# Group-Level Influence Maximization with Budget Constraint

Qian Yan[1], Hao Huang[1](✉), Yunjun Gao[2], Wei Lu[3], and Qinming He[2]

[1] State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China
{qy,haohuang}@whu.edu.cn
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{gaoyj,hqm}@zju.edu.cn
[3] DEKE, School of Information, MOE, Renmin University of China, Beijing, China
lu-wei@ruc.edu.cn

**Abstract.** Influence maximization aims at finding a set of seed nodes in a social network that could influence the largest number of nodes. Existing work often focuses on the influence of individual nodes, ignoring that infecting different seeds may require different costs. Nonetheless, in many real-world applications such as advertising, advertisers care more about the influence of groups (e.g., crowds in the same areas or communities) rather than specific individuals, and are very concerned about how to maximize the influence with a limited budget. In this paper, we investigate the problem of group-level influence maximization with budget constraint. Towards this, we introduce a statistical method to reveal the influence relationship between the groups, based on which we propose a propagation model that can dynamically calculate the influence spread scope of seed groups, following by presenting a greedy algorithm called GLIMB to maximize the influence spread scope with a limited cost budget via the optimization of the seed-group portfolio. Theoretical analysis shows that GLIMB can guarantee an approximation ratio of at least $(1 - 1/\sqrt{e})$. Experimental results on both synthetic and real-world data sets verify the effectiveness and efficiency of our approach.

## 1 Introduction

Given a social network, influence maximization aims at finding a subset of nodes (refer to as seeds) that could influence the largest number of nodes [7]. Over the last decade, this problem has received considerable attention due to its key importance in applications such as epidemic prevention, public opinion monitoring and viral marketing, in which local influence relationships between people may lead to an unexpectedly wide spread of disease, ideas, and product adoption [1,2].

Existing studies on influence maximization mostly focus on the influence of individuals [11], ignoring different costs required for infecting different seeds [13]. Nonetheless, analyzing the influence of specific individuals is trivial in many

S. Candan et al. (Eds.): DASFAA 2017, Part I, LNCS 10177, pp. 625–641, 2017.
DOI: 10.1007/978-3-319-55753-3_39

real-world scenarios, and cost budgets for infecting seeds are usually limited in practice. For example, to prevent and control epidemic diseases, establishing epidemic prevention stations is a common method. The siting of the stations depends on, firstly, the influence of crowds grouped in the same areas rather than the influence of individuals, and secondly, the cost of establishing a station in each area. The final goal, as a rule, is to make the best of the cost budget and prevent the diseases as much as possible. Analogous considerations also exist in advertising and promotional activities.

Driven by the practical applications above, in this paper, we study the problem of group-level influence maximization with budget constraint. A straightforward solution could be extending the existing individual-level approaches to solve this problem, since the influence of a group can be considered to be the sum of the influence of each group member [11]. Nonetheless, this solution has three drawbacks. (1) As the number of individuals is much greater than that of groups, analysing the influence of individuals instead of groups is much more computationally expensive. (2) Exact and clear influence relationship between individuals are hard to obtain. For example, although an epidemic prevention station can identify an infected person, it is difficult to find out which one infected him or her. (3) Cost estimation is based mostly on groups, e.g., crowds in the same geographical regions or human social communities, but few based on individuals.

To avoid the drawbacks of individual-level approaches, we propose to analyse the influence relationship at the level of groups. To this end, we adopt association probability to describe the influence relationship between a group pair. With historical infection data sets, we learn the association probability by checking the conditional independence between the infection statuses of each two groups, and construct an influence relationship graph. With the graph, we present an influence propagation model which can calculate the influence spread scope of any group or group set.

Based on the aforementioned group-level influence relationship graph and influence propagation model, we propose GLIMB (**G**roup-**L**evel **I**nfluence **M**aximization with **B**udget constraint) algorithm to approximate the optimal seed groups that maximize the influence spread scope with a limited cost budget. Towards this, GLIMB keeps searching the group that maximizes the incremental spread scope over cost ratio. Before the ending of searching, GLIMB checks whether there is an alternative group or group portfolio that can bring a greater incremental spread scope. Theoretical analysis shows that GLIMB provides at least a $(1 - 1/\sqrt{e})$-approximation. Experimental results on synthetic and real-world data sets verify the effectiveness and efficiency of our approach.

The remaining sections are organized as follows. In Sect. 2, we review the related work. In Sect. 3, we introduce how to construct the influence relationship graph and model the influence propagation at the level of groups, followed by presenting our GLIMB algorithm in Sect. 4. Experimental results and our findings are reported in Sect. 5 before concluding the paper in Sect. 6.

## 2 Related Work

The related work to group-level influence maximization with budget constraint can be classified into three categories, i.e., (1) influence relationship modeling, (2) individual-level influence maximization, and (3) influence maximization with budget constraint.

*Influence relationship modeling* aims at inferring the influence relationship between entities. Existing work focuses on inferring individual-level influence relationship based on historical infection statuses and infection time. Individuals that are sequentially infected within a time interval are regarded to have influence relationship [4,12]. Nonetheless, this idea is not appropriate to inferring group-level influence relationship. Because that in a group, the infection time of different individuals often vary a lot so that the time interval between the infections of two groups is hard to be determined. To model the group-level influence relationship, only a few approaches have been proposed. COLD model [5] carries out this work via subject analysis, but cannot construct an influence relationship graph for influence maximization. CSI model [11] regards the individual-level influence relationship across two groups as the group-level influence relationship, but requires the influence relationship between individuals, which is hard to obtain in practice. As CSI needs to calculate the strength of influence relationship between each two individuals, its time complexity is $O(n^2)$, where $n$ is the number of individuals.

*Individual-level influence maximization* tries to find top $k$ individuals that can maximize the expected influence spread scope. The existing approaches to this problem can be divided into two types, namely (1) greedy searching approaches [1,7,15,16], which utilize the submodularity of influence propagation model and keep selecting individuals that maximize current incremental spread scope, and (2) heuristic searching approaches, which efficiently identify the candidate seeds satisfying some heuristic rules, e.g., having the highest degree [3], influence ranking [6] or local influence [2]. Nonetheless, these approaches mostly still require the influence relationship between individuals.

*Influence maximization with budget constraint* considers the cost of each seed, and tries to make the best of a cost budget to maximize the influence spread scope. Only a few literature [10,13] addresses this problem with a basic idea of iteratively selecting a candidate seed that maximizes the incremental spread scope over cost ratio. Among the existing approaches, the best guarantee for the approximation ratio is $(1 - 1/\sqrt{e})$ [13], while our proposed GLIMB algorithm provides at least a $(1 - 1/\sqrt{e})$-approximation.

## 3 Group-Level Influence Propagation Model

In this paper, we assume that the underlying influence propagation between the individuals follows the IC (Independent Cascade) model, which is one of the most commonly used influence propagation models for individuals. Table 1 summarizes the notations. Given a set $\mathbf{S} \subseteq \mathbf{V}$ ($\mathbf{V} = \{v_1, \ldots, v_n\}$ refers to the set of

$n$ individuals) of seed individuals, IC model works as follows: Let $\mathbf{S}_t$ be the set of individuals newly infected at time $t$, with $\mathbf{S}_0 = \mathbf{S}$ and $\mathbf{S}_t \cap \mathbf{S}_{t+1} = \emptyset$. In round $t+1$, each infected individual $u \in \mathbf{S}_t$ tries to infect its uninfected neighbors (i.e., uninfected individuals having influence relationship with $u$) in $\mathbf{V} \setminus \bigcup_{0 \leq j \leq t} \mathbf{S}_j$ independently with probability $p_{u,v}$. When there are multiple infected individuals trying to infect the same uninfected individual simultaneously, the infections can be carried out in any order. If the influence relationship between individuals is given, IC model can help calculate the influence spread scope $\sigma(\mathbf{S})$ of $\mathbf{S}$, which is the expected number of infected individuals given seed set $\mathbf{S}$.

