Identification of Most Influential Spreaders in Twitter Social Network using Modified K Core Decomposition in Distributed Environment

ABSTRACT

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:

. 2018. Identification of Most Influential Spreaders in Twitter Social Network using Modified K Core Decomposition in Distributed Environment. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Social Networks have gained remarkable popularity in the past few decades. A huge number of people are using social networking sites like Facebook, Twitter, Google+, LinkedIn etc. The heavy reliance on social networking sites causes them to generate massive data. Online social networks play a major role in the diffusion of information [3]. Since social network has great influence on society, the recent focus is on extracting valuable information from this huge amount of data. Events, issues, interests etc. happen and evolve very quickly in social networks and their capture, understanding, visualization, and prediction are becoming critical expectations from both end-users and researchers. Therefore researchers in recent years have developed a variety of techniques and models to capture information diffusion in online social networks, analyze it, extract knowledge from it and predict it [1, 10].

Micro-blogging is an emerging form of communication. One of the most notable micro-blogging services is *Twitter*. It allows *twitterers* to publish tweets (with a limit of 140 characters). *Twitter* employs a social-networking model called "following", in which each *twitterer* is allowed to choose who she wants to follow without seeking any permission. Conversely, she may also be followed by others without granting permission first. In one instance of "following" relationship, the *twitterer* whose updates are being followed is called the "friend", while the one who is following is called the "follower". *Twitter* has gained huge popularity since the first day that it was launched [22]. It has also drawn increasing interests from research community [5]. Many analysis have been done on the dataset of twitter including identification of the most influential spreader [25, 31]. We discuss the state-of-art in details in the next section.

Information diffusion is a vast research domain and has attracted research interests from many fields, such as physics, biology, etc. The diffusion of innovation over a network is one of the original reasons for studying networks and the spread of disease among a population has been studied for centuries [2, 13, 15]. The knowledge of the spreading pathways through the network of social interactions is crucial for developing efficient methods to either hinder spreading in the case of diseases, or accelerate diffusing in the case of information dissemination. The information diffusion model is based on three fundamental questions: *(i) Which pieces of information diffuse the most, (ii) How, why and through which paths information is diffusing, and will be diffused in the future, (iii) Which members of the network play important roles in the spreading process?*

In this paper we are mainly focusing on the third question. Identifying the most influential spreaders in a network is critical for ensuring efficient diffusion of information, which allows to control the outbreak of any kind of epidemic [6, 27], utilize the limited resources to optimize the information propagation [20], successful e-commercial advertisements [18, 21], optimize the use of limited resources to facilitate information propagation [4] etc. In large social network graph the nodes having the largest degree are often considered as the most influential spreader [26]. However, recent studies have shown that the most connected people are not the most influential spreader. Kitsak et al. [16] showed that the best spreaders are not necessarily the most connected people in the network. They found that the most efficient spreaders are those located within the core of the network as identified by the k-core decomposition analysis [29]. Moreover, many centrality measures have been proposed to identify the most influential spreaders on a social network. All these metrics are mainly of two kinds: global and local metrics. Local metrics like degree centrality and K-core decomposition are of simple complexity but are less effective i.e. they fail to find the most influential spreaders accurately since of omitting the global structure of the network. Global metrics such as betweenness centrality and closeness centrality perform well in the identification of the influential nodes. However, these global measures incur high computational complexity which make them

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

infeasible to use in case of large graph [12, 24]. The state-of-the-art of different methods of finding most significant nodes in a network has been discussed in the paper of Lü et al. [20] reviewed the state of the art of different proposed methods and approaches dealing with detection of vital nodes in complex networks. Moreover, available methods are compared based on the different nature of the network.

Liu et al. [19] provided an improved method of using k-shell decomposition by removing the redundant links which have a low diffusion significance. Another improved method of k-shell decomposition was proposed by Wang et al.[30] who used k-shell iteration factor to evaluate the influence capability of a node. They used the iteration information of k-shell decomposition to distinguish among the nodes with the same maximum k-shell value.

