Balanced Seed Selection for Budgeted Influence Maximization in Social Networks

Shuo Han[1,2], Fuzhen Zhuang[1], Qing He[1], and Zhongzhi Shi[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{hans,zhuangfz,heq,shizz}@ics.ict.ac.cn

Abstract. Given a budget and a network where different nodes have different costs to be selected, the budgeted influence maximization is to select seeds on budget so that the number of final influenced nodes can be maximized. In this paper, we propose three strategies to solve this problem. First, Billboard strategy chooses the most influential nodes as the seeds. Second, Handbill strategy chooses the most cost-effective nodes as the seeds. Finally, Combination strategy chooses the "best" seeds from two "better" seed sets obtained from the former two strategies. Experiments show that Billboard strategy and Handbill strategy can obtain good solution efficiently. Combination strategy is the best algorithm or matches the best algorithm in terms of both accuracy and efficiency, and it is more balanced than the state-of-the-art algorithms.

Keywords: Budgeted Influence Maximization, Information Propagation, Social Networks.

1 Introduction

Influence maximization is a hot topic for viral marketing and has been heavily studied in the previous literature [5,3,4]. The traditional problem statement is to find a k-node set of seeds that propagate influence so that the number of resulting influenced nodes can be maximized. This definition is proposed by Kempe at al. [5] and implies an assumption that each node has an uniform cost to be chosen. Following his work, most of the existing works comply with the same assumption and focus on the k-node influence maximization problem [11,3,4]. However, this assumption does not accord with most real-world scenarios. For example, in the domain of online advertising service, different web sites have different advertising prices. If a company promotes its product by online advertisement, how to choose the web sites on budget? Spend much money on some few famous portal sites or choose less popular web sites to add the number of advertisements? Obviously, this problem is different from k-node influence maximization.

The problem statement of influence maximization can be extended to a generalized form. Given a social network where nodes may have different costs to be selected, an influence diffusion model and a budget, influence maximization is to find a seed set within the budget that maximizes the number of final influenced nodes. Nguyen et al. [10] call this problem budgeted influence maximization(BIM). In mathematics, k-node influence maximization and budgeted influence maximization have two different mathematical abstractions [8,6]. And it has been proved that the algorithms for the

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 65–77, 2014.

former will no longer produce satisfactory solutions for the latter [6]. Thus, most of the existing approaches for k-node influence maximization do not suit for the BIM problem.

The other challenge to influence maximization is the efficiency. A common method to this problem is Greedy algorithm [5], which does well in accuracy but has a bad performance in efficiency. Some researchers adopt different ideas to address this problem efficiently [3,2,4]. However, either they improve the efficiency at the cost of effectiveness, or the proposed approach can only address k-node influence maximization problem. To our knowledge, there are only two existing studies focusing on budgeted influence maximization [7,10], and both of them are based on Greedy algorithm. Although these studies concentrate on improving Greedy algorithm to reduce the runtime, they still can not overcome its intrinsic flaw of expensive computation.

In this paper, we tackle the problem of budgeted influence maximization and aim to propose an approach that has good performance in both accuracy and efficiency. We first analyze real networks empirically, including defining node roles and studying seed selection heuristics, which is the foundation of the proposed algorithm. Then, we advance three strategies to address this problem. The first one is Billboard strategy that chooses the most influential nodes as the seeds. The second one is Handbill strategy that chooses the most cost-effectvie nodes as the seeds. And the third one, called Combination strategy, uses Simulated Annealing to choose the best combination of nodes from the two nonoverlapping sets resulting from the previous two strategies. Experiments show that Billboard strategy and Handbill strategy can solve this problem efficiently and obtain good accuracies. Combination strategy is the best algorithm or matches the best algorithm in terms of both accuracy and efficiency, and is more balanced than the state-of-the-art algorithms.

2 Problem Statement and Preliminaries

2.1 Problem Statement

A social network can be modeled as a graph $G = (V, E)$, where vertices v represent individuals and edges E represent the relationship between two individuals, each vertice $v \in V$ has a cost $c(v)$ denoting the expense when it is selected. A diffusion model describes the spread of an information through the social network G. In this paper, we adopt Independent Cascade(IC) model for simulating influence propagation, which is a classical information propagation model and is widely used in the previous influence maximization research [5,7,4]. Given a network G, a diffusion model and a budget B, budgeted influence maximization is to find a seed set $S \in V$ such that subjecting to the budget constraint $\sum_{v \in S} c(v) \leq B$, the number of final influenced individuals $\sigma(S)$ can be maximized. The important notations used in this paper are declared in Table 1.

