Page Rank vs. Katz: Is the centrality algorithm choice relevant to measure user influence in Twitter?*

Hugo Rosa[1], Joao P. Carvalho[1,2], Ramon Astudillo[1], and Fernando Batista[1,3]

[1] INESC-ID
[2] Instituto Superior Técnico, Universidade de Lisboa
[3] ISCTE-IUL - Instituto Universitário de Lisboa
{hugo.rosa,joao.carvalho,ramon.astudillo,fmmb}@inesc-id.pt

Abstract. Microblogs, such as Twitter, have become an important socio-political analysis tool. One of the most important tasks in such analysis is the detection of relevant actors within a given topic through data mining, i.e., identifying who are the most influential participants discussing the topic. Even if there is no gold standard for such task, the adequacy of graph based centrality tools such as PageRank and Katz is well documented. In this paper, we present a case study based on a "London Riots" Twitter database, where we show that Katz is not as adequate for the task of important actors detection since it fails to detect what we refer to as "indirect gloating", the situation where an actor capitalizes on other actors referring to him.

Keywords: Page Rank, Katz, User Influence, Twitter, Data Mining

1 Introduction

Nowadays, there are 288 million active users on Twitter and more than 500 million tweets are produced per day [17]. Through short messages, users can post about their feelings, important events and talk amongst each other. Twitter has become so much of a force to be reckoned with, that anybody from major brands and institutions, to celebrities and political figures use it to further assert their position and make their voice heard. The impact of Twitter on the Arab Spring [6] and how it beat the all news media to the announcement of Michael Jackson's death [15], are just a few examples of Twitter's role in society. When big events occur, it is common for users to post about it in such fashion, that it becomes a trending topic, all the while being unaware from where it stemmed or who made it relevant. The question we wish to answer is: "Which users were important in disseminating and discussing a given topic?".

* This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under project PTDC/IVC-ESCT/4919/2012 and funds with reference UID/CEC/50021/2013.

Much like real life, some users carry more influence and authority than others. Determining user relevance is vital to help determine trend setters [16]. The user's relevance must take into account not only global metrics that include the user's level of activity within the social network, but also his impact in a given topic [18]. Empirically speaking, an influential person can be described as someone with the ability to change the opinion of many, in order to reflect his own. While [13] supports this statement, claiming that "a minority of users, called influentials, excel in persuading others", more modern approaches [4] seem to emphasize the importance of interpersonal relationships amongst ordinary users, reinforcing that people make choices based on the opinions of their peers.

In [2], three measures of influence were taken into account: "in-degree is the number of people who follow a user; re-tweets mean the number of times others forward a user's tweet; and mentions mean the number of times others mention a user's name.". It concluded that while in-degree measure is useful to identify users who get a lot of attention, it "is not related to other important notions of influence such as engaging audience". Instead "it is more influential to have an active audience who re-tweets or mentions the user". In [8], the conclusion was made that within Twitter, "news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users", while "celebrities with higher follower totals foster more conversation than provide retweetable content". The authors in [12] created a framework named "InfluenceTracker", that rates the impact of a Twitter account taking into consideration an Influence Metric, based on the ratio between the number of followers of a user and the users it follows, and the amount of recent activity of a given account. Much like [2], it also shows "that the number of followers a user has, is not sufficient to guarantee the maximum diffusion of information (...) because, these followers should not only be active Twitter users, but also have impact on the network".

In this paper, we analyze how two well known network analysis algorithms, PageRank and Katz, affect the computation of mention-based user influence in Twitter. Although these two methods have previously been compared [11] and found to have been equivalent, we show that the same conclusion does not apply in the context of social networks, and that PageRank is indeed more adequate. We base our conclusions on a real world case study of the 2011 London Riots, since it was an important social event where Twitter users were said to have played a role in its origin and dissemination.

