2018 IEEE International Conference on Data Mining Workshops (ICDMW)

The Propagation of Counteracting Information in Online Social Networks: A Case Study

Logan Rohde Somya Mohanty Jing Deng Fereidoon Sadri
Department of Computer Science, University of North Carolina at Greensboro, Greensboro, NC, U.S.A.
Email: ljrohde, mohanty.somya, jing.deng, f_sadri@uncg.edu

***Abstract*—Information propagation in online social networks has drawn a lot of attention from researchers in different fields. While prior works have studied the impact and speed of different information propagation in various networks, we focus on the potential interactions of two hypothetically opposite pieces of information, negative and positive. We experiment the amount of time that is allowed for the positive information to be distributed with wide enough impact after the negative information and different selection strategies for positive source nodes. Our results enable the selection of a set of users based on a limited operating budget to start the spread of positive information as a measure to counteract the spread of negative information. Among different methods, we identify that both eigenvector and betweenness centrality are effective selection metrics. Furthermore, we quantitatively demonstrate that choosing a larger set of nodes for the spread of positive information allows for a wider window of time to respond in order to limit the propagation of negative information to a certain threshold.**

***Index Terms*—information propagation; online social networks; centrality**

I. Introduction

Online social network sites are a place where information, opinions, thoughts, and feelings can be shared to millions of people almost instantaneously. With its 330 million users, of which 100 million are said to be active daily, Twitter has and will continue to play an instrumental role in this new way of information sharing [1]. Studies have been used to analyze the spread of information, sometimes called information propagation [2], through social networks such as Twitter, e.g., [3] [4], [5], [6]. Unlike these prior works, we look at the spread of two counter-acting sets of information propagating through a sub-network of Twitter simultaneously.

It is common for both positive and negative information[1] to spread through online social networks. Vosoughi et al. studied the spread of opposite types of information, although not the interactions among them [7]. There are entities and companies that run into the problem of negative, sometimes false, information about them spreading in online social networks, e.g. [8]. To counteract, positive information can be distributed and spearheaded throughout the network. The question here is how such a maneuver can be optimized with limited resources, e.g., monetary and time budget.

[1]Note that we leave the interpretation of "positive" and "negative" information to actual applications but just use two opposites to demonstrate their potential interactions.

Some prior works chose users' PageRank [3] [4], which require extra user information. In [9], node degree and centrality are used to make selections. Similar work has been proposed by Budak et. al. [10], where the authors compare greedy and degree centrality based node selection to limit the spread of mis-infromation across social networks. Instead, we look at eigenvector and betweenness centrality in this work. Our study is based on a snapshot of a small Twitter user subset pulled from a data set of 855,825 tweets on the topic discussed in [8]. The snapshot is then used to generate a directed graph for information flow. The edges are weighted based on our observed interactions, such as replying to a tweet and being mentioned in a tweet. The directed graph is then used to experiment two opposite sets of information flowing and counteracting with each other. We specifically focus on the positive information source selection strategies such as eigenvector and betweenness centrality.

Section II of this paper dives into the methods used to construct the network. It also takes a deeper look at the propagation model and algorithm that is used. Section III explains how the single and double information propagation experiments are executed and then analyzes the results with discussions. We conclude our work in Section IV.

II. Methodology

In order to observe the propagation of information across a social graph, a user-to-user connectivity network is needed. We explain the details of our work in the following.

A. Graph Construction

The data set of the tweets was collected from Twitter using Twitter's API [11] and it consists of 855,825 tweets. These tweets are searched within several days of the United Airlines incident [1] (search keyword "united"). The raw data contains username, date, retweets, favorites, text, geo, mentions, hashtags, id, and permalink, where each row represents a single tweet.

The graph is then constructed as follows: First, the set of unique usernames was obtained, and a node was assigned to each username. Then each row is read in and edges were created where applicable. Initially, we generated 4 types of edges as explained below. Then we converted multiple edges between a pair of nodes i and j to a single edge (i, j), and computed the probability of propagation along each edge $P(i, j)$.

2375-9259/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDMW.2018.00168

We use the *independent cascade model* [2] for information propagation. In order to help determine the probabilities of influence when running this model, 4 different types of directed edges were created. For each tweet

1) Each replied-to user was connected to the tweet owner. This represents the flow of information from the replied-to user to the tweet owner.
2) Each replied-to user was connected to other replied-to users that had replied to their tweet. This represents the flow of information passing to the users that have previously replied in the current thread of tweets that the tweet owner replied to.
3) The tweet owner was connected to each mentioned user in their tweet. This represents the flow of information from the tweet owner to the users they mention in their tweet.
4) Each user mentioned in the current tweet were connected to each other with a bi-directional edge.