Table 1. Important Notations

Notation	Description
$G = (V, E)$	A social network with vertex set V and edge set E
B	The total budget
$N(v)$	The out-neighbors of node v
$c(v)$	The cost of node v
$ce(v)$	The cost-efficient value of node v to influence propagate
$\sigma(S)$	The final influence of seed set S

2.2 Billboards and Handbills

There are two natural ideas commonly used to select seeds for the BIM problem in the real-life scenario. In this paper, we call them Billboard strategy and Handbill strategy. Billboard strategy is to choose the most influential nodes as seeds in the network. We call these nodes billboards, because they are like billboard advertisements that are located in high traffic area and can create the most impactful visibility to people. Handbill strategy is to choose the low-cost nodes as seeds. We call them handbills, because handbills are inexpensive to produce but can be distributed to a large number of individuals.

The two strategies select seeds from two different kinds of node candidates. They are billboard nodes and handbill nodes. In this paper, we distinguish them according to the degree. Since most real networks have scale-free property [1], we could adopt Pareto principle to distinguish between billboards and handbills, that is, the top 20% of nodes with the largest degree can be defined as billboards and the remainders are handbills. Fig.1 is a degree Pareto chart for a citation network (called HEP-PH network and its description is given in Section 4.1). From the chart, the degree distribution has a long tail and fits Pareto's law. The border degree for this network is 19, that is, the nodes whose degree is larger than 19 are billboards and the others are handbills.

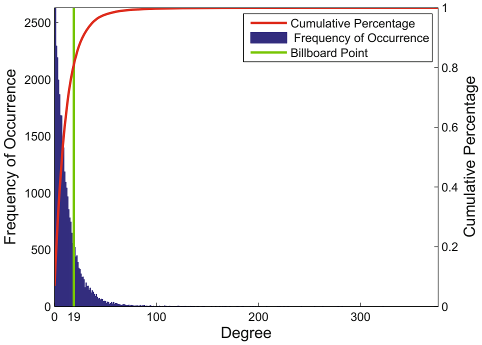


Fig. 1. Degree Distribution (HEP-PH)

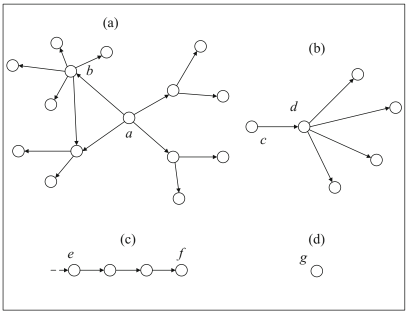


Fig. 2. Basic Structures of Social Network

2.3 Node Roles

A node role is a characterization of the part it plays in a network structure [11]. Nodes with different degrees or in different locations can be assigned different roles for describing their abilities to influence propagation. In our study, we first abstract four typical topologies from real networks, which are the basic structures to compose a network. They are respectively multi-cluster, single cluster, chain and loner, shown in Fig.2. Observing the nodes in these topologies, we define six roles to characterize different kinds of nodes, and they can cover all the nodes in the network. The definitions are follows.

- **King.** King is a global hub node in a multi-cluster topology, providing connections to many local communities. It has a wide influence by two characteristics. Firstly, it is a high-degree node that can influence many nodes directly. Secondly, many of its neighbors are also high-degree nodes. The influence can further propagate by the influential neighbors indirectly. In Fig.2, node a is a king node.

- **Seignior.** Seignior is a hub node in a local community. It is also a high-degree node. However, comparing with the king, it does not have that many influential neighbors and its influence is limited to a local region. In Fig.2, node b is a seignior.
- **Butterfly.** We call a node butterfly from the word "butterfly effect" meaning that a low-degree node can have wide influence by activating its influential neighbors. For example, in Twitter network, once a grassroot's tweet is forwarded by a celebrity, it will be popular explosively. In Fig.2, node c is a butterfly.
- **Leaf.** Leaf is the end of a communication chain. In Fig.2, node f is a leaf node. It only accepts information from spreaders, but can not further transmit it to others.
- **Loner.** Loner is an isolated node that has no connections with others, shown as node g in Fig.2.
- **Civilian.** Except for the above five types of nodes, the remaining nodes of a network can be categorized as civilians. Civilian nodes have no particular characteristics for influence propagation. In Fig.2, node e is a civilian.

