To Find Influential's in Twitter based on Information Propagation

Md Tabrez Nafis
Asst Professor, Department of Computer Science, Jamia Hamdard , Hamdard Nagar, New Delhi, Delhi 110062

Alok K Pathak
Student, M Tech., Department of Computer Science Jamia Hamdard , Hamdard Nagar, New Delhi, Delhi 110062

ABSTRACT
Identification of Influencers in Social Networks and targeting them for viral marketing is fast becoming the core idea in designing the strategies for brand building. The influential individuals stimulate "word-of-mouth" effects in their network[14]. This leads to triggering of long cascades of influence that convince their peers to perform a similar action (Participate in Survey , click on advertisement, 'like' a community etc).Targeting these influentials usually leads to a vast spread of the information across the network. Hence it is important to identify such individuals in a network.

Last five years have seen tremendous growth of social media People having access to Internet are spending considerable amount of time on various platforms generating huge data content by participating in survey, approving a thought, endorsing some view, building social profile and sharing knowledge. This has resulted in social influence emerging as a complex force, governing the diffusion of the influence in the network. Individual's social influence on the network tempts various companies to go beyond the conventional marketing for finding new customers, thereby making profitable marketing decisions.

In this work we have examined the identification of Influencials in the increasingly popular online social network, Twitter using Google's PageRank Algorithm [16] and have optimized the calculation of ranking score by taking into account the dynamism of 'Retweet' or 'Mention' functionality used by the twitter users which actually signify their actual participation in the idea propagation and subsequent cascade of influence in their network. The idea is to improve further on the intuition that the users with most number of followers are also the most influential in the given network. The improved PageRank algorithm not only considers the number of followers the user has , but also the number of times their tweet have been retweeted or their user name have been mentioned in the their follower's tweets.

General Terms
Algorithm, Online Social Network, Social Network Analysis

Keywords
Online Social Networks, Twitter, Influence, word of mouth marketing.

1. INTRODUCTION
People are increasingly becoming immune to conventional advertising mediums while the word of mouth diffusion is still regarded as one of the most effective mechanism of information propagation. Experimental study of measuring the word-of-mouth diffusion in a social network has suffered from two major roadblocks. First, it is difficult to tangibly observe the network over which word-of-mouth influence spreads, and hence influence is difficult to attribute accurately, especially in instances where diffusion propagates to multiple levels. And second, the observational data on diffusion are heavily biased towards the users who have higher degree of centrality by virtue of having more connections and peer reciprocity.

The micro-blogging social network Twitter presents a promising natural laboratory for the study of information propagation through the network unlike other Social Networks (e.g. Facebook) as Twitter is by design a tool to share or broadcast the information not only within the 'own network' but potentially to the whole world. Twitter data is exceptionally open and publically available as one does not require a prior approval from the content generator to 'Follow' them. Thus the network of "who listens to whom" can be reconstructed by crawling the corresponding "follower graph"[14].

Moreover since there is a limitation of 140 characters per tweet, the usual practice is to share the information using shortened URL services like bit.ly, TinyURL etc. Twitter exposes streaming APIs which could be used to retrieve the tweets, retweets and user mentions which can be used to quantify the Influence. Which makes the Twitter ecosystem is one of the best suited for studying the role of Influencers.

The individuals who have the capability to act as catalyst for the information to spread in the network are generally identified as influencers. This is a too broad definition because we may get influenced by friends, acquaintances, celebrities, popular media figures who use different medium to share their opinion or recommendation which is extremely different to compare empirically. Also, different kinds of Influencials end up influencing different categories of people which might be mutually exclusive. E.g. a sports personality may not influence a group of C++ programmers about a new textbook. Whereas a known person working in same field have more than decent chance of influencing one's decision to purchase the book.

Fortunately, Twitter's growing popularity has forced different types of potential influencers to communicate in exactly the same way under same limitations. The popular perception that the persons with more social connections ('Follower' in case of twitter) may also be validated by counting the number of 'Retweets' as an indication that the follower has been influenced by the user to share the information originally ideated by the influencer. This can also be used to compare the influencing capability of different types of Influencers and get them ranked.

It is submitted that our use of the term influencer is subject to assumption that user A is said to be influenced by user B if A shares (or Retweets) the seed URL originally posted by B which diffuse through the Twitter follower graph. The user B has not posted the URL which has been received through the follower graph. In the improved PageRank algorithm, we quantify the influence of a given post by the number of users who subsequently repost the URL, which can be traced back to the originating user through the follower graph.

2. RELATED WORK

A number of recent empirical papers have addressed the Subject of identification of influencers in Online Social Network. Adar and Admaic [11] in used an approach to reconstruct diffusion trees among bloggers. Leskovec et al [9] used referrals on an e-commerce site to infer how individuals are influenced as a function of how many of their contacts have recommended a product. Sun et al. [12] studied diffusion trees of fan pages on Facebook. Bakshy et al. [14] studied the diffusion of "gestures" between friends in Second Life. Weng et al [13] studied the measurement of influence taking both the topical similarity between users and the link structure into account.

