Timing Matters: Influence Maximization in Social Networks through Scheduled Seeding

Dmitri Goldenberg[1], Alon Sela[1,2], Erez Shmueli[1,*]

Abstract—One highly studied topic in the field of social networks is the search for influential nodes, that when seeded (i.e. activated intentionally), may further activate a large portion of the network through a viral contagion process. Indeed, various mathematical models were proposed in the literature to characterize the dynamics of such diffusion processes, and different solutions were suggested for maximizing influence under such models. However, most of these solutions focused on selecting a set of nodes to be seeded at the initial phase of the diffusion process.

This work suggests a scheduled seeding approach which aims at finding not only the best set of nodes to be seeded, but also the right timing to perform these seedings. More specifically, we identify three different properties of existing contagion models that can be utilized by a scheduled approach to improve the total number of activated nodes: (1) stochastic dynamics, (2) diminishing social effect, and (3) state-dependent seeding. By analyzing each of these properties separately, we demonstrate the advantages of the scheduled seeding approach over the traditional initial seeding approach, both by theoretical and empirical evaluation. Our analysis presents an improvement of 10%-70% in the final number of infected nodes, when using the scheduled seeding approach.

Our findings have the potential to open up a new area of research, focusing on finding the right timing for seeding actions, thereby helping both in improving our understanding of information diffusion dynamics and in devising better strategies for influence maximization.

Index Terms—Influence Maximization; Social Networks; Scheduling

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 resembles the spread of a virus. Along with their role as a major social communication channel in our lives, social networks redesigned the way individuals consume content, their personal preferences, and their decision making processes. These traits are also the reasons why political parties, commercial firms or even terror organizations are increasingly using social networks [31]. Indeed, social networks are often involved in shaping key historical events [56], [23], [25], [5], and are commonly used by marketers for market analysis and promotion of products.

One of the main mechanisms that enables information diffusion in social networks is social influence. In fact, it has been known for many years that one's social connections have an immense influence on his personal attitude and decision making process [3], [38], [22], [57]. Social proximity in a social network was even shown to predict tendencies that were once believed to be genetic, such as obesity, smoking, happiness [14].

Over the years, various mathematical models were proposed in the literature to characterize the dynamics of information diffusion in social networks (e.g., [30], [17]). One highly studied problem that arises in the context of such models is finding influential nodes, that if seeded (i.e. activated intentionally), may further activate a large fraction of the network through a viral contagion process. This problem is commonly referred to as the influence maximization problem. While many solutions were suggested for this problem (e.g., [11], [52], [54]), most of these solutions focus on selecting the set of nodes to be seeded at the initial phase of the contagion process.

This work suggests a scheduled seeding approach which aims at finding not only the best set of nodes to be seeded, but also the right timing to perform these seedings. In contrast to the existing initial seeding approach, which selects a subset of nodes to be seeded at the beginning of the contagion process, the scheduled approach provides the marketer an ability to adapt the seeding strategy as the process unfolds in order to gain a higher activation rate.

In particular, we identify three different properties, exhibited by some of the existing contagion models, that can be utilized by a scheduled approach to improve the total number of activated nodes: (1) when the contagion process exhibits stochastic dynamics, deferring seeding actions to later stages allows the marketer to gain additional knowledge as the contagion process evolves; (2) when the contagion process exhibits both a diminishing social effect over time and a complex contagion aspect (activation of a node requires it to have multiple infectious neighbors simultaneously), a scheduled approach allows the marketer to choose the right timing to perform a seeding action; and

[1] *Department of Industrial Engineering at Tel-Aviv University, Tel-Aviv, Israel.*
[2] *Department of Industrial Engineering and Management at Ariel University, Ariel, Israel*
[*] *Corresponding author, shmueli@tau.ac.il.*

(3) in cases where the current states of network nodes affect the success of the seeding action itself, a scheduled approach allows to improve the success rate by utilizing the effect of previous seeding actions.

For each one of these three properties, we demonstrate the advantages of the scheduled seeding approach over the traditional initial seeding approach, both by theoretical and empirical evaluation. The results of our evaluation show an improvement of 10%-70% in the final number of infected nodes, when using the scheduled seeding approach.

Our findings have the potential to open up a new area of research, focusing on finding the right timing for seeding actions, thereby helping both in improving our understanding of information diffusion dynamics and in devising better strategies for influence maximization.

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 models of viral diseases, followed by 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. Finally, we provide a categorization of these models along six different dimensions.

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 [1]. 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 [1]. 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.

In contrast to the SIR model that was designed to describe epidemiological phenomena, Bass diffusion model [37] was developed by Frank Bass in order to describe the process of how new products get adopted in a population. The main addition of this model is the innovation factor, namely, the ability of an individual to become infected (adopt a product), by himself, without interaction with other infected individuals. The model classifies the adopters of the product into two classes: innovators (first to adopt innovative products) and imitators (those who adopt the product after interacting with other adopters).

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 [22], [30]. 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 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. The exact calculation of $X_v(t)$ is provided in the next formula:

$$X_v(t) = \begin{cases} 1, & \text{if } X_v(t-1) = 1 \text{ or } \sum\limits_{w \in N(v)} b_{v,w} X_w(t-1) \geq \theta_v \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

It is important to note that under the assumptions that the threshold θ_v is known for all nodes, the dynamics and outcome of the model become deterministic, allowing to pre-calculate the exact result of any intervention strategy. This deterministic setting of the model is also known as the Tipping Model [51]. However, it is commonly assumed that these thresholds are unknown, and are uniformly distributed in the range $[0,1]$. While the dynamics of the model are still defined deterministically, the fact that the thresholds values are unknown during the process, prevents us from calculating the exact outcome, but only to estimate its probability.

Another well-studied individual-based information diffusion model is the Independent Cascade model [17], [18]. 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 in [15]. This model proposes a new contagion process which combines properties of SIR and Bass [37] 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 now, it has a positive probability to adopt the product in the nearest future. However, unlike the basic Bass model, Bass-SIR does not assume that an infected node will stay infective forever, and will eventually change its state to recovered.

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, [30] proposed two models which aimed at generalizing many of the 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 [19]. 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 [29].

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.

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 [7].

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 [41], the Betweenness centrality [8] and the Eigenvector centrality [6]. 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 [7] and [39].

With respect to influence maximization, several works investigated the efficiency of seeding central nodes. The work by [24], 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 [4], 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, [30] 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 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* [32] and *CELF++* [21] 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. [11], [10], [28], [12]. [11] 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. [10] 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. [28] 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.

2.2.2 Adaptive Seeding Strategies

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, [45] 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, [53] 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 each time step, is performed while taking into account the realization of the previous seeding attempts.

Jankowski et Al. [26], [27] 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.

In another study by Ni [40], the author propose a Markov decision process optimization within an "Incremental Chance" diffusion framework. According to the contagion model the author proposes, the probability of a node to get activated is proportional to the fraction of its infected neighbors, and once a node becomes active, it remains infective. The goal in this case is to minimize the time taken to reach a complete influence by selecting the seeding set, under the constraint that only a portion of the budget is available at each time-step.

[13] 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.

[35] 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:

$$p_v(t) = d_v + \sum_{w \in N(v)} b_{v,w} * X_w(t-1) \qquad (2)$$

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 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 [3]. [22] 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, [9] 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 [20] 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 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 [55]. 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 [34], 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

TABLE 1
A comparison of different contagion models

Model	Granularity	Stochasticity	Diminishing Social Effect	State-Dependent Seeding	Complex Contagion	Innovation
SIR[1]	C	SC	Y	N	N	N
Bass[37]	C	SC	N	N	N	Y
Linear Threshold[30]	I	SC	N	N	Y	N
Tipping Model[52]	I	D	N	N	Y	N
Independent Cascade [17]	I	SC	Y	N	N	N
Bass-SIR[15]	I	SC	Y	N	N	Y
Two-Stage Seeding[45]	I	SCS	N	Y	N	N
Dynamic Independent Cascade[53]	I	SCS	N	N	N	N
Incremental Chance[40]	I	SC	N	N	N	N
Scheduled Cascade[13]	I	SS	N	Y	Y	N
Push-Driven Cascade[35]	I	SS	N	Y	Y	Y
	Compartment based-(C) or Individual based-(I)	*Deterministic-(D), Stochastic Contagion-(SC), Stochastic Seeding-(SS), Fully Stochastic-(SCS)*	*Dimishing effect over time Yes-(Y)/ No-(N)*	*Seeding action success depends on current state - Yes-(Y)/ No-(N)*	*Activation triggered only by multiple sources of influence - Yes-(Y)/ No-(N)*	*Self activating nodes - Yes-(Y)/ No-(N)*

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 [16], 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.

In another study by Aral et al. [2], the authors study the influence and susceptibility to an innovative product, in a representative sample of 1.3 million Facebook users. The study showed that different individuals present different levels of influence and susceptibility, where age, gender and marital status play a key role in determining these levels. More specifically, younger users were found to be more susceptible than older users; men were more influential than women, while women had more influence on men than on other women; and married individuals were the least susceptible. Interestingly, the research also revealed that influential individuals are less susceptible than non-influential individuals, and that they tend to cluster in the network. These findings can help, for example, in obtaining a more accurate forecast of individuals' cascades-related properties, such as the threshold value θ_v or the weights of edges (influence relationships).

