Limiting the Spread of Misinformation While Effectively Raising Awareness in Social Networks

Huiyuan Zhang(✉), Huiling Zhang, Xiang Li, and My T. Thai

Department of Computer and Information Science and Engineering,
University of Florida, Gainesville, FL 32611, USA
{huiyuan,huiling,xixiang,mythai}@cise.ufl.edu

Abstract. In this paper, we study the Misinformation Containment (MC) problem. In particular, taking into account the faster development of misinformation detection techniques, we mainly focus on the limiting the misinformation with known sources case. We prove that under the Competitive Activation Model, the MC problem is NP-hard and show that it cannot be approximated in polynomial time within a ratio of $e/(e-1)$ unless $NP \subseteq DTIME(n^{O(\log \log n)})$. Due to its hardness, we propose an effective algorithm, exploiting the critical nodes and using the greedy approach as well as applying the CELF heuristic to achieve the goal. Comprehensive experiments on real social networks are conducted, and results show that our algorithm can effectively expand the awareness of correct information as well as limit the spread of misinformation.

1 Introduction

With the increasing popularity of online social networks (OSNs), such as Facebook, Twitter and Google+, OSNs have become the most commonly utilized vehicles for information propagation. However, along with genuine and trustworthy information, rumors and misinformation also spread all around the Internet through this convenient and quick dissemination channel, which results in undesirable social effects and even leads to economic losses [1–3]. The rumor of the earthquake in Ghazni province in August 2012 made thousands of people leave their home in panic and be afraid of returning back home [5]. And the rumor about Obama injured originated from Twiiter in June 2011 caused the instability in financial markets. Misinformation about diseases are often observed [6]. For instance, there were many Twitter tweets containing misinformation about swine flu at the outset of the large outbreak in 2009. And the misinformation about vaccinations makes parents withhold immunization from their children [8]. Thus, it is crucial to seek efficient ways to control the inadvertent and intentional spread of misinformation.

Furthermore, once users believe the misinformation they received, they are resistant to change their beliefs, even though there are clear retractions [8]. Thus, rather than making efforts to only eliminate misinformation after it causes users' misunderstandings, negative emotions and further disruptive effects, we

M.T. Thai et al. (Eds.): CSoNet 2015, LNCS 9197, pp. 35–47, 2015.
DOI: 10.1007/978-3-319-21786-4_4

want to disseminate "good information" so as to raise users' awareness, reshape their attitude, and thus reduce their vulnerabilities to misinformation. "Good information" could refer to something for the debunking of misinformation, such as specific recommendations, authorized announcements or valid news.

Related Work. The large size and complex topology of OSNs, and various users' characteristics make this problem more challenging. Some existing works focus on identifying the infected nodes [7], which shed light on how to further design algorithms to prevent the misinformation from disseminating to the whole network. There are some recent attempts on limiting misinformation by initially injecting some good information and letting this good information propagate in the same network to convince other users [1,2,5]. Budak et al. [2] formulated the problem as an optimization problem and gave a greedy solution with approximation guarantees. The β_T^I problem of limiting viral propagation of misinformation is investigated in [1]. Fan et al. [5] studied the containment of rumor originating from a community and obtain the minimum number of needed protectors. But they only aimed to limit the spread misinformation. [4,15] approach this problem in different ways, they want to limit the spread of misinformation by blocking some nodes so that the overall pairwise connectivity can be minimized. However, during the process of limiting the misinformation by using good information, we should also simultaneously propagate this good information to other users who are unaware of this misinformation as soon as possible.

In this paper, we study the problem of minimizing the cost to prevent the spread of misinformation and simultaneously disseminate good information. At first we assume that nodes being active of misinformation are detected. An effective and timely algorithm is proposed to identify the most important disseminators with the minimum total cost to inject correct information into the diffusion. In particular, we should detect a critical round in which we take full protection on them to limit the wide spread of misinformation in advance and also prompt the further propagation of good information. Extensive experiments on real datasets are conducted to evaluate the efficiency and effectiveness of our algorithms and the results show that our algorithms perform significantly well.

