Influence Blocking Maximization Under Refutation

Qi Luo (luoqi2018@mail.sdu.edu.cn)
Shandong University
Dongxiao Yu
Shandong University
Dongbiao Wang
Shandong University
Yafei Zhang
City University of Hong Kong
Yanwei Zheng
Shandong University
Zhipeng Cai
Georgia State University

Research Article

Keywords: Influence blocking maximization, refutation mechanism, competitive independent cascade model, reverse reachable set

Posted Date: July 25th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3184213/v1

Additional Declarations: No competing interests reported.

Influence Blocking Maximization Under Refutation

Qi Luo[1], Dongxiao Yu[1*], Dongbiao Wang[1], Yafei Zhang[2], Yanwei Zheng[1] and Zhipeng Cai[3]

[1]School of Computer Science and Technology, Shandong University, No. 72 Binhai, Qingdao, Shandong, P.R. China.
[2]Department of Media and Communication, City University of Hong Kong, Tat Chee Avenue, Hong Kong S.A.R., P.R. China.
[3]Department of Computer Science, Georgia State University, 25 Park Place, Atlanta, Georgia, USA.

*Corresponding author(s). E-mail(s): dxyu@sdu.edu.cn;
Contributing authors: luoqi2018@mail.sdu.edu.cn;
wangdongbiao@mail.sdu.edu.cn; yflyzhang@gmail.com;
zhengyw@sdu.edu.cn; zcai@gsu.edu;

Abstract

In social networks, a phenomenon termed the refutation mechanism arises when certain users spontaneously counter negative information based on their knowledge and experience. To the best of our knowledge, this paper focuses on the influence blocking maximization under the refutation mechanism for the first time. Specifically, incorporating the refutation mechanism with the Competitive Independent Cascade (CIC) model, we introduce the Refutation Competitive Independent Cascade (RCIC) model, while also considering real-time delay. Under the proposed model, we study the Joint Influence Blocking Maximization (JIBM) problem. The objective of JIBM is to maximize the expected number of nonnegatives by finding a set of positive seeds in a network. We show that the problem is NP-hard. We present a scalable approximation algorithm, named RR-JIBM, by making a non-trivial adaptation of the generation process of reverse reachable sets. We prove that the given algorithms achieve $(\mathbf{1 - 1/e - \varepsilon})$-approximation for any $\boldsymbol{\varepsilon > 0}$ for JIBM problem. An improved algorithm named RR-JIBM+ is also proposed to improve the efficiency of RR-JIBM in reality. Experiments on real-world social networks show that our algorithms

outperform other intuitive baselines in reducing the number of nodes influenced by negative seed nodes. Meanwhile, the RR-JIBM+ algorithm has a higher efficiency advantage than RR-JIBM on different datasets.

Keywords: Influence blocking maximization, refutation mechanism, competitive independent cascade model, reverse reachable set

1 Introduction

Online Social Networks (OSNs) such as Twitter, TikTok, Facebook, and Weibo have become essential tools for individuals to express opinions and disseminate information. While OSNs facilitate the wide dissemination of trustworthy information, they also promote the spread of rumors and other types of misinformation or negative information that can cause adverse social effects [1–7]. It is very necessary to find effective ways to block the spread of negative information, such as rumors and crises, to ensure the credibility and stability of online social networks.

Influence maximization is an important topic in social network research and could provide practical implications to promote the spread of positive behaviors and mitigate the circulation of rumors [8–17]. At present, existing studies have discussed the problem of rumor blocking from different perspectives: by removing a specific set of nodes from the network, the number of nodes to which negative information can be propagated is minimized [18–22], or removing some specific edges in the network to reduce the negative impact of the negative information on other normal nodes in the network [23–27], or selecting a specific seed set in the network to disseminate positive information to other nodes to reduce the number of nodes affected by negative information (which is called as influence blocking maximization problem) [28–31]. In this paper, we primarily concentrate on the Influence Blocking Maximization (IBM) problem. This problem involves the selection of a positive seed set, which propagates beneficial information to counteract the dissemination of negative information, ultimately minimizing the number of nodes affected by the activation of negative information.

However, almost all previous work on IBM focuses on the positive information propagation process from the selected positive seed nodes. In reality, blocking misinformation may not only come from the initial selection of positive seed users, but other users on the social network may also spontaneously refute misinformation and spread positive information to other users based on their knowledge background and experience. This phenomenon is called the refutation mechanism. For example, during the COVID-19 epidemic, a large amount of unconfirmed information was widely disseminated on social networks, but some users confirmed and refuted the misinformation and disseminated their positive views instead based on their knowledge and experience. Together with professional organizations that publish fact-checked

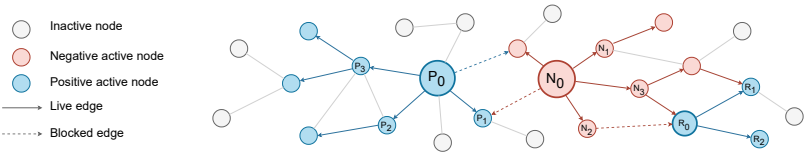

Fig. 1 An example of information spreading with refutation. N_0 is a negative seed node, P_0 is a positive seed node, and R_0 is the refuter. When R_0 is negatively activated by a negative node, it becomes a refuter and spread positive information to its neighbors.

information, these users act as a deterrent to the spread of misinformation. These users who refute rumors play a vital role in the early blocking of misinformation. However, this commonly seen phenomenon in OSNs is not well studied in the literature related to influence blocking maximization. In this paper, we incorporate the refutation mechanism into the competitive independent cascade model for the first time and study the impact of the refutation mechanism on the IBM problem.

