SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2013, Vol. 56 078102:1–078102:14
doi: 10.1007/s11432-013-4895-5

Influence maximization with limit cost in social network

WANG Yue[1]*, HUANG WeiJing[2], ZONG Lang[2],
WANG TengJiao[2] & YANG DongQing[2]

[1]*Department of Computer Science, School of Information, Central University of Finance and Economics,*
Beijing 100081*, China;*
[2]*Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education,*
Beijing 100871*, China*

Received August 9, 2012; accepted October 30, 2012

Abstract Social networking service (SNS) applications are changing the way information spreads in online communities. As real social relationships are projected into SNS applications, word of mouth has been an important factor in the information spreading processes of those applications. By assuming each user needs a cost to accept some specific information, this paper studies the initial "seed user" selection strategy to maximize information spreading in a social network with a cost budget. The main contributions of this paper are: 1) proposing a graphic SEIR model (gSEIR) by extending the epidemic compartmental model to simulate the dynamic information spreading process between individuals in the social network; 2) proposing a formal definition for the influence maximization problem with limit cost (IMLC) in social networks, and proving that this problem can be transformed to the weighted set-cover problem (WSCP) and thus is NP-Complete; 3) providing four different greedy algorithms to solve the IMLC problem; 4) proposing a heuristic algorithm based on the method of Lagrange multipliers (HILR) for the same problem; 5) providing two parts of experiments to test the proposed models and algorithms in this paper. In the first part, we verify that gSEIR can generate similar macro-behavior as an SIR model for the information spreading process in an online community by combining the micro-behaviors of all the users in that community, and that gSEIR can also simulate the dynamic change process of the statuses of all the individuals in the corresponding social networks during the information spreading process. In the second part, by applying the simulation result from gSEIR as the prediction of information spreading in the given social network, we test the effectiveness and efficiency of all provided algorithms to solve the influence maximization problem with cost limit. The result show that the heuristic algorithm HILR is the best for the IMLC problem.

Keywords data mining, social network, influence maximization, graph-based diffusion model, information propagation

Citation Wang Y, Huang W J, Zong L, et al. Influence maximization with limit cost in social network. Sci China Inf Sci, 2013, 56: 078102(14), doi: 10.1007/s11432-013-4895-5

*Corresponding author (email: eecswangyue@pku.edu.cn)

1 Introduction

With the development of technology, more and more social networking services (SNS) applications are thriving. The unique feature that differentiates them from traditional services is the social network structure of the users. This type of social network-based application is becoming a new and powerful media: as these applications permit SNS users to directly publish information to friends in an online community and exchange ideas with them, every user in the SNS becomes a news reporter or diffuser. This causes the era of "All-Citizen Media" and results in many complex information spreading patterns in SNSs, and thus it is a great challenge to predict the range of influence for certain information in SNSs.

In the e-commerce or media industry, researching the process by which that information spreads between people is important for adverting and the promoting new products. The basic rule for commerce activities is maximizing profit while limiting costs. Therefore, determining how to maximize information spreading in a social network with a limit cost is critical for SNS companies. The initial step in propagating information to a community is to select a collection of people in this community as the seeds to start the information propagation process. As the social network structure may greatly affect the information spreading process, the seed selection is very important for the maximizing the influence of information in a social network. Therefore, the influence maximization problem of information with limit cost in social networks is a meaningful and practical optimization problem.

To study the solutions of the influence maximization with limit cost problem in social networks, the following two key steps are: 1) simulates the information propagation process between users in the target social network and uses the simulate results to predict the information spreading task; 2) determines the best seed set that has the maximum influence in the target social network based on the prediction results. Thus in this paper, we study models that simulate the information propagation process and provide several strategies to solve the IMLC problem. Our main contributions in this paper are: 1) proposing a graphic SEIR model (gSEIR) by extending the epidemic compartmental model (SIR) to simulate the information propagation process between individuals in the SNS networks. As gSEIR can also simulate the dynamic change process of the statuses for all the individuals in SNS networks during the information spreading process, it can conveniently be applied as the prediction result for information spreading; 2) defining the influence maximization with limit cost problem (IMLC) formally, reducing this problem to the weighted set cover problem, and proving that the IMLC problem is NP-complete; 3) providing algorithms with several greedy strategies (minimum cost first, maximum influence first, maximum influence cost ratio first, and maximum cost effective ratio first) for the IMLC problem; 4) modeling the IMLC problem as an integer programming problem, and proposing the heuristic algorithm HILR to gain the maximum information influence for limit cost in a social network; 5) providing sufficient experiments to verify that the gSEIR model can generate similar macro-behavior as the SIR model for information spreading process in social network by combining the micro behaviors of all users. We also test the effectiveness and efficiency of all four greedy algorithms and the heuristic algorithm for the IMLC problem, with the results showing that HILR performs the best of the algorithms provided in this paper.

The remainder of this research is organized as follows. Section 2 discusses recent related work on epidemic models, data mining technology and influence maximization in social network. Section 3 gives basic definitions of the gSEIR model for simulating the information propagation process in social networks and formalizes the influence maximization with limit cost problem (IMLC). Section 3 also proves that IMLC can be reduced to the weighted set cover problem, and thus this problem is NP-complete. Section 4 proposes four different types of greedy algorithms and a heuristic algorithm based on the method of Lagrange multiplier for solving the IMLC problem. Section 5 presents experiments to compare gSEIR with SIR on real social network data, and testes the effectiveness and efficiency of all the algorithms in the paper. Section 6 provides conclusions and suggests the future research directions.

