Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 30 (2014) 81 – 87
Procedia Computer Science

1st International Conference on Data Science, ICDS 2014

A Note on Influence Maximization in Social Networks from Local to Global and Beyond

Chuan Zhou[a,*], Li Guo[a]

[a]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*

Abstract

In this paper, we study a new problem on social network influence maximization. The problem is defined as, given an activatable set A and a targeted set T, finding the k nodes in A with the maximal influence in T. Different from existing influence maximization work which aims to find a small subset of nodes to maximize the spread of influence over the entire network (i.e., from whole to whole), our problem aims to find a small subset of given activatable nodes which can maximize the influence spread to a targeted subset (i.e., from part to part). Theoretically the new frame includes the common influence maximization as its special case. The solution is critical for personalized services, targeted information dissemination, and local viral marketing on social networks, where fully understanding of constraint influence diffusion is essential. To this end, in this paper we propose a constraint influence maximization frame. Specifically, we point out that it is NP-hard and can be approximated by greedy algorithm with guarantee of $1 - 1/e$. We also elaborate two special cases: the local one and the global one. Besides, we present the works that are related and beyond.

Selection and peer-review under responsibility of the Organizing Committee of ICDS 2014.

Keywords: Influence maximization; Social network; United framework

1. Introduction

Social networks, as a popular and effective medium for information dissemination, play an increasingly important role in daily communication among individuals, groups and communities. Influence maximization is one of the fundamental problems in social networks, which has received significant attention in recent years [1, 2]. This research has been found useful in market recommendations, such as products, services, and innovative ideas, etc. through the powerful word-of-mouth effect in social networks.

The seminal work, by Kempe, Kleinberg and Tardos [3], first formulates influence maximization as a discrete optimization problem: Given a directed social graph with users as nodes, edge weights reflecting influence between users and a budget/threshold number k, finding k nodes in the graph, such that by activating these nodes, the expected spread of the influence can be maximized, based on a given stochastic influence propagation model.

Two popularly used stochastic influence propagation models are the *Independent Cascade* (IC) and *Linear Threshold* (LT) models [3]. In both models, at any time step, a user is represented as a binary variable with either

*Corresponding author. Tel.: +8610-82546710
E-mail address: zhouchuan@iie.ac.cn.

Selection and peer-review under responsibility of the Organizing Committee of ICDS 2014.
doi:10.1016/j.procs.2014.05.384

active (an adopter of the product) or inactive status, and influence propagates until no more users can become active. The major difference between the two models is the way of an active user propagating its influence to the neighbors: For the IC model when an inactive user becomes active at a time step t, it has exactly one chance to independently activate its currently inactive neighbors at the next time step $t + 1$; while in the LT model, the sum of incoming edge weights on any node is assumed to be at most 1, every user chooses an activation threshold uniformly at random from $[0, 1]$, and at any time step, a node becomes activated if the sum of incoming edge weights from the active neighbors exceeds the threshold.

Influence maximization under both IC and LT models is NP-hard, where a line of greedy/heuristic algorithms were proposed [3, 4, 5, 6, 7]. However, they all focus on finding global influential nodes over the entire social networks, through which the influence spread can be maximized. These work cannot answer the following question: given an activatable set A and a targeted set T, which nodes in A are the most influential ones to T?

To this end, in this paper we propose a constraint influence maximization frame. Specifically, we point out that it is NP-hard and can be approximated by greedy algorithm with guarantee of $1 - \frac{1}{e}$ (in Section 2). We elaborate its two special cases: the local one (in Section 3) and the global one (in Section 4). Finally, we present the related works (in Section 5) and conclude our paper with some future directions (in Section 6).

2. Constraint Influence Maximization

In this section we introduce the constraint influence maximization problem under the independent cascade model. Consider a directed graph $G = (V, E)$ with N nodes in V and edge labels $pp : E \to [0, 1]$. For each edge $(u, v) \in E$, $pp(u, v)$ denotes the propagation probability that v is activated by u through the edge. If $(u, v) \notin E$, $pp(u, v) = 0$. Let $Par(v)$ be the set of parent nodes of v, *i.e.*,

$$Par(v) := \{u \in V,\ (u, v) \in E\}. \tag{1}$$

Given an initially activated set $S \subseteq V$, the IC model works as follows. Let $S_t \subseteq V$ be the set of nodes that are activated at step $t \geq 0$, with $S_0 = S$. Then, at step $t + 1$, each node $u \in S_t$ may activate its out-neighbors $v \in V \setminus \cup_{0 \leq i \leq t} S_i$ with an independent probability of $pp(u, v)$. Thus, a node $v \in V \setminus \cup_{0 \leq i \leq t} S_i$ is activated at step $t + 1$ with the probability