2 User influence representation

We propose a graph representation of user's influence based on "mentions". Whenever a user is mentioned in a tweet's text, using the @*user* tag, a link is made from the creator of the tweet, to the mentioned user. For example, the tweet "*Do you think we can we get out of this financial crisis,* @*userB?*", from @*userA*, creates the link: @*userA* $\longrightarrow$ @*userB*. This is also true for re-tweets,e.g.

the tweet "*RT @userC The crisis is everywhere!*" from $@userA$, creates the link: $@userA \longrightarrow @userC$.

This representation not only is an exact structural replica of the communication web between users, but it also provides dynamism to how influence can be given and taken across the graph.

In graph theory and network analysis, the concept of centrality refers to the identification of the most important vertices's within a graph, i.e., most important users. We therefore define a graph $G(V, E)$ where V is the set of users and E is the set of directed links between them.

3 Network Analysis Algorithms

The computation of user influence is done by applying a centrality based algorithm to the graph presented in section 2. Here we present two of the most well-known and used centrality algorithms, Page Rank and Katz.

3.1 PageRank

Arguably the most well known centrality algorithm is PageRank [9]. It is one of Google's methods to its search engine and it was created as way for computing a ranking for every web page based on the graph of the web uses. In this algorithm, web pages are nodes, while back-links form the edges of the graph (Figure 1). It is defined by Equation 1 as $PR(v_i)$ of a page v_i.

$$PR_{vi} = \frac{1-d}{N} + d \sum_{v_j \in M(v_i)} \frac{PR(v_j)}{L(v_j)} \tag{1}$$

It can be intuitively said about Equation 1, that a page has high rank if the sum of the ranks of its back-links is high. In it, v_j is the sum ranges over all pages that has a link to v_i, $L(v_j)$ is the number of outgoing links from v_j, N is the number of documents/nodes in the collection and d is the damping factor. The PageRank is considered to be a random walk model, because the weight of a page v_i is "the probability that a random walker (which continues to follow arbitrary links to move from page to page) will be at v_i at any given time. The damping factor corresponds to the probability of the random walk to jump to an arbitrary page, rather than to follow a link, on the Web. It is required to reduce the effects on the PageRank computation of loops and dangling links in the Web." [11]. Dangling links are "simply links that point to any page with no outgoing links (...) they affect the model because it is not clear where their weight should be distributed" [9]. The true value that Google uses for damping factor is unknown, but it has become common to use $d = 0.85$ in the literature. A lower value of d implies that the graph's structure is less respected, therefore making the "walker" more random and less strict.

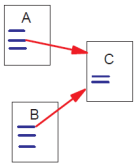

Fig. 1. A and B are back-links of C

3.2 Katz

Another well known method is the Katz algorithm [7]. It is a generalization of a back-link counting method where the weight of each node is "determined by the number of directed paths that ends in the page, where the influence of longer paths is attenuated by a decay factor" and "the length of a path is defined to be the number of edges it contains" [11]. It is defined by Equation 2 "where $N(v_i, k)$ is the number of paths of length k that starts at any page and ends at v_i and α is the decay factor. Solutions for all the pages are guaranteed to exist as long as α is smaller than $\lambda > 1$, where $1/\lambda$ is the maximum in-degree of any page" [11].

$$I_{vi} = \sum_{k=0}^{\infty} [\alpha^k N(v_i, k)] \tag{2}$$

It was shown in [11] that "Katz status index may be considered a more general form of PageRank because in can be modified, within a reasonable range, to be equivalent to PageRank" and that under a "relaxed definition of equivalence (...) PageRank and Katz status index is practically equivalent to each other" as long as the number of outgoing links from any vertex is the same throughout the graph, which is very unlikely for graph modeled from a social network. On the other hand, "it is also possible to modify PageRank to become completely equivalent to Katz status index", however, in that case, "the modified PageRank is no long a random work model because it can no longer be modeled from a probabilistic standpoint" [11].

