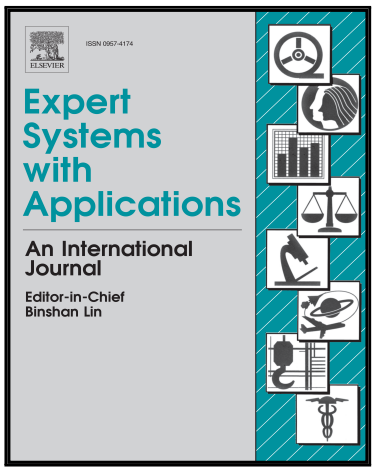
Accepted Manuscript



Active Viral Marketing: Incorporating Continuous Active Seeding Efforts into the Diffusion Model

Alon Sela, Dmitri Goldenberg, Irad Ben-Gal, Erez Shmueli

PII: S0957-4174(18)30246-X
DOI: 10.1016/j.eswa.2018.04.016
Reference: ESWA 11930

To appear in: *Expert Systems With Applications*

Received date: 7 December 2017
Revised date: 21 March 2018
Accepted date: 11 April 2018

Please cite this article as: Alon Sela, Dmitri Goldenberg, Irad Ben-Gal, Erez Shmueli, Active Viral Marketing: Incorporating Continuous Active Seeding Efforts into the Diffusion Model, *Expert Systems With Applications* (2018), doi: 10.1016/j.eswa.2018.04.016

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- A new diffusion model, which better fits real-world marketing scenarios, is proposed.
- Diffusion in this model relies on continuous active seeding efforts of the marketer.
- A scheduled seeding approach, which utilizes the states of nodes, is suggested.
- The importance of such an approach for the spread of products is demonstrated.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Active Viral Marketing: Incorporating Continuous Active Seeding Efforts into the Diffusion Model

Alon Sela[a,b], Dmitri Goldenberg[a], Irad Ben-Gal[a], Erez Shmueli[a,*]

[a] *Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv, Israel*
[b] *Department of Industrial Engineering, Ariel University, Ariel, Israel*

Abstract

Existing viral-marketing network models commonly assume a preliminary phase in which a marketer actively infects a subset of social network's users, represented by nodes, followed by a passive viral process, in which nodes infect other nodes without external intervention. However, in real-world commercial scenarios, substantial efforts are often invested by companies to promote their products, suggesting that the adoption of products is rarely the consequence of a viral spread alone.

Under this observation, this paper proposes a new diffusion model, named Active Viral Marketing, which better fits real-world marketing scenarios, where adoption of products relies on continuous active promotion efforts by the marketer. In the proposed model, the success of a marketing attempt to infect a potential customer (uninfected node), depends on the number of adopting friends (infected neighbors) of this user, assuming a user is more likely to adopt a product if more of his/her friends have already adopted it, while taking into account that social influence diminishes over time due to a memory-loss effect.

The paper further proposes a set of heuristics to schedule the marketing attempts. The main idea behind these heuristics is to utilize the information on the dynamic adoption-states of neighbor nodes, in addition to the static social network topology, when choosing the next node to seed. An extensive experimentation demonstrates how the proposed seeding heuristics improve

*Corresponding author

Email addresses: alonse2012@gmail.com (Alon Sela), dmitrig@mail.tau.ac.il (Dmitri Goldenberg), bengal@tau.ac.il (Irad Ben-Gal), shmueli@tau.ac.il (Erez Shmueli)

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

the adoption rate of products by 30%-75% in comparison to existing state-of-the-art methods that mainly rely on the network topology.

Keywords: Influence Maximization; Information Diffusion; Viral Marketing; Scheduled Seeding

1. Introduction

Online social networks offer a powerful tool for information sharing with friends, family and colleagues. In this aspect, they enable individuals to spread their messages passively through a viral process that might resemble the spread of a virus. Clearly, this property of online social networks also has a financial implication, since it can be utilized by companies and individuals that seek to advertise their products (we are using the terms products and services, interchangeably) to reach a large number of potential customers.

The importance of social influence in information spread processes was demonstrated in many studies (e.g., Asch (1951); Centola & Macy (2007)). One of the traditional models for describing diffusion of information in networks is the Linear Threshold model (Granovetter, 1978). In this model, a message spreads from one node to another if a fixed number of neighbors of the latter have already adopted it. The spreader of the message, who is interested to reach a large number of adopting nodes, has to wisely select a subset of network nodes, and actively infect (seed) them. Then, the assumption is that a passive viral process begins, in which nodes infect other nodes without any external intervention. Such a passive viral process can happen for example, if a Facebook user posts an exciting message or photo on his wall, which is then repeatedly shared by other Facebook users.

As shown by both analytical and simulative studies (Barthélemy et al., 2004; Khelil et al., 2002; Vespignani, 2012; Zhou et al., 2007), messages that propagate according to models similar to the Linear Threshold model, are expected to propagate into a substantial portion of the network. However, in recent years, several works (e.g., Leskovec et al. (2009, 2007b); Leskovec & Horvitz (2008); Goel et al. (2012)) have shown, based on real information cascades datasets, that the frequency of large information cascades in networks, are significantly lower than what was previously believed. In fact, it was shown that the vast majority of messages never spread beyond a few nodes.

ACCEPTED MANUSCRIPT

Since large information cascades are rare in reality, it is unlikely that a product or service will be spread only through a passive viral process. Indeed, in most real-world marketing scenarios, substantial additional efforts are invested in order to promote products. Companies cannot simply post Facebook messages on their products and expect them to spread passively, and therefore so many sales and marketing personnel are hired to actively promote these commercial products and services.

In this work, we propose a new information diffusion model, named Active Viral Marketing (AVM), which better reflects the need of commercial companies to invest continuous marketing efforts to promote their products. More specifically, nodes in our model cannot get infected by themselves through a passive viral process. Instead, they can get infected only through an active seeding attempt made by the spreader. The importance of social influence comes into play where the success of a seeding attempt depends both on the number of infected neighbors the node has and on the time frame in which they got infected.

As a motivating example, consider a tourism company that aims to promote a summer vacation through social networks advertising. The company can pop-up the advertisement to several social network's users that fit in age and social class and are considered to be influential (commonly estimated based on the network topology). These advertisements have a defined cost, which is paid to the social network company (and sometimes to the influencers as well). It is likely that a user that already clicked on the ad (and potentially booked a vacation), will discuss it with his social network's friends. However, in most cases it is unlikely that following such a discussion, these friends will initiate a call to the tourism company (as in the passive viral spread) in order to book a vacation. Rather it is more likely that the discussion with friends that already booked a vacation will have some positive impact on them. Now, consider a user that had several such discussions with his friends. If the company chooses to pop-up an ad to this user now, the likelihood of him clicking on the ad will be probably higher than in the case without previous discussions. Moreover, if the ad is presented to the user, long after the discussions with his friends, it is less likely that he will click on the ad, since the impact of these interactions weakens over time (for simplicity of exposition, we will term this phenomenon as "forgetting effect" or as "memory loss"). Therefore, in order for the tourism company to efficiently use its marketing budget (e.g., to maximize the click-through rate), it needs to schedule its advertisements in a way that balances the number of

ACCEPTED MANUSCRIPT

accumulated friends' clicks and memory loss (both grow over time and have an inverse effect).

Following the suggested AVM model and the observation above, this work develops a set of Scheduled Seeding Heuristics (SSH). The main idea behind SSH is to utilize the information on the dynamic states of nodes, in addition to the static network topology (that is commonly used by existing seeding heuristics), when choosing the next node to seed. This added information allows SSH to utilize better the social effect, by balancing between the number of infected neighbors of a node and its memory loss.

In order to evaluate the SSH heuristics, we conducted an extensive set of experiments, to compare them to other state-of-the-art seeding heuristics that rely on selecting central nodes (based on the network topology) prior to the seeding stage. The results of our experiments show that the SSH heuristics obtain an average adoption rate which is 30%-75% higher than the other benchmark heuristics, and that the superiority of SSH is consistent over a wide range of parameters' values selection.

The contribution of this work can be summarized along two axes:

- We propose a new diffusion model which, to our belief, better fits real-world scenarios of products adoption, where the spread of products relies on continuous active efforts of the sales or marketing departments.
- We demonstrate the importance and the high potential of a scheduled seeding heuristic, for the spread of trendy products, under a wide range of settings, and also point out the cases where such a heuristic is less effective.

The rest of this paper is structured as follows. Section 2 reviews the existing literature and provides the necessary background on information diffusion in networks. Sections 3 and 4 describe the proposed Active Viral Marketing model and Scheduled Seeding Heuristics, respectively. Section 5 details our evaluation methodology and the obtained results. Section 6 summarizes the paper, and presents directions for future research.

2. Background and Related Work

In this section, we provide the relevant background to the fields of contagion models and viral marketing. We start by presenting the basic theoretical

ACCEPTED MANUSCRIPT

models of viral diseases, followed by two well-known models, which capture the important aspects of viral marketing. These theoretical models are then inspected through the lens of real-world data evidences.

2.1. Contagion Models

Mathematical contagion models of diseases were historically developed by Epidemiology researchers as a tool to study the mechanisms by which diseases spread, to predict the future course of an outbreak and to evaluate strategies to control an epidemic (Anderson et al., 1992). Due to their success in the field of disease modeling, such models implied their wide usage in other fields as well, such as information diffusion and product adoption.

Existing contagion models can be broadly classified into two categories: (1) compartmental models and (2) individual-based models.

Compartmental models assume a fully interconnected population, in which the interactions and infections can occur between any pair of available individuals. This implies a homogeneous population in terms of their connectivity and chances of interaction. These models allow to observe different phenomena at the compartment level, such as the size of the compartment and the infection pace at different time periods of the contagion process. One of the most well-studied compartmental contagion models is the SIR model (Anderson et al., 1992). This model splits the population individuals into three compartments: S - susceptible, I - Infected and R - Recovered. The transitions between the states in this model are trivial - susceptible individuals have a probability β to become infected as a result of an interaction with infected individuals. Similarly, infected individuals recover (and therefore reassigned into the recovered compartment) with a constant pace γ.

Individual-based models assume the existence of a network structure that describe the potential interactions (network edges) between individuals (network nodes). In contrast to compartmental models, individuals cannot become infected from any member of the infected compartment, but only from their network neighbors.