$$1 - \prod_{u \in S_t \cap Par(v)} (1 - pp(u, v)) \tag{2}$$

where the subscript $u \in S_t \cap Par(v)$ means that node u, a parent node of v, is activated at step t. If node v is successfully activated, it is added into the set S_{t+1}. The process ends at a step τ with $S_\tau = \emptyset$. Obviously, the propagation process has $N - |S|$ steps at most, as there are at most $N - |S|$ nodes outside the seed set S. Let $S_{\tau+1} = \emptyset, \cdots, S_{N-|S|} = \emptyset$, if $\tau < N - |S|$. Note that each activated node only has one chance to activate its out-neighbors at the step right after itself is activated, and each node stays activated once it is activated by others.

In the IC model, the influence spread of a seed set S in the targeted set $T \subseteq V$, which is the expected number of activated nodes in T by S, is denoted by $\sigma_I(S \to T)$ as follow,

$$\sigma_I(S \to T) := \mathbb{E}^S\Big[\Big|\bigcup_{t=0}^{N-|S|} (S_t \cap T)\Big|\Big] \tag{3}$$

where $\mathbb{E}^S$ is the expectation operator with set S, and the subscript 'I' denotes the IC model.

Definition 1. *Given an integral $k \leq |V|$, the activatable set $A \subseteq V$ with $|A| \geq k$, and the targeted set $T \subseteq V$, the* ***constraint influence maximization*** *problem, under the IC model, is to find a subset $S^* \subseteq A$ such that $|S^*| = k$ and $\sigma_I(S^* \to T) = \max\{\sigma_I(S \to T)\ \big|\ |S| = k, S \subseteq A\}$, i.e.,*

$$S^* = \arg\max_{|S|=k, S \subseteq A} \sigma_I(S \to T). \tag{4}$$

From the definition, we can observe that, given a network structure $G = (V, E)$ and a diffusion model (IC model or LT model, etc.) which dominates the propagation process on the network, the constraint influence maximization problem is determined uniquely by its three inputs: the budget k, the activatable set A, and the targeted set T. Hence we can use a triple (k, A, T) to represent a constraint influence maximization problem. In especial, this frame includes two special cases.

- For a given node $w \in V$, let the activatable set $A = V \backslash \{w\}$ and the targeted set $T = \{w\}$, then the constraint influence maximization problem becomes the personal influence maximization, which was first proposed by Guo et al. [8].
- Let the activatable set $A = V$ and the targeted set $T = V$, then the constraint influence maximization problem becomes the influence maximization in the common sense, which was first proposed by Kempe et al. [3].

Theorem 1. *The constraint influence maximization problem under the IC model is NP-hard. For the given activatable set $A \subseteq V$ and targeted set $T \subseteq V$, let $f(S) := \sigma_I(S \to T)$ for all $S \subseteq A$, then the set function $f : 2^A \to \mathbb{R}^+$ is monotone and submodular with $f(\emptyset) = 0$.*

Under these observations, the constraint influence maximization problem in Eq. (11) can be approximated by the greedy algorithm as shown in Algorithm 1. Theoretically, a non-negative real valued function f on subsets of A is submodular, if $f(S \cup \{v\}) - f(S) \geq f(S' \cup \{v\}) - f(S')$ for all $v \in V$ and $S \subseteq S' \subseteq A$. Thus, f has diminishing marginal return. Moreover, f is monotone, if $f(S) \leq f(S')$ for all $S \subseteq S' \subseteq A$. For any submodular and monotone function f with $f(\varnothing) = 0$, the problem of finding a set S of size k that maximizes $f(S)$ can be approximated by the greedy algorithm in Algorithm 1. The algorithm iteratively selects a new seed u that maximizes the incremental change of f, to be included into the seed set S, until k seeds are selected. It is shown in [9] that the algorithm guarantees the approximation ratio $f(S)/f(S^*) \geq 1 - 1/e$, where S is the output of the greedy algorithm and S^* is the optimal solution.

