MEASURING NODE IMPORTANCE: A MULTI-CRITERIA APPROACH

Leila Weitzel[1, 3]
Universidade Federal do Pará[1]
Marabá, Pará, 68501-970, Brazil[1]

Paulo Quaresma[2]
Universidade de Évora[2]
Évora, 7000, Portugal[2]

José Palazzo M. de Oliveira[3]
Universidade Federal Rio Grande do Sul[3]
Porto Alegre, Rio Grande do Sul, 91501-970, Brazil[3]

ABSTRACT

Social Networks are created whenever people interact with other people. Online Social Networks, like Twitter and Facebook gained considerable popularity in the last years. With the popularity of Web applications and increasing reliance on mobile handheld devices, socializing over the Web has become an integral part of our daily lives. Twitter is a social networking and micro-blogging service; it creates several new interesting social network structures. In this sense, our main goals are study and analyze topological structure of retweet network and investigate the power of retweet mechanism. The findings suggest that relations of "friendship" at Twitter are important but not enough. Still, the centrality measures of a node importance do not show how important users are. We uncovered some other principles that must be studied like, homophily phenomenon, the tendency of individuals to associate and bond with similar others.

KEYWORDS

Social Network Analysis, Twitter, Retweet, Node Importance.

1. INTRODUCTION

The recent proliferation of Internet social media applications and mobile devices has made social connections more accessible than ever before. In the last few years the number of users of online social networks like Facebook, MySpace and Twitter gained considerable popularity and grown at an unprecedented rate (Kim et al. 2010). Twitter is a social networking and micro-blogging service. Twitter allows users to communicate and stays connected through the exchange of short messages, called tweets. These posts are brief (up to 140) and can be written or received with a variety of computing devices, including cell phones. Twitter creates several interesting social network structures. The most obvious network is the one created by the "follows" and "is followed by" relationships without approval, these create a different type of ties, where the directionality of tie is important (i.e. who is following whom)(Hansen et al. 2011). Unlike most other online social networking sites (like Facebook, etc), following on Twitter is not a mutual relationship. Any user can follow you and you do not have to follow back. Relationships at Twitter that are reciprocated are different and perhaps stronger than those that are not, and they are called "friendships". Twitter users follow someone, mostly because they are interested in the topics the user publishes in tweets, and they follow back because they find they share similar topic interest. When a user posts a message, if other users like it, they repost it (or "Retweet" - RT), and a large number of users can be potentially reached by a particular message. Based on this context, we looked at the problem through two perspectives: first, studying topological structure of user´s RT alter and ego-network, second, ranking nodes based on strength of RT ties. In particular, we investigate the influence of "retweeting" mechanism in health information messages context. The outline of this paper is

as follows: Section 2 presents the background of the research in the context of social network analysis; Section 3 we explain the data extraction technique and network modelling approach and data analysis and the methodological approach; Section 4 we discuss the results and future works and finally we present the acknowledgment, and References.

2. BACKGROUND

One common type of social analysis is the identification of communities of users with similar interests, and within such communities the identification of the most "influential" users. Efforts have been made to measuring the influence and ranking users by both their importance as hubs within their community and by the quality and topical relevance of their post. Some of these efforts are: (Balkundi & Kilduff 2005; Bar-Ilan & Peritz 2009; Bongwon Suh et al. 2010; Boyd et al. 2010; Cha et al. 2010; Gayo-Avello 2010a; Gayo-Avello 2010b; Gruhl et al. 2004; Nagarajan et al. 2010; Nagle & Singh 2009; Pal & Counts 2011; Romero et al. 2011; Sakaki & Matsuo 2010; Sousa et al. 2010; Welch et al. 2011; Yamaguchi et al. 2010; Ye & Wu 2010; Kwak et al. 2010). Most of these researches are based on: follower, tweet and mention count, co-follower rate (ratio between follower and following), frequency of tweets/updates, who your followers follow, topical authorities. Centrality measures such as Indegree/Outdegree, Eigen Vector, Betweenness, Closeness, PageRank (Page et al. 1999) and others have been used to evaluate node importance too. Each one of this metrics evidences a class of issue. For instance, Betweenness Centrality represents a node that occurs in many shortest paths among other nodes; this node is called "gatekeeper" between groups node. Closeness Centrality is the inverse of Average Distance (geodesic distance). Closeness reveals how long it takes information to spread from one node to others. Eigen Centrality measures take into account Hub-centrality (out links) and Authority-Centrality (in links). According Bonacich (Bonacich 2007), "Eigenvector Centrality can also be seen as a weighted sum of not only direct connections but indirect connections of every length, thus, it takes into account the entire pattern in the network. These measures are especially sensitive to situations in which a high degree position is connected to many low degree or vice-versa." Nevertheless, sometimes we must take node importance into full consideration based on several criteria that incorporate more global information. Evaluating node importance with a single metric can be considered incomplete and limited as it couldn't capture the specific differences among nodes.