We first combine the refutation mechanism with the Campaign-Oblivious Independent Cascade Model (COICM), which is a competitive independent cascade (CIC) model proposed in [28], to propose the Refutation Competitive Independent Cascade (RCIC) model. Taking Fig. 1 as an example: when R_0 is negatively activated by node N_3, R_0 becomes a refuter and propagates positive information to R_1 and R_2, which avoids R_1 and R_2 being influenced by negative information. Then, we consider the *Joint Influence Blocking Maximization* (JIBM) problem, which is an extension of the IBM problem on the RCIC model. The goal of the JIBM problem is choosing a positive seed set with a size of at most k to maximize the *joint influence blocking degree*, which is the number of nodes that are not activated by negative seed nodes (That is, the non-red nodes in Fig. 1). We show that the JIBM problem is NP-hard and the joint influence blocking degree is monotone and submodular, which makes it possible to implement efficient approximation algorithms using the reverse reachable (RR) set techniques [32, 33]. However, the complexity of our model makes the process of generating the RR set different from previous work significantly. The challenge of the JIBM problem is that we need to know which nodes have become refutation nodes when generating the RR set. To solve the problem, we make nontrivial adaptions to the generation of RR set, named RRGen, by using a two-phase generation algorithm. The first phase is the forwarding labeling phase for determining the refutation nodes, and the second one is the reverse sampling phase to generate a RR set from a random root node. Then, we design the approximation algorithm named RR-JIBM for the JIBM problem based on RRGen algorithm. We show that our algorithms solve the JIBM problem with an approximation ratio of $(1 - 1/e - \varepsilon)$ for any $\varepsilon > 0$. To further improve the efficiency of RR-JIBM, we propose the RR-JIBM+ algorithm by adding a reverse BFS phase to the generation process of the RR set, where a judgment is made on whether a reverse reachable set needs to be generated for a certain random root node. Finally, we compare RR-JIBM with the relevant baseline on real-world datasets. The experiments show that our

RR-JIBM algorithms can effectively minimize the number of nodes influenced by negative nodes. The RR-JIBM algorithm and the RR-JIBM+ algorithm on different datasets show that RR-JIBM+ has a higher efficiency advantage. The contributions of this paper are summarized as follows:

1. We combine the refutation mechanism with the CIC model for the first time and propose the RCIC model. On the basis of this model, we study the JIBM problem, whose goal is to maximize the expected number of nodes that are not activated by negative seed nodes by finding positive seed nodes with a size of at most k. We show that the JIBM problem is a NP-hard problem, and the joint influence blocking degree is monotone and submodular.
2. We design a two-phase generation algorithm, named RRGen, to generate the RR set in the RCIC model and propose the scalable approximation algorithm named RR-JIBM for the JIBM problem with an $(1 - 1/e - \varepsilon)$-approximation ratio for any $\varepsilon > 0$ respectively. The RR-JIBM+ algorithm is proposed to further improve the efficiency of the generation of RR set. Then, we compare them with other intuitive baselines on real-world social networks. The experiments show that our proposed algorithms can effectively reduce the number of nodes activated by negative seed nodes compared to the intuitive baselines and that RR-JIBM+ algorithm is more efficient than the RR-JIBM.

The rest of the paper consists of the following sections. In Section 2, we introduce the concept of the RCIC model and present the JIBM problem. In section 3, we analyze the properties of the joint influence blocking degree and design scalable algorithms for the JIBM problem. The experiment results are analyzed in Section 4. Finally, we summarize this paper and discuss further work in Section 5.

2 Diffusion Model and Problem Definition

In this section, we present the diffusion model used in this paper and propose the problems we study.

An online social network is usually modeled using a directed graph, named $G = (V, E)$, where V represents the users in the social network and E represents the relationship between users. For each node $v \in V$, let $N^{in}(v)$ and $N^{out}(v)$ denote v's *in-neighbors* and *out-neighbors* respectively. Moreover, we also use n and m to represent V and E, respectively.

2.1 Diffusion Model

Budak et al. [28] proposed the COICM, which address the competitive information propagation problem in a social network. In the COICM, a directed graph $G = (V, E)$ is used to represent social networks, where V is users in the social network and E is relationship between users. The COICM contains both positive and negative information cascade propagation processes from S_P and S_N, respectively. The positive and negative propagation probability of the edge

$(u, v) \in E$ are equal, denoted as $p(u, v)$. If a node is activated by positive information, its state is *positively active*, while its state is *positively active* if the node is activated by negative information, Otherwise, the state of the node is *inactive*. A node does not change its state after it has been activated. At time 0, the nodes in S_P are positively active while the nodes in S_N are negatively active. All other nodes are inactive. If a node u is activated by its neighbors positively (negatively) at exactly time t, then for each inaction out-neighbor node v it has one chance to activate a neighbor node with probability $p(u, v)$ at time $t + 1$, In COICM, the positively active neighbors have higher priority, that means when positively active and negatively active in-neighbors of v try to activate v at a given time, v always activated by positive information. The information propagation process terminates when no new nodes are activated positively or negatively.

We now extend the COICM to incorporate the refutation mechanism. The idea is that we allow every inactive node to have a single chance to refute the negative information and spread positive information when it is first activated by one of its negatively active neighbors. In detail, in our model, when a node is activated by one of its negatively active neighbors at time t, it has a single chance to determine whether it becomes a refuter. If it determines to become a refuter, that means it will become a positively active node after a random delay and spread positive information to its neighbors. In reality, whether a node refutes the negative information is based on differences in the nodes' personality, knowledge background, and experience. Here, we simplify the problem, the following set of parameters governs the refutation probability of a node.

Definition 1. *In graph $G = (V, E)$, every node $u \in V$ can be a refuter with refutation probability $q(u)$ when it is activated by one of its negatively active neighbors. If so, it becomes positively active after a random delay $\delta(u)$ sampled from a refutation delay distribution Δ and spread the positive information to its neighbors.*

We further consider the real-time delay on edges based on the refutation mechanism and obtain the refutation competitive independent cascade (RCIC) model. That means that for each edge $(u, v) \in E$, there is a propagation delay distribution $D(u, v)$ corresponding to it. We also assume that the distributions in both Δ and D are continuous. Thus, at any time t, for any node $u \in V$, at most one neighbor tries to activate it. Based on the above assumptions, the information dissemination process in the RCIC model is as follows.

1. For nodes that do not belong in $S_N \cup S_P$, if it is activated by one of its positively active neighbors at time t, it becomes a positively active node at time t.
2. For a node u that is not belong in $S_N \cup S_P$, if it is activated by one of its negatively active neighbors at time t, it becomes a negatively active node at time t with probability $1 - q(u)$. Otherwise, it becomes a positively active

node after a random delay $\delta(u) \sim \Delta(u)$ and spread positive information to its neighbors, unless it is activated by other positive active neighbors before time $t + \delta(u)$.

