Community Expansion in Social Network

Yuanjun Bi[1], Weili Wu[1,2], and Li Wang[2]

[1]Department of Computer Science, University of Texas at Dallas
Richardson, TX 75080, USA
[2]College of Computer Science and Technology, TaiYuan University of Technology
Taiyuan, Shanxi 030024, China
{yuanjun.bi,weiliwu}@utdallas.edu, wangli@tyut.edu.cn

Abstract. While most existing work about community focus on the community structure and the tendency of one individual joining a community; equally important is to understand social influence from community and to find strategies of attracting new members to join the community, which may benefit a range of applications. In this paper, we formally define the problem of community expansion in social network, which is under the marketing promotional activities scenario. We propose three models, Adopter Model, Benefit Model and Combine Model, to present different promotion strategies over time, taking into consideration the community structure characters. Specifically, Adopter Model is based on the factors that can make an individual come into a community. Benefit Model considers the factors that attract more new members. Combine Model aims to find a balance between Adopter Model and Benefit Model. Then a greedy algorithm *ETC* is developed for expanding a community over time. Our results from extensive simulation on several real-world networks demonstrate that our Combine Model performs effectively and outperforms other algorithms.

Keywords: community expansion, community, strategy, social network

1 Introduction

More and more people use online social sites, such as Facebook, Twitter and LinkIn to share interests and contact with each other. Its pretty impressive to see that by 2012, Facebook has more than 800 million active users, with Twitter 100 million and LinkedIn over 64 million in North America alone[1]. Due to their great social influence, some researchers study the structure of social networking, e.g.community detection, to simplify the representation. Some research work focus on the viral marketing analysis based on social medias. However, social medias or social communities need to develop themselves such that they can obtain greater influence and provide better service. In this paper, we argue that it is equally important to study the strategy for community expansion, which may carry significant benefits for a range of applications, such as

* This work was supported in part by the U.S. National Science Foundation under Grant CNS-0831579, CNS-1016320, and CCF-0829993.

i)*Broaden Sales Horizons*. Every company is looking for more customers to increase new sales. How to find potential customers is important especially when the marketing budget is limited. Community expansion problem analyzes how existing customers organized. And who should be potential new customers. It can provide strategy to increase customer community.

ii)*Political Campaign*. With the advent of Internet technology, the concept of community has less geological constraint. In some cases, it refers to a group people who have common will. Enlarging alliance is a good approach for political candidates to spread their influence towards decision making process. Modeling such influence based on community structure offers valuable insight in choices during campaign activities.

iii)*Boost Exhibition Participation*. Trade shows and exhibitions support a significant opportunity to enhance brand and product visibility. As the organizer of exhibitions, lager size and higher level participators can improve the fame of exhibition. How to choose excellent exhibit display participators that meet marketing needs and budgetary requirements can be found in our problem.

To illustrate our problem clearly, consider the following example. Suppose a company employs some salesmen to do promotional activities towards new customers. All the salesmen know the structure of the whole social network, that is who has relationship with whom. Since the cost of the promotional activities is limited, each salesman can only do the promotion to one person during a certain period. Our goal is to find a strategy for each salesman so that after several times there are new members as much as possible.

The challenged part is how to select the next potential customers to do the promotion. If the total number of new customers is our objective, we should not only consider those who are easily persuaded to come into the given community. Notice that each time when new customers come into the community, these new customers have influence on the structure of the community. In common sense, popular people may bring more customers but it is more difficult to persuade these people join the community. However, if more people came into the community, it will have more probability to attract those popular people.

Our paper engage in solving this problem. The main contributions of this paper can be listed as follows:

i)Formulate the problem of *Community Expansion* in social networks.

ii)Build three models for expanding the community, Adopter Model, Benefit Model and Combine Model. The first model considers factors that make an individual come into a community. The second model considers factors that make an individual attract more new members. The third model aims to find balance between Adopter Model and Benefit Model.

iii)Propose a greedy algorithm based on the above three models to present the community expansion progress. Based on the experiment results, an analysis is given to show which model is proper for which community structure.

The rest of this paper is organized as follows. In the next section, a brief overview of related work is introduced. Section 3 present the formal definition of our problem and several relative terms. We also compare our problem with other

similar problems in this section. Section 4 give three models for our problem. A greedy algorithm based on these models are proposed in section 5. The simulation results and conclusions are showed in section 6 and 7 respectively.