Nevertheless, in many real-world application scenarios of influence maximization, the influence relationship between individuals is hard to obtain. It is often the case that available data resources only include the historical infection state $s_i^k \in \{1, 0\}$ of each individual $v_i \in \mathbf{V}$ in the $k$-th ($k \in \{1, \ldots, \kappa\}$) outbreak of an infection event, forming a historical infection data set $\mathbf{D} = \left\{ s_i^k \in \{1, 0\} \mid 1 \leq i \leq n, 1 \leq k \leq \kappa \right\}$. In order to find out which seeds can bring the greatest influence spread scope, the influence propagation model should be constructed in advance. In this paper, we propose to carry out this work at the granularity

**Table 1.** Notations

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| $\mathbf{V}$ | The set of individuals | $\mathbf{M}$ | The set of groups of individuals in $\mathbf{V}$ |
| $n$ | The number of individuals in $\mathbf{V}$ | $m$ | The number of groups in $\mathbf{M}$ |
| $v_i$ | The $i$-th individual in $\mathbf{V}$ | $M_i$ | The $i$-th group in $\mathbf{M}$ |
| $\mathbf{S}$ | The set of seed groups, $\mathbf{S} \subseteq \mathbf{M}$ | $\kappa$ | The number of infection outbreaks |
| $s_i^k$ | The historical infection state of $v_i$ in the $k$-th infection outbreak | $\mathbf{D}$ | The historical infection data set that records each $s_i^k$, where $k \in \{1, \ldots, \kappa\}$ |
| $X$ | Any group in $\mathbf{M}$ | $\|X\|$ | The number of individuals in $X$ |
| $x$ | Infection status value of $X$, $x \in \{0, 1\}$ | $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$ | The group-level influence relationship graph |
| $\mathbf{E}$ | The set of directed edges in $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$ | $\mathbf{W}$ | The set of edge weights of edges in $\mathbf{E}$ |
| $W_{ij}$ | The edge weight of directed edge $(M_i, M_j)$ | $p_{ij}$ | The probability that $M_i$ can influence $M_j$ |
| $\mathbf{N}'_{M_i}$ | The set of groups that can directly influence $M_i$, $\mathbf{N}'_{M_i} = \{M_j \mid (M_j, M_i) \in \mathbf{E}\}$ | $\mathbf{N}_{M_i}$ | The set of groups that can be directly influenced by $M_i$, $\mathbf{N}_{M_i} = \{M_j \mid (M_i, M_j) \in \mathbf{E}\}$ |
| $\theta_i$ | The infection ratio of $M_i$ | $\theta_i^t$ | The infection ratio of $M_i$ at time $t$ |
| $\theta_{i \to j}$ | The pass ratio of influence on a directed edge $(M_i, M_j)$ from $M_i$ to $M_j$ | $\theta_{i \to j}^t$ | The pass ratio of influence on a directed edge $(M_i, M_j)$ from $M_i$ to $M_j$ at time $t$ |
| $\theta_{\mathbf{S} \to j}$ | The pass ratio of influence from $\mathbf{S}$ to $M_j$ | $\sigma(\mathbf{S})$ | The influence spread scope of $\mathbf{S}$ |
| $\mathbf{M}_{\mathbf{S} \to j}$ | The set of groups in all the path from any seed group in $\mathbf{S}$ to $M_j$ | $\mathbf{\Gamma}_{\mathbf{S} \to j}$ | The set of groups in $\mathbf{M}_{\mathbf{S} \to j}$ that can directly influence $M_j$ |
| $I_{i \to j}$ | The set of individuals in $M_j$ that are infected by $M_i$ | $I_{\mathbf{S} \to j}$ | The set of individuals in $M_j$ that are infected by $\mathbf{S}$ |
| $c(M_i)$ | The cost of infecting $M_i$ | $b$ | The cost budget |
| $\triangle\sigma(\mathbf{S}, M_i)$ | The incremental spread scope caused by adding $M_i$ in to $\mathbf{S}$ | $\delta(\mathbf{S}, M_i)$ | The ratio of incremental spread scope $\triangle\sigma(\mathbf{S}, M_i)$ over cost $c(M_i)$ |
| $\mathbf{L}$ | The set of candidate seed groups | $\mathbf{P}$ | The set of candidate seed group pairs |

of groups via the following two steps, namely (1) revealing influence relationship between groups, and (2) modeling influence propagation between groups.

### 3.1 Revealing Influence Relationship Between Groups

Statistically speaking, if a group of individuals can influence another group, there should be an association relationship between the two groups. The strength of the influence can be reflected by an association probability. To check association relationship, statistical independence testing is a commonly used approach. Formally, for any infection status values $x \in \{0, 1\}$ and $y \in \{0, 1\}$ of groups $X$ and $Y$, if relationship $p(x, y) = p(x)p(y)$ always holds, then $X$ and $Y$ are called independent to each other, denoted as $(X \perp\!\!\!\perp Y)$. Here, probabilities $p(x)$, $p(y)$ and $p(x, y)$ can be estimated based on the historical infection data set $\mathbf{D}$ under the assumption that individuals in a same group are homogeneous [12]. With this homogeneous assumption, the infection probability of each individual in a group is equal to each other, and can be regarded as the infection probability of the group. In other words, if we observed that 20% of individuals in a group were infected by a disease, it indicates that the disease has a 20 percent chance of infecting each individual of this group. Hence, we utilize $\mathbf{D}$ to estimate the infection probability of each group. For example, in the $k$-th outbreak of an infection event, $p_k(x = 1) = \sum_{v_i \in X} s_i^k / |X|$, where $|X|$ refers to the number of individuals in $X$; then, considering all $\kappa$ outbreaks of this infection event, we have $p(x = 1) = \sum_{k=1}^{\kappa} p_k(x = 1)/\kappa$ and $p(x = 0) = 1 - p(x = 1)$. Moreover, according to the definitions of joint probability and conditional probability, we can also estimate probabilities $p(x, y)$ and $p(x|y)$ based on $\mathbf{D}$.

Nonetheless, sometimes, the independence between $X$ and $Y$ is not enough to comprehensively express the association relationship between groups $X$ and $Y$. Because $X$ may influence $Y$ directly, or $X$ may influence $Y$ through group $Z$ even if $X$ cannot directly influence $Y$. Both cases result in $(X \not\perp\!\!\!\perp Y)$. To avoid this ambiguity, we adopt conditional independence to reveal the direct association relationship between groups $X$ and $Y$. For any infection status values $x \in \{0, 1\}$, $y \in \{0, 1\}$ and $z \in \{0, 1\}$ of groups $X$, $Y$ and $Z$, if $p(x, y \mid z) = p(y \mid z)p(x|z)$ always holds, then $X$ is independent of $Y$ conditioned on $Z$, denoted as $(X \perp\!\!\!\perp Y \mid Z)$. The physical interpretation of $(X \perp\!\!\!\perp Y \mid Z)$ is the independence between $X$ and $Y$ when the mediating effect of $Z$ is excluded.

In information theory, conditional mutual information is commonly used to quantify the conditional independence. Formally, given $Z$, the conditional mutual information of $X$ and $Y$, denoted by $Inf(X, Y \mid Z)$, is calculated as

$$Inf(X, Y \mid Z) = \sum\nolimits_{z \in \{0,1\}} p(z) \sum\nolimits_{x \in \{0,1\}} \sum\nolimits_{y \in \{0,1\}} p(x, y \mid z) log_2 \frac{p(x, y \mid z)}{p(x \mid z)p(y|z)}. \tag{1}$$

$Inf(X, Y \mid Z) = 0$ indicates that $(X \perp\!\!\!\perp Y \mid Z)$. A higher $Inf(X, Y \mid Z)$ indicates a stronger direct association relationship between $X$ and $Y$ given $Z$. Moreover, conditional mutual information has the following properties.

