Influence Maximization in Social Networks

Wu Chengyang
515030910415
wuchengyang@sjtu.edu.cn

***Abstract*—Influence maximization is the problem of choosing a small set of seed users within a large social network in order to maximize the spread of influence under certain diffusion models. Models and processes have been widely studied on the diffusion and propagation of influence in social networks. Specifically, a typical application is advertising products from enterprises. Previous work in this domain generally adopt data mining on the categories and historical behaviors of customers. Nevertheless, the above methods treated consumers as individual entities and omitted their relational bonds and mutual interactions. In fact, humans naturally pass experiences and options to each other. The long observed phenomenon is regarded as the *Word of Mouth* effect, i.e., oral communication from person to person. In social networks, users tend to share what articles they read and what products they purchase, or even what candidates they support in presidential elections. A fraction of neighbors are likely to adopt the same opinions and/or behaviors from the source and the same process may repeat for their neighbors as well. The aim is to determine a specific initial set so that the final influence is maximized when the propagation finishes. A variety of models already exist for applying to social networks, each with their own characteristics and functionalities. Since the issue of influence maximization is proved to be *NP-hard*, we here provide approximate solutions for efficient optimization. Additionally, new approaches to reduce cost and further enhance performance simultaneously are proposed. Corresponding algorithms illustrate a noticeable improvement of influence spread on social network data sets of various scales.**

I. Introduction

The rapid growth of online social networks in the past twenty years facilitated the spread of opinions and behaviors from individual to individual. Influence maximization aims to trigger the a large cascading diffusion by seeding an initial group of users. The allocation of this seed set is thus the critical focus of relevant research work.

The practical value of influence maximization can be best exemplified by companies in an attempt to promote their merchandise to clients. Traditionally, data mining and machine learning methods are employed to identify consumers with tags and labels thus predict the corresponding attitude and orders on the promoted products. Abundant work has already been accomplished to achieve applicable results in retailing and online shopping. Nonetheless, the previous solutions judged the users only on their own statistical information. The interactions between them, however, is absent from consideration. In reality, users are dependent of each other and may be influenced by the ideas and actions of their friends. If a consumer has purchased some specific products, it is reasonable that a number of neighbors tend to focus their attention and buy corresponding goods eventually.

The exhibition of influence diffusion exists not only in this scenario but in hugh amounts of other aspects in life as well. The phenomenon is denoted as the *Word of Mouth* effect, literally meaning the passing of information by oral communication.

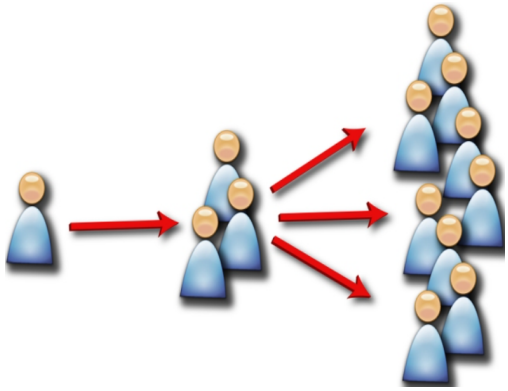

Fig. 1. the *Word of Mouth* effect

Upon feeling satisfied with what he has bought, a customer would probably recommend it to his intimate friends. The recommendation is usually more effective than directly advertising to the target people themselves, thus enterprises have made up the solution of *Free Trials*: Send samples of products free of charge to specific targets and hope they will share their thoughts and opinions with social neighbors. Another plausible approach is to invite influential individuals for endorsement of the products, i.e., to star in the corresponding advertisements. Practical options usually include celebrities like singers, actors, businessmen, etc. Regardless of which method to employ, the critical issue is to determine the selection of initial adopters. This problem of targeting groups of individuals to maximize the eventual influence originated from [1].

Another essential problem is limited access to the network topology. Previous work basically analyze the entire social network to exploit delicate bonds and relationships, while on the contrary retailers are only able to reach to those who previously subscribed as members or purchased some kind of products beforehand. Thus the possible range for selecting early adopters is limited to a small fraction of the nodes.