Algorithm 1: Greedy(k,f)

initial $S = \varnothing$
for $i = 1$ to k **do**
select $u = \arg\max_{w \in A \backslash S}(f(S \cup \{w\}) - f(S))$
$S = S \cup \{u\}$
end for
output S

3. The Local Case: Personal Influence Maximization

In this section we elaborate the constraint influence maximization problem under the activatable set $A = V \backslash \{w\}$ and the targeted set $T = \{w\}$ for a given node $w \in V$. This is the personal influence maximization introduced in [8]. Let X be a random activation result (consisting of live edges [3], through which all activated nodes can be reached from S) in the whole network G with the seed set $S \subseteq V \backslash \{w\}$. Then the influence degree from set S to the target w can be measured as shown in Eq. (5),

$$R_w(S) := \mathbb{P}^S(w \in X) \tag{5}$$

where $\mathbb{P}^S$ is the probability measure via seed set S, and the notation $w \in X$ represents that w is a node in the activation result X. Hence, the influence degree $R_w(S)$ is the probability that w is successfully activated by the propagation process when the initial seed set is S. In fact, from Eq. (3) and Eq. (5), it follows that

Proposition 1. *For the seed set $S \subseteq V \backslash \{w\}$, we have*

$$\sigma_I(S \to \{w\}) = R_w(S). \tag{6}$$

The personalized influence maximization problem aims to find a seed set $S^* = \{s_1, s_2...s_k\} \subseteq V\backslash\{w\}$, such that $R_w(S^*) \geq R_w(S)$ for any set $S \subseteq V\backslash\{w\}$ with k nodes in the network G. Thus, the ***objective function*** of personalized influence maximization problem can be formally described as in Eq. (7),

$$S^* = \arg\max_{|S|=k, S\subseteq V\backslash\{w\}} R_w(S). \tag{7}$$

The key point of our problem is how to calculate the probability $R_w(S)$, which is a $\#P-$hard problem [8]. An alternative method is to use Monte-Carlo simulation to evaluate $R_w(S)$. Based on Eq. (5), we have

$$R_w(S) = \mathbb{E}^S(1_{\{w\in X\}}), \tag{8}$$

where the indicative function $1_{\{w\in X\}} = 1$ if $w \in X$ stands, otherwise, $1_{\{w\in X\}} = 0$.

In particular, let Ω^S be the sample space of all possible activation results throughout the whole network G under the IC model with the seed set S, then by Eq. (8), the influence degree on w from set S can be calculated in Eq. (9),

$$R_w(S) = \sum_{x\in\Omega^S} \mathbb{P}(X = x) \cdot 1_{\{w\in x\}} \tag{9}$$

where $\mathbb{P}(X = x)$ is the probability of x in the sample space Ω^S, and $\sum_{x\in\Omega^S} \mathbb{P}(X = x) = 1$.

From Eq. (8), we can observe that $1_{\{w\in X\}}$ is an unbiased statistics to $R_w(S)$. However, its variance is somewhat large when it is used in Monte-Carlo simulation. To address this problem, Guo et al. [8] found a better random function to simulate the objective function.

Theorem 2. *The statistics $1 - \prod_{v\in Y}(1 - p_{vw})$ is unbiased for evaluating $R_w(S)$, and it has smaller variance than $1_{\{w\in X\}}$, i.e.,*

$$\mathbb{E}^S\Big[1 - \prod_{v\in Y}(1 - p_{vw})\Big] = R_w(S),$$

$$\mathbb{D}^S\Big[1 - \prod_{v\in Y}(1 - p_{vw})\Big] < \mathbb{D}^S\Big[1_{\{w\in X\}}\Big],$$

where Y is the random activation result throughout the whole network without the target node w.

The influence degree from S to the target w can be measured as in Eq. (10), which is demonstrated effective to measure the personalized influence degree with provable smaller variance by Theorem (2).

$$R_w(S) = \mathbb{E}^S\Big[1 - \prod_{v\in Y}(1 - p_{vw})\Big]. \tag{10}$$

4. The Global Case: Influence Maximization in Common Sense

In this section we elaborate the constraint influence maximization problem under the activatable set $A = V$ and the targeted set $T = V$. This is the (global) influence maximization problem in the common sense, which was first introduced in [3]. We denote $\sigma_I(S) := \sigma_I(S \rightarrow V)$. The optimization problem in Eq. (4) becomes

$$S^* = \arg\max_{|S|=k, S\subseteq V} \sigma_I(S) \tag{11}$$

The problem, as proved in a previous work [3], is NP-hard, and a constant-ratio approximation algorithm is feasible. In the work [3, 10], it is shown that the objective function $\sigma_I(S)$ in Eq.(11) has the submodular and monotone properties [3] with $\sigma_I(\varnothing) = 0$. Thus, the problem in Eq.(11) can be approximated by the greedy algorithm as shown in Algorithm 1 with $f = \sigma_I$ and $A = V$.