2 Related Work

There are a lot of work focus on social network structure and their diffusion. Newman, *et al*[2, 3] gave the definition of community that the nodes inside the community have tighter connections than nodes outside the community. Newman's Q value is considered as an important measurement for detecting community. Nguyen, *et al*[4] analysis four basic events occurring in dynamic network and propose adaptive algorithms separately to update the network community structure. Lars, *et al*[5]use decision tree to study the factors which influence an individual to join communities. They also study which community has more propensity to grow and how to measure the movement of individuals between communities. Kumar, *et al*[6] partition the nodes in network into three segments: singletons, middle region and giant component. Instead of using snapshot of network, Kumar put their experiments on entire lifetime of two large social network Flickr and Yahoo! 360 to study the overall properties of network and how these communities grow and merge. Their work are both based on the self development of a community which are different from the strategy that aims to enlarge a community as we study here. Other researchers [7–9] focus on marketing on social network. Domingos, *et al*[10] employ Markov random field to model the marketing value for each individual from collaborative filtering databases. The model use the influence between customers to increase the benefit. In [11] the authors extend their previous work by considering each customer's fund and reducing the computational cost. They apply the idea on knowledge-share sites. Generally, their work focus on the benefit which one single individual bring to a network but ignore the global group profit.

Information propagation problem is to find a set of initial set of users in a social network such that from this set the spread of influence in the network can be maximized[12]. Linear Threshold (LT) Model and Independent Cascade(IC) Model are two main approaches to formalize the influence maximization problem. Chen, *et al*[13] propose a MIA model and its heuristic algorithm to address the scalability and efficiency issue in large scale networks. Shaojie, *et al*[14] consider the links relationship impact on the information propagation. Saito, *et al*[15] predict the final influence over the whole network from a given initial set without modeling the diffusion process. They apply the expectation maximization(EM) algorithm to learn the influence probabilities. Goyal, *et al*[16] use action log to learn influence probabilities on each user. Their method can predict whether a user will take an action and tell when the action will be performed. Our work is different from all of the above. The comparison will be given in section 3.4.

3 Problem Formulation

3.1 Preliminaries

We start with introducing a set of fundamental concepts used throughout the paper. We denote the social network as a graph $G = (V, E, W)$, where V is a set of vertices, E is the set of edges and W is the weight matrix for edges. In a social network a vertex v corresponds to a person. An edge $e = (v, u)$ represents a connection between vertices v and u. w_{vu} represents the connection weight between v and u.

Definition 1. ***Target Community(TC)****: TC is a subgraph of G whose size we aim to enlarge. TC satisfies the definition of community that nodes inside the community are more densely connected internally than with the rest of the network. We consider the nodes inside TC as original customers(OC) while the nodes outside TC as potential customers(PC). Let $TCs = \{TC_1, TC_2, ..., TC_T\}$ be a series of target community, where TC_k is a snapshot of a target community TC at time t_k,($k \in [1, ..., T]$).*

Definition 2. ***Sales List(SL)****: SL is a subset of nodes which are outside the target community (i.e PC). Suppose there are M sales lists such that $PC = SL(1) \cup SL(2) \cup ... \cup SL(M)$. Note that we allow that different SL can have different number of nodes and one SL can have different versions over time, denoting the version at time t_k as $SL(m)_k$,($k \in [1, ..., T]$,$m \in [1, ..., M]$).*

Definition 3. ***New Customers(NC)****: Some customers will decide to join in TC after promotion. Among these new customers someone join because of promotional activity from salesmen. We define them as Mark Customers **MC**. While others receive no promotion but are influenced by their friends. We define them as Automatical Customers **AC**. The number of NC changes with each promotion time t_k and $NC = MC \cup AC$.*

3.2 Problem Definition

The progress of community expansion can be considered as the result of community attraction to individuals outside the community. The attraction can be departed to two parts.

1. Due to promotional activities from the target community, some potential customers are attracted. We denote the direct influence from target community to an individual i as $f_{TC \to i}$.

2. Through "word of mouth", some customers should be influenced by their neighbors. We denote the influence from one individual i to another individual j as $f_{i \to j}$.

The final influence from target community to one individual i can be described as

$$F_{TC \to i} = f_{TC \to i} + \sum_{k=1}^{d(i)} w_{ij_k} f_{i \to j_k} \tag{1}$$

where $d(i)$ denotes the number of neighbors of i. w_{ij_k} denotes the influence weight between i and j.

