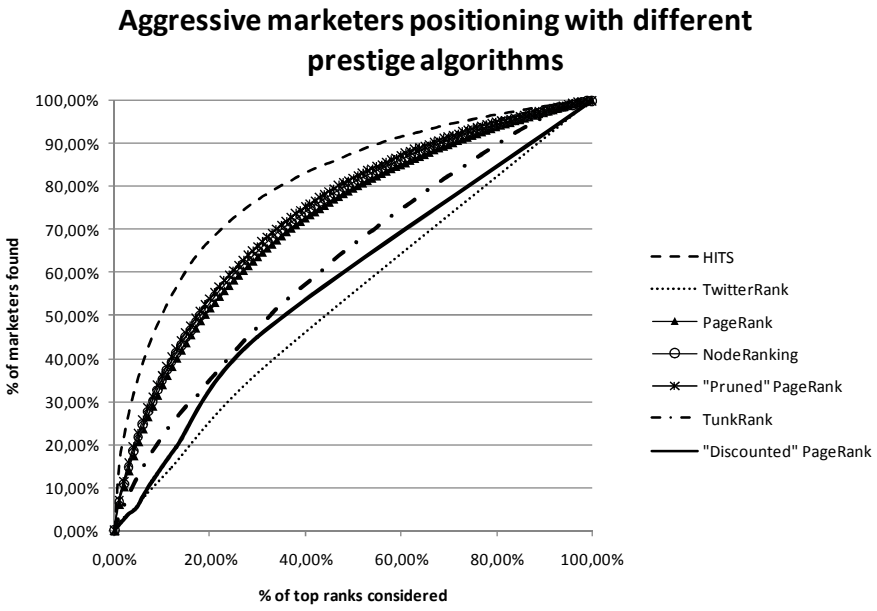

Fig. 2. Percent of aggresive marketers found for different slices of the users ranking.

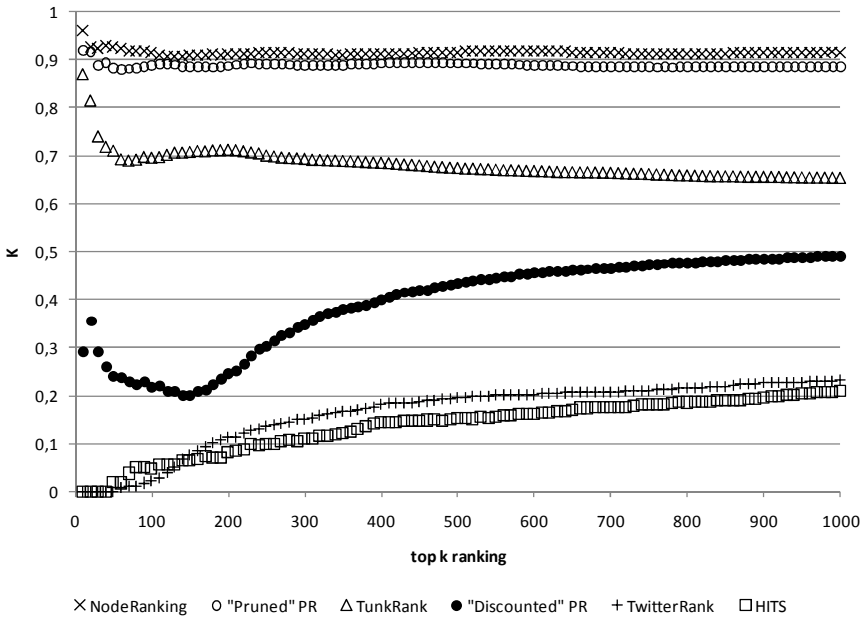

Fig. 3. Agreement between PageRank and the rest of rankings.

6 Discussion of Results

The analysis of the results obtained by PageRank when applied to the Twitter user graph supports our initial concern: users in social networks can easily game rank prestige algorithms –those described in this study. This is the most plausible explanation for spammers being much better positioned than aggressive marketers when the value of the contents provide by the former is virtually negligible.

The similarity between the results obtained by PageRank when applied to the complete graph and to the "pruned" version give support to another point of this author. Remember that the graph was "pruned" by removing those users with zero de-weighting which, in turn, was computed taking into account reciprocal links between users. One of the arguments of this author is that discounting such links is a fine way to separate users contributing value to the network from those with little or no value at all. Because the results obtained with both graphs are virtually the same we can take that as supportive of the goodness of our initial hypothesis.

There are two methods which greatly differ not only from PageRank but also from the other techniques, namely HITS and TwitterRank. Each of them exhibits different problems when applied to the Twitter graph.

HITS underperforms PageRank with respect to both spammers and marketers, and the induced ranking is very different from the other rankings. In fact, because of the very nature of HITS, this algorithm is virtually inoperative when confronted with a relatively small number of users weaving a tight network of reciprocal connections. Hub scores for users who massively follow other users tend to grow very fast and, then, those hub scores are used to compute authority scores for their followees (which are mostly spammers and are following them back). After just a few iterations those users with lots of reciprocal links earn an undeserved amount of authority. Hence, the HITS algorithm is not advisable to rank users within social networks without previously "cleaning" the graph.