Based on the above description, we distinguish the roles by quantification. We define kings and seigniors as billboard nodes, which have the top 20% largest degree in the network, and the other four roles are handbill nodes. We distinguish kings from seigniors by calculating the ratio of the sum of a node's neighbors' degree to its own degree and comparing it with a threshold. If the ratio is larger than a threshold, it is a king, otherwise, it is a seignior. For the four handbill roles, if a node's neighbors contain billboard nodes, we call it butterfly. If a node has precursors but has no followers, we call it leaf. If a node has neither precursors nor followers, we call it loner. Except for the five roles, the remaining nodes of a network are called civilians.

3 The Proposed Methods

In this paper, we propose three strategies to address the BIM problem. They are Billboard strategy, Handbill strategy and Combination strategy. Before designing algorithms, we first investigate seed selection heuristics through experimental analysis, which provides the basis for the proposed algorithms.

3.1 Seed Selection Heuristics

An commonly used method for the BIM problem is Modified Greedy algorithm [7,10]. In this section, we compute the seed set with Modified Greedy algorithm and obtain some heuristics from the seeds, which are the foundation of the proposed algorithms.

Firstly, we calculate the proportion of each role in a real network according to the role definition. The statistical result of a citation network (HEP-PH) is shown in Table 2.

Table 2. Proportion of Each Role in the HEP-PH Network

Role	King	Seignior	Butterfly	Leaf	Loner	Civilian
Proportion	0.0398	0.1652	0.3369	0.0721	0	0.3860

Then, we compute the seeds with Modified Greedy algorithm and calculate the proportion of each role in the seed set. The basic idea of Modified Greedy algorithm is to compute the solutions by using two different heuristics and return the better one as the result. Consequently, we can get two seed sets and each set is computed by one heuristic. The first heuristic is the basic greedy heuristic:

$$v_k = \underset{v \in V \setminus S_{k-1}}{\text{argmax}} \; \sigma(S_{k-1} \cup v) - \sigma(S_{k-1}), \tag{1}$$

where v_k is the target seed in step k, S_{k-1} is the seed set in step $k-1$, and $\sigma(S_{k-1})$ is the influence of seed set S_{k-1}. We perform the Greedy algorithm on HEP-PH network and calculate the proportion of each role in the seed set. The result is shown in Table 3.

The second heuristic takes cost into account and choose the most cost-effective nodes as the seeds.

$$v_k = \underset{v \in V \setminus S_{k-1}}{\text{argmax}} \; \frac{\sigma(S_{k-1} \cup v) - \sigma(S_{k-1})}{c(v)}, \tag{2}$$

where $c(v)$ is the cost of node v. The proportion of each role in the seeds computed by this heuristic is shown as Table 4.

Table 3. Proportion of Each Role in the Seed Set with Formula (1)

Budget	King	Seignior	Butterfly	Leaf	Loner	Civilian
1000	0.7273	0.0909	0.1818	0	0	0
3000	0.7857	0.1071	0.0714	0	0	0.0357
5000	0.8235	0.1373	0.0392	0	0	0

Table 4. Proportion of Each Role in the Seed Set with Formula (2)

Budget	King	Seignior	Butterfly	Leaf	Loner	Civilian
1000	0.1667	0.0513	0.6026	0	0	0.1795
3000	0.1216	0.0378	0.6622	0	0	0.1784
5000	0.1961	0.0784	0.6118	0	0	0.1137

From the above calculations, we can obtain two heuristic methods. The first one is role heuristic. Comparing Table 3 with Table 2, we could find that although there are very few king nodes in the network, they account for a large percentage in the seed set obtained by basic greedy heuristic. The basic greedy heuristic chooses the most influential seeds in the network, which is in accord with Billboard strategy. Then, the approach of Billboard strategy should distinguish kings from seigniors. Comparing Table 4 with Table 2, we could find that when we take cost into account, the butterfly node has more advantage than the other roles in influence propagation. Since, Handbill strategy is to choose the most cost-effective nodes as the seeds, the approach of Handbill strategy should distinguish butterfly from other roles.

The second heuristic method is distance heuristic. We measure the distance between each pair of the selected seeds, that is the length of the shortest path form a seed to another one. The average distance of HEP-PH network is larger than 2. Thus, in the proposed algorithms, we do not choose the nodes whose neighbors have already existed in the seed set, which can avoid the overlap of the seeds' influence.

In our study, we also do the same analyses on three other networks (their descriptions are declared in Section 4.1) and obtain similar conclusions. Due to the space limitation, we do not show the analysis results here.

