Accepted Manuscript

Seed Activation Scheduling for Influence Maximization in Social Networks

Mohammadreza Samadi, Rakesh Nagi, Alexander Semenov, Alexander Nikolaev

PII: S0305-0483(16)30357-7
DOI: 10.1016/j.omega.2017.06.002
Reference: OME 1786

To appear in: *Omega*

Received date: 30 June 2016
Revised date: 24 April 2017
Accepted date: 6 June 2017

Please cite this article as: Mohammadreza Samadi, Rakesh Nagi, Alexander Semenov, Alexander Nikolaev, Seed Activation Scheduling for Influence Maximization in Social Networks, *Omega* (2017), doi: 10.1016/j.omega.2017.06.002

Highlights

- Presented a new problem: influential Seed Activation Scheduling over a campaign horizon
- Blogger-centric marketing application is studied on two-level social networks
- Demonstrated the benefit of delayed seeds activation, even with unlimited budget
- Optimal policies investigated in the presence and absence of competition
- Column generation heuristic provides near-optimal solutions to large problems

Seed Activation Scheduling for Influence Maximization in Social Networks

Mohammadreza Samadi[a], Rakesh Nagi[b], Alexander Semenov[c], Alexander Nikolaev[d]

[a] Department of Operations Research and Advanced Analytics, American Airlines, Fort Worth, TX 76155

[b] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, IL 61801

[c] Department of Computer Science and Information Systems, University of Jyvaskyla, Finland

[d] Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY 14260

mohammadreza.samadi@aa.com, nagi@illinois.edu, alexander.v.semenov@jyu.fi, anikolae@buffalo.edu

June 20, 2017

Abstract

This paper addresses the challenge of strategically maximizing the influence spread in a social network, by exploiting cascade propagators termed "seeds". It introduces the Seed Activation Scheduling Problem (SASP) that chooses the timing of seed activation under a given budget, over a given time horizon, in the presence/absence of competition. The SASP is framed as a blogger-centric marketing problem on a two-level network, where the decisions are made to buy sponsored posts from prominent bloggers at calculated points in time. A Bayesian evidence diffusion model – the Partial Parallel Cascade (PPC) model – allows the network nodes to be partially activated, proportional to their accumulated evidence levels. The SASP under the PPC model is proven NP-hard. A mixed-integer program is presented for the SASP, along with an efficient column generation heuristic. The paper sets up its problem instances in real-world settings, taking web-based marketing as an application example. Favorable optimality gaps are achieved for SASP solutions on networks based on observed user interactions in pro-health discussion forums. The presented analyses highlight a trade-off between early and late seed activation in igniting and maintaining influence cascades over time. The results reveal the importance of early seeds for campaigns that favor longevity, e.g., in service industry, and the importance of late seeds for campaigns with deadline(s), e.g., in political competitions.

Keywords: social networks, scheduling, influence maximization, seed selection, marketing

1 Introduction and Motivation

Diffusion-driven information transfer is the key driver behind knowledge accumulation and opinion formation resulting from the communication between individuals in social networks [30, 45]. The viral spread of news, goods and judgments via such network-based information transfer is often referred to as "spread of influence" [4, 21]. Influence cascades are typically ignited through injection of new information, e.g., about new technologies or major social/political events, and grow through word-of-mouth (WOM). Coordinating the emergence and penetration depth of influence cascades is of much interest to practitioners, e.g., to the companies introducing new products or political parties running campaigns [61]. The problem of strategically exploiting the WOM effect in a given connected population motivates the Influence Maximization (IM) problem, where a set of early starters, termed seeds, is selected to start a diffusion-based cascade to achieve maximum spread [33]. There are two widely known ways of implementing the IM solutions in practice. The first one lies in offering the selected seeds free/discounted products to increase the probability that these seeds "adopt" the products/opinions (offered/supported by the decision-maker), in anticipation that their neighbors/peers would follow suit [60]. Mailing discount coupons to a pre-selected group of customers is an example of such a seed activation process [63]. The second one has a decision-maker paying the seeds for spreading a supporting sentiment about the products/opinions over time [36]. Hiring influential bloggers to post sponsored articles is an example of this practice. Platforms such as `PayPerPost.com` support it by providing online services for buyers to find the sponsored bloggers matching the buyers' needs.