Suppose at time slot t, each salesman chooses one customer from their sales list and form the promotion target customers set $L_t = \{i_1^t, i_2^t, ..., i_m^t\}$, recall that m is the number of sales lists. Then our problem can be defined as

Community Expansion Problem: Given social network $G(V, E, W)$, sales list SL, target community TC and time slot $t \in \{1, ..., T\}$, find nodes sets L_t, such that the influence from TC to L_t, $\sum_{t=1}^{T} F_{TC \to L_t}$ can be maximized.

$$\sum_{t=1}^{T} F_{TC \to L_t} = \sum_{t=1}^{T} (\sum_{a=1}^{m} f_{TC \to i_a^t} + \sum_{a=1}^{m} \sum_{b=1}^{d(i_a^t)} w_{i_a^t j_b} f_{i_a^t \to j_b}) \tag{2}$$

In our paper, the influence f refers to attracting new customers. Specifically, the result of $f_{TC \to i}$ can be seen as customer i who accepts promotion and chooses to be a new member of TC(i.e Mark Customer(MC)). The value of $f_{i \to j}$ can be considered as customer j who did not receive promotion but was influenced by friend i, decides to join TC as well(i.e Automatical Customer(AC)). To extend our problem, influence function f can be described in other ways, such as generating new connections or strengthening existing connections with TC.

3.3 Problem Assumptions

Our problem of expanding Target Community is based on the following assumptions:

i)**Specific potential client each time** That means in each time slot t_k each salesman can do the promotional activity to one and only one potential customer in their corresponding Sales List. Once one potential customer was chosen to be promoted, he or she should be removed from the Sales List.

ii)**Closed customers information open community information** We assume that Salesmen don't share their own potential customers information with each other. That means any two Sales Lists are not overlapping. However, each person in the network G can get the latest information about the Target Community structure. After each time slot, each person will obtain the TC information and check whether the changes will affect its action.

iii)**No new connection details** After each promotional activity, there might be some New Customers(NC) come into TC. However, in our paper we assume that the connections among the nodes in NC won't change in T time slots. That will be true since in real world, T time slots might be very small compared to the time which the Target Community uses to build up. The network may be very large and complex so that even there are some connections changed in T time slots, these changes won't affect the G's structure too much. On the other hand, the community organizers only care about enlarging the size of their community. They don't care about the new connections after new customers coming in TC.

3.4 Comparison with Influence Maximization Problem

Compare to Influence Maximization problem, our problem is different in that:

(1) In our problem, we focus on the influence from a community which has a specific structure, rather than the initial seed set in which nodes distributed randomly. The community structure constraints provide some factors which should be considered when building the models. While in Influence Maximization problem, the influence source has no constraints.

(2) In Influence Maximization problem there is only one interference in the diffusion progress which is choosing the initial seed set occurring at the beginning of diffusion. After choosing the seed set all the diffusion progress proceed automatically. While in our problem human interference occur during the whole progress. The salesmen do the promotion several times until the result is satisfied. Each time they will adjust the candidates list according to the current social graph structure.

(3) Our goal is to maximize the size of community not to spread the influence from the seed set as described in Influence Maximization problem.

4 Community Expansion Models

4.1 Intuitions

Before formally introducing the model, we first explain several key observations:

Observation 1 *In [5], the study shows that the tendency of an individual to join a community depends on the underlying network structure. The probability p of joining a community depends on the number of friends k who are already in the community. The relation between p and k is under the "law of diminishing returns". Besides k, how these friends connected in the community also affect an individual's decision. If an individual has no friends in the community, then "how far" from the people in the community will decide "how much" the community impacts on the person. [17] infers that everyone is approximately six steps away from others.*

Observation 2 *Approximately 25% of US advertisements employ celebrities in their media[18]. We cannot ignore that celebrities have positive impact on consumer attitudes towards the purchase intention. Considering of that, a company should consider the financial returns from celebrities. In the real world the celebrities are more likely be known by others. In the network graph, we can consider the nodes which have more connections with other nodes as the celebrities.*

Based on the intuitions and observations, we know that both the probability of an individual coming into a community and the benefit of an individual to a community depends on the current network graph structure. Now, we want to build the Adopter Model to present how easy a potential customer join TC and build the Benefit Model to denote how much benefit the customer can bring into TC. Then Combine Model considers these two factors both.

