CONTAINMENT OF MISINFORMATION PROPAGATION IN ONLINE SOCIAL NETWORKS WITH GIVEN DEADLINE

Nan Wang  
Renmin University of China, wangnan19900401@ruc.edu.cn

Li Yu  
Renmin University of China, buaayuli@ruc.edu.cn

Ni Ding  
Renmin University of China, dingni821@ruc.edu.cn

Dong Yang  
Renmin University of China, yangdong519@ruc.edu.cn

Follow this and additional works at: http://aisel.aisnet.org/pacis2014

Recommended Citation
Wang, Nan; Yu, Li; Ding, Ni; and Yang, Dong, "CONTAINMENT OF MISINFORMATION PROPAGATION IN ONLINE SOCIAL NETWORKS WITH GIVEN DEADLINE" (2014). PACIS 2014 Proceedings. 46. 
http://aisel.aisnet.org/pacis2014/46

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
CONTAINMENT OF MISINFORMATION PROPAGATION IN ONLINE SOCIAL NETWORKS WITH GIVEN DEADLINE

Nan Wang, School of Information, Renmin University of China, Beijing, P.R.China, wangnan19900401@ruc.edu.cn

Li Yu, School of Information, Renmin University of China, Beijing, P.R.China, State Key Laboratory of Digital Publishing Technology, Beijing, P.R.China, buaayuli@ruc.edu.cn (corresponding author)

Ni Ding, School of Information, Renmin University of China, Beijing, P.R.China, dingni821@ruc.edu.cn

Dong Yang, School of Information, Renmin University of China, Beijing, P.R.China, yangdong519@ruc.edu.cn

Abstract

In recent years, online social networks have become one of the most influential mediums of sharing information. However, alongside this promising feature comes the threat of rumors and misinformation propagation, which has aroused extensive attention from our society.

In this paper, we try to figure out how to select the smallest set of highly influential nodes decontaminated with good information to contain the propagation of misinformation effectively within given deadline and reach the ratio of positive nodes \( \omega^+ \) and the growth rate of newly added nodes with misinformation \( \varphi^- \) required in this paper. In the social networks with time delay, two competitive strategies of containing competitive propagating model (IC-CCPM) are proposed, respectively First Come, First Served strategy and Simultaneous Competitive strategy, considering not only Timeliness of competition between good information and misinformation, but also two kinds of information’s arrival at the same time, when users’ choices depend on the influence probability of neighbor users.

Two containing strategies are developed to select the smallest \( K \) positive seed nodes, respectively Greedy Containing strategy and Boundary Containing strategy.

A lot of experiments are made to test our proposed strategies by using three datasets, respectively twitter, Friendster and random. The experiment results show that Greedy Containing strategy and Boundary Containing strategy can be applied to most of the distributions of initial set of nodes contaminated by misinformation, which are superior to other available methods. In addition, boundary containing strategy is similar with greedy containing strategy containing effect, but better than greedy containing strategy in time efficiency.

Key words: online Social network, misinformation, competitive strategy, containing strategy
1 INTRODUCTION

In recent years, with the advent of social network sites, such as Facebook, Twitter et al, online social networks are developed greatly as one of the influential mediums of sharing information. Social networks bring huge influence to people’s all aspects of life, in which people can keep in touch with their friends and families and obtain latest information. However, many incidents that well-known twitter accounts were stolen and rumours were spread cause public panic and the fall of share prices. Therefore, a potential problem is posed that how to contain misinformation spread effectively in online social networks. In order to avoid huge loss caused by the cascaded spread of misinformation, an effective way of containing the spread of misinformation is needed.

Extensive research has been conducted on competition issues between positive and negative information in viral marketing, most of which are focused on how to minimize the influence of rumours or misinformation (chen.2011, Budak.2011, He.2012, Nguyen.2012). Two kinds of situations can be divided in the competitive information spreading environment, based on whether the negative seed nodes is known. In this paper, containing strategies based on competitive spread are studied, assuming that the set of negative seed nodes $S^-$ are known. Considering that sequence and speed are essential factors which affect what kind of strategies user will take, we try to figure out how to select the smallest set of highly influential nodes decontaminated with good information, initiated from the set of $S^+$, to contain the propagation of misinformation effectively within a given deadline $\tau$ and the ratio of positive nodes reaches $\omega^+$, and the growth rate of newly added nodes with misinformation in $\tau+1$ satisfies $\varphi^-$. As we all hope that information spreading in online social networks is credible and useful, our study aims to improve the environment of online social networks.

