NON-INTRUSIVE VIRAL MARKETING BASED ON PERCOLATION CENTRALITY

Complete Research

Mochalova, Anastasia, Katholische Universität Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany, Anastasia.Mochalova@ku.de

Nanopoulos, Alexandros, Katholische Universität Eichstätt-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Germany, Alexandros.Nanopoulos@ku.de

Abstract

Viral marketing can become an effective marketing technique in social networks. Initiating from a set of influential seed users, it can activate a "chain-reaction" driven by word-of-mouth. The effectiveness of viral marketing lies in the fact that it conveys an implied endorsement from social ties. However, not all viral marketing campaign become successful - some stop even before the number of activated users of the network reaches critical mass. In this paper we propose a new approach to viral marketing that will allow marketers to increase the performance of the stopped campaign by initiating new "waves" of the campaign. But in order to not overwhelm users that were already exposed to the initial campaign, the activation of seeds is performed in a non-intrusive way by suggesting users to follow recommendations of their friends. The proposed method for seed selection for the next "wave" is based on percolation centrality that takes into account already activated nodes and uses their social connections to boost the word-of-mouth propagation that already happens in the network.

Keywords: Social Networks, Viral Marketing, Influence Maximisation, Percolation Centrality.

1 Introduction

Our decisions are often influenced by other people: our friends, family, specialists and other people we trust (Muchnik, Aral, and Taylor, 2013). We often look for recommendations from trusted acquaintances because they convey an implied endorsement from social ties in the form of word-of-mouth (WOM) (Wang and Chang, 2013). This is a fundamental principle of viral marketing that relies on people passing along a marketing message, thus spreading WOM over an online social network (Bonchi et al., 2011). Viral marketing represents a low-cost high-impact marketing strategy that targets a small set of influential users (called *seeds*) that may cause a "chain-reaction" that spreads WOM to a large portion of a social network. For this reason, the resulting problem of optimising seed selection (a.k.a. the influence maximisation problem) has attracted a lot of attention (Kempe, Kleinberg, and Tardos, 2003).

One of the main cost of viral marketing campaign involves the activation of the seeds, which can become a hard task, considering that influential seeds attract a lot of attention and, thus, several marketers may attempt to use them for their campaigns. Second, these seeds might not be optimal to participate in a viral marketing campaign - research (Leskovec, Adamic, and Huberman, 2007) shows that providing excessive incentives for customers to recommend products could backfire by weakening the credibility of the very same links they are trying to take advantage of. There are also limits to how influential high centrality nodes are in the recommendation network: as a person sends out more and more recommendations for a product, the success of these recommendations inevitably declines, since an excessive number of product recommendations made by a user may be perceived as advertising by other users and, thus, lose the effect of word-of-mouth.

Due to the aforementioned problems, existing seed-selection strategies (e.g., Kempe, Kleinberg, and Tardos (2003)) might become sub-optimal in the long term because these strategies have been designed having a single application of a viral-marketing campaign. The reason is that with more attempts to activate the same (influential) seeds, they become harder to activate. Moreover, as explained above, the effectiveness of WOM originated from these seeds eventually decreases. Thus, marketers hoping to develop strategies for word-of-mouth advertising should analyse the topology and interests of the social network of their customers (Leskovec, Adamic, and Huberman, 2007). The seeds selected for the campaign should be, first of all, near the target group of customers, and second, the way the seeds are activated should change so the credibility of their recommendations does not change and their activation is less *intrusive*, i.e., not losing the WOM effect.

Another problem of existing viral marketing strategies is the fact that a campaign often stops before it reaches a critical mass of activations. How viral a message becomes depends on its content and its viral properties (Dobele, Toleman, and Beverland, 2005). However, viral properties on their own cannot guarantee the success of each message. For this reason, marketers should become able to boost a viral-marketing campaign that stopped, by engaging more members of a social network that will continue the propagation of WOM. This can be done by initiating new “waves” of the campaign that will reach a bigger audience.

In this paper, we propose a new *non-intrusive* approach to viral marketing that will help to overcome these problems. The idea of non-intrusive viral marketing is to incorporate the main benefit of standard viral marketing to seed activation, i.e., engage the power of user-to-user communication to activate seeds. In this case seeds are selected in the close proximity of the target group of the campaign and activated in a non-direct way though recommendations of already activated neighbours that are less central and, thus, cannot reach a wide audience as these seeds would. This can be done by showing what their friends are doing and suggesting to try this as well (a method that is already a part of many online social networks, such as Facebook’s “suggested events”). In this way, marketers can enforce the WOM propagation that already happens in the network and, thus, reduce the costs needed to activate the more central and expensive users as seeds.

Potential seeds are selected using *percolation centrality* score proposed by Piraveenan, Prokopenko, and Hossain (2013). This centrality score takes into account not only the network topology, but also already activated nodes. This measure allows marketers to use the knowledge they have about the viral marketing campaign and benefit from it to boost the performance of the campaign because nodes with high percolation centrality have high impact on further propagation of the WOM. This centrality score is also beneficial to marketing campaigns that have several “waves” because the centrality scores will be updated after each wave has finished, thus choosing the most suitable candidates for the next wave.

This approach to viral marketing has several advantages:

- It reduces the costs needed to activate seeds;
- It takes into account the target group;
- The seeds might be more willing to activate the others, because they perceive that their activation happened through social ties and not by being persuaded by marketers;
- The credibility of seeds is not affected;
- This type of marketing campaign will be perceived as less intrusive by users who already overwhelmed by the amount of advertisements they see - they might be put off if there is an excessive amount of advertising because they think that the more a company advertises, the worse the quality of advertised product is (Kirmani, 1997);
- It gives a possibility to reinforce a stopped viral marketing campaign.

