Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Susceptible User Search for Defending Opinion Manipulation

Wenyi Tang*, Ling Tian*, a, c, Xu Zheng a, c, Guangchun Luo b, c, Zaobo He d

a School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, PR China
b School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, PR China
c Trusted Cloud Computing and Big Data Key Laboratory of Sichuan Province, Chengdu, Sichuan, 611731, PR China
d Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, 45056, USA

Abstract
The development of cyberspace offers unprecedentedly convenient access to online communication, thus inducing malicious individuals to subtly manipulate user opinions for benefits. Such malicious manipulations usually target those influential and susceptible users to mislead and control public opinion, posing a bunch of threats to public security. Therefore, an intelligent and efficient searching strategy for targeted users is one prominent and critical approach to defend malicious manipulations. However, the major body of current studies either provide solutions under ideal scenarios or offer inefficient solutions without guaranteed performance. As a result, this work adopts the combination of unsupervised learning and heuristic search to discover susceptible and key users for defense. We first propose a greedy algorithm fully considering the susceptibilities of different users, then adopt unsupervised learning and utilize the community property to design an accelerated algorithm. Moreover, the approximation guarantees of both greedy and community-based algorithms are systematically analyzed for some practical circumstances. Extensive experiments on real-world datasets demonstrate that our algorithms significantly outperform the state-of-the-art algorithm.

Keywords: Cyberspace, Opinion manipulation, Manipulation defense, User search

1. Introduction
Recent years have witnessed the emergence of the online social network as a prevalent and pivotal platform for opinion spreading and confrontation [1], yet still growing. Within such cyberspace, overwhelming volumes of users discuss social topics [2] and share personal opinions [3], meanwhile providing a perfect platform for malicious opinion manipulations. For example, a panic buying might appear, if malicious sellers exploited the opinion manipulation on some rumors for benefits [4, 5], such that the salt is extremely effective for preventing the COVID-19 virus [6]. Moreover, the unprecedentedly convenient access to this cyberspace also diversifies the approaches for malicious opinion manipulations [7, 8]. Manipulators are possible to utilize human-like Artificial Intelligence (AI) agents [9, 10, 11] to achieve automated manipulation, making conventional defending methods rigid and incapable [8, 12, 13]. Therefore, more flexible and advanced AI-based countermeasures are imperatively needed for defending the opinion manipulation.

Within the opinion defending, one prominent and typical approach is to intelligently search for the susceptible and key users to execute purposeful prevention. These key users are likely targeted by malicious manipulators, as they are both susceptible [7] and will dominantly influence the public opinion [14]. Therefore, searching for such users for a given social platform is critical for both sides of opinion confrontation. Abebe et al. first formalized this searching task as the Opinion Dynamics with Varying Susceptibility (referred as ODVS) problem [7], where the user opinion is jointly determined by user innate opinion, susceptibility, and opinions of user’s friends. The objective of ODVS is to search a number of key users and merely vary their susceptibilities to certain degrees, in order to maximize or

*Corresponding author: Ling Tian.
Email: lingtian@uestc.edu.cn

Preprint submitted to Elsevier September 22, 2020
minimize the support of all users for a target opinion in an almost unaware manner. Therefore, manipulators are able to only impact a small group of users to manipulate the public by exploiting opinion diffusions, instead of directly targeting on all users.

Generally, ODVS is proved to be NP-Hard and categorized into the unbudgeted and budgeted conditions [7]. Under the unbudgeted condition, manipulators are able to vary susceptibilities of all users, thus key user search is needless. Chan et al. systematically investigated the unbudgeted condition, and proposed an effective strategy for choosing optimized susceptibility values [15]. The method provides state-of-the-art performance and also reduces the complexity of susceptibility determination to linear time. On the other hand, the budgeted condition (referred as BOVS) just allows manipulators to vary susceptibilities of a limited number of users. BOVS is actually more practical and valuable, since directly varying susceptibilities of all users (the unbudgeted condition) may be infeasible in reality. However, existing works lack the proper and remarkable study on BOVS. Existing greedy search algorithms for BOVS simply set susceptibilities to minimal, neglecting potential key users who hold maximum susceptibilities as optimal. Besides, these approximate algorithms request high computational consumption and providing no performance guarantee.

Therefore, this work reconsiders the BOVS problem, and proposes a novel greedy algorithm, together with an accelerated algorithm based on the community property. According to the comprehensive analysis, it is revealed that both algorithms provide a certain approximation guarantee under a special case of BOVS, referred as the Budgeted Opinion dynamics with Decreasing Susceptibility (abbreviated as BODS) problem. Compared with basic BOVS where user susceptibilities can be increased or decreased, BODS will only try to vary user susceptibilities in a descending direction. Generally, the descending assumption of BODS is actually more practical, as making users stubborn (decreasing the susceptibilities) is usually easier than making users susceptible. There are both psychological and environmental reasons for this assumption. From a psychological perspective, people prefer to trust what they innately believe in [16] and resist subsequent opinions. From an environmental perspective, Filter Bubbles in social networks [17] always show the content supporting users’ innate opinions, also strengthening their belief. Therefore, only decreasing user susceptibilities will be more effective and reasonable than varying in both directions.

In specific, this work proposes an advanced greedy algorithm for BOVS, which overcomes the shortage of existing algorithms [7] via an optimal strategy of determining susceptibility values. We also adopt the unsupervised learning and fully utilize the community property of a network to design an accelerated algorithm, which is developed on the intuition that the impact of a user within a proper community could be a good reflection of the impact in the whole network, while the former one could be computed much more efficiently. This community-based algorithm significantly outperforms the state-of-art algorithm by reducing the time consumption by more than one order of magnitude. Under the BODS scenario, the approximation guarantee of the proposed greedy algorithm is comprehensively analyzed and proved through the approximate submodularity [18]; then the community-based algorithm is proved to hold a close approximation ratio to the greedy algorithm. Finally, comprehensive evaluations on real-world datasets show the superiority of proposed algorithms on both effectiveness and efficiency. As far as we know, this is the first work that comprehensively investigates both susceptibility determination and key user search for BOVS. In summary, the contributions of this work are as follows.

- We propose an advanced greedy algorithm to overcome the shortage of the state-of-art algorithm on susceptibility determination.
- An accelerated algorithm is designed, which adopts unsupervised learning and fully utilizes the community structure of the network.
- We formally define the BODS as a more practical case of BOVS.
- Systematical analysis of approximation guarantee is performed on both algorithms with respect to the BODS scenario.
- Extensive evaluations on different real-world datasets reveal that our algorithms significantly outperform the state-of-the-art algorithm.

The remainder of this paper is organized as follows. In the second section, existing works related
to the target problem in this work are reviewed. In the third section, we introduce our preliminaries used for problem analysis and algorithm design. In the fourth section, we introduce our advanced greedy algorithm, and the analysis of its approximation guarantee under the BODS scenario. In the fifth section, we introduce our community-based algorithm, together with its performance analysis. In the sixth section, we analyze comprehensive evaluations. Section 7 concludes our work.

2. Related Work

As the cyber space becomes an indispensable part of people’s daily life, the online opinions impact the off-line marketing [19] and public security [20] of the society. This induces malicious individuals to develop various attacks for benefits [7, 21], meanwhile inspires researchers to engage in the corresponding defense [22, 23]. In this work, we concentrate on the scenario of opinion manipulation, where the manipulators attempt to control public opinion [7].

Dietrich et al. studied the manipulation that merely targets on a specific opinion leader [12], and presented such the manipulation could make the average opinion converge to a targeted value. Another study proposed a heuristic algorithm to control the public opinion with respect to the structural hole spanners, who are key connections between two almost unlinked communities [24]. Moreover, Amelkin et al. defined one kind of manipulation where manipulators use Influence Maximization to maximize the user support of the target opinion [8]. The relationship between the user linkage and manipulation was fully analyzed, thus providing a heuristic defense algorithm via linkage addition.