2 Related works

Epidemic models: The process of information spreading among people is similar to how diseases through a population. English scholar Thomas Robert Malthus provided the famous Malthus grown model to simulate the grown process of population in a region [1]. This model explains an exponential growing population by assuming that the growth rate is constant. In [2], this feature is applied to an epidemiology study assuming that the disease infects the same number of people periodically. As this method does not distinguish between infected and uninfected people, it cannot explain why the number infected with one disease can reach a peak and why there is always a portion of the population not infected. Thus, researchers incorporated the compartmental models [3, 4] to divide population into subsets with different status (infected, susceptible). This type of models includes SI, SIS, SIR [5] and some extending models [6, 7]. As epidemic models can analyze the overall infection status (e.g. calculate the number of infected) in a population for infectious events, recent research has applied them to study information spreading status in online communities [8,9]. However the current epidemic models ignore inner relations between the individuals and just treat the population as a whole, which make them difficult to apply to precisely simulate the spread of information spreading in social networks.

Data mining in social networks: SNSs map the social relationships between people in online communities, and allow direct information propagation between users to become an important information spreading process online social networks. Thus data mining technologies for discovering knowledge in SNSs are important issues these days. Anagnostopoulos et al. [10] applied several statistical tests to prove the existence of influence in SNSs; in [11, 12], social influence in Flickr and Arnetminer were quantified; Crandall et al. [13] discussed the correlation between homophily and influence in SNSs; Agarwal et al. [14] applied influence to discover the opinion leaders in SNSs. To analyze the propagation process between individuals in social networks, information spreading models have been proposed to simulate the micro-behavior of individuals when they are exposed to specific information. Representative methods are the independent cascading model (IC) [15] and linear threshold model (LT) [16, 17]. Based on information spreading models, Kempe et al. [18] proposed the influence maximization problem in a social network. This is an optimization problem for finding a set of initial individuals in a social network that maximizes the spread of information. Wei Chen et al. [19] studied how to enhance the efficiency of algorithms for solving this problem so as to apply them to large-scale data mining tasks in practical social networks. Models such as IC and LT are motivated by the social-behavior patterns of individuals when faced with specific information, but ignore the fact that the status of individuals may change as the information spreads, An information spreading process based on these two models cannot exhibit observed properties such as that there is always a peak and life span to the spread of information in a real social network. Therefore we propose an information spreading model that considers both the status and the social relationships of individuals to further study the spread of information in social networks.

Set cover problem: We prove that the influence maximization with limit cost problem (IMLC) in a social network can be reduced to the weighted set cover problem [20] in chapter 3. According to the discussion in Karp [21], the weighted set cover problem is NP-complete. With current approaches to solving the problem using greedy or heuristic methods [22,23]. Cormode et al. [24] proposed algorithms for avoiding high disk random access and thus to improving efficiency. Following the same method, we propose some greedy algorithms with various strategies and a heuristic algorithm in an integer programming framework to solve the IMLC problem.

3 Problem definition

Recent methods for studying social network structure assume that the social network is a graph $G = \langle V, E\rangle$, where the vertex set V represents all the individuals in G and the edge set E represents the relationships between these individuals. For example, if G is the social network for an email community, then for any $v_1,\ v_2 \in V(v_1 \neq v_2)$, an edge $e = \langle v_1, v_2\rangle$ means that v_1 has sent an email to v_2.

• **Information and influence.** Using the terminology of the topic model [25], a topic is a multinomial distribution of words from a cluster of documents, and thus information can be denoted as a collection of messages which contain the words sampled from the corresponding topic. In an email social network, information about a specific topic is a collection of emails which contain the given key words. In this setting, for any two users v_1 and v_2 ($v_1, v_2 \in V,\ v_1 \neq v_2$), if the email content sent by v_2 contains keywords originating from v_1, we say that the information spreads from v_1 to v_2 or that v_1 influences v_2.

As described in the introduction, the task in this paper is to find efficient strategies to maximize the influence in social network G while limiting the cost. The cost represents the degree of difficulty with which people accept specific information, and can be measured through monetary or other rewards required. For example, we may apply the page rank score as the cost of promoting certain types of information because page rank is generally proportional to social influence in the corresponding social network graph. So that we do not lose generality, we assume that there is a function cost$(\cdot)$ for mapping the corresponding cost values to all users in a given social network. With the setting of cost, we make the following assumption about the spread of information in a social network. The rest of the paper follows this assumption.

Assumption 1. Let $G = \langle V, E\rangle$ is a social network, suppose 1) every individual v in V needs a cost to be influenced for a given information; 2) v will keep sending information to all his neighbors with content that he has been just influenced by others.

Under Assumption 1, a user who is infected by information from the same topic may keep sending this kind of information to his neighbors. Thus the spread of information in the social network can be described as: 1) propagation of information to initial individuals who are influenced at a given cost; 2) the influenced individuals in the social network will conform Assumption 1 send this information to all their neighbors; 3) some of the individuals receiving information will be influenced. Steps 2) and 3) repeat until no new individuals in the given social network are influenced.

With the basic notions above, the main contributions of this paper: proposing a graph-based information spreading model to try to predict the pattern of information spreading and attempting to apply various techniques to solve the influence maximization problem with limit cost in our models. For efficiency reasons, we generally treat the graph-based spread of information as happening prior to our influence maximization algorithms. Detailed definitions and analysis of the graph-based model and optimization problems are in the following

3.1 Graphic information spreading model

To simulate the spread of information in the social network $G = \langle V, E\rangle$, we extend the compartmental epidemic model to simulate information propagation behavior for individuals in the network. Basically, our model specifies four statuses—susceptible (S), exposed (E), infectious (I), and recovered (R)—to model the behavior of the individuals in G when faced with given information.

We now describe our graphic SEIR information spreading model (gSEIR). When information from specific topic m propagates into a social network G, any individual v in G may have one of four statuses: 1) susceptible means that v has not yet received information m; 2) exposed means that v has been exposed to m (or is a neighbor of an infected node), but v does not yet accept or influenced by this information; 3) infectious means that v has accepted (is influenced by) information m and promotes the propagation of m by recommending the content of m to his friends (neighbors in G); 4) recovered means that v has recovered from infectious status, and now even if v is exposed to new information from the same topic, it cannot be infected (or accept the content of the new information). Note that nodes in infectious status may recover at specified rate r.