One of the fundamental individual-based models, commonly used to describe information diffusion in social networks is the Linear Threshold model (Granovetter, 1978; Kempe et al., 2003). The model assumes that the behavior of individuals greatly depends on the number of their network neighbors that are already engaged in that behavior. More formally, we denote the binary state of a node v (1 if active and 0 otherwise) at time t as $X_v(t)$ and

ACCEPTED MANUSCRIPT

the set of neighbors of node v as $N(v)$. A node v is influenced by each neighbor $w \in N(v)$ according to their edge weights $b_{v,w}$ which are set such that $\sum_{w \in N(v)} b_{v,w} = 1$. Each node v is assigned a threshold $\theta_v \in [0, 1]$, representing the fraction of v's neighbors that are required to be active in order for v to become active in the next time step. If the accumulated effect (sum of weights of active neighbors) on time step t on v is at least θ_v, v will become active at the next time step $t + 1$ and therefore will also begin to influence its own neighbors.

Another well-studied individual-based information diffusion model is the Independent Cascade model (Goldenberg et al., 2001a,b). In this model, a node v that was activated at time step t has a single chance to activate each of its currently inactive neighbors $w \in N(v)$. At the next time step, $t + 1$, v will not have any further influence on its neighbors. Similarly, if w becomes activated at time step $t + 1$, it will have one single chance to activate its inactive neighbors in time step $t + 2$.

A particularly interesting individual-based model, Bass-SIR, was recently suggested by Fibich (2016). This model proposes a new contagion process which combines properties of SIR and Bass (Mahajan et al., 1991) models, and applies them at the micro-level by utilizing a network structure. More specifically, as in the basic Bass model, if a node v did not adopt the product by time step t, it has a positive probability to adopt the product in the nearest future $(t, t + \Delta t)$:

$$P(v \text{ adopts in } (t, t + \Delta t)) = (p + q\frac{I_v(t)}{k_v})\Delta t + o(\Delta t)$$

Where p and q are bass coefficients of innovation and imitation accordingly, $I_v(t)$ is the number of infective neighbors of v at time step t, k_v is a normalization factor (usually $k_v = |N(v)|$) and $\Delta t \to 0$. Unlike the basic Bass model, Bass-SIR does not assume that an infected node will stay infective forever, and therefore the probability of an infective node to become recovered is:

$$P(v \text{ recovers in } (t, t + \Delta t)) = r\Delta t + o(\Delta t)$$

Where $\Delta t \to 0$, and r is the recovery pace.

The Linear Threshold and Independent Cascade models served as a basic setup to a wide range of works, and over the years many extensions were suggested to fit these models to special cases. In their seminal work, Kempe et al. (2003) proposed two models which aimed at generalizing many of the

ACCEPTED MANUSCRIPT

extensions into a unified framework. The introduction of these two general models served several goals. First, they present a unified framework for any arbitrary activation function that is consistent with the monotonicity condition. Second, they prove that these two models are equivalent, and provide a method to covert between them. Third, when limiting the discussion to sub-modular activation functions, Kempe et al. provide an approximation to the Influence Maximization problem, covered later in section 2.2.

2.2. Influence Maximization

An important field in the study of information diffusion through social networks is the identification of influential nodes with the goal of maximizing the adoption of products or ideas in the network. More formally, given a model of information diffusion (e.g., Linear Threshold, Independent Cascade, etc.) over a network G, the influence maximization problem deals with selecting a subset of the network nodes, whose intentional activation (often referred to as seeding) will ignite a viral contagion process that will impact a significantly large set of nodes. Often these models aim at optimizing a given target function related to the network adoption. The target function can have several forms, such as maximizing the number of adopters in a certain time period or budget (number of seeding actions), or minimizing the number of seeding actions required to reach a certain number of adopters.

For example, modern marketing efforts use social networks for market analysis and for defining promotion strategies. Unlike classical mass-marketing methods that address a wide market segment, social networks' promotion is often characterized by micro-segmentation, attempting to utilize detailed information about each of the involved individuals (Goldfarb & Tucker, 2011). The main motivation behind such an approach, is that influencing the opinion of only a few individuals may shape the opinion of the majority, by following a viral contagion process (Katz & Lazarsfeld, 1955).

The task of identifying influential nodes is still widely investigated, but the identification of influential nodes is not always easy. In many cases, nodes are referred to as "influential" when past evidence show that their involvement in the contagion process contributes significantly to the spread. Nonetheless, such detailed information is often absent, and most of the data available to the marketers is the topological structure of the social network and past adoption history.

ACCEPTED MANUSCRIPT

2.2.1. Initial Seeding Strategies for Influence Maximization

Identifying influential nodes, given only the network structure, can be addressed via graph-based metrics, such as the centrality measures (Borgatti, 2005).

One way to measure a node's centrality is by counting the number of its connections (known as the node degree). While calculating the degree of a node is a relatively trivial task, such an approach is limited since it takes into account only the first-order effect, without considering higher-order effects. Other frequently used centrality measures that take into account high-order effects include the PageRank (Page et al., 1999), the Betweenness centrality (Brandes, 2001) and the Eigenvector centrality (Bonacich, 2007). Each of these measures has its own attributes and represents a different type of importance that characterizes a node. For a good source on centrality measures, the reader is referred to (Borgatti, 2005) and (Newman).

With respect to influence maximization, several works investigated the efficiency of seeding central nodes. The work by Hinz et al. (2011), for example, investigated four seeding strategies: Hubs (Degree/EigenVector Centrality), Bridges (Betweenness Centrality), Fringes (Edge Nodes) and Random. The authors conducted three experimental studies of adoption using a small controlled network; a real social network of selected students; and a large-scale cellular network. The study found that targeting Hubs is the most effective strategy in terms of influence maximization, with the Bridges strategy right afterwards, both with a big gap above the Random strategy (150-200%) and a huge gap above the Fringes strategy. Similar results were obtained by Banerjee et al. (2013), where the authors investigated empirically the spread of financial loan systems within a social network of Indian villagers. The authors found that villagers with high Eigenvector centrality scores are more likely to influence others in their surroundings, in comparison to the other measures of centrality.

The performance of seeding strategies depends not only on the properties of the network topology and its nodes, but also on the information diffusion dynamics themselves. For example, Kempe et al. (2003) study the influence maximization problem under the linear threshold and independent cascade settings and their generalizations. The authors prove that finding the optimal solution to the problem is NP-hard in both settings and present a greedy algorithm which obtains a $(1 - 1/e)$ approximation of the optimal solution. While the greedy algorithm ensures a reasonably good result in terms of

ACCEPTED MANUSCRIPT

coverage, it is still very expensive in terms of runtime when executed on large-scale datasets.

The complexity of the problem and the non-scalability of the greedy approximation algorithm opened the chase after high performing and scalable seed selection heuristics. While many such heuristics were suggested in the literature, we focus on two well-studied groups of such heuristics.

One notable group of such heuristics are the *CELF* (Leskovec et al., 2007a) and *CELF++* (Goyal et al., 2011) algorithms, which are based on a "lazy-forward" optimization scheme for selecting the seeds. Their underlying idea is based on bounding the marginal contribution of a node in a future iteration, with its marginal contribution in a previous iteration due to monotonicity and sub-modularity properties of the influence maximization problem. These heuristics provide an efficient variation of the greedy approximation algorithm by improving the order of evaluating nodes to be added to the "seed set". Empirical evaluation showed that the proposed heuristics outperform (in terms of influence maximization) and run faster than the greedy algorithm, while still guaranteeing a constant factor approximation of the optimal solution.

Another notable group of heuristics was suggested by Chen at al. (Chen et al., 2009, 2010a; Jung et al., 2012; Chen et al., 2010b). Chen et al. (2009) presented an improved greedy algorithm for seeding outcome evaluation by reducing the search space per each evaluation, and showed a 700-times faster performance on the independent cascade model. Chen et al. (2010a) suggested the Maximum Influence Path (PMIA) algorithm. Using this method under the independent cascade model, the authors suggested to locate the nodes whose seeding will result in a long chain of cascades with the highest probability. Jung et al. (2012) proposed the Influence Rank Influence Estimation (IRIE) algorithm, which performs an estimation of the influence function for any given seed set, using precomputed influence estimated values for iterative seed set ranking. Empirical simulations have shown that the IRIE heuristic performance is similar to that of the Greedy, PMIA and Pagerank influence heuristics, while its memory consumption provides a significant improvement over that of the other heuristics.

While a large number of works in this field focused on the problem of maximizing influence with a given seeding budget, Long & Wong (2011) investigated the problem of minimizing the number of seeding actions to obtain a certain number of influenced nodes. The authors proved that the problem is NP-hard, and developed a greedy heuristic that provides error

ACCEPTED MANUSCRIPT

guarantees. They also studied the "Full-Coverage" setting, where the goal is to infect the entire network, and designed efficient algorithms for this purpose.

With the same spirit, Goyal et al. (2013) identified three orthogonal dimensions in the influence maximization problem: (1) the number of seed nodes activated at the beginning, (2) the expected number of activated nodes at the end of the propagation, and (3) the time taken for the propagation, claiming that it is possible to constrain either one or two of these dimensions and try to optimize the third. The authors then studied two of these variations and suggested approximated algorithms to solve them efficiently.

2.2.2. Adaptive Seeding Strategies for Influence Maximization

The majority of existing works that dealt with the influence maximization problem, focused on selecting a subset of network nodes, that if seeded simultaneously at the beginning of the process, would maximize the adoption rate at the end of the process. Recently, numerous works presented a new adaptive approach, which spreads the seeding actions over time, and therefore allows to reassess the contribution of the seeds' selection in each time step, in order to improve the overall adoption rate.

For example, Seeman & Singer (2013) present a two-stage framework for influence maximization. The underlying assumption of this model is that besides of the "non-active" (susceptible) and "active" (infective) states there is an intermediate state referred to as "available": a node v is considered available for seeding only if one of its neighbors $w \in N(v)$ is active. Given an initial set of available nodes $X \subseteq V$, the goal of the first stage is to select a seeding set $S \subseteq X$ in order to extend the set of available nodes, so that the seeding actions in the second stage will maximize the expected influence. The idea behind it relies on the known fact that selecting a neighbor of a random node v is likely to have a higher degree than v itself and thus one would like to include those higher-degree nodes in the set of available nodes for seeding.

In another study, Tong et al. (2017) suggest an adaptive seeding strategy for a variant of the Independent Cascade model. In this variant, referred to as "Dynamic Independent Cascade" model, the authors assume that the activation of a node v by seeding occurs with a probability p_v. Therefore, in contrast to the models surveyed above, a seeding action may fail, keeping the node in a non-active state. Under this setting, the authors suggest an adaptive seeding approach, in which the selection of nodes to be seeded at

ACCEPTED MANUSCRIPT

each time step, is performed while taking into account the realization of the previous seeding attempts.

Jankowski et al. (2017a,b) suggest an adaptive seeding approach to the influence maximization problem under the Independent Cascade model. The authors show that, regardless of the chosen strategy for selecting influential nodes, spreading the seeding actions along different time-steps of the diffusion process can improve the overall adoption rate. Moreover, they present an inherent trade-off between the obtained adoption rate and the duration of the diffusion process.