This work investigates the key user search for defending the opinion manipulation [7], where manipulators merely vary key users’ susceptibilities to maximize or minimize the sum of user opinions. On one hand, the target problem can be viewed as assigning the budget to partial users, thus showing the similarity to the task allocation problem for crowdsensing [25]. On another hand, the target problem presents the similarity to the famous Influence Maximization (IM) [26, 14], as both problems seek the key users. While the target problem requires an additional procedure of determining users’ susceptibility values, and it does not hold the submodularity, thus the methods for IM cannot be applied to our target problem. The target search problem was proved as NP-Hard and categorized into the unBudgeted and budgeted condition [7]. The unBudgeted condition was first shown to be a convex optimization problem [7]. However, Chan et al. corrected above conclusion, which actually presented to be a non-convex optimization problem [15]. Meanwhile, they systematically investigated the strategy of choosing susceptibility values to provide a state-of-the-art method, which reduces the complexity of susceptibility determination to linear time. The main improvement of our greedy algorithm (shown in Section 4) is built upon this study. As a result, the unBudgeted condition is comprehensively studied, while the more practical budgeted condition (referred as BOVS) lacks investigation.

The current study to search key users for BOVS follows a general greedy framework, as selecting users with maximal marginal gain [7]. It requests a high computational consumption, meanwhile without any performance guarantee. Besides, this algorithm simply sets susceptibilities to the minimum, thus may neglect potential key users who hold maximum susceptibilities as optimal. Therefore, this work proposes an advanced greedy algorithm to fill the gap of susceptibility choice, then fully utilizes the community property to accelerate the greedy algorithm.

3. Preliminary

The ODVS problem [7] aims at maximizing or minimizing user supports on a target opinion. Let an undirected graph $G = (V, E, \alpha, s)$ denote a network, where $V$ and $E$ is the user and edge set, respectively. The target problem is built on an opinion dynamic model [27], where each user $i \in V$ is associated with an innate opinion $s_i \in [0, 1]$ ($s = \{s_1, s_2, ..., s_{|V|}\}$) and a resistance parameter $\alpha_i \in [l_i, u_i]$ ($\alpha = \{\alpha_1, \alpha_2, ..., \alpha_{|V|}\}$). Note that, $1 - \alpha_i$ is the susceptibility value of user $i$ and $0 < l_i < u_i < 1$, thus a higher $\alpha_i$ value indicates that user $i$ is more stubborn. In the rest of this paper, we only discuss the resistance instead of susceptibility for simplicity. The expressed opinion of the individual user evolves in discrete time as follows,

$$z_i(t + 1) = \alpha_i s_i + (1 - \alpha_i) \frac{\sum_{j \in N_i} z_j(t)}{d(i)}, \quad (1)$$

where $N_i$ is the neighbor set and $d(i)$ is the degree of user $i$. 


Let a random walk matrix $P$ represent the interaction between users, in which each entry $P_{i,j} = \frac{1}{d_i}$ for each edge $e_{i,j} \in E$, or zero otherwise. The opinion evolution of the whole network converges to a unique equilibrium [28],
\[ z = (I - (I - A)^{-1}As, \]
where $z$ is the vector of expressed opinions, $s$ is the vector of innate opinions, $\alpha$ is the vector of resistances, and $A = \text{Diag}(\alpha)$ is a diagonal matrix holding each entry $A_{i,i} = \alpha_i$.

**Definition 1.** (Opinion Dynamics with Varying Susceptibility Problem [7]). Given a network $G = (V, E, \alpha, s)$, the goal is finding a seed set $Q \subseteq V$ and choosing the certain value of each user resistance $\alpha_i$ ($i \in Q$) for maximizing or minimizing the sum of all users’ expressed opinions $f(\alpha_Q)$ as,
\[ f(\alpha_Q) = \tilde{I}z = \tilde{I}(I - (I - A_Q)^{-1}A_Qs). \]

Note that, the subscript $Q$ in $f(\alpha_Q)$ and $A_Q$ indicates that the resistance values of users in seed set $Q$ are updated, while the rest are fixed. The minimization of expressed opinions can be solved by invoking a maximization algorithm using $\tilde{I} - s$ as the innate opinion vector [7]. Therefore, we concentrate on the maximization problem in this work.

**Unbudgeted and Budgeted Condition.** The ODVS problem is formalized as follow:
\[ \max \quad f(\alpha_Q) = \tilde{I}(I - (I - A_Q)^{-1}A_Qs) \]
\[ \text{s.t.} \quad \alpha_i \in [l_i, u_i] \quad \text{for} \quad i \in Q \]
\[ |Q| = k \]
\[ (4) \]
ODVS is categorized into the unbudgeted and budgeted condition, according to different sizes of seed set $Q$ [7]. The budgeted condition requests $|Q| = k < |V|$ referred as BOVS, and the unbudgeted condition requests $|Q| = |V|$. The objective function $f(\alpha_Q)$ is not a convex function of $\alpha_Q$ under both unbudgeted and budgeted conditions [15], thus the ODVS problem is NP-Hard under both conditions. The unbudgeted condition is essentially different from the BOVS, as it only considers the resistance choices of all users, which does not require to find the key users. As a result, this work focuses on the BOVS.

4. Greedy Algorithm

The main technical challenges of BOVS are searching key users meanwhile determining their optimal resistance values. Existing algorithms simply set resistance values to maximal, which may neglect potential key users who hold the minimal resistance values as optimal. Therefore, we propose an advanced greedy algorithm to fill the gap of susceptibility choice in this section.

**4.1. Algorithm Description**

**Algorithm 1 Greedy Algorithm**

**Input:** graph $G = (V, E, \alpha, s)$, size of seed set $k$

**Output:** seed set $Q$

1. $Q = \emptyset$, compute $f(\alpha_Q)$ by Equation 3
2. for $q = 1$ to $k$ do
3.  for $i = 1$ to $|V|$ do
4.  $\Delta f_Q(i) = \max(\Delta f_Q(i|\alpha_i = l_i), \Delta f_Q(i|\alpha_i = u_i))$
5.  $Q = Q \cup \{\text{arg max}_{i \in V} \Delta f_Q(i)\}$
6.  compute $f(\alpha_Q)$ by Equation 5
7. return $Q$

Generally, given the required size of seed set $|Q| = k$, our algorithm runs $k$ iterations to select $k$ seeds as shown in Algorithm 1. In each iteration, the user $i$ holding maximum marginal gain $\Delta f_Q(i) = f(\alpha_{Q \cup \{i\}}) - f(\alpha_Q)$ is chosen to be a seed. When computing the marginal gain $\Delta f_Q(i)$ of a single user, the algorithm determines the new resistance between two extreme values (maximum $u_i$ and minimum $l_i$) which holds the maximal marginal gain. For example, if $\Delta f_Q(i|\alpha_i = l_i) > \Delta f_Q(i|\alpha_i = u_i)$, we choose the lowest value $l_i$ as the new resistance.