We denote the probability that an exposed individual becomes infectious by p_i (the infected probability), and the probability that an infectious individual recover status by p_r (the recover probability: p_r is treated as constant for all individuals). According to Simon et al. [7], we suppose that a user may be affected by all his neighbors as to whether to accept given information. Therefore, following an independent Poisson process, the infected probability p_i can be defined as:

Definition 1. Infected probability. Let $G = \langle V, E\rangle$ be a social network. For $\forall v \in V$, $N(v)$ is the set for all neighbors of v, and $\mathrm{NI}(v)$ is the set for all infectious neighbors of v. For $\forall v_i, v_j \in V$, let w_{v_i,v_j} represents the frequency that j has been influenced by i in the past. Then the infected probability can

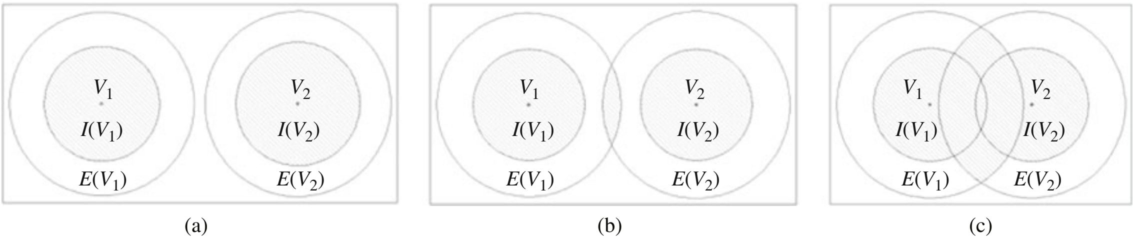

Figure 1 Possible results for $I(\{v_1,v_2\})$. (a) $E(\{v_1\})\bigcap E(\{v_2\})=\emptyset$, $I(\{v_1\})\bigcap I(\{v_2\})=\emptyset$; (b) $E(\{v_1\})\bigcap E(\{v_2\})\neq\emptyset$, $I(\{v_1\})\bigcap I(\{v_2\})=\emptyset$; (c) $E(\{v_1\})\bigcap E(\{v_2\})\neq\emptyset$, $I(\{v_1\})\bigcap I(\{v_2\})\neq\emptyset$.

be obtained from the following equation:

$$p_i = 1 - \exp\left[\delta t \frac{\sum_{v_i\in \mathrm{NI}(v)} w_{v,v_i}}{\sum_{v_j\in N(v)} w_{v,v_j}}\right]. \tag{1}$$

In Definition 1, δt is the accumulated time since the spread of information began. Thus, the information spreading process of the gSEIR model (gSEIR process) can be transformed to the following steps: 1) for given information m, select a set of individuals in G to be influenced by m, and treat them as the seeds individuals; 2) find the set of individuals in G who are exposed to m (in exposed status); 3) calculate the infected probability p_i for all individuals in the exposed status and change these individuals to infectious with probability p_i; 4) all infectious individuals recover with probability p_r; 5) repeat steps 2) and 3) until there are no infectious individuals.

3.2 Influence maximization with limit cost problem in social network

In any commercial activity, there is always a limit to the cost of popularizing products or advertising. Thus we study the influence maximization problem with limit cost in this paper. Usually, the cost for a specific user to accept an idea can be estimated by his reputation in the social network, so in this study we use the page rank value of an individual to evaluate the cost for him in our experiments. Before formalizing the influence maximization with limit cost problem, we provide some further notations.

- **Exposed set.** Let $G=\langle V,E\rangle$ be a social network, and m be the given information to be spread in G. Select the individuals in a subset V' ($V'\subseteq V$) as the seed nodes for information m. Suppose that all individuals in V follow the gSEIR model to propagate m. Then as time tends to infinity, the set of all the individuals that have been exposed to m is the exposed set of V', and it is denoted by $E(V')$. According to Definition 1, all individuals in $E(V')$ have a probability for being infected by m, so we denote the set of individuals in $E(V')$ that could be infected as $P[E(V')]$.
- **Maximum influence set.** According to the above setting, as time tends to infinity, call the final set of individuals that could be influenced by m the maximum influence set for V', and denote this set by $I(V')$. Because of the interaction between the individuals in G, the maximum influence set for V' ($|V'|>1$) may vary. We discuss the possible situations of the seed set using Figure 1.

Figure 1 shows the possible maximum influence sets for a subset V', where $V'=\{v_1,v_2\}$ ($v_1\neq v_2$).

1) As shown in Figure 1(a), when $E(\{v_1\})\bigcap E(\{v_2\})=\emptyset$ and $I(\{v_1\})\bigcap I(\{v_2\})=\emptyset$, all the individuals in exposed set $E(\{v_1,v_2\})$ have no chance of being infected, and thus the maximum influence set for V' is

$$I(V') = I(v_1)\bigcup I(v_2).$$

2) As shown in Figure 1(b), when $E(\{v_1\})\bigcap E(\{v_2\})\neq\emptyset$ and $I(\{v_1\})\bigcap I(\{v_2\})=\emptyset$, some individuals in the intersection of the exposed sets $E(\{v_1\})\bigcap E(\{v_2\})$ have a chance of being infected, and thus the maximum influence set for V' is

$$I(V') = I(v_1)\bigcup I(v_2)\bigcup P\Big[(E(v_1)-I(v_1))\bigcap(E(v_2)-I(v_2))\Big].$$

3) As shown in Figure 1(c), when $E(\{v_1\})\bigcap E(\{v_2\})\neq\emptyset$ and $I(\{v_1\})\bigcap I(\{v_2\})\neq\emptyset$, some individuals in the intersection of the exposed sets $E(\{v_1\})\bigcap E(\{v_2\})$ may also have a chance of being infected, and thus the maximum influence set for V' is

$$I(V')=I(v_1)\bigcup I(v_2)\bigcup P\left[(E(v_1)-I(v_1))\bigcap(E(v_2)-I(v_2))\right].$$

In the three situation above, for $\forall v_1,v_2\in V$, the maximum influence set $I(V')$ for $V'=\{v_1,v_2\}$ can be calculated using (2).

$$I(V')=I(v_1)\bigcup I(v_2)\bigcup P\Big[(E(v_1)-I(v_1))\bigcap(E(v_2)-I(v_2))\Big]. \tag{2}$$

We give the formal definition of the influence maximum problems in the following sections, and show that (2) is sufficient to calculate the results for all possible maximum influence set.