Our contributions in this paper are summarized as follows:

- This is the first attempt to limit the misinformation and also maximize the prevalence size of good information. And we introduce Competitive Activation model to represent the competition nature of misinformation and good information.
- For the MC problem, we prove its NP-hardness and show that it cannot be approximated in polynomial time within a ratio of $\frac{e}{e-1}$ unless $NP \subseteq DTIME\ (n^{O(\log \log n)})$.
- The DI algorithm has been developed to find the nodes which can effectively block misinformation and simultaneously expand the influence of good information. And this algorithm is shown to be scalable to large-scale networks and outperforms several other heuristics.

The rest of this paper is organized as follows. Section 1 introduces the competitive activation model. Section 2 and Section 3 give the definition of Misinformation Containment and analysis of its complexity. We propose Dominating Influence algorithm in Section 4, and evaluate the performance of our method in Section 5. Finally, Section 6 concludes this paper.

2 Competitive Activation Model

In this paper, an OSN is modeled as a directed graph $G = (V, E)$, where nodes in V represent users and edges in E represent social ties between each pair of users. The size of a given graph G is $n = |V|$. Starting with a seeding set, information can propagate along edges of the underlying network. It is very possible for a user to be exposed to both misinformation and good information. Negative dominance is used as the tie-breaking rule in Competitive Linear Threshold model [9]. However, considering various characteristics of users, they could make different decisions upon receiving same information. So, we introduce a new parameter *preference* to determine which activation will finally succeed. Our model for the simultaneous spread of misinformation and good information is as follows.

Each node $v \in V$ is associated with two thresholds θ_v^A and $\theta_v^B > 0$, and each edge $(u, v) \in E$ is assigned to two weights $w_{uv}^A, w_{uv}^B \geq 0$ corresponding to misinformation A and good information B. Let I_0^A and I_0^B denote the sets of initial A-active nodes, accepting the misinformation, and B-active nodes, believing good information, respectively. At time t, an inactive node v will become A-active if $\sum_{u \in I_{t-1}^A} w_{uv}^A \geq \theta_v^A$, or will become B-active if $\sum_{u \in I_{t-1}^B} w_{uv}^B \geq \theta_v^B$. When both thresholds have been satisfied, a node will decide to accept which one by its preference, $P_v^i = (\sum_{u \in N_a^{in}(v)} w_{uv}^i)/\theta_v^i$, where $i \in \{A, B\}$ and $N_a^{in}(v)$ is the set of activated in-neighbours of v . It will become A-active if $P_v^A \geq P_v^B$, and vice versa. After accepting one kind of information, a node will stay in this status and not change to accept another one till the end of diffusion process, reflecting the continued influenced effect of information perception.

3 Misinformation Containment and Inapproximability

3.1 Problem Definition

Definition 1. ***Misinformation Containment (MC).*** *Given misinformation A and good information B spread on a graph $G = (V, E, \theta^A, \theta^B, w^A, w^B)$, where $\theta^i = \{\theta_v^i\}$, $w^i = \{w_{uv}^i\}$ and $i \in \{A, B\}$, while set of I_0^A and k_B are given, this problem aims to find a seeding set for good information I_0^B of size k_B such that we can minimize the number of A-active nodes and simultaneously maximize the number of B-active nodes.*

3.2 Hardness of MC

In this section, we first show the NP-completeness of MC problem by reducing it from the **Maximum Coverage** problem. We further prove the inapproximability of MC which is NP-hard to be approximated within a ratio of $\frac{e}{e-1}$ unless $NP \subseteq DTIME(n^{O(\log \log n)})$.

Theorem 1. *The MC problem is NP-complete.*

Proof. We first consider the decision version of MC problem that asks whether the graph $G = (V, E, w^A, w^B, \theta^A, \theta^B, I_0^A, k_B)$ contains a set of vertices $I_0^B \subset V$ of size k_B such that the number of B-active nodes is at least t_B and the number of A-active nodes is at most t_A where t_A and t_B are positive integers. Given $I_0^B \subset V$, we can easily compute the influence spread of B as well as that of A in polynomial time under the CAM model. This implies MC is in NP.

To prove that MC is NP-hard, we reduce it from the decision version of Maximum Coverage problem defined as follows.