The main difference between our advanced greedy algorithm and the existing greedy algorithm [7] is the strategy of choosing the resistance (Line 4). Suppose a seed set $Q$ is already selected and the objective function $f(\alpha_Q)$ is computed, we plan to select one more seed. For each user $i \in V - Q$ with resistance $\alpha_i \in [l_i, u_i]$, we need to obtain the marginal gain via computing the objective function value $f(\alpha_{Q \cup \{i\}})$ under $\alpha_i = u_i$ and $\alpha_i = l_i$, respectively. The $f(\alpha_{Q \cup \{i\}})$ can be converted according to the Sherman-Morrison formula, which accelerates the procedure of choosing resistance (Line 4).
\[
\begin{align*}
f(\alpha_{Q \cup \{i\}}) &= \bar{I}(I - (I - A_{Q \cup \{i\}})P)^{-1}A_{Q \cup \{i\}}s \\
&= \bar{I}(I - (I - A_Q)P + x_i(x_i^T P)s] \\
&= f(\alpha_Q) + \bar{I}[Mx_i x_i^T s - U A_Q s - U x_i x_i^T s], \\
\end{align*}
\]
where \(M = (I - (I - A_Q)P)^{-1}\) is obtained in prior computation of \(f(\alpha_Q)\), and \(U = \frac{M x_i (x_i^T P) M}{\bar{I} x_i^T P M x_i}\). The value of marginal gain \(\Delta f_Q(i) = f(\alpha_{Q \cup \{i\}}) - f(\alpha_Q)\) is
\[
\Delta f_Q(i) = \bar{I}[M x_i x_i^T s - U A_Q s - U x_i x_i^T s],
\]
where \(x_i\) is a \(n\)-dimensional vector satisfying \(x_i(j) = 0\) if \(j \neq i\) and \(x_i(i) = \sqrt{\alpha_i - \alpha_i^0}\) otherwise, and \(\alpha_i^0\) denotes the original resistance value of user \(i\). Therefore, our advanced strategy of choosing resistance values fully considered two extreme values, instead of simply setting resistance values of all picked users to maximum [7]. Meanwhile, it avoids directly computing the complex matrix inversion, thus reducing the computation consumption.

Theorem 1 provides clues on why the two extreme values are sufficient for selection.

**Theorem 1.** (Extreme Values Are Sufficient). For arbitrary user \(i \in V\), fixing resistance values of all other users except \(i\), the objective \(f(\alpha_{\{i\}})\) is a monotone function in \(\alpha_i\). This implies that to maximize \(f(\alpha_{\{i\}})\), it is sufficient to consider two extreme values \(l_i\) and \(u_i\).

Theorem 1 is originally proposed for the unbudgeted condition [15]. It still holds under BOVS, since the situation that varying the resistance value of a specific user in BOVS is the same as the unbudgeted condition. Moreover, we show that a bigger seed set leads to a larger opinion sum at the equilibrium, which ensures the correctness of the proposed greedy algorithm.

**Theorem 2.** (Monotonicity). The objective \(f\) is a monotone ascending function respect to the seed set \(Q\), i.e., \(f(\alpha_{A \cup B}) \geq f(\alpha_A)\) when \(B - A \neq \emptyset\).

**Proof.** Given a network \(G = (V,E,\alpha,s)\), let \(f(\alpha)\) denotes the sum of expressed opinions without any resistance change. Then, \(f(\alpha_A)\) denotes the sum of expressed opinions, where every seed \(i \in A\) changes its resistance \(\alpha_i\) to the proper value \(\bar{u}_i\) as maximizing its marginal gain.

Consider \(f(\alpha_{A \cup \{i\}})\) and \(f(\alpha_A)\) (\(i \notin A\)), the difference between them is the user \(i\) contains the new resistance \(\bar{u}_i\) and the original \(\alpha_i^0\), respectively. When the new resistance \(\bar{u}_i\) is determined according to Theorem 1, it implies \(f(\alpha_{A \cup \{i\}}) \geq f(\alpha_A)\). For \(f(\alpha_{A \cup \{i\}, j})\) and \(f(\alpha_{A \cup \{i\}})\), the only difference is the resistance of user \(j\), thus \(f(\alpha_{A \cup \{i, j\}}) \geq f(\alpha_{A \cup \{i\}})\). By a certain number of iterations, we can obtain \(f(\alpha_{A \cup B}) \geq f(\alpha_A)\) when \(A \cap B = \emptyset\). Therefore, the ascending monotonicity of the objective function is proved.

The complexity of the proposed greedy algorithm is \(O(kn^4)\). It requires \(O(kn)\) loops in total, and the marginal gain is computed once in each loop (Line 4-5). For marginal gain corresponding to any user \(i\), the objective value \(f(\alpha_{Q \cup \{i\}}) = \bar{I}(I - (I - A_{Q \cup \{i\}})P)^{-1}A_{Q \cup \{i\}}s\) is required, thus taking \(O(n^3)\).

**4.2. Precision Analysis**

As mentioned in Section 1, there is a more practical case of BOVS. We first formalize the special case in this part, then provide the approximation guarantee of the advanced greedy algorithm under this case.

**4.2.1. Budgeted Opinion Dynamics with Decreasing Susceptibility**

The more practical and special case, referred as BODS, is defined as follows.

**Definition 2.** (Budgeted Opinion Dynamics with Decreasing Susceptibility). Given a network \(G = (V,E,\alpha,s)\), the goal is finding a seed set \(Q \subseteq V\) and choosing the certain value of each resistance \(\alpha_i\) \((i \in Q)\) in an ascending way for maximizing user expressed opinions \(f(\alpha_Q)\).

\[
\begin{align*}
\text{maximize } f(\alpha_Q) &= \bar{I}(I - (I - A_Q)P)^{-1}A_Qs \\
\text{s.t. } \alpha_i &\in [\alpha_i^0, u_i] \quad \text{for } i \in Q \\
|Q| &= k < |V|
\end{align*}
\]

where \(\alpha_i^0\) is the initial resistance value (without variation) of user \(i\).

Combined with Equation 4 and Equation 7, the difference between BODS and BOVS problem is the constraint of resistance choices for seeds. In BOVS problem, the newly chosen resistance \(\alpha_i\) of a seed \(i\)
can be varied in the range $[l_i, u_i]$. While in BODS problem, the newly chosen resistance $\alpha_i$ just can be set in the range $[\alpha^l_i, \alpha^u_i]$. Essentially, the BODS problem is a more practical and special case of the BOVS problem, as making users stubborn is easier than making them susceptible, due to both psychological and environmental reasons. The objective function of BODS problem is the same with BOVS, which is not a convex function of the resistance, thus BODS is also NP-Hard.

The proposed greedy algorithm is capable for BODS, as considering the marginal gain with respect to the original and maximum resistance of each user. Moreover, only the marginal gain of maximum resistance needs to be computed, since the marginal gain of the original resistance is zero. Thus, the greedy algorithm is more efficient for solving BODS than it for BOVS.

### 4.2.2. Approximate Ratio

This part provides the approximate guarantee of the advanced greedy algorithm under BODS.

**Theorem 3.** (Approximate Ratio.) The proposed greedy algorithm for BODS holds the $1 - e^{-1}$ approximate ratio.

**Proof.** Revisit that the objective function $f$ of BODS problem is a monotone function for the updating resistance $\alpha_i$, with respect to the seed set $Q : |Q| = k$. Therefore, the greedy algorithm holds the approximate submodularity, ensuring a $1 - e^{-\lambda}$ approximate ratio [18] where

$$\lambda = \min_{L \subseteq Q, |B| \leq k, L \cap B = \emptyset} \frac{\sum_{v \in B} f(\alpha_{L \cup \{v\}}) - f(\alpha_L)}{f(\alpha_{L \cup B}) - f(\alpha_L)}.$$  

(8)

Now, we analyze the bound of $\lambda$ by induction, i.e., whether $\lambda \geq 1$ for any $L \subseteq Q$ and $B : |B| \leq k$ satisfying $L \cap B = \emptyset$. Let $B_t$ denote the set $B$ in Equation 8 satisfying $|B_t| = t$.

