Journal of Computers Vol. 32 No. 4, 2021, pp. 187-200
doi:10.53106/199115992021083204015

A Maximizing Influence of Multiple Nodes Propagation Algorithm Based on Optimal Neighbor Discovery

Yun Liu[1], Ling Sun[1], Fei Xiong[1*], Junjun Cheng[2]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, 100044, China
{liuyun, 12111031, xiongf}@bjtu.edu.cn

[2] China Information Technology Security Evaluation Center, Beijing, 100085, China
chengjj@itsec.gov.cn

Received 1 June 2021; Revised 1 July 2021; Accepted 2 August 2021

Abstract. Individual influence is one of the important factors of the evolution of information dissemination in social networks. Usually the ranking of node influence reflects the ability of single node to diffuse information. And when a set of multiple nodes distributed at each location of the network is propagated as a propagation source of the information, the sum of the influence of the plurality of nodes is no longer the addition of the influence of each node. We propose a new influence maximization algorithm named Max Neighbor Heuristic (MNH) combining Greedy Algorithm with Heuristic Algorithm. In order to balance the accuracy and efficiency in MNH, we considered the strategy of optimal neighbor discovery. By comparing the efficiency and accuracy of the algorithm with the three classical algorithms, we find that MNH algorithm has obvious advantages. Although the performance of MNH algorithm has a little bit of volatility, it has certain advantages in the average time-consuming and precision of the algorithm, which provides a new method and idea to solve the problem of maximizing influence of multi-node communication in social network

Keywords: individual influence, information dissemination, social network

1 Introduction

The analysis of node influence in social network is an important research content in the study of complex network. In the specific network topology environment, the problem of influence ranking has been a more comprehensive study. For different network structures and different algorithm complexity requirements, the existing various algorithms have their own strengths [1-15]. However, in the actual production and life process, there has another more practical problem with node impact applications which is the influence maximization spreading problem. The problem is the derived from the viral marketing in the economics [16-17].Viral marketing refers the product promotion which only rely on a small number of people (Not single individual) in the early stages of product marketing, then re-use of these initial users to create the word of mouth effect to achieve widely spreading. Therefore, in order to maximize the effectiveness of publicity, how to select a limited number of people, and allow them to spread in different regions has become the problem. The issue of influence maximization spreading is one of the important research questions in the field of social network information dissemination. It is designed to discover the most informative set of nodes in a social network, which has a wide range of applications with high research and application value in the marketing, advertising, public opinion monitoring and other fields [18-20].

David firstly defined the problem of the maximization of influence as a discrete optimization problem [21]. It is proved that the propagation influence function $\sigma(\cdot)$ satisfies the submodular characteristics in both the independent cascade model and the linear threshold model, which are both NP-Hard problems.

* Corresponding Author

Based on this, a hill-climbing greedy algorithm is proposed, referred to as BasicGreedy algorithm. In 2007, Jure Leskovee proposed the CELF (Cost-Effective Lazy Forward) algorithm [22], and a preference mechanism is proposed based on the BasicGreedy algorithm, which does not calculate the influence gain of node *v*. As a result, it reduces the impact of each round of the marginal value of marginal revenue calculation. After that, Amit further proposed the CELF++ algorithm by improving the storage structure of the way, which improves the efficiency [23]. At the same time, Wei Chen proposed a New Greedy algorithm based on the BasicGreedy algorithm which reduces the number of simulations to *n* times [24-25].Yu Wang proposed a Community-Based Greedy Algorithm (CGA) based on the community characteristics of social networks [26], using the most influential node within the community to approximate the most influential users of the global network. Experimental results show the speed of the CGA algorithm has been significantly improved, but the accuracy of the same significant decline. The above algorithms are all belong to greedy algorithm that each step is to seek to produce the maximum impact of the spread of the scope of the node, and calculate the current local optimal solution based on the previous step. The advantage of greedy algorithm is that the algorithm has high precision which can approximate the optimal solution with the probability of $1-1/e$, but the disadvantages are also very obvious that large amount of computation time is required.

Although the efficiency of a large number of improved greedy algorithms has been significantly improved, it still cannot be applied to large-scale social network because of its high algorithm complexity. In order to pursue practicality and high efficiency, many heuristic algorithms have been proposed by researchers. The simplest heuristic algorithm is proposed by David Kempe [21], including heuristic algorithms based on Degree and Centrality. Wei Chen considered the overlap between node influence factors, and presented the Degree Discount algorithm for the independent cascade model [24]. The main idea is that if node *v* and its neighbor node *u* are simultaneously selected into the initial active node set, the influence of node v needs to be deducted. The experimental results show that the accuracy of the algorithm has been greatly improved after considering the node overlap factor, but the accuracy of the greedy algorithm is still not achieved. After that, Wei Chen proves that the problem of calculating the influence range of a given initial active node set are both essentially an NP-Hard problem in the independent cascade model and the linear threshold model, which means that the greedy class algorithm is fundamentally bound to lead to greater computational complexity. Besides, for the shortcoming of LDAG heuristic algorithm, Amit Goyal proposed SIMPATH heuristic algorithm [27-28], which can estimate the influence value of node more accurately by calculating the simple path of seed node. Kyomin Jung proposes IRIE heuristic algorithm in independent cascade model, which does not need to calculate the influence value of node, but based on trust propagation method. By combining influence ranking and influence estimation (Influence Ranking), the efficiency is obviously improved than the PMIA algorithm [29]. It is not difficult to find that a large number of heuristic algorithms improve the efficiency of the algorithm by sacrificing the precision of the algorithm [30-33].The deep reason is that the heuristic method always has an approximate estimation process. Therefore, the heuristic algorithm has the characteristics of low complexity and short operation time, but the accuracy of the heuristic algorithm is much lower than that of the greedy algorithm.