Chierichetti et al. (2014) introduce a different diffusion model in which there are two competing ideas, each aiming at maximizing its spread over a social network. More specifically, consider a marketer which addresses each one of the individuals in the network sequentially (the marketer has the ability to determine this sequence) and offers them a cause. The cause can either be accepted (Y) or denied (N) by each of the individuals, according to the following rule: the individual v accepts the offer if $|m_Y| - |m_N| \geq c$, deny it if $|m_N| - |m_Y| \geq c$ and chooses randomly between Y and N otherwise. m_Y and m_N represent the size of the group of v's neighbors who already decided to accept or deny the cause (Y or N), and c is a positive integer that serves as a decision threshold. The goal of the marketer in this setting is to determine the best order to address the individuals in order to maximize the amount of Y decisions. The authors also provide an efficient greedy algorithm that ensures the best achievable solution to the problem.

Lin et al. (2014) suggest the "Push-Driven Cascade" model in which the probability that a node will become active after a seeding action is determined by the activation state of its neighbors. More specifically, the probability of an individual v to become activated is:

$$p_v(t) = d_v + \sum_{w \in N(v)} b_{v,w} * X_w(t-1)$$

Where $X_w(t-1)$ is the binary state of node w (1 if active and 0 otherwise) at time $t-1$, the node v is influenced by each active neighbor $w \in N(v)$ according to their edge weights $b_{v,w}$ and d_v is v's own bias towards adoption. The role of the marketer in this setting is to choose a single node to seed at each time step in order to maximize the overall adoption in the network.

It is important to emphasize that in the two latter models, each node has an accumulated influence in favor of the product, but only the seeding act

ACCEPTED MANUSCRIPT

itself is considered to be the trigger for activation, where the viral spread serves only as a positive effect on the activation probability. This is in contradiction to classical diffusion models where nodes could become active as a result of a viral infection without any external intervening operation.

2.3. Information Diffusion in Real World Settings

As seen in the previous section, the dynamics of information diffusion in Social Networks were widely studied and many mathematical models which aim at describing these dynamics were suggested. In recent years, due to the increased availability of data, and the emergence of tools to store and process data at large-scale, a growing body of works have started to analyze the dynamics of information diffusion in real-world scenarios, and obtain better understanding of where existing models succeed and fail in describing these dynamics.

One of the principles behind many of these models is that of accumulated social effect. Already in 1951, the social psychologist Asch presented an experiment, in which he showed that the probability of a subject to change his opinion is proportional to the number of peers who are convincing him to do so (Asch, 1951). Granovetter (1978) in turn, presented a threshold behavior, in which an accumulated social effect is turned into an activation by reaching a personal threshold of the individual. Hence, since the threshold values are distributed randomly, the probability of an activation is proportional to the number of social influencers, similarly to Asch's findings. Later on, Centola & Macy (2007) had performed a large-scale empirical study of online social networks. He found that in contradiction to "Simple Contagion" in which a single interaction with an infected individual may lead to activation (e.g., like in the spread of infectious diseases), the activation of an individual often requires reinforcement from multiple infected sources, a phenomenon named by the author as "Complex Contagion".

A recent work by Goyal et al. (2010) studied the time effect of propagation of social influence in networks. Consequently, the authors suggested an extension to the General Threshold model by adding a diminishing time-dependency factor. More specifically, they considered three types of time-dependent models which reflect a lower ability of a node to spread the adopted idea as time passes: (1) A Static Model the influence of an infective node does not diminish over time; (2) A Discrete Model each activated node has a period of time in which it is infective. After that period, the node stops from being infective; and (3) Continuous Model the influence of an infective node

ACCEPTED MANUSCRIPT

v on a neighbor node w diminishes over time with an exponential rate. The authors found that the best fit to the data was obtained by the continuous (exponential decay) model. One explanation that was given to this diminishing influence effect in the scientific literature is the limited attention effect. According to this effect, a person which is exposed to multiple ideas during a single time period, is able to concentrate only on a few of them resulting in a forgetting effect (Weng et al., 2012). These findings, strengthen the usage of the recovery effect in several of the models mentioned above, such as SIR and Independent Cascade.

In another paper by Leskovec et al. (2007b), the authors investigate the cascading behavior of online information diffusion, by analyzing 45,000 blogs and about 2.2 million blog posts. The authors identified several cascade shapes that rule the majority of cascades, pointing out two specific shapes: star-shaped, reflecting the spread of information in different directions, and chain-shaped, presenting a chained sequence of information flow. Further investigating the degree-distribution of the cascades, they found that in-degree and out-degree distribution of bag-of-cascades follow power-law exponents of -2.2 and -1.92 respectively. Finally, by examining the distribution of cascade sizes for each shape of cascade, they found that all cascades follow a heavy-tailed distribution, and the probability of observing a cascade of n nodes follows a Zipf distribution. These findings emphasize that in real-world scenarios, highly viral information cascades rarely exist.

Another support for the above findings can be found in (Goel et al., 2012), where the authors analyze information cascades in seven different online domains. The authors observed that the vast majority of cascades are small, and that they usually terminate within one circle of neighbors of the initial adopting node.

3. The Proposed Active Viral Marketing Model

In this section, we propose a novel information diffusion model, named the Active Viral Marketing model, which better reflects the need of commercial companies to invest continuous marketing efforts to promote their products or services. According to the proposed model, at any given time-step t, a node v can only be at one of the following $X_v(t)$ states:

- $X_v(t) = 0$: Non-Infected
- $X_v(t) = 1$: Infected and Infectious

- $X_v(t) = 2$: Infected but not Infectious
- $X_v(t) = 3$: Seeding Failed

The possible transitions of a node v between these states are described in Figure 1:

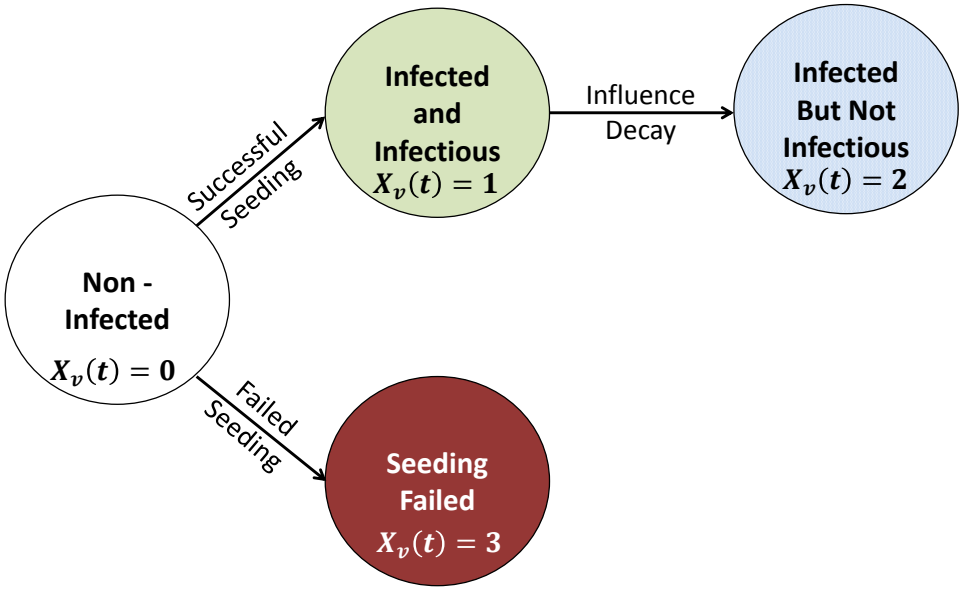


Figure 1: Infection states of nodes in the AVM model.

More specifically, if the spreader attempts to seed a non-infected node v at time-step t, the attempt may succeed with a probability $P_v(t)$. If the seeding attempt succeeds, then the node's state changes from $X_v(t) = 0$ to $X_v(t) = 1$. The probability of a successful seeding attempt is affected by v's individual preferences and the activation rate of v's neighbors (described in more details below).

If the seeding attempt fails, subsequent attempts to seed v are not allowed (since in a typical marketing scenario, subsequent seeding attempts may only annoy the potential customer and may lead to a negative attitude towards the spreader), and v is transitioned into a "Seeding Failed state ($X_v(t) = 3$). On the other hand, if the seeding attempt succeeds, v is transitioned into

ACCEPTED MANUSCRIPT

a "Infected and Infectious" state, and will influence its neighbors only for the next t_{inf} periods. After t_{inf} periods have ended, v's state changes to "Infected but not Infectious" ($X_v(t) = 2$).

The probability that an attempt to seed a node v at time-step t will succeed is given in Eq. 1:

$$P_v(t) = P_v^{ind} + P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v}) \quad (1)$$

This probability is composed of two factors: (1) the individual preferences of v, denoted by P_v^{ind} and (2) the social influence exerted on v by its infectious neighbors at time-step t, denoted by $P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v})$.

The social factor is calculated as the product of P_v^{soc} and $min(1, \frac{|N_v^1(t)|}{\theta_v})$. The maximal social effect that can be achieved is represented by P_v^{soc}, (note that $P_v^{ind} + P_v^{soc} \leq 1$). $min(1, \frac{|N_v^1(t)|}{\theta_v})$ represents the relative social effect, which increases proportionally with $|N_v^1(t)|$, denoting the number of infectious (state 1) neighbors of v, up to a certain level determined by the threshold θ_v. The min function assures that even if the number of active neighbors exceeds the threshold θ_v, the probability function would not exceed the value of 1, and therefore, the total social effect would not exceed P_v^{soc}.

The formulation of the social factor described above was inspired by the empirical results of Asch's conformity experiments (Asch, 1951). In his experiments, Asch inspected how the size of a group influences the probability of conforming to the opinion of the majority. He observed that as the size of the group grows, the conforming probability grows almost linearly until reaching a certain size, and after reaching that size, the probability doesn't grow further. We model these two properties by using the threshold θ_v and the maximum probability P_v^{soc}. A comparison of Asch's original findings and our simplified model (for the case of $p_v^{soc} = 0.6$, $p_v^{ind} = 0.1$ and $\theta_v = 4$) are depicted in Figure 2.

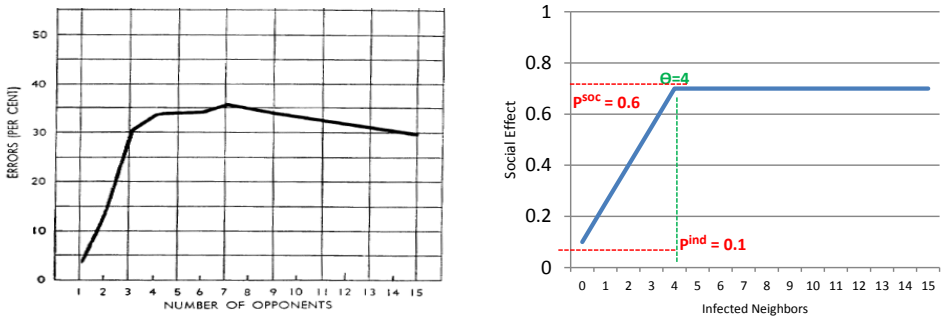


Figure 2: Social effect in Asch's conformity experiment (left) and its representation in the AVM model (right).