4.2 Adopter Model

In Adopter Model, an individual who wants to join in TC is affected by how many friends in TC and how close these friends are. We define the meaning of friends is the same as neighbors in the following context, who have direct connection with this individual. Let η denote the value of how easy an individual adopts the promotional activity from TC. We give the formulation of η by the value of k friends in TC

$$\eta = (a_1 \log k) + (a_2 * \frac{k}{n}) + (a_3 * d_{in})(k > 0) \tag{3}$$

$$\eta = \frac{a_4}{dis}(k = 0) \tag{4}$$

where, k is the number of friends in TC. n is the number of neighbors of the individual. d_{in} denotes the density of k friends connected in TC. dis is the distance between the individual and the first node which the individual meet in TC. a_1, a_2, a_3, a_4 are adjusted parameters to make the model to fit data set. Function 3 denotes the customers who already have friends in TC. Intuitionally, people are more likely join TC if they have more friends in TC. However, if there are enough people to affect the individual to decide to join TC, additional friends will have small effect. Such that η is not linearly changed with k. On the other hand, even if two persons have the same number of friends in TC, that does not mean they will both choose to accept TC. The high ratio of friends in TC will obtain greater influence from TC. So we consider $\frac{k}{n}$ here. The connection density of friends in TC is also important. The more mutual friends they have, the stronger power they have to affect another individual. Here, we compute d_{in} by

$$d_{in} = \frac{2 * \rho}{k * (k-1)} \tag{5}$$

where, ρ denotes the number of connections between k friends in TC. $\frac{k*(k-1)}{2}$ stands for the the number of connections if any two friends in TC have a relation. As showed in $Fig.1$, node V_1 has four friends V_2, V_3, V_4 and V_5 in TC. In $Fig.1(a)$, the connections between these four friends is 6 and their $d_{in} = 1$, while in $Fig.1(b)$, the four friends has no connection with each other so their $d_{in} = 0$.

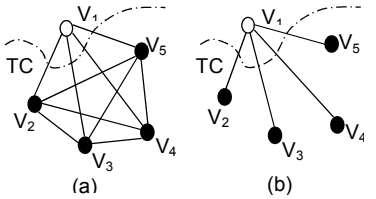

Fig. 1. Friends Connections in TC

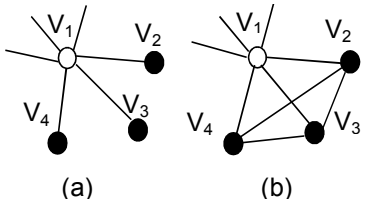

Fig. 2. Friends Connections not in TC

Function 4 give the η value if the individual has no friends in TC. The distance from TC determines how easy the individual to join in TC.

4.3 Benefit Model

Now we build Benefit Model to present the benefit that an individual can bring into TC. This model is important when the marketing strategy considers the cost constraints of promotional activities. Here to simplify the problem we assume that the promotional cost to each potential customer is the same. We will put the cost problem in the future extended work. Since our goal is to enlarge the size of Target Community as much as possible, the individual who knows more people will have more opportunities to introduce the Target Community to others, and attract more people coming into TC. On the other hand, if a popular node already has a lot of friends in TC, its benefit will be less than a node who have many friends that still are potential customers. Another factor that affects an individual's benefit is the structure of friends' connections. If the connections among friends is very density, it is difficult to persuade them to join TC since each individual is deeply influenced by the power of groups, not by a single person. As showed in $Fig.2$, node V_1 has three friends V_2, V_3 and V_4 who are potential customers. In $Fig.2(a)$, the connections between these three friends is 0 while in $Fig.2(b)$, these three friends have connections with each other so they are not that easily persuaded by V_1.

Based on the above analysis, we build the Benefit Model θ:

$$\theta = (b_1 * n) + (b_2 * \frac{n-k}{n}) - (b_3 * d_{out}) \tag{6}$$

where, n stands for the number of neighbors. $\frac{n-k}{n}$ presents the tendency of how many potential customers that the individual can attract. b_1, b_2, b_3 are adjusted parameters. d_{out} denotes the connections density among friends who are not in TC. It is computed by

$$d_{out} = \frac{2 * \sigma}{(n-k) * (n-k-1)} \tag{7}$$

where, σ denotes the number of connections among neighbors who are not in TC. $\frac{(n-k)*(n-k-1)}{2}$ stands for the the number of connections if any two neighbors who are not in TC have a relation with each other.