3. RESEARCH METHODOLOGY

3.1 Dataset Collection and Network Topological Structure

In this section we discuss about data collection by enlighten our data crawling methodology, the applied statistical data analysis, the statistical inference and its goal and also we detail the topology of two ego-networks. We have crawled with NodeXL (Smith, M et al. 2007) about 152 Tweeter's users in accordance with link "how to follow" and later "browse interests", and then we searched for topic "health" during March 2011. Afterward, we selected 100 users that have a website or a blog associated to health subject. From each of 100 seeds users, we extracted about 200 RT per user in a total of almost 4350 RT. Kwak et al. (Kwak et al. 2010) demonstrated that the median number of tweets per user stay between 100 and 1000, indicating that our RT size sample (200 RT) is suitable. The RTs are marked with characters RT or via @ + "screenname", therefore, we extracted either both replay tweets and mention. The RT ego-network G_{RT} was modeled as a direct graph where each node $u \in V$ represents a user (total 1237 nodes) and each edge $a_k = (u_i, u_j) \in A$ represents a RT relationship (total 1409 edges), i.e., an edge a_k from u_i to u_j stands that user u_i "retweet" user u_j. Every edge $\boldsymbol{a_k} \in \boldsymbol{G_{RT}}$ has an associated weight $\boldsymbol{w_{a_k}}$ defined by: $w_{a_k} = \frac{\sum RT_j}{RT_{max}} + \alpha$

Where $\sum \boldsymbol{RT_j}$ is the retweet count of $\boldsymbol{u_j}$, $\boldsymbol{RT_{max}}$ is the maximum number of retweet. The parameter $\boldsymbol{\alpha}$ is a sort of discount rate representing Twitter relationships: (a) following, (b) follower, (c) who are reciprocally connected and (d) when relationships - follower or following - are absent between users. Using this notation, if an individual $\boldsymbol{u_i}$ is a "follower" of $\boldsymbol{u_j}$, then $\boldsymbol{\alpha} \approx 0.07$ and if is "following" then $\boldsymbol{\alpha} \approx 0{,}14$, if is both follower

and following then $\boldsymbol{\alpha} \approx 0{,}15$ and if the relationship is absent then $\boldsymbol{\alpha} \approx 0{,}64$. The parameter $\boldsymbol{\alpha}$ was calculated in accordance with relationship ratio in a dataset (Figure 1). The parameter $\boldsymbol{\alpha}$ intend to discount the weight of the follow phenomenon, since many celebrities and mass media have hundreds of thousands of followers.

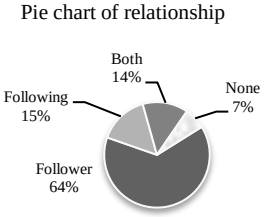

Figure 1: Pie chart of RT relationships in a dataset.

3.2 Ranking Node Approach

Most of such approaches discussed herein, rely on only single measure to determine influence among users. Jianwei Wang et al. (Jianwei Wang et al. 2008) described a method to discover influential users in Twitter. The work is really interesting; these researchers proposed a balanced method for evaluating node importance based on three Centrality measures: Degree, Betweenness and Closeness. They tested their method in a real-world network, the sexy relation network of the AIDS. Nevertheless, they have not taken into account any other additional information such as weighted ties. In this sense, motivated by their research our methodological approach is based on combining standard metrics with adjustable weighted parameters, considering not only the topological importance of a node, but also the strength of ties, i.e. the retweet power.