2.4 Summary and Comparison of Models

Table 1 compares the various contagion models discussed above along six dimensions:

- **Granularity:** Whether the model is Compartment-based (C) or Individual-based (I)
- **Stochasticity:** Whether the model is stochastic- Fully Deterministic (D), Only Stochastic Contagion (SC), Only Stochastic Seeding (SS) or Fully Stochastic - Both Stochastic Contagion and Seeding (SCS)
- **Diminishing Social Effect:** Yes (Y) or No (N)
- **State-Dependent Seeding:** The seeding action success depends on current state - Yes (Y) or No (N)
- **Complex Contagion:** Activation triggered only by multiple sources of influence - Yes (Y) or No (N)
- **Innovation:** Self activating nodes - Yes (Y) or No (N)

While the influence maximization problem has been studied in many works, the adaptive seeding approach (to which we call scheduled seeding hereinafter) has not gained a lot of focus yet. In the following sections, we focus on three aspects of the contagion model dynamics that emphasize the importance of scheduled seeding: (1) Stochasticity, (2) Diminishing Social Effect, and (3) State-Dependent Seeding.

3 Stochastic Dynamics

Diffusion in networks is usually modeled by stochastic processes. Stochasticity in such models is usually the result of either: (1) lack of knowledge about the actual values of parameters (such as the threshold values θ_v in the Linear threshold model), or (2) the realizations of events are of probabilistic nature (such as the probability of peer-activation in the Independent Cascade model). In fact, the

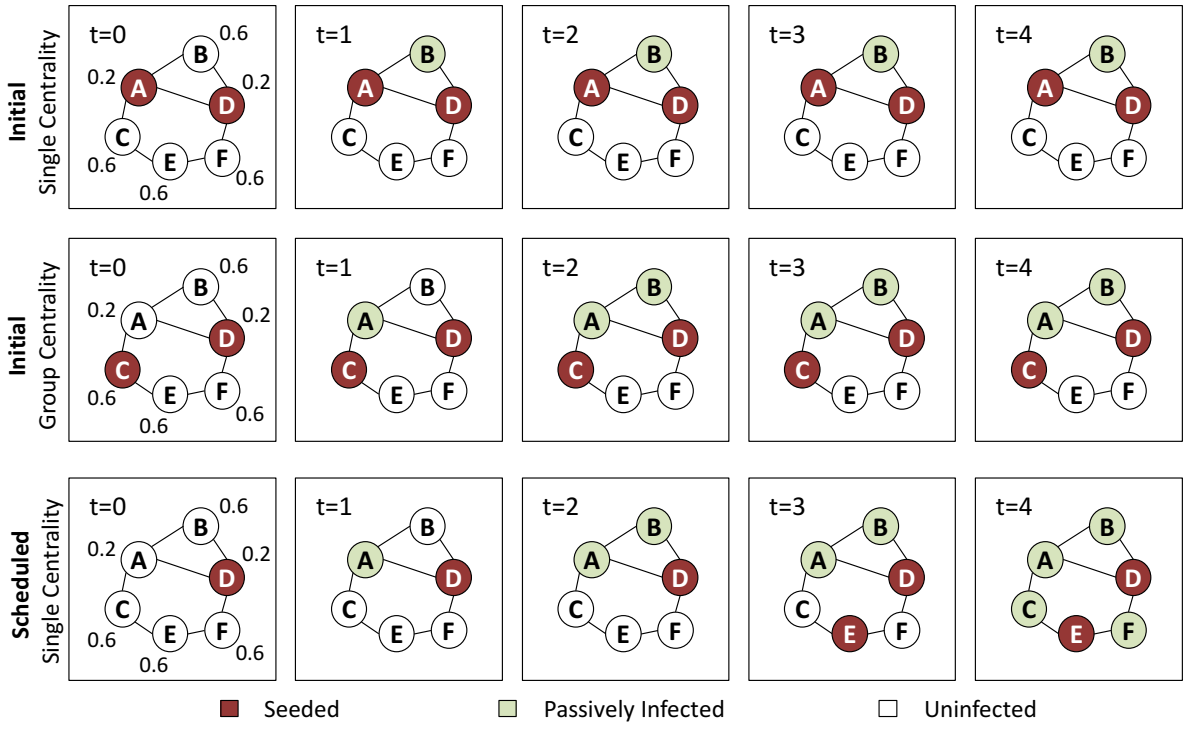

Fig. 1. Illustrating the advantage of scheduled seeding in a stochastic setting.

classical *SIR* model is often taught as one of the first basic stochastic models in undergraduate courses.

Therefore, introducing an adaptive seeding strategy that utilizes past knowledge on which nodes have already been activated in order to determine the next node to be seeded, can lead to a significantly improved adoption rate. For example, such a scheduled approach, can avoid wasting budget on seeding attempts of nodes that could become activated passively as part of the viral process.

3.1 A Toy Example

To illustrate the advantages of using a scheduled seeding approach under a stochastic setting, we consider a small network of six nodes, under the Linear Threshold contagion model (see Figure 1). In this network, nodes A and D have a high degree centrality ($d_v = 3$) and a low threshold value ($\theta_v = 0.2$). The rest of the nodes - B, C, E and F - have $d_v = 2$ and $\theta_v = 0.6$. It is worth noting that A and D also have the highest Eigenvector, Page-Rank and Betweenness centrality scores in this networks. We assume that although the network topology (and thus the centrality measures) is available to the marketer, the threshold values are unknown.

The figure illustrates the execution of three potential heuristics. The first heuristic (upper row of the figure) assumes that the marketer will assign a centrality score to each one of the nodes, and choose the two nodes with the highest scores, A and D, to be seeded at once at time-step $t = 0$. Afterwards, at time-step $= 1$, node B will reach its threshold value ($2 \div 2 > 0.6$) and will become activated passively. Nodes C, E and F will not reach their threshold value, and therefore the final number of activated nodes will be 3.

The second heuristic (middle row of the figure) takes a slightly more sophisticated approach in which we calculate the joint centrality score of different subsets of nodes (as opposed to individual scores). For example, if the marketer chooses to use the Group Betweenness Centrality (GBC) measure [43], it will choose one of the subsets $\{A, D\}$, $\{A, F\}$ or $\{C, D\}$, all have a GBC score of 0.4. The option of selecting the nodes $\{A, D\}$ was already presented in the first row of the figure. Selecting to seed one of the symmetric subsets $\{A, F\}$ and $\{C, D\}$ will result in a slightly higher number of activated nodes (i.e. 4 instead of 3).

The third heuristic (bottom row of the figure) takes a scheduled approach which chooses a single node (with the highest centrality score) to be seeded at time-step $t = 0$ and a second node at a later stage. In our case, assuming that the marketer chooses to seed the node D at time-step $t = 0$ (the option of choosing A is symmetrical), nodes A and B become activated passively at time-steps $t = 1$ and $t = 2$ respectively. At time-step $t = 3$, when no more passive activations are anticipated, the marketer chooses to seed node E which has the highest centrality score out of all uninfected nodes. Finally, at time-step $t = 4$, nodes C and D become activated passively, resulting in a final number of 6 activated nodes.

This example emphasizes two important points: (1) the scheduling approach gains useful knowledge by postponing seeding decisions to later time-steps, and (2) although selecting the same seed set (nodes D and E) at $t = 0$ would result in the exact same outcome (6 activated nodes), a seeding heuristic that is based only on the network topology (note that the threshold values are unknown) would not have a way to prefer these nodes over the others at time-step $t = 0$.

3.2 Theoretical Analysis

As demonstrated in the example above, the scheduled approach seems to have a significant advantage over the initial seeding approach, in settings of stochastic models. Inspired by this observation, and assuming the Independent Cascade or Linear Threshold stochastic models, we present below a general approach to transform an initial seeding heuristic to an incremental (scheduled) heuristic that guarantees the same performance, but may achieve this performance with less seeding actions. For the purposes of the following analysis (and similarly to [30]), we consider all stochastic outcomes to be drawn in-advance, however, they are kept unknown.

Algorithm 1 The Initial/Incremental Seeding Algorithm

Input:
$IsIncremental$ - whether INCREMENTAL or INITIAL seeding
$S = (v_1, v_2, \ldots, v_m)$ - an ordered set of potential nodes to seed (m is the seeding budget)
Output:
S^* - actual nodes that were seeded and their seeding time-steps

1: $t \leftarrow 0$
2: $S^* \leftarrow \emptyset$
3: $A \leftarrow \emptyset$
4: **for** v in S **do**
5: **if** $v \in A$ **then**
6: continue
7: **else**
8: $S^* \leftarrow S^* \cup \{(v, t)\}$
9: **if** $IsIncremental$ **then**
10: $t \leftarrow t + T(A \cup \{v\})$
11: $A \leftarrow F(A \cup \{v\})$
12: **end if**
13: **end if**
14: **end for**
15: return S^*

The incremental algorithm seeds a single node in S at a time, where after seeding the node, it waits for the viral process to end, and only then it seeds the next node in S that was not already activated by the viral process.