Some attempts on limiting misinformation have been made in earlier works (see the Related Work). The key difference between our work and theirs is that we think in the social networks with time delay, competition between good information and misinformation is time-efficient. In this paper, random number $t(u,w)$ are used as the time delay between user $u$ and user $w$, while “one hop time” is simply used in other studies between neighbour users. Moreover, current studies considered that when good information and misinformation arrive at the same time, misinformation is advantaged (chen.2011), or good information has superiority (Budak.2011, Nguyen.2012). However, we think that which kind of information is dominant depends on the influence probability of neighbour users. Based on this, we propose containing competitive propagating model (IC-CCPM), which extends the IC model and includes two competitive strategies: (1) First Come, First Served strategy, (2) Simultaneous Competitive strategy. If the first come, first served strategy is taken, user will choose to believe information from one side and refuse to believe another side, even though information from another side can reach later. If the Simultaneous Competitive strategy is taken, user will accept more credible information.

The most relevant work to our effort is the one suggested by Nguyen et al. (2012), in which the authors presented Decontamination Mechanism based on IC model and LT model, and believed that corrections to misinformation were carried with the propagating good information, thus as long as the user hasn’t been affected by misinformation, good information will be believed if arriving first and this will not change. Even if both kinds of information arrive at the same time later, user will still choose believe good information. The first come, first served strategy applied in social networks with time delay we
proposed in this paper is similar with their assumption, but we adopt Simultaneous Competitive strategy when two kinds of information arrive simultaneously, not considering that good information is advantaged. Moreover, they applied GVS and community computing to select the positive seed nodes, whereas we propose the Boundary Containing strategy which is better in time efficiency as well as Greedy Containing strategy. As for containing effectiveness, they think that the ratio of positive nodes reaches at least \(\beta\), while we consider that not only the coverage rate of positive nodes comes to \(\omega^+\), but also the growth rate of newly added nodes with misinformation in \(\tau+1\) satisfies \(p^-\).

At last, experiments are performed on Random Network Graph and Actual Network Graph, and the results indicate that Greedy Containing strategy and Boundary Containing strategy can be applied to most of the distributions of initial set of nodes contaminated by misinformation, which are superior to other available methods. Moreover, both strategies have similar containing effect, but Boundary Containing strategy is better than Greedy Containing strategy in time efficiency.

In a nutshell, our main contributions made in this work are summarized as follows. Firstly, the spreading competition between good information and misinformation is in the social networks with time delay. Secondly, two competitive strategies of containing competitive propagating model (IC-CCPM) are proposed, respectively First Come, First Served strategy and Simultaneous Competitive strategy, considering not only Timeliness of competition between good information and misinformation, but also two kinds of information’s arrival at the same time, when users’ choices depend on the influence probability of neighbour users. Thirdly, two containing strategies are developed to select the smallest \(K\) positive seed nodes, respectively Greedy Containing strategy and Boundary Containing strategy. Empirical results indicate that Boundary Containing strategy obtains relatively optimal containing effect.

The remaining parts of the paper are organized as follows. Related work is surveyed in next section. In section 3, containing competitive propagating model (IC-CCPM) is introduced. In section 4, containing strategies based on competition are proposed. In section 5, experiments are performed to compare the proposed containing strategies. Finally, conclusions are drawn and directions for future research are discussed.

2 RELATED WORKS

The information and influence propagation problem on social networks was first studied by Domingos and Richardon (2001) who also came up with the “Viral Marketing” concept in the field of data mining. Later, Kempe et al. (2003), tried to solve influence maximization problem (NP-hard) with discrete optimization method and proposed Linear Threshold (LT) and Independence Cascade (IC) models, which have become fundamental models of Viral marketing field in the following years. As the influence probability between users is difficult to decide, most of related studies are based on random number. However, Golbeck et al. (2011) not only verify that there is influence between users, but also applied Discrete Time (DT) model to predict whether and when users execute operations. In their study, network users’ logs of past behaviour were used as training set to predict influence probability.

Afterwards studies on viral marketing were emphasised on two aspects. One is single idea influence maximization problem. For instance, Kimura and Saito (2006) proposed shortest-path based on IC models and provided efficient algorithms to compute influence spread under these models. Chen et al.
(2010) showed that it is NP-hard to compute the exact influence of any node set in general graphs, proposed the MIA model and developed scalable algorithms to compute exact influence in trees and mine seed sets with equally good quality as those found by the approximation algorithm. Another is studies on competitive spreading influence. For instance, Bharathi et al. (2007) gave a \((1−1/e)\) approximation algorithm for computing the best response to an opponent’s strategy, and proved that the “price of competition” of this game is at most 2. Yu et al. (2012) proposed a game-theoretic model for competitive information dissemination in social network.