Given the Active Viral Marketing diffusion model and a seeding budget of size B, the goal is to find an ordered set of B nodes, denoted by $S = (v_1, v_2, .., v_B)$, such that seeding the node v_1 at time-step $t = 1$, the node v_2 at time-step $t = 2$, ..., the node v_B at time-step $t = B$, would maximize the total number of successful seeding attempts.

4. The Scheduled Seeding Heuristics

The influence maximization problem that was defined above for the Active Viral Marketing diffusion model is NP-hard and is not sub-modular (see Appendix A). Accordingly, in this section, we propose a set of seeding heuristics, named Scheduled Seeding Heuristics (SSH), that recommend which node to seed at each time-step. Similar to existing seeding heuristics, our heuristics utilize the static network topology when choosing the nodes to be seeded. However, in contrast to existing heuristics, our heuristics also take into account the information on the dynamic states of nodes at each time-step.

More specifically, at each time-step, our heuristics assign a utility score for each one of the non-infected (state 0) network nodes, with the idea that seeding a node with a higher utility score is worthier. The utility score is based on the expected value for each potentially seeded node, and is calculated as the probability of a successful seeding of the node itself, multiplied by the value of such an event.

ACCEPTED MANUSCRIPT

Given the vector of states of all network nodes at time-step t, denoted by $\vec{X}(t)$, the probability of a successful seeding of v at time-step t is denoted as:

$$P(\vec{X^v}(t+1))$$

Where $\vec{X^v}(t+1)$ is identical to $\vec{X}(t)$ with the additional assumption that node v changed its state to $X_v(t) = 1$ at time-step $t+1$.

The value of a successful seeding event of node v can be seen as the influence of v on future seeding attempts of its non-infected neighbors, formulated as:

$$\sum_{w \in N_v^0(t+1)} U(w, t+1, \vec{X^v}(t+1))$$

Where $w \in N_v^0(t+1)$ is a non-infected neighbor of node v at time-step $t+1$, and $U(w, t+1, \vec{X^v}(t))$ is the utility score of seeding w at time-step $t+1$, given that v was already seeded successfully at time-step t.

Finally, the utility score of a node v is calculated as the probability of a successful seeding of v, multiplied by the value of such an event:

$$U(v, t, \vec{X}(t)) = P(\vec{X^v}(t+1)) \cdot [1 + \sum_{w \in N_v^0(t+1)} U(w, t+1, \vec{X^v}(t+1))]$$

Note that the formulation of $U(v, t, \vec{X}(t))$ is recursive, and may involve successive iterations to evaluate the value of future seeding events beyond t+1. For practicality reasons, we limit the recursion to a depth of $k \in \{0, 1, 2\}$ iterations, as we found empirically that increasing the complexity of the algorithm by using higher k values has a diminishing return effect. The recursive computation of the score, for a depth of k iterations (k is provided as an input parameter), is shown in detailed in Algorithm 1.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Algorithm 1 The SSH Scoring Algorithm

Input:

t - time-step
$\vec{X}(t)$ - states of nodes in time-step t
v - node
k - recursion depth

Output:

Score of v

$P_v(t) \leftarrow P_v^{ind} + P_v^{soc} \cdot min(1, \frac{|N_v^1(t)|}{\theta_v})$
if $k = 0$ **then**
  $Score \leftarrow P_v(t)$
else
  $Score \leftarrow 1$
  for u in $N_v^0(t)$ **do**
    $Score \leftarrow Score +$ **SSH**$(t+1, \vec{X}^v(t+1), u, k-1)$
  end for
  $Score \leftarrow P_v(t) \cdot Score$
end if
return $Score$

To illustrate how Algorithm 1 works, consider the five-nodes network depicted in Figure 3.

ACCEPTED MANUSCRIPT

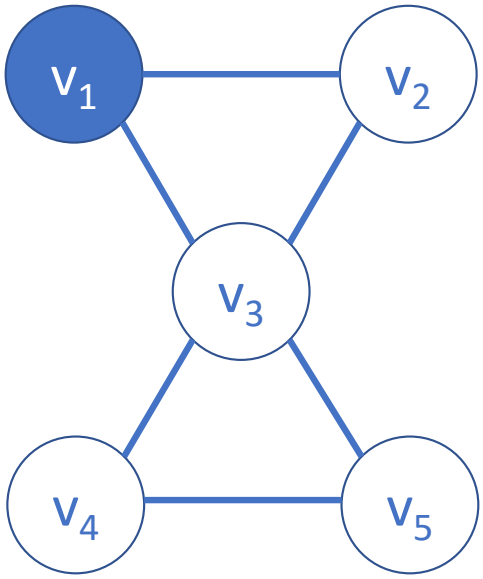


Figure 3: An illustration of a network with five nodes.

Assume that node v_1 was activated at the previous time-step $(t-1)$ and the following parameters: $P_v^{ind} = 0$, $P_v^{soc} = 1$ and $\theta_v = 0$, for all network nodes.

For a recursion depth of $k = 0$, we get the following utility scores for nodes v_2, v_3, v_4 and v_5 respectively:

$$SSH(t, \vec{X}(t), v_2, 0) = P_{v_2}(t) = 0 + 1 \cdot \frac{1}{2} = 0.5$$

$$SSH(t, \vec{X}(t), v_3, 0) = P_{v_3}(t) = 0 + 1 \cdot \frac{1}{2} = 0.5$$

$$SSH(t, \vec{X}(t), v_4, 0) = P_{v_4}(t) = 0 + 1 \cdot \frac{0}{2} = 0$$

$$SSH(t, \vec{X}(t), v_5, 0) = P_{v_5}(t) = 0 + 1 \cdot \frac{0}{2} = 0$$

Since both nodes v_2 and v_3 obtained the highest utility score, we will choose to seed either one of them at time-step t.

Alternatively, for a recursion depth of $k = 1$ we get the following utility scores for nodes v_2, v_3, v_4 and v_5 respectively:

$$SSH(t, \vec{X}(t), v_2, 1) = P_{v_2}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0)) = 0.5 \cdot (1+1) = 1$$

ACCEPTED MANUSCRIPT

$$SSH(t, \vec{X}(t), v_3, 1) = P_{v_3}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_2, 0) + SSH(t+1, \vec{X}(t), v_4, 0) +$$
$$+ SSH(t+1, \vec{X}(t), v_5, 0)) = 0.5 \cdot (1 + 1 + 0.5 + 0.5) = 1.5$$
$$SSH(t, \vec{X}(t), v_4, 1) = P_{v_4}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0) + SSH(t+1, \vec{X}(t), v_5, 0)) = 0$$
$$SSH(t, \vec{X}(t), v_5, 1) = P_{v_4}(t) \cdot (1 + SSH(t+1, \vec{X}(t), v_3, 0) + SSH(t+1, \vec{X}(t), v_4, 0)) = 0$$

Since node v_3 obtained the highest utility score, we will choose to seed it at time-step t.

Runtime Complexity Analysis: The higher time-consuming operations of Algorithm 1 are performed in steps 1 and 6-8. In step 1, the algorithm determines the number of infected neighbors of node v, and in steps 6-8, the algorithm determines the utility score of each one of the non-infected neighbors of node v, given that v was already seeded successfully. Line 7, in particular, includes a recursive call which reduces the recursion depth (k) by 1. Therefore, if we denote the maximum degree of a node by d, the runtime complexity of Algorithm 1 in the worst case is $O(d^{k+1})$. It is important to note that Algorithm 1 is executed for each one of the non-infected nodes in the network, every time a seeding decision has to be made. Therefore, denoting the number of nodes in the network as $|V|$ and the seeding budget as B, the total time spent on Algorithm 1 is $O(|V| \cdot B \cdot d^{k+1})$.

5. Evaluation

In this section, we present an extensive set of empirical experiments that compare the performance of the proposed SSH approach (that is state-based) with that of existing seeding heuristics that rely on the network topology without taking into consideration the states of the nodes.

5.1. Experimental Setting

All the experiments were implemented in Python 2.7 and executed on a Linux machine running Centos 7.1, with 128 GB of RAM and a single Intel 2.7 GHz CPU.

Each of the simulations was preceded with selecting a random set of nodes, served as an initially infected population of size F. The infection time-steps of the nodes in this initial population were drawn uniformly from the interval $[-t_{inf}, -1]$. Then, at each time-step of the simulation, a single node was seeded, where the selection of the seeded node was based on different

ACCEPTED MANUSCRIPT

heuristics (the set of examined seeding heuristics is described below). Each seeding attempt either succeeded or failed in accordance with Eq. 1. The transitions in states of nodes were re-calculated at each discrete time-step.

The simulation ended when the entire budget of seeding attempts, B, was used. At this point, the final seeding success rate was calculated for each of the heuristics.

5.1.1. Parameters' Space

In the experiments, we examined a variety of values for the different parameters. In each set of simulations that are reported below, all parameters except one were set to their default value (fixed in most cases to the median of their examined range of values), while a single remaining parameter was examined over a varying range of values. The parameters' space used in our experiments is detailed in Table 1. Each combination of parameters values was examined by executing 400 simulation runs, for each one of the compared heuristics.

Table 1: Simulation Parameter Space

Parameter	**Values**
Network Topology (see Table 2)	**Sampled Citation network**, Slashdot network, Sampled EuEmail network, WikiVote network, Epinions network, Enron network
Network size (# of sampled nodes)	5000, 10000, 50000, **100000**, 500000, 1000000
Initially infected population size (F)	50, 100, **200**, 500, 1000
Budget (B)	50, 100, **200**, 500, 1000
Threshold (θ_v)	3, 4, **5**, 6, 7
Maximal Social Effect (P_v^{soc})	0.1, 0.3, **0.5**, 0.7, 0.9
Infection Time (t_{inf})	10, 20, **50**, 100, 200
Individual Effect (P_v^{ind})	**0**, 0.1, 0.2, 0.3, 0.4, 0.5

* The default value of each parameter is marked in **bold**.