Maximum Coverage. Given a positive integer k, a set of m elements $\mathcal{U} = \{e_1, e_2, \cdots, e_m\}$ and a collection of sets $\mathcal{S} = \{S_1, S_2, \cdots, S_n\}$. The sets may have some elements in common. The *Maximum Coverage* problem asks to find a subset $\mathcal{S}' \subset \mathcal{S}$, such that $|\cup_{S_i \in \mathcal{S}'} S_i|$ is maximized with $|\mathcal{S}'| \leq k$. The decision version of this problem asks whether the input instance contains a subset $\mathcal{S}$ of size k which can cover at least t elements where t is a positive integer.

Reduction. Given an instance $\mathcal{I} = \{\mathcal{U}, \mathcal{S}, k, t\}$ of maximum coverage, we construct an instance $G = (V, E, \theta^A, \theta^B, w^A, w^B, I_0^A, k_B, t_A, t_B)$ of MC problem as follows.

The set of vertices: add one vertex u_i for each subset $S_i \in \mathcal{S}$, one vertex v_j for each element $u_j \in \mathcal{U}$, and a special vertex x.

The set of edges: add an edge (u_i, v_j) for each $e_j \in S_i$ and connect x to each vertex v_j.

Thresholds and weights: assign all vertices the same threshold $\theta^A = \theta^B = \frac{1}{2m}$, and each edges (u_i, v_j) has weight $w^A_{u_i v_j} = 0, w^B_{u_i v_j} = \frac{1}{m}$. In addition, for all edges leaving from x, we assign their weights as $w^A_{xv_j} = \frac{1}{2m}, w^B_{xv_j} = 0$.

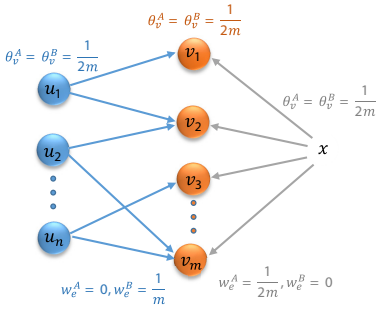


Fig. 1. Reduction from Maximum Coverage to Misinformation Containment

The construction is illustrated in Fig. 1. Finally, set $k_B = k$, $t_B = t + k_B$ and $t_A = m - t + 1$. Let $I_0^A = \{x\}$. We now show the equivalence between two instances.

Suppose that $\mathcal{S}^*$ is a solution to the maximum coverage instance, thus $|\mathcal{S}^*| \leq k$ and it can cover at least t elements in $\mathcal{U}$. By our construction, we can select all the nodes u_i corresponding to subset $S_i \in \mathcal{S}^*$ as a seeding set I_0^B. Thus $|I_B^0| = k = k_B$. Since $\mathcal{S}^*$ can cover at least t elements e_j in $\mathcal{U}$, then I_0^B can influence at least t vertices v_j corresponding to those e_j. Besides, for any v_j, both of A and B's total incoming influence exceed its threshold and $P_{v_j}^A \leq P_{v_j}^B$. Hence, there are at least $t + k_B$ B-active nodes in the MC problem and at most $m - t + 1 = t_A$ A-active nodes.

Conversely, suppose there is a B-seeding set $|I_0^{B*}| = k_B$ such that the number of B-active nodes is at least t_B. For any $v_j \in I_0^{B*}$, we replace it with its adjacent node u_i. This replacement does not reduce the number of B-active nodes. Then the $\mathcal{S}^*$ can be a collection of subset S_i corresponding to those $u_i \in I_0^{B*}$ after the replacement which has exactly size k and the number of elements which it can cover is at least $t_B - k_B = t$.

As MC problem is NP-complete, we further show that the above reduction implies a $\frac{e}{e-1}$-inapproximation factor in the following theorem.

Theorem 2. *The MC problem can not be approximated in polynomial time within a ratio of* $\frac{e}{e-1}$ *unless* $NP \subseteq DTIME(n^{O(\log \log n)})$.