In this paper we analyse the proposed novel approach to viral marketing by running simulation that use the data from a real social network. The approach implements and compares three centrality scores: degree (Freeman, 1979), betweenness (Freeman, 1980) and percolation (Piraveenan, Prokopenko, and Hossain, 2013). The empirical results show that non-intrusive marketing can generate a lot of activations and that percolation centrality outperforms other centrality scores.

The rest of the paper is organized as follows: In Section 2 the related work is presented. In Section 3 the examined problem and the proposed methodology is described. The design of the performance evaluation is presented in Section 4 which is followed by the presentation of empirical results and their discussion in Section 5. The paper is concluded by Section 6.

2 Related Work

Viral marketing has already proven to be an effective marketing technique in social networks (Leskovec, Adamic, and Huberman, 2007) and has attracted a lot of attention in recent research (Bonchi et al., 2011). Compared to more traditional ways, such as targeted marketing, viral marketing avoids the expense of contacting all members of a target group. In contrast, only a small number of influential seeds needs to be contacted in order for the message to spread widely. Often viral marketing campaigns are relatively inexpensive because customer networks take care of spreading the messages and no expensive media exposure needs to be purchased (Lans et al., 2010).

Viral marketing can be stated as an influence maximisation problem that is concerned with selecting the set of seed users in a social network who will initiate the spread of a piece of information and will cause the largest possible number of activations among remaining users (Kempe, Kleinberg, and Tardos, 2003). The seed set has a predefined size which corresponds to the cost of the viral marketing campaign: the larger the seed set size, the higher the cost. The problem of seed selection in online social networks is mainly a computing one – based on information available about the network, one needs to identify influential users to start the viral spread (Probst, Grosswiele, and Pfleger, 2013) and predict how this spread will diffuse in the network (Fang et al., 2013). Kempe, Kleinberg, and Tardos (2003) showed that the problem of selecting the optimum seed set is NP (non-deterministic polynomial time) and proposed an approximate greedy hill-climbing algorithm. Nevertheless, this approach requires knowledge of the strength of the connection of each connected pair of users u and v in the network. This strength is denoted as *influence factor* and expresses how much u can influence v and vice versa.

However, influence is intangible and a priori knowledge of influence factors is not possible in most real-world cases, thus the direct applicability of the greedy algorithm is limited. Influence factors can be estimated based on actions previously performed by users of the network (Goyal, Bonchi, and Lakshmanan, 2011). However, this approach can be used only when the recorded past actions are relevant to the current information cascade. Otherwise, the estimated influence factors will not be accurate. Additionally, for the same pair of users influence differs and depends on the information itself: a person might be an expert in one area but completely unaware of another. This factor is taken into account in topic-aware influence assessment (Barbieri, Bonchi, and Manco, 2013), where pair-wise influence factors are measured for different topics. Additionally, acquiring knowledge about influence factors may present problems related to privacy issues. For these reasons, recent research has also focused on selecting as seeds the users that have *central* position in the network structure, because information diffused by such users may have better chances to reach a larger part of the network (Hinz et al., 2011). To select users according to how central their position is, research in social network analysis has proposed to use a large set of centrality scores, such as degree, betweenness, closeness, and Eigenvector centrality (Newman, 2009). Centrality scores require knowledge only of the network structure (which is not difficult to obtain in online social networks) and not of influence factors.

Despite these advantages, existing approaches to viral marketing present some shortcomings. First of all, viral marketing relies heavily on seeds and their ability to activate and engage their social ties. However, research done by Leskovec, Adamic, and Huberman (2007) shows that providing excessive incentives for customers to recommend products could backfire by weakening the credibility of these seeds in eyes of their social ties making their recommendation less effective, thus, decreasing the effectiveness and spread of the viral marketing campaign. Second, users might become overwhelmed by the amount of advertisements they see - if there is a large amount of advertising coming from a company they start to think negatively about the quality of advertised product (Kirmani, 1997).

Third, most seed selection methods select seeds based on their importance in the whole network which might not always be the best approach because they treat all users as equally probable to get activated, thus, ignoring their inherent predisposition and the homophily - the fact that people with similar tastes tend to connect socially (Aral, Muchnik, and Sundararajan, 2013; Peres, Muller, and Mahajan, 2010). Recent research takes this into account, by introducing seed selection methods based on centrality of users relative to a specific part of the network known as target market (Mochalova and Nanopoulos, 2014; White and Smyth, 2003). Nevertheless, this method requires extensive knowledge of user preferences to identify the target market.

In this paper we consider a different situation when some users of the network (initiators) already became activated and the idea is to use this to company's advantage and try to activate their social ties by showing them the activations, that have already happened in the network, in form of a recommendation. In this case seeds will be selected among the extended neighbourhood of the initiators - the reason for this is, firstly, the fact that people are more likely to trust their social ties (Wang and Chang, 2013) and, secondly, the homophily (Aral, Muchnik, and Sundararajan, 2013; Peres, Muller, and Mahajan, 2010). Thus, the potential seeds will be selected among the extended neighbourhood based on their centrality scores - either, how central they are in the whole network (betwenness or degree centrality), or how important they are for the flow of information (percolation centrality). These centrality scores are described in Section 3.2.

3 Proposed Methodology

In this section we formally describe the problem and the methodology used.

3.1 Problem Description

A social network is represented by a simple undirected unweighted graph $G(V,E)$. V is the set of m nodes corresponding to the users of the network and E is the set of edges corresponding to the social ties among the users. Each node of V can be active or inactive. Initially, all nodes are inactive and once a viral campaign starts spreading, nodes exposed to propagated piece of information may become activated. The progressive spread of information is considered here, where activated nodes do not modify their opinions (Kempe, Kleinberg, and Tardos, 2003). For a node to become activated it is not enough just to be exposed to the information, but to accept it by performing an action, e.g. "Liking", posting, or sharing.