Some attempts have been made in the light of containing the spread of misinformation. Kostka et al. (2008) studied rumours spread using game theory method and showed that the opinion of propagating first is not always an advantage. Budak et al. (2011) applied the strategy of using “good” information dissemination to fight against misinformation propagation in social networks. Nguyen et al. (2012) studied the \(\beta^T\) Node Protectors problem, which aims to find the smallest set of highly influential nodes whose decontamination with good information helps to contain the viral spread of misinformation. The authors proposed GVS algorithm and a community-based heuristic method for the Node Protector problems. However, this algorithm requires the uniform protection fraction for each community. Kumar et al. (2013) proposed a novel information diffusion model for the spread of misinformation using Evolutionary game theory and Evolutionary graph theory. The model provides a framework to study the effects of multiple campaigns in the network which enable us to estimate the efficacy of launching counter campaigns against the spread of misinformation.

3 VIRAL MARKETING CONTAINING COMPETITIVE PROPAGATION

3.1 Viral Marketing Propagating Network

In our study, undirected network diagram \(G = (V, E, P, T)\) is built to indicate the social network with users connecting with each other, where the edges \(e(u, w)\) represent relationships between the users. There are two parameters on edges \(e(u, w)\), respectively the probability of mutual influence between users \(p(u, w)\) and propagation delay \(t(u, w)\). The probability of mutual influence \(p(u, w)\) simply represents the chance or probability of mutual influence between users, differing from the abstract representations of the influence user \(u\) has on user \(w\) in most viral marketing researches in the past, and neighbour users have same probabilities to influence each other. Propagation delay \(t(u, w)\) represents the delay of information transfer between users. Actually the information diffusion phenomenon with time delay has been verified in the traditional statistical physics.

3.2 User State

In addition to the relationship parameters between users, there are also state attributes on the individual users, which are used to indicate the probability \(P^+\) of accepting good information or the probability \(P^-\) of accepting misinformation. Assuming that \(P^+ + P^- = 1, P^- \in (0,1], P^+ \in (0,1]\), if \(P^+ > P^-\), \(P^+(v, t)\) can be used to indicate that user \(v\) is the accepter of good information and the attribute value is \(P^+(v, t)\). On the contrary, \(P^-(v, t)\) represents the accepter of misinformation and the
accepting probability. The state of user nodes is the superior state, and the sum of attribute value of the superior and inferior state equals 1.

As for a competitive campaign in online social networks, users’ state can be divided into three kinds, (1) stability, (2) statelessness, (3) instability. Therefore, in the initial situation of competition, negative seed nodes set \( S^- \) and positive seed nodes set \( S^+ \) are selected on undirected network graph \( G(V,E,P,T) \), and the state of negative seed nodes set is \( P^-(v,t) = 1, v \in S^-, t \in N \), which does not change at any time. At the same time, the state of positive seed nodes set is \( P^+(v,t) = 1, v \in S^+, t \in N \), which is fixed as well. In addition, the rest of user nodes are stateless.

### 3.3 Competitive Propagating strategies

After parameters of users’ relationship and state are discussed, we will talk about how the good information and misinformation compete among users in online social network. In this paper, competitive strategies aimed at two different situations are introduced, (1) if good information and misinformation do not reach at the same time, the first come, first served strategy will be taken, (2) if good information and misinformation reach at the same time, competitive strategy will be taken. The first strategy considers that once the influence probability of information from one side exceeds a threshold value \( \theta \), the user will choose to believe this information permanently. The reason why user chooses good information is that corrections of misinformation are carried with it, while the reason why user chooses misinformation is that it is difficult to turn negative nodes to positive nodes within a given deadline. However, the second strategy considers which kind of information user will accept depends on the influence probability of neighbour users.

We propose the competitive strategies which are inherited from Independent Cascade (IC) Model. In this model, any node is given only a single chance to influence its friends, thus if it fails to do so in time \( t \), it is not allowed to activate its friends again in time \( t + 1 \). If node \( w \) has multiple newly activated neighbors, they will try to activate \( w \) sequentially in an arbitrary order. Edge weight \( p(u,w) \) is the probability that user node \( u \) activate neighbour user node \( w \) successfully, which is random, instead of deducing from historical data.

Now, we are introducing the process of competitive propagation. Assuming that influence probability \( p(u,w) \) is ignored, the edge weight of social network \( G(V,E,T) \) only has propagation delay \( t(u,w) \).

![Figure 1. Competitive propagation simulated diagram](image)
As shown in figure 1, the time delay $t(A, B)$ between node A and node B equals random number 1. The green nodes are positive stable nodes with good information and the red nodes are negative stable nodes with misinformation. The node A, B, C, D are stateless when $T = 0$, and when $T = 1$, positive nodes and negative nodes start to spread information to neighbour nodes, forming good information and misinformation flow. Then node B and node C will stay in positive stable state $P^+ = 1$ and negative stable state $P^- = 1$ respectively, which do not change after that, because they are both affected by only one information flow. But when $T = 2$, node A are unstable, because two kinds of information are affecting node A at the same time, and the final state depends on the probabilities of good information and misinformation influence.