Problem 1 (Influence Maximization with Minimum Seeds in Social Network (IMMS))**.** Let $G=\langle V,E,W\rangle$ be a social network, where the vertex set V represents the individuals in G, the edge set E represents the direct relationships between individuals, and W is the closeness between pairs of individuals where for any $w(v_1,v_2)\in W$, $w(v_1,v_2)$ represents the frequency of interaction between v_1 and v_2. If m is the information to be spread in G, then IMMS can be described as follows: find the smallest subset $V_{\rm opt}$ of V such that $I(V_{\rm opt})=V$ or $|I(V_{\rm opt})|=r\%\times|V|$, where r is a target proportion of the population for this task.

Theorem 1. IMMS problem is equivalent to the set cover problem.

Proof. According to (2), IMMS can be converted into the following forms:

$$\bigcup_{v_i\in V_{\rm opt}} I(v_i)\bigcup P\left[\bigcup_{\forall v_a,v_b\in V_{\rm opt}} E(v_a)\bigcap E(v_b)\right]=V, \tag{3}$$

$$\left|\bigcup_{v_i\in V_{\rm opt}} I(v_i)\bigcup P\left[\bigcup_{\forall v_a,v_b\in V_{\rm opt}} E(v_a)\bigcap E(v_b)\right]\right|=r\%\times|V|. \tag{4}$$

Eq. (3) can be rewritten in the following form:

$$\bigcup_{\forall v_a,v_b\in V_{\rm opt}} I(v_a)\bigcup I(v_b)\bigcup P\Big[E(v_a)\bigcap E(v_b)\Big]=V. \tag{5}$$

If we construct a one-to-one map from a family of vertex sets s_i to all results of $I(v_a)\bigcup I(v_b)\bigcup P[E(v_a)\bigcap E(v_b)]$, then IMMS can be restated as finding a subset $S=\{s_1,s_2,s_3,\ldots\}$ of V with minimum cardinality such that $\bigcup_{s_i\in S}s_i$ contains all the individuals in V. Thus, IMMS is a set cover problem, and according to Karp [21], is NP-hard.

By following the descriptions of cost, we have a cost function $\mathrm{cost}(v)$ denoting the cost for $\forall v\in V$. With this setting, the main problem of IMLC can be described as follow.

Problem 2 (Influence Maximization with Limit Cost in Social Network (IMLC))**.** With the social network settings in Problem 1, given information m and a budget B for the information propagation task, find a subset $V'_{\rm opt}$ for which $I(V_{\rm opt})=V$ or $|I(V_{\rm opt})|=r\%\times|V|$, where r is a target proportion of the population and where $\sum_{v\in V_{\rm opt}}\mathrm{cost}(v)\leqslant B$.

As described in Problem 2, IMMS is a special case of IMLC with a cost function for each individual in the social network. We can prove in a similar manner that IMLC is equivalent to the weighted set cover problem (WSCP), which according to Karp [21], is NP-complete, and thus Problem 2 is also NP-complete.

```
Algorithm 1: greedy_MCF
Data: $G = \langle V, E\rangle$, influence rate threshold $\theta$, cost threshold $B$;
Result: $V_{\text{opt}}$, practical influence rate $r_c$
begin
    $C \leftarrow \emptyset$, $V_{\text{opt}} \leftarrow \emptyset$
    while $|C| \leqslant \theta \times |V|$ do
        if $\sum_{\forall v \in \text{seeds}} \text{cost}(v) \geqslant B$ break
        $v^* \leftarrow \arg\min_{v^* \in V} |\text{cost}(v^*)|$
        $V_{\text{opt}} \leftarrow V_{\text{opt}} \bigcup \{v^*\}$
        $C \leftarrow C \bigcup I(v^*)$
        Output $V_{\text{opt}}$
        Output practical influence ratio $r_c = |C|/|V|$
    end
end
```

```
Algorithm 2: greedy_MIF
Data: $G = \langle V, E\rangle$, influence rate threshold $\theta$, cost threshold $B$;
Result: $V_{\text{opt}}$, practical influence rate $r_c$
begin
    $C \leftarrow \emptyset$, $V_{\text{opt}} \leftarrow \emptyset$
    while $|C| \leqslant \theta \times |V|$ do
        if $\sum_{\forall v \in \text{seeds}} \text{cost}(v) \geqslant B$ break
        $v^* \leftarrow \arg\max_{v^* \in V} |I(\{v^*\})|$
        $V_{\text{opt}} \leftarrow V_{\text{opt}} \bigcup \{v^*\}$
        $C \leftarrow C \bigcup I(v^*)$
        Output $V_{\text{opt}}$
        Output practical influence ratio $r_c = |C|/|V|$
    end
end
```

```
Algorithm 3: greedy_MIC
Data: $G = \langle V, E\rangle$, influence rate threshold $\theta$, cost threshold $B$;
Result: $V_{\text{opt}}$, practical influence rate $r_c$
begin
    $C \leftarrow \emptyset$, $V_{\text{opt}} \leftarrow \emptyset$
    while $|C| \leqslant \theta \times |V|$ do
        if $\sum_{\forall v \in \text{seeds}} \text{cost}(v) \geqslant B$ break
        $v^* \leftarrow \arg\max_{v^* \in V} \frac{|I(\{v^*\})|}{\text{cost}(v^*)}$
        $V_{\text{opt}} \leftarrow V_{\text{opt}} \bigcup \{v^*\}$
        $C \leftarrow C \bigcup I(v^*)$
        Output $V_{\text{opt}}$
        Output practical influence ratio $r_c = |C|/|V|$
    end
end
```