In Algorithm 1, an important issue is that there is no efficient way to compute $\sigma_I(S)$ given a set S. Kempe et al. [3] run Monte-Carlo simulations of the propagation model for 10,000 trials to obtain an accurate estimate of the expected spread, leading to very expensive computation cost. Chen et al. [11] pointed out that computing $\sigma_I(S)$ is actually #P-hard, by showing a reduction from the counting problem of $s - t$ connectness in a graph.

However, The authors in [7] derive the upper bound for the spread $\sigma_I(S)$ under the IC model.

Theorem 3. *The upper bound for spread $\sigma_I(S)$ is*

$$\sigma_I(S) \leq \sum_{t=0}^{N-|S|} \Pi_0^S \cdot PP^t \cdot \boldsymbol{1} \tag{12}$$

where $PP = (pp_{ij})$ is the propagation probability matrix.

Furthermore, if the weight matrix PP satisfies the condition $\max_v \sum_u pp(u, v) < 1$ or $\max_u \sum_v pp(u, v) < 1$, the upper bound of $\sigma_I(S)$ can be relaxed to

$$\sigma_I(S) \leq \Pi_0^S \cdot (E - PP)^{-1} \cdot \mathbf{1}, \tag{13}$$

where E is a unit matrix and $(E - PP)^{-1}$ is the inverse of the matrix $(E - PP)$. By doing so, the upper bound in Eq. (13) is computationally tractable.

The upper bound can be used to prune unnecessary Monte-Carlo simulations in greedy algorithm, and Zhou et al. [7] proposed a new greedy-based algorithm *Upper Bound based Lazy Forward* (UBLF for short). The key idea behind classical CELF [12] is that the marginal gain of a node in the current iteration cannot be more than that in previous iterations. However, CELF demands N spread estimations to establish the initial bounds of marginal increments. In contrast, UBLF uses the upper bound given in Theorem 3 to rank all nodes in the initialization step, which eventually reduces the total number of spread estimations.

5. Related Works

Domingos and Richardson [13, 14] first formulated the influence maximization problem as an algorithmic problem in probabilistic methods. Later, Kempe et al. [3] first modeled the problem as the discrete optimization problem, as described in Section 4. A common limitation of greedy algorithm is computational inefficiency on large networks. Thus, two major types of solutions have been proposed.

First, many heuristic algorithms have been proposed to improve the efficiency of seed selection, *e.g.*, DegreeDiscount [4], MIA [11], DAG [15], SIMPATH [16], ShortestPath [17] and SPIN [18]. The heuristic algorithms proposed in these works can reduce computational cost in orders of magnitude, with competitive results of the influence spread level. However, none of these heuristic algorithms has a theoretical guarantee on the reliability of the results. In other words, it is unknown how far these heuristic solutions approximate the optimal solution. One can only borrow the simple greedy algorithm as the benchmark for performance testing.

Second, several optimized greedy algorithms have been proposed. A representative work, by Leskovec et al. [12], exploited the submodular property of the objective function, and proposed a Cost-Effective Lazy Forward selection (CELF) algorithm, which improves the running time of the simple greedy algorithm by up to 700 times. Following the same logic, Goyal et al. proposed CELF++ [19], an extension of CELF, that further reduces the number of spread estimation calls, leading to 35% – 55% faster than CELF. Besides, Chen et al. [4] proposed the NewGreedy and MixedGreedy algorithms in the IC model with uniform probabilities. However, their performance are non-steady, sometimes even worse than CELF. Recently, Zhou et al. in [7, 20] further enhanced the CELF by the upper bound based approach UBLF, in their method the Monte-Carlo calls in the first round are drastically reduced comparing with the CELF.

Besides the above two types of solutions, Wang et al. [21] discussed the influence maximization from the view of social community. They proposed a new community-based greedy algorithm for mining top-k influential nodes. Barbieri et al. [1] studied social influence from a topic modeling perspective. Guo et al. [22] investigated the influence maximization problem from the item-based data. Rodriguez et al. [23] studied the influence maximization problem in continuous time diffusion networks. Goyal et al. [24] proposed an alternative approach to influence maximization which, instead of assuming influence probabilities are given as input, directly uses the past available data. In the works [25, 26] the authors discussed the integral influence maximization problem when repeated activations are involved. The complementary problem of learning influence probabilities from the available data is studied in the works [27] and [28].

6. Conclusion

In this paper we propose the constraint influence maximization by considering the targeted activations from given activatable nodes set in social networks. We present its motivations, discuss its difference from the classical influence maximization problem, point out that it is NP-hard and can be approximated by greedy algorithm with guarantee of $1 - 1/e$. We also elaborate its two special cases: the local one and the global one.