Proof. We use the above mentioned reduction in the proof of Theorem 1. Suppose that there exists a $\frac{e}{e-1}$-approximation algorithm $\mathcal{H}$ for MC problem. Then $\mathcal{H}$ can return the number of B-active nodes in G with seeding size less than k_B. By our constructed instance, we can obtain the maximum coverage with size t if the the number of B-active nodes in the optimal solution for MC problem is $t + k_B$. Thus algorithm $\mathcal{H}$ can be applied to solve the Maximum Coverage problem in polynomial time. And this contradicts to the NP-hardness of Maximum Coverage problem [10].

4 Dominating Influence Algorithm

In this part, we propose our Dominating Influence (DI) algorithm for MC problem. DI algorithm consists of two sub-algorithms, which are DI-Gateway Nodes Detection and DI-Candidate Selection. DI-Gateway Nodes Detection helps us identify the gateway nodes, which are of significance in enlarging misinformation's influence. Before misinformation's diffusion naturally terminates, we use DI-Candidate Selection to find candidate seeding sets of different searching rounds, which are determined by the set of gateway nodes. Finally, we obtain the best seeds for good information from Dominating Influence algorithm.

4.1 Gateway Nodes Detection

In order to block the spread of misinformation, we should identify which nodes play an important role in its spreading out. In this paper, we use "gateway node" to refer to nodes which help misinformation propagate further. Knowing currently active nodes of misinformation, we can obtain the newly active nodes in each time stamp. Under CAM model, we have the following key observation.

Observation 1. *Each newly activated node in time t must be incident to at least one node that activated in time $t-1$.*

Algorithm 1. DI-Gateway Nodes Detection

Input: Two set of nodes I_t^A, I_{t-1}^A
Output: A set of gateway nodes C_{t-1}
$C_{t-1} \leftarrow \emptyset$
for $i = 1$ to $|I_{t-1}^A|$ **do**
  $\delta_{\max} = 0$
  for each $v \in I_{t-1}^A \setminus C_{t-1}$ **do**
    if $\delta_v(C_{t-1}) > \delta_{\max}$ **then**
      $\delta_{\max} \leftarrow \delta_v(C_{t-1})$
    end if
    if $\delta_{\max} = 0$ **then**
      Return C_{t-1}
    end if
    $C_{t-1} \leftarrow C_{t-1} \cup \{v\}$
  end for
end for
Return C_{t-1}

According to this observation, we utilize a trace back method to shrink the influence of misinformation step by step. Instead of starting from the inner-most round, we begin with the outer-most round. The reason is to avoid changes from happening in an earlier stage that may result in a cascading behavior to the later round. By simulating the propagation of misinformation, we record the set of nodes I_i^A that activated in round $i, i = 1, 2..., T$. To prevent the further propagation of misinformation to I_t^A, we should deal with nodes in I_{t-1}^A. Rather than targeting all nodes activated in round $t-1$, we want to find the gateway nodes which contribute to activating the most number of nodes in I_t^A. Thus, we use a greedy approach to sequentially select a node $u \in I_t^A$ maximizing the following marginal gain into set C_{t-1}:

$$\delta_u(C_{t-1}) = f(C_{t-1} + \{u\}) - f(C_{t-1}),$$

where $f(\cdot)$ is the number of newly activated nodes which are incident to the set of selected nodes.

The algorithm terminates and returns the set of gateway nodes C_{t-1} for a given set of A-active nodes I_t^A. The detail of this step is shown in Algorithm 1.

4.2 Candidate Selection

After obtaining the set of gateway nodes, we want to target those nodes and activate them before misinformation reaches. Meanwhile, we desire to enhance users' awareness of good information. To achieve both goals, we present the candidates selection in Algorithm 2, and the core is to iteratively choose a node that maximizes the following marginal gain:

$$\begin{aligned}\eta_u(I_0^B) =&\alpha[\psi(I_0^B + \{u\}) \cap C_{t-1} - \psi(I_0^B) \cap C_{t-1}]+\\ &\beta[\psi(I_0^B + \{u\}) - \psi(I_0^B)],\end{aligned}$$

where $\alpha + \beta = 1$. By adjusting the value of α, and β, we can change the effect on limiting misinformation's influence and expanding the influence of good information.