4.4 Combine Model

Combine Model wants to find a balance between Adopter Model and Benefit Model. Choosing customers who are not too easily joining TC but still have some benefit that will attract new customers in long term. We take some factors from Adopter Model and Benefit Model respectively to define Combine Model.

$$\gamma = c_1 \log k + c_2 * (n-k) + c_3 * d_{in} - c_4 * d_{out} \tag{8}$$

where c_1, c_2, c_3, c_4 are adjusted parameters. The definitions of k, n, d_{in} and d_{out} can be found in Adopter Model and Benefit Model.

5 The Algorithm

In this section, we propose an algorithm which includes three stages for Expanding Target Community (ETC). The first step is to find *Mark Customers(MC)* set after one promotional activity, in which we compute score for each potential customer and choose the one with highest score from each sales list. Then we compute its probability to see whether it will join TC or not. The second step is to update the graph information. The final step is to check whether there are *Automatical Customers(AC)* after graph was updated. After T times promotional activities we will get the total value MC and AC.

Algorithm 1: ETC Algorithm

Input: $G = (V, E)$,TC, m Sales Lists sl_1,$\cdots$,sl_m, T;
Output: NC,MC,AC;
$t = 0, MC = \emptyset, AC = \emptyset, NC = \emptyset$;
while $t < T$ **do**
  $t \leftarrow t + 1$;
  for *each* $sl_i, i < m$ **do**
    compute each n's score $S(n)$,$(n \in Sl_i)$;
    select the node v with the highest score in sl_i;
    compute v's joining probability $p(v)$;
    if $p(v) > \lambda$ **then**
      $MC \leftarrow MC \cup \{v\}$;
    end
  end
  for *each* $v \in MC$ **do**
    for *each v's neighbor* w **do**
      $w.k = w.k + p(v)$;
    end
  end
  for *each* $v \in MC$ **do**
    for *each v's neighbor* w **do**
      if $\frac{w.k}{w.n} > \lambda$ **then**
        $AC \leftarrow AC \cup \{w\}$;
      end
    end
  end
  $NC = MC \cup AC$;
end

5.1 Customer's Score and Joining Probability

To find Mark Customer set, we need to obtain the most reasonable potential customer in each Sale List. Based on three different models we discussed above

Adopter Model, *Benefit Model* and *Combine Model*, we compute η, θ or γ for each node as its score separately. Sort each Sale List by the score's value in descending order. The node with the highest score in each list will be severed the promotion. It comes into TC with some probability which has been studied by Lars, *et al*[5]. They find that the tendency of an individual to join a community is influenced by the number of friends within the community and by how those friends are connected to each other. In our algorithm, the probability of an customer v join in TC, $p(v)$, can be computed as Equation 9 for appropriate a, b, c.

$$p(v) = a \log k + b * d_{in} + c(k > 0) \tag{9}$$

For those customers who don't have friends in the community, we consider they still have probability to join in TC, with

$$p(v) = \frac{d}{dis}(k = 0) \tag{10}$$

The definitions of k, d_{in}, dis are the same as the description in Adopter Model. d is the parameter. Here we use algebraic function $f(x)$ to make sure the probability value is between 0 and 1.

$$f(x) = \frac{x}{\sqrt{1 + x^2}}$$

In our algorithm, whether a customer will join in TC is decided by threshold λ, $0 < \lambda < 1$. It is a factor reflects how easy an individual can join in a community. We will see how λ affects results in the experiments.

5.2 Graph Information Update and Automatical Customers

The coming of new Mark Customers will change their neighbor's information about friends number in TC k and connection density of friends. These updates will make some neighbors join TC automatically. We define that if there exists more than λ ratio of friends in TC, the customer will become *Automatic Customer*(AC). The nodes from AC will affect their neighbors as well, so the process of finding AC will not cease until no nodes from AC make their neighbors join TC. Figure 3 illustrates for V_1 how its k value is updated. After on promotional activity, node V_3 and V_4 join TC with the probability $p(3) = 0.65, p(4) = 0.6$. Node V_2 is an original customer in TC. Now node V_1 has $1+0.65+0.6 = 2.25$ friends in TC. Since $\frac{k}{n} = \frac{2.25}{4} > 0.5(\lambda = 0.5)$, node V_1 will join TC automatically.

