Evolutionary Influence Maximization in Viral Marketing

Sanket Anil Naik and Qi Yu

Abstract With the growth of social networks, significant amount of data is brought online that can benefit applications of many kinds if being effectively utilized. As a typical example, Domnigos proposed the concept of viral marketing, which uses the "word of mouth" marketing technique over virtual networks (Domingos, IEEE Intell Syst 20:80–82, 2005). Each user is associated with a network value that represents his/her influence in the network. The network value is used along with other intrinsic features that represent user shopping behaviors for the selection of a small subset of most influential users in the network for marketing purpose. However, most existing viral marketing techniques ignore the dynamic nature of the virtual network where both the features and the relationship of users may change over time. In this paper, we develop a novel framework for the selection of users by exploiting the temporal dynamics of the network. Incorporating temporal dynamics of the network would assist in selecting an optimal subset of users with the maximum influence over the network. This paper focuses on developing an algorithm for the selection of the users to market the product by exploiting the temporal and the structural dynamics of the network. Extensive experimental results over real-world datasets clearly demonstrate the effectiveness of the proposed framework.

Keywords Viral marketing · Subset selection · Evolutionary · Network value · Influence flow

1 Introduction

The exponential growth of the Internet has transformed the Web into a *virtual world*. As most people in the real world have become a part of this virtual world, their social experience has also been translated into the Web. The increasing popularity over social network sites, such as Facebook, LiveJournal, and Twitter indicates

S.A. Naik (✉) · Q. Yu
Rochester Institute of Technology, Rochester, NY, USA
e-mail: san8774@rit.edu

Q. Yu
e-mail: qi.yu@rit.edu

Ö. Ulusoy et al. (eds.), *Recommendation and Search in Social Networks*,
Lecture Notes in Social Networks, DOI 10.1007/978-3-319-14379-8_11

the immense interactions of the users on the Web. The large-scale data resulted from social interactions over the Web forms a rich information repository that has the potential to benefit various applications. Marketing is a typical example of such applications. Traditional Web-based marketing mechanisms are more inclined toward direct marketing, which identifies the most probable customers and then markets the product or service directly to them. Although direct marketing ensures that marketing is delivered directly to potential customers, it is a slow and expensive process especially when targeting thousands of millions of online users. If the market cannot be conducted in a timely fashion, valuable business opportunities may get lost.

Different from direct marketing that treats each customer as an isolated entity, *viral marketing* regards users as part of a connected network and aims at selecting a subset of users in the network to market with an ability to influence other members of the network [1]. The interactions between users in the network help achieve this objective via implicit or explicit recommendation. Individual's actions also contribute toward influencing people around a user. For example, people tend to look at what others around them are using or buying. A person who has a better understanding of the preference of their friends is more likely to make proper recommendation on products or services. These behavioral phenomena will further strengthen the effectiveness of viral marketing. Some recent statistics of social network user behavior provide clear evidence to justify the significant potential of viral marketing. For example, among all users of the major social network sites (e.g., Facebook, Twitter, and MySpace), 20 % of them share content of the network using the share option [2]. A 2009 research reveals that 32 % of the users share promotional offers inside a private social network [3] whereas 51 % of the users click "forward to a friend" in marketing emails [4].

In social networks, the ability of a user to influence others increases with the number of connections or interactions with other users in the network. Hence, viral marketing when implemented properly can grow exponentially. This exponential growth can be simplified and represented in the form of a pyramid, where users at every level are influenced by the users in the above level and have the responsibility to influence the users in the level below. Figure 1 represents the influence of viral marketing in the pyramid form where each user at every level influences two users in the level below. Hence, the number of users influenced at a particular level can be determined by $2^{\text{level}-1}$ and the total number of users influenced by a single user may be as large as $2^{\text{level}} - 1$.

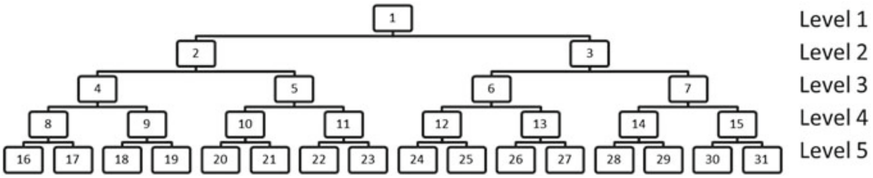


Fig. 1 Pyramid flow in viral marketing

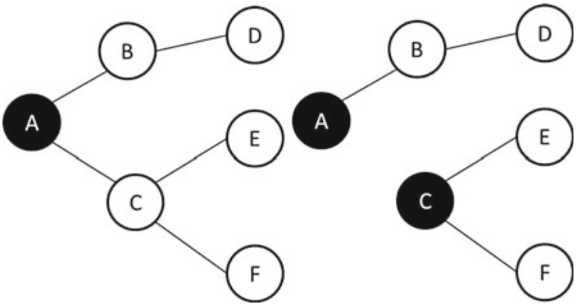


Fig. 2 Network graphs at time t_1 and t_2

A key limitation with existing viral market solutions is that they neglect the dynamic nature of the social relationships. As the social relationship changes over time, the relationship between users on the Web also changes. In social networks, new users register and existing users deactivate their accounts over time. The relationship between users also changes frequently. For example, a user A may be unknown to user B at time t_i, but may become good friend in time t_j for some $j > i$. Furthermore, the attributes of users can also change over time. For example, a user X's marital status might change between two time windows. The effectiveness of the transmission of knowledge depends upon recording these changes and adjusting accordingly to adapt to these changes. For example, the left graph in Fig. 2 representing a part of the social network at time t_1. Since there is a path to reach E and D from A, it is possible to influence E and D by influencing A. However, at time t_2, the network has a changed structure represented by the right graph in Fig. 2. In the new structure, the link between A and C is lost thereby making it impossible to influencing E from A and thus requiring C to be influenced separately at time stamp t_2. Thus, it is necessary to adapt to the changes in a dynamic network to maximize the influence of marketing.

In this paper, we propose a novel framework to effectively apply viral marketing in a dynamic social network.

2 Preliminaries

The selection of users in viral marketing is based on the concept of *network value*, where a user with a higher network value has more influence over other users in the network. The network value is determined by two key factors: the *intrinsic value* of a user and the *connectivity value* of the network. In this section, we detail these two important factors and then describe some other relevant concepts that are used throughout the paper.

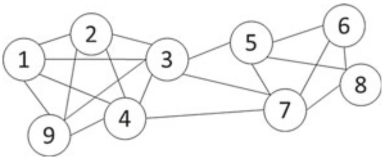


Fig. 3 Dense subgroups in the network

The intrinsic value is a normalized score calculated as a composition of various attributes of a user, which include things like recommendations made in a social network (e.g., Facebook), messages forwarded in an email system (e.g., Gmail), and so on. Some specific requirement of products or services may also be helpful to determine the intrinsic score. For example, the marketing of baby products may include the attribute of marital status, while marketing of ladies' perfumes may include the gender attribute. The activeness of the users in network also adds toward the calculation of the intrinsic value. The activeness in a social network could be calculated as the function of number of posts posted per day.

While the intrinsic value measures users' individual attributes, the connectivity value measures the network structure as a whole. It is a function of not only how well the user is connected in the network but also how well his/her neighbors are connected in the network. To effectively spread the influence in a network, it is necessary to hit the network from different ends, which is similar to the spread of an epidemic. The overall network can be usually regarded as a composition of a number of smaller strongly intraconnected subnetworks. Identification of such subnetworks and targeting users from each subnetwork is essential for fast spread of influence. As an example, in Fig. 3, nodes 1, 2, 3, 4 and 9 are strongly intraconnected while nodes 5, 6, 7, and 8 are strongly intraconnected, thereby forming two subnetworks, which are weakly interconnected. To effectively spread the influence, it would be necessary to select nodes from both subnetworks.

Another key concept used by the paper is *influence flow*, which is represented by a function of the live edges directed toward a user. An edge is regarded as live if its source is influenced. Thus, the probability of a user to get influenced in a particular time step is proportional to the number of influenced neighbors. Since the social network is inherently dynamic, we use subscript t to denote the temporal dynamics. Let S_t denotes the subset of users selected from the graph G_t at time step t. When there is a change on the network graph from $t-1$ to t, S_t should be changed accordingly. However, it is important that S_t does not deviate too much from the recent past due to a sudden change in a given time step. This actually is a reasonable expectation as a sudden change should mostly be due to an existence of some noise (e.g., a user accidentally removes a friend), which may be fairly common in a highly dynamic social network environment. Hence, the temporal knowledge obtained between the time windows provides an evolutionary outlook to system where it remains faithful to the current time window and not deviates significantly from the history [5].