**Theorem 1.** *For any variable sets $X$, $Y$ and $Z$, if the mutual information $Inf(X, Y)$ of $X$ and $Y$ is equal to 0, then relationship $Inf(X, Y \mid Z) = 0$ always holds.*

*Proof.* Since $Inf(X, Y) = Inf(X, Y \mid \emptyset)$, when $Inf(X, Y) = 0$, we have $Inf(X, Y \mid \emptyset) = 0$, indicating that $X$ is independent of $Y$ conditioned on $\emptyset$, i.e. $(X \perp\!\!\!\perp Y \mid \emptyset)$. Then, we can have $(X \perp\!\!\!\perp Y \mid \emptyset \cup Z)$ which is equal to $(X \perp\!\!\!\perp Y \mid Z)$, due to strong union property of conditional independence relation, indicating that $Inf(X, Y \mid Z) = 0$. ∎

**Theorem 2.** *For any variable sets $X$, $Y$, $Z$ and $W$, if $Inf(X, Y \mid Z) = 0$, then relationship $Inf(X, Y \mid Z \cup W) = 0$ always holds.*

*Proof.* $Inf(X, Y \mid Z) = 0$ indicates that $(X \perp\!\!\!\perp Y \mid Z)$. Then, due to strong union property of conditional independence relation, relationship $(X \perp\!\!\!\perp Y \mid Z \cup W)$ also holds, indicating that $Inf(X, Y \mid Z \cup W) = 0$. ∎

With the help of conditional mutual information, if group $X$ has a strong direct association relationship with group $Z$, then we can add a directed edge from $X$ to $Z$ and add $Z$ into the neighbor set $\mathbf{N}_X$ of $X$, indicating that $X$ can directly influence $Z$. When we check the direct association relationship between $X$ and group $Y \notin \mathbf{N}_X$, the mediating effect of groups in $\mathbf{N}_X$ should be excluded by calculating $Inf(X, Y \mid \mathbf{N}_X)$. Based on the above basic ideas, Algorithm 1 provides a construction approach for the group-level influence relationship graph $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$, in which $\mathbf{M}$ is the set of groups, $\mathbf{E}$ is the set of directed edges, and $\mathbf{W}$ is the set of edge weights.

Algorithm 1 takes as inputs the given group set $\mathbf{M}$, and the historical infection data set $\mathbf{D}$, which is used for calculating probabilities required in the computation of conditional mutual information. The algorithm first initializes $\mathbf{E}$, $\mathbf{W}$, and the neighbor set $\mathbf{N}_{M_i}$ for each group $M_i \in \mathbf{M}$ as empty sets (line 1), and calculates the mutual information $Inf(M_i, M_j)$ of each two groups in $\mathbf{M}$ (line 2), which is equal to the conditional mutual information $Inf(M_i, M_j \mid \emptyset)$ of the

**Algorithm 1.** Construction of Group-Level Influence Relationship Graph

```
Input  : Group set $\mathbf{M}$; historical infection data set $\mathbf{D}$.
Output: Influence relationship graph $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$.
1  Initial $\mathbf{E} \leftarrow \emptyset$, $\mathbf{W} \leftarrow \emptyset$, and an empty neighbor set $\mathbf{N}_{M_i}$ for each group $M_i \in \mathbf{M}$;
2  Calculate $Inf(M_i, M_j)$ for each two groups $M_i, M_j \in \mathbf{M}$ $(i \neq j)$;
3  for each $M_i \in \mathbf{M}$ do
4      for each $M_j \in \mathbf{M}$ $(j \neq i)$ having $Inf(M_i, M_j) > 0$ do
5          if $Inf(M_i, M_j \mid \mathbf{N}_{M_i}) > \varepsilon$ then
6              $\mathbf{N}_{M_i} \leftarrow \mathbf{N}_{M_i} \cup \{M_j\}$;
7      for each $M_j \in \mathbf{N}_{M_i}$ do
8          $\mathbf{E} \leftarrow \mathbf{E} \cup \{(M_i, M_j)\}$;            //$(M_i, M_j)$ is a directed edge from $M_i$ to $M_j$
9          $W_{ij} \leftarrow Inf(M_i, M_j \mid \mathbf{N}_{M_i} \backslash \{M_j\})$;      //$W_{ij}$ is the weight of edge $(M_i, M_j)$
10         $\mathbf{W} \leftarrow \mathbf{W} \cup \{W_{ij}\}$;
```

two groups conditioned on $\emptyset$. Then, for each group $M_i \in \mathbf{M}$, it identifies which of other groups $\{M_j \in \mathbf{M} \mid i \neq j\}$ have strong direct association relationship with group $M_i$ by checking whether $Inf(M_i, M_j \mid \mathbf{N}_{M_i})$ is greater than a threshold $\varepsilon$ (line 5). Instead of adopting a user-specified $\varepsilon$, we suggest to determine the $\varepsilon$ based on mutual information $Inf(M_i, M_j)$ of each two groups in $\mathbf{M}$. Specifically, by performing $K$-means with $K = 2$, the non-zero values of mutual information can be classified into two parts. Let $\varepsilon$ be the minimal value in the part containing greater mutual information values. Then, condition $Inf(M_i, M_j \mid \mathbf{N}_{M_i}) > \varepsilon$ helps find direct association relationship that are strong enough. Groups that satisfy the condition above will be added into the neighbor set $\mathbf{N}_{M_i}$ of $M_i$ (line 6). Finally, for each $M_j \in \mathbf{N}_{M_i}$, we add the directed edge $(M_i, M_j)$ into the edge set $\mathbf{E}$ (line 8), calculate the weight $W_{ij}$ of this edge (line 9), followed by adding $W_{ij}$ into the edge weight $\mathbf{W}$ (line 10).

The overall time complexity of Algorithm 1 is $O(mn\kappa + m^2 + 2^{\alpha}m^2)$, where $m$ is the number of groups in $\mathbf{M}$, $n$ is the number of individuals in all groups (usually $m \ll n$), $\kappa$ is the number of historical infection outbreaks recorded in $\mathbf{D}$, and $\alpha$ is the maximal number of groups that can be directly influenced by each $M_i \in \mathbf{M}$ (i.e., $\alpha = \max_{1 \leq i \leq m} |\mathbf{N}_{M_i}|$). To be specific, statistics for the probabilities used for calculating conditional mutual information take $O(mn\kappa)$ time. Then, calculating mutual information in line 2 requires $O(m^2)$ time. The time complexity of the loop of line 3 is $O(2^{\alpha}m^2)$. In the loop of line 3, the most computationally expensive step is in line 5, i.e., the computation of conditional mutual information. For each $M_i$ and $M_j$, it takes $O(2^{|\mathbf{N}_{M_i}|})$ time, where $|\mathbf{N}_{M_i}| \leq \alpha$ refers to the number of groups that can be directly influenced by group $M_i$. In practice, the influence of each group is usually limited, and only a few groups can be directly influenced by each group ($\alpha \ll m$). Furthermore, to reduce the time complexity of line 5, users can adopt a greater threshold $\varepsilon$ to reduce the cardinality $|\mathbf{N}_{M_i}|$ of each set $\mathbf{N}_{M_i}$ and obtain a smaller $\alpha$. Though, the compensation of a greater $\varepsilon$ is that the constructed graph will have less edges which only capture the strongest direct association relationship. Besides, to avoid unnecessary calculation in line 5, we carry out a pruning in line 4. According to Theorems 1 and 2, if $Inf(M_i, M_j) = 0$, then $Inf(M_i, M_j \mid \mathbf{N}_{M_i}) = 0$ always holds. Thus, it is not necessary to calculate $Inf(M_i, M_j \mid \mathbf{N}_{M_i})$ for each $M_j \in \mathbf{M}$ $(j \neq i)$ having $Inf(M_i, M_j) = 0$.

With the influence relationship graph constructed by Algorithm 1, in what follows, we introduce how the influence is propagated between the groups on the graph.

### 3.2 Modeling Influence Propagation Between Groups

In this section, we first introduce (1) the pass ratio of influence on each directed edge, based on which we elaborate (2) the rules of influence propagation on the influence relationship graph, followed by presenting (3) the function of influence spread scope, which calculates the expected number of of infected individuals given seed groups.

