Influence Maximization Based on True Threshold in Social Networks

Wei Hao[1,2], Qianyi Zhan[1,2(✉)], and Yuan Liu[1,2]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
6201613052@stu.jiangnan.edu.cn
[2] Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi, China
{zhanqianyi,lyuan1800}@jiangnan.edu.cn

Abstract. As E-marketing based on online social networks develops fast, the influence maximization problem draws attention from both academics and industries. This problem focuses on which subset of users should be selected as seed users so that based on the specific information diffusion model, the advertising companies can maximize word-of-mouth effect. Exisiting related work assume there is no cost to choose these seed users, or the cost is given in the problem setting. While in real situation, it is crucial but difficult to elicit users' true attitude over being seeds. Moreover, we notice "threshold" as users' private information in the Linear Threshold model can represent individual's preference. Thus we propose a new model in which users, willing to be seeds, are asked to report their threshold information. The method called TREE is designed to solve this model, especially the payment mechanism should make sure all users tell truth. Experiments on real social network data to verify the effectiveness of TREE.

Keywords: Social networks · Influence maximization · Linear threshold model

1 Introduction

Due to the success of online social networking (OSN) websites in recent years, the trend of using social networks as a marketing tool grows rapidly and spans diverse areas. Based on "word-of-mouth" diffusion process, *viral marketing* has received considerable attention from both field of research and industry. The idea behind viral marketing is using small marketing cost by targeting a set of most influential users in social network to make their aggregated influence reach a large portion of the network. Motivated by this, computer scientists focus on so-called *influence maximization problem*, which is finding such set of initial adopters that maximize the number of users influenced by them eventually.

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 235–247, 2021.
https://doi.org/10.1007/978-3-030-82562-1_21

The influence maximization problem is modeled as an algorithmic problem: given a social network graph $G = (V, E, w)$ and a budget k, where V denotes nodes in the social network, E is the set of directed edge representing the social relationship among nodes, and $w(u, v)$ represents the influence weight of node u on v and $0 \leq w(u, v) \leq 1$. The aim is to select a seed set S of k nodes such that the total influence effect is maximized through propagation.

Domingos and Richardson [11] first studied the problem of influence propagation and identification of influential users. The first systemic study of influence maximization problem was provided by Kempe et al. [3]. They proposed two basic stochastic influence models coming from mathematical sociology, namely, the *Independent Cascade (IC) model* and *Linear Threshold (LT) model*. Kempe et al. proved this problem is NP-hard for both two models, but the objective function of influence spread $\sigma(S)$ is *monotone* and *submodular*. The function is *monotone* if $\sigma(A) \leq \sigma(B)$ when $A \subseteq B$. Moreover it is modular if $\sigma(A \cup v) - \sigma(A) \geq \sigma(B \cup v) - \sigma(B)$ for all $A \subseteq B$ and $v \notin A$. Based on that Kempe et al. [3] proposed discrete optimization methodology and obtained the greedy approximation algorithms which achieved an approximation ratio of $1 - 1/e$.

Each node in both models is active or inactive. In the IC model, the simplest one of dynamic cascade models, an active node u is given a single chance to activate each currently inactive neighbor for example node v; the success probability is $w(u, v)$, which is independently of the history so far. While the linear threshold model is the core of most models based on the use of node-specific thresholds. Each node v holds a threshold θ_v uniformly at random from the interval $[0, 1]$, the node v is activated when the total weight of its active neighbors is no less than its threshold: $\Sigma w(u, v) \geq \theta_v$. The linear threshold model stresses the threshold behavior in propagation of influence, which makes it meets the actual situation better. For example, when enough people recommend a new restaurant or discuss a new book, we may also would like to follow it.

1.1 Related Work

A series of studies have been done to solve the influence maximization problem. The original greedy algorithm with constant-factor approximation was first proposed by Kempe et al. [3]. The Monte-Carol simulations on influence cascade to estimate the spread makes their algorithm not scalable. Leskovec et al. [5] developed an efficient algorithm based on "lazy forward" optimization in selecting new seeds. Experiments show it achieves near-optimal result with 700 times faster than the simple greedy algorithm.