6 Experiment

We conduct experiments on ETC algorithm as well as other two algorithms on four real-world networks. Our experiments aim at illustrating the performance of ETC algorithm from the following aspects: (1) Its capacity of attracting new members comparing to other algorithms; (2) Its efficiency of attracting new members comparing to other algorithms; (3) The tuning of its control parameter λ.

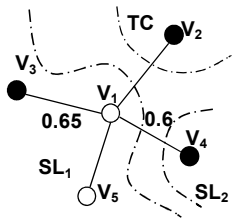

Fig. 3. Effect on neighbors

6.1 Experiment Setup

Datasets. We use three realistic data sets: American College Football, Arenas Email, NetHEPT and Facebook.

AmericanCollegeFootball(ACF) The network is a representation of the schedule of Division I games for the 2000 season, in which vertices represent teams (identified by their college names) and edges are regular-season games between the two teams they connect.

Arenas/email It comes from email interchange network, Univ. of Rovira i Virgili, Tarragona. The nodes are the members in this university and the edges represent email interchanges between members.

$NetHEPT$ This data set is an academic collaboration network taken from the "High Energy Physics (Theory)"section (from 1991 to 2003) of arXiv. The nodes in NetHEPT denote the authors and the edges represent the co-authorship.

Facebook The nodes in Facebook denote the facebook users and the edges represent the friendship. We choose these networks since it covers a variety of networks with size ranging from $1K$ edges to $1M$ edges. Some statistics about these networks' properties are given in Table 1. Close customer refers to individual who has friends in TC.

DataSets	NetHEPT/Author	Arenas/Email	ACF/Team	Facebook
Number of Nodes	15233	1133	115	63732
Number of Edges	62774	10903	1226	1634180
Number of Sale List	1819	70	12	210
Average Sale List Size	8.4	15.9	8.9	227.5
Target Community(TC) Size	1251	179	12	15963
Ratio of close customer	0.057	0.36	0.26	0.01
Average Friends of close customer	2.06	2.26	1.26	1.84
Average Degree	3.76	9.17	10.67	0.33

Table 1. Data Sets Properties

Generating Target Community.
To find TC and Sale Lists, we first partition the social network graph into several communities. We select the community with the maximum size as the Target Community TC. The rest communities are considered as Sale Lists.The partition

of NetHEPT and Facebook employ the methods in [19], and partition of Arenas and ACF use the methods in [20].

Algorithms.

We use our ETC algorithm based on the three models discussed above. Compare the three models with a baseline model and another algorithm which solves the similar problem. The following is a list of algorithms we evaluate in our experiments.

(1) ETC: Our algorithm is a greedy algorithm. Base on our Adopter Model, Benefit Model and Combine Model, we have methods $ETCA, ETCB, ETCC$ respectively.

(2) Random: As a baseline comparison, simply select node from each Sales List each time.

(3) TABI: TABI is a heuristic algorithm proposed by Tao, *et al*[21] to solve the participation maximization problem. This algorithm calculates participants' influence and allocate thread according to influence ranking, which is similar with our ETC algorithm. However, TABI only considers people who have participated in the forum, which means the algorithm only chooses candidates from people who have connections with the community. It computes every participant v's influence by

$$(1 - \prod_{u \in v.K} (1 - w_{u,v}))(1 + \sum_{x \in (v.N - v.K)} w_{v,x} \prod_{y \in x.K} (1 - w_{y,x}))$$

Here, $v.K$ refers to v's friends set in the community. $v.N$ refers to v's neighbor set. Since our data sets are unweighed social networking, each edges has the same weight.

To obtain each algorithm's capacity of attracting new members. We run the simulation 1000 times and take the average of results, which matches the accuracy of the greedy algorithm.

6.2 Experimental Results

Capacity of attracting new customers. We measure the capacity of attracting new customers by two measurements, Automatical Customer size and New Customer size. The promotion time T ranges from 1 to 10. The first measurement is for evaluating the performance of attracting people who can bring more benefit to the community. The second measurement is for evaluating the performance of attracting more new customers totally.