Algorithm 2. DI-Candidate Selection

Input: $G = (V, E, w^A, w^B\theta^A, \theta^B)$, C_{t-1} and k_B
Output: A candidate seed set $I_0^B(t-1)$ of size at most k_B
$P \leftarrow \emptyset$, $Q \leftarrow \emptyset$
for each $v \in C_{t-1}$ **do**
  Find node u that is $t-1$-hops away from v
  $P \leftarrow P \cup \{u\}$
end for
for $u \in P$ **do**
  Compute $\eta_u(I_0^B(t-1))$,
  Push u into Q
end for
while $|I_0^B| \leq k_B$ **do**
  repeat
    $u \leftarrow$ top of Q
    Recompute $\eta_u(I_0^B(t-1))$
  until u stays on top of Q
  if $\eta_u(I_0^B) \leq 0$ **then**
    Return $I_0^B(t-1)$
  end if
  $I_0^B \leftarrow I_0^B + \{u\}$;
  Return $I_0^B(t-1)$, $result(A, B, t-1)$
end while

Since greedy algorithms are always suffering from severe scalability problem, we use two techniques to effectively improve the running time. First, instead of selecting nodes from all over the network, we start from a candidate set P, which consists of nodes that are $t-1$ hops away from the targeted gateway nodes. Second, we employ CELF [11] heuristic to speed up the selection in each iteration. This approach can avoid the exhaustive update, which is extremely time consuming. This algorithm finally returns a candidate seeding set I_0^B as well as the total number of A-active and B-active nodes, respectively.

4.3 DI Algorithm

Incorporating above two algorithms, we obtain the DI algorithm, presented in Algorithm 3. First, we simulate the diffusion of misinformation and obtain termination round T along with the sets of activated nodes $I_t^A, t = 1, \cdots, T$ in each round. Starting with an arbitrary $I_t^A, t \in [1, T]$, by applying Gateway Nodes Detection, we are able to find the set of nodes C_{t-1} that contributed the most to activating nodes in I_t^A. Next, in order to limit the diffusion of misinformation, we should guarantee that the node $v \in C_{t-1}$ should be activated by good information no later than time $t-1$. This requires us to either let good information reach v earlier than $t-1$ or activate more of v's neighbors to be B-active nodes in order to make v's preference $P_v^B \geq P_v^A$ at $t-1$.

Algorithm 3. Dominating Influence Algorithm

Input: Graph $G = (V, E, w^A, w^B, \theta^A, \theta^B)$, I_0^A and k_B
Output: A seed set I_0^B of size k_B
Simulate A's influence starts with I_0^A
Get the termination round T and sets of active nodes $I_i^A, i = 1, ..., T$
for $t = T$ to 1 **do**
    $C_{t-1} \leftarrow$ DI-Gateway Nodes Detection (I_t^A)
    $(I_0^B(t), result(A, B, t)) \leftarrow$ DI-Candidate Selection (G, C_{t-1}, k_B)
end for
for $t = 1$ to T **do**
    Find τ where $argmax_{\tau \in [1,T]}\{B \setminus A | result(A, B, t)\}$
end for
Return $I_0^B(\tau)$

Considering the above time constraint, there will be a trade-off when selecting nodes into the seeding set. If we try to limit the propagation of misinformation at an early stage, the candidate set (which consists of nodes $t-1$ hops away from C_{t-1}) will be very limited, and thus may lead to decreasing the quality of seeds to disseminate good information. On the contrary, we are able to get a better candidate set by postponing the time to block misinformation, but this may result in increasing the number of A-active nodes dramatically. However, since the termination round of misinformation diffusion is usually a relatively small integer, and by applying the above mentioned enhancements to improve the running time, we are able to go through each C_t where t is from 1 to T searching round in order to find to best seeding set. Eventually, by measuring the difference between number of A-active and B-active nodes for every C_t, we can obtain the best seeding set to contain misinformation and maximally raise users' awareness.

5 Experiment and Evaluation

In this section, we perform various experiments based on the proposed algorithms and heuristics with real-world datasets, and evaluate the performance.