4 Dataset

In order to test the network analysis methods presented above, a database from the London Riots in 2011 [3] was used. The London Riots of 2011 was an event that took place between the 6th and 11th August 2011, where thousands of people rioted in several boroughs of London with the resulting chaos generated looting, arson, and mass deployment of police. Although Twitter was said to be a communication tool for rioting groups to organize themselves, there is little evidence that it was used to promote illegal activities at the time, though it was useful for spreading word about subsequent events. According to [5], Twitter played a big role spreading the news about what was happening and "was a

valuable tool for mobilizing support for the post-riot clean-up and for organizing specific clean-up activities". Therefore it constitutes a prime data sample to study how users exert influence in social networks, when confronted with such a high stakes event.

The Guardian Newspaper made public a list of tweets from 200 influential twitter users, which contains 17795 riot related tweets and an overall dataset of 1132938 tweets. Using a Topic Detection algorithm [1], we obtained an additional 25757 unhastagged tweets about the London Riots. It consists of a Twitter Topic Fuzzy Fingerprint algorithm [14] that provides a weighted rank of keywords for each topic in order to identify a smaller subset of tweets within scope. This method has proven to achieve better results than other well known classifiers in the context of detecting Topics within Twitter, while also being faster in execution. The sum of posting and mentioned users is 13765 (vertices) and it has 19993 different user mentions (edges), achieving a network connectivity ratio of $\frac{edges}{vertices} = 1.46$.

5 Experiments and Results

In this section, we compare the results of ranking the most influential users using Page Rank, Katz and a mentions based baseline. We proceed by performing an empirical analysis of the users in order to ascertain their degree of influence and their position in the ranks. The graphs and ranking were calculated using *Graph-Tool* [10].

Table 1 shows how both network analysis algorithms behave while highlighting the rank differences (shown by the arrows in the last column). A "Mentions rank" is used as a base line. Figure 2 provides a visual tool to the graph, as provided by PageRank.

There is an obvious relation between the number of mentions and the ranking provided by the application of both algorithms: the highest ranked users in either Katz and PageRank, are some of the most mentioned users in our dataset. In fact, the relation is more clear between Katz and the baseline Mentions based ranking: Table 1 shows that the rank in both approaches is always either identical (@guardian, @skynewsbreak, @gmpolice, etc...) or at most separated by two positions (@richardpbacon is ranked 27th based on mentions, and 29th based on Katz). In order to determine the relation between PageRank and "Mentions Rank", the Spearman correlation was calculated having achieved a value of $\rho = 0.9372$, which means they are heavily correlated. However, when limiting this calculation to the top 20, it changed to $\rho = 0.5535$, which implies that for the top users, just looking at the number of mentions, is not enough to determine influence.

An empirical analysis also shows that both Page Rank and Katz largely agree upon the ranking of most users, namely on the top two users: i) @guardian, Twitter account of the world famous newspaper "The Guardian"; ii) @skynewsbreak, Twitter account of the news team at Sky News TV channel. This outcome agrees with [8] previous statement, that, "news outlets, regardless of follower count, in-

Table 1. London Riots Top 20 most influential users according to Page Rank, and comparison with Katz. The arrows indicate most relevant rank differences.