For the moderate sized graph Arenas Email, as showed in Figure 4 and Figure 5, our ETCC performs best on both two measurements. When $T = 10$, for the New Customers measurement, ETCC is $4.1\%, 51\%, 81.8\%$ better, while for the Automatical Customers measurement, ETTC is $0.1\%, 32.8\%, 51.7\%$ better, comparing to ETCA, RANDOM and TABI respectively. ETTC performs even much better than ETCA, RANDOM and TABI when $T = 5$. The results of ETCB are very close to ETCC. TABI attracts more customers than RANDOM before $T < 8$. After that, RANDOM obtains more customers. That phenomena proves that TABI is an expanding algorithm only considering people who

have connection in the target community. TABI has weak capacity of attracting valuable customers who can bring automatical customers.

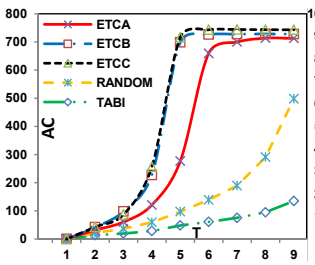

Fig. 4. Arenas_AC

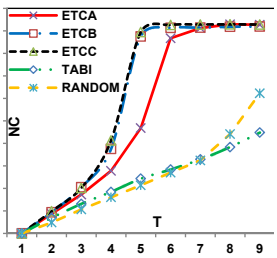

Fig. 5. Arenas_NC

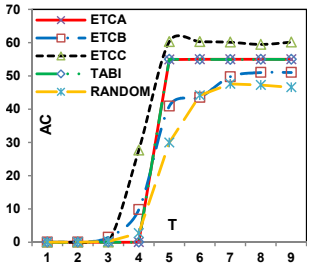

Fig. 6. ACF_AC

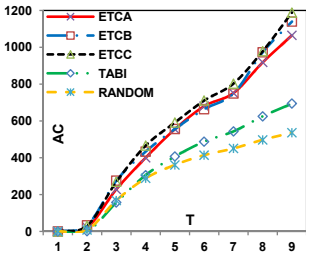

Fig. 7. NetHEPT_AC

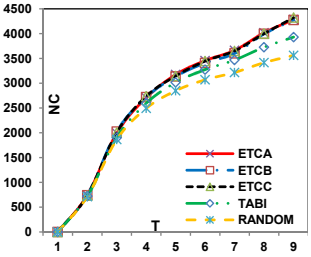

Fig. 8. NetHEPT_NC

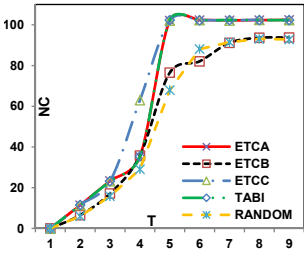

Fig. 9. ACF_NC

Figure 6 and Figure 9 show the results on ACF/Team dataset. ETCC still works the best on two measurements. For New Customers, ETCC is 8.4%, 9.3% better than ETCB, RANDOM respectively when $T = 10$. While for Automatical Customers, ETCC is 8.7%, 15.1%, 22.6% better than TABI, ETCB, RANDOM respectively. TABI which has the similar result as ETCA, performs better on this dataset. It is probably because ACF/Team is a small network with many people who has connection in TC(i.e a relatively large ratio of close friend). In this case, it seems choosing who can easily join TC is a better strategy for the community.

Next, for the 60 thousand edges NetHEPT dataset, Figure 7 and Figure 8 show ETCC performs slightly better than ETCA and ETCB, but consistently much better than TABI and RANDOM. For New Customers, ETCC is 9.1%, 17.6% better, while for Automatical Customers, ETCC is 41.6%, 54.9% better than TABI and RANDOM respectively when $T = 10$.

Finally, for the 1.6 million edge Facebook dataset, Figure 10 and Figure 11 show that this time ETCA performs much better than other algorithms. ETCB,

