Scalable Influence Maximization in Social Networks using the Community Discovery Algorithm

Jinshuang Li
Computing Center
Northeastern University
Shenyang, China
ljs@cc.neu.edu.cn

Yangyang Yu
Computing Center
Northeastern University
Shenyang, China
rainman3625@gmail.com

***Abstract*—Influence maximization is the problem of finding a small set of most influential vertices in a social network so that their aggregated influence in the network is maximized. Most social networks influence maximization problem are based on the following two basic propagation model: Independent Cascade Model and Linear Threshold Model. They all believe that the impact of all the vertices in a community is the same. It is inconsistent with the actual observed. In social networks, the influence of the different members in a community is not the same. Every community have some core members, their influence is far greater than the others. In view of this, a community discovery algorithm is proposed to find the core members of the community. Selecting the initial members from these core members will have the greatest influence.**

Keywords-social networks; influence maximization; community discovery; viral marketing

I. INTRODUCTION

It is a long history to research information dissemination in the social sciences. Influence maximization is the problem of finding a small set of most influential vertices in a social network so that their aggregated influence in the network is maximized. Research shows that people trust the information obtained from their close social circle far more than the information obtained from general advertisement channels such as TV, newspaper and online advertisements [1]. Thus many people believe that word-of-mouth marketing is the most effective marketing strategy. The increasing popularity of many online social network sites, such as Facebook, Flickr, and MySpace, present new opportunities for enabling large-scale and prevalent viral marketing online..

The identification of influential users in a social network is a problem that has received significant attention in recent research. For the influence maximization problem, given a probabilistic model of information diffusion such as the Independent Cascade Model, a network graph G, and a budget k, the objective is to select a set A of size k for initial activation so that the expected value of f(A) (size of cascade created by selecting set A) is maximized [2]. Early works relied on heuristics such as vertex degree and distance centrality. Although the problem of finding an optimal solution in this model is NP-hard, there is a greedy algorithm that yields a spread that is within $1 - 1/e$ of optimal [3]. This solution is computationally expensive. Work has been done to improve the performance of this greedy algorithm [4,5,6,7], but scalability remains a significant challenge.

Most social networks influence maximization problem are based on the following two basic propagation model: Independent Cascade Model and Linear Threshold Model. They all assume that each user is the same influence for other users. But in the real situations, the influence of opinion leaders is greater than ordinary users. We use community-discovery algorithm to find the opinion leaders, k members are selected from the opinion leaders so that we get the influence maximization set A.

II. BACKGROUND

Generally, a social network is modeled as a directed graph G=(V,E), where the vertices of V represent individuals and edges in E represent relationships and the orientations of the edges indicate the direction of influence. In [3] Kempe et al. proposed two basic stochastic influence cascade models, the independent cascade model(IC) and the linear threshold model(LT), which are extracted from earlier work on social networks analysis, interactive particle systems, and marketing.

A. *Independent Cascade Model*

Each edge e_{uv} in the graph is associated with a propagation probability p_{uv}, which is the probability that vertex u independently activates (influences) vertex v at step t+1 if u is activated at step t. Given a social network graph, the IC model, and a small number k, the influence maximization problem is to find k vertices in the graph such that under the influence cascade model, the expected number of vertices activated by the k seeds (referred to as the influence spread) is the largest possible. Roughly speaking, in the IC model each edge has an activation probability and influence is propagated by activated vertices independently activating their inactive neighbors based on the edge activation probabilities.

B. *Linear Threshold Model*

In the LT model, each edge has a weight, each vertex has a threshold chosen uniformly at random, and a vertex

This work is supported by the Fundamental Research Funds for the Central Universities under Grant No. N090504006.

978-0-7695-4763-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICGEC.2012.127

becomes activated if the weighted sum of its active neighbors exceeds its threshold.

The two models characterize two different aspects of social interaction. The IC model focuses on individual (and independent) interaction and influence among friends in a social network. The LT model focuses on the threshold behavior in influence propagation. The random threshold is to model the uncertainty of individuals' conversion thresholds.