With the rapid development of network technology, the evolution speed of on-line social network is obviously faster than before, and the relationship between nodes is influenced by more factors, which makes it change more frequently and adds a lot of uncertainty to the prediction based on network topology. Regarding the issue above, a new influence maximization spreading algorithm based on optimal neighbor discovery is proposed in this paper. The strategy of optimal neighbor discovery is adapted to find a random large influence node. Based on the random node, the speed time of Greedy Algorithm can greatly reduce in our algorithm.

2 Influence Propagation Maximization Algorithm

2.1 Definition of the Problem

Suppose a network $G=(V,E)$ with $|V|$ nodes and $|E|$ links. The constant $s<|V|$ represents the number of initial propagation nodes that need to be picked out from the network *G*. Any *s* nodes constitute the set $S=\{v_i \mid i\in[1,s]\}$ and propagate the information according to the given rule. At the end of the

propagation, $T(v_i)$ is the set of other nodes activated by v_i and $T(S)=T(v_1)\bigcup T(v_2)...\bigcup T(v_s)$. The problem of node influence maximization in network G can be defined as finding a set as following:

$$T(S_{\max})\geq T(S),\ |S_{\max}|=|S|=s \quad \textbf{(1)}$$

On the surface, the impact of maximizing the problem can be simply understood as the selection of multiple users with different influence to produce the largest diffusion of information. In fact, this problem is a NP (Non-deterministic Polynomial) difficult problem which is uncertainty in solving polynomials, and it is impossible to calculate the optimal solution quickly using the unified algorithm [34]. In contrast, it can quickly verify the answer if it has already been known. Because there is no quick solution to the direct solution except for the solution of violence, the influence maximization problem needs to balance the efficiency and approximate.

2.2 Max Neighbor Heuristic

The basic idea of MNH (Max Neighbor Heuristic) is to transform the multiple-node selection problem into probabilistic problem. By the heuristic method of randomly selecting, it can obtain to obtain a much smaller set of nodes than the original network size with less computation. Then, by using the local greedy algorithm, the relative optimal node set is obtained to approximate the optimal node set of the original network. MNH algorithm is divided into two parts:

Random heuristic process: Firstly, initialize the relative optimal node set V and randomly select a node u from the global scope. The all first-order neighbors of node u can be obtained which defined as v and select the largest degree node add into V. Then, the above random selection process is repeated until ηs different initial propagation nodes are picked, $\eta\in[1, N/s]$ is the size of candidate set.

Partial greedy process: Firstly, initialize the relative optimal node set S. According to the specific pre-set information dissemination model, count the average dissemination range of each node in set V in the spread of the spread of the range of nodes. Select the node which can create maximum spreading range to join the global optimal node set S. Then, traverse the remaining $\eta s-1$ nodes and calculate the marginal gain of each node. On the basis of the current set S, the marginal gain refers to the increase in the influence of adding an additional node v_i. Define $\sigma(\cdot)$ as an influence function, and the marginal gain created by v_i can be obtained as following:

$$\sigma_{v_i}(S)=\sigma(S\bigcup\{v_i\})-\sigma(S) \quad \textbf{(2)}$$

Finally, add the nodes that can produce the maximum marginal gain to S and remove it from the set V. This process is repeated $s-1$ times until s initialized propagation nodes are picked out from V.

```
Max Neighbor Heuristic
  initialize S=∅, V=∅
  while len(S)<ηs
       random pick vertex u∈G\s
      select u=MaxDegree{neighbor(u)}
  S=S⋃u
  for i=1 to s do
       for any node v in set G do
           s_v=0
           for i=1 to R do
              s_v=s_v+|RanCas(V∪{v})|
           end for
           s_v=s_v/R
       end for
       v=arg max_{v∈V}{s_v}
       V=V∪{v}
   end for
   output V.
```

From the logic above, we can find that the first part of MNH algorithm is the process of random selection of candidate seeds. Based on the random heuristic, it chooses the largest neighbor as the candidate seed, which avoids the huge computational complexity in large-scale network. The second part is the selection process of candidate seeds. The classical greedy algorithm with high precision is used to compute the marginal revenue one by one, so as to obtain a node set with the greatest influence.

According to the distribution characteristics of online social network nodes, the distribution of node degrees follows a power-law distribution [35-36]. Cohen's study indicated that the probability of a node with degree k is $p(k)$ in scale-free n-scale networks [37].

$$p(k) \sim ck^{-r} \tag{3}$$

Assume that the minimum and maximum node degrees are $k_{\min}$ and $k_{\max}$. Distribution parameter c satisfies the normalization condition $\int_{k_{\min}}^{\infty} p(k)dk = 1$, which it can be calculated $c = (\gamma - 1)k_{\min}^{\gamma-1}$. As the number of nodes in the top-s of the node degree is the smallest node degree in the top-s node $k_{\text{top-s}}$, we can see:

$$\int_{k_{\text{top-s}}}^{\infty} p(k)dk = \frac{s}{n} \tag{4}$$

Bring the normalization parameter $c = (\gamma - 1)k_{\min}^{\gamma-1}$ into Eqs.4:

$$\begin{aligned} \int_{k_{\text{top-s}}}^{\infty} p(k)dk &= \int_{k_{\text{top-s}}}^{\infty} ck^{-\gamma}dk \\ &= \int_{k_{\text{top-s}}}^{\infty} (\gamma - 1)k_{\min}^{\gamma-1}k^{-\gamma}dk \\ &= (\gamma - 1)k_{\min}^{\gamma-1}\int_{k_{\text{top-s}}}^{\infty} k^{-\gamma}dk \end{aligned} \tag{5}$$