There are several interesting future directions. First, the discrete formulation of propagation time can be further modified to a tractable continuous-time version; second, how to address the constraint influence maximization problem from real-world data is also challenging.

Acknowledgements

This work was supported by the NSFC (No. 61370025), 973 project (No. 2013CB329606), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences(No.XDA06030200).

References

[1] N. Barbieri, F. Bonchi, G. Manco, Topic-aware social influence propagation models, in: IEEE 12th International Conference on Data Mining (ICDM), IEEE, 2012, pp. 81–90.
[2] F. Bonchi, Influence propagation in social networks: A data mining perspective., IEEE Intelligent Informatics Bulletin 12 (1) (2011) 8–16.
[3] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: KDD 2003.
[4] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: KDD 2009.
[5] E. Even-Dar, A. Shapira, A note on maximizing the spread of influence in social networks, in: Internet and Network Economics, Springer, 2007, pp. 281–286.
[6] Y. Li, W. Chen, Y. Wang, Z.-L. Zhang, Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships, in: WSDM, 2013.
[7] C. Zhou, P. Zhang, J. Guo, X. Zhu, L. Guo, Ublf: An upper bound based approach to discover influential nodes in social networks, in: IEEE 13th International Conference on Data Mining (ICDM), IEEE, 2013, pp. 907–916.
[8] J. Guo, P. Zhang, C. Zhou, Y. Cao, L. Guo, Personalized influence maximization on social networks, in: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, ACM, 2013, pp. 199–208.
[9] G. L. Nemhauser, L. A. Wolsey, M. L. Fisher, An analysis of approximations for maximizing submodular set functionsłi, Mathematical Programming 14 (1) (1978) 265–294.
[10] E. Mossel, S. Roch, On the submodularity of influence in social networks, in: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, ACM, 2007, pp. 128–134.
[11] W. Chen, C. Wang, Y. Wang, 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, ACM, 2010, pp. 1029–1038.
[12] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: KDD 2007.
[13] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2001, pp. 57–66.
[14] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 61–70.
[15] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: ICDM 2010.
[16] A. Goyal, W. Lu, L. V. Lakshmanan, Simpath: An efficient algorithm for influence maximization under the linear threshold model, in: IEEE 11th International Conference on Data Mining (ICDM), IEEE, 2011, pp. 211–220.
[17] M. Kimura, K. Saito, Tractable models for information diffusion in social networks, in: PKDD 2006.
[18] R. Narayanam, Y. Narahari, A shapley value-based approach to discover influential nodes in social networks, IEEE Transactions on Automation Science and Engineering (99) (2010) 1–18.
[19] A. Goyal, W. Lu, L. V. Lakshmanan, Celf++: optimizing the greedy algorithm for influence maximization in social networks, in: WWW 2011.
[20] C. Zhou, P. Zhang, J. Guo, L. Guo, An upper bound based greedy algorithm for mining top-k influential nodes in social networks, in: 23rd International World Wide Web Conference (WWW), ACM, 2014, pp. 421–422.
[21] Y. Wang, G. Cong, G. Song, K. Xie, Community-based greedy algorithm for mining top-k influential nodes in mobile social networks, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 1039–1048.
[22] J. Guo, P. Zhang, C. Zhou, Y. Cao, L. Guo, Item-based top-k influential user discovery in social networks, in: IEEE 13th International Conference on Data Mining Workshops (ICDMW), IEEE, 2013, pp. 780–787.
[23] M. G. Rodriguez, B. Schölkopf, Influence maximization in continuous time diffusion networks, arXiv preprint arXiv:1205.1682.
[24] A. Goyal, F. Bonchi, L. V. Lakshmanan, A data-based approach to social influence maximization, Proceedings of the VLDB Endowment 5 (1) (2011) 73–84.

[25] C. Zhou, P. Zhang, W. Zang, L. Guo, Maximizing the long-term integral influence in social networks under the voter model, in: 23rd International World Wide Web Conference (WWW), ACM, 2014, pp. 423–424.
[26] C. Zhou, P. Zhang, W. Zang, L. Guo, Maximizing the cumulative influence through a social network when repeat activation exists, in: ICCS 2014.
[27] A. Goyal, F. Bonchi, L. V. Lakshmanan, Learning influence probabilities in social networks, in: Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, pp. 241–250.
[28] K. Saito, R. Nakano, M. Kimura, Prediction of information diffusion probabilities for independent cascade model, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2008, pp. 67–75.