By using properties of the IC model, much algorithms are designed for the IC model specifically. Kimura and Saito [6] proposed a model based on shortest-path. Chen et al. [12] develop a degree discount heuristics for the IC model with same edge probabilities. In [1], maximum influence arborescence (MIA) algorithm is proposed.

In contrast, the research on the LT model is much less. The above heuristic algorithms cannot be applied directly due to special features of the LT model.

To our knowledge, the algorithms in [2] and [4] are state of the art. In [2], Chen et al. observed that computing the spread is NP-hard in general graphs, while it costs linear time on directed acyclic graphs (DAGs). Based on that, they designed LDAG method which principle is similar with MIA in [1]. While [4] proposed an alternative algorithm SIMPATH that computes the spread by exploring simple paths in the neighborhood.

In recent years, a small number of research, as far as we know, only [13] and [9], introduce mechanism design into influence maximization problem. In [13], mechanism was designed to elicit individuals true costs of spreading information. However there is a obvious fault in the proof of budget control, which is the foundation of its other work. [9] designed so-called influencer model and influencer-influencee model to maximize message propagation. In these models, users are asked to report number of their friends, while in most social networks, this kind of information can be automatically and truthfully collected by the application itself.

Much progress has been obtained since the original greedy algorithm, but it still has large development space for research, especially for the LT model. We point out the following drawbacks from the brief summary of current work:

- Almost all the algorithms for the LT model set the nodes' threshold randomly. Moreover, greedy algorithms based on MC simulation assume the information of all nodes' threshold is perfectly known by the system. It is obviously unrealistic because many factors effect the user's threshold. Even to the same user, the value of threshold varies with different subject and different time. And the system will not be easy to get nodes' private information like value of its threshold.
- Traditional work does not take seed users' payment into consideration, instead, the number of seed users is always regarded as marketing budget. It just creates a gap between research and real situation, where budget of an advertising company is more related to how to pay the seed users.
- It is natural that a user with large influence, for example, a celebrity owning large number of followers, will be chosen as a seed node with a higher probability. But in real life, this celebrity may be not willing to spread advertisement information even offered by high payment. However current seed selection methods make decisions regardless of users' will.

1.2 Our Contribution

Our research is just centered around the above three problems.

To the first problem, we notice the necessity of reporting threshold, which is beneficial to both systems and users. Systems can do more accurate advertising according to users' private information, and users can show self preference through this action. Therefore we design a novel model that asks users to report their threshold and selects seed nodes on the basis of it.

To solve the second problem, we not only introduce the payment into the model, but also take incentive compatible condition into account. The payment

to the seed nodes is related to their reported threshold, so to maximize the interests of whole system, a truth-telling payment rule is of great importance.

The last problem is a dilemma. But system cannot force users being seeds to maximize the influence. So why not choose seeds from nodes that want to be. In our model, part of nodes are voluntarily to report their threshold and compete for being seeds.

We propose a seed selection algorithm called TREE TREE (Algorithm of **T**hreshold **R**eport and **E**valuation for **E**-marketing) for influence maximization problem, which contains seed selection and seed payment. Experiments show its good performance in practice.

2 Threshold Report Model

The novel model which asks nodes to report their threshold will be proposed in this section. Before detailed description of this model, we first explain the benefit of the system knowing nodes' threshold.

2.1 Behind "Threshold"

Granovetter was among the first to propose models to study the information propagation process through a social network. The LT model is generalized from his seminal paper [8], which laid the foundation of subsequent research on threshold-based model.

In [8], Granovetter pointed when people facing two mutually distinct and exclusive behavior alternatives, they have to make a binary decision. Deciding to do a thing or not to do depends on not only cost and benefit of this action but also many others make which choice.