More specifically, given a set of nodes S, a pseudo code of the initial and incremental seeding heuristics is provided in Algorithm 1. The algorithm takes as input a flag $IsIncremental$ stating whether to use the incremental or initial heuristic and an ordered set of potential nodes to seed S, and returns as output a subset of actual nodes to seed and their corresponding time-steps S^*. It is easy to see that when $IsIncremental = False$ (i.e., the case of initial seeding), the algorithm simply returns the nodes in S, all assigned with time-step $t = 0$. The incremental algorithm differs only in lines 9-11, where it applies the functions $F(A)$ and $T(A)$. $F(A)$ takes as input the current set of activated nodes A and

returns the final set of activated nodes as a result of the viral process only (i.e with no additional seedings). $T(A)$ returns the number of time-steps required for the viral process to converge.

In terms of runtime complexity, the algorithm iterates over the whole set of potential candidates (S), and therefore given a precomputed seeding set S, the runtime complexity is $O(|S|)$.

Now, let $Initial(S)$ denote the set of nodes returned by the initial algorithm and $Incremental(S)$ denote the set of nodes returned by the incremental algorithm. The following claim holds:

Claim: Given a set of nodes S, $F(Initial(S)) = F(Incremental(S))$.

Proof:
The incremental algorithm goes over each one of the nodes $v \in S$ and either adds v to the seeding set S^* (and therefore also to the set of activated nodes) or ensures that v was already passively activated. Therefore, necessarily $S \subseteq F(Incremental(S))$. Due to the monotonic property of both the Independent Cascade and Linear Threshold models, we get that $F(S) \subseteq F(F(Incremental(S)))$. However, since $S = Initial(S)$ and $F(F(A)) = F(A)$ for every set A, we get that $F(Initial(S)) \subseteq F(Incremental(S))$. Now, since $Incremental(S) \subseteq S = Initial(S))$, and again due to the monotonic property of both contagion models, we get that $F(Incremental(S)) \subseteq F(Initial(S))$. Therefore, $F(Initial(S)) = F(Incremental(S))$. ■

Since the incremental seeding heuristic manages to activate all the nodes that are activated by the initial seeding heuristics (i.e., $F(Initial(S)) = (Incremental(S))$), and since it manages to do it in at most the same number of seeding actions (i.e. $Incremental(S) \subseteq Initial(S)$), the incremental heuristic can utilize the remaining (unused) seeding actions to obtain additional activations (that are not obtained by the initial heuristic).

3.3 Evaluation

In the previous subsection we showed that under the Linear Threshold and Independent Cascade models, the scheduled approach obtains the same performance of the initial approach while using a potentially smaller number of seeding actions, and therefore if it utilizes the remaining unused seeding actions, it may obtain better performance. In this subsection, we report the results of simulations that we conducted to understand the extent of such improvement.

3.3.1 Experimental Setting

All of our simulations 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 simulation started with an uninfected network at time-step $t = 0$. Then, we simulated a contagious process where seeding actions were spread over time. The simulation ended when the seeding budget exceeded and there were no more potential infections.

Prior to running each simulation, actual values of parameters were assigned with each network node and edge, according to the contagion model used, similarly to the setup in [30]. More specifically, in the Linear threshold model, the weights of the edges $b_{w,v}$ were assigned such that $\sum_{w \in N(v)} b_{v,w} = 1$ and $b_{v,w} = \frac{1}{|N(v)|}$ for all $w \in N(v)$. The threshold value θ_v of a node v was uniformly drawn from the interval $[0, 1]$. In the Independent Cascade model, the probabilities of all edges, $p_{v,w}$, were set to a constant value.

The network topologies used in this experiment are based on real-world networks taken from Stanford Large Network Dataset Collection [33]. In order to maintain a roughly consistent size across networks, we obtained a sample of 10,000 nodes from each of the original networks. The sample was obtained by selecting a random seed node and performing a random walk from it until we reached 10,000 nodes (in the case where the connected component was not large enough, we selected another seed node and repeated the same process). The properties of the sampled

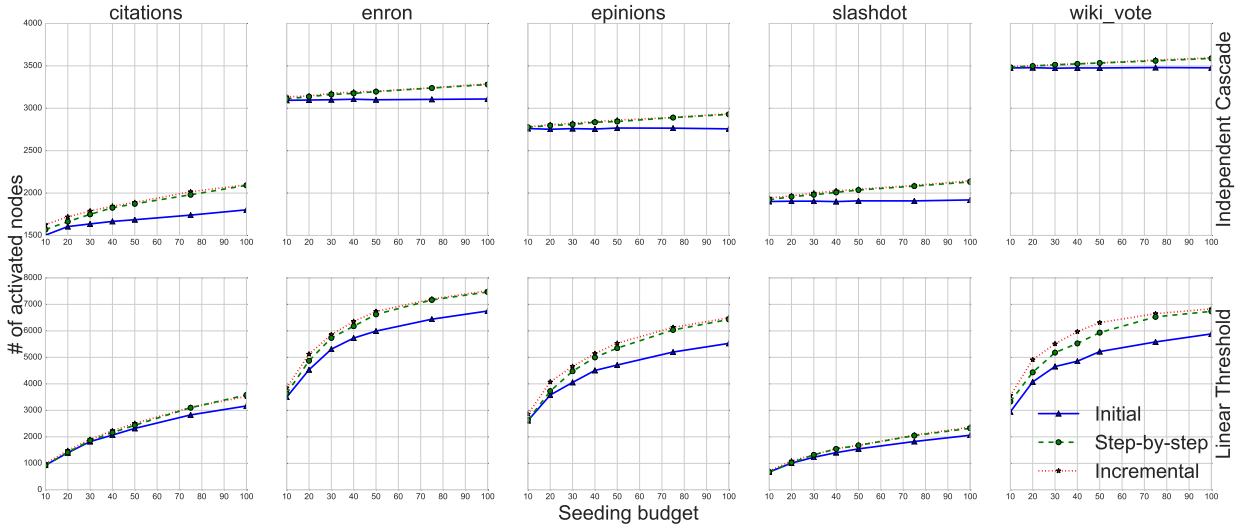

Fig. 2. Total number of activated nodes as a function of the seeding budget for various network topologies.

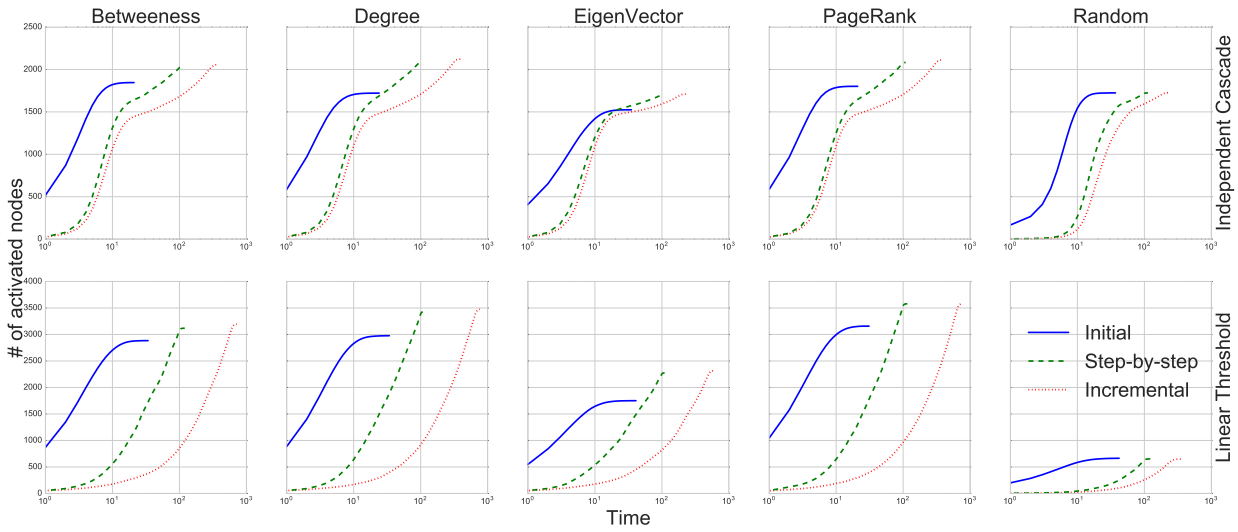

Fig. 3. Number of activated nodes as a function of time for various seeding scores.

networks are detailed in Table 2 (the properties of the original networks appear in brackets).

TABLE 2
Properties of the social networks used in our simulations

Network	Number of Nodes	Average Degree	Average Clustering
Citations	10000 (2244018)	6.81 (1.86)	0.20 (0.02)
Enron	10000 (36692)	10.13 (10.02)	0.44 (0.50)
Epinions	10000 (75879)	9.11 (10.69)	0.12 (0.14)
Slashdot	10000 (82168)	7.25 (14.18)	0.03 (0.06)
Wiki Vote	7115 (7115)	28.32 (28.32)	0.14 (0.14)

For selecting the nodes to seed and their time-steps, we used the following scheduling approaches:

- Initial - Performing all seeding actions at t=0
- Step-by-Step - Performing a single seeding action at each time-step.
- Incremental - Performing a single seeding action each time there are no more passive activations 3.2.

Our simulations compared the total number of activated nodes obtained by the different seeding heuristics for various parameters: network topology and, seeding budget, seeding score, simulation time-step, network size and infection probability. The reported results are the averages of 1000 simulations for each set of parameters.

3.3.2 Results

Seeding Budget

Figure 2 presents a comparison of the three seeding heuristics for different seeding budgets, over various network topologies, when fixing the seeding score to PageRank and $p_{v,w}$ to 0.1.