3. For a node $u \in V$ that is activated at time t, for each of its inactive outgoing neighbors v it has one chance to activate v with probability $p(u, v)$ after a delay $d(u, v) \sim D(u, v)$, unless v is activated by other neighbors before time $t + d(u, v)$.

Compared with the classical competitive independent cascade model, our RCIC model considers the refutation mechanism and the real-time delay in the negative and positive information propagation process. The refutation mechanism models realistic scenarios where some users may refute the negative information and spread positive information to others based on their knowledge, experience, and personality. The refutation delay parameters model the time difference between a user decides to refute negative information and disseminates positive information to other nodes. Note that when the refutation probabilities and refutation delays of all nodes are 0, and propagation delays of all edges are 1, the RCIC model falls back to the classical COICM [28].

To derandomize the influence propagation process and facilitate a better understanding of the RCIC model, we use *possible world model* similar with [34] to describe the RCIC model. We say that a direct edge $(u, v) \in E$ is *live* if u can activate v through (u, v) when u is active. In the RCIC model, the randomness of the influence propagation process comes from the following aspects: 1) the states of the nodes, where each node u can be a refuter with probability $q(u)$; 2) the refutation propagation delays of the refuters, which are drawn from the refutation delay distribution Δ; 3) the states of the edges, where each edge $(u, v) \in E$ can be live with probability $p(u, v)$; and 4) the propagation delays of the edges, which are drawn from propagation delay distribution D. We use $\mathcal{W}(p, q, D, \Delta)$ to denote the set of all possible worlds. To generate a possible world W, we first sample every edge in $(u, v) \in E$ with probability $p(u, v)$ and its corresponding propagation delay $d(u, v) \sim D(u, v)$ to generate a live graph $G_W = (V, E_W)$. For each node $u \in V$, we sample a probability $\alpha(u) \in [0, 1]$ and its corresponding refutation delay $\delta(u) \sim \Delta(u)$. If $\alpha(u) \leq q(u)$, u is called a candidate refutation node. A possible world $W = (L_W, C_W, d_W, \delta_W)$ sampled from $\mathcal{W}$ with probability $P[W|\mathcal{W}]$ can be seen as a tuple where L_W is the set of live edges E_W, C_W is the set of candidate refutation nodes, d_W is the propagation delays of edges in L_W, and δ_W is the refutation delays of the candidate refutation nodes in A_W.

We use F to represent that the source of the information a node receive is from S_N and that the information received is negative information, while T to represent that the source of the information a node receive is from S_N and that the information is positive information. In possible world W, for a node u, let $T_W(S_N, u, F)$ $(T_W(S_P, u))$ present the length of the shortest path from any negative (positive) seed to u consisting of entirely negatively (positively) active nodes, and $T_W(S_N, u, T)$ represent the length of the shortest path from any negative seed to u consisting of at least one positively active node. At

time 0, propagation starts from S_N and S_P and follows the direction of the live edges. For each step $t > 0$, an inactive node u is reachable by negative information if $T_W(S_N, u, F) = t$. The node u then becomes a positively active node at time $t + \delta(u)$ if u is a candidate refutation node and if not, u becomes a negatively active node at time t. Similarly, an inactive node u is reachable by positive information if $min(T_W(S_P, u), T_W(S_N, u, T)) = t$, and u becomes positively active at time t.

2.2 Joint Influence Blocking Maximization Problem

Next, we propose *Joint Influence Blocking Maximization* (JIBM) problem, which is an extrapolation of the IBM problem on the RCIC model. The objective function of the problem is to choose a specific set of positive seeds in the RCIC model to propagate positive information so that the number of nodes that are not activated by negative seed nodes is maximized.

Given the negative seed set S_N, for a specific possible world W, let $\sigma_W(S_P, S_N)$ represent the number of nodes are not negatively active when S_P is the positive seed set in W. Let $\sigma_W(S_P, S_N, u) = 1$ if u is not negatively active in W, $\sigma_W(S_P, S_N)$ can be further expressed as $\sum_{u \in V} \sigma_W(S_P, S_N, u)$. We call $\sigma(S_P, S_N) = \mathbb{E}_{W \sim \mathcal{W}}[\sigma_W(S_P, S_N)]$ the *joint influence blocking degree*, which is the expected number of nodes are not negatively active when S_P is the positive seed set under the RCIC model. Because we always use S_N to represent the negative seed set, so S_N will be omitted from $\sigma(S_P, S_N)$. Finally, $\sigma(S_P)$ can be calculated using the following formula:

$$\sigma(S_P) = \sum_{u \in V} \sum_{W \sim \mathcal{W}} P[W|\mathcal{W}] \sigma_W(S_P, u) \tag{1}$$

The objective function of the JIBM problem is to choose a positive seed set S_P of size at most k such that $\sigma(S_P)$ is maximized.

Definition 1 (Joint Influence Blocking Maximization Problem)**.** *Given a directed graph $G = (V, E)$, the negative seed set S_N, the refutation probability q of nodes, the refutation delay distribution Δ of nodes, the propagation probability p of edges, the propagation delay distribution D of edges, and the size of positive seeds to be selected k, the* Joint Influence Blocking Maximization *(JIBM) problem aims to find an optimal positive seed set S_P^* of size at most k, such that $\sigma(S_P^*)$ is maximized, i.e., $S_P^* = \text{argmax}_{S\_P \leq k} \sigma(S_P)$.*

The difference between the classical IBM problem [29] and our JIBM problem is that the JIBM problem considers the refutation nodes' role when selecting positive seed nodes. Because the nodes activated by refutation nodes do not need to be activated by positive seed nodes again. In this problem, in a fixed possible world, refutation nodes can make some nodes unaffected by negative information. Therefore, we use the term "joint" to specify that the refutation nodes and the selected positive seed nodes play a joint role in blocking the spread of negative information.

As mentioned above, when setting both the refutation probability and the refutation delay of each node to 0 and setting both the propagation delay on the edges to 1, the RCIC model falls back to the COICM. In this case, the JIBM problem and the IBM problem are equivalent, so the JIBM problem is NP-hard, as the IBM problem is NP-hard under the COICM.

3 Scalable Algorithm Design

In this section, we first show that the joint influence blocking degree satisfies the properties of monotone and submodular, which are crucial for the design of our algorithm. Then, we develop scalable algorithms for the JIBM problem.