**Pass Ratio of Influence.** For groups $M_i$, $M_j \in \mathbf{M}$ in the influence relationship graph $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$, if there is a directed edge $(M_i, M_j) \in \mathbf{E}$ from $M_i$ to $M_j$, then $M_i$ can directly influence $M_j$, and the strength of influence is proportional to the edge weight $W_{ij} \in \mathbf{W}$. Among all the groups $\mathbf{N}'_{M_j} = \{M_\ell \in \mathbf{M} \mid (M_\ell, M_j) \in \mathbf{E}\}$ that can directly influence $M_j$, group $M_i \in \mathbf{N}'_{M_j}$ can influence group $M_j$ with a probability $p_{ij} = W_{ij} / \sum_{M_\ell \in \mathbf{N}'_{M_j}} W_{\ell j}$.

Furthermore, according to IC model, uninfected individuals can only be influenced by infected ones. Thus, when the infection ratio $\theta_j$ of $M_j$ is less than 1, $M_i$ with a higher infection ratio $\theta_i$ (i.e., more infected individuals) has more chances to influence $M_j$.

In brief, the pass ratio of influence on a directed edge $(M_i, M_j) \in \mathbf{E}$, denoted by $\theta_{i \to j}$, is affected by infection ratio $\theta_i$ of the influence source $M_i$, the influence probability $p_{ij}$ from $M_i$ to $M_j$, and the infection ratio $\theta_j$ of the target group $M_j$. Formally, $\theta_{i \to j}$ can be calculated as $\theta_{i \to j} = \theta_i \times p_{ij} \times (1 - \theta_j)$. The physical interpretation of $\theta_{i \to j}$ is the newly-increased infection ratio of $M_j$ caused by the influence from $M_i$.

**Rules of Influence Propagation.** According to IC model, infected individuals at time $t$ (or round $t$) only have infectivity at time $t + 1$. Analogously, in the process of group-level influence propagation, we calculate and record the infection ratio $\theta_i^t$ of each $M_i \in \mathbf{M}$ at time $t$, based on which we can deduce each infection ratio $\theta_i^{t+1}$ at time $t + 1$.

At time $t = 0$, if the set $\mathbf{S}$ of seed groups is given, the expected infection ratio $\theta_j^0$ of $M_j \in \mathbf{S}$ can be predicted based on historical infection data.

At time $t > 0$, if the infection ratio of a group $M_i \in \mathbf{M}$ has increased at time $t - 1$, then $M_i$ will try to influence its neighbors $M_\ell \in \mathbf{N}_{M_i}$ through the directed edge $(M_i, M_\ell) \in \mathbf{E}$. For a target group $M_\ell \in \mathbf{N}_{M_i}$, (1) if $\mathbf{N}'_{M_\ell} = \{M_i\}$, i.e., $M_\ell$ will be only influenced by $M_i$, then the infection ratio $\theta_\ell^t$ of $M_\ell$ at time $t$ can be calculated as $\theta_\ell^t = \theta_\ell^{t-1} + \theta_{i \to \ell}^t$, where $\theta_{i \to \ell}^t = \theta_i^{t-1} \times p_{i\ell} \times (1 - \theta_\ell^{t-1})$; (2) if $\mathbf{N}'_{M_\ell} = \{M_i, M_k\}$ and infection ratio of $M_k$ has increased at time $t - 1$, then $M_i$ and $M_k$ will influence $M_\ell$ simultaneously at time $t$. Following the rules of IC model, the influences can be carried out in any order. Let $M_i$ execute the influence first, we can calculate $\theta_{i \to \ell}^t$ in the same way, and exclude this newly-increased infection ratio of $M_\ell$ in the calculation of $\theta_{k \to \ell}^t$ to avoid repeatedly infecting the same part in $M_\ell$. Specifically, $\theta_{k \to \ell}^t = \theta_k^{t-1} \times p_{k\ell} \times (1 - \theta_\ell^{t-1} - \theta_{i \to \ell}^t)$, and the infection ratio of $M_\ell$ at time $t$ can be updated by $\theta_\ell^t = \theta_\ell^{t-1} + \theta_{i \to \ell}^t + \theta_{k \to \ell}^t$. The above calculation rules can be easily extended to the cases that more groups influence $M_\ell$ simultaneously.

Moreover, we consider that the process of influence propagation is acyclic, i.e., once $M_i$ pass its influence to $M_j$, $M_j$ will not pass back its influence to $M_i$ any more. This consideration is commonly used in influence maximization [13]. When the infection ratio of each group does not increase any more, the influence propagation will end.

**Function of Influence Spread Scope.** According to the rules of influence propagation, each $M_j \in \mathbf{M} \backslash \mathbf{S}$ can be influenced by $\mathbf{S}$ iff there is at least one directed path from any seed group to $M_j$. Let $\mathbf{M}_{\mathbf{S}\to j}$ denote the union of groups in all the paths from any seed group in $\mathbf{S}$ to $M_j$, group set $\mathbf{\Gamma}_{\mathbf{S}\to j} = \mathbf{M}_{\mathbf{S}\to j} \cap \mathbf{N}'_{M_j}$ refers to the groups that can directly influence $M_j$ in $\mathbf{M}_{\mathbf{S}\to j}$. By combining the influence from each $M_k \in \mathbf{\Gamma}_{\mathbf{S}\to j}$ to $M_j$, we have the recursion formula for the pass ratio of influence from $\mathbf{S}$ to $M_j$, i.e.,

$$\theta_{\mathbf{S}\to j} = \begin{cases} 1 - \prod_{M_k \in \mathbf{\Gamma}_{\mathbf{S}\to j}} (1 - \theta_{\mathbf{S}\to k} \times p_{kj}), & M_j \notin \mathbf{S} \\ \theta_j^0, & M_j \in \mathbf{S} \end{cases} \tag{2}$$

Let $|M_j|$ denote the number of individuals in group $M_j$, the function $\sigma(\mathbf{S})$ of influence spread scope can be be formulated as follows.

$$\sigma(\mathbf{S}) = \sum_{M_j \in \mathbf{M}} |M_j| \times \theta_{\mathbf{S}\to j}. \tag{3}$$

Function $\sigma(\mathbf{S})$ has the following properties.