Using these methods to construct a network from the Twitter data, the end result had 479,543 unique nodes and 518,784 unique directed edges (of types 1, 2, 3, and 4). It was found that, of the 479,543 nodes, 243,471 of them were contained in the largest connected component of the network. The rest of the nodes were contained in connected components that ranged from 1 to 45 nodes. There were 190,085 nodes that had no adjacent nodes at all. For our analysis of information propagation, nodes that are not contained in the largest connected component are omitted.

B. Determining Edge Weights and Influence Probability

Towards the determination of information influence probability, each of the edge types (4 different edge types) needed to be associated with corresponding edge weights. More specifically, the weights would influence the probability of propagation of information based on the order of importance of different edge types. The reply to edge (type 1) was considered here to be of highest importance with an edge weight of 1000. This is due to the tweet owner taking the time to hit reply in a thread of tweets, so the tweet owner is clearly absorbing information contained in the original message. The edge with the next heaviest weight is the type 2 edge, with an associated weight of 100. As discussed above, these edges are similar to that of type 1, but they do not carry as much weight since they are from a previous time than that of the current tweet from the tweet owner. The type 3 edge follows up with an associated weight of 10. This edge type does not carry as much weight as type 1 and type 2 edges since when mentioned in a tweet, the mentioned user may or may not end up going to look at the tweet. Last, the edge that carries the least amount of weight is type 4 with a value of 1. This was determined since it will be less likely for information to pass between users that were mentioned in a tweet together.

In the creation of the network, it was found that there are instances where an edge connection between the same two users was repeated. When there are repetitions of the same edge, they could be of the same edge type or different. With this being the case, all edges, including their repeated instances, were stored in the database instead. Now, when the edges were read in from the database, the repeated edges with their respective edge weights were aggregated to form a unique edge with total weight of all similar edges.

In order to calculate each edge's associated influence probability (this term can be found in [2]), a normal distribution curve was used. All total edge weights were used in creating the normal distribution curve. An edge's weight was then inserted into a CDF to calculate the influence probability that information could spread to the destination node, thus infecting it. In this paper, we use the terms influence and infect interchangeably. The mean of the total edge weight distribution is 868.5, while the standard deviation is 1251.2.

C. Propagation models and algorithms

Within the created network of users, each edge now has an associated influence probability. The aim of our study was to analyze the propagation of information through such a network, where the goal is to maximize influence of the information to the nodes within the network. To identify the maximum nodes infected, a percentage was calculated by taking the number of nodes active and dividing it by the number of nodes in the largest connected component - 243,471. Since the rest of the connected components are of 45 nodes or less, we disregard these when calculating the percent of active nodes.

To do this, an independent cascade model was used. Chen et.al. explored the possibility of using the independent cascade model with the purpose being that the diffusion events of each arch in the graph were mutually independent [2]. This was important when running the propagation simulations because if a node were to be infected by multiple edges, the effect would be the same, resulting in a state of infection. Such propagation occurs at each unit time t and continues until our experiment stops.

III. Simulations and Results

As stated previously, the goal of the study is to observe the propagation within the network using multiple approaches towards node selection. For the purpose, we evaluated three key approaches using — 1) Random node, 2) Eigenvector Centrality node, and 3) Betweenness Centrality node selection.

A. Single Propagation

First, we investigate single propagation. In the case of Eigenvector and Betweenness centrality, a single node was chosen from top 5 nodes in both algorithms for start of propagation.

Figure 1 shows the effective spread of information to nodes in the network when a random node is chosen as the seed. With median scores of propagation aggregated across 100 simulations, we observe the spread of nodes starts at $t = 5$ in most cases and expands exponentially with spreading to maximum number of available nodes close to $t = 24$.

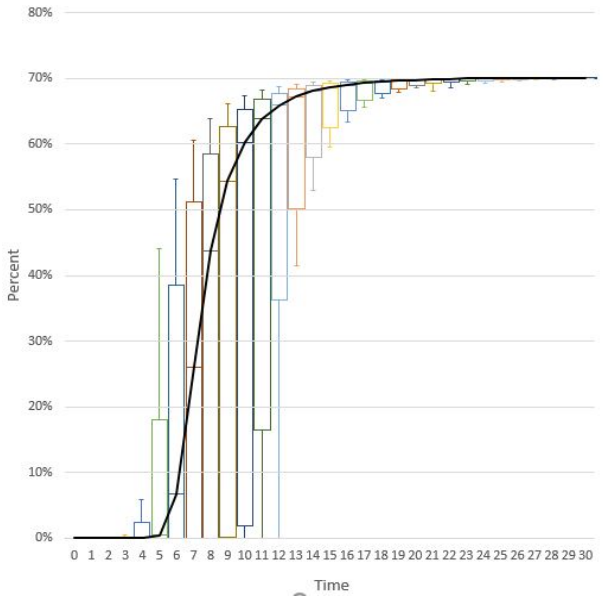

Fig. 1. Percentage of infected nodes within the graph at different time units (t) with Random selection of single start node.