We now take influence propagation in OSN as an example. When a user sees an advertisement of a skin care product in Twitter, she can forward this message and recommend to her followers or she can do nothing. We assume users are rational, that is, they choose actions to maximize their utility. There exists large individual differences, because for different individuals, cost of spreading the same ad varies and benefit which they expect to derive from the propagation is different too. "Threshold" is the concept to describe such variation among individuals. In this example, a user's "threshold" of spreading advertisement is defined as proportion of identity from her friends. Girls interested in cosmetics and skin care may hold a low threshold: it costs almost nothing for spreading such advertisement and they may even get extra bonus from the company. Some of them, who speak highly of this product after using, will be voluntary to recommend it. Then their threshold is near 0%. On the contrast, a boy who knows little about this will neglect this advertisement quickly even it has drawn much attention from his female friends. In this situation, his threshold is near 100%. Besides, celebrities are group of users who are very cautious about their remarks in OSN. Threshold on advertisement for them is very high, because they have to keep public images and take responsible for their recommend. In this

case, large cost of spreading advertisement always plays a more important role than high advertising income.

It is obvious in real situation, threshold varies considerably from person to person on the same object, on the other hand, one person holds different thresholds on variety things. Moreover, even for the same person and object, the value of threshold also changes with time. For example, people's ability of identifying rumors usually improves with age, which means the threshold of believing rumors increases as growing older. According to this diversity and uncertainty of threshold, we collect the users' information rather than make a baseless conjuncture. The above analysis has shown though "threshold" is a notation in the LT model, it measures the attitude of whether taking the same action when others do. In other words, it is a definition independent of the abstract model. Therefore practical applications can still ask users to report their psychological boundaries to the specific things.

2.2 Problem Formulation

After we explain the necessity of reporting threshold, the whole model can be given as following: The social network is represented by a weighted graph $G = (V, E, w)$. Here, V is the set of nodes in the social graph and $E \subseteq V \times V$ denotes the set of directed edges. The influence (weight) function $w : V \times V \rightarrow [0, 1]$ satisfies $\Sigma_{u \in V} w(u, v) \leq 1$. The system select a set of nodes S as initial seeds to spread the information after each node u reporting its threshold θ_u. Each seed node s is paid according to its performance and the threshold which it has reported before, denoted by $p_s = f(S, \theta_s)$. Given a budget B, the model's goal is to find a seed set S that maximizes $\sigma(S)$, which is the number of active nodes after propagation starting from S, in the condition of $\Sigma_{s \in S} p_s \leq B$.

Though this model takes threshold into consideration, this optimization problem is still the well-known NP-hard Max-k-Cover problem. It can also be easily proved that its influence function is monotone and submodular.

3 Proposed Method

To solve this problem, we need two steps: how to select seed nodes and how to pay for these seeds.

3.1 Seed Selection Rules

We first focus on seed selection methods to solve influence maximization problem in OSN. The Algorithm related to threshold, called TREE (Algorithm of **T**hreshold **R**eport and **E**valuation for **E**-marketing) is proposed.

Most existing influence maximization work based on LT model assume users' threshold information is given. While in a large scale social network, collecting threshold information of all users is impossible. But users' willingness of being seeds, which is sidestepped by most research work, shows significant importance. As analysis before, the selected nodes may not want to be seeds for various reasons. A more operable solution is choosing nodes from those intending to be seeds. Therefore we change the model that nodes enter the seed selection voluntarily. If node v wants to be a seed, it will be asked to report its threshold θ_v to the system. Based on it, we design a new sorting rule, which is only related to self threshold.

The threshold represents one user's wish of taking the same action with others. In the case of information propagation, high threshold means the user is not willing to spread this kind of message. It implies the advertising company has to pay more for this user's high cost of spread. Here threshold is also an indication of a node's cost. So this new measurement is the combination of one node's marginal contribution and its threshold. Then the greedy selection uses it sorting: the node selected in each round is the one that has the maximal weighted contribution given the previously chosen nodes. According to this ordering, the i^{th} seed node is:

$$\begin{aligned} s &\in argmax_{v \in N} \gamma_v \\ \gamma_v &= \frac{\sigma(S_{i-1} \cup v) - \sigma(S_{i-1})}{\theta_v} \end{aligned} \tag{1}$$

where S_{i-1} denotes the current seed set of $i-1$ selected nodes. This sorting rule a generalized version that used in greedy algorithm for submodular function maximization [10], and it guarantees constant factor approximation in [7] and [5].