When $t = 1$ for arbitrary $L \subseteq Q$, $B_1 = \{v\}$, it holds

$$\frac{f(\alpha_{L \cup \{v\}}) - f(\alpha_L)}{f(\alpha_{L \cup B}) - f(\alpha_L)} \geq 1.$$  

(9)

When $t \geq 2$, suppose that for $B_t : |B_t| = t$ and arbitrary $L \subseteq Q$,

$$\sum_{v \in B_t} \frac{f(\alpha_{L \cup \{v\}}) - f(\alpha_L)}{f(\alpha_{L \cup B_t}) - f(\alpha_L)} \geq 1.$$  

(10)

Consider $B_{t+1} : |B_{t+1}| = t+1$ and $B_{t+1} - B_t = v_{t+1}$ for arbitrary $L \subseteq Q$:

$$\sum_{v \in B_{t+1}} \frac{f(\alpha_{L \cup \{v\}}) - f(\alpha_L)}{f(\alpha_{L \cup B_{t+1}}) - f(\alpha_L)} \geq \frac{f(\alpha_{L \cup \{v_{t+1}\}}) - f(\alpha_L)}{f(\alpha_{L \cup B_{t+1}}) - f(\alpha_L)}$$

$$\geq f(\alpha_{L \cup \{v_{t+1}\}}) - f(\alpha_L) + f(\alpha_{L \cup \{v_{t+1}\}}) - f(\alpha_{L \cup B_{t+1}})$$

$$= f(\alpha_{L \cup \{v_{t+1}\}}) - f(\alpha_{L \cup B_{t+1}}) - f(\alpha_L).$$

(11)

We focus on the second part $f(\alpha_{L \cup B_{t+1}}) - f(\alpha_{L \cup B_t})$ of Equation 11, given the user set $V = \{v_1, v_2, ..., v_m, ..., v_n\}$. The set $L \cup B_t$ can be viewed as consisting of a basic set $L$ and an additional set $B_t$. For any user $v_{t+1}$ corresponding to $v_m \in V$, let $x_t + 1$ be a $p$-dimensional vector where $x_t + 1(j) = 0$ if $j \neq m$ and $x_t + 1(m) = \sqrt{\alpha_m - \alpha^o}$ otherwise. Let $M_{L,t} = (I - (I - A_{B_t})P)^{-1}$, then:

$$M_{L,t} = (I - (I - A_{B_t})P)^{-1}$$

$$= M_{L,t+1}P + x_t + 1(x_t + 1T)M_{L,t}$$

$$= M_{L,t} - \frac{M_{L,t}x_t + 1(x_t + 1T)M_{L,t}x_t + 1}{1 + (x_t + 1T)x_t + 1},$$

(12)

according to the Sherman-Morrison formula.

Let $U_{L,t+1} = \frac{M_{L,t}}{I + (x_t + 1T)x_t + 1}M_{L,t}$ for simplicity. Since all entries of $M_{L,t}$ and $P$ are non-negative and $x_t + 1T = \alpha_m - \alpha^o \geq 0$ in BODS problem, all entries of $U_{L,t+1}$ are non-negative. The objective function:

$$\tilde{I}_M(L,t + 1) = \tilde{I}_M(L,t) - U_{L,t + 1}$$

$$= \tilde{I}_M(L,t)As \leq \tilde{I}_M(L,t)As$$

(13)

Thus the second part of Equation 11 is:

$$f(\alpha_{L \cup B_{t+1}}) - f(\alpha_{L \cup B_t})$$

$$= \tilde{I}((M_{L,t} - U_{L,t + 1} + 1)(A_{B_t} + x_t + 1T))$$

$$- M_{L,t}A_{B_t}s$$

$$= \tilde{I}(M_{L,t}x_t + 1x_t + 1T) - U_{L,t + 1}x_t + 1x_t + 1T$$

$$= U_{L,t + 1}A_{B_t}s$$

$$\leq \tilde{I}(M_{L,t} - U_{L,t + 1})(x_t + 1x_t + 1T)$$

$$= \tilde{I}M_{L,t + 1}x_t + 1x_t + 1T.$$

Revisit that $B_{t+1} = L \cup \{v_1, v_2, ..., v_{t+1}\}$ consists of
a basic set $L$ and an additional set $\{v_1, v_2, \ldots, v_{t+1}\}$.  
\[
ML(t+1) = ML(t) - U_L(t+1) \\
= \ldots \\
= ML(0) - U_L(1) - \ldots - U_L(t+1), 
\]
where $ML(0) = (I - (I - A_L)P)^{-1}$. On another perspective, let $L' = L \cup \{v_{t+1}\}$ as the basic set, then the additional set is $\{v_1, v_2, \ldots, v_t\}$, i.e., $B_{t+1} = L' \cup \{v_1, v_2, \ldots, v_t\}$. Thus $ML(t+1)$ can be represented on the new perspective as,
\[
ML(t+1) = ML'(t) \\
= ML'(t-1) - U_L'(t) \\
= \ldots \\
= ML'(0) - U_L'(1) - \ldots - U_L'(t), 
\]
where $ML'(0) = ML(0) - U_{x_{t+1}} = \frac{ML(0) x_{t+1} (x_{t+1}^T P) M_L(0)}{1 + (x_{t+1}^T P) M_L(0)x_{t+1}^T}$. Therefore,
\[
f(\alpha_{L\cup B}) - f(\alpha_{L\cup B_i}) \\
\leq \Delta M_L(t+1)(x_{t+1}x_{t+1}^T) s \\
\leq \Delta M_L(0)(x_{t+1}x_{t+1}^T) s \\
= \Delta (ML(0) - U_{x_{t+1}})(x_{t+1}x_{t+1}^T) s \\
\leq \Delta (ML(0) - U_{x_{t+1}})(\lambda x_{t+1}x_{t+1}^T) s \\
= f(\alpha_{L\cup \{x_{t+1}\}}). 
\]
Combined with Equation 11, $\sum_{\alpha_{L\cup B_{t+1}}} (f(\alpha_{L\cup B_{t+1}}) - f(\alpha_{L})) \geq \lambda$ is proved, i.e., $\lambda \geq 1$. The $1 - \epsilon^{-1}$ approximate ratio is hold. \hfill $\Box$

Theorem 3 fills the gap between greedy algorithms and the approximation guarantee, thus theoretically ensuring the effectiveness of our greedy algorithms.

5. Community-based Algorithm

In this section, we propose a community-based algorithm for acceleration, since the advanced greedy algorithm in Section 4 still requests a high time consumption. The community structure is valuable and fundamental in social networks [29, 30], as users in the same community influence each other much more than users in different communities [31]. Therefore, it is reasonable to estimate the impact in the whole network by the impact in a community, thus reducing the time consumption. For simplicity, the accumulated change of opinions due to the variation of one’s susceptibility is denoted as her impact.

In specific, we adopt the unsupervised learning and fully utilize the community property to design a Community-Based Algorithm (referred as ComBA) for BOVS, which can also be applied to BODS. The algorithm consists of three parts, including the community partition, community absorption, and seed search. The first community partition divides the whole network into small communities by the impact-based unsupervised learning method. Our community partition takes both impacts and node connections into consideration, which is different from traditional community partition methods only considering node connections. The second community absorption refines the partition result, thus ensuring the whole algorithm holds an approximation guarantee. The third seed selection selects seeds in communities, which significantly reduce the time consumption compared with existing greedy algorithms.

5.1. Community Partition

The first community partition introduces unsupervised learning into the BOVS scenario, to divide the whole network into small communities according to impacts and user connections. Intuitively, the strongly impacted users with less hops should belong to the same community, since the goal of community partition is to estimate the impact of a user in the whole network by impact in a proper community. Therefore, we first design a measurement to combine the impact and connection relationships among users, then propose an unsupervised community partition algorithm based on the measurement.

**Definition 3. (Impact Distance).** Given a network $G = (V,E,\alpha, s)$, the impact distance $d_{i,j}$ measures the impact and connection relationships between user $i$ and $j$, formalized as
\[
d_{i,j} = \begin{cases} 
\text{hop}(i,j), & \Delta z_j(\alpha_i) > 0 \\
\infty, & \text{otherwise}
\end{cases} 
\]
where $\text{hop}(i,j)$ denotes the number of hops between user $i$ and $j$. The $\Delta z_j(\alpha_i)$ indicates the variation of user $j$’s expressed opinion at the equilibrium, when
the resistance $\alpha_i$ is varied (fixing the others) for maximizing the objective $f$.

The impact distance jointly measures the impact and connection relationships between users. According to this, positively impacted ($\Delta z_i(\alpha_i) > 0$) users with fewer hops are close users, that are more likely to belong to the same community with respect to our intuition.