Kempe *et al.,* [33] proposed the first discrete optimization problem for IM; they developed two time-independent stochastic diffusion models – the Independent Cascade (IC) model and the Linear Threshold (LT) model. The original IC and LT diffusion models are time-independent in that (1) influence is assumed to never diminish over time, and hence, (2) the final outcome of the diffusion process is only affected by the selection of seeds, but not the timing of seed activation. Kempe *et al.,* [33] pointed to the submodularity of the problem's objective function under the IC and LT models, and presented a greedy algorithm for IM seed selection: the algorithm features a theoretical bound on the problem optima, however, its practical application proved to be computationally challenging [10, 26].

Much research has emerged following up on the seminal work on IC and LT models. However, one question in the IM domain has remained unanswered thus far: *"Is it always most beneficial to activate the seeds all at once, and if not, then what modeling approach(es) can allow for optimizing the seed activation timing?"*. The IM efforts relying on the time-independent models, including the IC and LT models, cannot explicitly account for the dynamics of the information diffusion. Thus, any conclusions obtained using such models are valuable only if it can be assumed that the influence spread is instant or that its effect lingers forever and it is only this final (eventual) effect that matters. However, in many real-world situations, decision-makers plan for long campaigns that need not only be initiated but controlled over time, and whose maximal impact should be achieved at specific time point(s), i.e., meeting pre-determined deadlines.

Chierichetti *et al.,* [12] revised the Original IM Problem (OIMP) formulation to arrive at a cascade scheduling problem where the adoption status of a node is a function of two quantities: (1) the fraction of the current adopters in the node's neighborhood, and (2) the fraction of the current adopters in the whole network. This work has provided a basis for an emerging cascade scheduling literature [6, 27, 42]. Hajiaghayi *et al.,* [28] considered competition in cascade scheduling, under the assumption that the network nodes are market segments sharing certain similarities, and developed models for this problem's variations with and without recourse (when the state of the network may be observable or unobservable between sequential seeding actions). Still, all the above-mentioned works bypass the modeling of time, instead focusing only on the order, in which the seeds are exposed to a new opinion/product. Hence, the relaxation of the OIMP's requirement to select and activate all the seeds *prior to* the ensuing cascade formation opens a previously unexplored field of study. In addition to selecting the *early starters* ("initial seeds") for IM, the Seed Activation Scheduling Problem (SASP) presented in this paper allows one to select "late seeds", i.e., activate them in a calculated way during the diffusion process. An SASP solution informs a strategy that can be implemented by a decision-maker who seeks to gradually spend resources to maintain the activation of seeds over a given time window, e.g., to ensure the continued support of a marketing campaign by sponsored bloggers. In such a setting, the duration of activation and re-activation of seeds becomes important.

There are multiple practical considerations motivating the development of SASP. First, the resources required to activate seeds (campaign budget, free products, man-power, etc.) may not all be available to a decision-maker at once, early on. Second, any seed activation strategy becomes time-dependent when the Net Present Value (NPV) of money is taken into account, with the late seed additions being less costly. Third, the activation of late seeds may prevent the exposed population from forgetting the information they receive [41, 53–55], requiring regular "reminders" to keep a campaign going.

This paper relies on the latest findings in the experimental marketing literature to solve a practical two-level problem of initiating and controlling social cascades over blogger networks, such as Twitter, or online forums. The sponsored bloggers available to be hired (as ad-hoc paid-per-post marketing agents), are assumed to form the first level of a network under study, with the other users and lower-profile bloggers forming its second level. A decision-maker is tasked with activating and de-activating the first level nodes over time, with the goal to maximize the effect of the campaign on the second level nodes, under a given resource/budget constraint. The adopted influence spread model tracks the information exchanges between all the nodes, recording the levels of the evidence that the nodes accumulate in support or against the claim of interest, e.g., that the marketed product is worth purchasing. In what follows, the time-dependent Bayesian evidence diffusion model with partial activation (positive and negative) of nodes is detailed and a mixed-integer program is presented for the resulting SASP that maximizes the spread of positive evidence and minimizes the spread of negative evidence through the second level network. The computational investigations with two case studies provide insights on how the seed selection strategies depend on the time horizon and the problem's objective function form.

The rest of this paper is organized as follows. Section 2 reviews the IM literature. Section 3 presents the seed activation scheduling idea and the partial parallel cascade model, and then, develops a mixed-integer program for the SASP and investigates two case studies. An efficient heuristic solution methodology for SASP using column generation is presented in Section 4. Section 5 provides the computational results for the SASP instances formulated for several real-world social networks. Section 6 concludes the paper and suggests the directions for future research.