II. Model

Models are the basic components of influence spread. Previous work have proposed abundant available models for

representation of networks and diffusions. Basically, a social network is represented as a directed graph $G(V, E)$ with V nodes and E directed edges, the former denotes each user and the latter the social relationships between them. Upon advertising the products to a selected set of nodes, a fraction of the users accept it, i.e., they are *activated*. The *activated* nodes are hereby denoted as *seeds*, thus forming a *seed set*. As for the diffusion process, *active* nodes exert their influence on neighbors and the *inactive* ones among them will have a tendency to become *active*. Naturally considering, it is not hard to imagine that in a practical model resembling real world social networks the tendency for an *inactive* node to be activated is monotonically increasing with the number of neighbors that are already *active*. A typical model featuring the above characteristics is the *Susceptible-Infected model* widely adopted in medical and commercial affairs. *Susceptible-Infected-Recovered* and *Susceptible-Infected-Recovered-Susceptible* are models derived on the former one. These models approximate the ground truth well, but they fail to grab the minute changes concerning a small neighborhood. In order to capture the interactions between users rather than simply illustrating the gradual change of numbers of different types of nodes, we here adopt the following two models with each exhibiting different characteristics and functionalities [2]: *Linear Threshold Model* and *Independent Cascade Model*.

A. Linear Threshold Model

Any arbitrary edge **w** connecting two nodes **u** and **v** in a network consists of a weight and the sum of weights of the edges pointing to all nodes equals to 1.

$$\Sigma_{u \in pred(v)} w(u, v) = 1 \tag{1}$$

where **pred(v)** is the set of nodes with edges pointing to node **v**, i.e., the *predecessors* of node **v**. Correspondingly, an *inactive* node **v** is *activated* when the sum of weights from its *active predecessors* exceeds a certain value Θ

$$\Sigma_{u \in pred(v)} w(u, v) \geq \Theta \tag{2}$$

The threshold Θ for each node varies from individual to individual, which to some extent represents the level of activities and sensitivities for them. Once a node is *activated*, the same operations will be performed on its *successors*, i.e., the destiny nodes its edges are pointing to. Fig.2 illustrates a typical diffusion process for *Linear Threshold Model* where the *thresholds* of all nodes are equal to 50%. The influence spread of Linear Threshold Model has a stable solution, i.e., There is only one final stage. The weakness of it is bad adaptation to large scale networks as the effect of activated nodes may have to be taken into consideration for multiple steps when there exist remaining inactive successors.

B. Independent Cascade Model

In this model, there are no longer *thresholds* for individual nodes. Instead, each edge itself consists of an activation

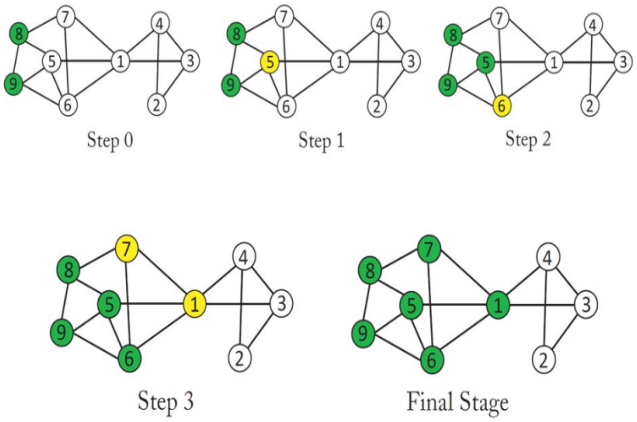

Fig. 2. Linear Threshold Model

probability **p(u,v)**. When a node **u** is *activated* at the end of a time slot, it will attempt to activate its *successors* at the beginning of next time slot where each *successor* **v** is *activated* with probability **p(u,v)** at the end of the time slot. The same operations are then performed again on the newly *activated* nodes.

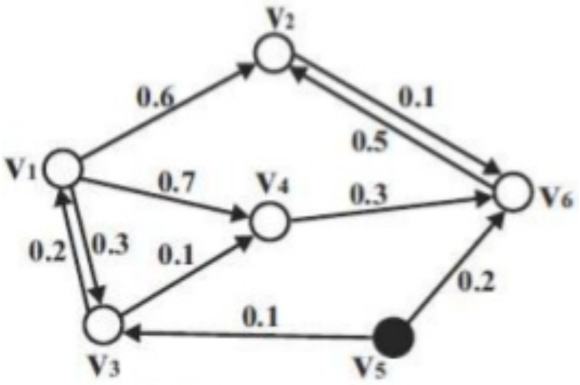

Fig. 3. Independent Cascade Model

The figure above illustrates an example of the influence diffusion for *Independent Cascade Model*. The effect of *activated* nodes will only be calculated once, so its performance fits well with large scale data sets The shortcomings of this model is the unpredictability of final stages, i.e., there could be numerous results. Due to the activation probability of each edge, it is random whether a node is activated or not. The figures 4 and 5 exemplify this randomness by showing two different final stages.