3.2 Billboard Strategy

Billboard strategy is the idea that chooses the most influential nodes in the network as the seeds. For example, companies put advertisements on major media and products are endorsed by celebrities, both of the marketing actions are Billboard strategy.

We have defined that Billboard strategy chooses seeds from billboard nodes, whose roles are kings and seigniors. And according to the role heuristic, the proposed approach of Billboard strategy should distinguish kings from seigniors. The distinguishing characteristic of kings from seigniors is that not only king nodes themselves but also most of their neighbors are high-degree nodes. It implies that when we evaluate the influence of a node, we need to consider its neighbors' abilities to influence propagation. Then, evaluating a node's influence can be converted as calculating the expected number of influenced nodes in its two-hop area.

Theorem 1. *Given a social network $G = (V, E)$ and the IC model with a small propagation probability p, let $N(u) = \{v|v \in V, e_{u,v} \in E\}$ be the out-neighbors of node u, $outD(v)$ be the out-degree of node v. The expected number of influenced nodes in seed u's two-hop area is estimated by*

$$1 + \sum_{v \in N(u)} (1 + outD(v) \cdot p) \cdot p - \sum_{\exists v_i, v_j \in N(u), e_{v_i,v_j} \in E} p^3. \tag{3}$$

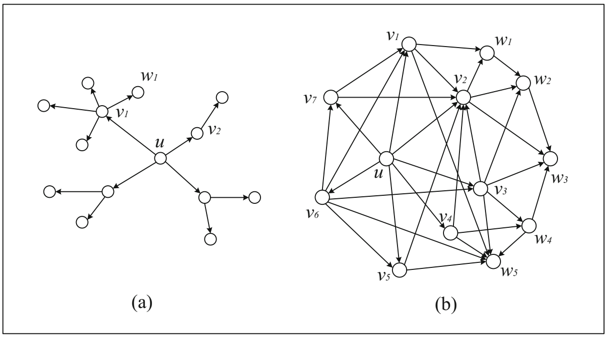


Fig. 3. The Topology of a Seed's Two-hop Area

Proof. Firstly, We consider a simple situation that there are no connections between the nodes in a seed's two-hop area, shown as Fig.3(a). Suppose node u is the seed, node $v_i \in N(u)$ is the seed's neighbor, and node $w_j \in \{w|w \in N(v_i), v_i \in N(u)\}$ is the seed's neighbor's neighbor. The probability that node v_i is influenced by seed u is p and the probability that node w_j is influenced by seed u is p^2. Then, the expected number of influenced nodes in seed u's two-hop area can be defined as:

$$1 + \sum_{v \in N(u)} (1 + outD(v) \cdot p) \cdot p, \tag{4}$$

where $outD(v)$ is the out-degree of node v. In this definition, 1 means node u itself, that is sure to be activated, and $\sum_{v \in N(u)} (1 + outD(v) \cdot p) \cdot p$ means the number of the potentially influenced nodes in its neighbors and neighbors' neighbors.

However, real-world networks usually have many tightly-knit groups that are characterized by a high density of ties [12], shown as Fig.3(b). In this situation, Formula (4) can not estimate a seed's influence accurately. For example, suppose that there are m edges connecting from node $v_i (i \neq 2)$ to node v_2. In the correct calculation, the probability p_{u,v_2} that node v_2 is influenced by node u is:

$$\begin{aligned} p_{u,v_2} &= 1 - (1-p)(1-p^2)^m \\ &= 1 - (1-p)[1 - mp^2 + C_m^2 p^4 + o(p^4)] \\ &= p + mp^2 - mp^3 + o(p^3). \end{aligned}$$

Suppose that there are n edges connecting from v_i to node w_5. In the correct calculation, the probability p_{u,w_5} that node w_5 is influenced by node u is:

$$\begin{aligned} p_{u,w_5} &= 1 - (1-p^2)^n \\ &= 1 - [1 - np^2 + C_n^2 p^4 + o(p^4)] \\ &= np^2 + o(p^3). \end{aligned}$$

Since the propagation probability p is usually very small, we can ignore $o(p^3)$.

However, in formula (4), we calculate the probability p_{u,v_2} as $p + mp^2$, since the probability that node v_2 is directly influenced by seed u is p and indirectly influenced by the m neighbors is mp^2. And we calculate the probability p_{u,w_5} as np^2, since node w_5 is indirectly influenced by the n neighbors. Comparing this calculation with the above derivation, for each edge $e \in \{e_{v_i,v_j} | e_{v_i,v_j} \in E, v_i \in N(u), v_j \in N(u)\}$, Formula (4) should minus p^3. Then, we can get Formula (3).