User	Mentions		PageRank		Katz		
	#	rank	score	rank	score	rank	
@guardian	160	2	0.0002854	1	0.022157	2	
@skynewsbreak	178	1	0.0002512	2	0.023479	1	
@gmpolice	122	4	0.0002128	3	0.019009	4	
@riotcleanup	107	6	0.0001767	4	0.017992	6	↘
@prodnose	67	14	0.0001761	5	0.014022	15	↘↘↘
@metpoliceuk	116	5	0.0001494	6	0.018709	5	
@marcreeves	69	11	0.0001476	7	0.014195	12	↘↘
@piersmorgan	78	8	0.0001465	8	0.014959	9	
@scdsoundsystem	69	12	0.0001442	9	0.014190	13	↘↘
@subedited	70	10	0.0001337	10	0.014278	11	
@youtube	48	20	0.0001257	11	0.012424	20	↘↘↘
@bbcnews	94	7	0.0001256	12	0.016426	8	↗↗
@mattkmoore	62	15	0.0001237	13	0.013614	16	↘
@richardpbacon	40	27	0.0001218	14	0.011771	29	↘↘↘
@lbc973	34	35	0.0001150	15	0.011432	34	↘↘↘↘
@skynews	74	9	0.0001113	16	0.014638	10	↗↗
@bengoldacre	61	17	0.0001055	17	0.013526	17	
@bbcnewsnight	68	13	0.0000988	18	0.014123	14	↗↗
@tom_watson	44	21	0.0000968	19	0.012107	22	↘
@paullewis	129	3	0.0000954	20	0.019602	3	↗↗↗↗
...							
@juliangbell	61	16	0.0000275	188	0.0166597	7	↗↗↗↗↗↗↗

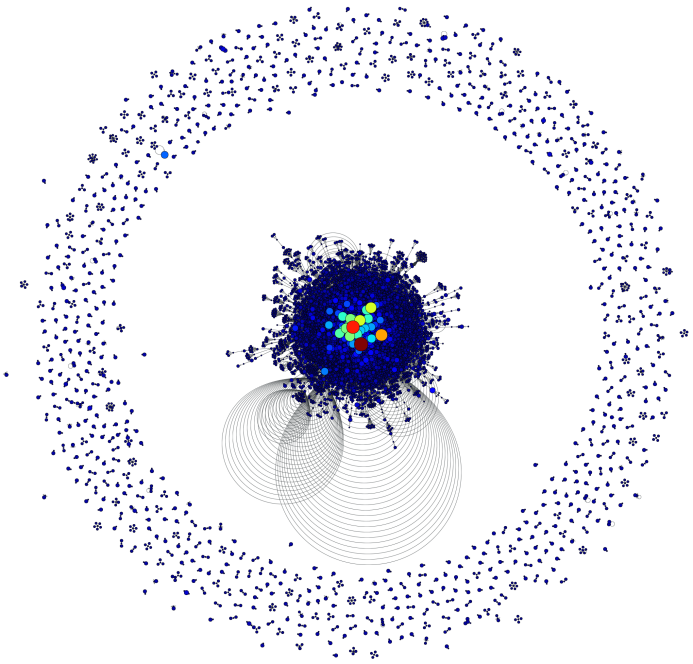

Fig. 2. User influence Page Rank Graph - larger circles indicate larger user influence.

fluence large amounts of followers to republish their content to other users" and can be justified by the incredibly high London Riots news coverage. Other users seem to fit the profile, namely @gmpoliceq, @bbcnews and @skynews. Most of the other users are either political figures, political commentators or jornalists (@marcreeves, @piersmorgan, and @mattkmoore).

However, when looking more closely at Page Rank vs. Katz rankings, it is also possible to realize some notorious differences: Katz's third and seventh top ranked users are not in PageRank's top users. The reasons behind these differences in the ranking positions should be thoroughly analyzed since they could highlight the strengths and weaknesses of each algorithm in what concerns their capability to express user influence in social networks. The two cases end up being different and should be treated separately: i) @paullewis, ranked 3rd by Katz shows up at 20th according to PageRank; ii) @juliangbell, ranked 7th by Katz shows up at 188th according to PageRank.