3 The Evolutionary User Selection Framework

This section describes the framework for the evolutionary selection of users to maximize the viral marketing influence. The initial step for the selection of influencers in the network is to determine their network values, which aggregate the intrinsic values of individual users and the connectivity value of the network. Then a number of smaller strongly intraconnected subnetworks are identified to enable the selection of the users in different parts of the network. We adopt a threshold-based approach to compute the influence flow in the network, where an inactive user (not influenced) gets influenced by an active user (influenced) if the number of direct active friends of that user goes beyond the threshold [6]. We introduce a novel evolutionary metric to monitor the influenced users after every time window to determine the need for the additional selection of users with respect to the change in the graph in that time window.

3.1 Network Value Calculation

The network value measures a user's capacity as an influencer in the network rather than a customer. The network value is determined as a function of intrinsic value and connectivity, where the former is the composition of various attributes related to the product or service to market along with other relevant data available from the network and the latter represents how well a user is connected with other users in the network.

3.1.1 Intrinsic Value

Intrinsic value (I) represents how well the user can be associated with a particular product or service as an influencer for marketing. The base calculation requires the computation of Feature Mapping and Recommendation Score.

Feature Mapping (M) is used to determine if a particular user can act as an influencer for a particular product or service. To determine this, every product or service is represented by a set of attributes and a set of normalized values $\{Nv_k\}$ are used to represent the strength of the features for that user, where normalization is used to scale all the required features to the same level. Weights $\{W_k\}$ are provided to give different importance to different features as required. More specifically, the feature mapping M of a user is defined as

$$M = \sum_k W_k \times Nv_k \tag{1}$$

Recommendation score (R) represents if a user can be viewed as a good recommender. The recommendation score is calculated based on not only the capability of the user itself but also his or her friends' capability of forwarding recommendations to others. Intuitively, a user with a highly influential friend tends to be influential as well. Specifically, the recommendation score R of a user is defined as

$$R = \frac{n}{N} \times \sum_{k=1}^{n} \frac{r_k}{Tr_k} \tag{2}$$

where n denotes the number of friends that receive recommends from the user, N is the total number of friends of the user, r_k is the recommendations forwarded by the kth friend, and Tr_k is the total recommendations received by the kth friend.

Finally, the *Intrinsic value* I_i of the ith user is calculated as

$$I_i = (W_M \times M_i) + (W_R \times R_i) \tag{3}$$

where W_M and W_R are weights against feature mapping and recommendation score of the user, respectively.

3.1.2 Connectivity Value

Connectivity value (C) represents not only on how well a user is connected in the network but also how his/her friends are connected in the network. It is necessary to select users whose friends are also well connected in the network to ensure the flow of influence beyond the secondary level. This is essential to achieve an exponential growth as illustrated in Sect. 1. Hence, we compute the connectivity value C_i of the ith user as

$$C_i = \frac{\sum_{k \in S_i} |S_k|}{|S_i|} \tag{4}$$

where S_i is the set of friends of the ith user.

3.1.3 Network Value

Network Value (Nv_i) of the ith user is computed by aggregating its intrinsic value and connectivity value. The intrinsic value and the connectivity first normalized before being aggregated.

$$I_i = \frac{I_i - I_{\min}}{I_{\max} - I_{\min}} \quad C_i = \frac{C_i - C_{\min}}{C_{\max} - C_{\min}} \tag{5}$$

$$Nv_i = (W_I \times I_i) + (W_C \times C_i) \tag{6}$$

where I_i is the normalized intrinsic value, C_i is the normalized connectivity value, W_I and W_C are weights against intrinsic and connectivity values, respectively.

3.2 Relationship

Relationships between the users in a network are represented by the edges connecting those users. Presence of an edge represents the existence of a relation. The strength of the relation is specified by the edge weight. A positive value shows a relation favorable for the flow of influence. The calculation of this strength is based on the information available from the network.

Communication weight (Wc) specifies the strength of the relation based on the amount of communication between the users. It is determined by the sum of the communication between the two users.

$$Wc_{ij} = \sum c_{ij} \tag{7}$$

where c_{ij} represents each communication sent from user i to user j.

Recommendation weight (Wr) represents the trust between the users. The trust can be positive or negative if the source user recommended or did not recommend the other user, respectively. This is a very strong relationship factor and also provides additional knowledge in semi-supervised grouping.

$$Wr_{ij} = \sum r_{ij} \tag{8}$$

where r_{ij} represents each recommendation/nonrecommendation of user j by user i.

The weight (W_{ij}) of the relation/influence from user i to user j can be calculated as follows:

$$W_{ij} = (W_1 \times Wc_{ij}) + (W_2 \times Wr_{ij}) \tag{9}$$

where W_1 and W_2 are weights against communication weight and recommendation weight between the users, respectively.

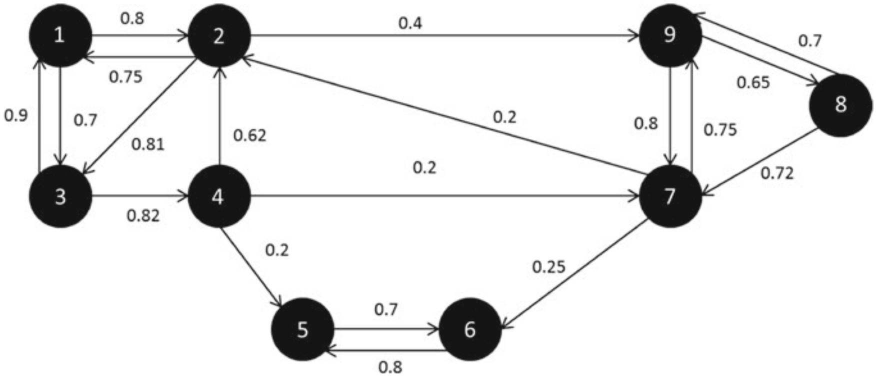


Fig. 4 A connected network example

3.3 Group Identification

A network as a whole can be represented as a composition of multiple dense subgraphs. These implicit groups have a strong association within the group and a lesser association between the groups. Dividing the network into such groups can be attained using graph-partitioning-based clustering algorithms, which identify groups of nodes that are strongly intraconnected and weakly interconnected. The intraconnectivity strength of a group bolsters the spread of influence in that group. Selecting influential candidates from each group allows attacking the network from different ends. This approach allows a faster spread of influence by dividing the network into smaller connected subnetworks.

Classical graph-partitioning-based clustering algorithms typically involve the calculation of the Eigen-decomposition of a graph [7], which has a cost of $O(N^3)$, where N is the total number of nodes in the graph [8]. Hence, it is computationally infeasible to directly apply these algorithms to large-scale social graphs. As our goal is not to precisely identify the clusters of nodes, we propose a fast graph partition algorithm for group identification based on the concept of connected components. The classical connected component algorithm uses a breadth-first (or depth-first) search algorithm to identify the connected subnetworks in a larger network [9]. However, this does not directly fulfill our requirement of identifying strongly connected groups or subnetworks in a network. For example, Fig. 4 shows a connected network. Directly applying the connected component algorithm on this network results in a single network as all the nodes in the network are connected.