Then, we add the influence probability $p(u, w)$ into the graph. In the past relevant research papers, there are many activation methods, such as dominance, random or beyond a certain threshold value. We think threshold $\theta$ here means the value of minimum impact probability acceptable, in order to exclude edges with low mutual influence. If good information and misinformation do not reach at the same time, the first come, first served strategy will be taken. For instance, if user $v$ is only affected by good information at a certain time $t$ and $\text{infl}u^+(v, t) \geq \theta$, the value of state attribute is $P^+(v, t') = 1, t' \geq t$. After that time, user $v$ is in positive stable state, vice versa.

If positive and negative information reach at the same time, competitive strategy will be taken. Influence probability can be computed as following: $\text{infl}u^+(v, t') = 1 - \prod_{u \in \text{Neigh}^+(v)}(1 - p(v, u))$ (1), $\text{infl}u^-(v, t') = 1 - \prod_{u \in \text{Neigh}^-(v)}(1 - p(v, u))$ (2).

Considering that when users receive good information and misinformation at the same time, which one to believe needs to think twice, traditional computing methods of information propagation are used to help us to make competitive strategies. As shown in formula (1) and (2), $\text{infl}u^+(v, t')$ means the probability that good information is affected at the time of $t'$ and $\text{infl}u^-(v, t')$ means the probability that misinformation is affected at the time of $t'$. User’s final state is shown as following.

1 ) If $\text{infl}u^+(v, t') \geq \text{infl}u^-(v, t') \geq \theta$, node $v$ is in unstable state.

\[
P^+(v, t') = \frac{\text{infl}u^+(v, t')}{\text{infl}u^+(v, t') + \text{infl}u^-(v, t')}
\]

2 ) If $\text{infl}u^-(v, t') \geq \text{infl}u^+(v, t') \geq \theta$, node $v$ is in unstable state.

\[
P^-(v, t') = \frac{\text{infl}u^-(v, t')}{\text{infl}u^+(v, t') + \text{infl}u^-(v, t')}
\]

3 ) If $\text{infl}u^-(v, t') < \theta, \text{infl}u^+(v, t') < \theta$, node $v$ is still stateless.

4 ) If $\text{infl}u^-(v, t') < \theta, \text{infl}u^+(v, t') \geq \theta$, node $v$ is in positive stable state. $P^+(v, t) = 1, t \geq t'$

5 ) If $\text{infl}u^-(v, t') \geq \theta, \text{infl}u^+(v, t') < \theta$, node $v$ is in positive stable state. $P^-(v, t) = 1, t \geq t'$

4 CONTAINING STRATEGIES BASED ON COMPETITIVE PROPAGATION

In this part, we will introduce the containing strategies (greedy containing strategy and boundary containing strategy), which are proposed to select the smallest $K$ positive seed nodes set $S^+$, so that the proportion of positive nodes can reach a satisfactory value $\omega^+$ within a given deadline $\tau$ and in $\tau + 1$. 
the growth rate of negative nodes also can reach a satisfactory result $\varphi^- : \omega^+ = \frac{\text{Num}(U_{s \in S^+} \sigma (s))}{\text{Num}(U_{s \in S^+} \sigma (s)) + \text{Num}(U_{s \in S^-} \sigma (s))}$ (3), $\varphi^- = \frac{\text{Num}(U_{s \in S^-} \sigma (s+1)) - \text{Num}(U_{s \in S^-} \sigma (s))}{\text{Num}(U_{s \in S^-} \sigma (s))}$ (4).

$\omega^+$ is the ratio of the number of positive nodes and the sum of positive and negative nodes within a given deadline $\tau$. $\sigma (s)$ is the set of nodes activated by seed node $s$, $U_{s \in S^+} \sigma (s)$ is the set of nodes successfully affected by the positive seed nodes and $U_{s \in S^-} \sigma (s)$ is the set of nodes successfully affected by the negative seed nodes. $\varphi^-$ is the growth rate of the number of newly added nodes with negative information in $\tau + 1$ and the number of nodes with negative information before $\tau$. $U_{s \in S^-} \sigma (s+1)$ and $U_{s \in S^-} \sigma (s)$ represent the sets of nodes affected by the negative seed nodes in $\tau + 1$ and $\tau$ respectively. $\omega^+$ and $\varphi^-$ are important standards which can measure the effectiveness of containment.