The graph also shows the variance in the observed result, where the exponential growth has a large deviation from the median (1st and 2nd quartile box size) in certain simulations, where the resulting spread is highly dependent on the network characteristics of the random seed node.

It is important to note that, in our simulations, the maximum amount of nodes that could be infected was found to be 170,772 out of the total 243,471 nodes in the largest connected component of the network (a maximum propagation coverage of about 70%). This is due to the fact that those nodes with zero in-degree and positive out-degree characteristics are considered a part of fully connected graph. With zero in-degree, information cannot be propagated to such nodes.

For comparison of random selection to eigenvector, and betweenness centrality methods of node selection, top 5 nodes from eigenvector and betweenness scores of the graph were chosen as seed nodes. Each was simulated through the network to observe median propagation rate (30 simulations conducted per node) to study the effectiveness of the algorithms to spread information through the network. Figure 2 compares the propagation characteristics for each of the algorithms, where both eigenvector and betweenness outperform random node selection by a fair margin. Eigenvector and betweenness are able to start exponential growth in their information propagation much earlier than random selection.

Table I outlines the performance of each algorithm in propagation metrics. Both eigenvector and betweenness based node selection consistently outperform random node selection, with 17.5 % of the nodes reached $0 \leq t \leq 2$, whereas random has a larger variation at $0 \leq t \leq 6$. Similarly, in the second quartile, 35% of the nodes are

TABLE I
COMPARISON OF PROPAGATION TIME (t)

Median Percent infected (p)	Random Time	Eigenvector Time	Betweenness Time
$0\% \leq p \leq 17.5$	$0 \leq t \leq 6$	$0 \leq t \leq 2$	$0 \leq t \leq 2$
$17.5\% \leq p \leq 35$	$t = 7$	$t = 3$	$t = 3$
$35\% \leq p \leq 52.5$	$t = 8$	$4 \leq t \leq 5$	$t = 4$
$52.5\% \leq p \leq 70$	$9 \leq t \leq 24$	$6 \leq t \leq 20$	$5 \leq t \leq 20$

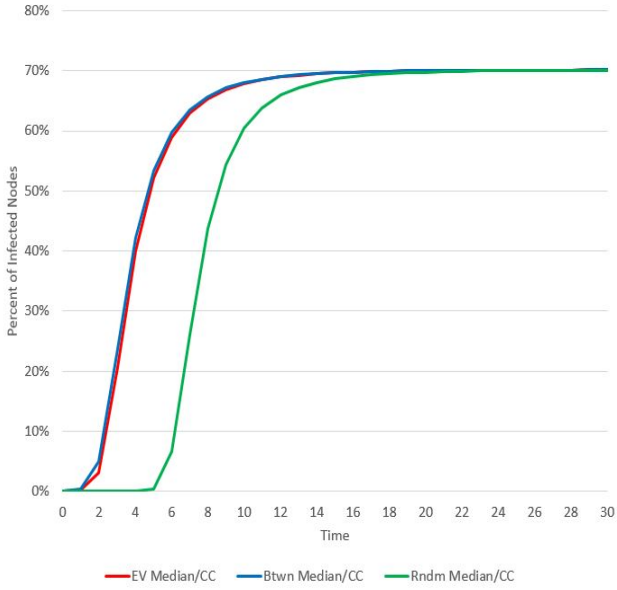

Fig. 2. Comparison of Random, Eigenvector, and Betweenness centrality selection algorithms for median infected nodes at t.

reached by both centrality algorithms at $t = 3$, whereas random is only able to reach the threshold at $t = 7$. For 52.5% of nodes, random takes double the time at $t = 8$ when compared to $t = 4$ for centrality. There is also a large variation in getting to the peak number of nodes at 70% with random taking $9 \leq t \leq 24$, and eigenvector at $6 \leq t \leq 20$ and betweenness at $5 \leq t \leq 20$.