The online social network is very similar to hyperlinked webpages on the Internet and hence the PageRank Algorithm[15] which was originally conceived for the Internet is also being used to determine the centrality of the users and since the most referenced webpages are considered to be the closest to the search string and is ranked first in the result set, this promotes the intuition that the user with highest centrality should also be the most influential because it has potentially the largest reach among all the users on the network. Hence many of the recently published papers that have examined the quantification of Influence specifically on twitter have focused the research around the PageRank Algorithm. Kwak et al [10] took into account the three pillars of social network influence : number of followers, page-rank, and number of retweets- and concluded that the rankings vary depending on the attribute under study. Similarly, Cha et al. [1] compared another set of three different measures of influence - number of followers, number of retweets, and number of mentions and also found that the most followed users did not necessarily score highest on the other measures. Weng et al. [13] proposed a modified PageRank algorithm that also accounted for topical similarity.

However, the PageRank Algorithm is based on the static nature of the network topology and hence it does not take into account the dynamic participation by the user while propagating the information in its follower graph

3. DATASET

In this section we describe the data retrieved from twitter for this study. During the crest of recent Swine Flu, the 'H1N1' topic was trending on twitter for a long time. We collected the 150 hours tweets during the period 23.03.2015 to 31.03.2015 related to 'H1N1' and filtered the tweets that contained URL (i.e., tweets referring bit.ly, tinyurl services). A Total of 5593 such tweets were collected for this study. The twitter handle was collected in a queue and for each twitter handle the followers were crawled to inspect if they have retweeted the original tweet from the primary contributor. All such twitter handles were inserted into the same queue. This process was iterated until all active users were crawled. The data was extracted using the standard twitter API available at http://dev.twitter.com. None of the nodes are inactive i.e., having not tweeted about H1N1 during the mentioned period.

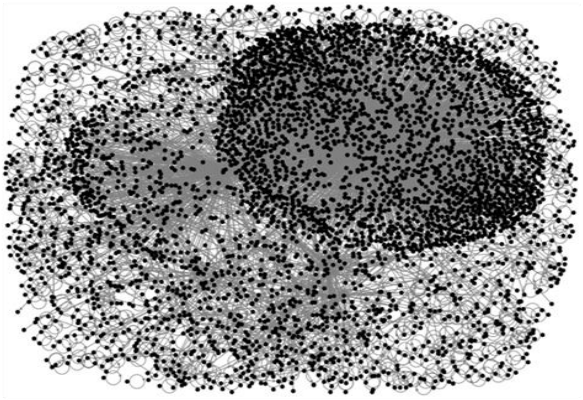

Fig 1: The Social Graph depicting the dataset under study

4. INFLUENCE COMPUTATION

4.1 PageRank Algorithm

The PageRank algorithm as described in the original Google paper[16] proposes the following equation –

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))
where

- PR(A) is the PageRank of page A,
- PR(Ti) is the PageRank of pages Ti which link to page A,
- C(Ti) is the number of outbound links on page Ti and
- d is a damping factor which can be set between 0 and 1.

The damping factor (d) may vary but eventually the pages get ranked in exactly the same order irrespective of the value of d. The relationships in social network is mode dynamic in nature and the simple edge between two nodes (i.e., users) doesn't not reflect the relationship clearly if it is a 'follower' or an active follower who has helped in broadcasting the information ideated by the primary contributor.

4.2 Improved PageRank Algorithm

In order to quantify the influence exerted by the 'special users' on their follower graph, we need to put appropriate weights on the edges connecting the user with its follower.

If a user A has replied to another specific user B or retweeted the content shared by user B more than to others, the user B can be considered as an influential person to A.

This can be calculated as under –

$$\text{Retweet Score (ui)} = \frac{\sum_{\text{follower(ui)}} \text{Retweet count (Follower)}}{\sum_{\text{follower(ui)}} \text{Tweet count (Follower)}}$$

The follower(ui) is the follower of user ui and the when the follower(ui) retweets the tweet originally posted by ui the Retweet Count increases. Similarly if the follower (ui) seldom retweets, and is an active user with valid number of tweets posted, the Retweet Score would decrease and hence the influence score would be lesser.

This retweet score could replace the PR(Tn)/C(Tn)) component of PageRank Algorithm and the Improved PageRank could take the following shape –

Influence(A) = (1-d) + d ((Retweet Score (u_1) + ... + d(Retweet Score (u_n))

Illustration-

If user A's tweet containing URL is retweeted by user B and B Є Follower(A). User B meanwhile tweets 10 more tweets not related to the previous one. The Influence exerted by A on B is –
Influence (A) = 1-0.85 +0.85(1/10) =0.235
Assume, the damping factor d = 0.85

The underlying principle behind the modified equation is as follows. If a person is more influential than others, her/his followers would be more responsive to her/his. There is strong likelihood that it the follower has been influenced by the user , there will be retweets/ reply to or appreciation of information and thereby generating more interaction between them. We can even ignore the non active nodes who hardly interact and have a simplistic network for visual analysis as against the complete social graph (refer fig 1).