The two rows of the figure represent the different contagion models (Independent Cascade and Linear Threshold), the five columns of the figure represent different network topologies (Citations, Enron, Epinions, Slashdot, and Wiki-vote), the three plots in each figure represent the different scheduling approaches (Initial, Step-by-Step and Incremental), the x-axis represent the seeding budget, and the y-axis represent the total number of activated nodes (at the end of the simulation).

As can be seen in the figure, larger seeding budgets lead to higher number of activated nodes as expected. Moreover, it seems that the two scheduled approaches utilize the increase in seeding budgets in a better way, and therefore increase their gap from the initial approach for larger seeding budgets.

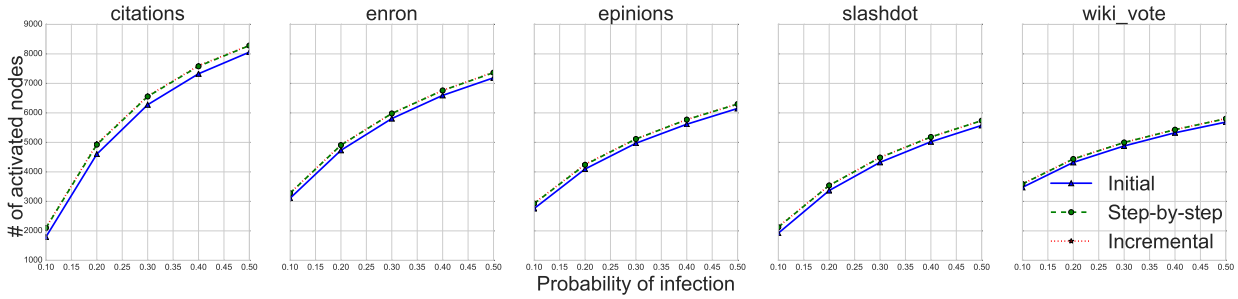

Fig. 4. Number of activated nodes as a function of the infection probability ($p_{v,w}$) for various network topologies.

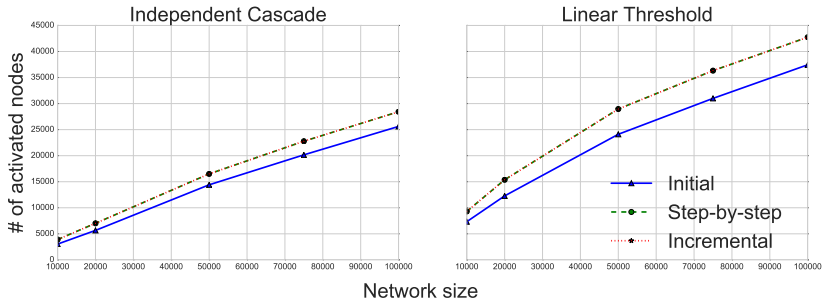

Fig. 5. Number of activated nodes as a function of network size.

A summary of the improvement rates obtained by the Step-by-Step scheduled approach (at the end of the simulation), for a fixed seeding budget of 100, is presented in Table 3. Note that the improvement rates of the Incremental approach are quite similar and were therefore omitted.

TABLE 3
Improvement rates of the Step-by-Step approach over the Initial seeding approach

Network	Independent Cascade	Linear Threshold
Citations	16.0%	13.1%
Enron	5.5%	10.6%
Epinions	6.2%	16.3%
Slashdot	11.1%	12.8%
Wiki Vote	3.2%	14.5%

Time

Figure 3 presents a comparison of the three seeding heuristics over time, for five different seeding scores, when fixing the network topology to Citations, the seeding budget to 100 and $p_{v,w}$ to 0.1.

The two rows of the figure represent the different contagion models (Independent Cascade and Linear Threshold), the five columns of the figure represent the seeding score function (Betweeness centrality, Degree centrality, EigenVector centrality, PageRank centrality and Random score), the three plots in each figure represent the different scheduling approaches (Initial, Step-by-Step and Incremental), the x-axis represent the time passed (in log-scale), and the y-axis represent the number of activated nodes.

As can be seen from the figure, the two scheduled approaches (Step-by-Step and Incremental) reach a significantly higher number of activated nodes, in comparison to the Initial approach, at the end of the simulation, as expected. We also observe a clear tradeoff between the total number of activated nodes and the time it takes to reach this number, where the initial approach ends much faster than the two others. It is also interesting to note that while the Incremental approach obtains the best performance, it is only slightly better than the Step-by-Step approach (as we also noted above) and its runtime is longer in an order of magnitude.

Network Size

Figure 5 presents a comparison of the three seeding heuristics for different network sizes, when fixing the network topology to Citations, the seeding budget to 100, the seeding score to PageRank and $p_{v,w}$ to 0.1.

The two columns of the figure represent the different contagion models (Independent Cascade and Linear Threshold), the three plots in each figure represent the different scheduling approaches (Initial, Step-by-Step and Incremental), the x-axis represent the network size, and the y-axis represent the number of activated nodes.

As can be seen from the figure, the number of activated nodes grows almost linearly with the network size, and the two scheduled approaches (Step-by-Step and Incremental) increase their gap from the initial approach for larger network sizes.

Infection Probability

Figure 4 presents a comparison of the three seeding heuristics for different infection probabilities, over five different network topologies, when fixing the seeding budget to 100, the seeding score to PageRank and the contagion model to Independent Cascade.

The five columns of the figure represent the different network topologies (Citations, Enron, Epinions, Slashdot, and Wiki-vote), the three plots in each figure represent the different scheduling approaches (Initial, Step-by-Step and Incremental), the x-axis represent the infection probability ($p_{v,w}$), and the y-axis represent the number of activated nodes.

As can be seen from the figure, the number of activated nodes grows sub linearly with $p_{v,w}$, and the gap between the scheduled approaches (Step-by-Step and Incremental) and the initial approach retains even for larger $p_{v,w}$ values.

4 Diminishing Social Effect

The limited attention phenomenon, namely the fact that an individual is capable to process only a relatively small amount of information in a given period of time [44], [55], was widely studied in the context of social networks. The fast-paced information stream that an average social-network user face daily, creates a scenario in which the opportunity of passing forward a recently acquired information by an individual to its peers has a limited potential time window [20]. We refer to this phenomenon as Diminishing Social Effect.

In fact, the diminishing social effect, is already supported by some of the existing contagion models. For example, in the SIR model, this phenomenon is represented by the $I \rightarrow$

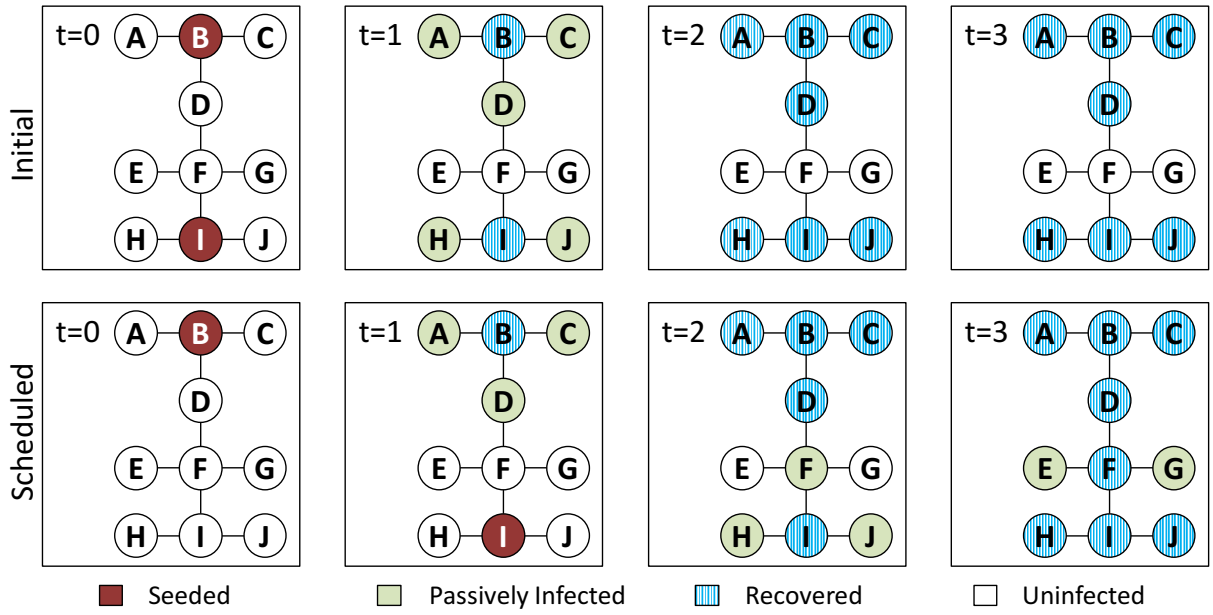

Fig. 6. Illustrating the advantage of scheduled seeding in a setting with a diminishing social effect.

R (Infectious $\rightarrow$ Recovered) transition. In the Independent Cascade model, this effect is reflected by the property that a node has a single opportunity only to activate its neighbors. A similar representation also exists in the General Cascade and Bass-SIR models.