To resolve this issue, we make the following modification of the classical algorithm. Specifically, before running the algorithm, we create an abstract view of the existing network by ignoring the weak edges. The resulting view is a network (V, E') where V represents the nodes in the original network and $E' \subset E$, i.e., subset of strong edges from the original set E. We use the box plot [10] approach to achieve outlier robustness when determining the threshold to assign edges into strong or

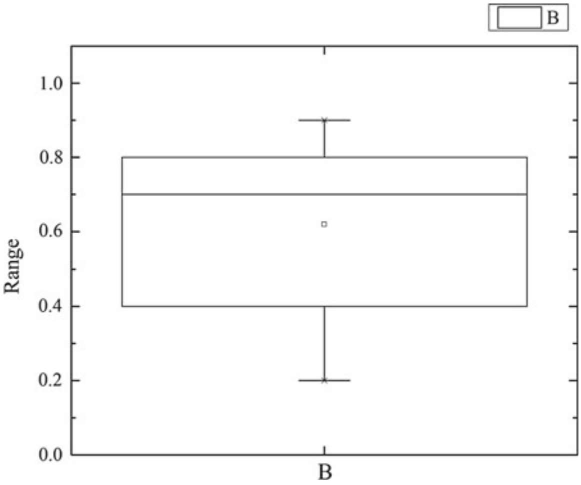


Fig. 5 Box plot of edge weights

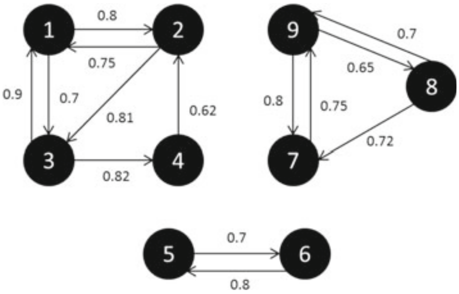


Fig. 6 Abstract view of the network

weak categories. As shown by the box plot in Fig. 5, the example network has a mean edge weight of 0.62. Ignoring edges below the mean results in an abstracted view shown in Fig. 6. Applying the connected component algorithm on this abstract view gives three distinct subnetworks. The nodes in each subnetwork are connected by strong edges, which represent a potential to direct an influence flow within that network. The graph traversal principle guarantees the time complexity to be linear in the number of nodes of the graph and hence achieves performance with orders of magnitude better than graph-partitioning clustering algorithms. The connected component philosophy also assures the identified groups to be internally connected.

Prior knowledge of users may also be leveraged to enhance the grouping process. For example, it is necessary to group users with close associations together. Close associations can be represented by coauthorship in the authorship dataset, trust–distrust in trust-based dataset, and followership in the micro blogging dataset. This prior knowledge might not always be conspicuously represented in the network. We adopt an award-penalty mechanism to incorporate the prior knowledge into the group identification process:

$$E'_{ij} = E_{ij} + R_{ij} W_{\text{reward}} - P_{ij} W_{\text{penalty}} \tag{10}$$

where E_{ij} denotes the original weight of the edge, $R_{ij} \in \{0, 1\}$ represents whether E_{ij} is a must link (i.e., i and j should be grouped together), $P_{ij} \in \{0, 1\}$ represents whether E_{ij} is a can-not link (i.e., i and j should not be grouped together), W_{reward} is the reward weight, and W_{penalty} is the penalty weight.

3.4 Influence Flow

The flow of the influence in the network is directed in accord to the threshold-based model specified by Kleinberg et al. [6]. A user is influenced if its influence value is higher than the threshold specified. This influence value (Iv) is a function of the number of influenced neighbors and the strength of the relation.

$$Iv = f(Ni, \sum_{k \in S_{\text{ie}}} W_k) \tag{11}$$

where Ni represents the total number of influenced neighbors, W_k denotes weight of the k^{th} edge, and S_{ie} represents the set of edges from influenced neighbors.

This function shows the ability of the user to get influenced based on the current state of the network. This function is represented by two functions viz. f_p and f_{s}. f_p is the percentile function of the *live edges* and f_{s} is the function of the total live edge strength with respect to the network.

An incoming edge is termed as a *live edge* if its source is an influenced user and the destination is not. The function f_p represents the percentile strength of the live edges for a user, i.e., the influence of the live edges with respect to the other incoming edges for a user

$$f_p = \frac{\sum_{k \in S_{\text{ie}}} W_k}{\sum_{k \in S_e} W_k} \tag{12}$$

where S_{ie} represents the live edge set. S_e denotes the set of all edges for the user, i.e., $S_{\text{ie}} \subset S_e$. W_k represents the weight of the k^{th} edge.

The value of f_p is compared with the threshold value θ set in the model. The selection of θ is based on empirical analysis of the dataset conforming closest to the natural flow of influence. The other function f_{s} represents the strength of live edges.

$$f_{\mathrm{s}} = \sum_{k \in S_{\mathrm{ie}}} \tag{13}$$

The value from f_{s} is compared with the average incoming edge strength in the network ($\mathrm{Te}_{\mathrm{avg}}$)

$$\mathrm{Te}_{\mathrm{avg}} = \mathrm{Me} \times \frac{\mathrm{Te}}{\mathrm{Tn}} \tag{14}$$

where Me is the current median edge weight in the network, Te represents the total edges in the network, and Tn denotes the total nodes in the network.

$$f_{\mathrm{s}} > \theta \text{ and } f_p > \mathrm{Te}_{\mathrm{avg}} \tag{15}$$

The impetus in using these two threshold measures is to give comparable opportunity for each user to be influenced. The function f_p mathematically favors the users with less number of neighbors, while function f_{s} favors one with large number of neighbors. The logical combination of these two functions in Eq. (15) provides a balance to the flow of influence.

3.5 Dealing with Dynamic Changes of the Network

The social network is not static. There are continuous changes like the addition of new users or deletion of the old ones. These changes include not only the change of relationship between users but also the attribute change of individual users. Nevertheless, these important changes go unrecorded in static graphs, which may significantly affect accuracy of user selection in viral marketing.

As the social network environment is highly dynamic and complete autonomous, changes should be treated as norms instead of exceptions. Meanwhile, many changes could be introduced due to various noises (e.g., accidentally adding or removing a friend). Hence, the framework should be robust enough to overcome the noises while being able to adapt to the changes. To achieve this dual objective, we propose to follow the temporal smoothness principle to cope with changes in the network, which demands the selection of users to respect the current snapshot of the network while not deviating dramatically from the recent past.

$$Nv = (1 - \alpha) \times Nv_{\mathrm{historical}} + \alpha \times Nv_{\mathrm{current}} \tag{16}$$

$$Wij = (1 - \alpha) \times Wij_{\mathrm{historical}} + \alpha \times Wij_{\mathrm{current}} \tag{17}$$

Adapting to this dynamic nature of the network is facilitated by growth rate (g). Growth rate measured at every time step represents the strength of the influence in the network. This strength is represented as

$$g_t = \frac{Vi_t}{Vu_t} \tag{18}$$

where g_t denotes the growth rate in time step t, Vi_t represents the influenced users in t, and $Vu_t = V - Vi_t$ representing the noninfluenced users in t with V specifying the total number of users in the network.

At each time window with the change in the network structure a new provisional growth rate is calculated against the new potential set of users for the next iteration

$$g'_t = \frac{Vi'_t}{(Vu'_t - Vi'_t)} \tag{19}$$

where g'_t represents the provisional growth rate at time step t with Vi'_t representing the potential list of influenced users and Vu'_t is the total number of noninfluenced users after t.

The potential list is based on the current visible users in the network with the network value (Nv) near threshold. If $g'_t < g_t$, then the growth rate would potentially decline in the next window. To counter this we calculate the number of users required to maintain the growth rate g_t.

$$Vd = (g_t \times (Vu'_t - Vi' - t)) - Vi'_t \tag{20}$$

where Vd is the additional number of users required.

We directly market the Vd best possible users from the network exclusive of Vi'_t to maintain the growth rate. This adaptability ensures the flow of influence maintained in the continually changing network.

3.6 The Algorithm

The algorithm of the proposed model can be broken down into three distinct blocks viz. *initialization*, *temporal update*, and *influence flow*. Algorithm 1 represents the evolutionary marketing function. This function selects the call to initialize or update the network based on time window. Algorithm 2 denotes the initialization block. This block is called when there is no prior network information or the previous network information needs to be overwritten with the current network. An existing network is updated using Algorithm 3. The flow of influence is reevaluated using block 4. The flow of influence in the network is described by Algorithm 5.