Lemma 1. *$\sigma(S_P)$ is monotone submodular for S_P under the RCIC model.*

Proof For a set function $f : 2^V \to \mathbb{R}$, given any subset V_1 and V_2 of V and satisfying $V_1 \subseteq V_2 \subseteq V$, for any $v \in V \setminus V_2$, if $f(V_1 \cup \{v\}) - f(V_1) \geq f(V_2 \cup \{v\}) - f(V_2)$, we say that f is *submodular*. While f is *monotone* if $f(V_1) \leq f(V_2)$ for any $V_1 \subseteq V_2 \subseteq V$. According to the above definition, given a positive seed set S_P and a possible world W, because $T_W(S_P, u)$ represent the shortest distance between S_P and u, when we select more nodes to add to S_P, $T_W(S_P, u)$ does not increase. So when more nodes are selected to join S_P, $\sigma_W(S_P)$ does not increase. So $\sigma(S_P)$ is monotone. Now we prove the submodularity. For any two positive seed set S_1 and S_2 selected from V and satisfying $S_1 \subseteq S_2 \subseteq V$, for any $v \in V \setminus S_2$, we need to prove $\sigma_W(S_1 \cup \{v\}, u) - \sigma_W(S_1, u) = 1$ when $\sigma_W(S_2 \cup \{v\}, u) - \sigma_W(S_2, u) = 1$. When $\sigma_W(S_2, u) = 0$, u is negatively activated by S_N, which means $T_W(S_N, u, F) < min(T_W(S_2, u), T_W(S_N, u, T))$. Meanwhile, $\sigma_W(S_2 \cup \{v\}, u) = 1$ means that when we add v to S_2, $T_W(S_2 \cup \{v\}, u) < T_W(S_N, u, F)$. Therefore, $T_W(\{v\}, u) < T_W(S_N, u, F)$ and consequently $\sigma_W(S_1 \cup \{v\}, u) = 1$. Furthermore, $\sigma_W(S_1, u) = 0$ because $S_1 \subseteq S_2$ and $\sigma_W(S_2, u) = 0$. So $\sigma_W(S_1 \cup \{v\}, u) - \sigma_W(S_1, u) = 1$ is proved. So $\sigma_W(S_1, u)$ is monotone submodular. According to 1, $\sigma(S_P)$ also monotone submodular. □

3.1 RR-JIBM

Based on the monotone submodular property and the reverse reachable set technique, we can design efficient approximation algorithms for the JIBM problem.

Definition 2 (Reverse Reachable Set)**.** *A reverse reachable (RR) set [33] $R(u)$ is the set of nodes that can be reached by reverse sampling from the root node u. Given a possible world W, the reverse reachable set rooted at $R_W(u)$ is the set of nodes that can reach u in W. We use R (R_W) to represent a random RR set if the root node of R (R_W) is selected at random from V.*

For a node set $S \subseteq V$ and a random RR set R, let $x(S, R)$ be a indicator function and $x(S, R) = 1$ iff the root node of R is not negatively active. We can

Algorithm 1 Forward labeling phase

Input: $G = (V, E)$, q, Δ, p, D, S_N, root node r
Output: The state of root node r
$Q \leftarrow S_N$;
for $v \in V$ **do**
    $delay[v] \leftarrow +\infty$;
    $state[v] \leftarrow inactive$;
end for
for $v \in S_N$ **do**
    $delay[v] \leftarrow 0$; $state[v] \leftarrow negative$;
end for
while $Q \neq \emptyset$ **do**
    $w \leftarrow argmin_{w' \in Q} delay[w']$;
    if $state[w] = shadow$ **then**
        $state[w] \leftarrow positive$; ▷ label w as a refuter
    end if
    if $state[w] = negative \wedge w \notin S_N \wedge w$ is refuted with probability $q(w)$ **then**
        $state[w] \leftarrow shadow$;
        $delay[w] \leftarrow delay[w] + \delta(w)$;
        continue;
    end if
    delete w from Q;
    for $v \in N^{out}[w]$ **do**
        if $state[w] = negative \wedge state[v] = shadow$ **then**
            continue;
        end if
        if (w, v) is none **then**
            label (w, v) as live in probability $p(w, v)$, otherwise blocked;
        end if
        if (w, v) is live **then**
            insert v into Q if $delay[v] = +\infty$;
            if $d(w, v) + delay[w] < delay[v]$ **then**
                $delay[v] \leftarrow delay[w] + d(w, v)$;
                $state[v] \leftarrow state[w]$;
            end if
        end if
    end for
end while
return $state[r]$;

obtain the following connection between the joint influence blocking degree $\sigma(S_P)$ and the random RR set R.

Lemma 2. $\sigma(S_P) = n * \mathbb{E}_{W \in \mathcal{W}}[x(S, R_W)]$.

Proof Let $g(S, R_W, u)$ be a indicator function and $g(S, R_W, u) = 1$ if u is the root node of R_W and $x(S, R_W) = 1$. It is clear that $g(S, R_W, u) = 1$ iff $\sigma_W(S, u) = 1$, so, we have the following formula:

$$\begin{aligned} \mathbb{E}_{W \in \mathcal{W}}[x(S, R_W)] &= \frac{1}{n} \sum_{u \in V} \mathbb{E}_W[g(S, R_W, u)] \\ &= \frac{1}{n} \sum_{u \in V} \mathbb{E}_W[\sigma_W(S, u))] \\ &= \frac{1}{n} \sigma(S) \end{aligned} \tag{2}$$

□

To generate the RR set, we design a novel sampling process that conceptually organized in three phases. The first phase samples a possible world W from $\mathcal{W}$. The second phase is a forward labeling phase starting from the negative seed nodes, which determines refutation nodes from candidate refutation nodes by running a Dijkstra algorithm following the outgoing edges. The last phase is a reverse sampling phase starting from a random root node r, which generates an RR set from a random root node r by running the backward Dijkstra algorithm following the incoming edges. However, sampling the entire possible world is not necessary because the propagation of positive and negative information is unlikely to reach the entire graph. In light of this observation, we do not need to generate the entire possible world in the first phase. Instead, we only need to reveal the state of an edge or node when that edge or node is visited for the first time during the forward labeling phase and the reverse sampling phase.