Two containing strategies we proposed, respectively Greedy Containing strategy and Boundary Containing strategy, will be introduced in details below.

4.1  Greedy Containing Strategy

According to the Greedy containing strategy proposed in this paper, optimal K positive seed nodes can be selected by greedy algorithm from the initial set of stateless nodes $N_{\text{stateless}} = V \setminus (S^+ \cup S^-)$ except nodes with misinformation to maximize the proportion of good information and contain the diffusion of negative information. The algorithm is shown by figure 2, and the set of $K$ positive seed nodes calculated by greedy algorithm is treated as the of seed nodes $S^+$, which is explained as Equation 5: $f(S^+) = U_{u \in S^+} \sigma (u)$ (5).

<table>
<thead>
<tr>
<th>Algorithm 1: greedy algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $S^+ \leftarrow \emptyset$</td>
</tr>
<tr>
<td>2. For each $i = 1$ to $k$ do</td>
</tr>
<tr>
<td>3. select $u \leftarrow \arg \max_{w \in V \setminus (S^+ \cup S^-)} { f(S^+) }$</td>
</tr>
<tr>
<td>4. $S^+ \leftarrow S^+ \cup { u }$</td>
</tr>
<tr>
<td>5. End do</td>
</tr>
<tr>
<td>6. Output: $S^+$</td>
</tr>
</tbody>
</table>

Figure 2. Greedy algorithm

4.2 Boundary Containing Strategy

Unlike greedy containing strategy, the key point of boundary containing strategy is that surrounding negative nodes successfully and forming a circle composed by positive nodes can contain the diffusion of misinformation effectively, and better time efficiency can be achieved (experiment 3).

In this strategy, we try to select the set of boundary nodes $N_\tau$ from nodes that negative seed nodes set $S^-$ can reach when $\tau = \tau$, and then select a set of optimal $K$ positive seed nodes $S^+$ from $N_\tau$ based on greedy algorithm. As is shown in figure 3, each of the negative seed nodes can form a circle of time.
distance \( \tau \) and nodes along the boundary of the circle is the set of boundary nodes \( N_\tau \). What’s more, the opinion of boundary circle is suitable for most of the distributions of negative seed nodes, increasing the feasibility of boundary containing strategy (experiment 1).

\[ \sigma \]

**Theorem 1:** In the containing competitive propagating model (IC-CCPM), the proportion of positive nodes \( \omega^+ \) and the growth rate of negative nodes \( \varphi^- \) are increasing and decreasing along with the increase of time \( T \) respectively. That is to say, the containing efficiency on the boundary \( \tau \) reaches the maximum. \( \omega^+_T \leq \omega^+_{T+1}, \varphi^-_T \geq \varphi^-_{T+1}, T \in (1, \tau] \).

**Proof:** As negative seed nodes are propagating rapidly in social network \( G(V, E, P, T) \), the increasing rule of negative boundary nodes is \( |N_t| \leq |N_{t+1}| \). Similarly, stronger influence and transmission capacity can be achieved with more same nodes.

When the number of positive seed nodes is the same with negative ones, both kinds of nodes affect stateless nodes independently. As a result, the number of positive nodes are more than negative nodes \( |U_{SE+}\sigma(s)| \geq |U_{SE-}\sigma(s)| \) at any time, because negative seed nodes are randomly distributed and positive seed nodes are selected by strategies, the number of positive nodes are more than negative nodes \( |U_{SE+}\sigma(s)| \geq |U_{SE-}\sigma(s)| \) at any time.

Therefore, choosing positive seed nodes on the spreading boundary of negative nodes will reduce \( n_t \) nodes out of boundary node set \( N_t \). The number of negative nodes on the boundary can be represented as \( |N_t| - n_t \), and \( n_t \) represents the number of nodes activated by positive seed nodes in \( N_t \) when \( T = t \). When \( T = 2 \), \( \omega^+_2 = \frac{n_1+n_2+n_3+n_4}{|N_1|+|N_2|+n_3+n_4} \), \( \omega^+_3 = \frac{n_1+n_2+n_3+n_4+n_5+n_6}{|N_1|+|N_2|+|N_3|+n_3+n_5+n_6} \), \( \omega^+_4 = \frac{n_1+n_2+n_3+n_4+n_5+n_6}{|N_1|+|N_2|+|N_3|+n_3+n_5+n_6} \).