3.2 Payment Rules

We then address the problem of seed nodes' payment, which mainly depends on seeds' performance. In this part, we discuss how different payment rules affect interests of both nodes and system based on the selection rule (1).

A. the Simple Rule. One of widely used payment rules, we call it *the simple rule*, only cares the seeds' marginal contribution. Under this rule, payment of i^{th} seed node v is

$$p_v^{sim} = k \times (\sigma(S_{i-1} \cup v) - \sigma(S_{i-1})) \tag{2}$$

$k > 0$ is constant value related to the budget. It is simple because the payment rule consists with the selection rule, and it means the system can calculate nodes' payment during the process of selection. While due to submodularity of influence spread function, one seed's payment varies with the order added in the seed set.

Another fatal disadvantage of this payment rule is that nodes may misreport its threshold to be seeds. Since all nodes involved hope to be seeds, and whether one node can be selected or not is related to threshold reported by itself, the

reality of elicited information can change the result significantly. The following example shown in Fig. 1 abstracted from a simple social network illustrates how false information harms the interests of the system.

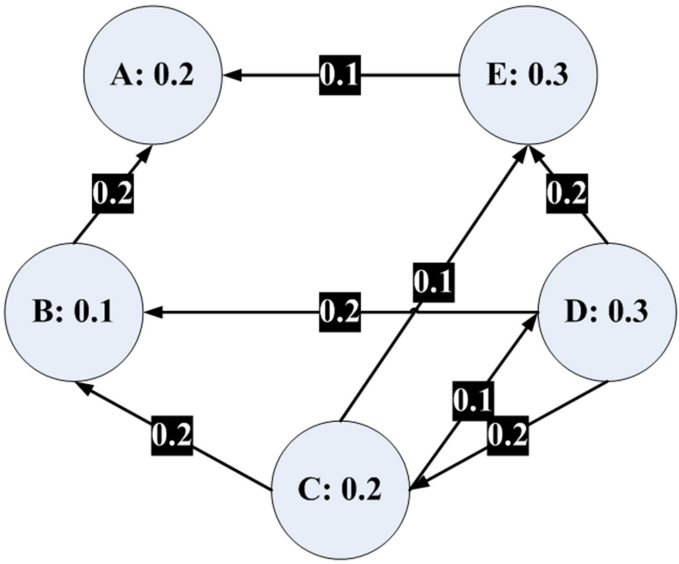

Fig. 1. A simple stylized social network.

In Fig. 1, there are five nodes A-E, and only one seed is needed. The weighted edges of this directed graph denote influence between nodes. True thresholds of nodes' are listed in the circle. It is obvious that node D will be chosen, because $\sigma(D) = 4$, which represents all other nodes can be activated by node D. If we choose node C, after two rounds, node B and A can be activated, thus $\sigma(C) = 2$.

Now all nodes are asked to report its threshold. If the system can collect true threshold information from nodes, then according to the selection rule 1,

$$\gamma_D = \frac{4}{0.3} > \frac{2}{0.2} = \gamma_C$$

So the seed should be node D, and payment is $p_D^{sim} = 4k$ and $p_C^{sim} = 0$. However if node C lies about its threshold as $\theta_C = 0.1$, then

$$\gamma_C = \frac{2}{0.1} > \frac{4}{0.3} = \gamma_D$$

Now node C will be the seed, its payment is $p_C^{sim} = 2k$. Node C increases its income, but this is definitely not the best choice for the system.