F-measure is generally accepted at Information Retrieval as evaluation performance methods and by far the most widely used. It has been past more than 15 years since the F-measure was first introduced by van Rijsbergen (Rijsbergen 1979) . He states, the F-measure (F) combines Recall (R) and Precision (P) in the following form:

$$F(R,P) = \frac{(\beta^2+1)P * R}{\beta^2 P+R} = \frac{1+\beta^2}{\frac{\beta^2}{R}+\frac{1}{P}} \quad \text{where } (0 \leq \beta \leq \infty)$$

Where β is a parameter that controls a balance between P and R. When $\beta = 1$ F comes to equivalent to the harmonic mean of P and R. If $\beta > 1$, F becomes more recall-oriented and if $\beta < 1$, it becomes more precision oriented $F_0 = P$. Hence, we assume the importance of each centrality measure Betweenness - BC, Closeness - CC, PageRank - PRANK (Page et al. 1999) and Eigen-Vector - EC. Then, Let Rank be a linear combination of metrics with associated weight defined by: $\text{Rank} = \frac{\sum_{k=1}^{m} w_k}{\sum_{k=1}^{m} \frac{w_k}{x_k}}$ and $(\sum_{k=1}^{m} w_k) = 1 \iff (\delta + \beta + \theta + \gamma) = 1$ is the weighted parameter and x_k is a set of four measures: $\{BC\,;\, CC\,;\, EC\,;\, PRANK\}$. The "control balance" is used in the same way as in F-measure. The first hypothesis is all of parameters have same value (line one in Table 1): $\delta = 0.25$; $\beta = 0.25$; $\theta = 0.25$; $\gamma = 0.25$, and afterward each of these is weighted according each line of Table 1. The Table 2 displays the top 20 ranked nodes using our approach the five weighted schemas (Table 1). Therefore, depending on issue involved, the weight can be modified to rank nodes. For example, if we want to identify importance of the node which acts like a "bridge[1]", in that case, it must be used the BC weighted scheme (line two in Table 1 and the result in Table 2 column 3). In order to gain insight of the ranking method, we associate each position (the top 20) with a value following this: the first top position received 20 points, the second position nineteen, and successively decrease one unity until the last one, that received one point. We perform that method for each of column results in Table 2.Then, we compute the sum of all nodes individually and the results of the recurring top 20 are displayed in Figure 1. We identified the relationships between top 20 recurring (Figure 1) and the users who replayed their tweet. As we discuss herein we had have hypothize that reciprocal relationships are perhaps stronger (at Twitter)

[1] An actor that connects two separated cliques.

than those that are not. However, that hypothesis does not prove itself. We do not ascertain this finding in the top 20 rank, the relationship was mostly follow.

Table 1. Weighted parameter

Measure / Weight	δ	β	θ	γ
1. Equal weighted	**0.25**	**0.25**	**0.25**	**0.25**
2. BC weighted	**0.7**	**0.1**	**0.1**	**0.1**
3. CC weighted	**0.1**	**0.7**	**0.1**	**0.1**
4. EC weighted	**0.1**	**0.1**	**0.7**	**0.1**
5. Prank weighted	**0.1**	**0.1**	**0.1**	**0.7**

Table 2. Weighted parameter

Top 20	Equal Weighted	BC Weighted	CC Weighted	EC Weighted	PRANK Weighted
1	**UC19**	**UC19**	**UC19**	**UC2**	**UC19**
2	**UC2**	**UC48**	**UC2**	**UC19**	**UC2**
3	**UC14**	**UC14**	**UC14**	**UC14**	**UC14**
4	**UC48**	**UC53**	**UC48**	**UC96**	**UC48**
5	**UC96**	**UC2**	**UC96**	**UC48**	**UC96**
6	**UC53**	**UC16**	**UC53**	**UC39**	**UC53**
7	**UC39**	**UC96**	**UC16**	**UC89**	**UC39**
8	**UC16**	**UC81**	**UC71**	**UC71**	**UC16**
9	**UC17**	**UC17**	**UC17**	**UC17**	**UC17**
10	**UC89**	**UC71**	**UC39**	**UC75**	**UC89**
11	**UC71**	**UC37**	**UC89**	**UC3**	**UC71**
12	**UC75**	**UC39**	**UC3**	**UC16**	**UC75**
13	**UC3**	**UC89**	**UC75**	**UC53**	**UC81**
14	**UC81**	**UC95**	**UC81**	**UC95**	**UC3**
15	**UC95**	**UC75**	**UC95**	**UC81**	**UC95**
16	**UC88**	**UC88**	**UC88**	**UC88**	**UC88**
17	**UC37**	**UC3**	**UC37**	**UC24**	**UC37**
18	**UC24**	**UC24**	**UC24**	**UC37**	**UC24**
19	**UC100**	**UC100**	**UC100**	**UC100**	**UC100**
20	**UC15**	**UC15**	**UC15**	**UC15**	**UC15**