Forward labeling phase

Algorithm 1 shows that the forward labeling phase. The main idea is that we introduce the novel *shadow* state for all nodes $u \in V$ to represent the intermediate state after a node decides to refute negative information and before the node propagates positive information to other neighbors. We use $state[u]$ represent the state of u and $delay[u]$ is represented as the shortest delay from negative seed nodes to node u. During forward labeling procedure, we use the Dijkstra algorithm to always select the node w with the shortest delay in the current set of candidate nodes Q (line 12). At this point, if the state of w is shadow, w will become positively active (line 14). If w is a negatively active node, and it determines to refute negative information (line 16) with probability $q(w)$, we first update $state[w]$ to shadow. Then, we sample the refutation delay $\delta(w)$, update $delay[w]$ and select the next candidate node. Otherwise, we visit all live outgoing edge (w, v) with edge delay $d(w, v)$, and do updates for $delay[v]$ and $state[v]$.

Reverse sampling phase

Algorithm 2 shows that the generation of RR set. After the forward labeling phase, we first check the state of the root node r. We do not need to generate

RR set for the root node if its state is not negatively active. This includes two cases, one that the refuters positively activate the root node, or that the root node is inactive. In both cases, we do not need to select a seed node to cover it. In this situation, we return $\emptyset$ as the RR set of r. If the state of r is negatively active, we apply the Dijkstra algorithm for incoming edges and generate a RR set. In this algorithm, we use $delay[u]$ to represent the delay from the random root node r to node u. In the process of generating the RR set, we always select the node w with the shortest delay from the candidate node set Q. We insert node w into R if w is not a negative seed node. For each incoming edge (v, w) of w, if the edge is not determined to be live or blocked, we first determine the state of the edge. If the edge is live, we update $delay[v]$ with edge delay $d(v, w)$. The procedure ends when a negative seed node is reached, and the set R is the RR set for the root node r.

The IMM algorithm [33] is a classical and effective algorithm to solve the influence maximization problem. In this paper, We adjust the IMM algorithm to solve the JIBM problem and get Algorithm 3. In Phase 1, we first calculate the lower bound and calculate the number of RR sets according to the lower bound to ensure high probability approximation. We generate random RR sets $\mathcal{R}$ according to the algorithm introduced above. In Phase 2, the greedy algorithm *NodeSelection* proposed in [33] is used to find at most k positive seed nodes that cover as many RR set in $\mathcal{R}$ as possible. Algorithm 3 differs from the IMM algorithm in that $\mathcal{R}$ only contains the RR set generated from the negatively activated root nodes. We use *protected* to denote the number of RR sets that $\emptyset$. By Eq. 2, $\sigma(S_P)$ can be estimated as n times the fraction of RR sets that their root node is not negatively active. After selecting a positive seed node set, the fraction can be calculated as $F_{\mathcal{R}}(S)$:

$$F_{\mathcal{R}}(S_P) = \frac{protected + \sum_{R \in \mathcal{R}} \mathbf{I}(S_P \cap R \neq \emptyset)}{protected + |\mathcal{R}|}. \tag{3}$$

As described in [33], IMM algorithm returns an $(1-1/e-\varepsilon)$-approximation ratio with at least $1 - \frac{1}{n^\ell}$ probability and runs in $O(\frac{(k+\ell)(n+m)\log n}{\varepsilon^2} \cdot \frac{EPT}{OPT})$ expected time, where $O(EPT)$ is the expected time for generating a random RR set. The big difference between IMM algorithm and RR-JIBM algorithm is the complexity of generating RRset. Our RR-JIBM algorithm need a two phase algorithm to generate the RRset, so the expected time for generating a random RR set is the sum of the expect time of each phase. So, we have the following theorem.

Theorem 1. *Let S_P^* be the optimal positive seed nodes of the JIBM problem, For every $\varepsilon > 0$ and $\ell > 0$, with probability at least $1 - \frac{1}{n^\ell}$, the positive seed nodes S_P^o selected by Algorithm 3 satisfies*

$$\sigma(S_P^o) \geq (1 - \frac{1}{e} - \varepsilon)\sigma(S_P^*).$$

Algorithm 2 RRGen for JIBM problem

Input: $G = (V, E)$, q, Δ, p, D, S_N, root node r
Output: RR set R
$state \leftarrow$ Algorithm 1; ▷ forward labeling phase
if $state \neq negative$ **then**
  return $\emptyset$; ▷ reverse sampling phase
end if
$R \leftarrow \emptyset$;
$Q \leftarrow \{r\}$;
for $u \in V$ **do**
  $delay[u] \leftarrow +\infty$;
end for
$delay[r] \leftarrow 0$;
while $Q \neq \emptyset$ **do**
  $w \leftarrow argmin_{w' \in Q} delay[w']$;
  if $w \in S_N$ **then**
    break;
  end if
  delete w from Q;
  $R \leftarrow R \cup \{w\}$;
  for $v \in N^{in}[w]$ **do**
    if (v, w) is none **then**
      label edge (v, w) as live with probability $p(v, w)$, otherwise blocked;
    end if
    if (v, w) is live **then**
      insert v into Q if $delay[v] = +\infty$;
      if $d(v, w) + delay[w] < delay[v]$ **then**
        $delay[v] \leftarrow d(v, w) + delay[w]$;
      end if
    end if
  end for
end while
return R;

In this case, the expected running time for Algorithm 3 is $O(\frac{(k+\ell)(n+m)\log n}{OPT\varepsilon^2} \cdot (EPT_F + EPT_B))$*, where* EPT_F *and* EPT_B *is the expected time in forward labeling phase and reverse sampling phase respectively.*

3.2 RR-JIBM+

Algorithm 2 may incur redundant computations when the forward labeling phase and the reverse sampling phase can be skipped entirely if the root node and the negative seed nodes are in different connected components. In this case, the negative seed nodes cannot affect the root node. To take advantage of