C. Applied Model

It is difficult to utilize the benefits of both *Linear Threshold Model* and *Independent Cascade Model* while deprecating their shortcomings. A compromise is accomplished as follows: In the social network graph **G(V,E)**, any arbitrary edge connecting two nodes **u** and **v** consists of a probability **p(u,v)**, and the sum of probabilities of all edges pointing to a node **u** is 1.

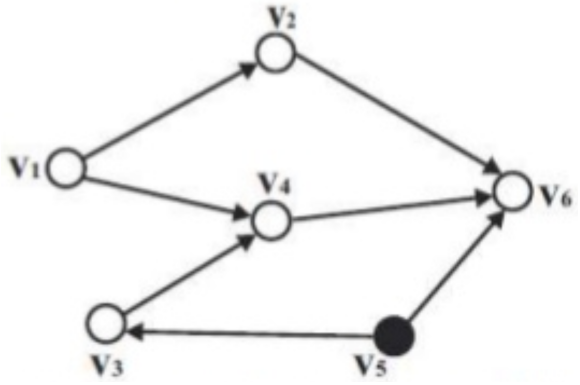

Fig. 4. A possible final stage for influence spread

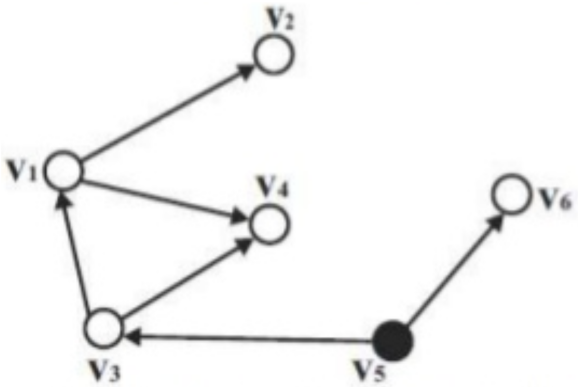

Fig. 5. Another possible final stage for influence spread

$$\Sigma_{u \in pred(v)} p(u, v) = 1 \tag{3}$$

For simplicity, the probabilities of these edges are hereby set to equal values

$$p(u, v) = \frac{1}{d_i(v)} \tag{4}$$

where $d_i(v)$ denotes the *in degree* of node **v**, i.e., the number of *predecessors* of this node.

D. Continuous and discrete influence maximization

Sending sample products to consumers is a risky operation as the possibility that a target refuses the promotion exists, in which case a lot of investment is actually wasted with no feedback. An improved solution is to provide discount coupons so that consumers can purchase products with a cheaper price. Thus when someone not interested ignored the coupons there is little loss for the enterprise. Apart from reducing the advertising budget, the company is now able to promote the products to more potential targets as well.

By changing free samples into discount coupons, the issue of influence maximization is now more flexible and leaves more space for optimization. Previously a target node can only be configured to be a seed, but now the discount can be of any value ranging from 0 to 1. The allocation can be further categorized into two subclasses: continuous allocation where the local optimal is achievable and discrete allocation implementing greedy algorithm at a little sacrifice of performance. Correspondingly, for a discrete approach where allocation can be selected from a set **S**, the optimal solution lies in one of the $|V|^{|S|}$ choices, which greatly reduces the complexity of this problem. Under this configuration, the optimization procedure is similar to the 0/1 selection in some ways.

E. Two stage seeding

Due to the limited access of nodes in the network, optimization within this range would probably render poor performance. Denote by **X** the nodes reachable in the beginning and **N(S)** the successors of group **S**, it is natural that some nodes in **N(x)** are likely to increase the influence spread at the final stage. An illustration of this mechanism is users inviting friends to clubs or websites. Only by activating corresponding nodes in **X** can nodes in **N(X)** become reachable. By applying this procedure in an attempt to reach nodes previously unaccessible, the seeding process can now be categorized into two stages: seeding in **X** and seeding in the reachable **N(X)**.