From the above, the algorithm of Billboard strategy can be stated as follows. We only take billboards that are the top 20% nodes with the largest degree in the network as the candidates for seed selection, and calculate their abilities to influence propagation by Formula (3). We in turn select the next best candidate with the largest ability value as the seed until the budget is exhaust. Based on the distance heuristic, when choosing a seed, we would judge whether its neighbors have already existed in the seed set. If not yet, we will choose it, otherwise, we will ignore it and take the next one.

3.3 Handbill Strategy

Handbill strategy is the idea that chooses the most cost-effective nodes as the seeds. In the real world, advertisers, limiting each location's cost to increase the advertising locations, can also broaden the awareness of product.

We have defined that Handbill strategy chooses seeds from handbill nodes, whose roles are butterfly, leaf, loner and civilian. And according to the role heuristic, the proposed approach of Handbill strategy should prioritize butterfly nodes. The distinguishing characteristic of butterfly nodes from other handbill roles is that there are some high-degree nodes existing in their neighbors. They can indirectly influence more nodes by their high-degree neighbors. Then, Formula (3) can also evaluate a node's influence for Handbill strategy. Moreover, since Handbill strategy is sensitive to cost, we divide the influence of a node by its cost and evaluate a node's cost-effective value as:

$$ce(u) = (1 + \sum_{v \in N(u)} (1 + outD(v) \cdot p) \cdot p - \sum_{\exists v_i, v_j \in N(u), e_{v_i,v_j} \in E} p^3)/c(u), \quad (5)$$

where $ce(u)$ is the cost-effective value of node u, $c(u)$ is the cost of node u, $N(u)$ is the out-neighbors of node u and $outD(v)$ is the out-degree of node v.

From the above, the algorithm of Handbill strategy can be stated as follows. We only take handbills that are the nodes with the bottom 80% largest degree in the network as the candidates for seed selection, and calculate their cost-effective value by Formula (5). We in turn select the next best candidate with the largest cost-effective value as the seed until the budget is exhaust. Based on the distance heuristic, when choosing a seed, we would judge whether its neighbors have already existed in the seed set. If not yet, we will choose it, otherwise, we will ignore it and take the next one.

3.4 Combination Strategy

After performing the above two strategies, we could get two nonoverlapping seed sets. One is billboard set, where the seeds have great influence. The other one is handbill set, where the seeds are cost-effective. Then, we proposed Combination strategy to select the "best" seeds from the two "better" seed sets and obtain a compositive solution. The proposed approach is based on Simulated Annealing algorithm [9], which taking influence maximization as the objective, searches an approximate solution in the two set of nodes.

We first give a brief introduction of the procedure of SA algorithm as follows.

(1) The algorithm firstly initializes an initial state S_0, an initial temperature T_0, an annealing schedule $T(t)$ and an objective function $E(S_t)$.

(2) At each step t, it produces a new state S'_t from the neighbors of the current state S_t. It probabilistically decides between moving the system to state S'_t or staying in state S_t. If $E(S'_t) \geq E(S_t)$, the system moves to the new state S'_t with the probability 1; otherwise, the system does this move with a probability of $exp(-(E(S'_t)-E(S_t))/T_t)$.

(3) The algorithm iteratively does step (2), until the system reaches a good enough state, or until the temperature T decreases to 0.

The Combination strategy is outlined in Algorithm 1. In this strategy, we evaluate the influence of a node by Formula(3) and define the objective function $E(S)$ as the sum of the influence of each node in the seed set. We set the initial seed set as billboard seed set. The algorithm has two levels of node replacements. The first one is billboard replacement(lines 3-10). In this level, we reduce a billboard $b \in S$ and add the equal cost of handbills $h_i \notin S$ to produce a neighbor set. Then, we judge whether to accept the new set(lines 4-10). If accepted, the algorithm comes into the second level of replacement: handbill replacement(lines 12-20). In the second level, we randomly replace a handbill $h \in S$ with the equal cost of other handbills $h_i \notin S$ to produce a neighbor set, and judge whether to accept the new set(lines 14-20). We repeat this replacement for q times. The billboard replacement is the outer iteration. We in turn try to replace each billboard. When all the billboard replacements have been executed, the algorithm is finished.