In models that present a combination of both a diminishing social effect, and a complex-contagion effect (see Table 1), the timing in which nodes get infected plays a key role on the likelihood of activating their neighbors. Therefore, a wise selection of the nodes to be seeded at different time-steps may significantly improve the final number of activated nodes.

4.1 Diminishing Social Effect Modeling

In order to further understand the diminishing social effect, and analyze it separately from the stochasticity property, we propose a simple extension to the deterministic Tipping model which adds diminishing social effect, as we proceed to explain.

As we explained in subsection 2.3, Goyal et Al. [20] suggested that the infectiousness of a node can be modeled with either a discrete or a continuous time model. In the continuous time model, the length of the infectiousness period of a node follows an exponential distribution (i.e., short periods are extremely more likely than long periods). In contrast, in the discrete time model, the length of the infectiousness period of all nodes, is limited by some fixed upper-bound IT.

Considering the discrete time model, we suggest to extend the Tipping model, by modeling the accumulated social effect (i.e., number of infectious neighbors) exerted on a node v at time-step t as follows:

$$\sum_{w \in N(v)} \sum_{\tau=Max(0,t-IT)}^{t-1} Y_w(\tau) \tag{3}$$

Where $Y_w(\tau)$ is a binary flag, representing whether the node $w \in N(v)$ got activated exactly at time-step τ. In fact, a similar extension to the tipping model was already considered in [49], [50], [46].

As for the continuous time model, instead of determining the infectiousness time period, we suggest a variation in which the "infectiousness strength" of a node diminishes exponentially over time. Therefore, the accumulated social effect exerted on a node v at time-step t is modeled as follows:

$$\sum_{w \in N(v)} \sum_{\tau=0}^{t-1} \alpha^{t-\tau-1} \cdot Y_w(\tau) \tag{4}$$

Where $\alpha \in [0, 1]$ represents the exponential decay pace.

4.2 A Toy Example

In this subsection, we provide a toy example that illustrates the importance of a scheduled seeding approach in the extended Tipping model that we suggested in the previous subsection. Intuitively, such a scheduled approach would search for nodes to be seeded "Just-In-Time", so that their neighbors will reach their threshold activation boundaries.

Figure 6 compares between an initial seeding approach and a scheduled seeding approach, under the extended Tipping model.

We consider a small network of 10 nodes, where the threshold value is set to $\theta = 0.5$, meaning that each node would require at least half of its neighbors to be infectious in order to become infected. We also consider a discrete time model with an infectiousness period of $IT = 1$ (this is equivalent to a continuous model with $\alpha = 0$), and a seeding budget of $B = 2$.

Considering an initial seeding approach (upper row of the figure) in which the entire seeding budget must be utilized at $t = 0$, the maximum possible number of activated nodes at the end of the contagion process is seven. For

example, choosing to seed the nodes B and I at time-step $t = 0$ will trigger the activation of nodes $\{A, C, D, H, J\}$ at time-step $t = 1$ (they all require a single infectious neighbor to reach their threshold and become activated). All other nodes $\{E, F, G\}$ do not get infected at time-step $t = 1$ since they do not reach their threshold. At time-step $t = 2$, nodes $\{B, I\}$ stop being infectious, and at time-step $t = 3$, nodes $\{A, C, D, H, J\}$ stop being infectious, and therefore the final number of activated nodes remain seven.

Allowing to schedule the seeding actions (bottom row of the figure), we can reach a higher number of activated nodes. For example, seeding the node B at time-step $t = 0$, triggers the activation of nodes $\{A, C, D\}$ at time-step $t = 1$. Then, the node I is seeded at time-step $t = 1$, which triggers the activation of nodes $\{F, H, J\}$ at time-step $t = 2$. Finally, E and G will become activated at time-step $t = 3$, resulting in the maximal number of activated nodes (i.e., 10).

4.3 Theoretical Analysis

In the previous subsection, we saw an example in which a scheduled seeding approach obtained a higher number of activated nodes than an initial approach, under the extended Tipping model. However, as we claim next, this is a direct result of the diminishing social effect property.

Claim: Under the basic Tipping model ($\alpha = 1$), there exists an initial seeding solution that obtains the optimal number of activated nodes.

Proof:
Let S and T denote the set of seeded nodes and their corresponding time-steps in one of the optimal solutions, and let A be the final set of activated nodes at the end of the contagion process. As we will show next, an initial strategy that will seed all nodes in S at time-step $t = 0$, will result in a final set of activated nodes $A^* = A$.

Let $v \in S$ be a node that was seeded at time-step $t = T(v)$ in the optimal solution. Let's denote A_v as the set of all nodes that were activated at time-step $t > T(v)$ and would not be activated if v would not be seeded at time-step $t = T(v)$. Since there is no diminishing social effect (i.e., nodes are infectious forever), seeding v at time-step $t = 0$ instead of time-step $t = T(v)$, will still ensure that all nodes in A_v will be active at time-step $t > T(v)$ (or earlier).

By repeating the same process for all $v \in S$, we get an initial strategy that ensures that all nodes in A are activated by the end of the contagion process (i.e., $A^* \supseteq A$). Since A is an optimal solution, it necessarily means that $|A^*| = |A|$, and therefore $A^* = A$ is an optimal solution. ■

It is important to note, that since the Tipping model and its extended version are both deterministic, it is possible to find the optimal solution (scheduled or not) in advance (i.e., before the contagion process begins). However, finding such an optimal solution is NP-hard.

While several optimization methods and approximation heuristics were suggested for influence maximization under the basic Tipping model (see [51] for further details), the extended version, which requires to determine not only the network nodes to seed, but also when to seed each of them, was not investigated at all.

Next, we suggest to model the influence maximization problem under the extended Tipping model as the following Integer Linear Programming (ILP) optimization problem:

$$
\begin{aligned}
&\textbf{Maximize } \sum_{v \in V} \sum_{t=0}^{T} Y_{v,t} \\
&\textbf{subject to} \\
&1. \sum_{t=0}^{T} Y_{v,t} \leq 1 && \forall v \in V \\
&2. \sum_{v \in V} S_v \leq B \\
&3. Y_{v,t} \leq S_v + \frac{1}{\theta D(v)} \sum_{w \in N(v)} \sum_{\tau=0}^{t-1} \alpha^{t-\tau} Y_{w,\tau} && \forall v \in V, t = 1..T \\
&4. D(v) \sum_{k=0}^{t} Y_{v,k} \geq \sum_{w \in N(v)} \sum_{\tau=0}^{t-1} \alpha^{t-\tau} Y_{w,\tau} - \theta D(v) && \forall v \in V, t = 1..T \\
&5. Y_{v,0} \geq S_v + IsInit - 1 && \forall v \in V \\
&6. |V| \cdot \sum_{v \in V} Y_{v,t} \geq \sum_{\tau=t+1}^{T} \sum_{u \in V} Y_{u,\tau} && \forall t = 0..T \\
&S_v \in \{0,1\}; Y_{v,t} \in \{0,1\} && \forall v \in V, t = 1..T
\end{aligned}
\tag{5}
$$

Decision Variables:

- $Y_{v,t}$ - Whether node v got activated exactly at time-step t (Binary)
- S_v - Whether node v was seeded at any time-step (Binary)

Target Function:

$Max \sum_{v \in V} \sum_{t=0}^{T} Y_{v,t}$ - Maximize the amount of activations during the entire contagion process

Model Parameters:

- θ - threshold value (in the range $[0, 1]$)
- α - diminishing social effect factor (in the range $[0, 1]$)
- V - set of all network nodes
- $D(v)$ - degree of node v
- $N(v)$ - set of v's neighbors
- T - upper-bound on the length of contagion process
- B - number of seeding actions (budget)
- $IsInit$ - whether the entire seeding budget must be utilized at time-step $t = 0$ (Binary)

Model Constraints:

1) Every node can get activated once (at most)
2) The number of seeded nodes is bounded by the budget
3) A node can get activated only if seeded or if the social effect exerted on it passed the threshold
4) Threshold passing implies that the node becomes activated in the current or previous time-steps
5) When the IsInit flag is on, the entire seeding budget must be used at time-step $t = 0$
6) For simplicity reasons - "idle times" (i.e., time-steps with no activations) in the middle of the process are not allowed

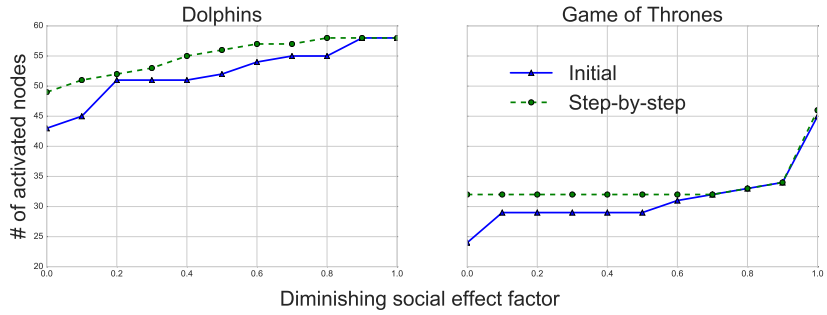

Fig. 7. The best obtained number of activated nodes as a function of the diminishing social effect factor.

As can be seen from the problem formulation, the set of possible solutions for the Initial approach, is a subset of the Scheduled approach (since it practically adds constraint number 5).