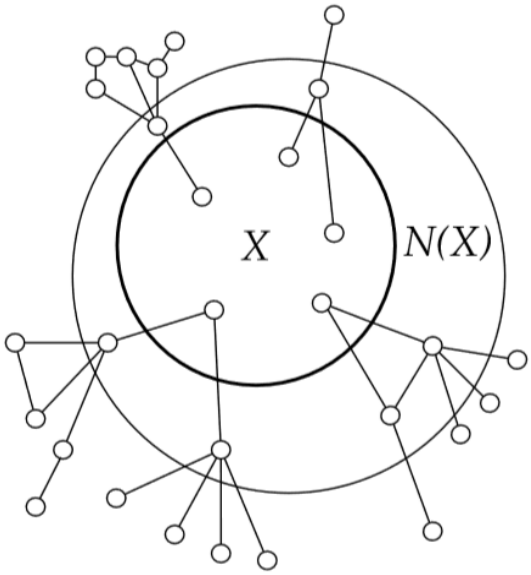

Fig. 6. Two stage seeding

The above figure illustrates a typical two stage seeding process where we:

1) Select seed nodes in the initial reachable set **X**.
2) Promote discount coupons to the corresponding users.
3) The users make their own decisions to accept or refuse the promotion.
4) For users who accept the promotion **S**, their successors in the social network **N(S)** are now accessible.
5) Perform influence maximization in **N(S)** to select a number of initial seed set.
6) Begin the diffusion process based on the early adopters given.

F. NP-hardness of influence maximization

Much work is already done to exploit the issue of influence maximization. Nevertheless, since the procedure can be proved to be NP-hard [2], the most plausible optimization mechanism is still nonexistent.

THEOREM 1. *The influence maximization problem is NP-hard for the Independent Cascade Model.*

Proof. Consider an instance of the NP-complete Set Cover problem, defined by a collection of subsets $S_1, S_2, ...S_m$ of a ground set $U = (u_1, u_2, ..., u_n)$; we wish to know whether there exist k of the subsets whose union is equal to U. Assuming $k < n < m$, we show that this can be viewed as a special case of the influence maximization problem.

Given an arbitrary instance of the *Set Cover* problem, we define a corresponding directed bipartite graph with $n + m$ nodes: there is a node i corresponding to each set S_i, a node j corresponding to each element u_j, and a directed edge (i, j) with activation probability $p(i, j) = 1$ whenever $u_j \in S_i$. The *Set Cover* problem is equivalent to deciding if there is a set A of k nodes in this graph with $\delta(A) \geq n + k$. Note that for the instance we have defined, activation is a deterministic process, as all probabilities are 0 or 1. Initially activating the k nodes corresponding to sets in a *Set Cover* solution results in activating all n nodes corresponding to the ground set U, and if any set A of k nodes has $\delta(A) \geq n + k$, then the *Set Cover* problem must be solvable.

THEOREM 2. *The influence maximization problem is NP-hard for the Linear Threshold Model.*

Proof. Consider an instance of NP-complete *Vertex Cover* problem defined by an undirected n-node graph $G = (V, E)$ and an integer k; we want to know if there is a set S of k nodes in G so that every edge has at least one endpoint in S. We show that this can be viewed as a special case of the influence maximization problem.

Given an instance of the *Vertex Cover* problem involving a graph G, we define a corresponding instance of the influence maximization problem by directing all edges of G in both directions. If there is a vertex cover S of size k in G, then one can deterministically make $\delta(A) = n$ by targeting the nodes in the set $A = S$; conversely, this is the only way to get a set A with $\delta(A) = n$.

III. Problem Formulation

With the collection of models mentioned in the previous section, we are now able to optimize the expected influence spread denoted as $UI(C)$, we will derive the representation of $UI(C)$ and prove its monotonicity [4], facilitating our proposal of solution in the next section.

Assume that different targets are activated with individual probabilities. Denote by $C = (c_1, c_2, ..., c_n)$ a configuration of discounts to all reachable nodes V . It is thus natural that the probability of a subset $S \in V$ of nodes become the seed set is

$$Pr(S; V, C) = \Pi_{u \in S} p_u(c_u) \Pi_{v \in V - S}(1 - p_v(c_v)) \quad (5)$$

Now we define the influence maximization problem as follows. Given a social network $G = (V, E)$, a budget B, an activation probability function $p_u(c_u)$ for every user u, and a diffusion model with influence function $I(S)$, find the configuration C that is the optimal solution to the following optimization problem:

$$\begin{aligned} maximize \quad & UI(C) \\ s.t. \quad & 0 \leq c_u \leq 1, \forall u \in V \\ & \Sigma_{u_v} c_u \leq B \end{aligned}$$

A configuration satisfying the above constraints is denoted as a *feasible configuration*.