Algorithm 1 Evolutionary User Selection

```
  function EVOLUTIONARY($G'$, $w$)
Input: The Time window $w$
Input: The Network Graph update $G'$
     if $w = 1$ then
        $G = G'$
        call function InitializeInfluence($G$)
     else
        $G$ = UpdateNetwork(G,G')
        call function EvolutionaryInfluence(G,w)
     end if
  end function
```

Algorithm 2 Initialize_Influence

```
Input: The network graph $G$
  $\mathcal{D} = \phi$                                          ▷ direct market list
  $\mathcal{C}$ = IdentifyGroups($G$)
  for each $c \in \mathcal{C}$ do
     $n$ = # top users to select from $c$
     if  $n > 0$ then
        while $i < n$ do
           $v$ = getTop($c$, $i$)                          ▷ extract $i$th ranked $v$
           if $v$ is not influenced then
              $\mathcal{D} = \mathcal{D} \cup v$
              if state of $v$ not visited then
                 set $v$ state to visiting
              end if
           end if
           $i \leftarrow i + 1$
           if $i > |c|$ then
              break
           end if
        end while
     end if
  end for
  $\mathcal{V} = \mathcal{D}$                                   ▷ Initialize the visiting list
  for each node $v \in \mathcal{D}$ do
     set $v$ as influenced
     if state of $v$ is visiting then
        add neighbors of $v$ to $\mathcal{V}$
     end if
  end for
  update the potential list $\mathcal{P}$ with $\mathcal{V}$ using eq (15)
  call function InfluenceFlow
```

Algorithm 3 Update_Network

```
  function UPDATENETWORK($G$, $G'$)
Input: The network Graph $G(V, E)$
Input: The network Graph update $G'(V', E')$
Output: The updated network Graph G
    for each $v \in V \cup V'$ do
        if $v \in V$ & $v \in V'$ then
            update Nv using eq (16)
        else if $v \in V$ then
            update Nv with $\mathrm{Nv}_{\text{current}} = 0$ using eq (16)
        else if $v \in V'$ then
            add $v$ to $V$
        end if
    end for
    for each $e_{ij} \in E \cup E'$ do
        if $e_{ij} \in E$ & $e_{ij} \in E'$ then
            update $Wij$ using eq (17)
        else if $e_{ij} \in E$ then
            update $Wij$ with $Wij_{\text{current}} = 0$ using eq (17)
        else if $e_{ij} \in E'$ then
            add $e_{ij}$ to $E$
        end if
    end for
  return $G$
  end function
```

Algorithm 4 Update_Influence

```
  function EVOLUTIONARYINFLUENCE($G$,$w$)
Input: The network Graph $G$
Input: The time window $w$
    update the potential list $\mathcal{P}$ eq (15)
    calculate potential growth rate $g'$ eq (19)
    if $g' < g$ then
        calculate additional users $Vd$ using eq (20)
        directly market $Vd$
    end if
    call function InfluenceFlow
  end function
```

Algorithm 5 Influence_Flow

```
function InfluenceFlow
    $\mathcal{V} = \phi$                                                          ▷ visiting list
    $\mathcal{P} = \phi$                                                          ▷ Potential list
    $\mathcal{T} = \phi$                                                          ▷ temp visiting list
    while $(size(\mathcal{V}) > 0 \| size(\mathcal{P}) > 0)$ & $(nW \neq$ true$)$ do    ▷ $nW$ =new window availability
        for each $v \in \mathcal{P}$ do
            if $f_s > threshold$ & $f_p > threshold$ then
                set $v$ as influenced
            end if
        end for
        update growth rate $g$ using eq (18)
        for each $v \in \mathcal{V}$ do
            if state of $v$ not visited then
                add $v$ neighbor to $\mathcal{T}$
            end if
        end for
        append $\mathcal{V}$ with $\mathcal{T}$
        update $\mathcal{P}$ with $\mathcal{V}$ using eq (15)
    end while
end function
```

4 Experiments

We conduct a set of experiments to assess the effectiveness of the proposed evolutionary user selection framework for viral marketing. For comparative purpose, we include two nonevolutionary user selection approaches. More specifically, the first approach, referred to as *Non Evol Group*, uses the group identification feature, which allows it to attack the network from different ends by selecting most influential candidates from each group. The second approach, referred to as *Non Evol No Group*, selects the most influential candidates from the entire network independent of their group presence. Following the same naming convention, we refer to the proposed approach as *Evol Group*. The key metrics we evaluate include the total influences in the network and the ability to sustain the flow of influence in the network.

4.1 Dataset and Experiment Setup

The experiments are conducted over a real-world dataset collected from large-scale social network. As the data is collected over a long period of time, the temporal dynamics are clearly captured by the dataset.

4.1.1 FriendFeed

The dataset that we have used is a well-known micro blogging and social network service "FriendFeed" (http://friendfeed.com). The microblogging feature in association with user's ability to follow a particular *entry* (like in well-known Twitter) or "like" or "comment" on one (like Facebook) provides a vast pool of social data [11]. The structure of the dataset available comprised of followers, entries, comments, likes, users, and networks for the date monitored between August 1, 2010 and September 30, 2010. Mining the information based on the contents of the entries and comments is ignored as it is out of scope for current research work. Current research monitors the quantitative interaction between the users on the micro blogging service. The users are represented as the nodes of our graph.

Directed edge is present from user B to user A if

A follows B or
A comments on an entry by B or
A likes an entry of B

These interactions represent the ability of the user B to influence user A directly or indirectly. The data was processed to ignore orphaned entries and users with no interactions in the network for our experimental needs. The resulting network was represented by 22,817 nodes and 303,785 edges. For evolutionary analysis the dataset between August 1, 2010 and September 30, 2010 was divided into 12 windows, thus each window representing data for 5 days.

Network Value Calculation: Friendfeed provides extensive information on the social activity of a user. The quantitative information like number of followers, total entries in the social network, comments made and received, likes made and received for the entries constitute toward the network value of the user in our graph. These quantitative measures are represented in a normalized format and incorporated with the associated weights. This measure provides intuitive information on the activeness of the user in the network and the overall influence they represent.

Edge Weight Calculation: The edges in our graph exhibit the social relationship between the users in the network. The number of comments and likes shared between the users along with the follower information is used to represent this relationship. The edges representing fellowship are identified as *must-link edges*. In parallel to HEP-Th these *must-link* edges are used in the semi-supervised grouping.

The evolutionary model reads each window at a time in confidence of the evolutionary principles, whereas the nonevolutionary models look at a single snapshot of aggregated data, there by losing the temporal knowledge. The evolutionary algorithm updates the graph after each window by conforming to the evolutionary concept of maintaining the low history cost and representing high snapshot quality [5]. The nonevolutionary-based algorithms are allowed to run till they converge, i.e., either *influence* the entire network or no more *influence* update available. Keeping with the scope of this paper, for simplicity the algorithms are allowed to run at three equally spaced out threshold levels ($\theta = 0.25, \theta = 0.5, \theta = 0.75$).

Figure 7 shows the total number of users influenced by each model at different threshold levels θ and Fig. 8 compares the number of iterations the flow persisted in the network. Both graphs show the supremacy of the evolutionary model over the nonevolutionary-based model. The evolutionary model influences far more users than its nonevolutionary counterpart and runs for large number of iterations. This high performance can be attributed to the capture of the temporal change in the network missed by the nonevolutionary-based models that looked at static aggregates' snapshot. The large number of iterations provides the confidence of a longer run of *influence* as well as larger coverage of the influence in the network.

Fig. 7 FriendFeed-influenced users

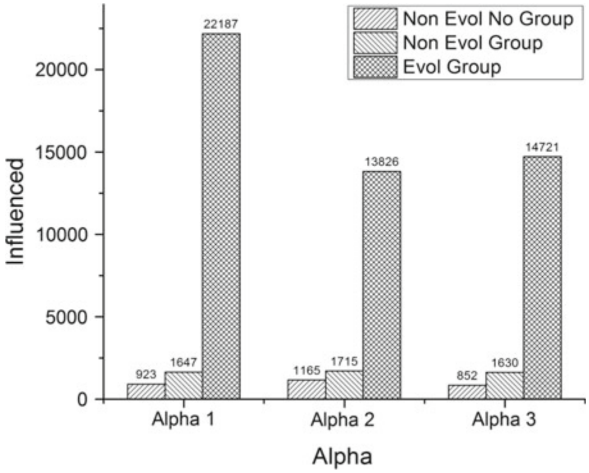
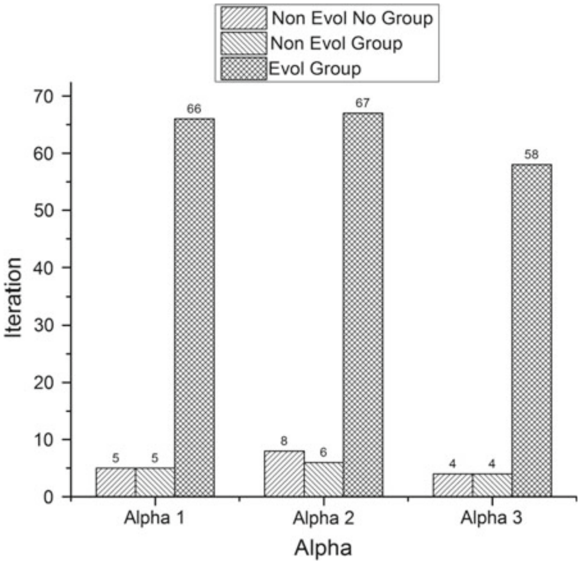


Fig. 8 FriendFeed-iterations



This ability to run for a longer period ensures the *influence* flow kept alive which is evident from the growth rate representation in Fig. 9. It can be observed that the growth rate for the nonevolutionary model represents a single spike. The growth rate increases till it discovers the network from the initial selection whereby after reaching the saturation level the growth rate steadily decreases. In contrast, the growth rate for the evolutionary model is represented by multiple spikes. The multiple spike results from the continual learning of the network as the new data comes in at each time window and self-adjusting the growth rate to keep the *influence* flowing in the network, thereby restoring the confidence of the *influence* flow in the network. The inability to grow from initial selection could be the result of loss of relational information in the aggregated graph and improper selection of the users. Figure 10 demonstrates a similar analysis for the total users' influences against the iterations. The nonevolutionary models grow quickly and reach the saturation level, whereas the evolutionary model steadily increases in step mode. The step level represents the knowledge of new data flowing-in for a window. This suggests that the supremacy of the proposed evolutionary algorithm in influencing the users in coauthorship network with an assurance of maintaining the flow of *influence* by adjusting to the incoming data.