C. *Influence Maximization Vertices in Social Networks*

Many social networks exhibit the property of containing community structure [8]. They naturally divide into groups of vertices with denser connections inside each group and fewer connections crossing groups, where vertices and connections represent network users and their social interactions, respectively. Members in each community of a social network usually share things in common such as interests in photography, movies, music or discussion topics and thus, they tend to interact more frequently with each other than with members outside of their community. Reference [9] proved that in contrast to common belief, there are plausible circumstances where the best spreaders do not correspond to the most highly connected or the most central people. Instead, they find that the most efficient spreaders are those located within the core of the network as identified by the k-shell decomposition analysis.

Every community has one or more core members. Obviously the core members in the communities will have greater influence than other members. Community detection in a network is the gathering of network vertices into groups in such a way that vertices in each group are densely connected inside and sparser outside. Detecting communities in a network provides us meaningful insights to its internal structure as well as its organization principles. Furthermore, many communities detection algorithm records the core vertices in the process of partition community. If choosing these vertices as the information dissemination set, it will undoubtedly produce the greatest influnce in the network. One can possibly run any of the community detection methods. The improved K-means algorithm is proposed in this paper.

III. Finding K Influence Maximization Set Using improved K-means algorithm

A. *Vertex Association Matrix*

There are n vertices and k edges in the graph G, and d (i, j) represents the shortest path between vertex i and vertex j. ε_{ij} is the efficiency of information dissemination (EID) of vertex i to j, which is defined as (1). That is the longer the shortest path is, the lower the efficiency of information communication is.

$$\varepsilon_{ij} = 1/d(i,j) \tag{1}$$

NE(G) is defined as the means of the EID of the graph G, which show in (2).

$$NE(G) = \sum_{i \neq j \in G} \varepsilon_{ij} / n(n-1) = \sum_{i \neq j \in G} (1/d(i,j))/n(n-1) \tag{2}$$

Centrality of edge is defined as (3).

$$C_{ij} = \Delta NE / NE = (NE(G) - NE(G'))/NE(G) \tag{3}$$

Vertex association indicates similar extent between the two vertices. The shorter the shortest path between two vertices is, the greater their vertex association value is. On the other hand, if the longer the shortest path between two vertices, the smaller their vertex association is.

$$nodelink(i,j) = 1 - C_{ij} \tag{4}$$

In (4), $nodelink(i,j)$ is the vertex association between two vertices which associate each other directly. If the two vertices which do not associate each other directly, their shortest path is greater than 2, the vertex association value become lower with the shortest path become longer. The equation is defined as (5).

$$nodelink(i,j) = MAX \prod nodelink(i',j') \tag{5}$$

According to (4) and (5), the vertex association is defined as follow.

$$nodelink(i,j) = \begin{cases} 1 - C_{ij} & i \leftrightarrow j \\ MAX \prod nodelink(i',j') & i \leftrightarrow \cdots \leftrightarrow j \end{cases} \tag{6}$$

Vertex association matrix is defined as (7).

$$L = \begin{cases} \deg(i) & i = j \\ nodelink(i,j) & i \neq j \end{cases} \tag{7}$$

In matrix, if vertex i equal to vertex j, the value is the degrees of the vertex, otherwise, the value is the value of the vertex i and vertex j's association value.

B. *Modularity*

Modularity Q is a function of the particular division of the network into groups, with larger values indicating stronger community structure. This quantity measures the fraction of the edges in the network that connect vertices of the same type minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices. If the number of within-community edges is no better than random, we will get Q = 0. Values approaching Q = 1, which is the maximum, indicate strong community structure. Q is defined as (8).

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{8}$$

A_{ij} is the element in the adjacency matrix. k_i is degrees of vertex i. $\delta(c_i, c_j)$ is a equation which value is one when $c_i = c_j$, otherwise, the value is zero. m is the count of the edges in the networks. The probability existing an edge between the vertex i and vertex j is $\frac{k_i k_j}{2m}$ if it is a random networks.