Therefore, the number of nodes in \( N_3 \) equals the sum of the number of positive and remained negative nodes, \( |N_3| = n_3 + \Delta(n_3) \). The number of remained negative nodes \( \Delta(n_3) \) can be ignored in IC-CCPM, \( \omega^+_2 \leq \frac{n_1+n_2+n_3+n_4}{|N_1|+|N_2|+n_3+n_4} \), \( \omega^+_3 = \frac{n_1+n_2+n_3+n_4+n_5+n_6}{|N_1|+|N_2|+|N_3|+n_3+n_5+n_6} \). Assuming that when \( T = k \),

\[ \omega^+_k \leq \omega^+_{k+1} \text{ is true, } \omega^+_{k+1} = \frac{\sum_{i=k+1}^{2(k+1)} n_i}{\sum_{i=1}^{2(k+1)} n_i}, \omega^+_{k+2} = \frac{\sum_{i=k+2}^{2(k+2)} n_i}{\sum_{i=1}^{2(k+2)} n_i} \], when \( T = k + 1 \). As \( \omega^+_k \leq \omega^+_{k+1} \) is true, the growth rate of negative nodes is

\[ \varphi^-_k \geq \varphi^-_{k+1} \text{ is true, } \varphi^-_{k+1} = \frac{\sum_{i=1}^{2(k+1)} n_i}{\sum_{i=1}^{2k+1} n_i} \text{, when } T = k + 1 \].

**Figure 3. Boundary containing graphs**

Theorems following are proposed to answer the question why the set of K positive seed nodes should be selected on the time boundary \( T = \tau \).

\[ \text{Figure 3. Boundary containing graphs} \]
Finally, when \( T = k + 1 \), then we got \( \omega^{+}_{k+1} \leq \omega^{+}_{k+2} \), similarly, according to the more the number of nodes, the more the total degree, in addition of \( \omega^{+}_{T} \leq \omega^{+}_{T+1} \), the chance of successful information spread will increase, so we can get \( \varphi^{+}_{T} \geq \varphi^{+}_{T+1} \), and the proof is complete.

Differing from greedy containing strategy, boundary containing strategy not only select K positive seed nodes in \( N_{t} \), but also consider the time distance of the positive seed node and other boundary nodes is shortest. \( \text{AvgTime}(u) \) is the average time distance from node \( u \) to other boundary nodes, as shown in the formula 6: 
\[
\text{AvgTime}(u) = \frac{\sum_{v \in N_{T}} \text{Dijkstra}(u,v)}{|N_{T}|} \quad (6).
\]

If \( |\text{Activated}(v)| = |\text{Activated}(w)| \) and \( \text{AvgTime}(u) < \text{AvgTime}(w) \), it is better to select user node \( u \) which has the shortest time distance with other boundary nodes. As a result, user node \( v \) can activate more boundary nodes than user node \( w \).

The algorithm is shown by figure 4. Dijkstra algorithm is used to select nodes which have the shortest time distance \( T \) from step 1 to step 5, \( \text{Dis}(u) = \min_{v \in S} \text{Dijkstra}(s,u) = T \), and eligible user node \( u \) will be added into boundary set \( N_{T} \). Each boundary node \( \forall v \in N_{T} \) is traversed and the set of nodes activated by this node \( \text{Activated}(v) \) and the average time distance \( \text{AvgTime}(v) \) can be obtained from step 6 to step 14. K optimal seed nodes will be selected based on \( \text{Activated}(v) \) and \( \text{AvgTime}(v) \) calculated earlier from step 15 to step 17.

**Algorithm 2: boundary algorithm**

1. \( N_{t} \leftarrow \emptyset, S^{+} \leftarrow \emptyset \)
2. **For each** \( u \in V/S^{-} \) **do**
   3. \( \text{Dis}(u) \leftarrow \min_{v \in S^{-}} \text{Dijkstra}(s,u) \)
   4. **If** \( \text{Dis}(u) = \tau \) **Then** \( N_{t} \leftarrow N_{t} \cup \{u\} \)
5. **End do**
6. **For each** \( v \in N_{t} \) **do**
   7. \( \text{Activated}(v) \leftarrow \emptyset, \text{AvgTime}(v) \leftarrow 0 \)
   8. **For each** \( w \in N_{t} \) **do**
   9. **IF** \( \text{ExpendTime}(v,w) \leq \tau \) **and** \( \text{Influ}(v,w) \geq \theta \) **Then**
      10. \( \text{Activated}(v) \leftarrow \text{Activated}(v) \cup \{w\} \)
   11. **Compute** \( \text{AvgTime}(v) \) Formula 6
12. **End IF**
13. **End do**
14. **End do**
15. **For each** \( i = 1 \) to \( k \) **do**
16. \( s \leftarrow \max_{w \in N \setminus S^{+}} |\text{Activated}(v) \cup \text{Activated}(w) - \cup_{v \in S^{+}} \text{Activated}(v)| \)
17. \( S^{+} \leftarrow S^{+} \cup \{s\} \)
18. **End do**
19. **Output:** \( S^{+} \)

*Figure 4. Boundary algorithm*
5 EXPERIMENTS

In this section, three group experiments are made to show the performance of containing strategies proposed in this paper and select at smallest $K$ positive seed nodes to achieve that the ratio of nodes with positive information comes to $\omega^+$ and the growth rate of nodes with negative information reaches $\varphi^-$. We assume that $\omega^+ \geq 0.6, \varphi^- \geq -\ln 0.2 = 1.6$ satisfies value. $\varphi^-$ has been processed by natural logarithm for convenience, so $-\ln \varphi^-$ becomes our experimental analysis data.