Combining the results of Eqs.4 and Eqs.5, we can get the minimum degree of node satisfying top-s:

$$k_{\text{top-s}} = k_{\min}(\frac{n}{s})^{\frac{1}{\gamma-1}} \tag{6}$$

When $s = 1$, the degree of the largest node in the network can be obtained:

$$k_{\text{top-1}} = k_{\max} = k_{\min}(n)^{\frac{1}{\gamma-1}} \tag{7}$$

Since the MNH algorithm is aimed on the node with the highest degree of the neighbor, it is assumed that a node v is selected at random and the node of the first degree of the node is u and the node degree is k. The probability that node u is randomly selected as an independent node is $p(k)$. If the node u is selected based on the maximum neighbor selection rule v is selected first and then the node u is selected as a neighbor), the probability that node u is selected will be $kp(k)$, which is equivalent to selecting a node from an edge. It can be seen that, according to the random selection method in the MNH algorithm, the nodes with larger node degrees are more likely to be randomly selected. Therefore, in a network that the maximum and minimum node degrees is $k_{\max}$ and $k_{\min}$, the probability of selecting the degree k from the neighbors of any node can be:

$$p_l(k) = kp(k) = c_l k^{1-r} \tag{8}$$

c is normalization parameter $c_l = \left(k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}\right)^{\gamma-2}$ and $\int_{k_{\min}}^{k_{\max}} c_l k^{1-\gamma}dk = 1$. After random selecting a node q, define $p_{\text{top-s}}$ is the probability that the maximum degree neighbor of q belongs to the top-s node set. $p_{\text{top-s}}$ is equal to the integral of the probability distribution for all nodes whose degree is greater than

$k_{\text{top-s}}$ is selected.

$$\begin{aligned} p_{\text{top-s}} &= \int_{k_{\text{top-s}}}^{k_{\max}} kp(k)dk \\ &= \int_{k_{\text{top-s}}}^{k_{\max}} c_l k^{1-\gamma} dk \\ &= \frac{k_{\max}^{2-\gamma} - k_{\text{top-s}}^{2-\gamma}}{k_{\max}^{2-\gamma} - k_{\min}^{2-\gamma}} \end{aligned} \tag{9}$$

According to the rules of multiple selection nodes in MNH algorithm, the single-node case in (9) can be extended to ηs node cases, which is the probability of randomly selecting ηs nodes that contain top-s nodes. In the case that the randomly selected node set ηs is far smaller than the network size n, the reduction of the candidate set n after the node is selected can be neglected. Based on the permutation and combination, the probability $p_{\text{top-s}}$ is as following:

$$P = \binom{\eta s}{N} \left(p_{\text{top-s}}\right)^{\eta s} \left(1 - p_{\text{top-s}}\right)^{N-\eta s} \tag{10}$$

Bring Eqs. (6)(7)(9) into Eqs. (10), the final probability function can be calculated. In order to observe the effect of candidate set coefficient η on the solution results, we use a scale-free random network with $N = 1500$ and $s = 50$ to observe the accuracy of MNH algorithm under different η, and the result is shown in Fig. 1.

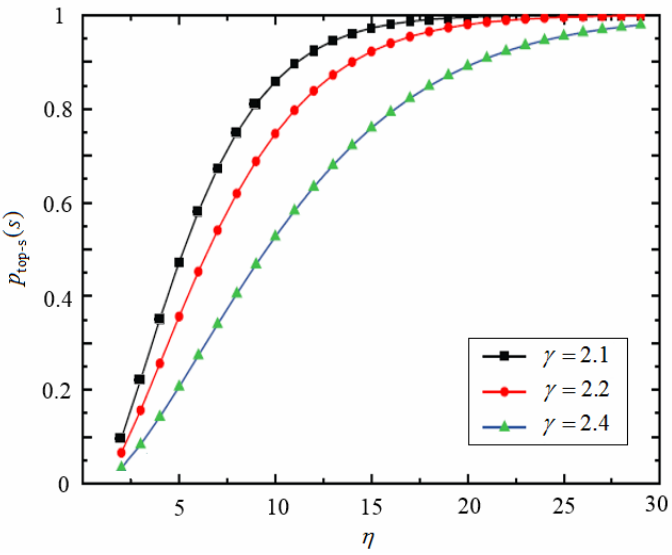

Fig. 1. The effect of Candidate set coefficient η and power exponent γ on the accuracy of MNH algorithm

It can be seen that as the candidate set coefficient η increases, the accuracy of MNH is gradually increasing. At the same time, with the reducing of power index, the curve in Fig. 1 is steeper, and the algorithm is also accurate. In the line of $\gamma = 2.1$, when $\eta > 20$, the accuracy of the algorithm is close 1. The social network power exponent γ actually reflects the heterogeneity of the node's degree. The larger γ, the heterogeneity of node degree is stronger and. Therefore, we can consider that the MNH algorithm has stronger practicability in the network with large node heterogeneity.

The complexity of the MNH algorithm is equal to the sum of the random heuristic process and the local greedy process. In the random heuristic process, the candidate node set size is ηs , and the neighbor degree of each nodes need to be counted. According to the probability density distribution of social network nodes, we can obtain that the complexity of a single node is:

$$\langle k \rangle = \sum_{1}^{n} kp(k) = c\sum_{1}^{n} \frac{1}{k^{\gamma-1}} < c\sum_{1}^{n} \frac{1}{k} = c\ln(n) \tag{11}$$

Thus, the computational complexity of the heuristic process with ηs node is $O(\eta s \log(n))$. On the other hand, as the time complexity of the global greedy algorithm with $n=N$ is $O(sNRm)$ [24], the complexity of the algorithm is greatly reduced when $n = \eta s$, which would be $O(\eta s^2 Rm)$. Combined with the complexity of the above two processes, the time complexity of MNH algorithm is obtained. The comparison results with other three kinds of classical algorithm are shown in Table 1 with N node and m edge. s is the number of nodes that need to be acquired. The average number of cycles is R.