**Theorem 3.** *Function $\sigma(\mathbf{S})$ is monotone.*

*Proof.* We first proof that $\theta_{\mathbf{S}\to j}$ is monotone increasing, i.e., given $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$ and $\mathbf{S} \subseteq \mathbf{T} \subseteq \mathbf{M}$, for any $M_j \in \mathbf{M}$, relationship $\theta_{\mathbf{T}\to j} \geq \theta_{\mathbf{S}\to j}$ always holds. When $M_j \in \mathbf{S}$, according to the definition of $\theta_{\mathbf{S}\to j}$, we have $\theta_{\mathbf{S}\to j} = \theta_{\mathbf{T}\to j} = \theta_j^0$. Hence, relationship $\theta_{\mathbf{S}\to j} \leq \theta_{\mathbf{T}\to j}$ holds for $M_j \in \mathbf{S}$. When $M_j \notin \mathbf{S}$, we can proof relationship $\theta_{\mathbf{S}\to j} \leq \theta_{\mathbf{T}\to j}$ by induction. (1) For each $M_k \in \mathbf{M}_{\mathbf{S}\to j}$ which is directed influenced by $\mathbf{S}$, if there is a directed path from $\mathbf{T}\backslash\mathbf{S}$ to $M_k$, i.e., $\theta_{\mathbf{T}\backslash\mathbf{S}\to k} > 0$, then $\theta_{\mathbf{T}\to k} > \theta_{\mathbf{S}\to k}$; otherwise $\theta_{\mathbf{T}\to k} = \theta_{\mathbf{S}\to k}$. (2) For each $M_\ell \in \mathbf{M}_{\mathbf{S}\to j}$ which is directed influenced by $M_k$, since $\theta_{\mathbf{T}\to k} \geq \theta_{\mathbf{S}\to k}$, we have relationship $1 - \prod_{M_k \in \mathbf{\Gamma}_{\mathbf{S}\to \ell}} (1 - \theta_{\mathbf{T}\to k} \times p_{k\ell}) \geq 1 - \prod_{M_k \in \mathbf{\Gamma}_{\mathbf{S}\to \ell}} (1 - \theta_{\mathbf{S}\to k} \times p_{k\ell}) = \theta_{\mathbf{S}\to \ell}$. Moreover, since $\mathbf{S} \subseteq \mathbf{T} \subseteq \mathbf{M}$, we have $\mathbf{M}_{\mathbf{S}\to \ell} \subseteq \mathbf{M}_{\mathbf{T}\to \ell}$, and thus $\mathbf{\Gamma}_{\mathbf{S}\to \ell} \subseteq \mathbf{\Gamma}_{\mathbf{T}\to \ell}$. Then, relationship $\theta_{\mathbf{T}\to \ell} = 1 - \prod_{M_k \in \mathbf{\Gamma}_{\mathbf{T}\to \ell}} (1 - \theta_{\mathbf{T}\to k} \times p_{k\ell}) \geq 1 - \prod_{M_k \in \mathbf{\Gamma}_{\mathbf{S}\to \ell}} (1 - \theta_{\mathbf{T}\to k} \times p_{k\ell})$ holds, and hence relationship $\theta_{\mathbf{T}\to \ell} \geq \theta_{\mathbf{S}\to \ell}$ holds. (3) By induction, we can proof that for each group $M_i \in \mathbf{M}_{\mathbf{S}\to j}$, relationship $\theta_{\mathbf{T}\to i} \geq \theta_{\mathbf{S}\to i}$ holds. (4) If there is at least one path from $\mathbf{S}$ to $M_j$, then relationship $\theta_{\mathbf{T}\to j} \geq \theta_{\mathbf{S}\to j}$ holds; otherwise, $\theta_{\mathbf{S}\to j} = 0$, and relationship $\theta_{\mathbf{T}\to j} \geq \theta_{\mathbf{S}\to j}$ also holds since $\theta_{\mathbf{T}\to j} \geq 0$. In summary, $\theta_{\mathbf{S}\to j}$ is monotone increasing.

Since a non-negative linear combination of monotone increasing functions is still a monotone increasing function, function $\sigma(\mathbf{S})$ is a monotone increasing function. ∎

**Theorem 4.** *Function $\sigma(\mathbf{S})$ is submodular.*

*Proof.* We first proof the submodularity of $\theta_{\mathbf{S}\to j}$, i.e., given $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$ and $\mathbf{S} \subseteq \mathbf{T} \subseteq \mathbf{M}$, for any $M_i, M_j \in \mathbf{M}$, relationship $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} \geq \theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j}$ always holds. (1) When $M_i \in \mathbf{S}$, we have $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} = \theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j} = 0$. Hence, the relationship $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} \geq \theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j}$ holds for $M_i \in \mathbf{S}$. (2) When $M_i \in \mathbf{T}\backslash\mathbf{S}$, we have $\theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j} = 0$. Since $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} \geq 0$, the relationship $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} \geq \theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j}$ holds for $M_i \in \mathbf{T}\backslash\mathbf{S}$. (3) When $M_i \notin \mathbf{T}$, we proof the submodularity of $\theta_{\mathbf{S}\to j}$

at the granularity of individuals. Let $\theta_{i\to j}$ denotes the newly-increased infection ratio of $M_j$ caused by $M_i$ regardless of the influence of any other seed groups. As $M_j \notin \mathbf{S} \subseteq \mathbf{T}$, $\theta_j^0 = 0$. Hence, $|M_j| \times \theta_{i\to j}$ is the expected number of individuals (directly and indirectly) infected by $M_i$. We denote these individuals by set $I_{i\to j}$. Similarly, sets $I_{\mathbf{S}\to j}$ and $I_{\mathbf{T}\to j}$ refer to the individuals (directly and indirectly) infected by $\mathbf{S}$ and $\mathbf{T}$, respectively, regardless of the influence of any other seed groups. $I_{\mathbf{S}\to j} \subseteq I_{\mathbf{T}\to j}$ since $\mathbf{S} \subseteq \mathbf{T}$. With sets $I_{i\to j}$, $I_{\mathbf{S}\to j}$ and $I_{\mathbf{T}\to j}$, we have $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} = \big(|I_{i\to j}| - |I_{i\to j} \cap I_{\mathbf{S}\to j}|\big)/|M_j|$, and $\theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j} = \big(|I_{i\to j}| - |I_{i\to j} \cap I_{\mathbf{T}\to j}|\big)/|M_j|$. As $I_{\mathbf{S}\to j} \subseteq I_{\mathbf{T}\to j}$, we have $|I_{i\to j} \cap I_{\mathbf{S}\to j}| \leq |I_{i\to j} \cap I_{\mathbf{T}\to j}|$, and thus $\big(|I_{i\to j}| - |I_{i\to j} \cap I_{\mathbf{S}\to j}|\big)/|M_j| \geq \big(|I_{i\to j}| - |I_{i\to j} \cap I_{\mathbf{T}\to j}|\big)/|M_j|$, indicating that relationship $\theta_{\mathbf{S}\cup\{M_i\}\to j} - \theta_{\mathbf{S}\to j} \geq \theta_{\mathbf{T}\cup\{M_i\}\to j} - \theta_{\mathbf{T}\to j}$ holds for $M_i \notin \mathbf{T}$, $M_j \notin \mathbf{S}$.

In brief, $\theta_{\mathbf{S}\to j}$ is submodular. Since a non-negative linear combination of submodular functions is still a submodular function, function $\sigma(\mathbf{S})$ is also submodular. ∎

**Corollary 1.** *Given* $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$*, relationship* $\sigma(\{M_i\}) \geq \sigma(\mathbf{S} \cup \{M_i\}) - \sigma(\mathbf{S})$ *holds for any group set* $\mathbf{S} \subset \mathbf{M}$ *and group* $M_i \in \mathbf{M}\backslash\mathbf{S}$.

*Proof.* Since $\sigma(\mathbf{S})$ is submodular, i.e., given $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$ and $\mathbf{T} \subseteq \mathbf{S} \subseteq \mathbf{M}$, for any $M_i \in \mathbf{M}$, relationship $\sigma(\mathbf{T} \cup \{M_i\}) - \sigma(\mathbf{T}) \geq \sigma(\mathbf{S} \cup \{M_i\}) - \sigma(\mathbf{S})$ always holds, we have $\sigma(\{M_i\}) = \sigma(\emptyset \cup \{M_i\}) - \sigma(\emptyset) \geq \sigma(\mathbf{S} \cup \{M_i\}) - \sigma(\mathbf{S})$. ∎

## 4 The GLIMB Algorithm

Given the constructed influence relationship graph $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$, the function $\sigma(\mathbf{S})$ of influence spread scope, and the cost $c(M_i)$ for each group $M_i \in \mathbf{M}$, in this section, we address the problem of finding a set $\mathbf{S}$ of groups that maximizes $\sigma(\mathbf{S})$ under a cost budget constraint $b$, i.e., $\max_{\mathbf{S}\subseteq\mathbf{M}} \sigma(\mathbf{S})$, s.t. $\sum_{M_i\in\mathbf{S}} c(M_i) \leq b$.