4 Experiments

4.1 Datasets

We use four real-world networks in our experiments. The first one is HEP-PH citation network, where nodes are papers and an directed edge means one paper cites another. The second one is an Email network, which records one day of email communication in a school, where nodes are email addresses and edges are communication records. The third one is a P2P network, where nodes represent hosts and edges represent the

Algorithm 1. Combination Strategy

Input: Graph $G = (V, E)$, budget B, billboard seed set S_B, handbill seed set S_H, initial temperature T_0, temperature drop ΔT, objective function $E(S)$ and the number of loop q.
Output: The final seed set S.

```
$t \leftarrow 0, T_t \leftarrow T_0, S \leftarrow S_B, |S_B| = k$;
for $i \leftarrow 1$ to $k$ do $flag \leftarrow$ false;
    create a neighbor set $S'$; /*replace a billboard $S_B(i)$ with the equal cost of handbills*/
    calculate the influence difference $\Delta E \leftarrow E(S') - E(S)$;
    if $\Delta E \geq 0$ then
        $S \leftarrow S'$, $flag \leftarrow$ true;
    else
        create a random number $\xi \in U(0, 1)$;
        if $exp(\Delta E/T_t) > \xi$ then
            $S \leftarrow S'$, $flag \leftarrow$ true;
    if $flag$ is true then
        while $q > 0$ do $q \leftarrow q-1$
          create a neighbor set $S'$; /*replace a handbill with the equal cost of other handbills*/
            calculate the influence difference $\Delta E \leftarrow E(S') - E(S)$;
            if $\Delta E \geq 0$ then
                $S \leftarrow S'$;
            else
                create a random number $\xi \in U(0, 1)$;
                if $exp(\Delta E/T_t) > \xi$ then
                    $S \leftarrow S'$;
    $T_t \leftarrow T_t - \Delta T$
return $S$;
```

connections between two hosts. And the last one is a Web network, where nodes are web pages and edges are links. Some of them are used in recent influence maximization research [3,4,10]. The basic statistics of the datasets are given in Table 5.

Table 5. Statistics of Datasets

Dataset	HEP-PH	Email	P2P	Web
Nodes	34,546	27,018	62,586	148,468
Edges	421,578	66,012	147,892	356,294

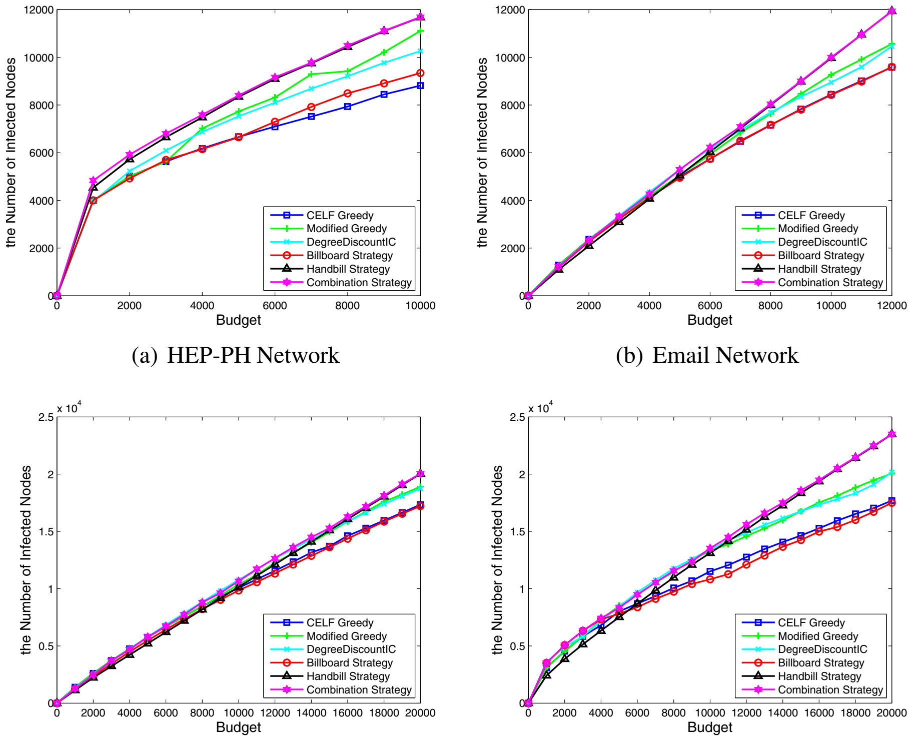
4.2 Experiment Setup

In the experiment, we adopt simulation method [5] to compute the influence propagation of the resulting seed set. Given a seed set, we repeat the simulations for R=$10,000$ times and take the average. We use IC model to simulate the influence propagation and set the propagation probability to be 0.1. In Combination strategy, we set T_0=$1,000,000$ and ΔT=$1,000$. All the experiments are performed on a server with Intel(R) Core i7-4770K CPU, 32G memory. All the codes are written in Java.