4.4 Evaluation

In the previous subsection we presented an algorithm that finds the optimal seeding solution (both initial and scheduled) under the extended Tipping model. Moreover, we showed that when the social effect does not diminish over time (i.e., $\alpha = 1$), the optimal solution for the initial approach obtains the same performance of the scheduled approach. Here, we aim at understanding better the gap between the optimal solutions of the initial and scheduled approaches for different magnitudes of the diminishing social effect ($\alpha \in [0, 1]$).

4.4.1 Experimental Setting

All of our optimization problems were preprocessed in Python 2.7, and executed using the IBM ILOG CPLEX engine (solution time was limited to 6,000 seconds per each instance). All executions were done on a Linux machine running Centos 7.1, with 128 GB of RAM and 32 Intel 2.7 GHz CPUs.

Our simulations compared the total number of activated nodes by the two optimal solutions (initial and scheduled) for various values of α, over two network topologies. The threshold value was set to $\theta = 0.5$ for all network nodes, and the seeding budget was set to $B = 5$.

In order to achieve a nearly-optimal solution in a reasonable amount of time, we had to use relatively small network topologies. For this purpose, we used the Game of Thrones network [42] and the Dolphins network [36] The properties of the two network topologies are detailed in Table 4.

TABLE 4
Properties of social networks

Network	Number of Nodes	Average Degree	Average Clustering
Dolphins	60	4.77	0.27
Game of Thrones	107	6.58	0.55

4.4.2 Results

Figure 7 presents a comparison of the two optimal solutions.

The two columns of the figure represent the different network topologies (Game of Thrones and Dolphins), the two plots in each figure represent the different seeding approaches (Initial and scheduled), the x-axis represent the diminishing social effect factor (α), and the y-axis represent the best obtained number of activated nodes.

As can be seen in the figure, the gap in performance between the two approaches increases for smaller values of α, reaching 15%-33% improvement for $\alpha = 0$. This finding

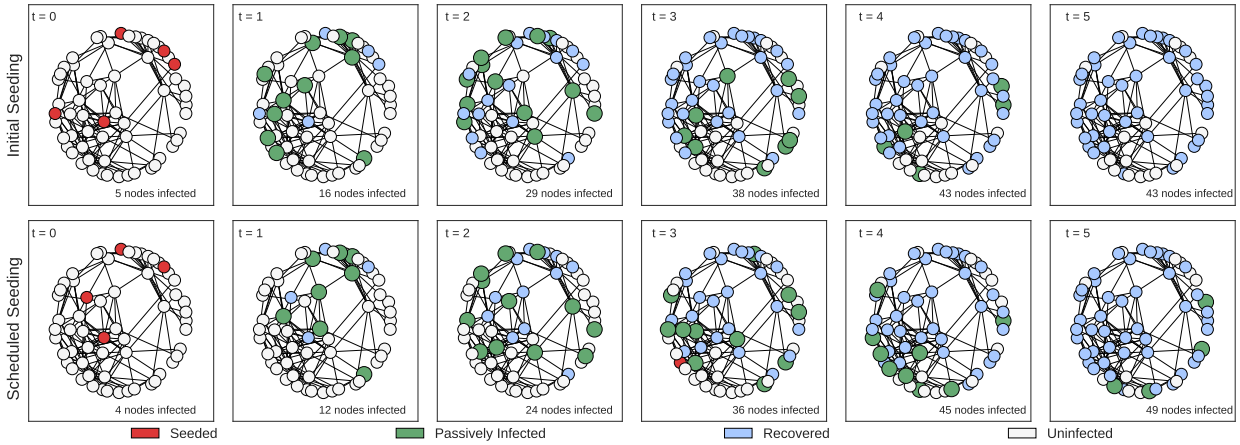

Fig. 8. The dynamics of the diffusion process for the two approaches, on the Dolphins network topology with $\alpha = 0$

is also consistent with our analytical observation in the subsection above - when the social effect does not diminish at all (i.e., $\alpha = 1$), both approaches obtain the same result.

To further illustrate the strategies (and obtained results) of the two approaches, Figure 8 demonstrates the dynamics of the diffusion process on the Dolphins network topology with $\alpha = 0$ and $B = 5$.

The two rows of the figure represent the two different seeding approaches (Initial and scheduled), the x-axis represent the time-step and the different color of nodes represent their states (Seeded, Passively Infected, Recovered, Uninfected).

When taking the initial seeding approach, the optimal solution that seeds 5 nodes at time-step $t = 0$ led to 11 new infections at time-step $t = 1$, and a total of 43 infections by time-step $t = 4$ (the last time-step of the diffusion process). The optimal solution of the Scheduled strategy, chose to seed four nodes at time-step $t = 0$, leading to 8 new infections at time-step $t = 1$ and 24 new infections at time-steps $t = 2, 3$. Then, at $t = 3$, the last node was seeded, contributing to a joint social effect together with the other passively infected nodes, leading to 9 new infections at time-step $t = 4$, and reaching a total of 49 infections at time-step $t = 5$ (the last time-step of the diffusion process).

5 State-Dependent Seeding

In the previous sections, we considered models in which the social influence on a network node v by its neighboring nodes can lead to v's activations. These dynamics describe a passive contagion, in which network nodes can become activated without an external intervention.

However, in many real-world cases, the option that a node gets activated depends solely on external active intervention (i.e., seeding actions). Moreover, such seeding actions may sometimes fail, in contrast to classical models such as Linear Threshold and Independent Cascade, where the success of a seeding action is guaranteed. In such cases, the success of a seeding attempt often depends on the states of neighboring nodes, leading to a "State-Dependent Seeding" setting. Similar settings were described in [13] and [35].

Such a setting provides an inherent advantage to the scheduled seeding approach, due to the accumulation of the social effect over the time-steps of the diffusion process.

5.1 State-Dependent Seeding Modeling

In order to isolate the effect of the state-dependent seeding, we suggest the following simplified model. Given a social network and a set of previously activated nodes, a non-activated node v will get activated at time-step t if and only if it was seeded at time-step t and the fraction of v's previously infected neighbors passes the threshold value θ_v, namely:

$$\sum_{w \in N(v)} X_w(t-1) \geq \theta_v \cdot |N(v)| \tag{6}$$

It is important to note that a node in this setting can get activated only if it was actively seeded, and therefore the total number of activated nodes is at most the number of seeding attempts B.

5.2 A Toy Example

The following example demonstrates the advantage of the scheduled seeding approach under the State-Dependent Seeding model. Figure 9 depicts a network of five nodes at time-step $t = 0$, where two of the network nodes, $\{C, E\}$, were already previously activated. The rest of the nodes, $\{A, B, D\}$ are inactive at time-step $t = 0$, and may get activated at later time-steps only if actively seeded by the marketer and at least half ($\theta = 0.5$) of their neighbors are already active.

Assuming a seeding budget of $B = 3$, the initial seeding approach (upper row of the figure) will try to seed all of the inactive nodes, $\{A, B, D\}$, at the next time-step $t = 1$. In this case, only one node, D, will get activated, since two out of four neighbors it has ($\{C, E\}$) are already active, resulting

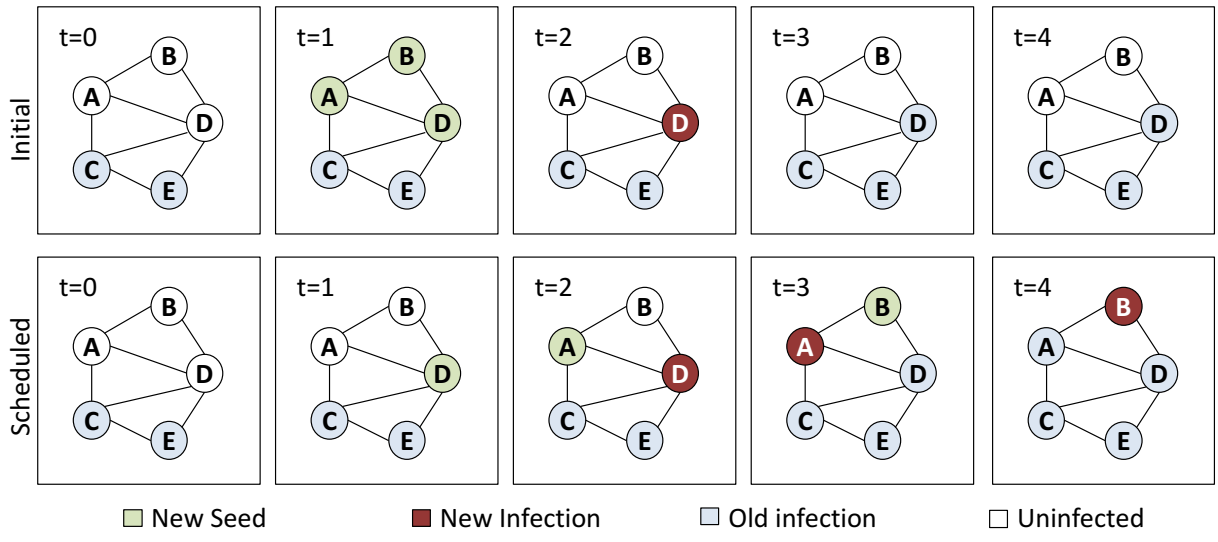

Fig. 9. Illustrating the advantage of scheduled seeding in a setting with state-dependent seeding.

in a total of three activated nodes at the end of the contagion process.