Initially few users of a network get activated (with or without direct influence from marketers). They, in turn, might activate some of their friends, who can later activate their friends, and so on until the initial wave of activations stops. Other approaches to viral marketing (e.g., Hinz et al. (2011) and Kempe, Kleinberg, and Tardos (2003)) focus only on the initial spread of activations but do not propose any actions when it stops. The problem is how to enforce this viral marketing campaign based on the knowledge available about the previous wave. This could be done by starting a new wave of information diffusion in the network and selecting new seeds for the next wave. However, because the campaign was already happening in the network, the potential seeds could have been already exposed to it. Thus, direct persuasion from marketers might have a negative effect on the willingness of these seeds (Kirmani, 1997) and their effectiveness (Leskovec, Adamic, and Huberman, 2007).

The seed selection for the next wave of the viral marketing campaign is done by, firstly, considering already activated nodes that form a set of *initiators* I - however, they are not the initiators of the new wave of the marketing campaign, they are all users that became activated initially. A non-intrusive viral-marketing campaign is based on these initiators: people are more likely to trust a personal recommendation given by a friend or trusted acquaintance (Jurvetson, 2000), in this case friends and friends of friends - they form an *extended neighbourhood N*. This extended neighbourhood represents a set of user that could potentially become activated next.

Non-intrusive approach is based on providing notifications to the extended neighbourhood N about the activations of initiators. These notifications can be perceived mostly as a suggestion from a friend or acquaintance and not as a direct advertising. Such notification mechanisms have been already implemented by some social networking sites, e.g. Facebook's "suggested events", Twitter's notifications about friends' following or retweeting. In this case the non-intrusive persuasion done by providing users in the extended neighbourhood with information about what their neighbourhood recommends will have better results, as these potential seeds will not consider it as advertisement but as a friendly suggestion.

Users in the extended neighbourhood of initiators $N(I)$ are potential seeds - among them the seed set is chosen by addressing the influence maximisation problem (Kempe, Kleinberg, and Tardos, 2003) that seeks a set of potential seeds S of k users of a social network that will be targeted and will maximize the expected spread of a given piece of information through the network. It is assumed that the size of S is controlled by the number k of members in it, where k is predefined and represents the cost to target the members of S. The total number of users of the network that will be activated during the spread initiated by S, is denoted as $A_k(S)$. Given the graph $G(V,E)$ that represents the structure of the network, the influence maximisation problem can be defined as the problem of finding the seed set S of size k with the maximum possible $A_k(S)$ value; i.e.:

$$S = \underset{U \subseteq V}{argmax} A_k(U), s.t. |S| = k \tag{1}$$

In contrast to seeding strategies examined in the related work (e.g. Hinz et al. (2011), Kempe, Kleinberg, and Tardos (2003)) not all potential seeds will become activated, because they may not have any direct incentive to do that. This reflects the non-intrusive nature of the proposed approach because the seeds selected to continue a stopped campaign are not forced but could be persuaded by the friendly suggestions. Activation of seeds in all simulation cases will follow the information diffusion process (Kempe, Kleinberg, and Tardos, 2003) similar to the other users and they will adopt the propagated information according to the following probability:

$$p_a = \gamma(1-\theta) + (1-\gamma)\frac{N_a}{N_n}, \tag{2}$$

where N_a represent number of activated users in the extended neighbourhood of the node, N_n is a total number of users in the extended neighbourhood of the node, and θ corresponds to *predisposition* of the user.

The probability of activation depends on two factors. First factor is the predisposition that a user has about the diffused information (product, brand, etc.), which is represented by θ that takes values in the range $[0,1]$, with values closer to 0 indicating more positive predisposition; i.e., node is more likely to get activated. Thus, the higher the value of $1-\theta$, the more probable is that a user gets activated.

Second factor is *peer pressure* represented by N_a/N_n - fraction of neighbours of the node that got activated. The more nodes in the neighbourhood are activated, the higher the N_a/N_n and the higher the pressure that node gets from his peers, and, thus, the more probable he is to get activated.

Variable γ is used to combine these two factors that influence probability of activation (predisposition and peer pressure). In our study we follow the natural assumption that γ is equal 0.5, which make both factors equally important for the activation probability. However, other values can be examined as well.

Factor $\gamma(1-\theta)$ is a value in the range $[0,0.5]$ because $1-\theta$ takes values in the range $[0,1]$ and γ equals to 0.5. Similarly, factor $(1-\gamma)N_a/N_n$ is a value in the range $[0,0.5]$ because N_a/N_n takes values in the range $[0,1]$ (0 - none of the neighbours are activated and 1 - all neighbours are activated) and $1-\gamma$ equals to 0.5. Thus, probability p_a is always in the range $[0,1]$.

The process of seed selection can be repeated several times after the previous wave of activations ceased. The currently activated nodes than become the set of *initiators* for the next wave of the viral marketing campaign, thus allowing the campaign to keep running.

The objective is to re-enforce the existing stopped viral marketing campaign by selecting the nodes from the extended neighbourhood to whom the suggestions will be shown (a suggestion is a notification that users from the initiator set have been activated in previous waves). To control the non-intrusive aspect of the approach we set a maximum number of users that will be exposed to the suggestion as 30% of the extended neighbourhood (this is refereed to as *exposure factor* later) - it is an upper bound because not all of them will follow the suggestion. The problem is to optimise the selection of nodes from the extended neighbourhood that will increase the spread in the following wave of the campaign.