5.1 Dataset Description

We use three real-world networks, which are widely used for information diffusion process analysis, their basic statistics are summarized in Table 1, including:

Gnutella. The snapshot of the Gnutella peer-to-peer file sharing network in August 2002. Nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts [13].

Facebook. This dataset contains friendship information among New Orleans regional network on Facebook, spanning from September 2006 to January 2009, where nodes represent users and edges among them are friendship.

Amazon. This network is collected by crawling Amazon webpages. In this graph, an edge (i, j) indicates that product i is frequently co-purchased with product j by customers [14].

Table 1. Basic Information of Investigated Networks

Network	Gnutella	Facebook	Amazon
Nodes	6,301	61,096	262,111
Edges	20,777	1,811,130	1,234,877
Avg. Degree	3.29	29.64	4.71
Type	Directed	Directed	Directed

For graphs we tested on, nodes' thresholds for accepting misinformation and good information are generated uniformly at random in the range $[0, 1]$. To assign the influence weights on each edge, we adopt the method in [12], where we uniformly generate edge weights at random in the range $[0, 1]$, and then normalize the weights of all incoming edges of a node v to let it satisfy that $\sum_{u \in N^{in}(v)} w_{u,v} \leq 1$. Furthermore, for the seeding set of misinformation, we employ the greedy algorithm proposed by Kempe et al. [12], where in each iteration, the node with maximum marginal gain is chosen into the seeds. We are most likely to be able to detect misinformation and take action to contain its spread after it has propagated for a while and leads to undesirable effect [3]. Considering this observation, we introduce a delay d to model the time difference of disseminating good information and misinformation starting out. Compared with random selection, assigning seeds set for misinformation in this way can guarantee the high quality of misinformation initiators, and makes our problem of choosing seeds set for good information so as to limit the influence of misinformation more challenging.

Algorithms Compared. In our experiments, we compare our algorithm with several other heuristics listed as follows:

- Random: Randomly select k_B nodes from $V \setminus I_0^A$ as the seeds for good information in the graph.
- MaxDegree: We choose top k_B nodes from $V \setminus I_0^A$ with highest degree as the seeding set for good information.
- MaxGreedy: The greedy algorithm focuses on maximizing the influence of good information, in which the node with the maximum influence of good information is iteratively picked[16].
- MinGreedy: The greedy algorithm targets on minimizing misinformation propagation; the node with maximum number of A-active nodes blocked is selected in each iteration [2,9].