4 Algorithms for IMLC problem

4.1 Greedy algorithms for IMLC

Naive greedy algorithms: As IMLC is NP-complete, we present three naive greedy algorithms for solving the problem. Given a social network $G = \langle V, E, W\rangle$ and a cost budget B, the three strategies of these greedy algorithms are as follows. 1) Minimum cost first (MCF), by iteratively selecting the individual with the minimum cost and adding it to the seed set while the sum of the cost for the seeds is less than B. 2) Maximum influence first(MIF), by iteratively selecting the individual with the maximum influence as a seed while $\sum_{\forall v \in \text{seeds}} \text{cost}(v) \geqslant B$. MIF extends the greedy algorithm in [19] by considering the cost limit. We implement MIF for the purpose of comparison. 3) Maximum influence-cost ratio first (MIC), by iteratively selecting the individual with the maximum influence-cost ratio and adding it to the seed set while $\sum_{\forall v \in \text{seeds}} \text{cost}(v) \geqslant B$. The pseudo-code for each greedy algorithm is listed as Algorithms 1–3

Algorithm 4: greedy_MCE

```
Data: $G = \langle V, E\rangle$, influence rate threshold $\theta$, cost threshold $B$;
Result: $V_{\rm opt}$, practical influence rate $r_c$
begin
    $C \leftarrow \emptyset$, $V_{\rm opt} \leftarrow \emptyset$
    while $|C| \leqslant \theta \times |V|$ do
        if $\sum_{\forall v \in {\rm seeds}} {\rm cost}(v) \geqslant B$ break
        $v^* \leftarrow \arg\max_{v^* \in V} \frac{{\rm cost}(v^*)}{|I(v^*) - C|}$
        $V_{\rm opt} \leftarrow V_{\rm opt} \bigcup \{v^*\}$
        $C \leftarrow C \bigcup I(v^*)$
        Output $V_{\rm opt}$
        Output practical influence ratio $r_c = |C|/|V|$
    end
end
```

respectively.

These greedy Algorithms 1, 2 and 3 can all obtain feasible solutions to IMLC. In the experiments, we found that the cardinality of the maximum influence set of the seed sets obtained by greedy_MIC is the largest. As these naive greedy algorithms lack theoretical support, it is hard to discuss the relationship between these feasible solutions and the globally optimal solution. Therefore, we provide a greedy algorithm which is an H_n ($H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$) factor approximation algorithm for IMLC.

As we have proved that IMLC can be reduced to the weighted set cover problem, the greedy algorithm for the weighted set cover problem can be applied to IMLC. Suppose that OPT is the optimal solution. It was proved in [20] that the solution found by the greedy algorithm with the maximum cost effective ratio strategy is at most $(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}) \cdot {\rm OPT}$. The idea is to iteratively add the most cost-effective individual into the seed set until all the individuals of V are covered. In IMLC, the cost-effectiveness α can be determined by the following equation:

$$\alpha = \frac{{\rm cost}(v^*)}{|I\{v^*\} - C|}. \tag{6}$$

In (6), $I\{v^*\}$ is the maximum influence set for the current individual, C is the individuals in V that are already covered, and ${\rm cost}(v^*)$ is the cost of choosing v^* as a seed. With this measure, we propose a greedy algorithm using that "maximum cost-effectiveness (MCE)" strategy. The pseudo-code is listed in Algorithm 4.

We apply the Monte Carlo method to obtain the maximum influence set for the individuals. During the running of the greedy algorithms, obtain the maximum influence set using the Monte Carlo method at each step is inefficient. Therefore, we preprocess the gSEIR process for every individual in the social network, and store the corresponding maximum influence set in a database. This pre-processing, we can reduce the time for the greedy algorithms to obtain the solution to IMLC.

4.2 Integer programming framework for IMLC

As proved in Section 3, IMLC can be reduced to the weighted set cover problem. According to [26], IMLC can be converted into following integer programming forms:

$$\begin{aligned} &Z = \min\,(c_1x_1 + c_2x_2 + c_3x_3 + \cdots + c_nx_n), \\ &\text{s.t.} \ \sum_{j=1}^{n} a_{ij}x_j \geqslant 1, \quad i = 1, 2, \ldots, m, \quad x_j = \{0, 1\}, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{7}$$

In (7), $a_{ij} \in \{0, 1\}$ with $a_{ij} = 1$ representing that the maximum influence set including the jth individual as a seed can influence the ith individual and c_i ($i = 1, 2, \ldots, n$) represents the value of the cost function for each individual in the social network. In the experiments, we use page rank to estimate the cost for these individuals in the social network.

Table 1 A summary of data sets

Source	Node number	Edge number	Ave. degree	Topics
Enron	86,857	362,483	4.17	146,649
Hartford	212	284	1.34	196

In the integer programming form of IMLC, we can apply Lagrange relaxation methods [26] to simplify the problem. By setting a Lagrange multiplier for each constraint in (7), IMLC can be converted into

$$\begin{aligned} &Z=\min\left(\sum_{j=1}^{n}c_jx_j+\sum_{i=1}^{m}\lambda_i\left(1-\sum_{j=1}^{n}a_{ij}x_j\right)\right),\\ &\text{s.t.}\ \ x_j=\{0,1\},\ \ j=1,2,\ldots,n,\ \ \lambda_i>0,\ \ i=1,2,\ldots,m. \end{aligned}\tag{8}$$

Let $d_j=c_j-\sum_{i=1}^{n}\lambda_ia_{ij}$, then (8) can be converted to

$$\begin{aligned} &Z=\min\left(\sum_{j=1}^{n}d_jx_j+\sum_{i=1}^{m}\lambda_i\right),\\ &\text{s.t.}\ \ x_j=\{0,1\},\ \ j=1,2,\ldots,n,\ \ \lambda_i>0,\ \ i=1,2,\ldots,m. \end{aligned}\tag{9}$$

Thus the solution to IMLC becomes finding the seed set for obtaining the minimum value in (9). By this setting, when $d_j\leqslant 0$, $x_j=1$, Z will obtain the minimum value, or when $d_j>0$, $x_j=0$, Z will obtain the minimum value.