ACCEPTED MANUSCRIPT

In most of our experiments, we assumed that the values of the parameters θ_v and P_v^{soc} are known. In another dedicated experiment, we assumed that the distributions of these parameters' values are normal, and we only know their mean and standard deviation. These means are denoted by μ_θ and $\mu_{P^{Soc}}$, while the standard deviations are denoted by σ_θ and $\sigma_{P_v^{Soc}}$, respectively. The actual values of these parameters for each node, were randomly generated prior to each simulation run, and were not used in any way by the SSH heuristics.

5.1.2. Network Topologies

The simulations were executed on different network topologies, as detailed in Table 2. These topologies represent snapshots of real-world social networks, with some adaptations to our experimental framework, such as converting the networks to undirected, or sampling a subset of nodes. The original social network datasets are publicly available at (Leskovec & Krevl, 2014)

Table 2: Networks Used in Simulation

Network	**Number of Nodes**	**Average Degree**	**Average Clustering**	**Sampled?**
Citations	1000000	2.83	0.04	Yes
Citations	500000	4.06	0.06	Yes
Citations	100000	7.60	0.14	Yes
Citations	50000	8.20	0.16	Yes
Citations	10000	6.81	0.20	Yes
Enron	36692	10.02	0.50	No
WikiVote	7115	28.32	0.14	No
Slashdot	82168	14.18	0.06	No
EuEmail	100000	1.57	0.03	Yes
Epinions	75879	10.70	0.14	No

5.1.3. Seeding Heuristics

We compared three variations of the proposed SSH approach (SSH-0, SSH-1 and SSH-2, where the levels of recursion were $k = 0$, $k = 1$ and $k = 2$ respectively) with four benchmark approaches as we proceed to describe. These benchmark approaches included both a state-of-the-art network-centrality-based approach (GEC), and a simple random selection of nodes (Random).

ACCEPTED MANUSCRIPT

Furthermore, for each of these two benchmark approaches we added a variation which considered as optional seeding candidates, only nodes that have a non-zero probability to become infected (i.e., nodes that have at least one infected neighbor). These additional variations were named Picky-GEC and Picky-Random.

The seven heuristics mentioned above are described in further details below:

Random Randomly seeds one uninfected node at each time-step.

GEC Chooses the uninfected node with the highest Eigenvector Centrality measure at each time-step.

Picky-Random Randomly chooses an uninfected node from the nodes that have a non-zero probability to become infected.

Picky-GEC Chooses the uninfected node with the highest Eigenvector Centrality from the nodes that have a non-zero probability to become infected.

SSH-0 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 0).

SSH-1 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 1).

SSH-2 - Chooses the uninfected node with the highest value of $P_v(t)$ at each time-step (i.e., Algorithm 1 with k = 2).

5.2. Results

5.2.1. Overall Comparison of SSH with the other Benchmark Methods

Figure 4 presents an overall comparison of the SSH approach to the other benchmark methods. Figure 4 (top) presents this comparison for different network topologies whereas Figure 4 (bottom) focuses on different sample sizes of the Citation network topology. In these experiment, all other parameters that are mentioned in Table 1 except for the network topology and size were set to their default values.

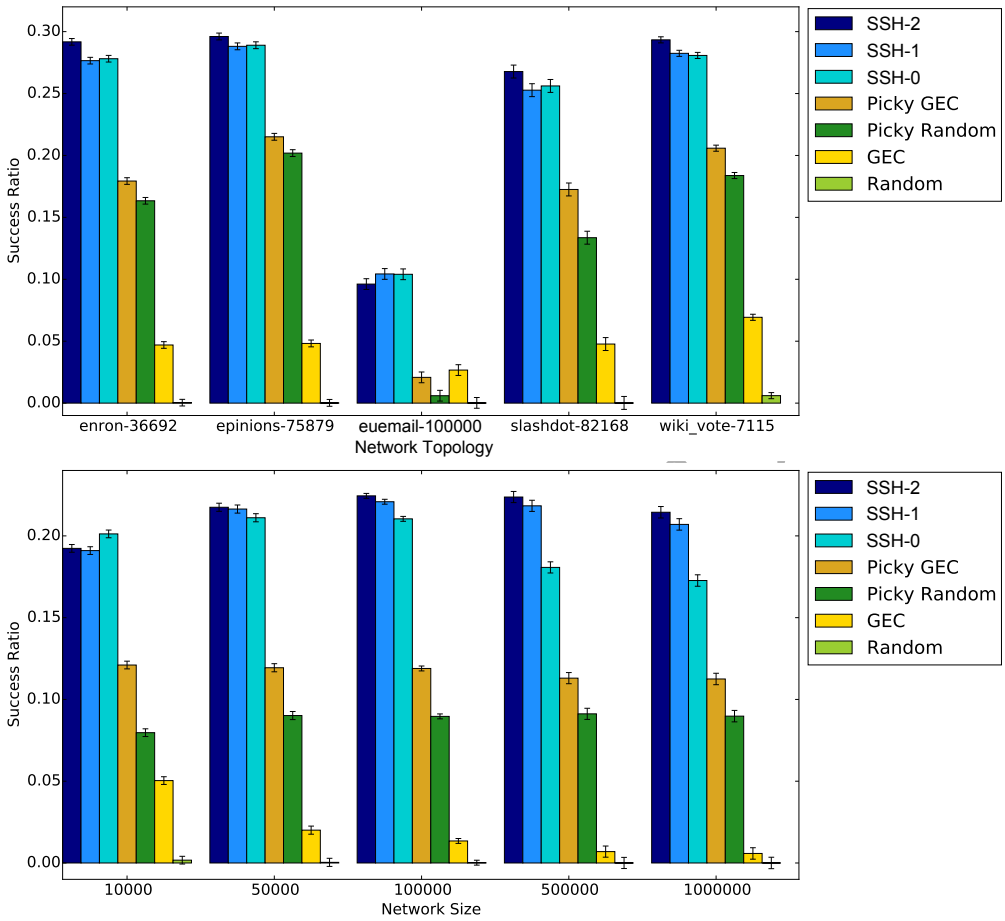


Figure 4: An overall comparison of SSH (the three blue bars) with the benchmark methods, across different network topologies (top) and network sizes (bottom).

As can be seen in the figure, the three SSH heuristics (blue bars) significantly outperform the other benchmark methods. More specifically, comparing SSH-0 (the worst out of the three SSH heuristics) to Picky-GEC (the best out of the other benchmark methods), the improvement ranges from 30% to 75%.

With regard to the different SSH heuristics, it seems that in most cases SSH-2 achieves the best performance, followed by SSH-1 and then SSH-0. This is in accordance with the amount of information that each of those heuristics uses to evaluate the scores of potential nodes to seed. However, it

ACCEPTED MANUSCRIPT

is worth mentioning that the differences in performance between these three heuristics are relatively low in comparison to the other benchmark methods.

As expected, the worst performing heuristic (by far) is the Random heuristic, which does not utilize any information about the network topology nor the states of the nodes. The GEC heuristic, performs slightly better than Random heuristic, since it utilizes information about the network topology.

Two interesting heuristics are Picky-Random and Picky-GEC that utilize partial information about the states of the nodes (i.e., which nodes have non-zero probability to be seeded successfully). As can be seen in the figure, these two heuristics perform better than the basic Random and GEC heuristics but worse than the SSH heuristics. We can also see that Picky-GEC performs slightly better than Picky-Random since it also utilizes information on the network topology.

5.2.2. Centrality of Seeded Nodes

In the previous experiment, we saw that the SSH heuristics perform significantly better than the GEC heuristic. In order to understand better why this is the case, we compared the centrality of nodes that were chosen by each of the two approaches. We were mainly interested to know if the SSH heuristics select to seed central nodes, or if it chooses to seed less central nodes. Note that in real-world marketing scenarios that involve seeding, not all seeding actions have the same cost. In fact, highly central nodes in social networks often represent celebrities, and the cost of seeding such celebrities is likely to be higher than that of less known individuals. Figure 5 presents the Eigenvector centrality of nodes that were chosen for seeding by the SSH-1 and GEC heuristics, along time.

ACCEPTED MANUSCRIPT

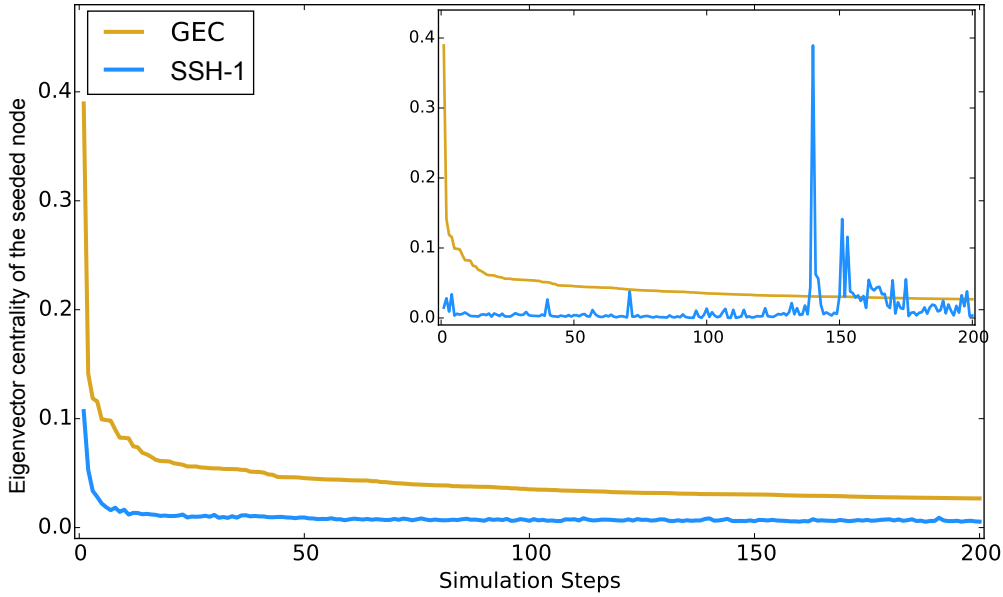


Figure 5: Eigenvector centrality of the nodes chosen for seeding along time.

The exterior figure presents the average Eigenvector centrality of nodes over 400 executions, where all parameters from Table 1 were set to their default values. As can be seen from the figure, both heuristics tend to start with nodes that have a higher Eigenvector centrality score and continue with nodes with lower and lower Eigenvector centrality scores. While this observation is expected for the GEC heuristic, it is less expected for the SSH-1 heuristic, since it does not make an explicit use of the network topology. It can also be seen that the average centrality score of the nodes selected by the SSH-1 heuristic is substantially lower than that of the GEC heuristic.