2 Literature Review

Social influence, the phenomenon where the revealed/observed judgments of people make their peers adjust their judgments, has received much attention in marketing [3, 61], health care [13, 56], and political science [2, 29]. Prescriptive social influence research works towards calculated campaign planning in support of an opinion, behavior or product of interest. The earliest methods for finding influential nodes in social networks were based on centrality calculations. One limitation of the centrality-based heuristics is that they only exploit the information about the network positions of the nodes, while ignoring the information about the individual differences in the node characteristics [52]. Recent developments in centrality-based heuristics look to incorporate the data of such differences and of overlapping communities into the centrality scores [43, 47].

The idea of modeling the social influence mechanisms to formulate and solve IM problems was first introduced by Domingos and Richardson [15, 50]. Kempe *et al.,* [33] posed a discrete IM problem, introducing two time-independent diffusion models (IC and LT diffusion models) to describe how influence may spread in a social network from a set of seeds. These models have long served as a basis for the algorithmic developments [9, 10]. However, the basic IC and LT diffusion models do not incorporate time in describing influence propagation. A stream of literature on algorithmic seed selection has focused on addressing this issue over the last few years. Goyal *et al.,* [23] were among the first who addressed the need for time-aware diffusion models. Ever since, the IM problems where a certain objective is to be achieved by some preset deadlines has received much attention [8, 17, 24, 39].

The aspect of seed selection timing in IM, and the notion of cascade scheduling where the seeds fueling a cascade are added iteratively, have been considered most recently [6, 12, 28, 55]. It is important to distinguish the idea of late seed activation planning, presented in this paper, from the work on adaptive seed selection [22]. To be more specific, in the latter efforts, seed selection is viewed as a multi-stage decision-making process where the decision-maker observes the results (reward) of earlier actions prior to making decision of the subsequent actions. Seeman and Singer [57] presented a two-stage IM problem: in the first stage, the decision-maker spends a portion of the budget in seed activation to make nodes “accessible”, and in the second stage, spends the rest of budget for activating some of those accessible nodes. The adaptive submodularity property of the multi-stage IM problem helps designing efficient greedy algorithms to find the near optimal solutions with guaranteed bounds [11]. Note, however, that in a typical real-world setting, one can only observe the actions performed by an exposed population (e.g., purchases), as opposed to the spread of judgments in it [53, 54].

The original diffusion models for IM were designed for single campaign problems, with the activation status of each node represented by a binary variable. Such models interpreted each node activation as an event of a completed purchase or technology adoption: it was assumed that the once activated nodes never lose their activation [33]. The later studies of competition – between multiple parties – in IM relaxed this assumption and introduced the concept of activation loss/reversal [5]. The experimental marketing literature, meanwhile, models the process of buying decision formation as a gradual transition through multiple states, between the completely inactive one to the fully active one [1, 46, 62]. The current paper introduces a new evidence-based diffusion model that permits such partial activation, which is also particularly suitable for describing opinion adoption: the adoption/influence level of a node ranges from zero (inactive state) to one (fully activated state). Additionally, from a marketing IM perspective, the problem formulations and methods developed in this paper allow the decision-maker to plan not only for increasing the immediately realized sales, but also, for increasing the overall awareness about the product in the population, potentially leading to an indirect (delayed) profit.

In summary, compared to the presented literature, this paper introduces the concept of seed activation scheduling and employs the exact solution methodology using mathematical programming to explore and exploit the effects of seed activation timing on the IM strategies. Moreover, the idea of modeling partial activations in a two-level blogger network serves as a generalization of the previously existing diffusion models for marketing, healthcare and political applications to new settings.

3 Influence Propagation Models for Seed Activation Scheduling

The SASP is designed to schedule the seed selection/activation over the time window of a given IM problem. Section 3.1 describes an SASP formulation for campaign planning in the blogger-centric marketing setting, and explains the practical appeal of seed activation scheduling from a theoretical perspective. A two-level IM problem is developed, where the decision-maker pays sponsored bloggers (seeds), by time period, to maintain the influence they deliver over time. Section 3.2 presents an evidence-based diffusion model with partial activation of nodes, called Partial Parallel Cascade (PPC) model. In Section 3.3, the NP-hardness of the IM problem under the PPC diffusion model is proved, and a mixed-integer program is designed to solve the NP-hard SASP under the PPC diffusion model, optimally. Finally, two case studies are presented to showcase the utility of the SASP and to computationally explore the impact of the objective function variations on the optimal seed activation strategy.