Afterward, we evaluated the recurring top 20 with: (a) our methodology with RT capability (displayed in Figure 1) and the results are shown in column 1-4 in Table 3 and (b) without RT weighted links, they are respectively calculated from four methods: BC, CC, EC and Prank and the results are shown in column 5-8 in Table 3.

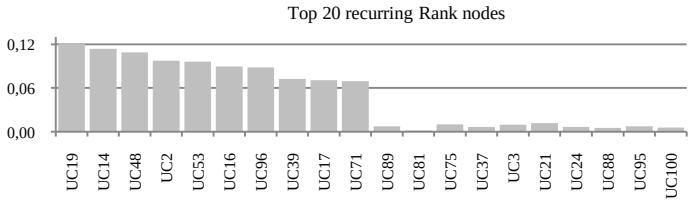

Figure 1. Bar chart of recurring top 20 nodes

As shown in Table 3, the nodes with highest RT values gained enhanced position, for example, nodes UC19, UC48 and UC53. In spite of their location are not more central in the network, their importance was more significant as a credible source. We uncovered that top k ranked nodes do not necessarily have highest values of RT. The UC2 node act like a bridge ("gatekeeper") and it is in fourth ranking in Figure 1. This findings suggests that centrality measures associated with our weighted ties approach controls the node importance, i.e., despite top k node in Table 3 column one are considered a key nodes, other influential nodes play an important role as being a believable information provider too; for instance, UC19 and UC48.

Table 3. Top 10 ranked nodes with two approaches

Rank	user ID without RT weighted ties	user ID with RT weighted ties
1	UC37	UC19
2	UC14	UC14
3	UC2	UC48
4	UC16	UC2
5	UC19	UC53
6	UC88	UC16
7	UC100	UC96
8	UC53	UC39
9	UC81	UC17
10	UC24	UC71

4. DISCUSSION

Our goal was mostly to analyze and evaluate the power of retweeting. Hence, in order to address this goal, we proposed a topological network structure to represent the strength of RT; and also a weighted parameter to estimate this influence. Based on this approach, we ranked the nodes by its authority and we tested the F-measure method to control the top ranked positions. As a case of study we used Twitter community relationships; particularly the ego-network relationships who have an interest in healthcare. The experimental results offer an important insight of the relationships among Twitter users. The findings suggest that relations of "friendship" or follows are important but not enough to find out how important nodes are. Further, the study also gives us a clear understanding of the how measure selection can affect the rank. Choose the most appropriate measure depends on what we want to represent; for example, in/out degree, Eigen-Vector and even PageRank operate look alike “edges counts” as the “popularity” measures. Conversely, closeness and betweenness centrality measures specify the key position that a node occupies in a graph. The results also shown that centrality measures associated with our weighted ties approach controls suitably the node rank. Moreover, we have observed that in Twitter community, trust plays an important role in spreading information; it motivates a user to reply messages to other users, thus, the culture of “Retweeting” demonstrate the potential to reach trust for dissemination of information. We uncovered other some principles that must be studied like for instance, homophily phenomenon. According to Macpherson (McPherson et al. 2001), homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people. Homophily suggests that people with similar backgrounds with regard to their socio-demographic, behavioral and intrapersonal and others characteristics tend to established ties. Thus, we plan to expand our study by incorporating this approach. Last but not least, we aimed to extend our experiments combining topological structure, others relationships besides RT and homophily and evaluated the results with other approaches proposed in literature.