An intuitive strategy, which can be denoted as `NaiveGreedy`, is to select at each step a group $M_i$ that maximizes the incremental spread scope over cost ratio $\delta(\mathbf{S}, M_i) = \big(\sigma(\mathbf{S}\cup\{M_i\}) - \sigma(\mathbf{S})\big)/c(M_i)$ if $c(M_i)$ is less than the remaining budget [10,13]. However, this strategy is easy to plunge into local optima. For example, assume there are three equal-sized groups $M_1$, $M_2$ and $M_3$, $\theta_1^0 = \theta_2^0 = \theta_3^0$, $c(M_1) = 0.9$, $c(M_2) = c(M_3) = 2$, and cost budget $b = 2$. Let $M_1$ be an isolated group, while $M_2$ and $M_3$ are connected with influence probability one to each other. Then, although $M_2$ or $M_3$ can maximize the spread scope under the cost budget constraint, `NaiveGreedy` will select $M_1$ as the seed group and stop, since $\theta_1^0|M_1|/c(M_1) > (\theta_2^0|M_2| + \theta_3^0|M_3|)/c(M_2) = (\theta_2^0|M_2| + \theta_3^0|M_3|)/c(M_3)$.

To avoid the drawbacks of `NaiveGreedy` method, an improved solution known as `ImprovedGreedy` records the set $\mathbf{S}_{naive}$ of seed groups selected by `NaiveGreedy`, and identifies the group $M_{max}$ having the largest influence scope and a cost no more than $b$, followed by returning set $\mathbf{S} = \arg\max\big(\sigma(\mathbf{S}_{naive}), \sigma(\{M_{max}\})\big)$ as the seed group set. It has been proven that `ImprovedGreedy` provides an approximation ratio of $(1 - 1/\sqrt{e})$, when the function of influence spread scope is monotone and submodular [13].

**Algorithm 2.** The GLIMB Algorithm

```
Input  : $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$; budget $b$; cost $c(M_i)$ and infection ratio $\theta_i^0$ for each
         $M_i \in \mathbf{M}$.
Output: Set $\mathbf{S}$ of seed groups.
Initial $\mathbf{S} \leftarrow \emptyset$, $\mathbf{L} \leftarrow \mathbf{M}$;        //$\mathbf{L}$ records the candidates for seed groups
Calculate $\sigma(\{M_i\})$ for each $M_i \in \mathbf{M}$, $M_\lambda \leftarrow \arg\max_{M_i \in \mathbf{M}} \sigma(\{M_i\})$;
while $\mathbf{L} \neq \emptyset$ do
    $\mathbf{L} \leftarrow \mathbf{L} \backslash \{M_i \in \mathbf{L} \mid c(M_i) > b - \sum_{M_j \in \mathbf{S}} c(M_j)\}$;
    $\ell \leftarrow 0$, $MaxRatio \leftarrow 0$;
    for $M_i \in \mathbf{L}$ do
        if $\frac{\sigma(\{M_i\})}{c(M_i)} > MaxRatio$ then
            $\delta(\mathbf{S}, M_i) \leftarrow \triangle\sigma(\mathbf{S}, M_i)/c(M_i)$; //$\triangle\sigma(\mathbf{S}, M_i) = \sigma(\mathbf{S} \cup \{M_i\}) - \sigma(\mathbf{S})$
            if $\delta(\mathbf{S}, M_i) > MaxRatio$ then
                $MaxRatio \leftarrow \delta(\mathbf{S}, M_i)$, $\ell \leftarrow i$;
    if $\forall M_k \in \mathbf{L} \backslash \{M_\ell\}$, $c(M_k) > b - \sum_{M_j \in \mathbf{S} \cup \{M_\ell\}} c(M_j)$ then
        if $c(M_k) \geq c(M_\ell)$ then
            $M_\ell \leftarrow \arg\max_{M_i \in \mathbf{L},\ b \geq \sum_{M_j \in \mathbf{S} \cup \{M_i\}} c(M_j)} \triangle\sigma(\mathbf{S}, M_i)$;
            $\mathbf{S} \leftarrow \mathbf{S} \cup \{M_\ell\}$, $\mathbf{L} \leftarrow \mathbf{L} \backslash \{M_\ell\}$;
        else
            $\mathbf{P} \leftarrow \{< M_i, M_j >\mid M_i, M_j \in \mathbf{L} \backslash \{M_\ell\},\ i \neq j,\ \triangle\sigma(\mathbf{S}, M_i) +$
            $\triangle\sigma(\mathbf{S}, M_j) > \triangle\sigma(\mathbf{S}, M_\ell),\ c(M_i) + c(M_j) \leq b - \sum_{M_j \in \mathbf{S}} c(M_j)\}$;
            while $\mathbf{P} \neq \emptyset$ do
                $< M_{i^*}, M_{j^*} > \leftarrow \arg\max_{< M_i, M_j > \in \mathbf{P}} \big(\triangle\sigma(\mathbf{S}, M_i) + \triangle\sigma(\mathbf{S}, M_j)\big)$;
                if $\triangle\sigma\big(\mathbf{S} \cup \{M_{i^*}\}, M_{j^*}\big) > \triangle\sigma(\mathbf{S}, M_\ell)$ then
                    $\mathbf{S} \leftarrow \mathbf{S} \cup \{M_{i^*} \cup M_{j^*}\}$, $\mathbf{L} \leftarrow \mathbf{L} \backslash \{M_{i^*} \cup M_{j^*}\}$;
                    break;
                else
                    $\mathbf{P} \leftarrow \mathbf{P} \backslash \{< M_{i^*}, M_{j^*} >\}$;
    else
        $\mathbf{S} \leftarrow \mathbf{S} \cup \{M_\ell\}$, $\mathbf{L} \leftarrow \mathbf{L} \backslash \{M_\ell\}$;
    if $\mathbf{L} = \emptyset$ and $\sigma(\{M_\lambda\}) \geq \sigma(\mathbf{S})$ then
        $\mathbf{S} \leftarrow \{M_\lambda\}$, $\mathbf{L} \leftarrow \mathbf{M} \backslash \{M_\lambda\}$;
```

Nevertheless, both `NaiveGreedy` and `ImprovedGreedy` have a high risk of waste budge. For example, (1) Case 1: when the remaining budget is 4, and there are still two candidates $M_1$ and $M_2$, of which the costs are 2 and 3, respectively, assume that $\delta(\mathbf{S}, M_1) = 1$ while $\delta(\mathbf{S}, M_2) = 0.8$, `NaiveGreedy` and `ImprovedGreedy` will select $M_1$ and stop, although candidate $M_2$ can bring a greater incremental spread scope, which is 2.4 (it is 2 for $M_1$); (2) Case 2: when the remaining budget is 4, and there are still two candidates $M_1$, $M_2$ and $M_3$, of which the costs are 3, 2 and 2, respectively, assume that $\delta(\mathbf{S}, M_1) = 1$ while

$\delta(\mathbf{S}, M_2) = \delta(\mathbf{S}, M_3) = 0.8$, `NaiveGreedy` and `ImprovedGreedy` will select $M_1$ and stop, although selecting the candidate portfolio of $M_2$ and $M_3$ can bring a greater incremental spread scope, which is 3.2 (it is 3 for $M_1$).

To avoid the above waste-budget cases, in Algorithm 2, we propose a novel algorithm called GLIMB for the problem of influence maximization with budget constraint.

The GLIMB algorithm takes as inputs the influence relationship graph $G(\mathbf{M}, \mathbf{E}, \mathbf{W})$, the cost budget $b$, the cost $c(M_i)$ for each $M_i \in \mathbf{M}$ (assume that $\forall M_i \in \mathbf{M}$, $c(M_i) \leq b$), and the expected infection ratio $\theta_i^0$ for each $M_i \in \mathbf{M}$ which is used in the calculation of influence spread scope. It first initializes an empty set $\mathbf{S}$ to record the seed groups, and a set $\mathbf{L}$ which is initially set as $\mathbf{M}$ to record which candidate groups are left for $\mathbf{S}$ (line 1), followed by calculating $\sigma(\{M_i\})$ of each $M_i \in \mathbf{M}$ (line 2). Then, it iteratively searches the currently best candidate or candidate portfolio for seed groups. Each iteration has two *routine phases*, namely (1) removing each group having a cost greater than current remaining budget from set $\mathbf{L}$ (line 4), and (2) finding the group $M_\ell \in \mathbf{L}$ that can maximize the incremental spread scope over cost ratio $\delta(\mathbf{S}, M_\ell)$ for current $\mathbf{S}$ (lines 5–10), and two *extra phases* which are carried out before ending, namely (3) searching an alternative candidate or candidate portfolio (if any) that can bring a greater incremental spread scope (lines 11–23), and (4) checking whether the algorithm plunges into local optima with the strategy used by `ImprovedGreedy` (lines 26–27).