- The first group experiment is made to test and compare the performance of four different methods, respectively greedy containing strategy, boundary containing strategy, high degree containing strategy and random containing strategy, to select positive seed nodes for random distribution of nodes with misinformation. These latter two results are mainly used to compare with the first two. (see Graph 5, 6, 7)
- The second group experiment is made to verify theorem 1 and compare the containing efficiency of greedy and boundary containing strategies when selecting positive seed nodes in the set of boundary node from 1 to $\tau$. We assume that $\tau = 4$. (see Graph 8)
- The third group experiment is made to compare algorithm’s time complexity of greedy and boundary containing strategies when $K$ positive seed nodes are selected within a given deadline $\tau$. (see Table 2)

5.1 Experiment Dataset

In this paper, three datasets are used to make experiment. In addition to Twitter (11316811 nodes and 85331846 edges) and Friendster (11316811 nodes and 85331846 edges) dataset, a Random dataset is randomly generated for our experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Dataset</th>
<th>Neg_10</th>
<th>$\tau$</th>
<th>$\theta$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>Twitter2000</td>
<td>Scattered_10</td>
<td>4</td>
<td>0.4</td>
<td>5-10</td>
</tr>
<tr>
<td></td>
<td>Aggregate_10</td>
<td>4</td>
<td>0.4</td>
<td>5-10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scattered_10</td>
<td>4</td>
<td>0.4</td>
<td>5-10</td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>Twitter2000</td>
<td>Scattered_10</td>
<td>4</td>
<td>0.4</td>
<td>5-10</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>Twitter2000</td>
<td>Scattered_10</td>
<td>2,4,6</td>
<td>0.4</td>
<td>4,6,8,10</td>
</tr>
<tr>
<td></td>
<td>Friendster2000</td>
<td>Scattered_10</td>
<td>2,4,6</td>
<td>0.4</td>
<td>4,6,8,10</td>
</tr>
<tr>
<td></td>
<td>Random2000</td>
<td>Scattered_10</td>
<td>2,4,6</td>
<td>0.4</td>
<td>4,6,8,10</td>
</tr>
</tbody>
</table>

Table 1. Three group experiments

Considering that if the edge is too few, it is hard to propagate from seed nodes, so we extract sub-dataset with 2000 user nodes including dense edge from original Twitter and Friendster dataset. They are named according to the size, respectively named as Twitter2000, Friendster2000. In addition, another random dataset named Random2000 according to the size are generated for check the performance of two strategies. For each dataset, influence probability and influence time corresponding to each edge are generated randomly, ranging from 0 to 1.
5.2 Experiment Result

In order to prove the effectiveness of two containing strategies we proposed in this paper, the comparison of node degree and random containing strategies is complemented in the experiment. In addition, considering that the effectiveness of containing strategy is related to the distribution of negative seed nodes, the set of different distributions of negative seed nodes is found through experimental system on the basis of the experimental data.

The distribution of negative seed nodes can be divided into three classes, the first category can be called Scattered Class, in which the distance of any two seed nodes is more than the deadline ∀T(u, w) ≥ τ, u ∈ S−, w ∈ S−. The second is called Aggregate Class, in which the distance of any two seed nodes is less than the deadline ∀T(u, w) ≤ τ, u ∈ S−, w ∈ S−. The third is called Irregular Class, in which there are the set of nodes w with different states and abnormal nodes ∃T(u, w) > τ, w ∈ W−. For instance, in experiment 1, three different distributions of 10 negative nodes can be called as Scattered_10, Aggregate_10 and Irregular_10.

The first group experiment is conducted to compare the containing performance of greedy containing strategy, boundary containing strategy, high degree containing strategy and random containing strategy in three different distributions of negative seed nodes. (See Graph5, 6, 7). From the view of overall trend, Scattered Class is more difficult to contain than the other two. Because of the scattered distribution, it is difficult to influence stateless nodes as much as possible with limited positive seed nodes K. As the number of seed nodes increase, greedy and boundary containing strategies are better than the other two and the first two strategies are similar as for the value of ω+ and ϕ−. However, as is shown in the figure b of figure 4, 5, 6, the increase of ϕ− in boundary containing strategy is fluctuant, while greedy containing strategy shows a smooth curve.