B. Mean Payment. In (1), the selection measurement is

$$\gamma_v = \frac{\sigma(S_{i-1} \cup v) - \sigma(S_{i-1})}{\theta_v} \tag{3}$$

which implies the system prefers the node with small threshold. While from the previous analysis, we see one node's threshold may be affected by its influence, in other words, large influence if one node will leads to its higher threshold. On the other side, one node's too low threshold over different messages makes it followers feeling it cannot stand for its position, so that too low threshold will also harm the influence of nodes.

Neither nodes with too low or too high threshold are the best choice of seeds. Therefore payment rules will punish these extreme threshold. Seed node v's payment is denoted by

$$\begin{aligned} p_v^{mean} = B \times \frac{Con(v)}{\sigma(S)} + \frac{\gamma}{2}(\frac{n-1}{n}(\theta_v - A_v) \\ - \sum_{u \in S, u \neq v} \frac{1}{n-2}(\theta_u - A_u)^2) \end{aligned} \tag{4}$$

where

$$\Sigma_{v \in S} Con(v) = \sigma(S)$$

$$A_v = \frac{1}{n-1} \Sigma_{u \in S, u \neq v} \theta_u$$

and n is the number of seeds. Variations of the punishment parameter, γ, changes the penalty that is imposed on an individual for deviating from the mean of other seeds' threshold.

Given θu, which $u \in S$ and $u \neq v$, node v reports θv to maximize the income$f(v)$, which equals to

$$\begin{aligned} f(v) = B \times \frac{Con(v)}{\sigma(S)} + \frac{\gamma}{2}(\frac{n-1}{n}(\theta_v - A_v) \\ - \sum_{u \in S, u \neq v} \frac{1}{n-2}(\theta_u - A_u)^2) \end{aligned} \tag{5}$$

The first order condition is

$$f'(v) = \frac{1}{n} + \frac{\gamma(n-1)}{n}(\theta_v - A_v) \tag{6}$$

A Nash equilibrium is ${\theta_v}_{v=1}^n$, such that for each node v, θ_v satisfies the above first order condition given ${\theta_u}_{u \neq v}$.

Theorem 1. *The Nash equilibrium 6 satisfies the Samuelson Condition.*

Proof. To see this, we simply sum the first order conditions over v, and obtain

$$\sum_{v=1}^{n} f'(v) = 1 + \frac{\gamma(n-1)}{n}(\theta_v - A_v) = 1 \tag{7}$$

which is the Samuelson efficiency condition.

Theorem 2. *The payment rules balance the budget both on and off the equilibrium path.*

Proof. We sum all seeds' payment, which is

$$\sum_{v=1}^{n} p_v^{mean} = B \sum_{v=1}^{n} \frac{Con(v)}{\sigma(S)} + G \tag{8}$$

where

$$G = \frac{\gamma(n-1)}{2n} \sum_{v=1}^{n} (\theta_v - A_i)^2 - \sum_{v=1}^{n} \sum_{u \neq v} \frac{1}{n-2} (\theta_u - A_i)^2 \tag{9}$$

It is trivial to get the calculation result that $G = 0$. And according to

$$\Sigma_{v \in S} Con(v) = \sigma(S)$$

We get

$$\sum_{v=1}^{n} p_v^{mean} = B$$

Actually, the budget balance is achieved by the last term in the payment rule, the squared standard error of the mean of others' information.

4 Experiment

After the introduction of the new algorithm, we are now interested in understanding its behavior in practice, and comparing its performance with other methods.

4.1 Experiment Setup

Some preparation of experiments are listed as following, including network data, other methods used to compare with and threshold setting.

Network Data: We use cit-HepTh, arxiv HEP-TH (high energy physics theory) citation network, as experiment data. This graph is from the e-print arXiv and covers all the citations within a dataset of 27,770 papers with 352,807 edges.

Comparison Algorithms: As discussed above, we transplant TREE into High Degree and Page Rank algorithm. Here we compare information propagation range and runtime of different methods, including Random Selection, High

Degree, Page Rank, TREE on High Degree (High TREE) and TREE on Page Rank (Page TREE).