Table 1. The time complexity and type of four kinds of algorithm

Algorithm	Time complexity	Type
MNH	$O(\eta s \log(n) + \eta s^2 Rm)$	Hybrid
DegreeHeuristic	$O(m)$	Heuristic
BasicGreedy	$O(sNRm)$	Greedy
NewGreedy	$O(sRm)$	Greedy

When the candidate set factor η is small, $1 \le \eta s << N$, MNH algorithm complexity is much smaller than Basic Greedy algorithm, but greater than DegreeHeuristic algorithm complexity, the closest NewGreedy algorithm. While ηs is large and close to N. The algorithm complexity of MHN is basically the same as that of the BasicGreedy. The above characteristics reflect the flexibility of MNH algorithm, which the appropriate candidate factor η can be selected according to the specific requirements of timeliness and accuracy.

3 Simulation Comparison and Analysis

Due to the different topological features of online social networks, the solution to the problem of maximizing influence is dependent on the network topology, which the accuracy and operation speed of the same algorithm in different network topologies are quite different. The measurement of the final influence scope depends on the propagation model. Experiments were carried out in two representative propagation models: the linear threshold model [38] and the independent cascade model [39], which the real network data is selected as the experimental data set in our simulations.

3.1 Data set

Twitter relationship data: Twitter is a US social networking and microblogging service site, is the world's top ten sites on the Internet one of the visits. It is a typical micro-blog application. Users connected with each other through mutual attention, which they can receive the first time users are concerned about the tweet. If we only observe the twitter network structure from the relationship among the users, when the two users are concerned about each other, we can consider a connection between the two nodes. When any user sends the information, influences will be generated to other users.

Friendfeed relationship data: Friendfeed is a Web site for aggregating Web 2.0 services. It can aggregate various social networking services (such as Blog, Twitter, Del.icio, Digg, Flickr, etc.). Based on its services, we can easily know published information of friends in various social networks. Therefore, Friendfeed user relationships constitute a network to more fully reflect the user in the network world and the interaction between other users. The statistics are shown in Table 2.

Table 2. Basic statistical characteristics of network datasets

Data name	V	$\lvert E \rvert$	$\langle k \rangle$	$k_{\max}$	$k_{\min}$	γ
Twitter	4053	69436	19.34	329	1	2.21
Friendfeed	6244	127491	58.21	221	1	2.33

V indicates the number of nodes. The total number of edges is $|E|$. The average degree, maximum and minimum degree is $\langle k \rangle$, $k_{\max}$ and $k_{\min}$. γ is the power exponent.

3.2 Simulation Analysis in Linear Threshold Model

A random function is used to assign the impact propagation thresholds randomly to all nodes in the dataset. The impact propagation threshold is defined as $\theta(v) \in [0,1]$, which reflects critical point on choosing to put forward the information after being affected by it. We remove non-connected nodes and the nodes with $\theta(v) = 0$, which provides the connectivity foundation for the information dissemination of seed nodes. Influence weight of node v affected by neighbors *i* is defined as $b_{i,v}$:

$$\sum_{i=0}^{k_v} b_{i,v} = 0 \tag{12}$$

k_v is the degree of *v*. The edge assignments for each node's neighbors are affected by the random function, and the sum of weights is 1. Actual influence value I_v is defined as affect from the any active node around *v*:

$$I_v = \sum_{i \in N'} p_{i,v} b_{i,v} \tag{13}$$

N' is the active node set of *v* neighbor. When the node *v* in the inactive state is affected by $I_v > \theta(v)$ at time *t*, the node *v* enters the active state and has an effect on its neighbor as an active node at $t+1$. The experiment is then repeated until the new active node is no longer created in the network. In order to distinguish the importance of nodes, and to ensure that the diffusion time between the algorithms are comparable, the probability of information diffusion here is a unified mean value $p_{i,v} = p = 0.01$.

In the MNH, BasicGreedy and NewGreedy algorithms, the above-mentioned models are used to calculate the influence marginal gain of nodes. The initial seed node is the active node and the remaining nodes are the inactive nodes. In the Twitter data set, the maximum number of influence nodes $s = 10$ is presented as an example, which the results of the four algorithms in the Twitter network are shown in Table 3.

Table 3. The top 10 influence nodes result from four algorithms in the linear threshold model with Twitter data set

MNH	DegreeHeuristic	BasicGreedy	NewGreedy
432	432	2241	2241
56	727	432	432
2241	3088	1801	1231
762	530	56	1801
1801	96	121	2332
530	442	2332	1691
24	132	3231	56
181	1722	762	181
2332	587	181	24
3231	1532	587	3231

U_R, U_D, U_B and U_N are defined as the maximum impact node sets of four kinds of algorithms. From Table 3 we can be found that the results obtained by BasicGreedy and NewGreedy have a highest coincidence. Seven of the ten nodes are the same, $U_B \bigcap U_N = \{2241, 432, 1801, 2332, 3231, 181, 56\}$. The intersection of MNH algorithm and BasicGreedy algorithm is $U_R \bigcap U_N = \{432, 56, 762, 1801, 181, 2332, 3231\}$, which indicate that if we consider the BasicGreedy as a benchmark, we can initially determine the accuracy of MNH is close to NewGreedy. Besides, The intersection of BasicGreedy and DegreeHeuristic is less, $U_D \bigcap U_B = \{432, 587\}$, which tell us that the heuristic algorithm and the greedy

algorithm are obviously different in solving the accuracy. Based on the comparison between MNH and DegreeHeuristic, $U_D \bigcap U_B = \{432, 530\}$, it can be considered that MNH algorithm, which is a hybrid heuristic method, is more biased towards the result of greedy algorithm in accuracy.

Based on the conclusions above, we use the propagation model to further validate the results of the MNH algorithm. As the result of the analysis in Section 2.2, we get the conclusion that the probability of the MNH algorithm approach to the optimal solution will increase when the number of initial seeds get larger. Therefore, for the two types of data sets, respectively, the experiments with different size sets are performed, $1 \le S \le 60$. We use all the nodes in U at the same time as the active node to propagate in the early stage of diffusion, and record the time elapsed from the beginning to the end of the information diffusion and final diffusion range.