The supremacy of the evolutionary model, specifically in this data, can be contributed to the fact that the data is highly volatile as expected from any social network. This temporal volatility is successfully captured by evolutionary model as it by-par outperforms the nonevolutionary model.

4.1.2 HEP-Th

The second dataset we use in our experiments is the HEP-Th dataset. HEP-Th dataset represents the information on papers in theoretical high-energy physics from arXiv (www.arxiv.org). The collaboration graph represents the relationship between journals, papers, and authors as represented in Fig. 11. The structure of this dataset provides a key feature of coauthorship in a social network. The data captures the relationship between the authors. The authors are represented as the nodes of network graph. An edge is created from author B to author A if (1) A refers a paper by author B or (2) A coauthors a paper with author B. These interactions represent the ability of the author B to influence author A directly or indirectly. The data was preprocessed to remove missing data and corresponding authors with no connection to the network were ignored. Authors of the papers not present in the author list were added with synthetic identifiers. The resulting graph is a close approximation of the original graph. Amongst the 49 distinct research areas available in the dataset the experiments were run only with reference to High-Energy Physics, Atomic Physics, History of Physics, Computational Physics, Numerical Analysis, and Classical Physics.

Network Value Calculation: The network value of the author in this network is calculated with reference to (1) Number of papers in the targeted area of influence, (2) Number of citations received, (3) Total number of downloads in the first 60 days,

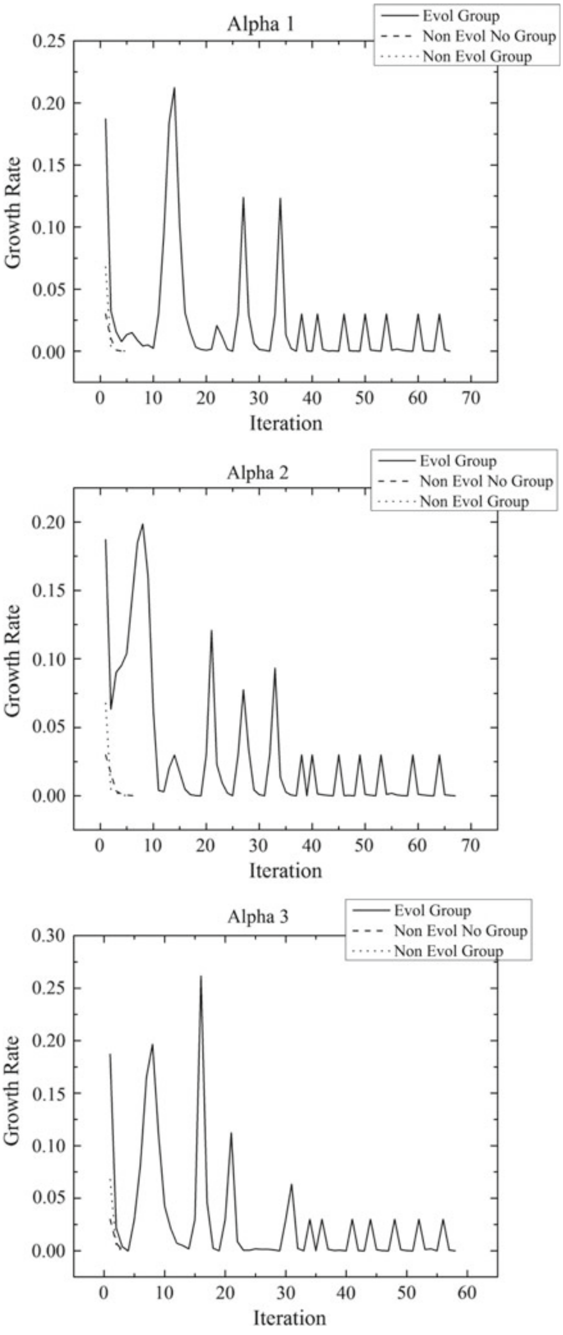


Fig. 9 FriendFeed-growth rate against iteration at each θ

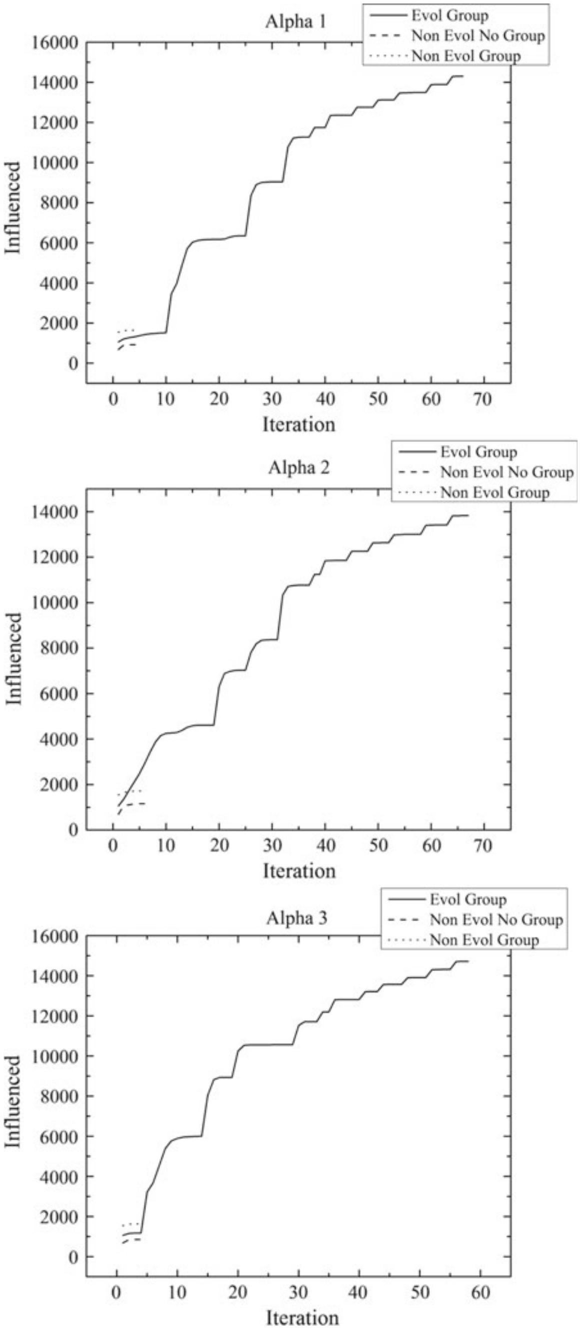


Fig. 10 FriendFeed-influenced against iteration at each θ

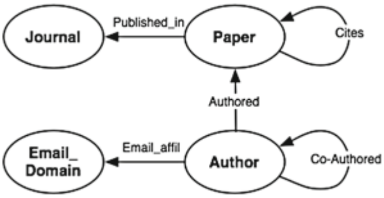


Fig. 11 HEP-Th schema [12]

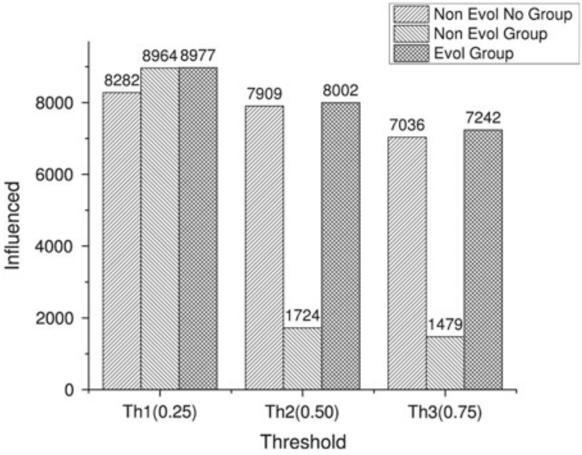


Fig. 12 # Influenced users versus threshold θ

and (4) Total number of papers published. Each factor is normalized based on the overall network and combined with the corresponding weight.