C. Detecting Community Core Vertex Algorithm

We improve the K-means community detection algorithm. K-Means algorithm input parameters k, the given network is divided into k clusters, the vertices within the cluster have high similarity and the vertices in the different clusters have low similarity.

The parameters k is replaced by the modularity because it is impossible to know exactly the count of the communities in the social network. When a new vertex is added in the community, if the value $\Delta Q = Q' - Q$ (Q' denote the new modularity value after the edge is added) greater than 0, the edge is accepted, otherwise, the edge is denied. The algorithm is as follows.

Algorithm 1 DetectKIM(G,k)

```
1: initialize S =null , m=2;
2:random select a vertex vi, S=S∪vi;
3:select vj which nodelink(i,j) is minimized, S=S∪vj;
4: while m<k do
5:  for each vertex v∈G\S do
6:    select vk which nodelink(a,vk) is maximized, a∈S
7:    Cv=Cv∪vk
8:    For each vertex v in Cv
9:       calculate sum of nodelink(a,v) , a∈S
10:      select vk which sum is maximized, update it to S
11:   end for
12:  end for
13:  calculate $\Delta Q = Q_k - Q_{k-1}$, if it less than zero, goto step 16
14:   m=m+1
15: End while
16: output S:
```

IV. EXPERIMENT

A. Experiment Setup

We use a real-world network dataset DBLP. The DBLP Computer Science Bibliography Database maintained by Michael Ley. It provides bibliographic information on major computer science journals and proceedings. We extracted 1814 data about author and his cooperator for the experiment. The diagram graph is as Figure 1.

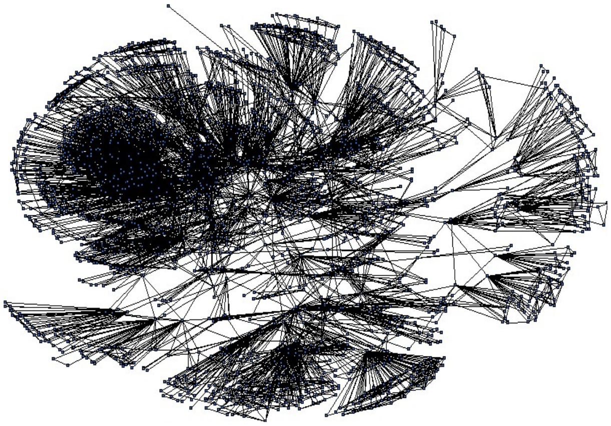

Figure 1. The author and his cooperator's relationships in DBLP.

We use the detecting community core vertices algorithm to test it. The curve of modularity shows as Figure 2.

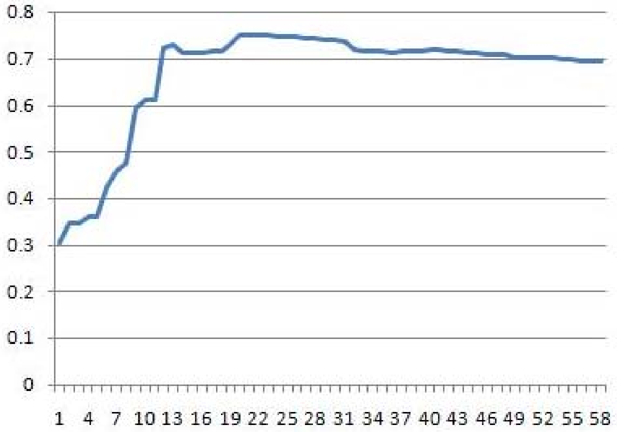

Figure 2. The curve of modularity

It can be seen from Figure 2 that the network modularity reaches a maximum value when k equal to the value 13. The community cutoff is shown in Figure 3.