Threshold Setting: The key factor in LT model is how to set users' threshold, which is beyond this paper's topic. But we observe from realworld dataset that one user's threshold maybe relates to the number of his followers or the number of his following people, which is this node's out degree and in degree in network graph. A node with large out degree always hold a high threshold, because this user will be more cautious about his views and comments than common people. On the hand, a node with large in degree also has a high threshold, and the reason is that they can receive different opinions on the same thing, so the probability of being persuaded easily for him is low. Thus we set three kinds of nodes' threshold: Random Threshold, Out Degree Threshold and In Degree Threshold. Due to $\theta \in [0, 1]$, we normalize the node v's out degree and in degree, denoted by:

$$\theta_v^{out} = \frac{out_v - out_{min}}{out_{max} - out_{min}} \tag{10}$$

$$\theta_v^{in} = \frac{in_v - in_{min}}{in_{max} - in_{min}} \tag{11}$$

4.2 Results and Analysis

In the experiments, algorithms' performance lies in the information propagation range and running time. The result is the mean value of 10 times' computation of using one of algorithms in a specific network.

Figure 2 shows different algorithms' performance with three kinds of threshold setting. Naturally, all other algorithms exceed the baseline method, Random Selection. Besides this, the other common conclusion from these three figures is that TREE enlarges the information propagation range in all conditions. It is obvious that using PageTREE (or HighTREE) achieves better results than Page Rank (or High Degree).

We also notice the differences between three figures tell us more information about TREE. Comparing with Random Threshold (shown in (a)) and In Degree Threshold (shown in (c)), TREE in Out Degree Threshold (shown in (b)) does not have evident superiority, especially HighTREE. This is mainly because when node's threshold is close related to its out degree, the two measurements in different algorithms represent the same character of the node. In this situation, the combination of two methods is of lesser significance.

The mean value of algorithms' runtime of selecting 100 seeds is presented in Fig. 3. It is rational that TREE needs more time because it adds another sorting measurement, and more computation is inevitable. However, the time consumption increases slowly. Therefore we think it is still worthy of enlarging the propagation range at the cost of runtime slowdowns.