Edge Weight Calculation: The edges in our graph exhibit the relationships between authors. The weight of the edges between the nodes in this graph is presented as a combination of the coauthorship count between the two authors and cites reference count of node A to node B for an edge directed from node B to node A. The edges representing coauthorship is termed as *must-link* edges. *Must-link* edges are used for semi-supervised grouping. Similar to the network value calculation, the factors contributing toward the edge weight are normalized based on the network and combined with the associated weight

Figure 12 shows that the evolutionary algorithm outperforms both the nonevolutionary algorithms with respect to the total number of users *influenced* at all threshold levels. The difference in the total users influenced by the group-based nonevolutionary can be directed toward the alteration of the flow based on the change in the *influence* threshold level as pointed out earlier. Figure 13 demonstrates that the nonevolutionary model runs over a longer period of time as it learns about the new data after each window, thereby providing the confidence of a longer run of *influence*. Figure 14 shows that the growth rate for the nonevolutionary model represents a single spike. The growth rate increases till it discovers the network from the initial selection whereby after reaching the saturation level the growth rate steadily decreases. In contrast, the growth rate for the evolutionary model is represented by

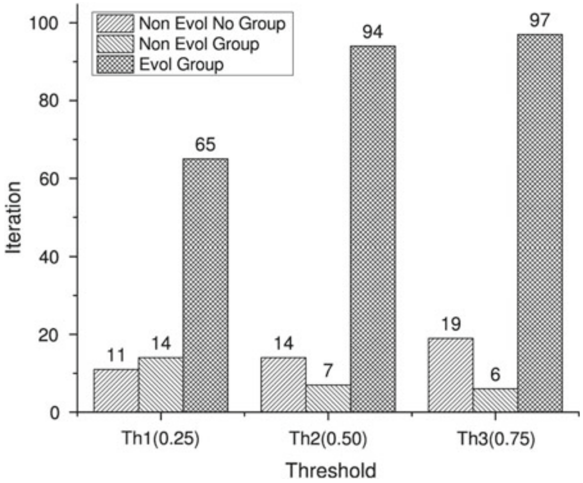


Fig. 13 # Iterations versus threshold θ

multiple spikes. The multiple spike results from the continual learning of the network as the new data comes in at each time window and self-adjusting the growth rate to keep the *influence* flowing in the network, thereby restoring the confidence of the *influence* flow in the network. Figure 15 demonstrates a similar analysis for the total users' influences against the iterations. The marginal gain in total influenced users of evolutionary over nonevolutionary model can be contributed toward the slowly changing nature of the network.

4.1.3 Epinion

Epinion is the third and final dataset that we used for our experimental analysis. It is one of the best known knowledge-sharing sites and termed as a "web of trust" for its trust relationship-based network. It allows users to post reviews in addition to rating. Users interact with each other by rating reviews and also by listing reviewers they trust. The "web of trust" employs the service to present reviews from trusted users first [13]. The architecture of the service provides information on authorship of articles, trust/distrust information, and product ratings by users which can be invaluable for any social network analysis model. For our experiment we captured the data on this service from January 2001 to September 2003. For evolutionary model this information was split into buckets of quarters thereby providing 11 windows for our analysis.

The data was processed to capture the information relevant to the relationship between the users in the network. The users are represented as the nodes in our corresponding graph representation with the edges symbolizing the relationship between the users. We represent a strong edge from user A to user B, if user A lies in the trust list of user B. An edge lies from user A to user B, if user B provides a positive

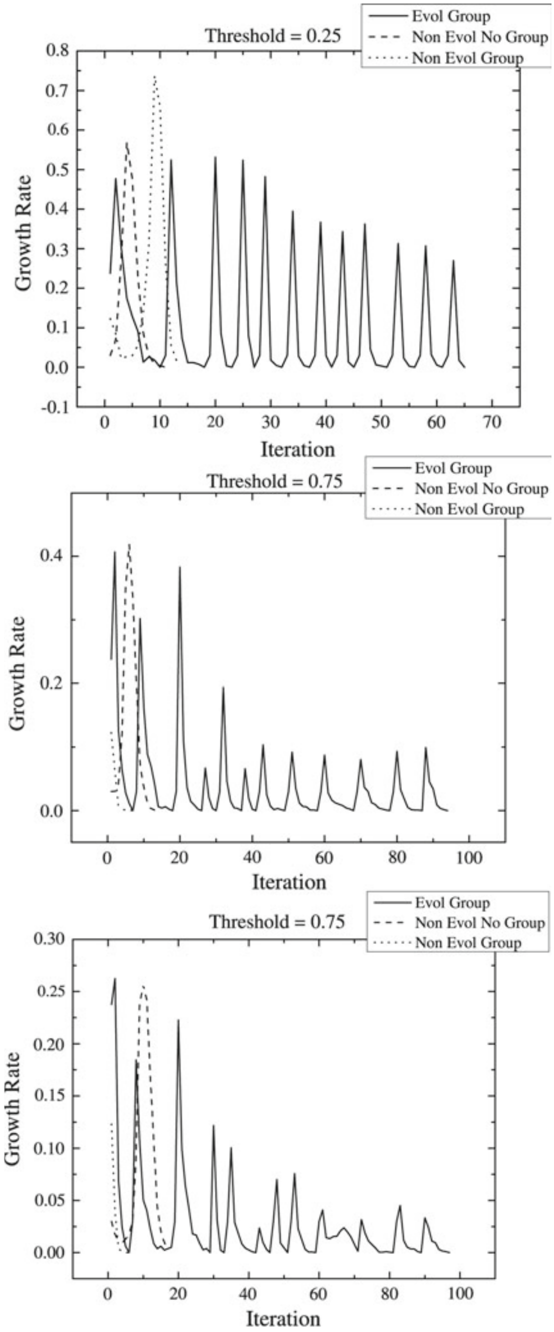


Fig. 14 Growth rate versus threshold θ

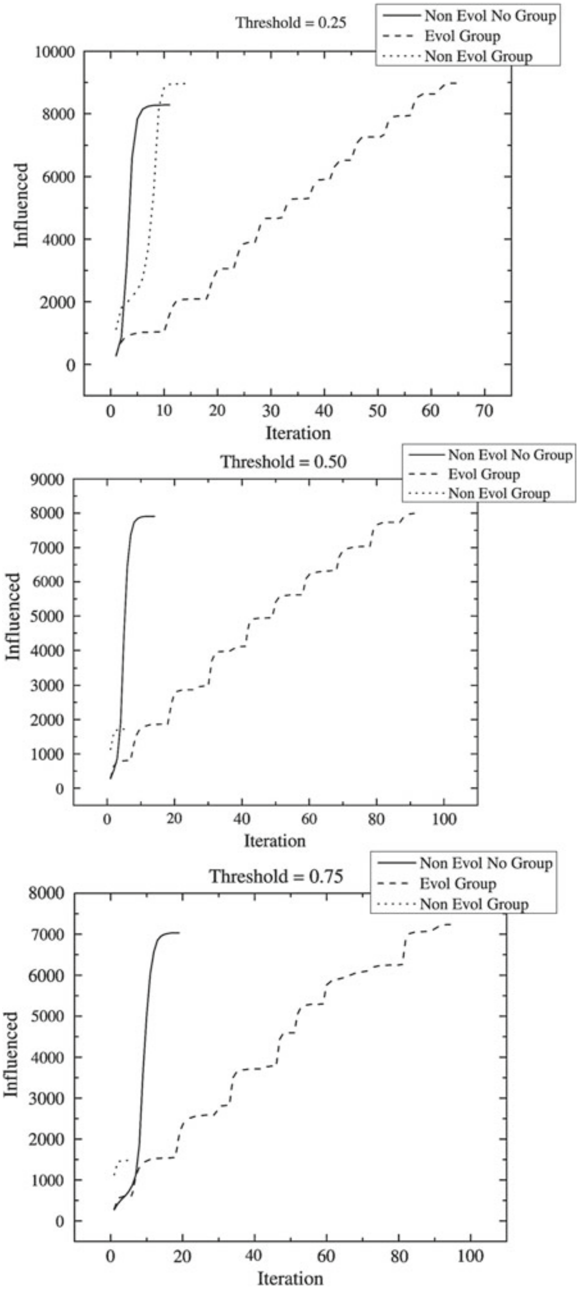


Fig. 15 # Influenced user versus threshold θ

rating to A's review. A negated edge lies if user B provides a negative rating to A's review. This approximate representation of the Epinion service resulted in 158,142 users and 5,053,088 edges.