Fig. 2 shows the actual propagation range that can be generated by four influence-maximizing algorithms under the Friendfeed data set. It can be seen that with the increasing of initial seed nodes, the range of maximization influence produced by the four set shows a positive increase. The BasicGreedy algorithm curve is higher than the other three curves, and the growth is relatively flat when $s > 20$. This indicates us that the greedy algorithm adopts the actual propagation range as the reference value in the selection of the first initial seed node, which makes the result more accurate. In contrast, MNH and DegreeHeuristic only rely on node degree as the criterion of node influence, ignoring the topological position and pivotal effect of nodes in the whole network, so that the final results deviate from the true value.

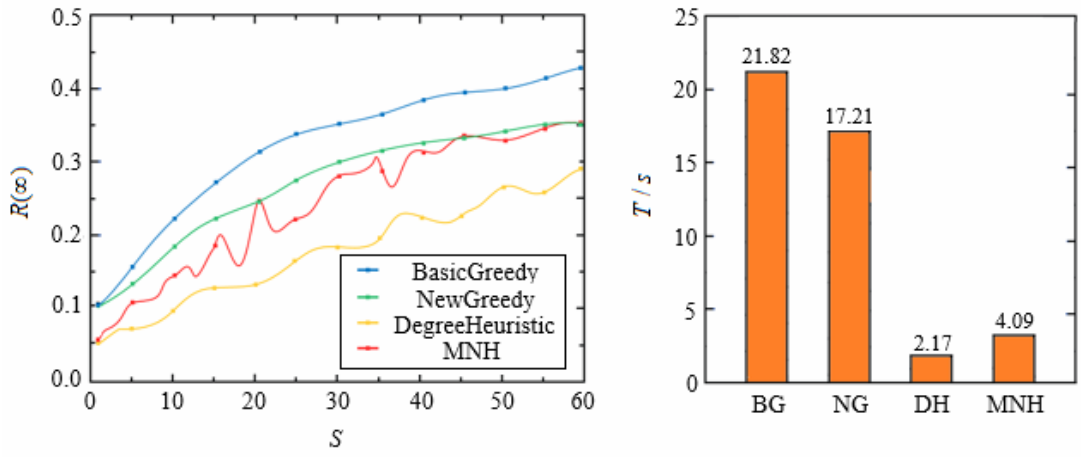

Fig. 2. The simulation results of four algorithms in the linear threshold model in Friendfeed network

On the other hand, MNH curve has obvious fluctuation which the range is basically between NewGreedy and DegreeHeuristic curves. This indicates that the randomness of the MNH algorithm in the selection of candidate seed sets gives the instability of the final influence range. With the increasing of s, the fluctuations gradually slow down. This phenomenon is consistent with previous accuracy analysis results on MNH algorithm. At the same time, it is obvious that the increase of the MNH curve is highly related to the initial seeds numbers, and basically coincides with the NewGreedy curve after $s > 45$. However, there is still about 7% gap of the spread range between MNH and BasicGreedy. It can be assumed that such gap is obtained by sacrificing the complexity of algorithm. With the number of initial propagation nodes increases, the gap will become smaller and smaller. In the comparison of computation time, the whole computation time of MNH algorithm is only $1/5$ of NewGreedy algorithm, and $1/2$ for the DegreeHeuristic algorithm. Therefore, from a practical perspective, MNH algorithm has the best performance balance time-consuming and accurate.

It can be seen in the experimental results (Fig. 3) of in the twitter data set that the gap between the four curves are significantly reduced, and the random fluctuation amplitude of MNH also narrowed significantly. But, the overall trend remains consistent with the Friendfeed data set. When the initial seed number is small, four final influence ranges are at a relatively low level. Within $s < 20$, final influence range R increases rapidly and has a positive correlation with s. This is due to the heterogeneity of the nodes in the Twitter data set, $\gamma_T < \gamma_F$. As a result, the local community concentration phenomenon in the

whole network is more prominent. When the initial number of seed nodes is small, the range of influence spread is limited to local communities. For the effective coverage of the global network, it is necessary to distribute more initial propagation nodes. When $s > 30$, four curves slowdown significantly, where MNH surpasses NewGreedy and gradually close to BasicGreedy around $s = 40$. It can be seen that the accuracy of the MNH algorithm on the Twitter data set is better than on the Friendfeed data set. The experimental results show that the MNH algorithm has higher heuristic in the network with low power exponent. The main reason is that the process of randomly selecting initial seeds in MNH algorithm effectively avoids the obvious topology of the local community concentration and makes the distribution of initial seeds more dispersive.

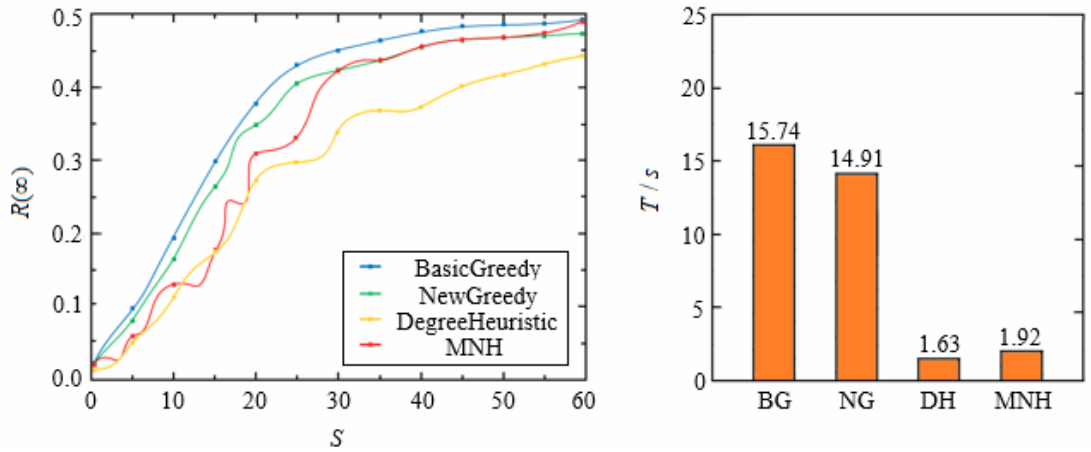

Fig. 3. The simulation results of four algorithms in the linear threshold model in Twitter network