ACKNOWLEDGEMENT

Leila Weitzel and José Palazzo M. de Oliveira would thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brazilian National Research Council) for the partial financial support of this work.31/2010-9. Leila Weitzel would also like to thank Universidade Federal do Pará for providing partial financial support of this work.

REFERENCES

Balkundi, P. & Kilduff, M., 2005. The ties that lead: A social network approach to leadership. *The Leadership Quarterly*, 16(6), p.941-961.

Bar-Ilan, J. & Peritz, B.C., 2009. A method for measuring the evolution of a topic on the Web: The case of “informetrics”. *Journal of the American Society for Information Science and Technology*, 60(9), p.1730-1740.

Bonacich, P., 2007. Some unique properties of eigenvector centrality. *Social Networks*, 29(4), p.555-564.

Bongwon Suh et al., 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. Social Computing (SocialCom), 2010 IEEE Second International Conference on. p. 177-184.

Boyd, D., Golder, S. & Lotan, G., 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Hawaii International Conference on System Sciences*. Los Alamitos, CA, USA: IEEE Computer Society, p. 1–10.

Cha, M., Haddadi, H. & Gummadi, P.K., 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *International Conference on Weblogs and Social Media*.

Gayo-Avello, D., 2010a. Detecting Important Nodes to Community Structure Using the Spectrum of the Graph. *Cornell University Library*.

Gayo-Avello, D., 2010b. Nepotistic relationships in twitter and their impact on rank prestige algorithms. *Arxiv preprint arXiv:1004.0816*.

Gruhl, D. et al., 2004. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, p. 491-501.

Hansen, D., Smith, M.A. & Shneiderman, B., 2011. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, MA, USA: Morgan Kaufmann.

Jianwei Wang, Lili Rong & Tianzhu Guo, 2008. A New Measure of Node Importance in Complex Networks with Tunable Parameters. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on. p. 1-4.

Kim, W., Jeong, O. & Lee, S.W., 2010. On social Web sites. *Information Systems*, 35(2), p.215-236.

Kwak, H. et al., 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*. Raleigh, North Carolina, USA: ACM, p. 591-600.

McPherson, M., Smith-Lovin, L. & Cook, J.M., 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27, p.415–444.

Nagarajan, M., Purohit, H. & Sheth, A., 2010. A Qualitative Examination of Topical Tweet and Retweet Practices. In *ICWSM 2010*. International AAAI Conference on Weblogs and Social Media. Washington, DC.

Nagle, F. & Singh, L., 2009. Can Friends Be Trusted? Exploring Privacy in Online Social Networks. In *2009 International Conference on Advances in Social Network Analysis and Mining*. 2009 International Conference on Advances in Social Network Analysis and Mining (ASONAM). Athens, Greece, p. 312-315.

Page, L. et al., 1999. *The PageRank Citation Ranking: Bringing Order to the Web.*, Stanford InfoLab.

Pal, A. & Counts, S., 2011. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong, China: ACM, p. 45-54.

Rijsbergen, C.J. van, 1979. *Information retrieval* 2^{o} ed., London: Butterworths.

Romero, D.M. et al., 2011. Influence and passivity in social media. In *Proceedings of the 20th international conference companion on World wide web - WWW '11*. the 20th international conference companion. Hyderabad, India, p. 113.

Sakaki, T. & Matsuo, Y., 2010. How to Become Famous in the Microblog World. *2010*.

Smith, M et al., 2007. NodeXL: a free and open network overview, discovery and exploration add-in for Excel.

Sousa, D., Sarmento, L. & Mendes Rodrigues, E., 2010. Characterization of the twitter @replies network. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents - SMUC '10*. the 2nd international workshop. Toronto, ON, Canada, p. 63.

Welch, M.J. et al., 2011. Topical semantics of twitter links. In *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong, China: ACM, p. 327-336.

Yamaguchi, Y. et al., 2010. TURank: Twitter User Ranking Based on User-Tweet Graph Analysis. In *Web Information Systems Engineering – WISE 2010*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, p. 240-253.

Ye, S. & Wu, S., 2010. Measuring Message Propagation and Social Influence on Twitter.com. In *Social Informatics*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, p. 216-231..