Regarding the apparently contradictory results obtained by TwitterRank, they are due to the highly biased way in which it distributes prestige: the top 10 users account for 77% of the total prestige and the top 25 users for 95.5%. Virtually all of the users in the network achieve no prestige at all and, in spite of that, spammers manage to be "one-eyed kings in the land of the blind". This is pretty disappointing but, to be fair, modifying a topic-sensitive method to operate globally is, perhaps, pushing too hard the technique. Anyway, given that even the simplified version used for this research is (1) much more computationally expensive than the rest of methods surveyed, and (2) it requires much more data (namely, the tweets) to obtain the rankings, it seems not recommendable, especially when other available methods (e.g. TunkRank) are faster and provide much better results.

Lastly, there is one method clearly outperforming PageRank with respect to penalization of abusive users while still inducing plausible rankings: TunkRank. It is certainlty similar to PageRank but makes a much better job when confronted with "cheating": aggressive marketers are almost indistinguishable from common users, and spammers just manage to grab a much smaller amount of the global available prestige and reach lower positions than those achieved when using PageRank. In

addition to that, the ranking induced by TunkRank certainly agrees with that of PageRank, specially at the very top of the list, meaning that many users achieving good positions with PageRank should also get good positions with TunkRank. All of this makes TunkRank a highly recommendable ranking method to apply to social networks.

With regards to the proposal of this author, the results are not conclusive. It seems to outperform PageRank –and even TunkRank– because the amount of prestige grabbed by abusive users is smaller and their rankings lower than when applying standard PageRank. Nevertheless, it has two issues which deserve further research.

On one hand the induced ranking could be labeled as "elitist" because 70% of the users tie for the last position. One could argue that this is unsurprising given that 16% of the users from the graph have got a zero de-weighting factor; and, in fact, such results are consistent with the well-known participation inequality [10], and with a recent study revealing that 75% of the users just publish a tweet every 9 days, and 25% of the users do not tweet at all [11]. Thus, this could be considered a minor issue.

On the other hand, "discounted" PageRank exhibits a fairly distinctive curve when comparing its agreement with PageRank (see Figure 3). The agreement is much lower than, for instance, that found between PageRank and TunkRank, but the most striking behavior is the local maximum at the top positions, followed by a relatively large trough, to eventually stabilize. We found several lesser-known users at top ranks and, after studying them, we concluded that most of them have one or more "famous" followers who, in many cases, they manage to outrank. We have denoted this as the "giant shoulders" effect and it explains not only the trough at the head of the list but the smaller agreement for the rest of the ranking: many of the top users from PageRank or TunkRank are a little behind of lesser-known users they are following. This is aesthetically displeasing, at least, and the effect it can exert in the applications of the ranking is still to be explored. Nevertheless, tackling with this and the former issue is left for future research.

7 Conclusions and Future Work

This study makes four main contributions. First, graph prestige in social networks can be "gamed" by means of relationship links. The fact that spammers are always better positioned than marketers supports this assert.

Second, evaluating ranking in itself should not be the point; it should, instead, be evaluated within an objective context. Avoiding abusive users to reach undeserved rankings is a good metric to compare the performance of different algorithms.

Third, TunkRank is an obvious candidate to rank users in social networks. Although highly related to PageRank, TunkRank outperforms it with respect to penalizing abusive users while still inducing plausible rankings. In addition to that, it is simple to implement and computationally cheap –at least as cheap as PageRank.

And fourth, de-weighting the influence of a user by discounting reciprocal links seems to be a good way to separate those users contributing valuable contents to the global ecosystem from those with little to no value at all. This is supported by the fact

that when applying PageRank to both the complete version of the Twitter graph and to a "pruned" version we obtained virtually the same results.

The study opens several lines of research. First, the rankings induced by the different methods should be analyzed in other contexts, for instance, as a way to rank content providers in order to find relevant information within a social network. Second, TunkRank can for sure be manipulated and, thus, its vulnerabilities should be thoroughly studied. And third, a deeper analysis of the role of nepotistic links, in general, and the discounted ratio described in this paper, in particular, is needed.

Acknowledgements

The authors would like to thank F. Zapico and D. Guerra for their help during the Twitter dataset collection, and to M. Fernández for comments on an early draft of this paper. This work was partially financed by grant UNOV-09-RENOV-MB-2 from the University of Oviedo.

8 References

1. Talbot, D. How Google Rank Tweets, http://www.technologyreview.com/web/24353/
2. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, http://dbpubs.stanford.edu/pub/1999-66
3. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, pp. 668—677 (1998)
4. Pujol, J.M., Sangüesa, R., Delgado, J.: Extracting Reputation in Multi Agent Systems by Means of Social Network Topology. In: Proceedings of the first international joint conference on Autonomous agentsand multiagent systems, pp. 467—474 (2002)
5. Tunkelang, D.: A Twitter Analog to PageRank, http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/
6. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers. In: WSDM'10: Proceedings of the third ACM international conference on Web Search and Data Mining, pp. 261—270 (2010)
7. Yardi, S., Romero, D., Schoenebeck, G., boyd, d.: Detecting spam in a Twitter network. First Monday, vol. 15, no. 1—4 (2010)
8. McCown, F., Nelson, M.L.: Agreeing to Disagree: Search Engines and Their Public Interfaces. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital Libraries, pp. 309—318 (2007)
9. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. In: Proceedings of the 14th annual ACM-SIAM symposium on Discrete algorithms, pp. 28—36 (2003)
10. Nielsen, J.: Participation Inequality: Encouraging More Users to Contribute. http://www.useit.com/alertbox/participation_inequality.html
11. Heil, B., Piskorski, M.: New Twitter Research: Men Follow Men and Nobody Tweets. http://blogs.hbr.org/cs/2009/06/new_twitter_research_men_follo.html