The community partition algorithm is shown in Algorithm 2, as adopting the K-MEANS [32] framework. It first makes some initializations on impact distances and the community set (Line 1 - Line 7), then assigns participations based on impact distances in total $\tau$ iterations (Line 8 - Line 12). For each iteration, arbitrary user $i \in V$ finds the closet community center, and updates its community label $i.c$ according to the closet center (Line 9 - Line 10). Then each community $C_j \in C$ updates its center by selecting the user with minimum sum of impact distances among community members (Line 11 - Line 13). After sufficient iterations, the community partition $C$ is obtained.

Algorithm 2 Community Partition

| Input: | graph $G = (V,E,\alpha,s)$, community number $r$, iteration threshold $\tau$ |
|--------|--------------------------------------------------|
| Output: | community set $C$ |
| 1: | for each $i \in V$ do |
| 2: | for each $j \in V$ do |
| 3: | compute the impact distance $d_{i,j}$ |
| 4: | initialize community set $C = \{C_1,C_2,...,C_r\}$ |
| 5: | for each $C_i \in C$ do |
| 6: | $C_i = \emptyset$ |
| 7: | select $r$ random community centers $c = \{c_1,c_2,...,c_r\}$ |
| 8: | for $t = 1$ to $\tau$ do |
| 9: | for each $i \in V$ do |
| 10: | $i.c = \arg\min_{c_j \in C} d_{c_j,i}$ |
| 11: | update each $C_i \in C$ by each $j.c$ ($j \in V$) |
| 12: | for each community $C_i \in C$ do |
| 13: | center $c_j = \arg\min_{a \in C_i} \sum_{b \in C_i} d_{a,b}$ |
| 14: | return $C$ |

5.2. Community Absorption

The impact distance is not symmetric since the impact $\Delta z_i(\alpha_j)$ may not equal to $\Delta z_j(\alpha_i)$ for any user pair $i$ and $j$. Such the asymmetry may create improper partitions, so that the impact in these communities cannot successfully approximate the impact in the whole network. Furthermore, the effectiveness of the first community partition also relies on the preset iteration threshold $\tau$, which may lead to unstable community partitions with respect to different networks. Therefore, we utilize the second community absorption to refine the potential improper community partition, which follows a community combination framework [33].

Revisit that our intuition is to utilize the impact of a user in a community to estimate the impact in the whole network, we first define the absorption ratio to measure the impact difference of a user in its community and outside its community.

Definition 4. (Absorption Ratio). Given two communities $C_m$ and $C_l$, we define the absorption ratio as

$$AoR(C_m,C_l) = \max_{i \in C_m} \frac{R_l(i)}{R_m(i)}.$$

where $R_m(i) = \sum_{j \in C_m} \Delta z_j(\alpha_i)$ denotes the variation of expressed opinion sum of users in community $C_m$ with varying the resistance $\alpha_i$ for opinion maximization.

Specifically, the computations of impact $R_m(i)$ and $R_l(i)$ are a little different, when we compute the absorption ratio $AoR(C_m,C_l)$. For $R_m(i)$, only the community $C_m$ (a subgraph of $G$) is used to determine the resistance choice $\alpha_i$ and compute the variation of opinion sum $R_m(i)$. Combined with the objective in Equation 3, both resistance matrix $A$ and interaction matrix $P$ are $|C_m| \times |C_m|$ dimensional, where $|C_m|$ is the size of community $C_m$. While for $R_l(i)$, we first compute the expressed opinion vector $z$ for the whole network given the determined $\alpha_i$, then obtain the opinion sum $R_l(i)$. The used resistance matrix and interaction matrix are $|V| \times |V|$ dimensional, thus the computation on outside impact $R_l(i)$ has a higher complexity than the in-community impact $R_m(i)$.

Based on partitioned communities, we preset an absorption threshold $\theta$ to judge the impact level between two communities, as shown in Algorithm 3. Communities holding a higher absorption ratio than the absorption threshold are deemed to "strongly" impact each other, thus being combined by community absorption. In particular for each community $C_m \in C$, it checks the absorption ratio with other communities $C_l \in C - C_m$ (Line 4 - Line 6). If the absorption ratio is not less than the absorption threshold $\theta$, we combine these communities (Line 7 - Line 8). The algorithm continues.
above procedures, until there are no satisfied communities to be combined.

**Algorithm 3 Community Absorption**

**Input:** graph $G = (V, E, \alpha, s)$, community set $C$, absorption threshold $\theta$

**Output:** community set $C$

1: $AbsortFlag = 1$
2: while $AbsortFlag == 1$ do
3: $AbsortFlag = 0$
4: for each community $C_m \in C$ do
5: for each community $C_l \in C - C_m$ do
6: compute the $Aor(C_m, C_l)$
7: if $Aor(C_m, C_l) \geq \theta$ then
8: $C_m = C_m \cup C_l$, $AbsortFlag = 1$
9: return $C$

5.3. Seed Search

After the second community absorption, we properly divide the whole network into a set of communities. The third seed search selects seeds among communities for maximizing the expressed opinion sum of the whole network.

At first, our seed search algorithm finds the most impact user of each community, and conserves them in the candidate set $T$ (Line 4 - Line 7). The candidate $top_i \in T$ (with respect to each community $C_i$) consists of the index of the most impacted user $u_i$ and the opinion sum of corresponding community $i$, given the resistance $\alpha_u$, is varied to the proper value. The proper resistance value of each user $i$ is determined via Theorem 1 (Line 5), which is as same as in Algorithm 1. Moreover, a set $q = \{q_1, q_2, ..., q_r\} = \arg\max\{\alpha_{u_i}\}$ is initialized to maintain the determined seeds in each community (Line 3). The algorithm executes $k$ iterations to select the total $k$ seeds (Line 8 - Line 14), in which the user with maximal value of impact $j$ is chosen as the seed (Line 9). Suppose the current seed is $u_m$, it is associated with a certain resistance value $\alpha_{u_m}$ which is determined when we compute its impact (Line 5). Then seed $u_m$ is recorded by set $q_m$ to avoid reselecting. The initial resistance $\alpha_{u_m}^0$ is updated to the determined value $\alpha_{u_m}$ (Line 11). Subsequently, the most impact user in community $m$ is updated without considering the former selected seeds (Line 12 - Line 14). After $k$ iterations, the seed set $Q = |Q| = k$ is derived.

The complexity of our seed search is $O(km^4)$, where $m$ is the maximal community size among all partitioned communities. In specific, the initialization of the candidate set $T$ (Line 4 - Line 7) takes $O(rm^4)$, where $r$ is the number of communities. The $k$-step search (Line 8 - Line 14) takes $O(km^4)$, as utilizing the impact in a community to approximate the impact in the whole network. The approximation analysis in Subsection 5.4 shows that the algorithm is more effective when the community number decreases. Therefore, we modify the absorption threshold $\theta$ to make $r < k$ in practical use of the ComBA algorithm, thus containing the $O(km^4)$ complexity.