The interior figure presents a single execution, out of these 400 executions, for each of the two heuristics. As expected, the GEC heuristic performs the same in the single execution case and in the average case. However, with regard to the SSH-1 heuristic, we notice that central nodes are chosen somewhere at the middle of the contagion process and not necessarily at the initial stages. In other words, at any given time, the SSH-1 heuristic might prefer to choose a non-central node over a central node as long as its expected utility (its likelihood to be seeded successfully and its impact on its neighbors) is considered higher. This observation, together with the superiority of the SSH approach (as demonstrated in the previous experiment), emphasize

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

the importance of utilizing the states of the nodes and not only the network topology when assessing their ability to spread information. This is especially interesting since, centrality measures of a node, such as Eigenvector centrality, which take into account the network topology only, are often considered in the literature as a good proxy for the node's ability to spread information.

5.2.3. Sensitivity Analysis of the Model's Parameters

Figure 6 shows the total number of successful seeding attempts as a function of the seeding budget B. As expected, the number of successful seeding attempts grows with the budget size for all heuristics, but this growth presents a "diminishing return" effect. The figure also demonstrates the superiority of the SSH approach (blue plots), where its gap from the other heuristics increases with the budget size.

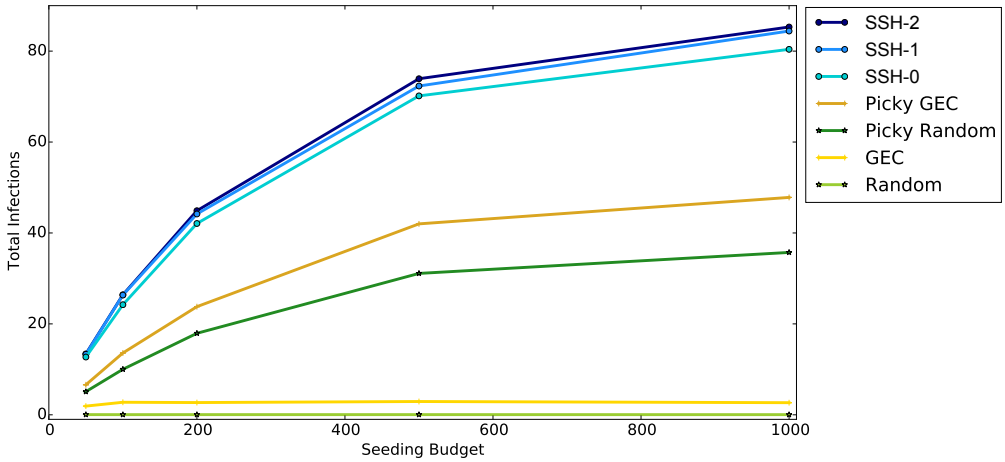


Figure 6: The number of successful seeding attempts as a function of the seeding budget B.

As described in the previous section, we assume the existence of an initially infected population of size F, prior to the beginning of the seeding attempts. Figure 7 reports the influence of F on the success rate of the different seeding heuristics. As expected, larger F values lead to higher success rates for all of the heuristics. While this increase exists, but is barely noticeable for the Random and Picky Random heuristics, it is clearly evident in the case of the SSH heuristics. Here as well, the SSH heuristics outperform

the other heuristics, even for small values of F, and the gap becomes larger as F grows.

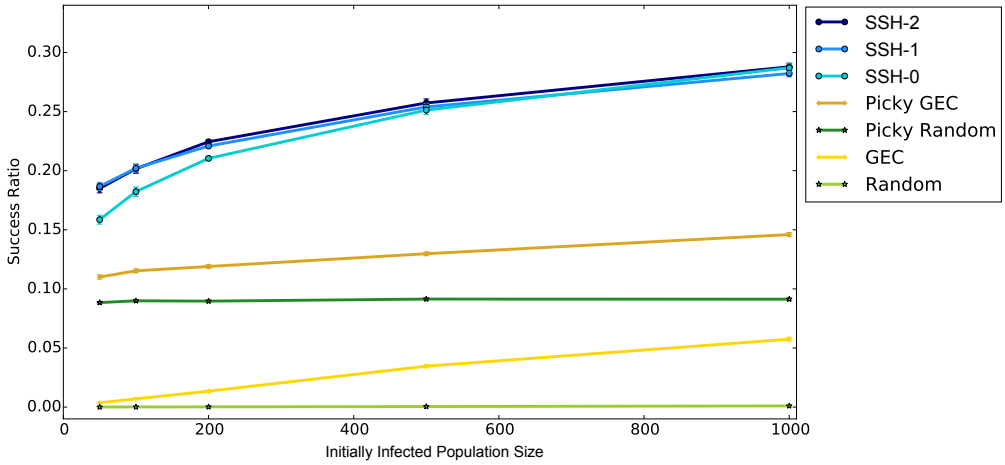


Figure 7: The proportion of successful seeding attempts as a function of the initially infected populations size F.

Figure 8 reports the influence of the infection time t_{inf} on the success rate of the different seeding heuristics. As can be seen in the figure, larger t_{inf} values lead to higher success rates for all of the heuristics. This is quite expected since lower t_{inf} values imply shorter infectious period of newly infected nodes, leading to lower social influence in the network at any given time. When the infection time is significantly short (around 5-10 time-steps), all of the heuristics suffer from poor performance. However, infection times of 50 time-steps and above result in high performance, where the improvement in performance gradually decreases with higher values of t_{inf}. Again, we see that the SSH approach (blue plots) significantly outperforms the other heuristics, for all of the examined values of t_{inf}.

ACCEPTED MANUSCRIPT

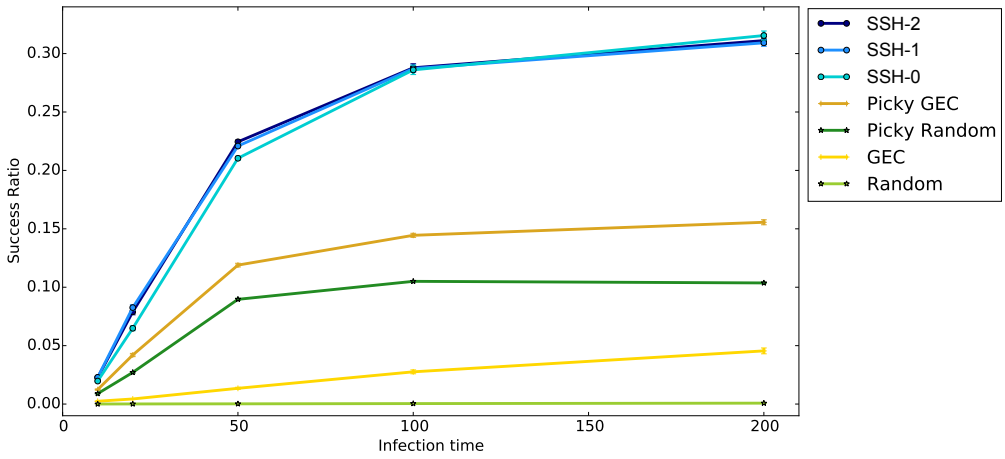


Figure 8: The proportion of successful seeding attempts as a function of the infection time t_{inf}.

The effect of the maximal social effect p_v^{soc} and the social threshold θ_v on the success rate of the different seeding heuristics is demonstrated in Figure 9. As can be seen in Figure 9 (top), higher values of p_v^{soc} are associated with higher success rates for all heuristics, as expected. Interestingly, the SSH approach grow super-linearly with p_v^{soc}, whereas all other heuristics grow roughly linearly. This causes the gap between the SSH approach (blue plots) and the other heuristics to become larger with higher values of p_v^{soc}. Indeed, when the social forces are stronger, the SSH approach, which better utilizes the information about the social influence is expected to reach better results. A similar (though inversed) trend of what was observed in Figure 9 (top) is presented in Figure 9 (bottom). This inversed trend is quite expected due to the $\frac{P_v^{soc}}{\theta_v}$ element in Eq. 1.

ACCEPTED MANUSCRIPT

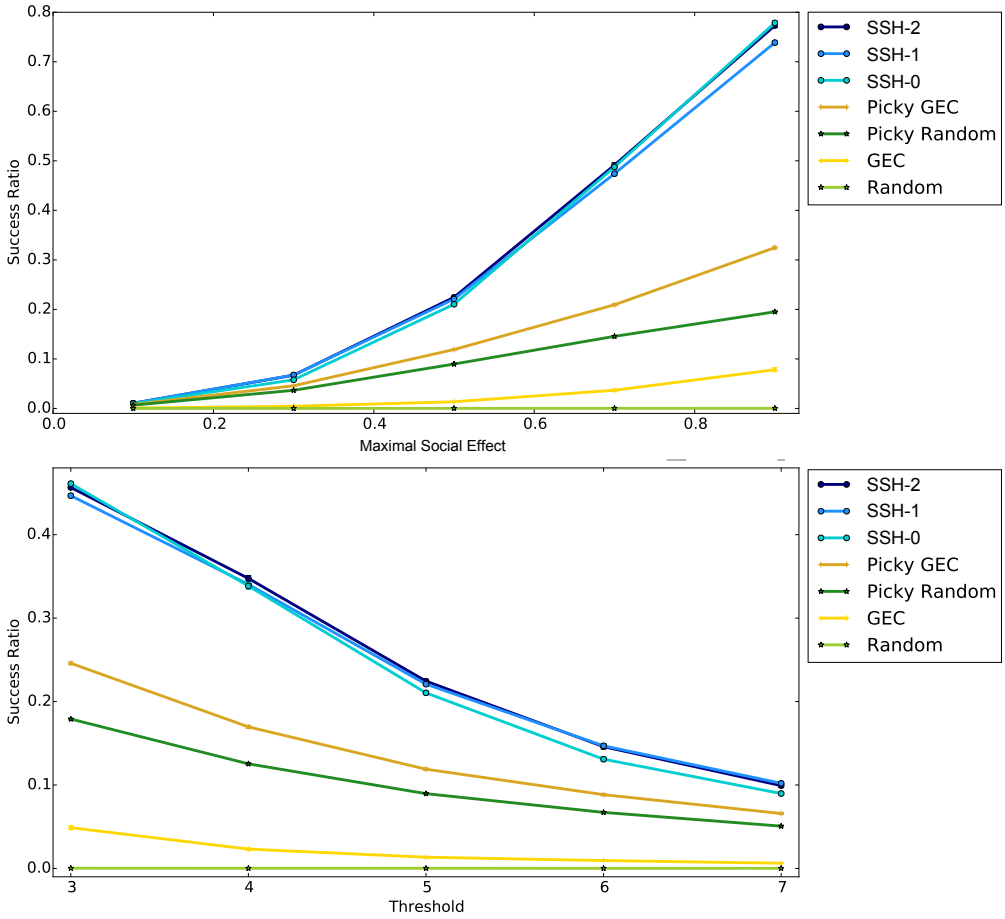


Figure 9: The proportion of successful seeding attempts as a function of the maximal social effect P_v^{soc} (top) and the social threshold θ_v (bottom).