B. Double Propagation

The network was also evaluated for spreading of positive and negative information through the network simultaneously. In the simulation, the negative information gets a head start at propagation with positive information following shortly after at multiple time units t. This is to mimic real-world situations where negative information starts and then positive information or announcements react to those. Observations were recorded for examination of time t at which positive information starts to spread among the nodes of the network and the cross-point of positive versus negative where positive information has no longer infected more nodes than the negative information. The spread of the negative information will still start at time $t = 0$. The start node for the negative information is always chosen randomly. The start node for

TABLE II
COMPARISON OF MAX START TIME t

Random Selection		EigenVector Selection	
Number of Nodes	Max Positive Start Time	Number of Nodes	Max Positive Start Time
1	0	1	4
2	1	2	4
5	1	5	6
10	2	10	9

the positive information is chosen randomly, or by having a good eigenvector centrality value.

The interactions of these two opposite information deserve some discussions. Normally, people react to negative information or opinion quickly and it takes time for them to change their mind. The same can be said for positive information. Therefore, while other more advanced interaction techniques are possible, we focus on a simple method in which each user gets one information and sticks with it even though the opposite information may arrive later. We will discuss other types of interactions as part of our future work in Section VI.

Figures 3 and 4 represent examples of the results from running the positive and negative information propagation simulations. Figure 3 consists of the simulations where the positive start node was chosen randomly, while in Figure 4 the positive start node was chosen to be the node with the highest eigenvector centrality. At first glance of these two figures, it is clear that selecting the top eigenvector node instead of choosing the node randomly, allows for a wider window of time to respond. Similarly, simulations were ran by choosing the positive start nodes by choosing 2 randomly, 5 randomly, 10 randomly, the top eigenvector node, the top 2 eigenvector nodes, the top 5 eigenvector nodes, and the top 10 eigenvector nodes. For simplicity, the results of these simulations are shown in Table II.

For each method of choosing the positive start node(s), 100 simulations were ran for each positive start time t on the x-axis of Figure 3, and t ran from 0 to 50. For each 100 simulations at each point t on the x-axis, the total number of infected nodes for both positive and negative information were calculated. The median value of each 100 simulations was then taken and this is the value represented at each t. The results shown in Table II. Figure 5 lets us take a look at the trends of total positive infected nodes for each method of choosing the start nodes. We can observe higher number of infected nodes (throughout t) based on eigenvector centrality in comparion to random node selection at each budget sizes.

Now, we want to take a deeper look into the results from the double propagation simulations. Taking a look at columns two and four in Table II, "Max positive start time", we can see the maximum time t at which the positive information can start to spread and still infect more nodes than negative information infects. This column allows us make two very important conclusions. First, we can say that the higher the budget one has for picking their start set of nodes, the wider

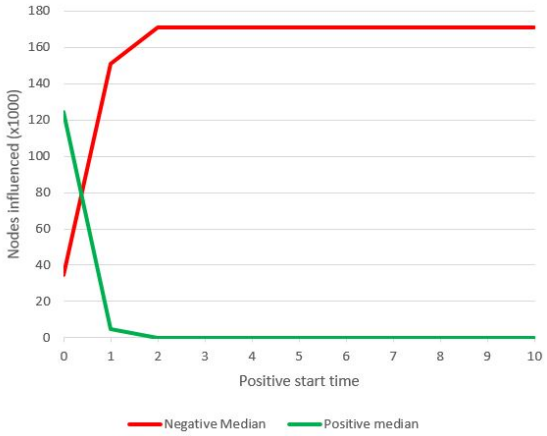

Fig. 3. Median infected nodes by positive and negative propagation at each time t with both start nodes randomly selected.

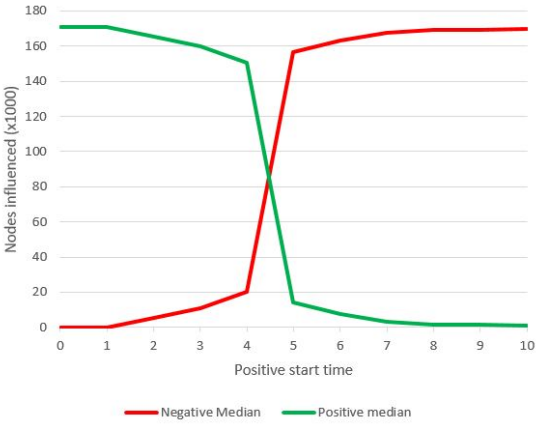

Fig. 4. Median infected nodes by positive and negative propagation at each time t with positive start node based on highest eigenvector centrality value (negative based on random).

the window they will have to start the spread of positive information. When choosing the start nodes randomly, if one only has a budget of 1 nodes, they must start the spread of positive information exactly when the spread of negative information starts. This time being $t = 0$. But, if there is a budget of 10 nodes, positive information can start spreading when $0 \leq t \leq 2$. When choosing the start nodes based off their eigenvector centrality, having a higher budget helps as well. If the budget is one node, positive information can start spreading when $0 \leq t \leq 4$. If the budget is ten nodes, the positive information now has a higher window to start spreading, $0 \leq t \leq 9$.