3.3 Simulation Analysis in Independent Cascade Model

S is set as the initial propagated seed node set. All nodes are activated $t = 0$ and have an impact on the neighboring nodes. When $t = t_0$, if v is an active node, any neighbor w of node v will be activated at time with probability $p_{v,w}$ and turn into activated node. Different from the linear threshold model, the probability $p_{v,w}$ is the influence attribute of the node v itself, the higher the value $p_{v,w}$, the greater the influence of node v. In order to ensure that activate the node can be infect at least one neighbor even if the number of activate the nodes is at a low degree. The node activation probability is uniformly set to constant $p_{v,w} = p = 0.05$.

At any time $t = t_0$, when there are multiple active nodes around the node w at the same time, the active node v_i is selected in a random order, and activates the node w with a probability of $p_{v,w}$ at $t = t_0 + 1$. It can be considered that in an independent cascade model, the probability that any node is activated by its neighbors is independent of each other and the probability that the node is successfully activated depends on the number of active nodes in the neighbor. Thus, starting from the first active node appearing in the neighbor of node w until w is successfully activated, the probability is:

$$1 - \prod_{round\ n} \prod_{i=0}^{k_n} (1 - p_{v_i,w}) \tag{14}$$

n is the round number, and k_n is the number of active neighbors in each round of activation event. After the active node tries to activate all neighbor nodes, it will lose has the ability to activate other nodes and change to an inactive node. The entire propagation process ends after the activation node does not exist in the network. The number of inactive nodes reflects the size of the influence range of the initial seed node set S.

Based on the above experimental model, the maximum number of influential nodes $s = 10$ in Twitter data set is taken as an example. The results of the four algorithms are shown in Table 4.

Table 4. The top 10 influence nodes result from four algorithms in the linear threshold model with Twitter data set

MNH	DegreeHeuristic	BasicGreedy	NewGreedy
2241	432	2241	2241
522	727	432	1753
432	3088	1421	157
762	530	1753	1801
991	96	762	2332
24	442	372	1421
2028	132	2332	56
181	1722	522	612
2332	587	612	291
56	1532	157	429

U_R, U_D, U_B and U_N are defined as the maximum impact node sets of four kinds of algorithms. Comparison of Table 3 and Table 4, it can be found that there is a certain difference of the selected results under the two models. Taking BasicGreedy as an example, the intersection of the solution is $U_B^{'} \bigcap U_B = \{2241, 432, 762, 2332\}$, only four nodes are the same. NewGreedy also has only four nodes belong to the largest impact nodes. However, the result of the DegreeHeuristic has no change in the two models. This is not difficult to understand, because the degree of heuristic algorithm to the node level of the evaluation of the impact of the node itself, and ignore the overlap of the influence of nodes, so no matter how the dissemination of information in the network diffusion, the node itself The topology feature will not change, so the DegreeHeuristic algorithm results will not change. This is not difficult to understand, because the degree of heuristic algorithm is based on the node degree to evaluate the impact of nodes, and ignore the overlap of the influence of nodes. As a result, no matter how the dissemination of information in the network diffusion, the topology of the node itself will not change, so that the result of the DegreeHeuristic algorithm will not change. The result of MNH algorithm has a large coincidence part under the two propagation models, $U_R^{'} \bigcap U_R = \{432, 2241, 762, 181, 2332, 24, 56\}$. However, some new nodes are selected, such as node 991, node 2028. Compared with the other algorithms, we can see that these nodes are not the solutions of other algorithms. This is because the MNH algorithm randomly selects 2*s* candidate propagating seed nodes, not only follow the principle of maximum neighbor, but also have a certain randomness.

For each of these changes, a set of different maximum impact nodes is taken $1 \le s \le 60$. To observe the accuracy of the MNH algorithm under the independent cascade model, the further validation of Twitter and Friendfeed data sets are conducted, and the results are shown in Fig. 4 and Fig. 5.

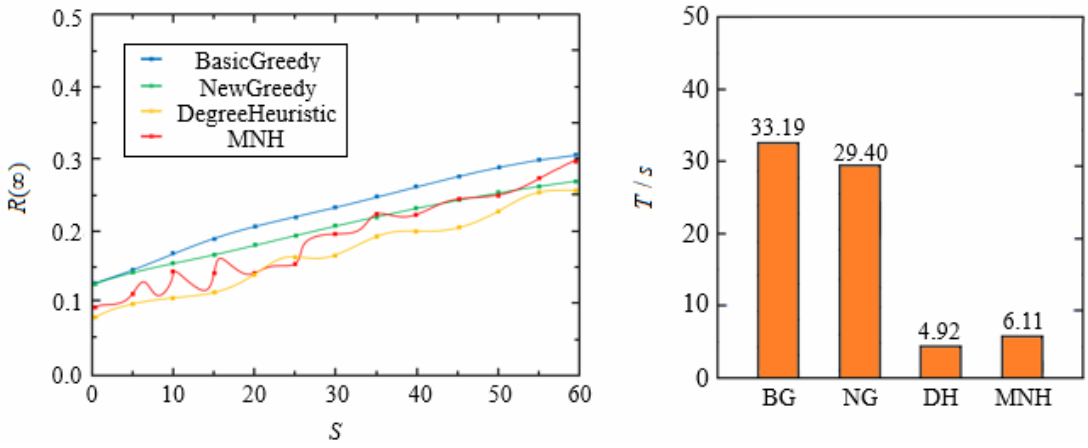

Fig. 4. The simulation results of four algorithms in the independent cascade model in Friendfeed network