The traditional influence maximization employing the 0/1 selection can be simplified to the following expression:

$$\begin{aligned} maximize \quad & UI(C) \\ s.t. \quad & c_u = 0 \quad or c_u = 1, \forall u \in V \end{aligned}$$

Assume that $p_u(c_u)$ is monotonic with respect to c_u, it is now hard to figure out that the inequality constraint $\Sigma_{u \in V} c_v \leq B$ can be replaced by an equation $\Sigma_{u \in V} c_u = B$. Apparently, an optimal C for influence maximization uses up the budget B. Thus the influence maximization uses up the budget B. Thus the influence maximization problem is now expressed as follows:

$$\begin{aligned} maximize \quad & UI(C) \\ s.t. \quad & 0 \leq c_u \leq 1, \forall u \in V \\ & \Sigma_{u_v} c_u = B \end{aligned}$$

IV. Coordinate Descent Algorithm

Based on the problem formulated above, we now deploy a coordinate decent algorithm to solve the influence maximization problem.

Upon allocating discount to target users in X and waiting for them to accept it, a new set of users are now reachable.

1) Coordinate Descent in Stage 2: Denote by S the seed set in stage 1 and B_2 the budget for resource allocation in this stage. The initial allocation in $N(S)$ can be arbitrary, e.g. equally split between all nodes. The coordinate descent now aims to optimize the allocation of discount C_2 in $N(S)$, denoting the target function as $Q(C_2; N(S))$. The process iterates a fixed amount of times or carries on until convergence. The optimization in each iteration is a random selection of two users u and v with allocated discounts c_u and c_v. They are rearranged to improve the performance of $Q(C_2; N(S))$ while those of the others are fixed.

Denote by $B_2^{'} = c_u + c_v$, the target function $Q(C_2; N(S))$ can now be rewritten with regard to c_u:

$$\begin{aligned} Q(C_2; N(S)) = \Sigma_{T \subseteq N(S) \backslash \{u,v\}} P_r(T; C_2, N(S)) \backslash \{u, v\}) \\ \{[1 - p_u(c_u)][1 - p_v(B_2^{'} - c_u)] I(T) \\ + [1 - p_u(c_u)] p_v(B_2^{'} - c_u) I(T \cup \{v\}) \\ + p_u(c_u)[1 - p_v(B_2^{'} - c_u)] I(T \cup \{u\}) \\ + p_u(c_u) p_v(B_2^{'} - c_u) I(T \cup \{u, v\})\} \end{aligned}$$

Which is a function of c_u. Since the discounts of nodes $0 \leq c_u \leq 1$ and $0 \leq c_v \leq 1$, the constraints on c_u is $max(0, B_2^{'}) \leq c_u \leq min(B_2^{'}, 1)$. Thus each iteration renders the new discount of u and v. The optimization problem is formulated as follows:

$$\begin{aligned} max \quad & Q(c_u) \\ s.t. \quad & max(0, B_2^{'} - 1) \leq c_i \leq min(B_2^{'}, 1) \end{aligned}$$

The optimization over this single variable in a close interval is guaranteed to achieve a close form solution. Since $p_u(\cdot)$ and $p_v(\cdot)$ are both continuous and differentiable, the optimal solution of c_u is either a stationary point in the range $(max(0, B_2^{'} - 1), min(B_2^{'}, 1))$ or one of the above border points.

2) Coordinate Descent in Stage 1: The coordinate descent in the stage is quite similar to the former one. The initial allocation can be arbitrary on the reachable set of nodes X, and the objective function $f(C_1; X)$ is optimized iteratively. In each iteration two random users i and j are selected. Denote by $B_1^{'} = c_i + c_j$ and the discount of the rest of the nodes are fixed. Now $f(C_1; X)$ can be expressed as follows.

$$\begin{aligned} f(C_1; X) = \Sigma_{S \subseteq X \backslash \{i,j\}} P_r(S; C_1, X \backslash \{i, j\}) \\ \{[1 - p_i(c_i)][1 - p_j(B_1^{'}) - c_i] maxQ(C_2; N(S)) \\ + [1 - p_i(c_i)] p_j(B_1^{'} - c_i) maxQ(C_2; N(S \cup \{j\}) \\ + p_i(c_i)[1 - p_j(B_1^{'})] maxQ(C_2; N(S \cup \{i\}))\} \\ + p_i(c_i) p_j(B_1^{'} - c_i) maxQ(C_2; N(S \cup \{i, j\}))\} \end{aligned}$$