Algorithm 3 RR-JIBM

Intput: $G = (V, E)$, q, Δ, p, D, k, S_N, ε, ℓ
Output: positive seed set S
$\mathcal{R} \leftarrow \emptyset$; $LB \leftarrow 0$; $\varepsilon' \leftarrow \sqrt{2}\varepsilon$; $protected \leftarrow 0$;
find γ that satisfies $\lceil \lambda^*(\ell) \rceil / n^{\gamma+\ell} \leq 1/n^{\ell}$; ▷ Workaround 2 in [35]
$\ell \leftarrow \ln 2 / \ln n + \ell + \gamma$;
for $j = 1$ to $\log_2(n-1)$ **do**
    $z_j \leftarrow n/2^j$; $\theta_j \leftarrow \lambda' / z_j$;
    while $|\mathcal{R}| + protected < \theta_j$ **do**
        randomly select a node from V as the root node r;
        generate the RR set R from r using Algorithm 2;
        if $R \neq \emptyset$ **then**
            insert R into $\mathcal{R}$;
        else
            $protected \leftarrow protected + 1$;
        end if
    end while
    $S_j \leftarrow NodeSelection(\mathcal{R}, k)$;
    if $n \cdot F_{\mathcal{R}}(S_j) \geq (1 + \varepsilon') \cdot z_j$ **then**
        $LB \leftarrow n \cdot F_{\mathcal{R}}(S_j) / (1 + \varepsilon')$;
        break;
    end if
end for
$\theta \leftarrow \lambda^*(\ell) / LB$;
while $|\mathcal{R}| + protected \leq \theta$ **do**
    randomly select a node from V as the root node r;
    generate the RR set R from r using Algorithm 2;
    if $R \neq \emptyset$ **then**
        insert R into $\mathcal{R}$;
    else
        $protected \leftarrow protected + 1$;
    end if
end while
$S \leftarrow NodeSelection(\mathcal{R}, k)$;
return S;

this observation, we propose Algorithm 4 to further improve the computation efficiency. The critical thought is to run a reverse BFS starting from the random root node r to determine whether the negative seed nodes can influence r. We can skip the forward labeling phase and reverse sampling phase entirely if the root node r cannot be influenced by the negative seed nodes. If we need to perform the forward labeling phase and reverse sampling phase, we only need to consider the live edges visited in the reverse BFS phase.

Backward BFS phase.

In this phase, we use a FIFO queue Q to store the nodes that will be visited in the future. We first enqueue the root node r into Q. Then we dequeue a node w from Q and test the live/blocked status for each incoming edge (v, w) independently by its propagation probability $p(v, w)$ independently. If the state of edge (v, w) is live and node v has not been reached before, we enqueue v into Q. Let U represent the nodes explored in this phase. If $U \cap S_N = \emptyset$, that means that no negative information can be propagated from negative seed nodes to root node r, so that the forward labeling phase and reverse sampling phase can be entirely skipped. Otherwise, we run Algorithm 2 only from $U \cap S_N$, along the explored live edges.

By replacing Algorithm 2 in Algorithm 3 with Algorithm 4, we can obtain the RR-JIBM+ algorithm, where we omit the pseudocode of the algorithm. These two algorithms have the same approximation guarantee. The analysis on expected time complexity is similar. Because the expected time for generating a random RR set is the sum of the expect time of each phase, the expected running time for the RR-JIBM+ algorithm is $O(\frac{(k+\ell)(n+m)\log n}{OPT\varepsilon^2} \cdot (EPT_{B1} + EPT_{F1} + EPT_{B2}))$, where EPT_{B1} and EPT_{B2} is the expected time of backward BFS phase and reverse sampling phase respectively, and EPT_{F1} is the expected time of forward labeling phase.

Algorithm 4 RRGen+ for JIBM problem

Input: $G = (V, E)$, q, Δ, p, D, S_N, root node r
Output: RR set R
$Q \leftarrow \{r\}$;
$U \leftarrow \emptyset$;
while $Q \neq \emptyset$ **do**
    dequeue w from Q;
    for $v \in N^{in}[w]$ **do**
        if (v, w) is none **then**
            label (v, w) as live with probability $p(v, w)$, otherwise blocked;
        end if
        if (v, w) is live $\wedge$ v is not visited **then**
            enqueue v into Q;
            $U \leftarrow U \cup \{v\}$; ▷ label v visited
        end if
    end for
end while
$R \leftarrow \emptyset$;
if $U \cap S_N \neq \emptyset$ **then**
    $R \leftarrow$ Algorithm 2 with negative seed set $U \cap S_N$;
end if
return R;

It should be noted that there is no theoretical guarantee that the RR-JIBM+ algorithm will always save time compared to the RR-JIBM algorithm. The worst case of Algorithm 4 used in RR-JIBM+ is that $U \cap S_N = S_N$, which means that the reverse BFS phase is wasted. In this case, the RR-JIBM+ algorithm will be worse than the RR-JIBM. But if the negative seed node is not able to affect the randomly selected root node, then RR-JIBM+ is better, because in this case the simple BFS is better than the complex Dijkstra algorithm.

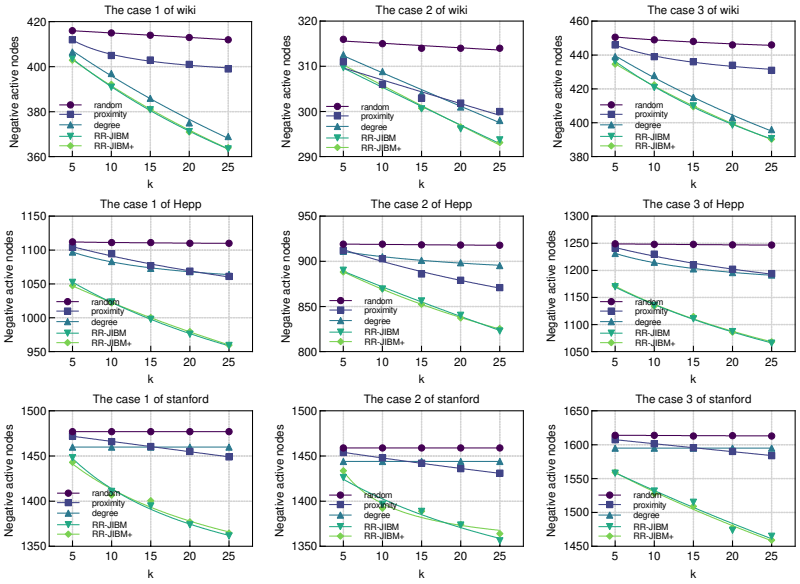

Fig. 2 The number of negatively active nodes for different algorithms in different cases

4 Empirical Evaluation

To evaluate the effectiveness and efficiency of our algorithm, we conducted experiments on three real-world networks. All experiments are carried out on a computer running Ubuntu 20.04 under Intel Xeon Platinum 8176 Processor and 62G memory. All experiments code is written in C++ and compiled with clang++ with -O2.