3.2 Seed Selection

Existing research proposes selecting seeds according to how central their position is with respect to the network structure represented by graph G (Hinz et al., 2011). The intuition is that the more central a node is, the more likely it is to reach many other nodes. Different algorithms for computing centrality scores have been proposed, such as degree, betweenness, closeness, and Eigenvector centrality (Borgatti and Everett, 2006). Experimental comparison of some of these centrality scores indicated that they have comparable performance (Hinz et al., 2011). For this study we selected two global centrality scores: betweenness centrality, that showed one of the best performances (Goyal, Bonchi, and Lakshmanan, 2011), and degree centrality that has the lowest complexity (Borgatti and Everett, 2006).

Degree centrality (Freeman, 1979) computes the number of paths of length one that emanate from a node. The nodes with high degree centrality usually have increased activity and, thus, are more likely to engage in diffusion of information through the network.

Betweenness centrality (Freeman, 1980) computes the share of times that one node need another node to reach the third node via the shortest path. Nodes with high betweenness centrality are usually the nodes that connect otherwise unconnected parts of the network. Thus, they allow the access and propagation of the idea in several parts of the network at the same time.

However, to incorporate the knowledge available about the previous run of the campaign, a new centrality score is needed that is more specific and will take already activated nodes into account. *Percolation centrality* proposed by Piraveenan, Prokopenko, and Hossain (2013) satisfies this requirement. This centrality measure quantifies relative impact of nodes based on their topological connectivity, as well as their states. Thus, this measure takes into account not only the network structure, but also the position of potential seeds relative to already activated nodes. Nodes with high percolation centrality play a vital role in the diffusion of information.

Percolation centrality is calculated for each node v at time t using the formula (Piraveenan, Prokopenko, and Hossain, 2013):

$$PC^t(v) = \frac{1}{(m-2)} \sum_{s \neq v \neq r} \frac{\sigma_{s,r}(v)}{\sigma_{s,r}} \frac{x_s^t}{|\sum x_i^t| - x_v^t}, \tag{3}$$

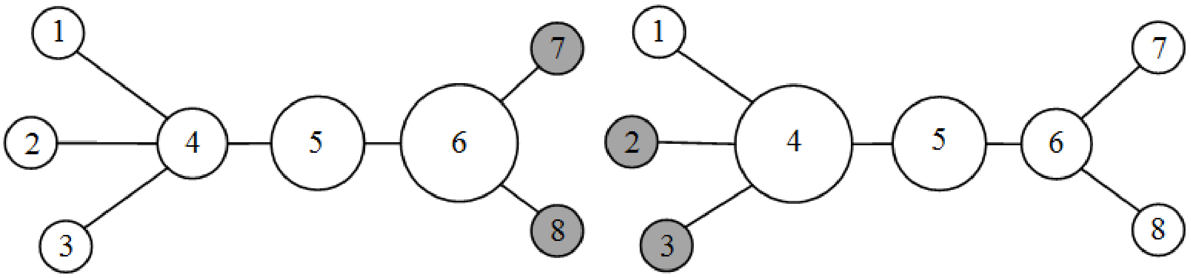
where $\sigma_{s,r}$ is number of shortest path from node s to node r and $\sigma_{s,r}(v)$ is number of shortest path from node s to node r that pass through v; x_s^t is state of node s at time t (in this case 1 is active and 0 is inactive).

Percolation centrality averaged over all possible single contagion sources reduces to betweenness centrality. In this case all shortest path become percolated paths, since all nodes are potential “sources” of percolation.

Let us consider an example of how percolation centrality assesses the importance of different nodes by looking at the Figure 1 (example is based on the sample graph provided in Piraveenan, Prokopenko, and Hossain (2013)). There a sample social network is given with 8 nodes connected by edges. In this network two nodes have been activated (7 and 8 for Figure 1a and 2 and 3 for Figure 1b). Now consider nodes 4 and 6. Both of them are centrally located and would have high betweenness centrality. If we calculate percolation centrality based on Figure 1a[1], the node 6 will have a much larger percolation centrality value. The opposite is true for Figure 1b - there node 4 has the largest percolation centrality value. As it can be

[1] The calculation can be done manually but it is quite tedious, that is why we do not include it here.

seen from the example, percolation centrality can vary significantly based on what nodes got activated and, therefore, it is more dynamic compared to other centrality measures, such as betweenness and degree centralities that are also considered in this paper.



(a) The nodes 7 and 8 in the right side of the network are activated.

(b) The nodes 2 and 3 in the left side of the network are activated.

Figure 1: Example of calculation the percolation centrality on a sample network. The size of the nodes corresponds to their percolation centrality values.

The seed selection method for all centrality measures is following: for each node $v \in N$ (extended neighbourhood) centrality score (degree, betweenness or percolation) $c(v)$ is computed based on the network structure (and already activated nodes for percolation centrality), the nodes then ranked based on their centrality scores (the higher the score, the more central the node is), and we select the top-k nodes with the highest scores as potential seeds. However, for percolation centrality only the shortest path of length 3 are considered because recent findings showed that influence propagation in viral processes in social networks often happens only within close proximity of the seeds (Cha, Mislove, and Gummadi, 2009).

3.3 Information Diffusion Model

To test the performance of the proposed multi-stage seed selection approach (See Section 3.2), diffusion processes are implemented based on the widely-studied diffusion model called Linear Threshold (LT) (Kempe, Kleinberg, and Tardos, 2003). In *Linear Threshold (LT) model* each node w is influenced by each of its neighbours v with a weight $i_{v,w}$ that is equal to the (normalized) influence factor that v has on w. Additionally, each node w has a threshold θ_w in the range between 0 and 1, which corresponds to the attitude of w. Attitude refers to the predisposition of a user in a social network to respond positively or negatively towards a diffused piece of information. Therefore, the higher θ_w is, the harder it is to activate w. The diffusion process of LT unfolds in discrete steps. In each step all nodes that were active before, remain active. An inactive node w is activated only if the total weight of its active neighbours is equal or exceeds its threshold θ_w. The process runs until no more activations are possible. Therefore, LT models the collective influence that each node receives from all its neighbours.