The second conclusion we can make from this, is that choosing the start nodes based off of their eigenvector centrality, allows for a larger window as well. Looking at having a budget of one node, choosing the node randomly basically leaves no room for error. Choosing the node that has the best eigenvector centrality allows for up to 4 extra ticks of t

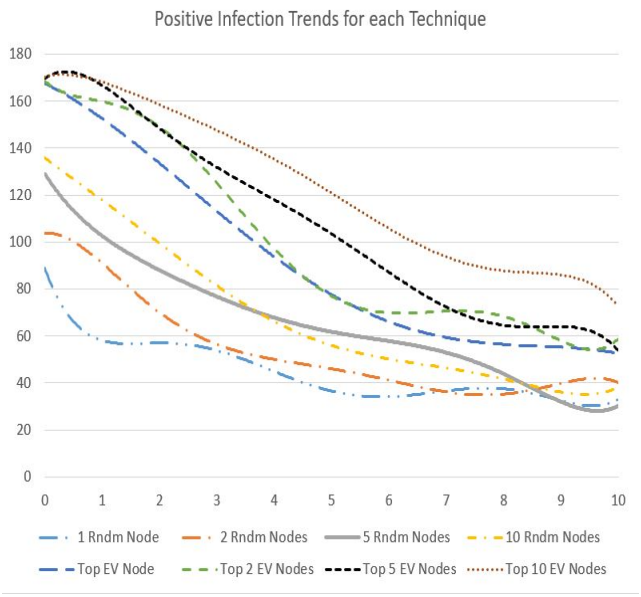

Fig. 5. Comparison of average infected nodes for each positive start time t (from 0 to 10) with different selection criteria.

to discover the negative information, choose the start node, and start the propagation of positive information through the network. This continues when comparing methods for a budget of two, five, and ten nodes. Choosing the nodes based on eigenvector centrality gives an extra 3, 5, and 7 ticks of t respectively.

IV. Conclusion

In this work, we have investigated information propagation through the snapshot of a small subset of Twitter that was constructed using a specific topic query. Our investigation focused on the propagation of negative information and counteracting with positive information that is expected to be unleashed with a delay. We have studied the selection of positive source nodes based on random selection, eigenvector centrality, and betweenness. As expected, our simulation results based on a so-called "one-off" infection mechanism showed that random selection of positive source nodes has a much slower infection or counteracting performance than the selections based on eigenvector centrality and betweenness. For example, when we give positive information a delay of 5 unit times, the selection of eigenvector centrality almost doubles the number positive infections compared to that of choosing positive source nodes randomly.

Our study, while interesting and indicative with strong results, leaves some future work directions. Different infection mechanisms should be investigated (other than the "one-off" method where nodes/users are infected by either negative or positive information and never change their mind). Such infection mechanisms can include change-of-mind based on number of positive information received, latency between two different information reaching the user, and others. Furthermore, a more realistic propagation of information other than "ticks" should be investigated as well.

References

[1] S. Aslam, "Twitter by the numbers (2018): Stats, demographics & fun facts," accessed: 2018-04-25. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/

[2] W. Chen, L. V. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synthesis Lectures on Data Management*, vol. 5, no. 4, pp. 1–177, 2013.

[3] E. Sadikov and M. M. M. Martinez, "Information propagation on twitter," *CS322 Project Report*, 2009.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[5] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 519–528.

[6] S. Ye and S. F. Wu, "Measuring message propagation and social influence on twitter. com," in *International Conference on Social Informatics*. Springer, 2010, pp. 216–231.

[7] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[8] D. Victor and M. Stevens, "United airlines passenger is dragged from an overbooked flight," accessed: 2018-04-25. [Online]. Available: https://www.nytimes.com/2017/04/10/business/united-flight-passenger-dragged.html

[9] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[10] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the spread of misinformation in social networks," in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, 2011, pp. 665–674. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963499

[11] Twitter, "Twitter public api documentation," accessed: 2018-04-25. [Online]. Available: https://developer.twitter.com/en/docs