Detecting User Influence in Twitter: PageRank vs Katz, a case study*

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 [16]. 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 [14], 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?".

Determining user relevance is vital to help determine trend setters [15]. 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 [17]. 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.

* 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.

While [12] supports this statement, claiming that "a minority of users, called influentials, excel in persuading others", more modern approaches [5] seem to emphasize the importance of interpersonal relationships amongst ordinary users, reinforcing that people make choices based on the opinions of their peers. In [3], 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 [1], 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 [11] 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 [3], it also shows that "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".

With the previous definitions of influence in mind, 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, regardless of it being a retweet or a conversation. 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*.

2 Network Analysis Algorithms

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. Arguably the most well known centrality algorithm is PageRank [8]. It is one of Google's methods to its search engine and uses web pages as nodes, while back-links form the edges of the graph. It is defined by Equation 1 as $PR(v_i)$ of a page v_i.

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

In Equation 1, 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.” [10]. 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.

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” [10]. 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” [10].

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

3 Experiments and Results

In order to test the network analysis methods presented above, a database from the London Riots in 2011 [4] was used. 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 [2], we obtained an additional 25757 unhastagged tweets about the London Riots. It consists of a Twitter Topic Fuzzy Fingerprint algorithm [13] that provides a weighted rank of keywords for each topic in order to identify a smaller subset of tweets within scope. The sum of posting and mentioned users is 13765 (vertices) and it has 19993 different user mentions (edges), achieving a network connectivity ratio of $\frac{edges}{vertices} = 1.46$.

The remainder of this section presents the results of each algorithm’s ranking for most influential users. An empirical study of the users is made, in order to ascertain their degree of influence. The graphs and ranking were calculated using *Graph-Tool* [9].

Table 1 shows how both network analysis algorithms behave with our graph representation, while highlighting the changes in rank between them, as shown by the arrows in the last column. Figure 1 provides a visual tool to the graph, as provided by PageRank. There is a relation between the number of mentions and the ranking in both algorithms, since these users are some of the most mentioned users in our dataset.

When comparing PageRank with Katz, several differences arise, but the top two users are agreed upon: i) @guardian, Twitter account of the world famous

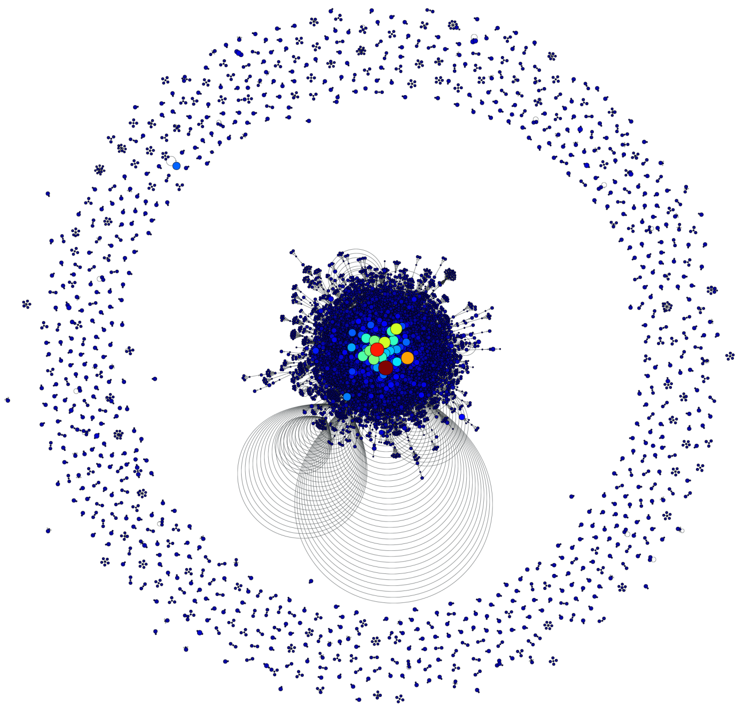

Fig. 1. User influence Page Rank Graph - larger circles indicate larger user influence.

Table 1. Most influential users according to Page Rank, and comparison with Katz.

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	↘
...							
@paullewis	129	3	0.0000954	20	0.019602	3	↗↗↗↗
...							
@juliangbell	61	16	0.0000275	188	0.0166597	7	↗↗↗↗↗↗↗

newspaper "The Guardian"; ii) @skynewsbreak, Twitter account of the news team at Sky News TV channel. This outcome agrees with [1] previous statement, that, "news outlets, regardless of follower count, influence large amounts of followers to republish their content to other users". Other users seem to fit the profile, namely @gmpoliceq and @bbcnews. Most of the other users are either political figures, political commentators or jornalists (@marcreeves, @piersmorgan, and @mattkmoore).

However, Katz's third and seventh top ranked users, are not in PageRank's top users. These are two very different cases: 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 user @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. Contrary to what is mentioned in previous works, it is our comprehension that Katz is not adequate to detect a user's importance in social media such as Twitter.

4 Conclusions and Future Work

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. Alex Leavitt, Evan Burchard, D.F.S.G.: The influentials: New approaches for analyzing influence on twitter (2009)
2. 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)
3. 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)
4. Crockett, K, S.R.: Twitter riot dataset (tw-short) (2011)
5. 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
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. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web (1999)
9. Peixoto, T.: https://about.twitter.com/company
10. 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
11. Razis, G., Anagnostopoulos, I.: Influencetracker: Rating the impact of a twitter account. CoRR (2014), http://arxiv.org/abs/1404.5239
12. Rogers, E.M.: Diffusion of innovations (1962)
13. 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)
14. 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
15. 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
16. Twitter: https://about.twitter.com/company
17. 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