**Algorithm 4 Seed Search**

**Input:** graph $G = (V, E, \alpha, s)$, community set $C = \{C_1, C_2, ..., C_r\}$, size of seed set $k$

**Output:** seed set $Q$

1: $Q = \emptyset$, $T = \{top_1, top_2, ..., top_r\}$, $q = \{q_1, q_2, ..., q_r\}$
2: $top_1 = top_2 = ... = top_r = \emptyset$
3: $q_1 = q_2 = ... = q_r = \emptyset$
4: for each community $C_i \in C$ do
5: $f_i = \max_{j \in C_i} R_i(j)$
6: $u_i = \arg\max_{j \in C_i} R_i(j)$
7: $top_i = u_i, f_i$
8: for $a = 1$ to $k$ do
9: $top_m = u_m, f_m = \max_{top_i \in T} f_i$
10: $Q = Q \cup \{u_m\}$, $q_m = q_m \cup \{u_m\}$
11: update $\alpha_{u_m}$ corresponding to $f_m$
12: $f_m = \max_{j \in C_m - q_m} R_m(j)$
13: $u_m = \arg\max_{j \in C_m - q_m} R_m(j)$
14: $top_m = u_m, f_m$
15: return $Q$

5.4. Performance Analysis

Three parts of the community-based algorithm hold different time complexities. The first community partition generally takes $O(n^3)$, as there are $n^2$ node pairs requiring the computation of impact distance, where computing one impact distance costs $n^3$. The second community absorption takes $O(r^2mn^3)$, as iteratively computing the absorption rate between community pairs, where $r$ is the number of communities, and $m$ is the size of the largest community. Since the first two parts could be regarded as pre-processing steps in practice, the general complexity of the community-based algorithm can be deemed to $O(km^4)$, i.e., the complexity of the third seed search.
Combined with Lemma 1, we derive the performance guarantee of the ComBA under BODS by the analysis of its approximation ratio.

**Lemma 1.** Given a user \( i \) in community \( C_m \), the impact \( R_m(i) \) of user \( i \) in its community \( C_m \) is \( \frac{1}{1+(r-1)\theta} \) approximate to the impact in the whole network \( G \), formalized as

\[
R_m(i) \geq \frac{1}{1+(r-1)\theta} R(i),
\]

where \( \theta \) is the absorption threshold used in Algorithm 3 and \( r \) is the number of partitioned communities.

**Proof.** Revisit \( R_m(i) = \sum_{j \in C_m} \Delta z_j(\alpha_i) \) denotes the impact of user \( i \) in community \( C_m \) and \( R(i) = \sum_{j \in G} \Delta z_j(\alpha_i) \) denotes the impact in the whole network \( G \). As the community absorption is based on the absorption ratio and threshold \( \theta \), we have

\[
R(i) - R_m(i) = \sum_{C_i \in \mathcal{C} \cap R_m} R_i(i) \leq \sum_{C_i \in \mathcal{C} \cap R_m} \theta \cdot R_m(i) \leq (r-1) \cdot \theta \cdot R_m(i)
\]

Thus, we obtain the relationship between \( R_m(i) \) and \( R(i) \),

\[
R_m(i) \geq \frac{1}{1+(r-1)\theta} R(i).
\]

**Theorem 4.** Let \( Q^* : |Q^*| = k \) be the optimal seed set maximizing the objective \( f \), and \( Q : |Q| = k \) be the seed set obtained by our community-based algorithm. \( R(Q^*) \) and \( R(Q) \) denote the impacts (objective values) of these two seed sets, respectively. The proposed community-based algorithm holds the \( 1 - e^{-\frac{1}{k(1+(r-1)\theta)}} \) approximation ratio.

**Proof.** Let \( Q(j) \) be the first \( j \) users and \( v_j \) be the \( j \)-th user in \( Q \). Considering the difference \( R(Q^*) - R(Q(j-1)) \), it is covered by the \( j \) subsets of \( R(Q^*) \). Thus, one of the \( j \) subsets in the optimal seed set \( Q^* \) is at least \( \frac{R(Q^*) - R(Q(j-1))}{j(1+(r-1)\theta)} \), because of the pigeonhole principle [34]. Combined with Lemma 1, our community-based algorithm selects the user with maximum marginal gain as the seed, thus holding \( R(v_j) \geq \frac{R(Q^*) - R(Q(j-1))}{j(1+(r-1)\theta)} \). When \( j = 1 \), we have \( R(Q(1)) = R(v_1) \geq \frac{R(Q^*)}{1+(r-1)\theta} \). As a result, the following deduction is derived:

\[
R(Q(k)) = R(Q(k-1)) + R(v_k)
\]

\[
\geq R(Q(k-1)) + \frac{R(Q^*) - R(Q(k-1))}{k(1+(r-1)\theta)}
\]

\[
= (1 - \frac{1}{k(1+(r-1)\theta)}) R(Q(k-1))
\]

\[
\geq \frac{R(Q^*)}{k(1+(r-1)\theta)}
\]

\[
\geq \frac{1}{k(1+(r-1)\theta)} R(Q(k-2))
\]

\[
\geq \frac{1}{k(1+(r-1)\theta)} R(Q(k-3))
\]

\[
\geq \frac{1}{k(1+(r-1)\theta)} R(Q(k-4))
\]

\[
\geq \cdots
\]

\[
\geq \frac{1}{k(1+(r-1)\theta)} R(Q(1))
\]

\[
\geq \frac{1}{k(1+(r-1)\theta)} \sum_{i=0}^{k-2} \frac{1}{k(1+(r-1)\theta)} R(Q^*)
\]

\[
\geq \frac{1}{1 - \frac{1}{k(1+(r-1)\theta)}} R(Q^*)
\]

The \( \frac{1}{1 - \frac{1}{k(1+(r-1)\theta)}} \) is a monotone increasing function in \( k \), thus its upper bound is \( e^{-\frac{1}{k(1+(r-1)\theta)}} \) when \( k \to \infty \). We obtain that \( R(Q) \geq (1 - e^{-\frac{1}{k(1+(r-1)\theta)}}) R(Q^*) \).

According to Theorem 4, it implies that the absorption threshold \( \theta \) in Algorithm 3 controls the tradeoff between approximate precision and algorithm efficiency. When \( \theta = 0 \), the community absorption algorithm combines all community partitions into one network, thus the community-based absorption algorithm degenerates to the greedy algorithm with \( (1 - e^{-1}) \)-approximation. As \( \theta \) becomes bigger, the number of generated communities increases, thus increasing the complexity, meanwhile decreasing the approximate precision. Therefore, we always set a small value of the absorption threshold \( \theta \) to make \( r < k \) in practical use, thus ensuring the performance of the community-based algorithm.

**6. Evaluations**

In this section, we perform comprehensive experiments to evaluate our proposed algorithms with the state-of-art algorithm. Evaluation results are systematically analyzed, showing the outperformance of our proposed algorithms.
6.1. Datasets and Setup

Table 1: The statistics of datasets.

| Datasets       | Number of Users | Number of Edges |
|----------------|-----------------|-----------------|
| Weibo-general  | 1,382           | 6,745           |
| ego-Facebook   | 4,039           | 88,234          |
| com-Youtube    | 1,134,890       | 2,987,624       |
| com-LiveJournal| 3,997,962       | 34,681,189      |

Two “small” and two “large” real-world datasets are used to evaluate all algorithms, as shown in Table 1.

- **Weibo-general** is a public available dataset [35], containing the tweet cascades and user comments about topics in Sina Weibo, a Chinese Twitter-like website. We select a subgraph associated with the topic #Mi phone# (a mobile phone brand) in this evaluation. The user opinions on this topic are measured by a short-text classifier, and we use these opinions as the innate opinions of users. All directed edges are transferred into undirected.

- **ego-Facebook**: this is a public available dataset containing a subgraph of Facebook. Both ego-Facebook and Weibo-general are the “small” social networks. We synthesize the innate opinions of users following the uniform random distribution.

- **com-Youtube** and **com-LiveJournal** are “large” datasets containing Youtube and academic co-author relationships, respectively. The innate opinions of users are synthesized via the uniform random distribution.

With respect to above datasets, we synthesize the initial resistance following the uniform distribution and the normal distribution, respectively. All algorithms are performed on a Linux machine with 64-G memory and Intel Xeon 3.0-GHz CPU.

6.2. Comparisons

We compare our proposed greedy and community-based algorithms with two baselines and the state-of-art greedy algorithm.

- **Baseline I** [7] selects top $k$ users with highest innate opinions as seeds. For each seed, the resistance is set to be maximum.

- **Baseline II** [7] assigns $score(i) = \frac{d(i)}{2m} \times \frac{s_i}{\sum_{j \in N_i} s_j}$ to each user $i$, and selects the top $k$ users with highest scores as seeds, where $d(i)$ is the degree, $m$ is the edge number of the graph, $N_i$ is the neighbor set of user $i$. For each seed, it fixes the resistance as maximum.