In contrast, using the scheduled seeding approach (bottom row of the figure), nodes can be seeded one-by-one allowing the marketer to utilize the accumulated social effect over time. For example, seeding the node D at time-step $t = 1$ will get it activated. Then, seeding the nodes A and B at time-steps $t = 2$ and $t = 3$ will get each of them activated, since they pass their threshold values, resulting in a total number of five activated nodes at the end of the contagion process.

5.3 Theoretical Analysis

In this subsection, we present a greedy algorithm for selecting the nodes to be seeded under the State-Dependent Seeding model, and prove its optimality for both the initial and scheduled variations.

Algorithm 2 presents a single greedy iteration of selecting nodes to be seeded. It returns all the nodes that can potentially become active if seeded, subjected to the seeding budget limits.

Algorithm 2 Greedy Seed Selection Algorithm

Input:
$\vec{X}(t-1)$ - States of nodes at time-step $t-1$
B - Remaining seeding budget
Output: S - Set of nodes to be seeded at time-step t
1: $Q \leftarrow \{v | X_v(t-1) = 0 \wedge \sum_{w \in N(v)} X_w(t-1) \geq \theta_v \cdot N(v)\}$
2: $N \leftarrow min(B, |Q|)$
3: $S \leftarrow$ a subset of N random nodes from Q
4: **Return** S

The suggested algorithm can be used by both an initial seeding approach and a scheduled seeding approach. More specifically, given a seeding budget of size B, the initial version executes the algorithm once to retrieve a set of nodes, all of which will be seeded at time $t = 0$ (The rest of the budget, is not useful at time-step $t = 0$ and therefore is not used). In the scheduled version, the algorithm is executed once to retrieve a set of nodes, all of which will be seeded at time $t = 0$. Then, the algorithm is executed repeatedly at each time-step to retrieve a set of nodes to be seeded at that time-step, until the budget is fully utilized or no more new activations are possible.

In terms of runtime complexity, the algorithm evaluates the activation potential of all nodes, where the potential of each node is evaluated by iterating over all of its' neighbors. This is equivalent to iterating over all edges, and therefore, the runtime complexity of Algorithm 2 is $O(|E|)$.

The following claim holds with respect to the greedy algorithm:

Claim: The greedy algorithm G described above, ensures the optimal number of activated nodes at the end of the contagion process.

Proof: As described above, G seeds all the possible nodes for activation at each time-step, subjected to the budget limit B.

If the total number of activated nodes by G equals to the budget B, then G necessarily achieved the best possible number of activations.

Otherwise, denoting the set of activated by a seeding algorithm A by time-step t as $R_A(t)$, we will show by mathematical induction, that for any seeding algorithm A, and for any time-step t, $R_A(t) \subseteq R_G(t)$.

As for the base case, at $t = 0$, G selects all possible nodes that can get activated and therefore any node $v \in R_A(0)$ necessarily belongs to $R_G(0)$, resulting in $R_A(0) \subseteq R_G(0)$.

For the inductive step, suppose that $R_A(t-1) \subseteq R_G(t-1)$, every node v that is activated by algorithm A at time step t is also available for activation by G, since the set of activated neighbors of v in $R_A(t-1)$ are necessarily also in $R_G(t-1)$ (by applying the inductive assumption), and therefore will necessarily be seeded by G.

Since both the basis and the inductive steps have been proven, by mathematical induction, the statement $R_A(t) \subseteq R_G(t)$ holds for any time-step t, and therefore G is optimal. ■

5.4 Evaluation

In the previous subsection we showed that under State-Dependent Seeding settings the scheduled approach outperforms the initial approach in terms of the final number of infected nodes. In this subsection, we report the results of simulations that we conducted to understand the extent of such improvement on real-world network topologies.

5.4.1 *Experimental Setting*

All of our simulations 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.

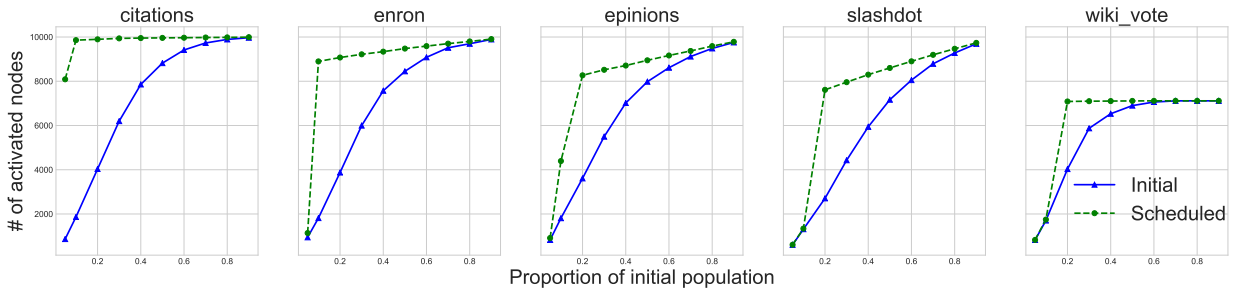

Fig. 10. Number of activated nodes as a function of the proportion of initially infected population.

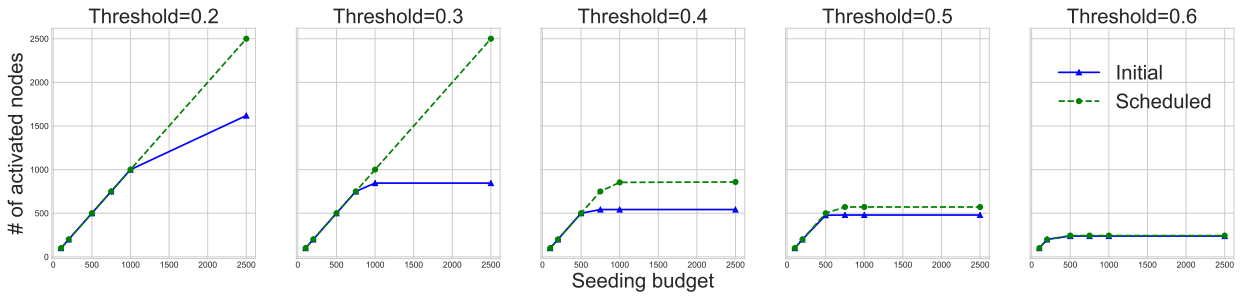

Fig. 11. Number of activated nodes as a function of the seeding budget.

Each simulation started with a random selection of an initially infected population of a given size F. In order to select the nodes to seed and their time-steps, we used the greedy algorithm with the Initial and Scheduled seeding heuristics mentioned above. Under the Initial strategy, all of the possible (in terms of successful activation) seeds were performed at $t = 0$. Under the Scheduled strategy, each time-step we executed the greedy algorithm once again until there were no more new activations.

As the underlying network topologies for our simulations, we used the same network topologies as in subsection 3.3.1 (see Table 2).

Our simulations compared the total number of activated nodes and runtime obtained by the two seeding strategies for various parameters: proportion of initially infected population, network topology, seeding budget, threshold and network size. The reported results are the averages of 1000 simulations for each set of parameters.

5.4.2 Results

Proportion of Initially Infected Population

Figure 10 presents a comparison of the two seeding heuristics for different proportions of initially infected populations, over five different network topologies, when fixing the seeding budget to "unlimited" (the end of the contagion process is defined by the availability of activation of further potential nodes) and θ to 0.3.

The five columns of the figure represent the different network topologies (Citations, Enron, Epinions, Slashdot, and Wiki-vote), the two plots in each figure represent the different scheduling approaches (Initial and Scheduled), the x-axis represents the proportion of initially infected population F, and the y-axis represents the number of activated nodes.

The figure presents several interesting trends: First, for intermediate values of F, the scheduled approach performs significantly better (up to 150% improvement) than the initial approach. Second, in the vast majority of cases, for extreme values of F (i.e., 0.1 and 0.9) the plots of the initial and scheduled approach merge. These extreme cases represent either a very low likelihood for a successful seeding action, or an extremely high one (due to the number of already active nodes in the network). Third, the initial and scheduled plots form an S-shape curve, which resembles that of the SIR and Bass models. The main reason for that is that there exists a critical-mass point, after which the activation rate climbs rapidly.

Seeding Budget

Figure 11 presents a comparison of the two seeding heuristics for different seeding budgets, over five different threshold values, when fixing the network topology to Citations and F to 1000.

The five columns of the figure represent the different threshold values ($\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$), the two plots in each figure represent the different scheduling approaches (Initial and Scheduled), the x-axis represents the seeding budget, and the y-axis represents the number of activated nodes.

As expected, the number of activated nodes grows with the budget, until a certain point where no more activations are possible. We also observe that this point arrives earlier for higher threshold values since higher threshold values reduce the number of options for a successful seeding attempt. Moreover, this point arrives later in the case of the Scheduled approach, since previous successful seeding

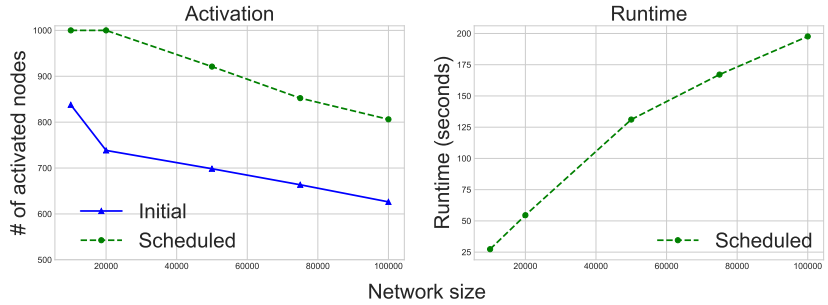

Fig. 12. Number of activated nodes and runtime as a function of network size.

attempts increase the number of options for a successful seeding attempt in the future.