Network Value Calculation: Epinion signifies trust-based influence information. This information can be perceived to determine the influence level of an Epinion user. Knowledge of trust/distrust count along with overall ratings and article counts is utilized to determine the network value of the user. These characteristics are normalized at the network level and combined with their corresponding weights.

Edge Weight Calculation: The edges provide a trust-based relation. Total ratings plus the trust/distrust information are used in the calculation of the edge weight. An edge representing a higher trust level is termed as *must-link* edge and one with higher distrust level is termed as *can-not link* edge. Semi-supervised grouping exploits this additional information.

Evolutionary model performs better than its comparative model in the total number of users influenced in different threshold levels as shown in Fig. 16. However, the nonevolutionary group-based model also performs significantly better than the greedy model. Approximately 97,000 and 85,000 more users are influenced by evolutionary and nonevolutionary group-based model in contrast to the greedy model. This huge difference between group- and nongroup-based model can be based on two possible reasons: (1) The influential nodes being concentrated at one end of the network thereby obstructing the opportunity to grow, (2) The graph being highly disconnected thereby reducing the possibility to reach different parts of the network. The group-based model works well against such cases. Figure 17 demonstrates that the total number of iterations used by evolutionary is higher as compared to the other models in line with the earlier observations. Growth rate exhibits the same behavior

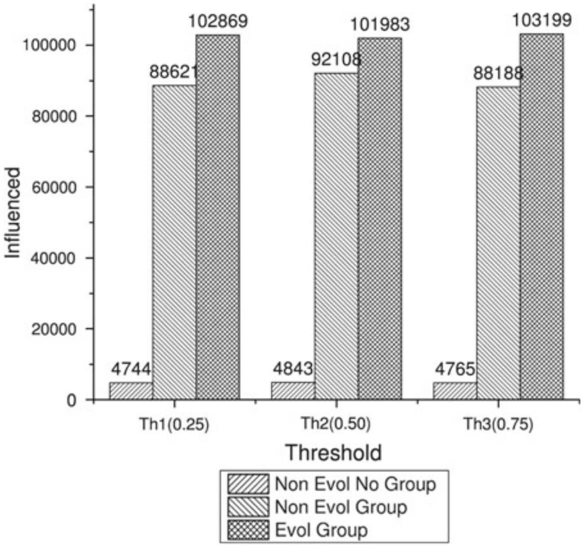


Fig. 16 # Influenced users versus threshold θ

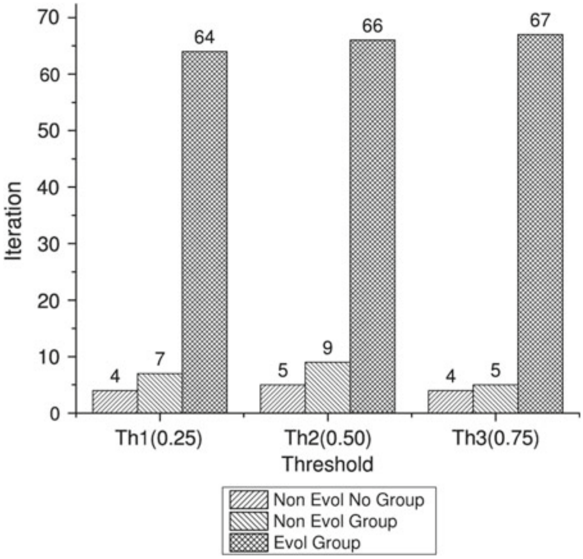


Fig. 17 # Iterations versus threshold θ

as FriendFeed for all the models as shown in Fig. 18. Figure 19 represents the total number of users influenced after each iteration. Likewise information on the early saturation of the nonevolutionary-based model and step-based growth of the evolutionary model can be inferred. However, another interesting aspect to capture here is the large number of users initially influenced for the group-based model as compared to the greedy model. This is only possible if the network is highly disconnected, which results in large number of groups and hence the large initial subset.

5 Related Work

The traditional marketing looks at customer as an individual and not as a part of the society with an ability to influence others. Viral marketing uses the network value of the users in the network in contrast to the customer value used by the direct marketing concept [1]. Domingos and Richardson [14] described network value as the composition of the connectivity of the user in the network and the users ability to influence other users in the network. The selection of the users can also be supported by the concept of predictive rating. Domingos and Richardson [15] proposed probabilistic approach similar to predictive rating determination. The model analyzes similar users liking and actions to determine the feasibility of marketing to a particular user in analogous to predict rating of the user from the analysis of similar users.

The virtual network for the viral marketing analysis can be represented by graph with nodes representing the users and edges representing the relationship between the users. Transaction logs and the event lists provide a large amount of data for