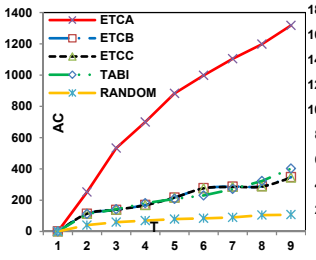

Fig. 10. Facebook_AC

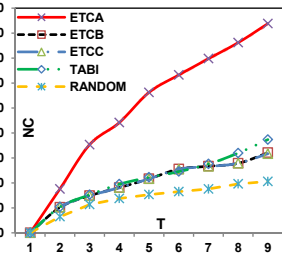

Fig. 11. Facebook_NC

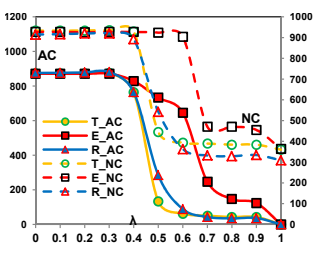

Fig. 12. Tuning of λ

ETCC and TABI have close results. For New Customers, ETCA is 55.4%, 75.4% better, while for Automatical Customers, ETCA is 69.5%, 91.9% better than TABI and RANDOM respectively when $T = 10$. This phenomena is quite different from phenomena of other datasets result as we have seen so far. Note that there are two unique features for this dataset: (a)the average degree of each node is small, which means the distribution of nodes in the network is scattered. As a result there are more nodes are easy to persuaded to join TC, which means ETCA is a good choice; (b)TABI seems to consider the nodes that can join in TC easily as well. However, TABI only considers nodes that have connections in TC while the ratio of close friend in this dataset is rather small. So TABI will ignore some nodes which in the view of ETCA is better choice.

Overall, we see that ETCC significantly outperforms the rest algorithms in most cases. ETCB and ETCA still have better results than TABI and RANDOM.

Efficiency of attracting new customers. The efficiency of attracting new customers is another important evaluation criterion, especially when the community considers the promotion time cost. In Figure 5 and Figure 9, we can see that when $T = 5$, ETCC curve has reach its peak value, which means it has attracted most new customers. While TABI and RANDOM need more time to reach their peak value.

Tuning of parameter λ. We investigate the effect of the tuning parameter λ on the capacity of attracting new customers. λ ranges from 0 to 1. We compare the results of ETCC, TABI and RANDOM when $T = 10$. Since λ decides how easy an individual can join in TC, Figure 12 show that the capacity of attracting new customers increases when the λ value decreases, as expected. For both attracting NC and AC, ETCC(E_NC,E_AC) keep high value in lager range of λ than TABI(T_NC,T_AC) and RANDOM(R_NC,R_AC), indicating that ETCC performs more stable than TABI and RANDOM.

7 Conclusions

In this paper, we formally define the problem of expanding community. A greedy Expanding Target Community (ETC) algorithm is proposed, which employs three models Adopter Model, Benefit Model and Combine Model. These models

consider the factors that affect an individual to join community and the factors that attract new members. Experiment results based on four real world datasets shows that our models perform better than RANDOM and TABI algorithm.

References

1. Infographic: Social media statistics for 2012. `http://www.digitalbuzzblog.com/social-media-statistics-stats-2012-infographic`.
2. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *PNAS*, 2002.
3. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev.*, 2004.
4. N. P. Nguyen, T. N. Dinh, Y. Xuan, and M. T. Thai. Adaptive algorithms for detecting community structure in dynamic social networks. In *INFOCOM*, 2011.
5. L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.
6. R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, 2006.
7. R. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
8. D. McKenzie Mohr and W. Smith. *Fostering Sustainable Behavior: An Introduction to Community Based Social Marketing*. New Society Publishers, 1971.
9. J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 2007.
10. P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, 2001.
11. M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.
12. D. Kempe, J. Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.
13. C. Wang, W. Chen, and Y. Wang. Scalable influence maximization for independent cascade model in large-scale soical networks. *Data Mining and Knowledge Discovery*, 2012.
14. S. Tang, J. Yuan, X. Mao, X. Li, W. Chen, and G. Dai. Relationship classification in large scale online social networks and its impact on information propagation. In *INFOCOM*, 2011.
15. K. Satio, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. *KES*, 2008.
16. A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.
17. Stanley Milgram. The small world problem. *Psychology Today*, 1967.
18. Terence A. Shimp. *Advertising promotion: Supplemental aspects of integrated marketing communications*. South-Western College Pub, 2002.
19. V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008.
20. Y. Hu, H. Chen, P. Zhang, Z. Di M. Li, and Y. Fan. Comparative definition of community and corresponding identifying algorithm. *Phys. Rev.*, 2008.
21. T. Sun, W. Chen, Z. Liu, Y. Wang, X. Sun, M. Zhang, and C. Lin. Participation maximization based on social influence in online discussion forums. In *ICWSM*, 2011.