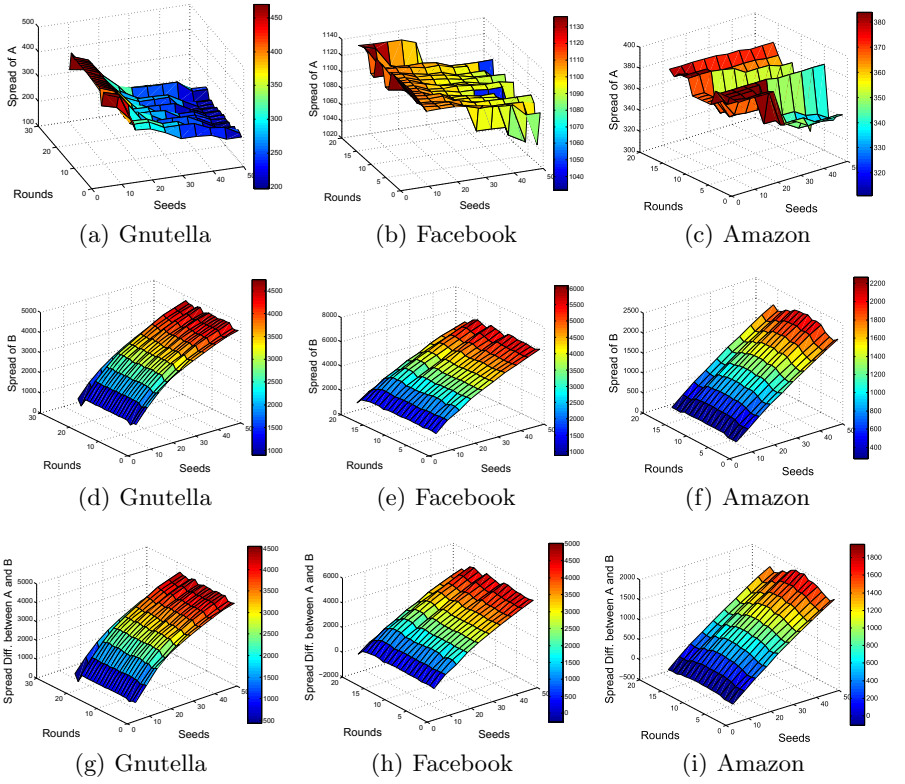
5.2 Experimental Results

In this part, we first measure the performance of our algorithm, in which we evaluate the number of A-active nodes and B-active nodes as well as their difference across three real world datasets with different number of seeds and rounds. Secondly, we compare the the results from all above mentioned algorithms. Next, we evaluate how time delay impacts the overall performance.

Seeding Set. We first present the spread of misinformation A and good information B achieved by selecting 50 B-seeds at different rounds. We evaluate them based on the number of A-active and B-active nodes, along with the difference between them. Fig. 2 shows two types of information of selecting 50 seeds with initial set $|I_0^A| = 10$ and time delay $d = 2$. The initiators of misinformation are selected by above described method, and before we disseminate good information in the network, misinformation has already activated 83, 205 and 50 nodes in Gnutella, Facebook and Amazon, respectively.

Fig. 2(a), 2(b), 2(c) show that the number of A-active nodes keeps dropping with a larger size of good information seeds. For example, in Gnutella, without adding any B-seeds, the spread of misinformation could reach as many as 851 nodes. However, by adding 50 seeds of good information selected by DI, the active size of misinformation can be limited to only 208 nodes. Conversely, Fig. 2(d), 2(e), 2(f) depict that the amount of B-active nodes increases dramatically with more B-seeds. For the seeds chosen from round 14 in Gnutella, the total number of B-active nodes can be 4749, eventually. Furthermore, we find that the difference between B-active nodes and A-active nodes is steadily increasing with larger budget of the seeding set of good information. It is also fluctuating with different targeting rounds.

Different Methods. Next, we compare the spread of both kinds of information achieved from different heuristics. The comparison is based on the number A-active nodes and B-active nodes and their difference. Fig. 3 shows the spread of misinformation and good information when there are 50 B-seeds and 10 initial A-active nodes, and the time delay $d = 2$ obtained from different heuristics. For limiting the spread of misinformation, MinGreedy is the best among those five methods across three datasets, while Random hardly blocks it. Except for MinGreedy, DI outperforms other heuristics as it effectively prevents the further



(a) Gnutella (b) Facebook (c) Amazon

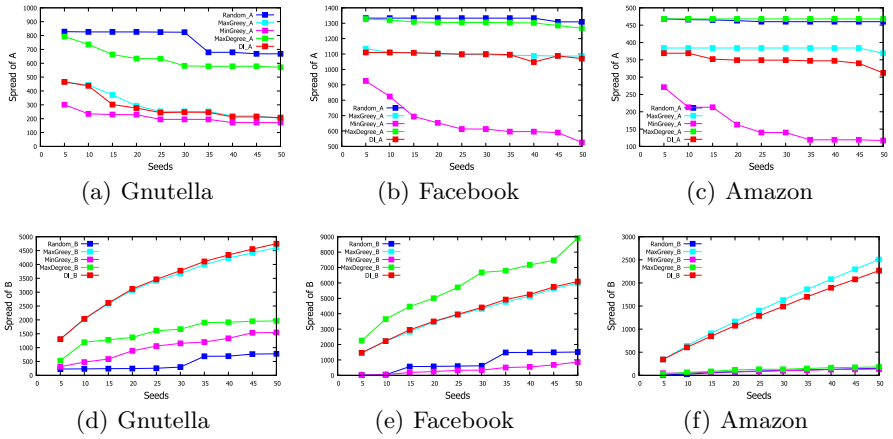
(d) Gnutella (e) Facebook (f) Amazon

(g) Gnutella (h) Facebook (i) Amazon

Fig. 2. Influence spread in three networks

propagation of misinformation. As shown in Fig. 3(c), the amount of A-active nodes nodes goes down to 312 finally while it could be 468 without any B-seeds, which means that a 33% of A-active nodes has been reduced.