We evaluate the accuracy and efficiency of the three proposed strategies with three state-of-the-art algorithms. The first baseline algorithm is CELF Greedy algorithm(CELF G) [7], which is the basic greedy algorithm with CELF speed optimization. The second one is Modified Greedy algorithm(Modified G) [7], which has been introduced in Section 3.1. The third one is DegreeDiscountIC algorithm(DDIC) [3]. It is a fast algorithm for influence maximization problem, and we modify it to make it suit for BIM problem.

4.3 Experiment Results

Accuracy When Varying Budget. In this experiment, we evaluate the accuracy of each algorithm by varying the budget. In the real world, the advertising price of a web site is always relevant to its popularity. Then we define the cost of a node u as $outD(u) \cdot p + 1$, where $outD(u)$ is the out-degree of node u and p is the propagation probability.



(a) HEP-PH Network

(b) Email Network

(c) P2P Network

(d) Web Network

Fig. 4. Influence Propagation of Different Algorithms as Budget is Varied

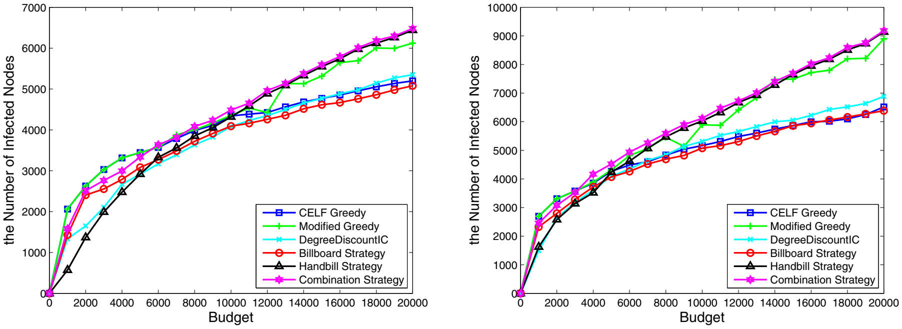
The experiment results are shown in Fig. 4. For HEP-PH network, Combination strategy has the best accuracy. It outperforms the second best algorithm Handbill strategy by 1.19% averagely, but this advantage reduces with the increase of budget. Combination strategy and Handbill strategy are superior to Billboard strategy obviously, which implies it is better to choose cost-effective nodes rather than influential nodes as seeds for HEP-PH network. Modified G and DDIC have similar accuracies. They are respectively inferior to Combination strategy by 10.36% and 13.3% averagely. And CELF G has the worst performance in accuracy.

For Email network, when the budget $B \in [0, 5, 000]$, Billboard strategy is better than Handbill strategy in accuracy. But when $B \in (5, 000, 12, 000]$, the latter overtakes the former. When $B \in [0, 2, 000]$, Modified G and CELF G have the best accuracies and outperform Combination strategy by 2.7%. When $B \in (2, 000, 4, 000]$, DDIC has the best accuracy and outperforms Combination strategy by 1.45%. When $B \in (4, 000, 12, 000]$, Combination strategy becomes the best algorithm in accuracy.

For P2P network, when the budget $B \in [0, 8, 000]$, Billboard strategy has a better performance than Handbill strategy in accuracy. But when $B \in (8, 000, 20, 000]$, the latter overtakes the former. When $B \in [0, 4, 000]$, Modified G and CELF G have the best accuracies and outperform Combination strategy by 2.6% averagely. When $B \in (4, 000, 10, 000]$, DDIC overtakes Greedy algorithm to be the best algorithm and slightly outperforms Combination strategy by 1.15%. When $B \in (10, 000, 20, 000]$, Combination strategy becomes the best algorithm in accuracy. It outperforms DDIC by 3.03% and outperforms Modified G by 3.71% averagely.

For Web network, when the budget $B \in [0, 5, 000]$, Billboard strategy is better than Handbill strategy in accuracy. But when $B \in (5, 000, 20, 000]$, the latter overtakes the former. When $B \in [0, 4, 000]$, Combination strategy has the best accuracy and outperforms the second best algorithm Modified G by 2.18%. When $B \in (4, 000, 9, 000]$, DDIC overtakes Combination strategy and becomes the best algorithm in accuracy. It slightly outperforms Combination strategy by 1.72%. When $B \in (9, 000, 20, 000]$, Combination strategy once again becomes the best algorithm. It outperforms Modified G by 10.85% and outperforms DDIC by 10.89% averagely.