4.1 Experiment setting

We describe the datasets, algorithms, and parameter settings in our experiments.

4.1.1 Datasets

We use the following three datasets. *wiki*, *Hepp*, and *stanford*. They all come from SNAP [1]. The statistical information for datasets is given in Table 1.

4.1.2 Algorithms

We test our RR-JIBM algorithm and RR-JIBM+ algorithm with the following baselines.

- **Random:** selecting nodes at random as positive seed nodes.
- **Degree:** selecting the k nodes with the greatest degrees as positive seed nodes.
- **Proximity:** selecting the k nodes with the greatest negative activation probability among the out-neighbors of the negative seed nodes as positive seed nodes.
- **CIMM:** similar with Algorithm described in [36], find the positive seed nodes on competitive independent cascade model without refutation mechanism.

Table 1 Dataset description

Dataset	Node	Edge	Avg out-Degree	Max out-Degree
wiki	7K	103K	14.5	893
Hepp	34K	421K	12.2	411
stanford	281K	2M	8.2	255

4.1.3 Parameters

The negative seed nodes are the 25 nodes with the highest degree, and we test positive seed nodes sizes of 5,10,15,20, and 25. We use the same parameters setting for RR-JIBM and RR-JIBM+ algorithms: $\ell = 1$, $\varepsilon = 0.1$. For the datasets we use, the parameters cannot be estimated, so we use synthetic settings. For the propagation probabilities, we follow *WC model* [9]: $q(u, v) = 1/\mathsf{N\hat{}-in"[v]}$ for edge $(u, v) \in E$. We first choose a random value $c(v)$ from $\{0.1, 0.05, 0.01\}$ as the base value of the refutation probability $q(v)$ of node v. To consider various scenarios of refutation probability, we further consider the following three cases:

1. **Case 1:** $q(v) = c(v)$.
2. **Case 2:** $q(v) = min\{\beta c(v)\mathsf{N\hat{}-out"[v]}, 1\}$, where β is a constant value and $\beta = 0.01$ in all experiments. This means $q(v)$ is positively correlated with v's out-degree $\mathsf{N\hat{}-out"[v]}$.
3. **Case 3:** $q(v) = c(v)/\mathsf{N\hat{}-out"[v]}$. If v's out-degree is 0 we set $q(v) = 0$. This means $q(v)$ is negatively correlated with v's out-degree $\mathsf{N\hat{}-out"[v]}$.

[1] http://snap.stanford.edu/data/index.html

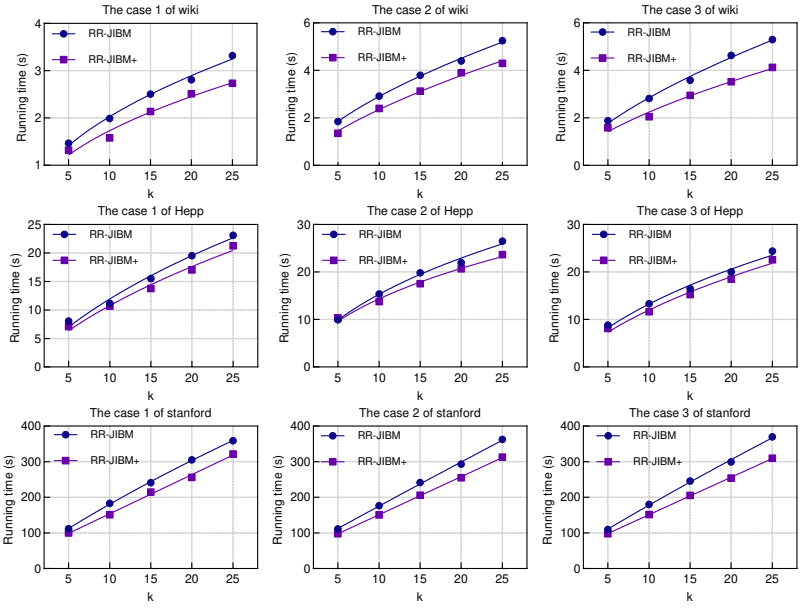

Fig. 3 Running time of *RR-JIBM* and *RR-JIBM+* in different cases

For refutation delays and propagation delays, we use exponential distribution with rate 1 for all nodes and edges.

4.2 Experiment results

We evaluate our algorithms from two metrics: blocking nodes and running time.

4.2.1 Influence Blocking

As we can see in Fig. 2, our algorithms significantly outperform other intuitive baselines in all cases across various datasets. Furthermore, the advantage of our algorithm becomes increasingly apparent as k increases, as demonstrated on the *Hepp* and *stanford* datasets. For CIMM algorithms that do not consider the refutation mechanism, they are in most cases worse than those that consider the refutation mechanism. This is because after considering the refutation nodes, those nodes that will be activated by the refutation nodes do not need to select additional seed nodes. In case 3 of each dataset, the gap between CIMM and the other considered refutation mechanism is smaller than the other cases, because in this case, the refutation probability of nodes with larger degree will be smaller than that of nodes with smaller degree, which is reflected in the experiment that the effect of refutation is weaker than the other two cases.

4.2.2 Running Time

We compared the efficiency of the RR-JIBM and RR-JIBM+ algorithms in different cases. As shown in Figure 3, running time increases with the number of selected seed nodes increases, and RR-JIBM+ algorithm outperforms RR-JIBM algorithm in almost all situations. This is because if the selected root node and the negative seed node are not connected, then it is much faster to determine whether they are connected by a reverse BFS than to simulate information propagation using the Dijkstra algorithm. Furthermore, as the number of seed nodes increases, the gap between the two gradually becomes larger.