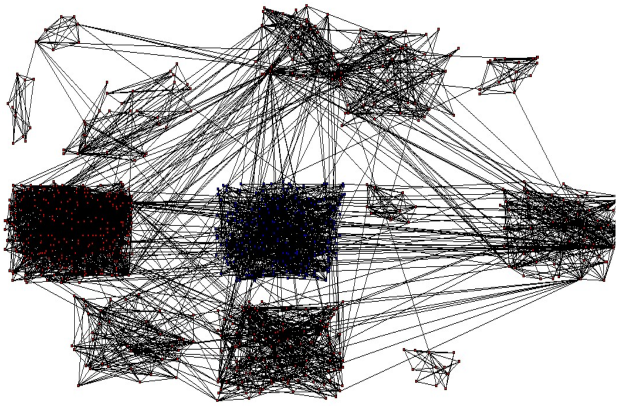

Figure 3. The community cutoff in DBLP

B. Experiment results

It can be clearly seen 1814 the author can be divided into 13 community Using our algorithm, and record the community's core vertices. Obviously, the core vertices are selected in every community as the initial vertex set, we only need to select the 13 vertices to maximum the influence of the entire network.

Due to the uncertainty of the influence of information dissemination, in order to ensure the reliability of transmission, it is a good choice selecting a number of core vertices in each community as the spread of initialization vertices set. Analyzing our algorithms can be found that it has already recorded two core vertices, and it is very easy to remember any number of core vertices.

It provides scientific guidance to set the appropriate value k researching on influence maximization problem using community detection algorithm, and it can avoid subjective and arbitrary to set the value k. It is consistent with the cognition that the influence of the core members is much greater than ordinary members' whether they locate in the real life or in the virtual networks.

V. CONCLUSIONS

No matter what Information Cascade Model or Linear Threshold the Model, they all believe that the influence of all the vertices in a community is the same. It is inconsistent with the actual observed. In social networks, the influence of the different members in a community is not the same. Every community have some core members, their influence is far greater than the others. In view of this, this paper proposed a community detection algorithm to find the core members of the community. Selecting the initial members from these core members must have the greatest influence. The contributions of this paper are mainly reflected in three aspects.

1) Clearly pointed out that the influence of the vertices in the network is different, the more close to the core community members have greater influence.

2) Pointed out that community detection algorithm can be used to find the greatest influence vertices in the social networks.

3) For a given network, calculate the appropriate value k which let influential throughout the network and without getting wasted.

REFERENCES

[1] J. Nail. The consumer advertising backlash, May 2004. Forrester Research and Intelliseek Market Research Report.

[2] M. Richardson and P. Domingos. "Mining knowledge-sharing sites for viral marketing". In KDD, 2002:pp. 61-70.

[3] D. Kempe, J. M. Kleinberg, and É. Tardos. "Maximizing the spread of influence through a social network". In KDD, 2003:pp. 137-146.

[4] W. Chen, Y. Wang, and S. Yang. "Efficient influence maximization in social networks". In KDD, 2009:pp. 199-208.

[5] W. Chen, Y. Yuan, and L. Zhang. "Scalable influence maximization in social networks under the linear threshold model". In ICDM, 2010:pp. 88-97.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance. "Cost-effective outbreak detection in networks". In KDD, 2007:pp. 420-429.

[7] TIAN Jia-Tang, WANG Yi-Tong, FENG Xiao-Jun. "A New Hybrid Algorithm for Influence Maximization in Social Networks". Chinese Journal of Computers, Oct. 2011, vol. 34, pp.1956-1965.(in chinese)

[8] S. Jimeng, C. Faloutsos, S. Papadimitriou, and Philip S. Yu. "Graphscope: parameter-free mining of large time-evolving graphs". In KDD, 2007:p687-696.

[9] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, etc. "Identification of influential spreaders in complex networks". Nature Physics, August 2010, vol.6, pp. 888-893, doi:10.1038/NPHYS1746