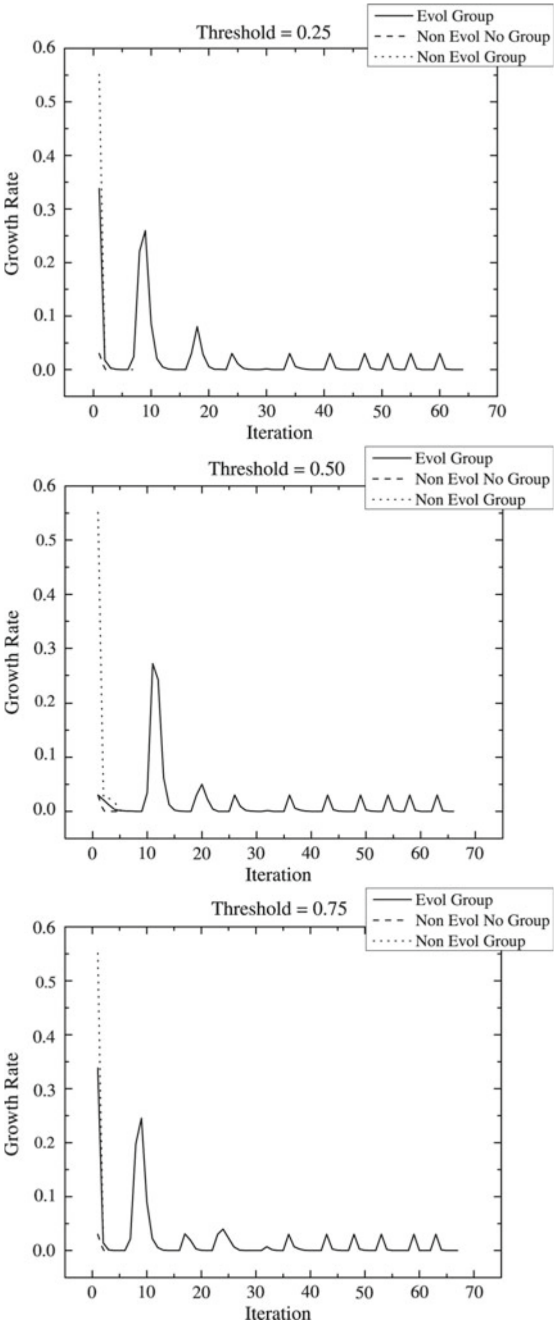


Fig. 18 Growth rate versus threshold θ

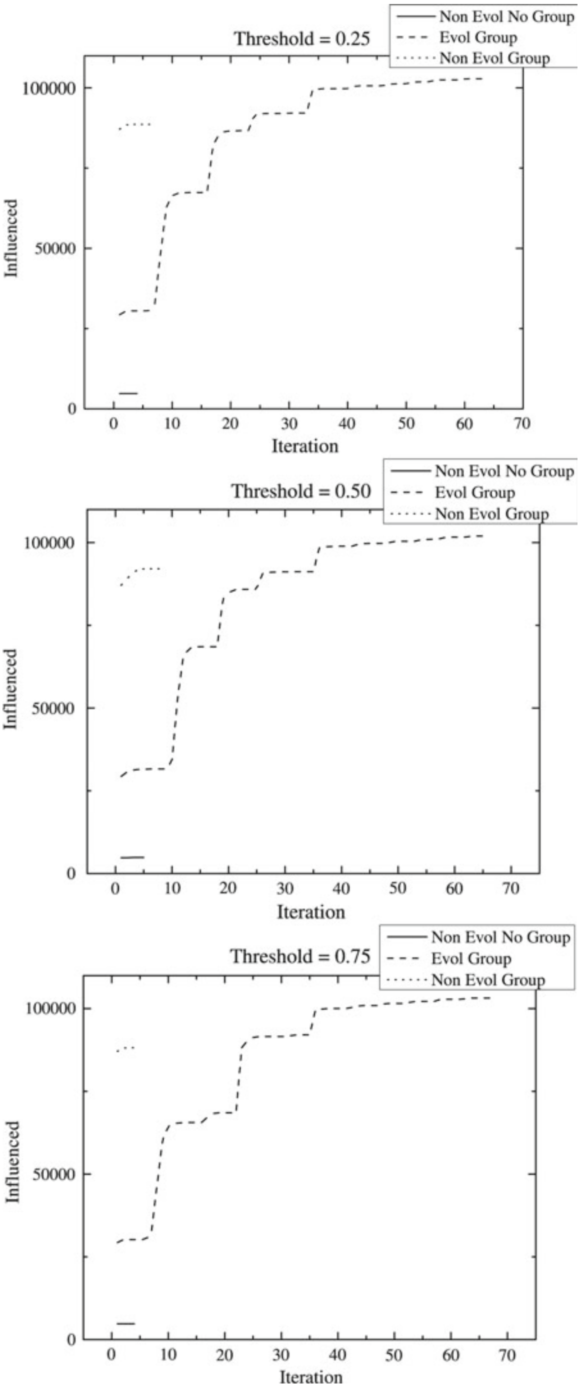


Fig. 19 # Influenced user versus threshold θ

analyzing the relationship between the users. Aalst and Song [16] proposed the concept of process mining in social network analysis to determine the relationship between the users based on their interactions. The concept makes the use of a triplet consisting of case, activity, and user. If an event (c1, a1, u1) followed by (c1, a2, u2) represents an interaction between user u1 and u2 for the case c1. However, if there is another user u3 who shares the same responsibility as u2 but there is no event of u3 on the same case following u1's activity then it depicts that u1 and u2 are closely bonded in the network as compared to u1 and u3.

A user interacts with only a small subset of the users in the entire network. It is necessary to identify the dense relationship between different subset of users in the network. These dense relationships can be represented as a group in the network. Leskovec et al. [17] analyzed that 77 % of the recommendations comes from within the group. The identification of such groups is necessary to select users from each group to facilitate the flow of influence in viral marketing. Long et al. [18] proposed a technique of identifying such groups in a network by addition of a virtual node. This node represents group features identified from the empirical data. Using Gaussian similarity, the users in the network that share those features are connected to those virtual nodes forming a group. Chi et al. [19] proposed a concept for the group identification using community factorization method. They proposed the method on social networks like blogosphere where the structural and temporal dynamics of the graph is different than the Web with short life time dens subgraph. The factorization method extracts the communities and their temporal behavior and assists in identifying the long-term graph structure from a series of short-term graphs.

Previous analyses of the social network for marketing were based on static graphs with data captured over a period of time. These analyses do not take into account the temporal changes. Social relationships change with time, with the addition or removal of users from the network and change in the relationships between the users. These changes affect the flow of influence in the network. Thus, integrating the temporal nature of the graph is of prime importance. Evolutionary clustering proposed by Chakraborti et al. [5] provided a new dimension for clustering in a dynamic graph. The gist of this concept is that the clustering should be faithful to the current data and should not deviate dramatically from the previous time step. The concept proposes the computation of sequence of clusters in each time window. The cost of the clustering is represented as the combination of the snapshot quality and the history cost. The model considers the object feature similarity along with the time-series similarity function. The model penalizes for the deviation in clustering with respect to the history data but not with respect to the new data. The input for the model is the matrix representing the relation between each pair of objects at each time step and the output is the clustering with respect to the new matrix and history. Chakraborti et al. [5] model is restricted to clustering which can be adapted for various dynamic graph analyses like user selection for viral marketing in dynamic graph. Sharan and Neville [20] worked on temporal relationships for predictive analysis in dynamic graphs. They summarized the dynamic graph with the weighted static graphs and then incorporated the weighted links in the Relational Bayes Classifier to moderate the influence of the attributes. The model tries to exploit both the relational and temporal aspects in the domain of predictive data mining. It is based on the concept of homophily in relational domains to mine inference about nature of relationship with the recent ones conferring more homophily than earlier one.

It is necessary to efficiently predict the flow of the influence in a dynamic graph to compare the states of the network at different time steps to conform to the evolutionary concept. Kempe et al. [6] provided an approximation model to predict the flow of influence in the network. They proposed two approaches viz. linear and probabilistic. The linear approach maintains a threshold value of influence for each user. An inactive user (not influenced) gets influenced by an active user (influenced) if the summation of the weights of direct active friends goes beyond the threshold value for that user. Here the weight represents the ability of an active user to influence an inactive user. The independent cascade model follows a probabilistic approach. This model allows an active user to influence an inactive user in only one step. The probability of the inactive user getting influenced because of the active user is based on empirical data. If the inactive user does not get influenced in the following steps that active user does not get another chance to influence that inactive user.

6 Conclusion and Future Work

We considered the problem of selection of the users in a dynamic network to maximize the flow of influence and proposed a threshold-based evolutionary framework. This framework selects the users in a network by conforming to the evolutionary principle and prolonging the flow of influence by dynamically adapting to the change in the network. Experiments on HEP-Th, Friendfeed, and Epinion demonstrated the superiority of this model in comparison to the nonevolutionary models. The proposed model not only maximizes the influence flow but also dynamically adapts to the change in the network, thereby maintaining the flow in the network.

The current framework was implemented with threshold-based models. It would be interesting to study the implementation of probability-based models. The calculation of the network values was based on quantitative parameters. This can be extended to include subjective parameters. The framework can be extended to incorporate text mining on the data to provide more intuitive information for determining the relationship between the users in the network.

References

1. Domingos P (2005) Mining social networks for viral marketing. IEEE Intell Syst 20:80–82
2. Merkle (2010) View from the social inbox 2010. http://www.merkleinc.com/user-assets/Documents/WhitePapers/Social%20Inbox%202010%20WPaper%20Final.pdf
3. Gen Y (2009) Study shows Gen Y wants more control in email exchanges
4. Epsilon (2008) Asia Pacific consumer email survey. Technical report

5. Chakrabarti D, Kumar R, Tomkins A (2006) Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 554–560
6. Kempe D, Kleinberg J, Tardos V (2003) Maximizing the spread of influence through a social network. In: International conference on knowledge discovery and data mining, pp 137–146
7. Xu X, Long B, Zhang Z, Yu PS (2007) Community learning by graph approximation. In: Proceedings of the seventh IEEE international conference on data mining, pp 232–241
8. Xiang T, Gong S (2008) Spectral clustering with eigenvector selection. Pattern Recognit 41:1012–1029
9. Hopcroft J, Tarjan R (1973) Efficient algorithms for graph manipulation. Commun ACM 16:372–378
10. Michael F, David HC, Boris I (1989) Some implementations of the boxplot. Am Stat 43:50–54
11. Celli F, Di Lascio FML, Magnani M, Pacelli B, Rossi L (2010) Social network data and practices: the case of friendfeed. In: International conference on social computing, behavioral modeling and prediction, Berlin (2010)
12. Hep-th—kdl—umass amherst. https://kdl.cs.umass.edu/download/attachments/3440884/hepth-schema.png?version=1&modificationDate=1345733950033
13. Massa PAP (2006) Trust-aware bootstrapping of recommender systems. In: Proceedings of ECAI 2006 workshop on recommender systems, pp 29–33
14. Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Eighth international conference on knowledge discovery and data mining, pp 61–70
15. Domingos P, Richardson M (2001) Mining the network value of customers. In: Seventh international conference on knowledge discovery and data mining, pp 57–66
16. Aalst WMvd, Song M (2004) Mining social networks: uncovering interaction patterns in business processes. In: Desel J, Pernici B, Weske M (eds) Business process management. Springer, Berlin, pp 244–260
17. Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web (TWEB) 1:5-228–5-237
18. Long B, Xu X, Yu PS, Zhang Z (2007) Community learning by graph approximation. In: Proceedings of the 2007 seventh IEEE international conference on data mining, pp 232–241
19. Chi Y, Zhu S, Song X, Tatemura J, Tseng BL (2007) Structural and temporal analysis of the blogosphere through community factorization. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pp 163–172
20. Sharan U, Neville J (2007) Exploiting time-varying relationships in statistical relational models. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, pp 9–15