4.3 Comparison between PageRank and Improved PageRank algorithm

In many ways the Improved PageRank Algorithm works on the similar lines like the original PageRank algorithm. However, the results are obtained by applying the equation on a simplified social graph where the edges correspond to active links connecting the nodes only when there is a retweet/mention of the user id of the original contributor. This excludes the score (howsoever miniscule they may be) originating from the links pointing to the central node which seldom interact. Thus the analysis is more on live data which points to the more correct result set.

5. RESULTS

As per our findings, the most influential person who shared information related to H1N1 during the timeframe under consideration, medya_sondakika has 1909 Retweets (Refer figures listed in Table 1) and has roughly 9K followers. The twitter handle Timesofindia has 4.15 million followers and yet the H1N1 related news/ information shared by the id has not been retweeted the most.

We have listed the top 10 twitter handles ranked as per the improved PageRank algorithm. The table also shows the number of followers the user has.

Table 1. Statistics based on the Retweet relationship

S No	Twitter Handle	Retweets	Followers
1	medya_sondakika	1909	9034
2	tuhafamagercek	289	848431
3	anadoluajansi	55	317026
4	about_electric	41	77547
5	Timesofindia	40	4169128
6	ronankelly13	38	716
7	h1n1_1	35	656
8	tevhidigundem	32	6700
9	hurriyet	30	2203119
10	hussamrabialgha	25	81

This goes to show that the dynamism of Retweet is the real measurement of quantifying the influence rather than the static attributers form network theory used by PageRank method. And hence the PageRank method is not most suited for identifying the influential users in the online Social Network.

6. CONCLUSION

The social media marketers are constantly buzzing with talk about influencers who are generally targeted for fast propagation of idea or brand promotion. Their profile of being an expert in the given domain induces cascaded actions among the people they influence.

We have studied that as against the popular notion that the users with maximum social connection might not be the actual Influencers. The PageRank algorithm does not take into account the enthusiasm of users actually contributing in the information propagation by retweeting the posts shared by the content generator. More the number of retweets by multiple users in the follower graph better is the Influencing capability of the user.

The current work may be improved and the result set could be made more accurate by taking into account the other factors that make a social network user the influencer like his attitude (perhaps , could be captured through his profile contents) , and the ability to create opinion or promote some sentiment. This would require more careful scrutiny than counting diffusion of tweets on the network.

7. REFERENCES

[1] M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM), (2010)

[2] D. Tunkelang. A Twitter Analog to PageRank. Available at http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-PageRank

[3] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In Proceedings of the 3rd conference on Online social networks, WOSN'10, 2010

[4] Daniel M.Romero, Sitaram Asur, Influence and Passivity in Social Media, ACM

[5] Watts D, Dodds P (2007) Influentials, Networks, and Public Opinion Formation. Journal of Consumer research 34(4): 441–458.

[6] Juyup Sung, Seunghyeun Moon, Jae-Gil Lee. The influence in Twitter – Are they really influenced. Springer International Publishing. 0302-9743

[7] Boyd D, Golder S, Lotan G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. Hawaii International Conference. PP 1-10 Honolulu, USA 2010

[8] S. Aral and D. Walker. Identifying Influential and Susceptible Members of Social Networks. Science, vol. 337, pp. 337-341, 2012

[9] J Leskovec, L Adamec, B Huberman. The Dynamics of Viral Marketing, 2005

[10] Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? In the proceedings of the 19th International Conference on WWW'10, pp 591-600

[11] E Adar and L Adamic. Tracking information epidemics in blogspace. Web Intelligence, pp207-214, IEEE 2005

[12] Sun E, Rossen I, Marlow C, Lento T. Gesundheit! modeling contagion through facebook news feed. In the proceedings of Third International AAAI Conference on Weblogs and Social Media, 2009

[13] Weng J, Lim E.P., Jiang H, He Q. Twitterrank: Finding Topic-Sensitive Influential Twitterers. In WSDM, pp261-270, 2010

[14] E. Bakshy et al. Everyone's an influencer: quantifying influence on twitter,‖ in WSDM '11 Proceedings of the fourth ACM international conference on Web Search and data mining, 2011

[15] Amy N. Langville, Carl D. Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. The American Mathematical Monthly Vol. 115, No. 8 (Oct., 2008), pp. 765-768

[16] The original PageRank paper by Google's founders Sergey Brin and Lawrence Page - http://www-db.stanford.edu/~backrub/google.html