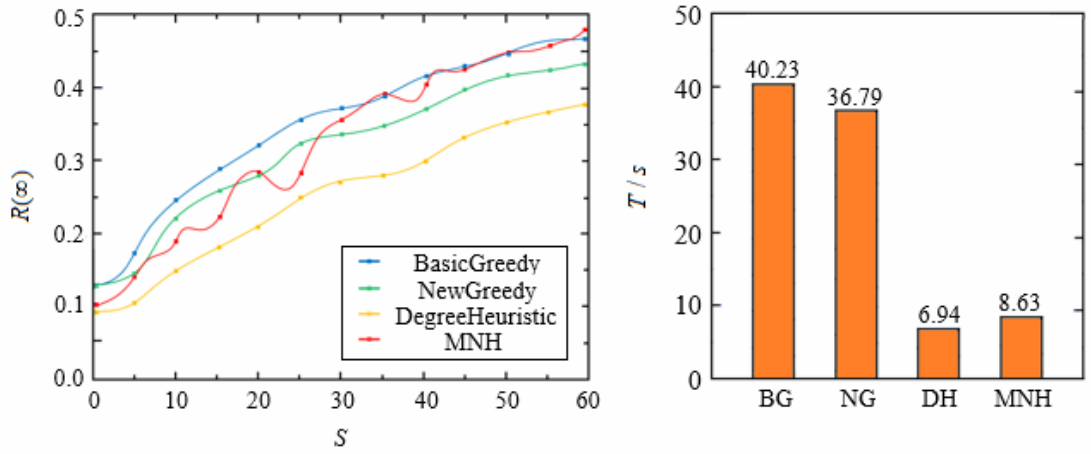

Fig. 5. The simulation results of four algorithms in the independent cascade model in Twitter network

We can find that, in two types of data sets, the overall curves growth rate has slowed down compared with the linear threshold model. The BasicGreedy algorithm still has the greatest influence range, and the DegreeHeuristic algorithm has the smallest influence range. Similar trend is shown in BasicGreedy and NewGreedy curves, and the average gap is about 6% between two of them. It is worth noting that, in the initial propagation of seed node $s=1$, the results of each algorithm have significantly improved. The results show that the independent cascade model has more advantages in propagating information when the initial number of seed nodes is small. In contrast, when the number of initial seed nodes is small, it is difficult for the linear threshold model to make the influence value of the influence of node I_v greater than threshold $\theta(v)$.

MNH algorithm is still better performance in the Twitter data set which has a higher heterogeneity. In Fig. 5, MNH curve fluctuation peak has been close to BasicGreedy curve at $s=35$ in twitter data set, but in Friendfeed data set, this phenomenon is postponed to $s=60$, which the MNH algorithm in Twitter data set can be considered more accurate.

In addition, by comparing the trend of MNH algorithm in different propagation models, we can find that the trend of MNH curve is similar to that of linear threshold model. This result also confirms the fact that the MNH algorithm has a large overlapping part in the intersection of the results of the two propagation models. In contrast, the results of BasicGreedy in the two propagation model have significant differences. In linear Threshold Model, the BasicGreedy curve fluctuates within the $[10\%, 43.5\%]$ in Fig. 2. While, in independent threshold model, it fluctuates within $[12.2\%, 31\%]$. Refer to the overlapping results of BasicGreedy in Table 3 and Table 4, it indicates that the accuracy of the BasicGreedy algorithm has declined in independent threshold model. Therefore, it can be seen from the side that the results calculated by MNH algorithm have higher usage similarity, and Reliance on the propagation model is relatively low, which shows the superiority of MNH.

Compare the operation time of the four algorithms in two datasets, MNH algorithm is still keeping a low computational time-consuming. Compared with NewGreedy and BasicGreedy algorithm, the operation speed advantage is obvious. Although in the two datasets, the DegreeHeuristic and MNH algorithm has a similar time-consuming, the results of MNH algorithm are more accurate. Therefore, we can consider that MNH algorithm has the best efficiency and practicability by balance the time consuming and the accuracy of the algorithm.

4 Conclusions

Social networks provide great convenience for people to spread information. The efficient mining of high-impact node sets in social networks has important implications for both the maximization of information diffusion benefits and the complex network theory itself. The problem of maximizing influence is a tradeoff between efficiency and accuracy. How to design a time-consuming, high-precision and strong-adapting algorithm becomes the goal of this problem.

In the analysis and evaluation of all kinds of greedy algorithm and heuristic algorithm, BasicGreedy

algorithm in most cases has higher operational accuracy. The theoretical result can approximate the optimal solution with the probability of $1-1/e$ approximately. However, BasicGreedy needs to traverse the characteristics of global nodes one by one, which greatly increases the computational complexity of the algorithm in large-scale network environment. Several new algorithms based on BasicGreedy always focus on balance the time-consuming and precision, but not fundamentally solve the problem. At the same time, the heuristic algorithm has higher practicability and flexibility in the real production life, and it can obtain the efficient solving process for different topological features. But it is undeniable that the heuristic based on the network topology features limits the accuracy of the algorithm itself.

After combining the advantages of the greedy algorithm and the heuristic algorithm, this paper proposes a hybrid optimal impact-maximization algorithm based on optimal neighbor discovery, named MNH algorithm. The MNH obtains the candidate seed nodes more than the required number by first obtaining and mining the data randomly, then obtains the optimal solution by comparing the marginal revenue. After analyzing and comparing this algorithm with other three kinds of algorithms in different propagation models, it can be found that although the MNH algorithm has obvious instabilities in the two real networks. But in the integrated accuracy, MNH is close to NewGreedy algorithm, and the operation time-consuming than NewGreedy algorithm to significantly reduce, especially when the set of nodes to be acquired is large.