On the contrary, the number of B-active nodes is increasing sharply for both of the DI and MaxGreedy algorithms. Fig. 3(d) demonstrates that the number of B-active nodes climbs to 4749 and 4608 after selecting 50 nodes by DI and MaxGreedy, while for other three methods, the total number for A-active nodes is less than 1500, similar results can be obtained in Amazon. However, the MaxDegree in Facebook achieves the largest number of nodes accepting misinformation. By digging into the data, we find that there are some super nodes with massive outgoing edges are chosen by MaxDegree, while missed by MaxGreedy. Considering the greedy approach in selecting seeds, some of those super nodes may have less gain than other nodes due to the way we assign edge weights. However, the combination of them could lead to a large cascading influence. Hence, MaxDegree even outperforms MaxGreedy on Facebook. However, seldom nodes accepting misinformation have been reduced compared to our DI.



(a) Gnutella (b) Facebook (c) Amazon

(d) Gnutella (e) Facebook (f) Amazon

Fig. 3. The number of nodes activated by misinformation and good information achieved by different algorithms

6 Conclusions

In this paper, in order to protect users from potential influence of misinformation, we aim to block misinformation and also raise users' awareness. We formulate the MC problem, and then prove it is NP-complete and cannot be approximated in polynomial time within a ratio of $\frac{e}{e-1}$ unless $NP \subseteq DTIME(n^{O(\log \log n)})$. An efficient algorithm DI is proposed, and extensive experiments on three real-world datasets are conducted. Experiments results show that our algorithm outperforms several other heuristics and well scalable to large-scale social networks.

Acknowledgment. This work is supported in part of NSF Career Award 0953284 and NSF CCF-1422116.

References

1. Nguyen, N.P., Yan, G., Thai, M.T., Eidenbenz, S.: Containment of misinformation spread in online social networks. In: Proceedings of the 4th Annual ACM Web Science Conference (2012)
2. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web (2011)
3. Tripathy, R.M., Bagchi, A., Mehta, S.: A study of rumor control strategies on social networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (2010)
4. Ventresca, M., Aleman, D.: Efficiently identifying critical nodes in large complex networks. Computational Social Networks **2**, 6 (2015)

5. Fan, L., Lu, Z., Wu, W., Thuraisingham, B., Ma, H., Bi, Y.: Least cost rumor blocking in social networks. In: IEEE International Conference on Distributed Computing Systems (ICDCS) (2013)
6. Jin, F., Dougherty, E., Saraf, P., Cao, Y., Ramakrishnan, N.: Epidemiological modeling of news and rumors on twitter. In: Proceedings of the Workshop on Social Network Mining and Analysis (2013)
7. Lim, Y.S., Ribeiro, B., Towsley, D.: Classifying latent infection states in complex networks. Computational Social Networks **2**, 8 (2015)
8. Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J.: Misinformation and its correction continued influence and successful debiasing. Psychological Science in the Public Interest **13**(3), 106–131 (2012)
9. He, X., Song, G., Chen, W., Jiang, Q.: Influence blocking maximization in social networks under the competitive linear threshold model. In: SDM (2012)
10. Feige, U.: A threshold of ln n for approximating set cover. Journal of the ACM (JACM) **45**(4), 634–652
11. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (2007)
12. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
13. Leskovec, J.: Gnutella peer-to-peer network (2002). http://snap.stanford.edu/data/p2p-Gnutella08.html
14. Leskovec, J.: Amazon co-purchasing network (2003). http://snap.stanford.edu/data/amazon-0302.html
15. Zhang, H., Alim, A.A., Thai, T.M., Nguyen, T.H.: Monitor placement to timely detect misinformation in online social networks. In: Proceedings of the International Conference on Communications (2015)
16. Carnes, T., Nagarajan, C., Wild, S.M., Van Zuylen, A.: Maximizing influence in a competitive social network: a follower's perspective. In: Proceedings of the Ninth International Conference on Electronic Commerce (2007)