- **Greedy** [7] follows the same greedy framework with our proposed Greedy+ algorithm, as choosing the user with maximum marginal gain as the seed. The main difference between Greedy and Greedy+ is the determination of the resistance value. In Greedy, it simply sets the resistance values of all seeds to be maximum.

- **Greedy+** is our proposed greedy algorithm. It selects seeds with respect to the marginal gain, meanwhile chooses the proper resistance based on two extreme values.

- **ComBA** is our proposed community-based algorithm, which fully utilizes the community property of the network to reduce the computation complexity, meanwhile guarantees a desirable performance.

These algorithms are evaluated on both BOVS and BODS problems. All algorithms can be directly run for the BODS problem, since BODS just ties the bound of the resistance value.

6.3. Results

We evaluate the effectiveness and efficiency of all comparisons in this part, to show the superiority of proposed algorithms.

6.3.1. Effectiveness

The effectiveness of each algorithm is illustrated in this part through the expressed opinion sum of different algorithms. This work focus on the maximization problem, so a better algorithm results in a larger opinion sum at the equilibrium.

We synthesize the innate opinion $s_i$ for each user $i$ following the uniform distribution with the range $[0, 1]$ on the ego-Facebook, com-Youtube, and com-LiveJournal dataset. The innate opinion in the Weibo-general dataset equals to the user original opinion polarity, i.e., each $s_i \in \{0, 1\}$. The initial
resistances follow the uniform distribution and the normal distribution, respectively. The size of seed set is assigned to 4%, 8%, 12%, 16%, 20% of the node number.

**BOVS.** For BOVS, two extreme values of the resistance need to be tried when Greedy+ and ComBA compute the marginal gain of a user. The evaluation results in BOVS are shown in Figure 1 and Figure 2 with respect to different resistance distributions.

In general, the proposed ComBA present a similar performance with Greedy+ on Weibo-general and ego-Facebook datasets, significantly outperforming other comparisons. This indicates that our community-based search maintains the performance of the greedy one, meanwhile reducing the time consumption.

Specifically, associated with uniform resistances, Greedy+ algorithm averagely outperforms the Baseline I by 22.57% and Baseline II by 20.15% on Weibo-general and ego-Facebook. When focusing on the opinion increment caused by the seed set, Greedy+ outperforms the state-of-art Greedy algorithm by 30.79% and 18.49% on Weibo-general and ego-Facebook. The underlying reason is that our resistance determination strategy can fully consider the minimum and maximum value when determining the resistance. Meanwhile, greedy algorithm does not consider the choice of minimal resistance, thus neglecting some potential vital users. As for two large datasets, both Greedy and Greedy+ cannot accomplish the computation within a day, thus omitted with respect to these datasets. However, the community-based algorithm ComBA can obtain the solution within a day, indicating the efficiency of our community-based searching strategy. Furthermore, the ComBA algorithm outperforms two baselines by 28.25% and 18.16% respectively with respect to the opinion sum, which shows a significant effectiveness for BOVS.

**BODS.** The evaluation settings for BODS are the same as BOVS, except the resistances is merely allowed to be increased. Generally, our Greedy+ and ComBA algorithm significantly outperform two baselines, meanwhile presenting a comparable performance with the Greedy algorithm, as shown in Figure 3. In specific, we analyze the opinion increment caused by the seed set. On Weibo-general and ego-Facebook datasets, the ComBA algorithm averagely outperforms the Baseline I by 124.79% and Baseline II by 92.46% associated with uniform resistances; outperforms the Baseline by 198.12% and Baseline II by 76.98% associated with normal resistances. This outperformance is owed to the effectiveness of the seed search in ComBA.

Evaluation results show that the seed set of Greedy+ is almost the same as the seed set of Greedy under BODS scenario, although Theorem 1 theoretically offers a better way for Greedy+ to make choice of the resistance. This is because Greedy+ computes the maximal marginal gain between the original and maximum resistance of user $i$. Then the marginal gain will be zero if the resistance is chosen to the original resistance. Thus in most cases, Greedy+ actually selects the user who holds the maximum marginal gain, and varies its resistance to the maximum value, which is the same with Greedy. The scenario where Greedy+ selects a seed with maintaining its original resistance occurs when increasing the resistance for every user decreases the sum of opinion. However, such a scenario is unlikely to appear in common networks.

### 6.3.2. Efficiency

Finally, the time consumption is adopted to evaluate the efficiency of proposed algorithms. Two baseline methods are omitted in this part, since they are simple algorithms with ground truth results on effectiveness. We vary the size of seed set to evaluate the efficiency among comparisons, as the complexity of the algorithm is dominated by the seed set size. Moreover, different distributions of initial resistances do not influence the time consumption. Therefore, we only present the evaluation results with resistances under uniform distribution in Figure 4.

In general, all algorithms present similar performances with respect to BODS and BOVS problem, that the community-based algorithm significantly outperforms two greedy algorithms. Comparing the performances on Weibo-general and ego-Facebook datasets, the ComBA algorithm significantly outperforms both Greedy and Greedy+ by one and two orders of magnitudes, respectively. Such outperformances indicate the superiority of our community-based accelerating strategy, as estimating the impact of a user in the whole network by that in a proper community. Moreover, the time consumptions of Greedy+ and Greedy result in the same order of magnitude, as holding the same computation complexity. While the Greedy+ presents the 3.09% and 3.8% increment of time consumption...
Figure 1: Sum of expressed opinions with uniform resistances for BOVS.

Figure 2: Sum of expressed opinions with normal resistances for BOVS.

Figure 3: Sum of expressed opinions for BODS.
than Greedy. This is due to that Greedy+ adopts a more complex strategy on choosing resistance value to achieve a better performance on the selection of seed set.

7. Conclusion

This paper studies the problem of opinion dynamics with varying susceptibility, which aims at discovering susceptible and key users to defend malicious opinion manipulations in cyber space. Specifically, the paper focuses on the scenario where the defender can only search for a limited number of users under a given budget. As the problem is proved to be NP-hard, two heuristic searching algorithms are proposed. The first algorithm overcomes the shortage of existing algorithms by an optimal strategy of determining susceptibility values. The second algorithm adopts unsupervised learning methods and fully utilizes the community property, to achieve a high efficiency. Moreover, the approximation guarantees of both greedy and community-based algorithms are comprehensively analyzed under the budgeted scenario. Evaluations on real-world datasets show that our algorithms significantly outperform the state-of-the-art algorithms on both effectiveness and efficiency.

Declaration of Competing Interest

We confirm that there is no known conflict of interest associated with this publication. There has been no significant financial support for this work that could have influenced its outcome.

Acknowledgments

This research work was partly supported by the National Key R&D Program of China (No. 2018YFC0807500), by Ministry of Science and Technology of Sichuan Province Program (No. 2018GZDZX0048, 20ZDYF0343), by National Natural Science Foundation of China (NSFC) under Grant NO. U19A2059, and by the Young Scientists Fund of the National Natural Science Foundation of China under Grant NO. 61802050.