At the same time, compared with different propagation model, MNH algorithm also has a better applicability, this advantage is mainly reflected in the face of the network that the topology is unknown, and MHN will have a more stable performance. This feature is also should be the future researchers in the algorithm design needs to be an important factor in practical considerations. This is an important factor that researchers need to consider in the algorithm design in future.

At present, most nodes influence analysis is based on the network topology to approximate the influence of nodes. However, the influence of the individual is by no means depended on the topological characteristics alone in the real social network. The user's interest preference, the information content, and the active degree are the focal points of the node influence analysis. At the same time, in the evaluation standard of the node's real influence, the scope of information dissemination and the speed of transmission are only one aspect. From multiple dimensions to analyze the impact of the node will be the future direction of this study.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2018YFC0832304, the National Natural Science Foundation of China under Grant No. 61872033 and U1836118, the Beijing Nova Program from Beijing Municipal Science & Technology Commission under Grant Z201100006820015, and the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant No. 18YJCZH204.

References

[1] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the web, Technical Report, 1999.

[2] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, Journal of Mathematical Sociology 2(1)(1972) 113-120.

[3] D. Wei, X. Deng, X. Zhang, Y. Deng, S. Mahadevan, Identifying influential nodes in weighted networks based on evidence theory, Physica A: Statistical Mechanics and its Applications 392(10)(2013) 2564-2575.

[4] C. Gao, D. Wei, Y. Hu, S. Mahadevan, Y. Deng, A modified evidential methodology of identifying influential nodes in weighted networks, Physica A: Statistical Mechanics and its Applications 392(21)(2013) 5490-5500.

[5] L.C. Freeman, Centrality in social networks conceptual clarification, Social networks 1(3)(1978-1979) 215-239.

[6] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40(1)(1977) 35-41.

[7] M.G. Everett, S.P. Borgatti, The centrality of groups and classes, The Journal of mathematical sociology 23(3)(1999) 181-201.

[8] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical review E 64(1)(2001) 016132.

[9] E. Estrada, J.A. Rodriguez-Velazquez, Subgraph centrality in complex networks, Physical Review E 71(5)(2005) 056103.

[10] K.I. Goh, B. Kahng, D. Kim, Universal behavior of load distribution in scale-free networks, Physical Review Letters 87(27)(2001) 278701.

[11] S. Brin, L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems 30(1-7)(1998) 107-117.

[12] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, Computer networks 56(18)(2012) 3825-3833.

[13] S.J. Kim, S.H. Lee, An improved computation of the pagerank algorithm, in: Proc. Advances in Information Retrieval, 2002.

[14] L. Zhang, T. Qin, T.Y. Liu, Y. Bao, H. Li, N-step PageRank for web search, in: Proc. Advances in Information Retrieval, 2007.

[15] L. Lü, Y.C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PloS one 6(6)(2011) e21202.

[16] P. Domingos, M. Richardson, Mining the network value of customers, in: Proc. the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.

[17] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proc. the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[18] V. Mahajan, E. Muller, F.M. Bass, New product diffusion models in marketing: A review and directions for research, The journal of marketing 54(1)(1990) 1-26.

[19] J. Goldenberg, B. Libai, E. Muller, Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata, Academy of Marketing Science Review 9(2001) 1.

[20] F.M. Bass, Comments on “a new product growth for model consumer durables the bass model”, Management science 50(12_supplement)(2004) 1833-1840.

[21] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proc. the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[22] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, Cost-effective outbreak detection in networks, in: Proc. the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

[23] A. Goyal, W. Lu, L.V.S. Lakshmanan, Celf++: optimizing the greedy algorithm for influence maximization in social networks, in: Proc. the 20th international conference companion on World wide web. ACM, 2011.

[24] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proc. the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[25] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[26] Y. Wang, G. Cong, G. Song, K. Xie, Community-based greedy algorithm for mining top-k influential nodes in mobile social networks, in: Proc. the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

ACM, 2010.

[27] A. Goyal, W. Lu, L.V.S. Lakshmanan, Simpath: An efficient algorithm for influence maximization under the linear threshold model, in: Proc. 2011 IEEE 11th International Conference on Data Mining (ICDM), 2011.

[28] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: Proc. 2010 IEEE 10th International Conference on Data Mining (ICDM), 2010.

[29] K. Jung, W. Heo, W. Chen, IRIE: A scalable influence maximization algorithm for independent cascade model and its extensions, CoRR abs/1111.4795, 2011.

[30] X. Liu, S. Li, X. Liao, S. Peng, L. Wang, Z. Kong, Know by a handful the whole sack: efficient sampling for top-k influential user identification in large graphs, World Wide Web 17(4)(2014) 627.

[31] T. Carnes, C. Nagarajan, S. Wild, A.V. Zuylen, Maximizing influence in a competitive social network: a follower's perspective, in: Proc. the ninth international conference on Electronic commerce. ACM, 2007.

[32] S. Bharathi, D. Kempe, M. Salek, in: Proc. Competitive influence maximization in social networks, Internet and Network Economics, 2007.

[33] W.S. Yang, S.X. Weng, Application of the ant colony optimization algorithm to the influence-maximization problem, International Journal of Swarm Intelligence and Evolutionary Computation 1(1)(2012) 235566.

[34] Wikipedia, NP-hardness. <https://en.wikipedia.org/wiki/NP-hardness>.

[35] S. Milgram, The small world problem, Psychology today 1(1)(1967) 61-67.

[36] E. Y. Daraghmi, S.M. Yuan, We are so close, less than 4 degrees separating you and me!, Computers in Human Behavior 30(2014) 273-285.

[37] R. Cohen, S. Havlin, Scale-free networks are ultrasmall, Physical review letters 90(5)(2003) 058701.

[38] M. Granovetter, Threshold models of collective behavior, American journal of sociology 83(6)(1978) 1420-1443.

[39] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing letters 12(3)(2001) 211-223.