While all of the above analyses focused on the social effect, where we set the individual effect to $P_v^{ind} = 0$, we now turn to analyzing the effect of the individual (non-social) effect on the success rate of the different seeding heuristics (see Figure 10). First, we observe that the success rates of all seeding heuristics increase with the individual effect P_v^{ind}. We can also see that the growth rate is similar in all heuristics, including the Random heuristic. This observation makes sense, since large values of P_v^{ind}, significantly reduce the importance of the social effect, and therefore make the scheduled approach less necessary. Similarly, we also see that for larger val-

ACCEPTED MANUSCRIPT

ues of individual effect (i.e., $P_v^{ind} \geq 0.05$), the SSH-0 heuristic outperforms the SSH-1 and SSH-2 heuristics.

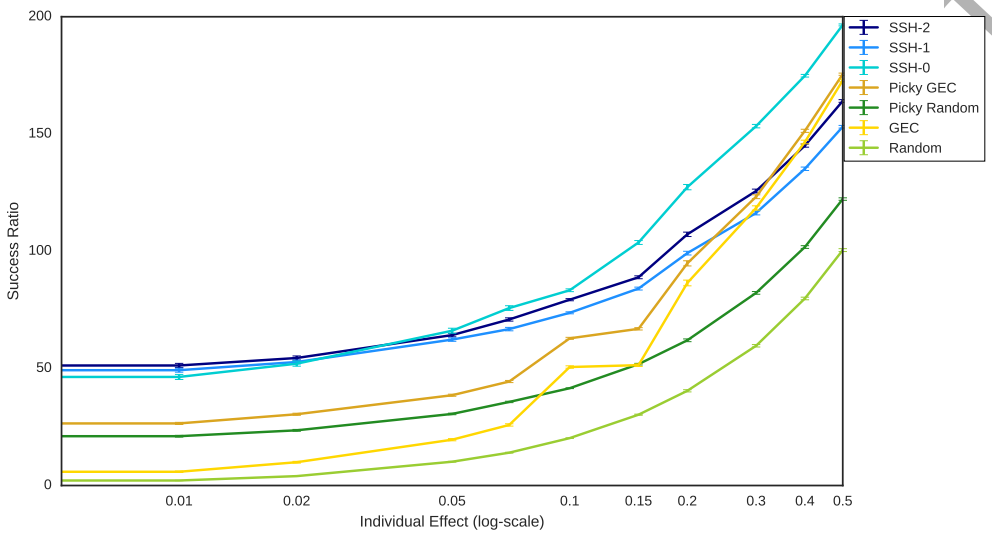


Figure 10: The number of successful seeding attempts as a function of the individual effect P_v^{ind}.

5.2.4. Introducing Uncertainty

The results described in the section above were obtained by assuming that the values of P_v^{soc} and θ_v are known. In most cases however, this is not a realistic assumption. At best, the distribution of these parameters can be estimated from previous marketing campaigns, but the specific parameter value for each person is still considered unknown. Based on this understanding, we conducted another experiment to inspect the performance of the proposed SSH approach within a more realistic scenario, in which P_v^{soc} and θ_v are assumed to be normally distributed and their means and standard deviations are assumed to be known; however, the actual values for each node are considered unknown.

Accordingly, in each set of executions, we first chose the mean and standard deviation. Then, we generated the "real" values for P_v^{soc} and for θ_v for each node based on the chosen distributions. Finally, we ran the different seeding heuristics where the means of the distributions were given as inputs,

ACCEPTED MANUSCRIPT

instead of their actual values. Note that in these experiments, the real values are only used in the simulative process, but is not used by the seeding node selection process.

Figure 11 reports the success rate of the different heuristics as a function of uncertainty (reflected by SD/mean).

The interior figure shows the success rate of Picky-Random as a function of uncertainty. As can be seen from the figure, the success rate increases moderately with uncertainty. The explanation for this is that high uncertainty values lead to a larger number of nodes with high P_v values (due to high P_v^{soc} and low θ_v values).

The exterior figure reports the relative success rate of the different heuristics, normalized with respect to Picky-Random, as a function of uncertainty. As can be seen from the figure, while the GEC, Picky-GEC and Random heuristics preserve the same relative success rate when uncertainty increases, the success rate of the SSH approach decreases. This is quite expected, as the SSH approach explicitly relies on the values of P_v^{soc} and θ_v for calculating the scores of nodes. Thus, an inaccurate estimation of these values due to a large standard deviation, leads to poorer selection of nodes and to a reduced performance. In contrast, all other heuristics which do not rely on the values of P_v^{soc} and θ_v, and therefore are not affected by inaccurate values of P_v^{soc} and θ_v. Nevertheless, even in relatively high uncertainty levels, the success rate of the SSH approach is still significantly higher than that of the other methods.

ACCEPTED MANUSCRIPT

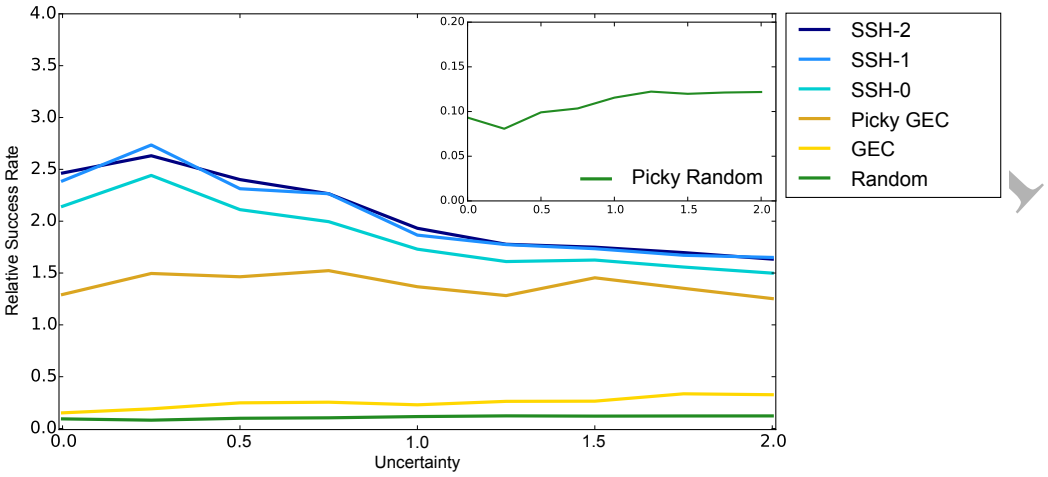


Figure 11: The improvement rate as a function of the degree of uncertainty (measured as the standard deviation of P_v^{soc} and θ_v).

5.2.5. Runtime

The different SSH heuristics represent growing degrees of future planning effort. While SSH-0 is fully greedy, in terms of planning only the current step, SSH-1 tries to plan one step ahead, and SSH-2 method plans two steps ahead. Although the SSH approach can be used with even higher number of planning steps (i.e., higher than 2), we did not find such large number of planning steps more effective. This observation is of high importance since the computational cost of planning ahead significantly increases with the network size, and due to the tremendous sizes of real-world social networks.

Figure 12 reports the runtime of the different heuristics as a function of the network size (different sample sizes of the Citation network). The runtime of SSH-0 and Picky-Random are roughly the same since they require to perform $O(1)$ operations for each one of the network nodes in each iteration. The runtime of SSH-1 is slightly higher since it requires some calculations of the first social circle of each network node in each iteration. The runtime of SSH-2 is again significantly higher than the runtime of SSH-1, since it requires some calculations on the first and second social circles of each network node (which cover a large fraction of the entire network) in each iteration. The runtime of GEC and Picky-GEC is also very high since it requires to calculate the Eigenvector centrality score for each of the network nodes (this is done once for each node, but the calculation is still expensive). Finally, we observe

ACCEPTED MANUSCRIPT

that starting from a certain network size (700,000), the runtime of GEC and Picky-GEC becomes even higher than that of SSH-2. Since the run-times of SSH-0 and SSH-1 seems to be reasonable, and since their success rate is almost as good as that of SSH-2, we will probably prefer to use them in future applications of real-world scenarios that involve large-scale networks.

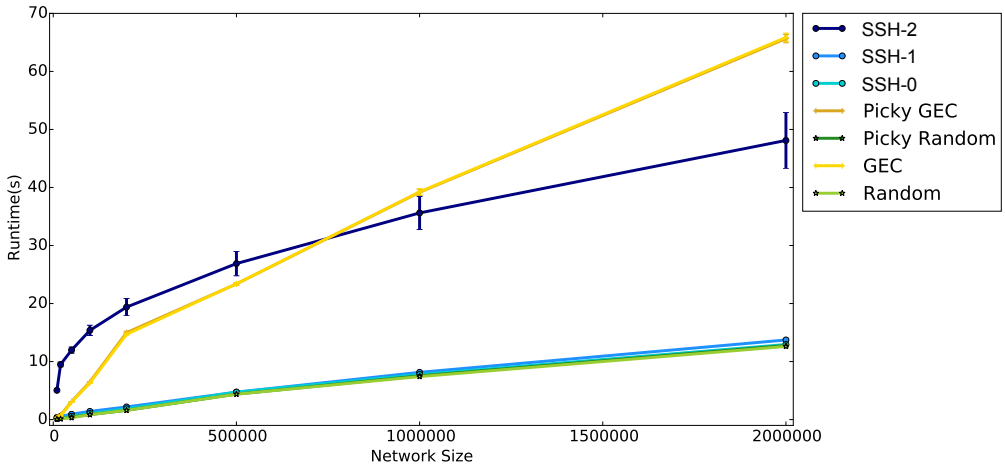


Figure 12: Runtime of the different heuristics as a function of the network size.

6. Summary and Future Work

Many works that study information diffusion in social networks consider a phenomenon by which information spreads virally through the network. Yet, unlike the spread of biological viruses that can be carried passively by agents and infect a significant portion of the network, information cascades are known to be shorter while long cascades are rather rare. These results do not necessarily imply that social impacts lost their importance, but rather that people spread information in a more selective way, which does not necessarily fit the assumptions of traditional models of infectious diseases.

We propose a new information diffusion model, named Active Viral Marketing (AVM), in which agents, e.g., sales representative of a company, communicate with network users, e.g., potential clients, and offer them a new product or service. The probability that a user accepts such an offer is based on the previous adoption rate of his/her friends, as well as his/her own tendency toward the product.

ACCEPTED MANUSCRIPT

Since promotion actions often incur some financial cost (limiting the number of clients that can be approached), the company has to select which users to approach and at what time, in order to increase the total adoption rate in the network. The need for a correct timing of approaching a customer is a direct result of memory retention loss, where new products quickly become an old habit and therefore the likelihood of influencing a peer node to purchase the new product quickly decays. The proposed Scheduled Seeding Heuristics (SSH) for user selection, chooses nodes that are most likely to accept an offer at any given time-step, and thus are more likely to influence their own non-infected neighbors at the next time-step.