4.3 A Lagrange relaxation heuristic algorithm

With the former descriptions, we proposed a Lagrange relaxation-based heuristic algorithm for IMLC. Furthermore, by considering that an increasing number of infected neighbors for an individual will increase the infected probability, we use a matrix $P=\{p_{ij}\}$ to store the exposed individuals (the neighbors of the infected individuals), where p_{ij} represents that the jth individual is exposed when the ith individual is chosen as a seed. Thus d_j can be written in the following form:

$$d'_j=c_j-\sum_{i=1}^{m}\lambda_ia_{ij}-w\left(1-\exp\left[\delta t\frac{\sum_{i=1}^{m}p_{ij}}{\sum_{i=1}^{m}N_{ij}}\right]\right).\tag{10}$$

In (10), δt is the iterative time, N_{ij} represents the neighborhood relation of i and j where $N_{ij}=1$ indicates that i and j are neighbors and w is a parameter adjusting the number of infected neighbors for an exposed individual. With these settings, we present the pseudo-code for this Lagrange relaxation-based algorithm as Algorithm 5.

5 Experiments and discusses

5.1 Data sets and experiments configuration

We test our algorithms on the Hartford data from a real social network [27] and the Enron email data[1]. The details of the data used are listed in Table 1, where "topics" indicates the number of topics for the information tracked in the dataset.

The experiments in this section include: 1) experiments to compare the gSEIR process on the social network data with the SIR epidemic model; 2) experiments to compare the effectiveness and efficiency of all the algorithms for IMLC.

All the experiments are performed on a PC with a 2.93 GHz CPU and 2 GB RAM. The prototype is implemented in Python, and the results for the maximum influence set are stored in MySQL.

[1] http://www.cs.cmu.edu/∼enron/.

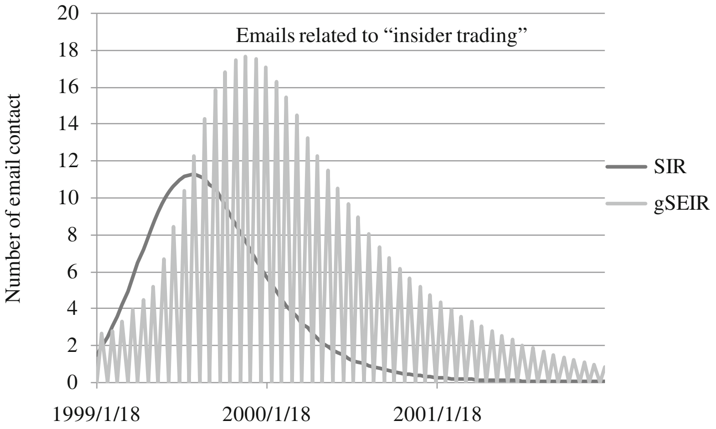

Figure 2 Simulated results of gSEIR & SIR.

```
Algorithm 5: HILR(LR)
Data: $G = \langle V, E\rangle$, influence rate threshold $\theta$, cost threshold $B$;
Result: $V_{\rm opt}$, practical influence rate $r_c$
begin
    Calculate cost($v$) for $\forall v \in V$
    $c_i \leftarrow$ cost($v_i$) for $\forall v_i \in V$
    Preprocess the gSEIR process, obtained matrix $A = \{a_{ij}\}$ and $P = \{p_{ij}\}$
    Initialized Lagrange multiplier $\lambda_i$ for all nodes $v$ ($\forall v \in V$) $C \leftarrow V$
    while $|C| \leqslant \theta \times |V|$ do
        foreach $v_j$ in $C$ do
            $d'_j = c_j - \sum_{i=1}^m \lambda_i a_{ij} - w \left(1 - \exp\left[\delta t \frac{\sum_{i=1}^m p_{ij}}{\sum_{i=1}^m N_{ij}}\right]\right)$
            if $d'_j \leqslant 0$ then
                $V_{\rm opt} = V_{\rm opt} \bigcup \{v_j\}$ and $C = C - V_j$
            end
            if $\sum_{\forall v \in {\rm seeds}} {\rm cost}(v) \geqslant B$ break
            Update all Lagrange multipliers
        end
    end
    Output $V_{\rm opt}$
    Output practical influence ratio $r_c = |C|/|V|$
end
```

5.2 Comparison of gSEIR and SIR

In this section, we compared the information spreading process with gSEIR and the SIR epidemic model on the Enron email data. We apply a binomial distribution to simulate the accumulated process with the SIR model, and set the parameter for the SIR model using the following method. Depending on the different topics of the emails, we choose those topics with interaction frequencies greater than 150 to track. For each topic, set the population number N as the number of involved email users. By obtaining the related email contact number R for the topics, set the propagation ratio $\beta = R/N$. We set the parameters for gSEIR as follows. Generated a graph to store the related social network $G = \langle V, E\rangle$, adding all the involved email users to the vertex set V. Obtain the email contact relations on the whole dataset to obtain the relations between all users in V, and add these relations to the edge set E. Select the earliest email senders of each topic as the seed nodes for the information spreading process, and then start the gSEIR process.

Figure 2 compares the information spreading processes simulated by gSEIR and SIR. In the early years of this century, the famous insider trading case of Enron made inestimable losses and thus directly led to the bankruptcy of Enron. Thus, we analyzed the information spreading process of the topic with the keyword “insider trading” in the Enron email dataset. According to our prior description, we set the

Table 2 Peak value comparison for topics

Topic keywords	gSEIR	SIR	Real
Insider trading	17.66	11.31	19
Bankruptcy	111.40	7.79	105
Kenneth Lay became CEO	47.05	57.17	42

Table 3 Average practical costs for algorithms

Algorithm	MIF	MIC	MCF	MCE	LR
Average practical cost	0.0079	0.0091	0.0087	0.0089	0.0085

parameters for SIR as: $N = 97$, $\beta = 0.38$, $\mu = 0.1$. For gSEIR, the recover rate is $\mu = 0.1$, and the related subgraph for this topic includes 97 nodes and 173 edges. We run the SIR and gSEIR model 100 times each, and graph the average simulated numbers of email with the keyword "insider trading" in Figure 2. As it is shown in Figure 2, the simulated distribution of the gSEIR model is similar to the result for the SIR model. Thus, the overall information spreading process for the gSEIR model is similar to that for the traditional epidemic model.