The first *extra phase* (lines 11–23) plays the central role to help avoid waste-budget cases. It works as follows. If each remaining candidate has a cost no less than $c(M_\ell)$, it is easy to prove that the remaining cost budget cannot afford more than one alternative candidate. Thus, in the groups that can be afforded by the remaining cost budget, we select the one that can bring the greatest incremental spread scope (line 13) as the latest seed group (line 14). Otherwise, we find out the set $\mathbf{P}$ of alternative candidate portfolios (a portfolio consists of two candidates) that can be afforded by the remaining cost budget (line 16). In $\mathbf{P}$, if there is a candidate portfolio that can bring a greater incremental spread scope than current $\triangle\sigma(\mathbf{S}, M_\ell)$ (line 19), we add the two candidates in this portfolio into $\mathbf{S}$ (line 20), and go to the next iteration of the loop of line 3 (line 21).

The time complexity of GLIMB is dominated by the time complexity of the second *routine phase* (lines 4–10) and the first *extra phase* (lines 11–23). Let $\tau$ be the maximum time required by calculating $\delta(\mathbf{S}, M_i)$ for each candidate group $M_i \in \mathbf{L}$. The second *routine phase* requires $O(m^2\tau)$ time since there are at most $m \times (m-1)$ times of calculation on $\delta(\mathbf{S}, M_i)$, where $m$ is the number of groups in $\mathbf{M}$. To avoid redundant computations in the second *routine phase*, we adopt a pruning method based on Corollary 1, which indicates that the incremental spread scope $\triangle\sigma(\mathbf{S}, M_i)$ of $M_i$ will not be greater than $\sigma(\{M_i\})$. In other words, for each incremental spread scope over cost ratio $\delta(\mathbf{S}, M_i)$, the upper bound is $\sigma(\{M_i\})/c(M_i)$. Hence, if this upper bound $\sigma(\{M_i\})/c(M_i)$ is less than $MaxRatio$ which records current maximal ratio of incremental spread scope over cost, then this $M_i$ definitely cannot maximize the incremental spread scope over cost ratio, and thus can be excluded to calculate the $\delta(\mathbf{S}, M_i)$. The first *extra phase* requires $O\big((|\mathbf{L}|-1)^2\tau\big)$ time since there are at most $|\mathbf{L}|-1$ candidate groups for the calculation of line 13 and $(|\mathbf{L}|-1)^2$ candidate portfolios for the calculation of line 19 ($|\mathbf{L}| \leq m$).

Moreover, GLIMB algorithm has the following performance guarantees.

**Theorem 5.** *GLIMB provides at least a* $(1-1/\sqrt{e})$*-approximation.*

*Proof.* Without the execution of the first *extra phase*, the result of GLIMB is equal to that of `ImprovedGreedy`. If (1) the first *extra phase* is executed and finds an alternative candidate or candidate portfolio that can bring a greater incremental spread scope, and (2) after this execution of the first *extra phase*, the second *extra phase* is not executed till ending, then $\sigma(\mathbf{S}) > \sigma(\mathbf{S}')$; otherwise, $\sigma(\mathbf{S}) = \sigma(\mathbf{S}')$. In brief, we have $\sigma(\mathbf{S}) \geq \sigma(\mathbf{S}')$.

With the monotone and submodular function $\sigma(\cdot)$, `ImprovedGreedy` provides a $(1-1/\sqrt{e})$-approximation. Thus, GLIMB can provide at least a $(1-1/\sqrt{e})$-approximation. ∎

## 5 Experimental Evaluation

In this section, we first describe the data sets used for experiments, and then verify the efficacy of the two algorithms proposed in this paper.

### 5.1 Experimental Setup

We adopt (1) LFR benchmark graphs [8] and (2) Amazon product co-purchasing network [9], respectively, as the underlying individual-level influence relationship graphs. A LFR benchmark graph can be generated by setting the number $n$ of nodes, the number $m$ of communities (groups), the average size *avg-s* of communities, and the average degree *avg-d* of each node. If there is an edge between two nodes, we regard that these two nodes can directly influence each other. We generate three series of LFR benchmark graphs with properties summarized in Table 2. Amazon product co-purchasing network was crawled from Amazon website. It contains 262111 nodes, each of which refers to a product, and 1234877 directed edges. A directed edge from node $i$ to node $j$ indicates that the $i$-th product is frequently co-purchased with the $j$-th product. Community detection approaches [14] can help divide the nodes into different number of groups.

**Table 2.** Properties of LFR benchmark graphs used for experiments

| Graph data sets | Group number $m$ | Average group size *avg-s* | Average degree *avg-d* |
|---|---|---|---|
| LFR1.1–1.5 | 50, 100, 150, 200, 250 | 200 | 3 |
| LFR2.1–2.5 | 200 | 50, 100, 150, 200, 250 | 3 |
| LFR3.1–3.5 | 200 | 150 | 3, 5, 7, 9, 11 |

The historical infection data set $\mathbf{D}$ can be obtained by simulating $\kappa$ times of infection outbreaks on each underlying individual-level influence relationship graph with randomly selected seed nodes in each simulation. In each infection outbreak, each infected node tries to activate its uninfected neighbors with probability $p$. In all the experiments, $\kappa$ is set to 50, the proportion of the seed nodes is set to 10%, and $p$ is set to 0.2.

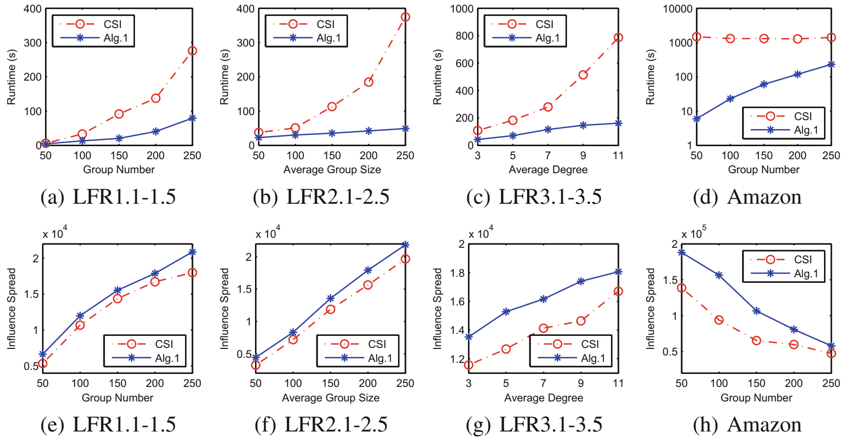
### 5.2 Performance Study of Influence Relationship Graph Construction

In this experiment, we carry out performance study on the construction of group-level influence relationship graph by comparing our proposed approach, i.e., Algorithm 1, with the existing algorithm known as CSI [11] in terms of (1) runtime for construction, and (2) effect to influence maximization. Since CSI requires the influence relationship between individuals to estimate group-level influence relationship, we give it a privilege that the underlying individual-level influence relationship graphs are available for CSI.

**Runtime for Graph Construction.** To evaluate the effects of (1) group number $m$, (2) average group size *avg-s*, and (3) the compactness of individual-level influence relationship (reflected by average degree *avg-d*) to the runtime for the construction of group-level influence relationship graph, we carry out runtime comparisons on graphs LFR1.1–1.5 that have varying $m$, graphs LFR2.1–2.5 that have varying *avg-s*, and graphs LFR3.1–3.5 that have varying *avg-d*, respectively. For the Amazon product co-purchasing network that has a fixed number of nodes and a fixed degree for each node, we only vary the number $m$ of groups returned by community detection from 50 to 250 with an interval of 50 (the corresponding *avg-s* will also vary with the varying $m$).