Network Size

Figure 12 presents a comparison of the two seeding heuristics for different network sizes, when fixing the network topology to Citations, F to 1000, the seeding budget to 1000 and θ to 0.3. The two columns of the figure represent two different measures (number of activated nodes and runtime), the two plots in each figure represent the different scheduling approaches (Initial and Scheduled), the x-axis represents the network size, and the y-axis represents the measure (number of activated nodes or runtime).

As can be seen from the figure on the left, the number of activated nodes decreases roughly linearly with the network size. The reason for that is that when the size of the network grows, the 1000 initially infected nodes are more spread, and therefore it becomes more difficult to find nodes that can be seeded successfully (i.e., nodes that have enough activated neighbors). Nevertheless, we see that the scheduled approach retains its superiority over the initial approach even in such cases.

From the right side of the figure, we observe that the execution time for the scheduled approach grows sub-linearly with the network size, making it a viable option for real-world applications.

6 Summary and Conclusions

This work suggested and analyzed a scheduled seeding approach for influence maximization. In particular, we identified three properties of existing contagion models, namely stochastic dynamics, diminishing social effect and state-dependent seeding, that are beneficial to the scheduled seeding approach. For each of these properties, we considered a simplified contagion model, allowing us to isolate the contribution of the particular property to the success of the scheduled seeding approach. We then presented a minimalistic toy-example that demonstrated the benefits gained by the scheduled seeding approach. We also conducted a theoretical analysis to compare the performance of the scheduled seeding approach and the traditional initial seeding approach. Finally, we conducted a set of empirical simulative experiments, which compared the performance of the scheduled and initial seeding approaches over numerous real-world social networks topologies. The results of the empirical evaluation presented an improvement of 10%-70% in the final number of activated nodes, when using the scheduled seeding approach, which was consistent over a wide variety of parameters' values.

Throughout this work, we tried to keep the analyses as simple as possible. For example, when presenting the three properties that are beneficial for the scheduled seeding approach, we isolated each property from the others. Future work may further analyze different combinations of these properties, which may present non-trivial interaction between them. For example, while the existence of stochasticity was shown as a strong argument in favor of delaying seeding actions, the presence of a diminishing social effect may require to seed the nodes as soon as possible. A preliminary exploration of such a case can be found in [47], [48], where the dynamics of state-dependent seeding are analyzed together with a presence of stochasticity and diminishing social effect.

The main goal of our analyses was to show that under certain circumstances, the scheduled seeding approach can outperform the initial seeding approach. While the results of our analysis provided a sense of scale for the improvement obtained by the scheduled seeding approach, further theoretical analysis may try to provide some guarantees on the bounds of such improvement.

Our evaluation was mainly based on theoretical analysis and simulation that utilized real-world network topologies. In the future, it would be interesting to enrich our 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 a scheduled seeding approach and an initial seeding approach.

To conclude, we believe that our findings have the potential to open up an entirely new area of research, focusing on finding the right timing for seeding actions, thereby helping both in improving our understanding of diffusion dynamics and in devising better strategies to shape social behavior.

Acknowledgements

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

References

[1] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[2] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[3] Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men. S*, pages 222–236, 1951.

[4] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

[5] JM Berger and Heather Perez. Occasional paper the islamic states diminishing returns on twitter: How suspensions are limiting the social networks of english-speaking isis supporters. 2016.

[6] Phillip Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564, 2007.

[7] Stephen P Borgatti. Centrality and network flow. *Social networks*, 27(1):55–71, 2005.

[8] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[9] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[10] Wei Chen, Chi Wang, and Yajun Wang. 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*, pages 1029–1038. ACM, 2010.

[11] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[12] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 88–97. IEEE, 2010.

[13] Flavio Chierichetti, Jon Kleinberg, and Alessandro Panconesi. How to schedule a cascade in an arbitrary graph. *SIAM Journal on Computing*, 43(6):1906–1920, 2014.

[14] Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown, 2009.

[15] Gadi Fibich. Bass-sir model for diffusion of new products in social networks. *Physical Review E*, 94(3):032305, 2016.

[16] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. ACM, 2012.

[17] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[18] Jacob Goldenberg, Barak Libai, and Eitan Muller. 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, 2001.

[19] Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

[20] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[21] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011.

[22] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[23] Kerric Harvey. *Encyclopedia of social media and politics*. Sage Publications, 2013.

[24] Oliver Hinz, Bernd Skiera, Christian Barrot, and Jan U Becker. Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6):55–71, 2011.

[25] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. Opening closed regimes: what was the role of social media during the arab spring? 2011.

[26] Jarosław Jankowski, Piotr Bródka, Przemysław Kazienko, Boleslaw K Szymanski, Radosław Michalski, and Tomasz Kajdanowicz. Balancing speed and coverage by sequential seeding in complex networks. *Scientific reports*, 7(1):891, 2017.

[27] Jarosław Jankowski, Piotr Bródka, Radosław Michalski, and Przemysław Kazienko. Seeds buffering for information spreading processes. In *International Conference on Social Informatics*, pages 628–641. Springer, 2017.

[28] Kyomin Jung, Wooram Heo, and Wei Chen. Irie: Scalable and robust influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 918–923. IEEE, 2012.

[29] Elihu Katz and Paul Lazarsfeld. Personal influence, 1955.

[30] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[31] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[32] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[33] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[34] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 551–556. SIAM, 2007.

[35] Shuyang Lin, Qingbo Hu, Fengjiao Wang, and S Yu Philip. Steering information diffusion dynamically against user attention limitation. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 330–339. IEEE, 2014.

[36] David Lusseau. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(Suppl 2):S186–S188, 2003.

[37] Vijay Mahajan, Eitan Muller, and Frank M Bass. New product diffusion models in marketing: A review and directions for research. In *Diffusion of technologies and social behavior*, pages 125–177. Springer, 1991.

[38] Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

[39] Mark Newman. Networks: an introduction. 2010. *United Slates: Oxford University Press Inc., New York*, pages 1–2.

[40] Yaodong Ni. Sequential seeding to optimize influence diffusion in a social network. *Applied Soft Computing*, 56:730–737, 2017.

[41] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[42] Andrew Pitts. Polinode: A web application for the collection and analysis of network data. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 1422–1425. IEEE, 2016.

[43] Rami Puzis, Yuval Elovici, and Shlomi Dolev. Fast algorithm for successive computation of group betweenness centrality. *Physical Review E*, 76(5):056709, 2007.

[44] Xiaoyan Qiu, Diego F.M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. Limited individual attention and online virality of low-quality information. *Vaccine*, 32(33), 2017.

[45] Lior Seeman and Yaron Singer. Adaptive seeding in social networks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 459–468. IEEE, 2013.

[46] Alon Sela, Irad Ben-Gal, Alex Sandy Pentland, and Erez Shmueli. Improving information spread through a scheduled seeding approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, pages 629–632. IEEE, 2015.

[47] Alon Sela, Dmitri Goldenberg, Irad Ben-Gal, and Erez Shmueli. Latent viral marketing, concepts and control methods. *arXiv preprint arXiv:1704.01775*, 2017.

[48] Alon Sela, Dmitri Goldenberg, Irad Ben-Gal, and Erez Shmueli. Active viral marketing: Incorporating continuous active seeding efforts into the diffusion model. *Expert Systems with Applications*, 107:45–60, 2018.

[49] Alon Sela, Dmitri Goldenberg, Erez Shmueli, and Irad Ben-Gal. Scheduled seeding for latent viral marketing. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 642–643. IEEE, 2016.

[50] Alon Sela, Erez Shmueli, Dima Goldenberg, and Irad Ben-Gal. Why spending more might get you less, dynamic selection of influencers in social networks. In *Science of Electrical Engineering (ICSEE), IEEE International Conference on the*, pages 1–4. IEEE, 2016.

[51] Paulo Shakarian, Abhinav Bhatnagar, Ashkan Aleali, Elham Shaabani, and Ruocheng Guo. *Diffusion in social networks*. Springer, 2015.

[52] Paulo Shakarian, Sean Eyre, and Damon Paulo. A scalable heuristic for viral marketing under the tipping model. *Social Network Analysis and Mining*, 3(4):1225–1248, 2013.

[53] Guangmo Tong, Weili Wu, Shaojie Tang, and Ding-Zhu Du. Adaptive influence maximization in dynamic social networks. *IEEE/ACM Transactions on Networking (TON)*, 25(1):112–125, 2017.

[54] Thomas W Valente. Network interventions. *Science*, 337(6090):49–53, 2012.

[55] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Scientific reports*, 2, 2012.

[56] Darrell M West. *Air wars: Television advertising and social media in election campaigns, 1952-2012*. Sage, 2013.

[57] Philip Zimbardo. The lucifer effect-understanding how good people turn evil, 2007.