According to LT, a user becomes activated when enough users connected to it, become active, which represents a “social pressure”. The attitude of a user w in a social network corresponds to the threshold θ_w that is used in LT and quantifies the “resistance” of w to “social pressure”. This model was chosen because the concept of “social pressure” is used in seed activation process and, thus, suits the LT model better.

4 Performance Evaluation

In this section, we describe the design of our empirical study that compares the centrality scores presented in Section 3.2.

4.1 Data Set

For the empirical study, a real world data set was used that contains data crawled on Dec, 2008 from YouTube (Zafarani and Liu, 2009). The data set contains $15,088$ nodes connected by $153,520$ edges in the social network. The average node degree is 10.2. The data contains the information about contacts (i.e., connections) between nodes. For the tasks of seed selection based on centrality scores, an undirected unweighted graph representing the structure of this network was constructed.

In order to test the performance of proposed methods using the diffusion model described in Section 3.3, a corresponding directed and weighted graph was built. This graph contains influence factors and is used only to measure the performance of each seed selection method and not for selecting potential seeds themselves, which are selected based on the undirected unweighted graph mentioned above. To build the weighted graph, information about the contacts was used. Following the approach commonly used in related research by Kempe, Kleinberg, and Tardos (2003), it was assumed that influence that node v has on node w is proportional to number of contact that w has. The weighted graph with influence factors i_{vw} is used to calculate the incoming weights in LT model during the performance evaluation (see Section 3.3 for more details).

4.2 Performance Measures and Model Parameters

The performance of each method is calculated based on $A_k(S)$, i.e., the total number of users of the network that will be activated during the spread initiated by the seed set S with exposure factor k (see Section 3.1). Higher values of $A_k(S)$ correspond to better performance.

Three seed selection methods were tested[2]:

- Based on degree centrality
- Based on betweenness centrality
- Based on percolation centrality

Since each application of LT involves a probabilistic element (i.e., in each trial LT activate nodes with some probability determined by the influence factors in the way that has been described above), each measurement is repeated $1,000$ and the averages are reported.

The following parameters were used for the diffusion process:

- *Predisposition* θ_w of each node w in LT model is a random variable that follows Beta distribution, i.e., $\theta_w \sim Beta(\alpha, \beta)$. The reason why the Beta distribution was selected is that it is very flexible and by tuning its parameters α, β one can represent social networks with entirely different overall attitudes of their users. Therefore, within the same network, users have various attitudes and the Beta distribution allows to choose the tendency of the varying attitude scores. In the results the following sets of values were examined: i) $\alpha = 2, \beta = 2$ representing a network with most members having neutral attitude, as the mean value is 0.5, and some members having positive and some having negative attitude; ii) $\alpha = 5, \beta = 2$ representing a network with most members having negative attitude, because the mean value is high, thus making them harder to activate; iii) $\alpha = 0.5, \beta = 0.5$ representing a network with members that either have positive or negative attitude and almost no members with neutral attitude.
- *Exposure factor* k represents a fraction of the total number of users in the extended neighbourhood. The smaller the exposure factor, the more the non-intrusive aspect is preserved. Exposure factor is reported as a percentage of the overall number of nodes in the extended neighbourhood of initiators. The default value for exposure factor is 20%. However, the value of exposure factor is not a percentage of the whole network, but percentage of total number of users in the extended neighbourhood. Moreover, not all selected seeds will become activated (and, thus, participate in the initiation of the new wave) - the

[2] We also tested Eigenvector centrality, however it showed similar performance to betweenness centrality. Thus, for brevity, we omit results for Eigenvector centrality, which also offers a more clear comparison of centrality scores.

probability of them activating depends their inherent preference (see Section 3.1 for more details on activation probability), that in difficult network will make seeds difficult to activate.

- *Number of waves n* examined were 1, 2 and 3. The number of waves was chosen to be rather small because otherwise the campaign will become more intrusive and importunate; and the probability of activation decreases with repeated interactions (Leskovec, Adamic, and Huberman, 2007). The default value for number of waves is set to 3.

The results are delivered as a percentage of total number of users in the network.

4.3 Simulation Procedure

The following procedure (depicted in Figure 2) has been used to measure the performance of the selected methods:

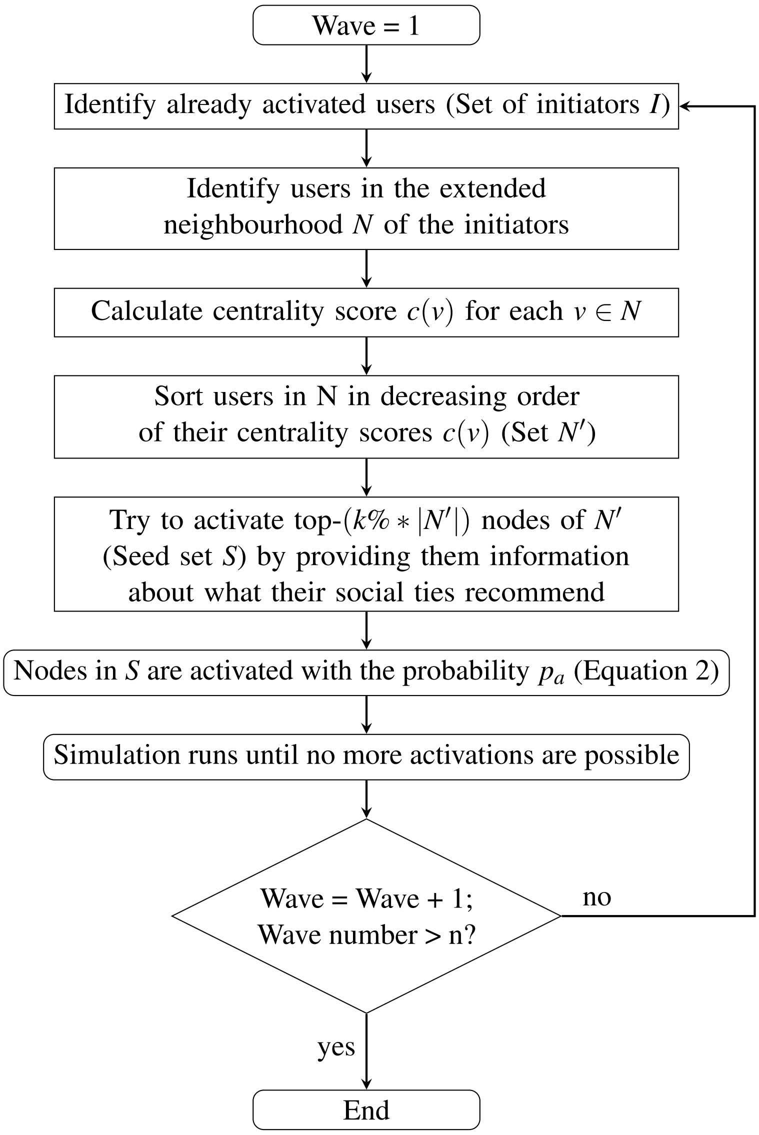

Wave = 1
Identify already activated users (Set of initiators I)
Identify users in the extended neighbourhood N of the initiators
Calculate centrality score $c(v)$ for each $v \in N$
Sort users in N in decreasing order of their centrality scores $c(v)$ (Set N')
Try to activate top-($k\% * |N'|$) nodes of N' (Seed set S) by providing them information about what their social ties recommend
Nodes in S are activated with the probability p_a (Equation 2)
Simulation runs until no more activations are possible
Wave = Wave + 1; Wave number > n?
no
yes
End

Figure 2: Flow-chart depicting the simulation procedure

1. The process starts with the identification of already activated users that form a set of initiators I. These nodes were activated either without direct influence from marketers or during the precious wave of the viral marketing campaign.

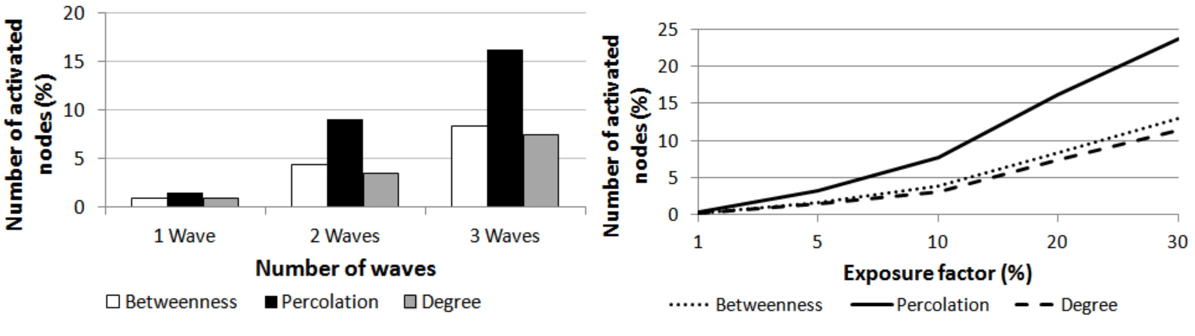
2. Once initiators are identified, the social network is analysed to find users in the extended neighbourhood of the initiators (friends and friends of friends) that form set N of neighbours.
3. For each node in the extended neighbourhood N, its centrality score is calculated according to the chosen centrality measure (see Section 3.2).
4. The users in extended neighbourhood N are sorted in the decreasing order based on their centrality scores, such that the most central node in the set will be on the top of the list. This way the set N' is formed.
5. From set N' top-$(k\% * |N'|)$ nodes, where k is exposure factor and $|N'|$ is the size of the set, are selected forming the seed set S. In real life scenario the activation attempt would happen by showing the users in the seed set suggestions and recommendations of users in their corresponding extended neighbourhoods. Selected seeds would become activated with the probability modelled by Equation 2 (see Section 3.1). For the simulation the suggestions were imitated by direct attempts to activate the users in the seed set that would succeed with the same probability (Equation 2).
6. The chain reaction of activations started by the activated seeds happens in the network following the Linear Threshold Model (see Section 3.3) with the predisposition distribution $\theta \sim Beta(\alpha, \beta)$ selected for this simulation. The process runs until no more activations are possible - that marks the end of the wave.
7. The procedure is repeated until the desired number of waves was reached.

5 Empirical Results

In this section, the performance of *degree*, *betweenness* and *percolation* centrality (see Section 3.2 for more details) are analysed for different number of waves and exposure factors and then compared to each other.

Neutral network ($\alpha = 2, \beta = 2$). The results for the network where most of the nodes have neutral predisposition towards the propagated idea are shown in Figure 3. Figure 3a denotes the results for different number of waves: method based on percolation centrality shows better performance compared to degree and betweenness centrality that show similar performance: the more waves there are, the bigger difference in performance between three scores there is.

Simulations with different exposure factors shown in Figure 3b demonstrate the advantages of percolation centrality once again - with the increase of the exposure factor increases the advantage of percolation centrality relative to betweenness and degree.