The reason behind @paullewis high placement in the Katz rank is the number of mentions. As said previously, Katz is a generalization of a back-link counting method, which means the more back-links/mentions a user has, the higher it will be on the ranking. This user has 129 mentions, but PageRank penalizes it, because it is mentioned by least important users, which means a less sum weight is being transfered to it in the iterative process. This logic also applies to users @bbcnewsnight, @skynews and @bbcnews. Additionally, @paullewis is also an

active mentioning user, having mentioned other users a total of 14 tweets, while @skynewsbreak and @guardian have mentioned none. As a consequence, Paul Lewis transfers its influence across the network while the others simply harvest it. There are several users that drop in ranking from PageRank to Katz for the very same reason. Users such as @prodnose, @marcreeves and @youtube do not have enough mentions for Katz to rank them higher.

User @juliangbell, despite mentioned often (61 times), is down on the PageRank because of indirect gloating, i.e., he retweets tweets that are mentioning himself: *"@LabourLocalGov #Ealing Riot Mtg: @juliangbell speech http://t.co/3BNW0q6"* was posted by @juliangbell himself. The user is posting somebody else's re-tweet of one of his tweets. As a consequence a link/edge was created from @juliangbell to @LabourLocalGov, but also from @juliangbell to himself, since his username is mentioned in his own tweet. Julian Bell is a political figure, making it acceptable that he would have a role in discussing the London Riots, but the self congratulatory behavior of re-tweeting other people's mentions of himself, is contradictory with the idea of disseminating the topic across the network. While Katz is not able to detect this effect, PageRank automatically corrects it, which is why, contrary to what is mentioned in previous works [11], it is our comprehension that Katz is not equivalent to PageRank in the task of detecting user relevance in social networks such as Twitter

6 Conclusions

With this study, we have shown that in the context of user influence in Twitter, PageRank and Katz are not equal in performance, thus disproving previous claims. PageRank has proved a more robust solution to identify influential users in discussing and spreading a given relevant topic, specially when considering how it deals with indirect gloating, an item Katz fails to penalize.

References

1. Carvalho, J.P., Pedro, V., Batista, F.: Towards intelligent mining of public social networks' influence in society. In: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS). pp. 478 – 483. Edmonton, Canada (June 2013)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: in ICWSM '10: Proceedings of international AAAI Conference on Weblogs and Social (2010)
3. Crockett, K, S.R.: Twitter riot dataset (tw-short) (2011)
4. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 57–66. KDD '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/502512.502525
5. Guardian, T.: http://www.theguardian.com/uk/2011/dec/07/twitter-riots-how-news-spread
6. Huang, C.: Facebook and twitter key to arab spring uprisings: report. http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report (June 2011), accessed: 2014-05-02

7. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (March 1953), http://ideas.repec.org/a/spr/psycho/v18y1953i1p39-43.html
8. Leavitt, A., Burchard, E., Fisher, D., Gilbert, S.: The influentials: New approaches for analyzing influence on twitter (2009)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web (1999)
10. Peixoto, T.: https://about.twitter.com/company
11. Phuoc, N.Q., Kim, S.R., Lee, H.K., Kim, H.: Pagerank vs. katz status index, a theoretical approach. In: Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology. pp. 1276–1279. ICCIT '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/ICCIT.2009.272
12. Razis, G., Anagnostopoulos, I.: Influencetracker: Rating the impact of a twitter account. CoRR (2014), http://arxiv.org/abs/1404.5239
13. Rogers, E.M.: Diffusion of innovations (1962)
14. Rosa, H., Batista, F., Carvalho, J.P.: Twitter topic fuzzy fingerprints. In: WCCI2014, FUZZ-IEEE, 2014 IEEE World Congress on Computational Intelligence, International Conference on Fuzzy Systems. pp. 776–783. IEEE Xplorer, Beijing, China (July 2014)
15. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: News in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 42–51. GIS '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1653771.1653781
16. Tinati, R., Carr, L., Hall, W., Bentwood, J.: Identifying communicator roles in twitter. In: Proceedings of the 21st International Conference Companion on World Wide Web. pp. 1161–1168. WWW '12 Companion, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2187980.2188256
17. Twitter: https://about.twitter.com/company
18. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 261–270. WSDM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1718487.1718520