For each $S \subseteq X \backslash \{i, j\}, maxQ(C_2; N(S))$ is derived by performing coordinate descent in stage 2. Similar to the expressions of stage 2, $f(C_1; X)$ is only dependent on c_i, so the optimization in each iteration can be formulated as follows:

$$\begin{aligned} max \quad & f(c_i) \\ s.t. \quad & max(0, B_1^{'} - 1) \leq c_i \leq min(B_1^{'}, 1) \end{aligned}$$

Similar to the previous proposals, the optimization problem here is guaranteed to have a closed form solution. It is also natural that the values of $f(\cdot)$ and $Q(\cdot)$ are upper bounded by the number of nodes in the corresponding seed set. As long as the optimization iteration exhibits a monotonic increase in the performance, the proposed algorithm is bound to converge. However, the possibility that this optimization procedure falls into a local optimum exists. A plausible solution is to adopt several random initial allocations.

V. EXPERIMENTS

The performance of the above proposed algorithms are assured to be non-decreasing during the optimization process, but its scalability and efficiency in practical scenarios have not yet been validated.

The data sets for experimental simulations are obtained from SNAP[5]. *Wiki-Vote* is the information of voting in public discussions for *Wikipedia*. *Ca-CondMat* covers the scientific collaborations between authors who submit papers to *Condense Matter* category. The scales of the two data sets are provided as follows.

TABLE I
STATISTICAL DATA OF DATA SETS WIKI-VOTE AND CA-CONDMAT

Data set	Nodes	Edges
wiki-Vote	7115	103689
ca-CondMat	23133	93497

Note that the former one is a directed graph while the later is undirected. The solution for converting to a directed network is replacing every undirected edge (u, v) into two directed ones (u, v) and (v, u). For simplicity, we randomly select 100 nodes as the initially reachable nodes denoted as X. The information diffusion model, as proposed before, is adopted from the *Independent Cascade Model*. Each directed edge (u, v) consists of an activation probability

$$p_{uv} = \frac{1}{d_i(v)} \tag{6}$$

where $d_i(v)$ denotes the in degree of node v.

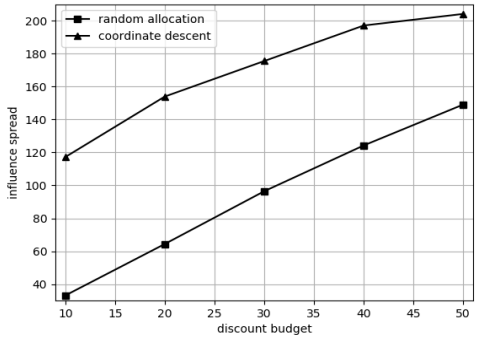

Fig. 7. Results on wiki-Vote

Results of the experiments are shown in the two figures above. It can be observed that coordinate descent algorithm provides a significant improvement on the performance of influence spread in the final stage for both data sets.

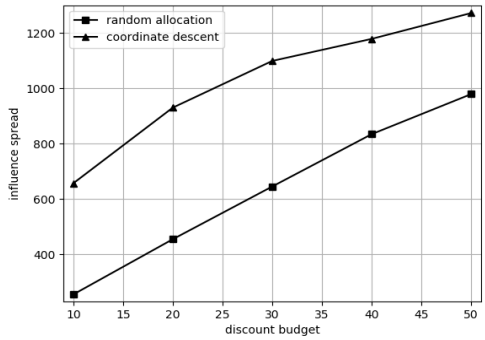

Fig. 8. Results on ca-CondMat

VI. CONCLUSION

The influence maximization problem is addressed with limited access to the users in a social network. The main contribution is to establish a two stage seeding model and design the corresponding coordinate descent algorithms to optimize the resource allocation. Experiments carried out on real-world data sets indicate the improvement of performance and thus verify the practicability of proposed algorithms.

REFERENCES

[1] Milad Eftekhar, Yashar Ganjali, and Nick Koudas. "Information cascade at group scale." knowledge discovery and data mining(2013): 401–409.

[2] Kempe, David, Jon M. Kleinberg, and Eva Tardos. "Maximizing the spread of influence through a social network." knowledge discovery and data mining (2003): 137-146.

[3] Chen, Wei, Yajun Wang, and Siyu Yang. "Efficient influence maximization in social networks." knowledge discovery and data mining (2009): 199-208.

[4] Yu Yang, Xiangbo Mao, Jian Pei, and Xiaofei He. "Continuous Influence Maximization: What Discounts Should We Offer to Social Network Users?" management of data(2016): 727-741.

[5] Jure Leskovec and Andrej Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection." http://snap.stanford.edu/data (2014)