To further compare the simulation of gSEIR and SIR, we analyze the information spreading processes of three real topics in the Enron email data using these two models. We estimate the parameters for gSEIR and SIR by making their processes last for the same life span as the real process, and compared the peak value for each models with the real value in Table 2. As shown in Table 2, gSEIR is more precise than SIR in predicting the peak value of email contact.

As the gSEIR model performs similarly to the SIR model, and gSEIR is more precise than SIR in predicting peak the email contact, we apply gSEIR in the remaining experiments to predict the maximum influence set and the exposed set of the seed nodes.

5.3 Comparison of algorithms about IMLC

• **Effectiveness experiments.** We test the effectiveness of all the proposed algorithms for IMLC, with the results shown in Figures 3–6. Figure 3 shows the practical cost of the results found by all the algorithms on the Hartford data; Figure 4 shows the influence ratio results for the nodes found by all the algorithms on the Hartford data; Figures 5 and 6 show the infection-cost ratio for results found by all the algorithm on the Hartford data and the Enron data respectively.

We note from Figure 3 that practical costs are often lower that the thresholds, because the sum of costs for all users may not always accurately equal the threshold. Thus the practical cost is an approximation to the cost threshold. The closeness of the practical cost and the cost threshold may be interpreted as the cost efficiency of a specific algorithm. As shown in Figure 3, with the increasing of cost limit, each algorithm adaptively adjusts its practical cost. The average practical cost for each algorithm is listed in Table 3. Table 3 shows that MIF has the lowest cost efficiency, and that MIC is the most cost efficient as it has the highest budget and practical cost ratio. This reflects the maximum influence-cost ratio strategy of MIC.

To compare the effectiveness of the algorithms for IMLC, we compared the simulated results for the maximum influence sets for the seeds found by each algorithm. This experiment has the following steps: 1) apply each algorithm to the Hartford data set and obtained a set of seeds; 2) performed the gSEIR process 100 times on the seed sets obtained to simulate information spreading, and use the spreading result as the prediction of the maximum influence set for each seed set; 3) obtain the influence ratio γ (denote γ as the ratio of the maximum influence number and the number of users in the whole social network) for each algorithm. As the task is to find a feasible seed set with the maximum influence in the social network, the algorithm with the largest influence ratio may be the most effective for solving IMLC.

As shown in Figure 4, algorithm HILR (LR) has the maximum influence ratio in this experiment, and algorithm greedy_MCE and greedy_MCE also have high influence ratio for their cost-effective strategy.

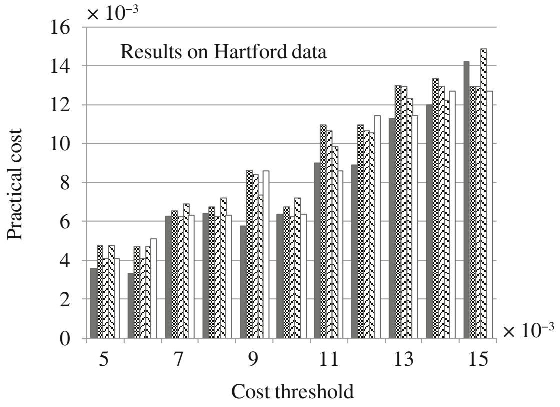

Figure 3 Practical cost for algorithms.

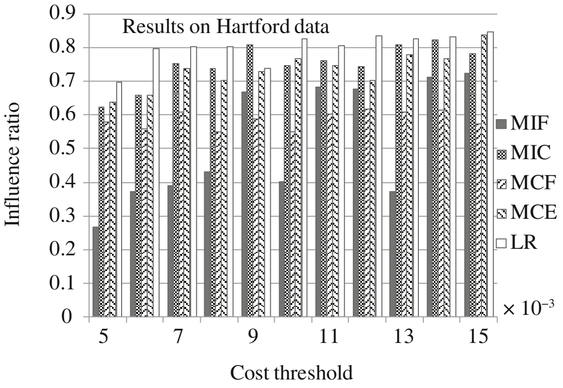

Figure 4 Influence ratio comparison for algorithms.

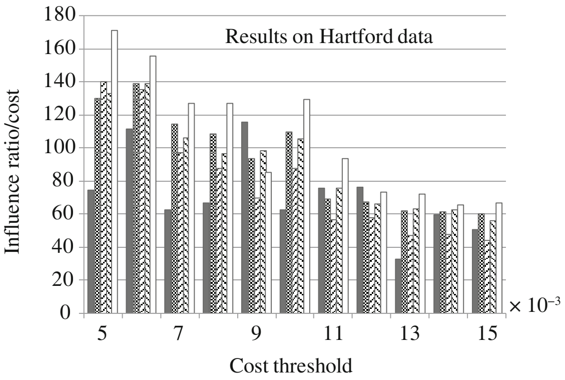

Figure 5 Infected-cost ratio for Algorithm 1.

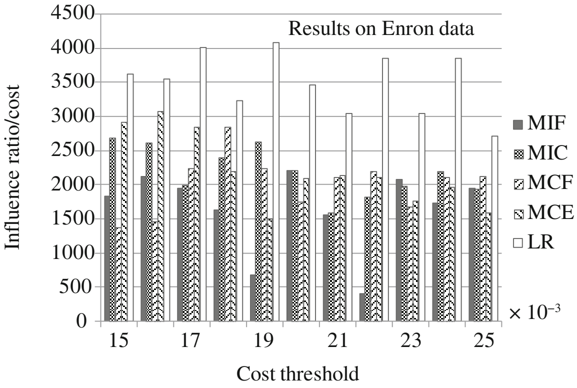

Figure 6 Infected-cost ratio for Algorithm 2.