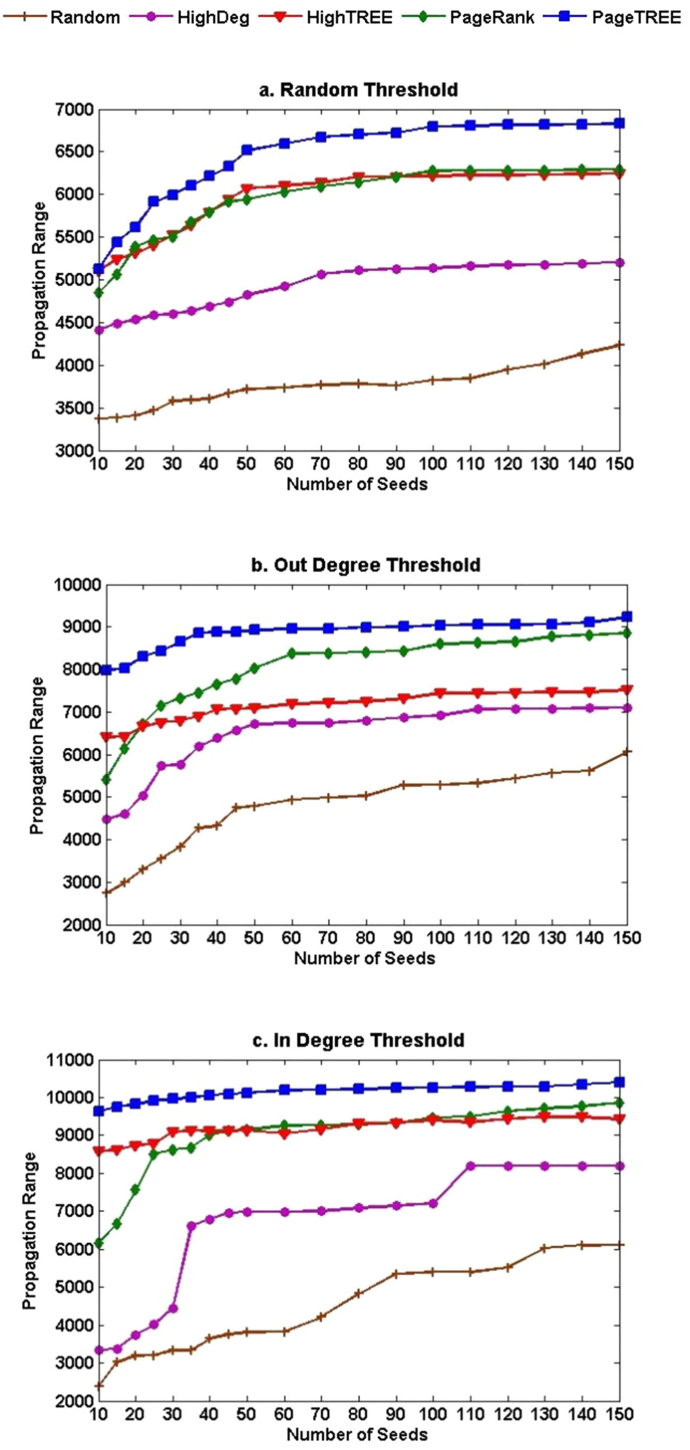

Fig. 2. Propagation range comparison in cit-HepTh.

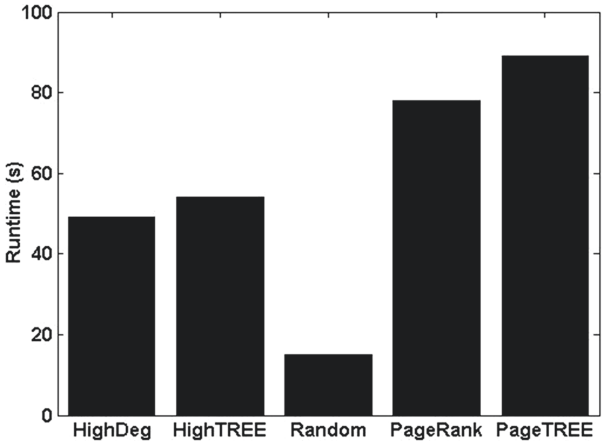

Fig. 3. Runtime comparison in cit-HepTh.

5 Conclusion

In this paper, based on LT model, we propose a new model in which users are voluntarily to report their private information - threshold to be viral marketing seeds. The method, called TREE, is designed to solve this model. Influence maximization problem and mechanism design are involved in the whole process. Experiments verify the good performance of TREE.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (NSFC, project 61902152, 61972182) and the Natural Science Foundation of Jiangsu Province (BK20180600).

References

1. 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)
2. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 88–97. IEEE (2010)
3. 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, pp. 137–146 (2003)
4. Goyal, A., Lu, W., Lakshmanan, L.V.: Simpath: an efficient algorithm for influence maximization under the linear threshold model. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 211–220. IEEE (2011)
5. Leskovec, J., Krause, A., Guestrin, C.et al.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429 (2007)

6. Kimura, M., Saito, K.: Tractable models for information diffusion in social networks. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 259–271. Springer, Heidelberg (2006). https://doi.org/10.1007/11871637_27
7. Krause, A., Guestrin, C.: A note on the budgeted maximization of submodular functions (2005)
8. Granovetter, M.: Threshold models of collective behavior. Am. J. Sociolo. **83**(6), 1420–1443 (1978)
9. Mohite, M., Narahari, Y.: Incentive compatible influence maximization in social networks and application to viral marketing. arXiv preprint arXiv:1102.0918 (2011)
10. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions. Math. Program. **14**(1), 265–294 (1978)
11. P. Domingos, M. Richardson: Mining the network value of customers. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66 (2001)
12. 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 (2009)
13. Singer,Y.: How to win friends and influence people, truthfully: in-fluence maximization mechanisms for social networks. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, pp. 733–742 (2012)