Node of the existing methods take the user info into consideration while identifying the most influential spreaders which obviously play vital role for information spreading. On top of that, with the ubiquitous use of social network, social networks are producing large amount of data. The higher the number of nodes and their friend-follower relationship data is available, the more the graph becomes suitable to identify the real most influential spreaders. However, we need special kind of

We propose a variant of k-core decomposition which is distributed and also incorporates the node information. Summarizing, our main contributions are:

-

2 TECHNICAL BACKGROUND

2.1 Centrality Measurement

2.1.1 *Degree Centrality.*

2.1.2 *Closeness Centrality.*

2.1.3 *Betweenness Centrality.*

2.1.4 *Eigen Vector Centrality.*

2.2 K-Core Decompostion

2.3 Susceptible-Infected-Recovered (SIR) Model

The SIR model is one of the simplest compartmental models in epidemiology. In an SIR model, every mode of a network can be at one of the following three states:

- S (Susceptible): These denotes those people who have not been infected with the disease yet. However, they are not immune to it and therefore they are under thread of being infected with the disease in the future.
- I (Infected): These are people who have already been infected with the disease. Moreover, the infected people can transmit the disease to the susceptible neighbors with a probability of β.
- R (Recovered): These people have been recovered from the disease with a probability of γ and are immune now. Therefore, they are no longer under any thread to be infected with the disease in future.

The SIR model runs the simulation of disease diffusion based on the topology of the network and the probability parameters β and γ. The simulation stops when the network has no more infected nodes. This SIR model is capable of finding the most influential spreaders based on their sreadability.

3 RELATED WORK

An online social network (*OSN*) results from the use of a social network site (*SNS*) that allows the users to publish messages and connect to other users which result in creating social relationships. An OSN is formally represented by a graph, where nodes are users and edges are relationships that can be either directed or undirected. In this graph the nodes play an important role to disseminate information. Finding the most influential spreader in an OSN has caught attention of the researchers. Recently more and more attentions have been paid to microscopically study the *spreadability* for each node. The knowledge of node *spreadability* is crucial for developing efficient methods to either decelerate spreading in the case of diseases, or speed up spreading in the case of information flow. Moreover, it can be helpful for identifying the initial spreader of certain disease or information.

Though the most connected nodes (hubs) and the nodes with high betweenness centrality are commonly believed to be the most influential spreaders in networks, the k-core (also called k-shell) method is found to perform better in identifying the best individual spreaders [16**?**]. Basically, the principle of the k-core decomposition is to assign a core index k_s to each node such that nodes with the lowest values are located at the periphery of the network while nodes with the highest values are located in the center of the network. We shall discuss the details about the algorithm shortly.

Cataldi et al. [**?**] proposed to use the well known PageRankalgorithm [**?**] to assess the distribution of influence throughout the network. The *PageRank* value of a given node is proportional to the probability of visiting that node in a random walk of the social network, where the set of states of the random walk is the set of nodes. Both the methods only exploit the topology of the network, and ignore other important properties, such as nodes' features and the way they process information. Lü et al. [**?**] proposed the *LeaderRank* algorithm to identify influential spreaders in directed networks, which is a simple variant of *PageRank*, namely a so-called ground node connected with every other node by a bidirectional link is introduced into the original network, and then the standard random walk process is applied to dig out influential spreaders. Li et al. [**?**] further improved the *LeaderRank* by allowing nodes with more fans get more scores from the ground node. With almost the same converging speed (we have checked by simulations), this so-called $WeightedLeaderRank$ performs better than *LeaderRank*.

Romero et al. [**?**] develop a graph-based approach similar to the well known HITS algorithm, IP(i.e. *Influence-Passivity*), that assigns a relative $influence$ and a $passivity$ score to every users based on the ratio at which they forward information. However, no individual can be a universal influencer, and influential members of the network tend to be influential only in one or some specific domains of knowledge. Therefore, Palet al. [25] developed a non-graph based, topic-sensitive method. To do so, they define a set of nodal and topical features for characterizing the network members. Using probabilistic clustering over this feature space, they rank nodes with a within-cluster ranking procedure to identify the

most influential and authoritative people for a given topic. Weng et al. [31] also develop a topic-sensitive version of the *PageRank* algorithm dedicated to Twitter, *TwitterRank*. They presented the phenomenon of *homophily* in a community of Twitter. By making use of this phenomenon, a PageRank-like algorithm, called *TwitterRank*, is proposed to measure the topic-sensitive influence of the *twitterers*.