Considering that the containing effect of boundary containing strategy depends on the time distance between nodes on the boundary τ, if seed node S+ cannot traverse the boundary once withinτ, nodes with negative information will have the opportunity to escape from the containing circle. According to the given containing objectives (ω+ ≥ 0.6, ϕ− ≥ −ln0.2 = 1.6) and the set of negative seed nodes, the minimum k of greedy and boundary containing strategies is 7, while the minimum k of other two strategies is 10, which doesn’t reach the standard of effective containment. The other two distributions are similar.

![Graph](image_url)

*Figure 5. The result of ω+ and −lnϕ− in scattered distribution of negative seed nodes*
The result of $\omega^+$ and $-\ln \varphi^-$ in aggregate distribution of negative seed nodes

The result of second group experiment verifies theorem 1 and shows that the optimal effect of containment can be achieved in the time boundary $\tau$ when boundary containing strategy is taken (see Graph 7), $\omega^+$ and $-\ln \varphi^-$ obtain the maximum in the time boundary $N_4$, presenting the monotone increasing tendency ($\tau = 4$). As is shown in the graph about $\omega^+$, $\omega_1^+ > \omega_2^+$ can be found and the ratio of the number of nodes with positive information when $\tau = 1$ is greater than the ratio when $\tau = 2$.

Positive seed nodes are selected from the set of time boundary nodes $N_1$, reducing $K$ neighbour nodes which negative seed nodes can contaminate. In this case, the ratio of positive nodes $\omega_1^+$ will increase, but the growth rate of negative nodes will increase as well. As for deadline $\tau$, we think that time limit
customers requires must far less than the complete time of negative nodes, which is the reason why we choose $\tau = 4$ as the deadline of experiment.

The third group experiment is made to compare algorithm’s time complexity of greedy and boundary containing strategies when K positive seed nodes are selected within a given deadline. Boundary containing strategy is better than greedy containing strategy in time efficiency, especially when the longer the deadline and the greater the number of nodes with positive information. Because of boundary containing strategy, K positive nodes will be selected from the set of boundary nodes $N_\tau$, in order to form a circle composed by nodes of positive nodes as soon as possible, instead of maximize the influence rate of positive nodes, which the greedy containing strategy focuses on.

<table>
<thead>
<tr>
<th>Tactics</th>
<th>Deadline</th>
<th>$\tau=2$ days</th>
<th>$\tau=4$ days</th>
<th>$\tau=6$ days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of K</td>
<td>6</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>boundary</td>
<td>Twitter2000</td>
<td>0.311</td>
<td>0.202</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>Friendster2000</td>
<td>0.199</td>
<td>0.189</td>
<td>0.192</td>
</tr>
<tr>
<td></td>
<td>Random2000</td>
<td>0.199</td>
<td>0.201</td>
<td>0.208</td>
</tr>
<tr>
<td>greedy</td>
<td>Twitter2000</td>
<td>0.097</td>
<td>0.089</td>
<td>0.086</td>
</tr>
<tr>
<td></td>
<td>Friendster2000</td>
<td>0.073</td>
<td>0.073</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>Random2000</td>
<td>0.099</td>
<td>0.101</td>
<td>0.101</td>
</tr>
</tbody>
</table>

Table 2. Runtime (seconds) about the boundary containing strategy and greedy containing strategy

6 CONCLUSION AND DISCUSSION

In the competitive online social networks, we propose containing competitive propagating model (IC-CCPM) based on Independent Cascade Model, which has two competitive strategies. Two containing strategies are proposed to select the smallest set of highly influential nodes decontaminated with good information to contain the propagation of misinformation effectively within given deadline.

A lot of experiments are made to test our proposed strategies by using three datasets, and the experiment results show that Greedy Containing strategy and Boundary Containing strategy can be applied to most of the distributions of initial set of nodes contaminated by misinformation, which are superior to other available methods. In addition, boundary containing strategy is similar with greedy containing strategy containing effect, but better than greedy containing strategy in time efficiency, and the optimal effect of containment can be achieved in the time boundary $\tau$ when boundary containing strategy is taken.

In the future, the feasibility of competitive and containing strategies will be increased by considering the situation that negative seed nodes are unknown. As the propagation of good information and misinformation is based on influence probability between users, which is mutual and equivalent in this paper, the situation that the influence probability can be studied to enrich the IC-CCPM.

Acknowledgements

This work was supported by the grants from National Natural Science Foundation of China (No.71271209, No.71331007), Beijing Municipal Natural Science Foundation (No.4132067),
References