As the strategy of algorithm greedy_MIF only considered to choose the nodes with biggest influence, the final influence ratio of it is the lowest.

To compare the price-performance ratio for all algorithms in this paper, we calculated the ratio of the maximum influence set and the sum of the cost for each seed set of the algorithms respectively, and the results are shown in Figures 5 and 6. Both results in Figures 5 and 6 show HILR (LR) may be the best option to solve the influence maximization with limit cost problem.

5.4 Scalability

We tested the scalability of our algorithms for IMLC on the entire Enron email data set. The results are shown in Table 4, where time is measured in milliseconds. The results in Table 4 do not include the time costs for pre-processing in each algorithm. As shown in Table 4, as the number of edges increases, the greedy algorithms are all more efficient than the Lagrange relaxation based heuristic algorithm HILR. Therefore, considering the effectiveness results from the last section, the greedy algorithms may still be an good option for solving IMLC.

6 Conclusions

The influence maximization problem with limit cost in social networks (IMLC) is a common problem in the e-commerce field, where it is useful for finding the best advertising partners or agents in SNS applications. This problem may also have many other applications in the media industry. We have proposed a formal definition of IMLC, and have proved that it is NP-complete. By applying several greedy strategies and integer programming methods, we have provided several algorithms to solve this

Table 4 Time efficiency for algorithms (ms)

Edge No.	MIF	MIC	MCF	MCE	LR
100	11445	12447	72310	14231	143210
200	608552	682270	431477	59800	174845
500	895977	869365	689895	698340	568392
5000	696957	669949	796597	846215	325415
10000	808489	120818	543086	299167	395628
20000	566508	489866	713102	632737	1295628

problem. To predict the maximum influence set of seed users in the social network, we propose a graph based SEIR model, gSEIR, to simulate the information spreading process in the social network. Our experiments show that gSEIR can simulate both the macro- and micro-behaviors of the individuals in the social network. We use the gSEIR model as a pre-processing step to solve IMLC, and thus the results from our algorithms for IMLC can also be treated as predictions of the possible seed selections for maximizing the influence in a given social network.

Our future work will focus on improving the efficiency of our algorithms to make them suitable for IMLC in large-scale social networks, and trying to apply the methods in this paper in the practical applications.

Acknowledgements

This work was supported by National Key-tech R&D Program of China (Grant No. 2009BAK63B08), National High-tech R&D Program of China (863 Program) (Grant No. 2009AA01Z150), Graduate Student Education Innovation Project of Central University of Finance and Economics (CUFE), National Nature Science Foundation of China (Grant Nos. 60970143, 61272398), and Beijing-funded Joint Development Program of the Central Universities in Beijing.

References

1 Oded G, Weil D N. Population, technology and growth: from malthusian stagnation to the demographic transition and beyond. Amer Econ Rev, 2000, 90: 806–828

2 Murray J D. Mathematical Biology. 2nd ed. Berlin: Springer-Verlag, 1993

3 Gibson G J, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain Monte Carlo methods. IMA J Math Appl Med Biol, 1998, 15: 19–40

4 Khelil A, Becker C, Tian J. An epidemic model for information diffusion in MANETs. In: Proceedings of the 5th ACM International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems. New York: ACM, 2002. 54–60

5 Jacquez J A, Simon C P. The stochastic SI model with recruitment and deaths I, comparison with the closed SIS model. Math Biosci, 1993, 117: 77–125

6 Lekone P E, Finkenstadt B F. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics, 2006, 62: 1170–1177

7 Simon P L, Taylor M, Kiss I Z. Exact epidemic models on graphs using graph-automorphism driven lumping. J Math Biol, 2011, 62: 479–508

8 Adar E, Adamic L A. Tracking information epidemics in blogspace. In: Conference on Web Intelligence, Compiegne, 2005. 207–214

9 Saito K, Kimura M, Motoda H. Discovering influential nodes for SIS models in social networks. In: Proceedings of the 12th International Conference on Discovery Science. Berlin/Heidelberg: Springer-Verlag, 2009. 302–316

10 Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 7–15

11 Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010. 241–250

12 Tang J, Sun J M, Wang C, et al. Social influence analysis in large-scale networks. In: Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 807–816

13 Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 160–168

14 Agarwal N, Liu H, Tang L, et al. Identifying the influential bloggers in a community. In: Proceedings of the 1st ACM International Conference on Web Search and Data Mining. New York: ACM, 2008. 207–217

15 Saito k, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model. In: Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Berlin/Heidelberg: Springer-Verlag, 2008. 67–75

16 Kleinberg J. Cascading behavior in networks: algorithmic and economic issues. In: Nisan N, Roughgarden T, Tardos E, et al., eds. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007

17 Saito K, Kimura M, Ohara K, et al. Selecting information diffusion models over social networks for behavioral analysis. In: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin/Heidelberg: Springer-Verlag, 2010. 180–195

18 Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003. 137–146

19 Chen W, Wang Y J, Yang S Y. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Paris, 2009. 199–208

20 Vazirani V V. Approximate Algorithms. Springer, 2001

21 Karp R M. Reducibility among combinatorial problems. In: Miller R E, Thatcher J W, eds. Complexity of Computer Computations. New York: Plenum Press, 1972. 85–103

22 Caprara A, Fischetti M, Toth P. Algorithms for the set covering problem. Ann Oper Res, 1998, 92: 353–371

23 Gao B J, Ester M, Cai J Y, et al. The minimum consistent subset cover problem and its applications in data mining. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007. 310–319

24 Cormode G, Karloff H, Wirth A. Set cover algorithms for very large datasets. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010. 479–488

25 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022

26 Fisher M L. The Lagrangian relaxation method for solving integer programming problems. Manage Sci, 1981, 27: 1–18

27 Weeks M R, Clair S, Borgatti S P, et al. Social networks of drug users in high-risk sites: finding the connections. AIDS Behav, 2002, 6: 193–206