Accuracy with Different Cost Definitions. This experiment is to evaluate the accuracy of each algorithm with different cost definitions. We respectively define the cost of a node u as $outD(u) + 1$ and $outD(u) \cdot c + 1$, where $outD(u)$ is the out-degree of node u, p is the propagation probability and c is a random number in $(0, 1]$. Due to the space limitation, we only show the results of HEP-PH network in Fig.5.



(a) $c(u) = outD(u) + 1$

(b) $c(u) = outD(u) \cdot c + 1$

Fig. 5. Influence Propagation of Different Algorithms as Cost is Varied

With these two cost definitions, when the budget is small, Billboard strategy always has a better accuracy than Handbill strategy. But with the increase of budget, Handbill strategy overtakes the former and even outperforms most other algorithms. Combination strategy has an excellent and stable performance in accuracy. When the budget is small, it may fall behind with the best algorithm slightly. However, as the budget increases, it narrows the gap rapidly and becomes the best algorithm, which outperforms Handbill strategy slightly and outperforms other algorithms obviously. Modified G has a good accuracy when the budget is small. But it is inferior to Combination strategy and Handbill strategy when the budget is large. We can find that the curve of Modified G have some drops, which implies the Greedy algorithm falls into local optimums. CELF G performs well when the budget is small, but it can not maintain the advantage when the budget is increasing. Comparing with other algorithms, DDIC has a low accuracy.

Efficiency We compare the runtimes of the six algorithms, when the budget is 5,000 and the propagation probability is 0.1 on the four networks.

From Table 6, we observe that Billboard strategy, Handbill strategy, Combination algorithm and DDIC have the runtime in the same order of magnitude. And Combination algorithm is slightly slower than the other three. By contrast, CELF G has a much longer runtime and it is slower than the fastest algorithm by three orders of magnitude. Modified G has the longest runtime, it is even slower than CELF G by 2 times at least.

Table 6. Runtime(seconds) of Each Algorithm on Different Networks

Network	Billboard	Handbill	Combination	CELF G	Modified G	DDIC
HEP-PH	5.89	8.23	23.21	4.95×10^3	1.23×10^4	6.85
Email	4.91	6.96	19.78	2.27×10^3	7.63×10^3	5.08
P2P	6.32	9.61	27.76	5.05×10^3	1.26×10^4	7.53
Web	11.03	13.49	40.25	7.36×10^3	2.09×10^4	13.42

5 Conclusions

Unlike most of the existing works on k-node influence maximization, this paper focuses on budgeted influence maximization. We address the BIM problem by three strategies: Billboard strategy, Handbill strategy and Combination strategy. From the comparison experiments with the state-of-the-art algorithms, we can conclude that Combination strategy is the most balanced algorithm. On one hand, it has the best performance or matches the best algorithms in accuracy. On the other hand, it has the runtime in the same order of magnitude with the fastest algorithm. Billboard strategy and Handbill strategy have the best efficiencies. They can obtain good solutions in some situations.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No.61175052, 61203297, 61035003), National High-tech R&D Program of China (863 Program) (No.2014AA012205, 2013AA01A606, 2012AA011003), National Program on Key Basic Research Project (973 Program) (No.2013CB329502).

References

1. Bollobás, B., Riordan, O., Spencer, J., Tusnády, G., et al.: The degree sequence of a scale-free random graph process. Random Structures & Algorithms 18(3), 279–290 (2001)
2. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1029–1038. ACM (2010)

3. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208. ACM (2009)
4. Jiang, Q., Song, G., Cong, G., Wang, Y., Si, W., Xie, K.: Simulated annealing based influence maximization in social networks. In: AAAI (2011)
5. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
6. Khuller, S., Moss, A., Naor, J.S.: The budgeted maximum coverage problem. Information Processing Letters 70(1), 39–45 (1999)
7. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429. ACM (2007)
8. Megiddo, N., Zemel, E., Hakimi, S.L.: The maximum coverage location problem. SIAM Journal on Algebraic Discrete Methods 4(2), 253–261 (1983)
9. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087 (1953)
10. Nguyen, H., Zheng, R.: On budgeted influence maximization in social networks. IEEE Journal on Selected Areas in Communications 31(6), 1084–1094 (2013)
11. Scripps, J., Tan, P.N., Esfahanian, A.H.: Node roles and community structure in networks. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 26–35. ACM (2007)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)