5 Conclusion and Future Work

This paper studies the influence blocking maximization problem named the JIBM problem under a competitive independent cascade model that incorporates a refutation mechanism. We design the scalable approximation algorithm RR-JIBM based on a reverse reachable set and propose RR-JIBM+ based on the new generation strategy. Experiments on several real datasets show that the effectiveness and efficiency of proposed algorithms. Our model is designed for information maximization with a refutation mechanism. However, information propagation in networks can be further considered for more complex situations, such as the propagation of positive messages that may also turn into negative messages during message propagation, which can be considered in future studies. Finally, incorporating refutation mechanisms into other rumor-blocking tasks or influence propagation models and designing more efficient algorithms still need to be determined as future work.

Declarations

Ethical Approval: not applicable.
Competing Interests: We declare that the authors have no known competing interests or personal relationships that might be perceived to influence the discussion reported in this paper.
Authors' contributions: Qi Luo and Dongxiao Yu wrote the manuscript, Yafei Zhang and Yanwei Zheng collected the data, Dongbiao Wang designed the experiments, and Zhipeng Cai analyzed the results. All authors reviewed the results and approved the final version of the manuscript.
Funding: National Key Research and Development Program of China under Grant 2020YFB1005900, National Natural Science Foundation of China (NSFC) under Grant 62122042, Shandong University multidisciplinary research and innovation team of young scholars under Grant 2020QNQT017.
Availability of data and materials: All the datasets can be accessed from http://snap.stanford.edu/data/index.html.

References

[1] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)

[2] Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O.: The diffusion of microfinance. Science **341**(6144), 1236498 (2013)

[3] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the National Academy of Sciences **113**(3), 554–559 (2016)

[4] Paluck, E.L., Shepherd, H., Aronow, P.M.: Changing climates of conflict: A social network experiment in 56 schools. Proceedings of the National Academy of Sciences **113**(3), 566–571 (2016)

[5] Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.-C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature Communications **9**, 4787 (2018)

[6] Jones, N.M., Thompson, R.R., Schetter, C.D., Silver, R.C.: Distress and rumor exposure on social media during a campus lockdown. Proceedings of the National Academy of Sciences **114**(44), 11663–11668 (2017)

[7] Bovet, A., Makse, H.A.: Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications **10**(1), 7 (2019)

[8] Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66. Association for Computing Machinery, ??? (2001)

[9] 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)

[10] Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208. ACM Press, ??? (2009)

[11] Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., Zhou, T.: Identifying influential nodes in complex networks. Physica A: Statistical Mechanics and its Applications **391**(4), 1777–1787 (2012)

[12] Morone, F., Makse, H.A.: Influence maximization in complex networks through optimal percolation. Nature **524**(7563), 65–68 (2015)

[13] Ren, X.-L., Gleinig, N., Helbing, D., Antulov-Fantulin, N.: Generalized network dismantling. Proceedings of the National Academy of Sciences **116**(14), 6554–6559 (2019)

[14] Fan, C., Zeng, L., Sun, Y., Liu, Y.-Y.: Finding key players in complex networks through deep reinforcement learning. Nature Machine Intelligence **2**(6), 317–324 (2020)

[15] Braunstein, A., Dall'Asta, L., Semerjian, G., Zdeborová, L.: Network dismantling. Proceedings of the National Academy of Sciences **113**(44), 12368–12373 (2016)

[16] Mugisha, S., Zhou, H.-J.: Identifying optimal targets of network attack by belief propagation. Phys. Rev. E **94**, 012305 (2016)

[17] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nature Physics **6**(11), 888–893 (2010)

[18] Wang, B., Chen, G., Fu, L., Song, L., Wang, X.: Drimux: Dynamic rumor influence minimization with user experience in social networks. IEEE Transactions on Knowledge and Data Engineering (2017)

[19] Yan, R., Li, D., Wu, W., Du, D.Z.: Minimizing influence of rumors by blockers on social networks. In: CSoNet (2018)

[20] Yan, R., Li, D., Wu, W., Du, D.Z., Wang, Y.: Minimizing influence of rumors by blockers on social networks: Algorithms and analysis. IEEE Transactions on Network Science and Engineering (2020)

[21] Yao, Q., Guo, L.: Minimizing the social influence from a topic modeling perspective. In: ICDS (2015)

[22] Yao, Q., Shi, R., Zhou, C., Wang, P., Guo, L.: Topic-aware social influence minimization. Proceedings of the 24th International Conference on World Wide Web (2015)

[23] Kimura, Masahiro and Saito, Kazumi and Motoda, Hiroshi: Minimizing the spread of contamination by blocking links in a network. In: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2 (2008)

[24] Masahiro Kimura and Kazumi Saito and Hiroshi Motoda: Blocking links to minimize contamination spread in a social network. ACM Trans.

Knowl. Discov. Data (2009)

[25] Medya, S., da Silva, A.L., Singh, A.K.: Approximate algorithms for data-driven influence limitation. IEEE Transactions on Knowledge and Data Engineering (2020)

[26] Yao, Q., Zhou, C., Xiang, L., Cao, Y., Guo, L.: Minimizing the negative influence by blocking links in social networks. In: ISCTCS (2014)

[27] Zhu, J., Ni, P., Wang, G.: Activity minimization of misinformation influence in online social networks. IEEE Transactions on Computational Social Systems (2020)

[28] 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)

[29] He, X., Song, G., Chen, W., Jiang, Q.: Influence blocking maximization in social networks under the competitive linear threshold model. In: SDM (2012)

[30] Tong, G., Du, D.-Z.: Beyond uniform reverse sampling: A hybrid sampling technique for misinformation prevention. (2019)

[31] Wu, P., Pan, L.: Scalable influence blocking maximization in social networks under competitive independent cascade models. Computer Networks (2017)

[32] Borgs, C., Brautbar, M., Chayes, J., Lucier, B.: Maximizing social influence in nearly optimal time. In: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (2014)

[33] Tang, Y., Shi, Y., Xiao, X.: Influence maximization in near-linear time: A martingale approach. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (2015)

[34] Lu, W., Chen, W., Lakshmanan, L.V.S.: From competition to complementarity: Comparative influence diffusion and maximization (2015)

[35] Chen, W.: An issue in the martingale analysis of the influence maximization algorithm imm. In: Computational Data and Social Networks (2018)

[36] Tong, G., Wu, W., Guo, L., Li, D., Liu, C., Liu, B., Du, D.-Z.: An efficient randomized algorithm for rumor blocking in online social networks. IEEE Transactions on Network Science and Engineering **7**(2), 845–854 (2020)