In a large set of simulations, we show that the proposed heuristics increase the adoption rate in 30%-75% (depending on the initial conditions), over a state-of-the-art method that seeds the nodes according to their Eigenvector centrality score.

Having indicated that, it is important to note that the proposed method is mainly applicable to products that have a viral characteristic. These are products or services where a substantial part of the purchasing decision is based on social influence. In products or services for which the social forces are significantly less influential, it might be better to use the existing state-of-the-art methods of selecting nodes based on the network's topological properties.

Most diffusion models, including the proposed model, assume that all seeding actions have the same cost. As mentioned in Section 5.2.2, highly central nodes in social networks often represent celebrities or influencers, and the cost of seeding such entities is likely to be higher than that of less known individuals. Future studies should take into account different seeding costs for different nodes, depending for example on the network topology.

An interesting future extension to this work would be to study diffusion models that combine both the traditional passive infection together with the proposed continuous active seeding. Such a combined model is expected to be applicable for a wider range of real-world scenarios than each one of the two isolated models. Furthermore, it would be interesting to extend the proposed utility-based heuristics to support such a combined model.

The evaluation of this study is mainly based on simulations that utilize real-world network topologies. In future works, it would be interesting to enrich these simulations with additional real-world data such as purchasing history of users. In addition, it would be insightful to conduct a live experiment to compare the adoption rate obtained by the scheduled seeding

ACCEPTED MANUSCRIPT

approach versus the non-scheduled seeding approach.

Acknowledgements

This work was funded by the Kamin grant of the Israeli Chief Scientist (file number 58073).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Appendix A. Properties of the Influence Maximization Problem under the Active Viral Marketing Diffusion model

Appendix A.1. NP-Hardness

Claim: The influence maximization problem is NP-hard for the Active Viral Marketing diffusion model.

Proof: Consider an instance of the NP-hard Set Cover problem Garey & Johnson (1979): Given a collection of subsets $\{S_1, S_2, ..., S_m\}$ of a ground set $U = \{u_1, u_2, ..., u_n\}$, we wish to know whether there exist k of the subsets whose union is equal to U. We show that this can be viewed as a special case of the influence maximization problem for the Active Viral Marketing diffusion model. (We can assume that $k < n < m$.)

Given an arbitrary instance of the Set Cover problem, we define a corresponding directed graph as follows. The graph contains $1 + m + n$ nodes: a single node A, a node v_{S_i} for each subset S_i, a node v_{u_j} for each element u_j, and $m + \sum_{S_i} |S_i|$ directed edges: a directed edge (A, v_{S_i}) from A to each one of the v_{S_i} nodes and a directed edge (v_{S_i}, v_{u_j}) whenever $u_j \in S_i$.

In addition, consider the following parameters: $\theta = 1$, $t_{inf} = k$, $P_{ind} = 1$ and $P_{soc} = 0$ for node A, $\theta = 1$, $t_{inf} = 1 + k + n$, $P_{ind} = 0$ and $P_{soc} = 1$ for all other nodes, and a seeding budget of size $B = 1 + k + n$.

We note the following:

1. For the instance we have defined, activation is a deterministic process, as all probabilities are either 0 or 1.
2. A solution to the influence maximization problem must choose to seed node A at time-step $t = 0$ (seeding the node A at time-step $t = 0$ is assured to succeed while trying to seed any other node is assured to fail).
3. At least k out of the m nodes of type v_{S_i} must be seeded (a direct result of the seeding budget size).
4. Assuming that node A was seeded at time-step $t = 0$, seeding each one of the v_{S_i} nodes at time-steps $1 \leq t \leq k$ is assured to succeed (they only need one infected neighbor for the seeding action to succeed). Similarly, seeding each one of the v_{S_i} nodes at time-steps $t > k$ is assured to fail ($t_{inf} = k$ for node A).

ACCEPTED MANUSCRIPT

5. Following the four bullet points above, it stems that a solution to the influence maximization problem must choose to seed node A at time-step $t = 0$, k out of the m nodes of type v_{S_i} at time-steps $1 \leq t \leq k$ and all of the n nodes of type v_{u_j} at time-steps $k+1 \leq t \leq k+n$.
6. Assuming that node A was seeded at time-step $t = 0$ and k out of the m nodes of type v_{S_i} were seeded at time-steps $1 \leq t \leq k$, seeding a node v_{u_j} at time-steps $k+1 \leq t \leq k+n$ will succeed only if there exists a node v_{S_i} for which $u_j \in S_i$ and v_{S_i} is one of the k chosen nodes at time-steps $1 \leq t \leq k$.
7. The maximum number of nodes that can be seeded successfully is $1 + k + n$ (due to the budget size).

The answer to the Set Cover problem is True if and only if the solution to the corresponding influence maximization problem led to the successful seeding of exactly $1 + k + n$ nodes. ($1 + k + n$ successful seedings mean that we managed to seed successfully node A, k out of the m nodes of type v_{S_i} and all n nodes of type u_j, which further imply that there exists k subsets that cover the entire set U).

Since the Set Cover problem is known to be NP-hard, then so is the influence maximization problem for the Active Viral Marketing diffusion model.

Appendix A.2. Sub-Modularity

Consider the Active Viral Marketing diffusion model defined above and the function F, which receives an ordered subset of network nodes to be seeded (at consecutive time-steps) as input, and returns the expected number of successful seedings as output. By definition, F is not sub-modular, since sub-modular functions receive a set rather than an ordered set as input. Moreover, even if we extend the definition of sub-modular functions to the case of ordered sets, F would still not satisfy the sub-modularity condition. To illustrate why, consider a network composed of two nodes v_1 and v_2 and a single edge between them, and the following parameters: $P_v^{ind} = 0.1$, $P_v^{soc} = 0.9$, $\theta_v = 1$ and $t_{inf} = 2$, for all network nodes. Now, consider the two ordered sets $X = ()$ and $Y = (v_1)$. The sub-modularity condition requires (among the rest) that adding v_2 to Y will result in a lower gain in F than adding it to X (since $X \subset Y$). More specifically, it is required that $F((v_1, v_2)) - F((v_1)) < F((v_2)) - F(())$ However, it is easy to see that $F(()) = 0$, $F((v_1)) = 0.1$, $F((v_2)) = 0.1$, and $F((v_1, v_2)) = 0.1 + (0.1 \cdot 1 + 0.9 \cdot 0.1) = 0.29$. Therefore, $F((v_1, v_2)) - F((v_1)) = 0.19 > F((v_2)) - F(()) = 0.1$ and the sub-modularity condition is violated.

ACCEPTED MANUSCRIPT

References

Anderson, R. M., May, R. M., & Anderson, B. (1992). *Infectious diseases of humans: dynamics and control* volume 28. Wiley Online Library.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men. S*, (pp. 222–236).

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *Science*, *341*, 1236498.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2004). Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters*, *92*, 178701.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, *29*, 555–564.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, *27*, 55–71.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, *25*, 163–177.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American journal of Sociology*, *113*, 702–734.

Chen, W., Wang, C., & Wang, Y. (2010a). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1029–1038). ACM.

Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 199–208). ACM.

Chen, W., Yuan, Y., & Zhang, L. (2010b). Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 88–97). IEEE.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Chierichetti, F., Kleinberg, J., & Panconesi, A. (2014). How to schedule a cascade in an arbitrary graph. *SIAM Journal on Computing*, *43*, 1906–1920.

Fibich, G. (2016). Bass-sir model for diffusion of new products in social networks. *Physical Review E*, *94*, 032305.

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability: A guide to the theory of npcompleteness (series of books in the mathematical sciences), ed. *Computers and Intractability*, (p. 340).

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 623–638). ACM.

Goldenberg, J., Libai, B., & Muller, E. (2001a). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, *12*, 211–223.

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

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, *30*, 389–404.

Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2010). Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 241–250). ACM.

Goyal, A., Bonchi, F., Lakshmanan, L. V., & Venkatasubramanian, S. (2013). On minimizing budget and time in influence propagation over social networks. *Social network analysis and mining*, *3*, 179–192.

Goyal, A., Lu, W., & Lakshmanan, L. V. (2011). Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web* (pp. 47–48). ACM.

ACCEPTED MANUSCRIPT

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*, 1420–1443.

Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, *75*, 55–71.

Jankowski, J., Bródka, P., Kazienko, P., Szymanski, B. K., Michalski, R., & Kajdanowicz, T. (2017a). Balancing speed and coverage by sequential seeding in complex networks. *Scientific reports*, *7*, 891.

Jankowski, J., Bródka, P., Michalski, R., & Kazienko, P. (2017b). Seeds buffering for information spreading processes. In *International Conference on Social Informatics* (pp. 628–641). Springer.

Jung, K., Heo, W., & Chen, W. (2012). Irie: Scalable and robust influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on* (pp. 918–923). IEEE.

Katz, E., & Lazarsfeld, P. (1955). Personal influence.

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

Khelil, A., Becker, C., Tian, J., & Rothermel, K. (2002). An epidemic model for information diffusion in manets. In *Proceedings of the 5th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems* (pp. 54–60). ACM.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 497–506). ACM.

Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web* (pp. 915–924). ACM.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007a). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 420–429). ACM.

Leskovec, J., & Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., & Hurst, M. (2007b). Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining* (pp. 551–556). SIAM.

Lin, S., Hu, Q., Wang, F., & Philip, S. Y. (2014). Steering information diffusion dynamically against user attention limitation. In *Data Mining (ICDM), 2014 IEEE International Conference on* (pp. 330–339). IEEE.

Long, C., & Wong, R. C.-W. (2011). Minimizing seed set for viral marketing. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 427–436). IEEE.

Mahajan, V., Muller, E., & Bass, F. M. (1991). New product diffusion models in marketing: A review and directions for research. In *Diffusion of technologies and social behavior* (pp. 125–177). Springer.

Newman, M. (). Networks: an introduction. 2010. *United Slates: Oxford University Press Inc., New York*, (pp. 1–2).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.*. Technical Report Stanford InfoLab.

Seeman, L., & Singer, Y. (2013). Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on* (pp. 459–468). IEEE.

Tong, G., Wu, W., Tang, S., & Du, D.-Z. (2017). Adaptive influence maximization in dynamic social networks. *IEEE/ACM Transactions on Networking (TON)*, *25*, 112–125.

Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nature physics*, *8*.

ACCEPTED MANUSCRIPT

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific reports*, *2*.

Zhou, J., Liu, Z., & Li, B. (2007). Influence of network structure on rumor propagation. *Physics Letters A*, *368*, 458–463.

ACCEPTED MANUSCRIPT