All these methods described are summarized in Table 1. We can see that none of the above approach is distributed. Most famous approach for finding influential spreader is the k-core decomposition which considers only the network topology. We propose a distributed variant of the algorithm which also considers user info like no. of followers, no. of friends, whether the user is verified etc.

Table 1: Summary of influential spreaders identification methods

Algorithm	Network Topology	User Info	Topic Info	Distributed
k-core decomposition	Y			
PageRank	Y			
Topic-sensitive PageRank	Y		Y	
IP		Y		
Topical Authorities		Y	Y	

4 EXPERIMENTAL SETUP

4.1 Datasets

5 EVALUATION

6 EVALUATION METRICS

6.1 Rank Correlation Coefficient

In general, correlation analyses are bi-variate analyses that measure the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and −1. A value of ±1 indicates that there exists a perfect degree of association between the comparing two variables. As the correlation coefficient value goes towards 0, this relationship between the two variables gets weaker. The ± sings of the coefficient indicated he direction of the relationship; a + sign indicates a positive relationship while a − sign indicates a negative relationship.

In Section **??**, we measure the correlation of the ranked list of users based on their spreadability generated by our modified K core decomposition against the rankings generated by other methods. We use two non-parametric rank correlations: Kendall's tau and Spearman's rank correlation coefficient.

6.1.1 Kendall Tau Correlation Co-efficient. The Kendall tau rank correlation co-efficient is used to test the similarities in the ordering of data when it is ranked by quantities. While other types of correlation coefficients use the observations as the basis of the correlation, Kendall's correlation coefficient uses pairs of observations and determines the strength of association based on the patter on concordance and discordance between the pairs. Assume that L_1 and L_2 are the two rankings that are to be compared. Then Kendall analysis takes the following two properties into consideration:

- **Concordant**: Any pair of items (x_1, y_1) in L_1 and (x_2, y_2) in L_2 are considered as concordant if and only if they meet one of the following two conditions:
  - $(rank\_in\_L_1(x_1) > rank\_in\_L_2(x_2)$ and $rank\_in\_L_1(y_1) > rank\_in\_L_2(y_2))$
  - $(rank\_in\_L_2(x_2) > rank\_in\_L_1(x_1)$ and $rank\_in\_L_2(y_2) > rank\_in\_L_1(y_1))$
- **Discordant**: Any pair of items (x_1, y_1) in L_1 and (x_2, y_2) in L_2 are considered as discordant if and only if they meet one of the following two conditions:
  - $(rank\_in\_L_1(x_1) > rank\_in\_L_2(x_2)$ and $rank\_in\_L_1(y_1) < rank\_in\_L_2(y_2))$
  - $(rank\_in\_L_2(x_2) > rank\_in\_L_1(x_1)$ and $rank\_in\_L_2(y_2) < rank\_in\_L_1(y_1))$

Kendall Tau correlation co-efficient is denoted by τ. If L_1 and L_2 are two different rankings with n similar elements, $N(C)$ and $N(D)$ represent the number of concordant and discordant pair respectively, then τ can be calculated using the following equation:

$$\tau(L_1, L_2) = \frac{N(C) - N(D)}{\frac{1}{2}n(n-1)} \tag{1}$$

6.1.2 Spearman's Rank Correlation Co-efficient. Spearman's Rank correlation coefficient, R_s is a technique which can be used to summarize the strength and direction (negative or positive) of a relationship between two variables. The result will always be within 1 and minus 1. The closer R_s is to +1 or −1, the stronger the likely correlation. A perfect positive correlation is +1 and a perfect negative correlation is −1. Assume that L_1 and L_2 are two rankings of same n elements. For any element x, if the rankings of x in L_1 and L_2 are $rank\_in\_L_1(x)$ and $rank\_in\_L_2(x)$ respectively, then the distance of ranks, $d = rank\_in\_L_1(x) - rank\_in\_L_2(x)$. This value is squared to remove any negative values and When written in mathematical notation the Spearman Rank formula looks like this:

$$R_s = 1 - \frac{6\sum d^2}{n^3 - n} \tag{2}$$

6.1.3 Calculating Statistical Significance using Co-efficient Values. Statistical significance is a measure of whether any research outcome are meaningful or not. In the field of hypothesis testing of statistics, the term *null hypothesis* is the default assumption that there is no association or relationship between two measured phenomena [11]. A result has statistical significance when it is very unlikely to have occurred given the null hypothesis [23]. To be specific, a significance level, α is set for the experiment which denotes the probability of the experiment rejecting the null hypothesis, given that the null hypothesis were assumed to be true [8]; and the p-value of a result, p is the probability of obtaining a result at least as extreme, given that the null hypothesis were true. The result is statistically significant, by the standards of the study, when the condition $p \leq \alpha$ holds[7, 9, 14, 17, 28]. for the evaluation of our ranking method, we set the significance level to 5%.

To measure the statistical significance of the result, we use the following formula to compute a z-value:

$$z = \frac{3 \times T\sqrt{n(n-1)}}{\sqrt{2(2N+5)}} \tag{3}$$

where T is the correlation co-efficient measured by the previously described techniques. Using the z-score, an area is found from a z-table. This area value is considered as the p-value, of a result.

6.2 Normalized Discounted Cumulative Gain, *NDCG*

In this paper, we use Normalized Discounted Cumulative Gain, *NDCG* which is one of the widely used techniques to evaluate ranking systems. Let X represents the set of all users who generate a graph ne

have a specific influence value, t is a term in the query and our task is to retrieve top M similar keywords $\{w_1, w_2, \cdots, w_M\}$.

- **Normalized Discounted Cumulative Gain,** *NDCG* For term t, we define cumulative gain at rank position M ($CG^t@M$) as:

$$CG^t@M = \sum_{i=1}^{M} sim_{jcn}(t, w_i) \tag{4}$$

Discounted cumulative gain (*DCG*) penalizes each rating (similarity score) based on its rank in the results.

$$DCG^t@M = \sum_{i=1}^{M} \frac{sim_{jcn}(t, w_i)}{\log(i+1)} \tag{5}$$

NDCG is obtained by dividing *DCG* by Ideal DCG (*IDCG*), which normalizes the gain within $[0, 1]$. *IDCG* is the *DCG* of the best possible results based on the ground truth similarity scores. If for a term t, the top r similar keywords are $w_1^{(I)}, w_2^{(I)}, \cdots, w_r^{(I)}$ based on sim_{jcn} metric, then

$$IDCG^t@M = \sum_{i=1}^{r} \frac{sim_{jcn}(t, w_i^{(I)})}{\log(i+1)} \tag{6}$$

Here r is the number of top keywords with positive similarity score ($r \leq M$).

$$NDCG^t@M = \frac{DCG^t@M}{IDCG^t@M} \tag{7}$$

Finally,

$$NDCG@M = \frac{1}{|X|} \sum_{t \in X} NDCG^t@M \tag{8}$$

- **Mean Average Precision,** *MAP* Let's define a binary similarity function between two terms t_1 and t_2:

$$sim^b_{jcn}(t_1, t_2) = \begin{cases} 1, & \text{if } sim_{jcn} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

For a term t, the number of true positive results upto rank i, $TP^t@i = \sum_{j=1}^{i} sim^b_{jcn}(t, w_j)$.

The total number of true positive results, $TTP^t = \sum_{w \in X} sim^b_{jcn}(t, w)$.

The average precision upto rank M is defined as $AP^t@M$:

$$AP^t@M = \frac{1}{TTP^t} \sum_{i=1}^{M} \frac{TP^t@i}{i} \tag{10}$$

Thereby,

$$MAP@M = \frac{1}{|X|} \sum_{w \in X} AP^w@M \tag{11}$$

6.3 Infection Scale on SIR Model

REFERENCES

[1] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 183–194.