Figures 1(a)–(d) illustrate the runtime comparison result on each graph data set, from which we can have the following observations. (1) Our approach is significantly faster than CSI on the graph construction. This is because the time complexity of Algorithm 1 is $O(mn\kappa + m^2 + 2^{\alpha}m^2)$, which is linear to the number $n = m \times$ *avg-s* of nodes (individuals), while the time complexity of CSI is quadratic to $n$. (2) The gradients of runtime curves of Algorithm 1 in Figs. 1(a) and (d) are greater than that in Figs. 1(b) and (c), indicating that the number of groups is dominant affecting factor for the efficiency performance of Algorithm 1. (3) The compactness of individual-level influence relationship (i.e., *avg-d*) can also slightly affect the runtime of Algorithm 1. This is because with a higher average degree, each node is expected to spread its influence to more neighbors, and thus more groups may have significant influence relationship, resulting in that Algorithm 1 needs to execute more computation of conditional mutual information.

**Effect to Influence Maximization.** On the group-level influence relationship graphs constructed by Algorithm 1 and CSI, we perform our GLIMB algorithm with the same paraments. Figures 1(e)–(h) illustrate the corresponding influence spread scopes of the seed groups selected by GLIMB. From the figures, we can



(a) LFR1.1-1.5 (b) LFR2.1-2.5 (c) LFR3.1-3.5 (d) Amazon

(e) LFR1.1-1.5 (f) LFR2.1-2.5 (g) LFR3.1-3.5 (h) Amazon

**Fig. 1.** (a)–(d): Runtime for the construction of group-level influence relationship graph on different graph data sets. (e)–(h): The influence spread scopes of seed groups selected by GLIMB on the corresponding constructed group-level influence relationship graphs.

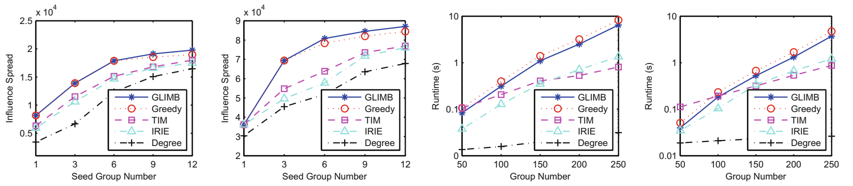
observe that our proposed Algorithm 1 can help our GLIMB algorithm to find better seed groups that bring larger influence spread scopes.

In brief, compared with CSI which learns the influence relationship between groups based on individuals, our proposed group-level approach can not only achieve a better efficiency performance, but also help improve the results of influence maximization.

### 5.3 Performance Study of GLIMB Algorithm

In this experiment, we verify the effectiveness and efficiency of our GLIMB algorithm for the problem of influence maximization with budget constraint. For the purpose of comparison, we modify the existing individual-level influence maximization approaches, including a state-of-the-art greedy searching method TIM [15] and two canonical heuristic searching methods DegreeDiscount [3] and IRIE [6], to search optimal seed groups with budget constraint by (1) considering the number of individuals in each group as a weight during their estimation of influence spread scope, and (2) adopting an `ImprovedGreedy`-like strategy. Moreover, the `ImprovedGreedy` approach [13] (denoted as Greedy in short) is also involved in the comparison.

**Comparison on Influence Spread Scope.** We compare our GLIMB algorithm with the other tested algorithms on all the LFR benchmark graph data sets listed in Table 2 and all the Amazon graph data sets used in Sect. 5.2, and record the influence spread scope when different number of seed groups are selected. Figures 2(a) and (b) illustrate the comparison results on LFR1.4 (in which $m = 200$)



(a) LFR1.4 (b) Amazon (c) LFR1.1-1.5 (d) Amazon

**Fig. 2.** (a)–(b): Influence spread scopes of the seed groups selected by different tested algorithms. (c)–(d): Scalability to number of groups.

and the Amazon graph data set containing 200 groups. From the figures, we can observe that (1) the seed groups selected by GLIMB and Greedy always have the significantly larger influence spread scopes than the other tested algorithms; (2) when the budget is sufficient to afford more seed group portfolios, the seed groups selected by GLIMB can bring a larger influence spread scope than the seed groups selected by Greedy. Similar observations can be observed on the rest of tested data sets.

**Scalability Study.** To investigate the scalability of GLIMB to the number $m$ of groups, in Figs. 2(c) and (d), we report the runtimes of GLIMB and the other tested algorithms on LFR1.1 to 1.5 and the Amazon graph data sets, in which the number $m$ of groups varies from 50 to 250 with an interval of 50. From the figures, we can have the following observations. (1) GLIMB is slightly more efficient than Greedy. This is due to the pruning method adopted in GLIMB, which reduces some redundant computation. (2) The gradient of IRIE's runtime curve is close to that of GLIMB's and Greedy's runtime curves, while IRIE's runtime is less than the runtimes of GLIMB and Greedy. This advantage in runtime of IRIE may be from a lower coefficient of the dominant item in its time complexity (3) The gradient of TIM's runtime curve is slightly smaller than that of GLIMB's and Greedy's runtime curves, since it has a $O(m \log m)$ time complexity. (4) DegreeDiscount is the most efficient, although its performance on influence maximization is often not comparable to GLIMB and Greedy.

In summary, with a compensation of more runtime, our GLIMB algorithm can make a better use of cost budget to achieve a larger influence spread scope, compared against the other tested approaches.

## 6 Conclusion

In this paper, we have studied the problem of group-level influence maximization with budget constraint. Towards this, we have proposed an efficient construction approach for group-level influence relationship graphs, introduced how to model influence propagation on the graph, and presented the GLIMB algorithm to search the optimal seed groups with at least a $(1 - 1/\sqrt{e})$-approximation. Experimental results on both synthetic and real-world data sets have demonstrated the efficacy of our approaches.

**Acknowledgements.** This work was supported in part by the NSFC Grants (61502347, 61522208, 61502504, and 61472359), and the Nature Science Foundation of Hubei Province (2016CFB384).

## References

1. Borgs, C., Brautbar, M., Chayes, J., Lucier, B.: Maximizing social influence in nearly optimal time. In: SODA, pp. 946–957 (2014)
2. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: KDD, pp. 1029–1038 (2010)
3. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: KDD, pp. 199–208 (2009)
4. Gomez-Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. ACM Trans. Knowl. Disc. Data **5**(4), 1019–1028 (2012)
5. Hu, Z., Yao, J., Cui, B., Xing, E.: Community level diffusion extraction. In: SIGMOD, pp. 1555–1569 (2015)
6. Jung, K., Heo, W., Chen, W.: IRIE: scalable and robust influence maximization in social networks. In: ICDM, pp. 918–923 (2012)
7. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD, pp. 137–146 (2003)
8. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
9. Leskovec, J., Adamic, L., Adamic, B.: The dynamics of viral marketing. ACM Trans. Web **1**(1), 5 (2007)
10. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: KDD, pp. 420–429 (2007)
11. Mehmood, Y., Barbieri, N., Bonchi, F., Ukkonen, A.: CSI: community-level social influence analysis. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8189, pp. 48–63. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40991-2_4
12. Myers, S., Leskovec, J.: On the convexity of latent social network inference. In: NIPS, pp. 1741–1749 (2010)
13. Nguyen, H., Zheng, R.: On budgeted influence maximization in social networks. IEEE J. Sel. Areas Commun. **31**(6), 1084–1094 (2013)
14. Shang, R., Luo, S., Li, Y., Jiao, L., Stolkin, R.: Large-scale community detection based on node membership grade and sub-communities integration. Phys. A Stat. Mech. Appl. **428**, 279–294 (2015)
15. Tang, Y., Xiao, X., Shi, Y.: Influence maximization: near-optimal time complexity meets practical efficiency. In: SIGMOD, pp. 75–86 (2014)
16. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-*k* influential nodes in mobile social networks. In: KDD, pp. 1039–1048 (2010)