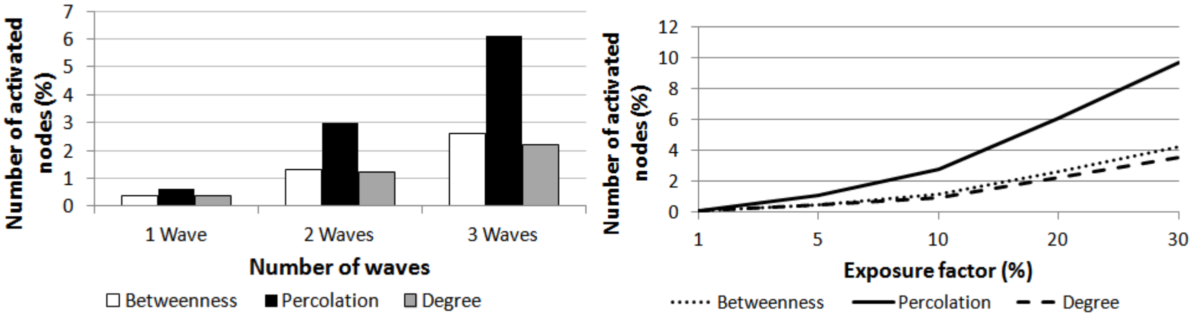


(a) Percentage of activated nodes vs. Number of waves (b) Percentage of activated nodes vs. Exposure factor

Figure 3: Activated nodes in neutral network ($\alpha = 2, \beta = 2$) for different centrality scores.

Negative network ($\alpha = 5, \beta = 2$). The empirical results for the network where most of the users are negative towards the propagated piece of information are shown in Figure 4. Simulations with different number of waves (Figure 4a) show similar dynamics to neutral network - with the increase of number of waves, the performance of all methods increases, however increase in performance of percolation centrality is noticeably larger than of betweenness and degree centrality. It is also noteworthy that the overall number of activated nodes decreased which is expected in the network where nodes are much less probable to become activated.

The results for different exposure factors in negative network are presented in Figure 4b and are also similar to the results of simulations run in the neutral network - with the increase of the exposure factor, more nodes become activated in the end and increase in the performance of percolation centrality is much larger than increase in the performance of degree and betweenness centralities.

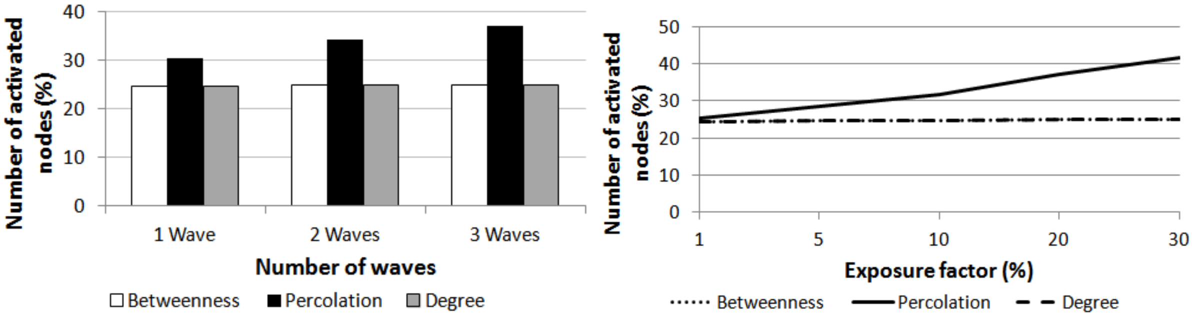


(a) Percentage of activated nodes vs. Number of waves (b) Percentage of activated nodes vs. Exposure factor

Figure 4: Activated nodes in negative network ($\alpha = 5, \beta = 2$) for different centrality scores.

Polarized network ($\alpha = 0.5, \beta = 0.5$). The results of the simulations performed in the network where most of the users are either positive or negative towards the propagated idea are presented on the Figure 5. Figure 5a shows the results of simulations with different number of waves. For betweenness and degree centrality increase in number of waves does not bring almost any improvement, but percolation centrality benefits from increase of number of waves.

Results of simulations with different exposure factors are shown on the Figure 5b. They show similar trend - betweenness and degree centrality are almost unaffected by increase of exposure factor, while percolation centrality shows increase in performance with increasing exposure factor.



(a) Percentage of activated nodes vs. Number of waves (b) Percentage of activated nodes vs. Exposure factor

Figure 5: Activated nodes in polarized network ($\alpha = 0.5, \beta = 0.5$) for different centrality scores.

6 Conclusions

In this paper we have presented a non-intrusive approach to viral marketing, which instead of activating seeds directly, does it by employing the power of user-to-user interactions. Non-intrusiveness stems from two facts: the first fact is that users are indirectly receive suggestions according to the actions of their neighbours, and the second fact is that exposure factor is kept low (small number of users in the network will see the recommendations). This way the quality of activated seeds and their credibility and, thus, their influence on other users of the network does not diminish (Leskovec, Adamic, and Huberman, 2007). Second, the cost of activating these seeds reduces. Additionally, this approach takes into account the network topology and the interests of the users which positively affects the success of the viral campaign. Lastly, by using non-intrusive viral marketing company reduces the probability of the campaign backfiring - this can happen if there is too much advertising causing people to start thinking negatively about the quality of the advertised product (Kirmani, 1997).

To use and benefit from knowledge available from the initial run of the viral campaign and already activated users in the network, seed selection based on percolation centrality (Piraveenan, Prokopenko, and Hossain, 2013) is used. This centrality measure give high scores to nodes that are central in the network topology but also important for the propagation of WOM from already activated user to yet unaffected areas. Percolation centrality is also a dynamic centrality score because it takes into account already activated nodes, thus, making it suitable for marketing campaigns with several waves.

The presented empirical results show that non-intrusive viral marketing is a viable approach to viral marketing and that it is possible to achieve a significant increase in performance of by adopting a novel centrality score - percolation centrality. The reason is that users with high percolation centrality act as bridges and vital for further propagation of information. However, although the consideration of more waves results in better performance, the campaign might become more intrusive with the increasing number of waves (Leskovec, Adamic, and Huberman, 2007).