[2] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious diseases of humans: dynamics and control*. Vol. 28. Wiley Online Library.

[3] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 519–528.

[4] Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. 2013. Information and influence propagation in social networks. *Synthesis Lectures on Data Management* 5, 4 (2013), 1–177.

[5] Alex Cheng, Mark Evans, and Harshdeep Singh. 2009. Inside Twitter: An in-depth look inside the Twitter world. *Report of Sysomos, June, Toronto, Canada* (2009).

[6] Reuven Cohen, Shlomo Havlin, and Daniel ben Avraham. 2003. Efficient Immunization Strategies for Computer Networks and Populations. *Phys. Rev. Lett.* 91 (Dec 2003), 247901. Issue 24. https://doi.org/10.1103/PhysRevLett.91.247901

[7] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. 27–28 pages.

[8] Peter Dalgaard. 2008. Introductory Statistics with R. Springer, Chapter Power and the computation of sample size, 155–262.

[9] Jay L Devore. 2011. *Probability and Statistics for Engineering and the Sciences* (8th ed.). Cengage learning. 300–344 pages.

[10] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Commun. ACM* 55, 6 (2012), 70–75.

[11] Norman R. Draper. 2011. The Cambridge Dictionary of Statistics, Fourth Edition by B. S. Everitt, A. Skrondal. *International Statistical Review* 79, 2 (2011), 273–274. https://doi.org/10.1111/j.1751-5823.2011.00149_2.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-5823.2011.00149_2.x

[12] S Fortunato. 2010. Community detection in graphs. Phys. Rep. (2010).

[13] JAP Heesterbeek. 2000. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Vol. 5. John Wiley & Sons.

[14] Valen E Johnson. 2013. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* 110, 48 (2013), 19313–19317.

[15] Matt J Keeling and Pejman Rohani. 2008. *Modeling infectious diseases in humans and animals*. Princeton University Press.

[16] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. 2010. Identification of influential spreaders in complex networks. *Nature physics* 6, 11 (2010), 888–893.

[17] Martin Krzywinski and Naomi Altman. 2013. Points of significance: Significance, P values and t-tests. , 1041–1042 pages.

[18] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. 2007. The Dynamics of Viral Marketing. *ACM Trans. Web* 1, 1, Article 5 (May 2007). https://doi.org/10.1145/1232722.1232727

[19] Ying Liu, Ming Tang, Tao Zhou, and Younghae Do. 2015. Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Scientific reports* 5 (2015), 13172.

[20] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. 2016. Vital nodes identification in complex networks. *Physics Reports* 650 (2016), 1–63.

[21] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics reports* 519, 1 (2012), 1–49.

[22] Sarah Milstein, Ben Lorica, Roger Magoulas, Gregor Hochmuth, Abdur Chowdhury, and Tim O'Reilly. 2008. *Twitter and the micro-messaging revolution: Communication, connections, and immediacy–140 characters at a time*. O'Reilly Media, Incorporated.

[23] Jerome L Myers, Arnold D Well, and JR Lorch. 2010. Developing the fundamentals of hypothesis testing using the binomial distribution. *Research design and statistical analysis* (2010), 65–90.

[24] Mark Newman. 2010. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

[25] Aditya Pal and Scott Counts. 2011. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 45–54.

[26] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical review letters* 86, 14 (2001), 3200.

[27] Romualdo Pastor-Satorras and Alessandro Vespignani. 2002. Immunization of complex networks. *Physical review E* 65, 3 (2002), 036104.

[28] Carol K Redmond and Theodore Colton. 2001. *Biostatistics in clinical trials*. John Wiley & Sons. 35–36 pages.
[29] Stephen B Seidman. 1983. Network structure and minimum degree. *Social networks* 5, 3 (1983), 269–287.
[30] Zhixiao Wang, Ya Zhao, Jingke Xi, and Changjiang Du. 2016. Fast ranking influential nodes in complex networks using a k-shell iteration factor. *Physica A: Statistical Mechanics and its Applications* 461 (2016), 171–181.
[31] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 261–270.