References

[1] K. Xu, F. Wang, H. Wang, B. Yang, Detecting fake news over online social media via domain reputations and content understanding, Tsinghua Science and Technology 25 (2019) 20–27.
[2] C. Wang, X. Xin, J. Shang, When to make a topic popular again? a temporal model for topic rehottup prediction in online social networks, IEEE Transactions on Signal and Information Processing over Networks 4 (2017) 202–216.
[3] C. Vairetti, E. Martínez-Cámara, S. Maldonado, V. Luzón, F. Herrera, Enhancing the classification of social media opinions by optimizing the structural information, Future Generation Computer Systems 102 (2020) 838–846.
[4] Z. He, Z. Cai, J. Yu, X. Wang, Y. Sun, Y. Li, Cost-efficient strategies for restraining rumor spreading in mobile social networks, IEEE Transactions on Vehicular Technology 66 (2016) 2791–2800.
[5] Y. Liu, Z. Cai, X. Wang, F. Hao, Incentive mechanisms for crowdblocking rumors in mobile social networks, IEEE Transactions on Vehicular Technology 68 (2019) 9220–9232.
[6] Wikipedia. Panic buying, 2020. URL: https://en.wikipedia.org/wiki/Panic_buying.
[7] R. Abebe, J. Kleinberg, D. Parkes, C. E. Tsurakakis, Opinion dynamics with varying susceptibility to persuasion, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1089–1098.
[8] V. Amelkin, A. K. Singh, Fighting opinion control in social networks via link recommendation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 677–685.
[9] W. Nie, N. Narodytska, A. Patel, Relgan: Relational generative adversarial networks for text generation, in: International conference on learning representations, 2018.
[10] H. Wang, K. Qin, G. Lu, G. Luo, G. Liu, Direction-sensitive relation extraction using bi-sdp attention model, Knowledge-Based Systems (2020) 105928.
[11] Q. Zhao, J. Niu, H. Chen, L. Wang, M. Atiquzzaman, An indicative opinion generation model for short texts on social networks, Future Generation Computer Systems 86 (2018) 1471–1480.
[12] F. Dietrich, S. Martin, M. Junger, Opinion dynamics control by leadership with bounded influence, in: 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 1–6.
[13] T. Mihaylov, G. Georgiev, P. Nakov, Finding opinion manipulation trolls in news community forums, in: Proceedings of the nineteenth conference on computational natural language learning, 2015, pp. 310–314.
[14] W. Tang, G. Luo, Y. Wu, L. Tian, X. Zheng, Z. Cai, A second-order diffusion model for influence maximization in social networks, IEEE Transactions on Computational Social Systems 6 (2019) 702–714.
[15] T. H. Chan, Z. Liang, M. Sozio, Revisiting opinion dynamics with varying susceptibility to persuasion via non-convex local search, in: The World Wide Web Conference, 2019, pp. 173–183.
[16] I. Hernandez, J. L. Preston, Disfluency disrupts the confirmation bias, Journal of Experimental Social Psychology 49 (2013) 178–182.
[17] U. Chitra, C. Musco, Analyzing the impact of filter bubbles on social network polarization, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 115–123.
[18] A. Das, D. Kempe, Approximate submodularity and
its applications: subset selection, sparse approximation and dictionary selection. The Journal of Machine Learning Research 19 (2018) 74–107.

[19] Y. Li, H. Zhou, Z. Lin, Y. Wang, S. Chen, C. Liu, Z. Wang, D. Gifu, J. Xia, Investigation in the influences of public opinion indicators on vegetable prices by corpora construction and wechat article analysis. Future Generation Computer Systems 102 (2020) 876–888.

[20] Z. He, Z. Cai, X. Wang, Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks, in: 2015 IEEE 35Th international conference on distributed computing systems, IEEE, 2015, pp. 205–214.

[21] Z. Cai, Z. He, X. Guan, Y. Li, Collective data-sanitization for preventing sensitive information inference attacks in social networks. IEEE Transactions on Dependable and Secure Computing 15 (2016) 577–590.

[22] Z. He, Z. Cai, J. Yu, Latent-data privacy preserving with customized data utility for social network data. IEEE Transactions on Vehicular Technology 67 (2018) 1087–1096.

[23] M. Siddula, Y. Li, X. Cheng, Z. Tian, Z. Cai, Anonymization in online social networks based on enhanced equi-cardinal clustering. IEEE Transactions on Computational Social Systems 6 (2019) 809–820.

[24] C. Gong, Y. Du, X. Li, X. Chen, X. Li, Y. Wang, Q. Zhou, Structural hole-based approach to control public opinion in a social network. Engineering Applications of Artificial Intelligence 93 (2020) 103690.

[25] Z. Cai, Z. Duan, W. Li, Exploiting multi-dimensional task diversity in distributed auctions for mobile crowdsensing. IEEE Transactions on Mobile Computing (2020).

[26] J. Li, Z. Cai, M. Yan, Y. Li, Using crowdsourced data in location-based social networks to explore influence maximization. in: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE, 2016, pp. 1–9.

[27] N. E. Friedkin, E. C. Johnsen, Social influence and opinions. Journal of Mathematical Sociology 15 (1990) 193–206.

[28] A. Das, S. Gollapudi, R. Panigrahy, M. Salek, Debiasing social wisdom. in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 500–508.

[29] Y. Zhang, B. Wu, Y. Liu, J. Lv, Local community detection based on network motifs. Tsinghua Science and Technology 24 (2019) 716–727.

[30] X. Zheng, Z. Cai, G. Luo, L. Tian, X. Bai, Privacy-preserved community discovery in online social networks. Future Generation Computer Systems 93 (2019) 1002–1009.

[31] Z.-B. Dong, G.-J. Song, K.-Q. Xie, J.-Y. Wang, An experimental study of large-scale mobile social network. in: Proceedings of the 18th international conference on World wide web, 2009, pp. 1175–1176.

[32] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations. in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, Oakland, CA, USA, 1967, pp. 281–297.

[33] Y. Wang, G. Cong, G. Song, K. Xie, Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 1039–1048.

[34] H. Ma, H. Yang, M. R. Lyu, I. King, Mining social networks using heat diffusion processes for marketing candidates selection. in: Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 233–242.

[35] J. Zhang, B. Liu, J. Tang, T. Chen, J. Li, Social influence locality for modeling retweeting behaviors. in: Twenty-Third International Joint Conference on Artificial Intelligence, 2013.
Author biography

Wenyi Tang received the B.S. degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently a Ph.D. student in the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include data privacy, social network and machine learning.

Ling Tian received the bachelor, master and Ph.D. degrees from the school of computer science, University of Electronic Science and Technology of China in 2003, 2006 and 2010, respectively. She is currently a Professor in UESTC. She had been a visiting scholar in Georgia State University during 2013 in United States. Her research interests include image and video coding, signal processing.

Xu Zheng received his B.S. and M.S. degree from School of Computer Science and Technology at Harbin Institute of Technology. Mr. Zheng is currently an assistant professor in School of Computer Science and Engineering, University of Electronic Science and Technology of China, and a PhD student in the Department of Computer Science at Georgia State University. Mr. Zheng’s research areas focus on wireless network and Big Data.

Guangchun Luo received the Ph.D. degree in computer science from University of Electronic Science and Technology of China, Chengdu, China, in 2004. He is currently a professor and the Secretary of School of Information and Software Engineering at the UESTC. He has published over sixty journal and conference papers in his fields. His research interests include computer networks and big data.

Zaobo He received his Ph.D. degree from the Department of Computer Science at Georgia State University. Dr. He is currently an Assistant Professor in the Department of Computer Science at Miami University. His research interests include data privacy, cyber-security, and big data analytics.
Highlights

Susceptible User Search for Defensing Opinion Manipulation
Wenyi Tang, Ling Tian, Xu Zheng, Guangchun Luo, Zaobo He

- This work adopts advanced AI-based countermeasures to search susceptible and key users for purposefully defensing the opinion manipulation on social networks.
- We propose an advanced greedy algorithm to overcome the shortage of the state-of-art algorithm on susceptibility determination.
- An accelerated algorithm is designed, which adopts unsupervised learning and fully utilizes the community structure of the network.
- Systematical analysis of approximation guarantee is performed on both algorithms for some practical circumstances.
- Extensive evaluations on different real-world datasets reveal that our algorithms significantly outperform the state-of-the-art algorithm.
Author statement

Wenyi Tang: Conceptualization, Methodology, Software, Validation, Writing-Reviewing and Editing.

Ling Tian: Advice and Supervision.

Xu Zheng: Writing-Reviewing and Editing.

Guangchun Luo: Advice and Supervision.

Zaobo He: Software, Validation.
Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.