The implications for marketers would be that using the non-intrusive marketing will allow them to reduce costs needed for activating seeds and get seeds that are more willing to activate the others, because they perceive that their activation happened through social ties and not by being persuaded by marketers. This type of marketing campaign will be perceived as less intrusive by users who already overwhelmed by the amount of advertisements they see. Additionally, non-intrusive marketing based on percolation centrality takes into account the target group and already activated nodes, thus, making the campaign more specific and gives a possibility to reinforce a stopped viral marketing campaign.

It is important, however, to address the ethical implications of viral marketing. The positive aspect of WOM-effect is that it gives users the power to create and diffuse the information in the more democratic way and not to be manipulated by the media that promotes what they think is important. However, viral marketing uses the social connections and tries to orchestrate the WOM propagation and, thus, it no longer happens on its own. The fact that many approaches can use the power of influential users to persuade the crowd can have ethical issues. However, our approach is non-intrusive and tries to minimise these ethical issues. The influential users in this case are not forced to engage, they are just given information about the marketing message and they are not manipulated into accepting it and spreading it further. We leave it to their free choice whether to accept the marketing message or not.

Another problem of viral marketing is privacy issues. Some approaches need to monitor activities of users to identify the most influential users. However, the approaches that are based on centrality scores (including our approach) are usually less privacy-sensitive as they just need the data about the existence of social ties which is often publicly available.

This work has some limitations. First of all, it is based on the idea that somebody in the network got activated without direct influence from marketers, which might be challenging for new products. Second, it is assumed that all activated nodes can be identified and used for the later waves of seed-selection, but it is not always true, e.g. some communications stay private. It is planned to extend this work, by enabling only a partial knowledge about activated nodes. Also this study will be extended to consider more centrality scores and more factors that affect the diffusion of information in the network.

References

Aral, S., L. Muchnik, and A. Sundararajan (2013). "Engineering social contagions: Optimal network seeding in the presence of homophily." *Network Science* 1 (02), 125–153.

Barbieri, N., F. Bonchi, and G. Manco (2013). "Topic-aware social influence propagation models." *Knowledge and information systems* 37 (3), 555–584.

Bonchi, F., C. Castillo, A. Gionis, and A. Jaimes (2011). "Social network analysis and mining for business applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2 (3), 1–37.

Borgatti, S. P. and M. G. Everett (2006). "A graph-theoretic perspective on centrality." *Social Networks* 28 (4), 466–484.

Cha, M., A. Mislove, and K. P. Gummadi (2009). "A measurement-driven analysis of information propagation in the flickr social network." In: *Proceedings of the 18th international conference on World wide web*. WWW '09, pp. 721–730.

Dobele, A., D. Toleman, and M. Beverland (2005). "Controlled infection! Spreading the brand message through viral marketing." *Business Horizons* 48 (2), 143–149.

Fang, X., P. J.-H. Hu, Z. Li, and W. Tsai (2013). "Predicting adoption probabilities in social networks." *Information Systems Research* 24 (1), 128–145.

Freeman, L. C. (1979). "Centrality in social networks conceptual clarification." *Social networks* 1 (3), 215–239.

— (1980). "The gatekeeper, pair-dependency and structural centrality." *Quality and Quantity* 14 (4), 585–592.

Goyal, A., F. Bonchi, and L. V. S. Lakshmanan (2011). "A data-based approach to social influence maximization." *Proceedings of the VLDB Endowment* 5 (1), 73–84.

Hinz, O., B. Skiera, C. Barrot, and J. U. Becker (2011). "Seeding strategies for viral marketing: an empirical comparison." *Journal of Marketing*.

Jurvetson, S. (2000). "What exactly is viral marketing?" *Red Herring* 78, 110–111.

Kempe, D., J. Kleinberg, and E. Tardos (2003). "Maximizing the spread of influence through a social network." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 137–146.

Kirmani, A. (1997). "Advertising repetition as a signal of quality: If it's advertised so much, something must be wrong." *Journal of advertising*, 77–86.

Lans, R. Van der, G. Van Bruggen, J. Eliashberg, and B. Wierenga (2010). "A viral branching model for predicting the spread of electronic word of mouth." *Marketing Science* 29 (2), 348–365.

Leskovec, J., L. A. Adamic, and B. A. Huberman (2007). "The dynamics of viral marketing." *ACM Trans. Web* 1 (1).

Mochalova, A. and A. Nanopoulos (2014). "A targeted approach to viral marketing." *Electronic Commerce Research and Applications* 13 (4), 283–294.

Muchnik, L., S. Aral, and S. J. Taylor (2013). "Social influence bias: A randomized experiment." *Science* 341 (6146), 647–651.

Newman, M. (2009). *Networks: an introduction*. Oxford University Press.

Peres, R., E. Muller, and V. Mahajan (2010). "Innovation diffusion and new product growth models: A critical review and research directions." *International Journal of Research in Marketing* 27 (91), 91–106.

Piraveenan, M., M. Prokopenko, and L. Hossain (2013). "Percolation centrality: Quantifying graph-theoretic impact of nodes during percolation in networks." *PloS one* 8 (1), e53095.

Probst, F., L. Grosswiele, and R. Pfleger (2013). "Who will lead and who will follow: Identifying Influential Users in Online Social Networks." *Business & Information Systems Engineering* 5 (3), 179–193.

Wang, J.-C. and C.-H. Chang (2013). "How online social ties and product-related risks influence purchase intentions: A Facebook experiment." *Electronic Commerce Research and Applications*.

White, S. and P. Smyth (2003). "Algorithms for estimating relative importance in networks." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 266–275.

Zafarani, R. and H. Liu (2009). *Social Computing Data Repository at ASU*. URL: `socialcomputing.asu.edu`.