3.1 The Seed Activation Scheduling Problem with Partial Activation

This section presents the SASP for planning a blogger-centric social campaign, where sponsored bloggers, whose posts get exposed to many readers in an online social network, are paid to publish the posts supporting a marketing or ideological campaign. The idea of blogger-centric advertising is well-discussed in marketing and experimentally proven to be effective for increasing product awareness [18, 40]. However, when a group of sponsored bloggers with overlapping/separate areas of influence are available and each blogger has its own cost for providing sponsored posting, selecting the bloggers and scheduling their actions over the time horizon of a budget-constrained campaign can be challenging.

The conventional IM problem formulation views all the nodes in a network as potential seeds, and assumes that no node can reject a seed nomination. However, in a real-world setting, a blogger cannot be expected to blindly accept one's proposition to serve as a seed. Also, decision-makers typically consider only a few most prominent bloggers as potential seeds; the network positions and characteristics of blogger nodes inform this preliminary selection. Indeed, on the sponsored blogger service platforms, such as PayPerPost.com, a segmentation system allows one to filter bloggers based on their profiles, taking into account their Google page rank and Alexa rank: only some bloggers qualify as good candidates to be employed for a given campaign [36]. By separating the potential seeds (influencers) from the regular nodes (influencees) in SASP, one reduces the solution space of the seed selection problem.

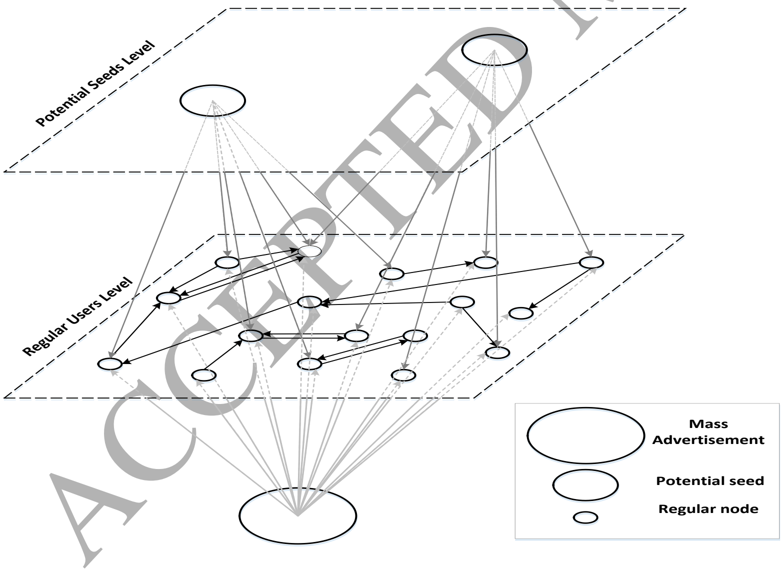

Figure 1: Modeling evidence diffusion in a two-level network: the first level contains the potential seeds (influential nodes) and the second level contains the regular nodes.

Figure 1 shows a network separated into two levels. The first level contains the nodes that can be hired to support a campaign – the activation status of these nodes does not figure in the objective function of the SASP. The second level contains the nodes that are the targets of the campaign. A cascade is to be ignited at the first level, consequently spreading to the second level, e.g., through (re)sharing/(re)tweeting. No arcs are assumed to go *from* the second level *to* the first: the nodes in the first level are assumed not to be influenced by their audiences – they spend their time writing, on a paid basis. The model does not count on the voluntary activation of potential seeds, without being paid by the decision-maker, as the first level nodes are assumed to represent marketing agents/accounts, not personal blogs. The assumption that the first-level nodes are sponsored bloggers also reduces the chance of the rejection of a seed nomination – such nodes are assumed to have already made themselves available as potential seeds, albeit at a certain price. Finally, the source of negative influence in the SASP is assumed to access and affect all the nodes uniformly, as mass media (e.g, television) advertisements do. The objective of the SASP is to schedule seed activations (in the first level) over the time horizon of the problem to maximize the spread of positive evidence and minimize the spread of negative evidence among the second level nodes, under a budget constraint.

Each potential seed in the first level network has its own cost of service for one time period. Without loss of generality, the service cost of the bloggers is assumed to be proportional to the number of their followers [37]. While scheduling the activation of seeds at each time period, the decision-maker has to meet the resource constraint for that time period, and also, the resource constraint over the whole time horizon. The lack of flexibility in defining resource constraints has long posed a challenge in applied IM [37, 60]. The resource constraints in the existing IM formulations (those which seek to activate all the seeds at once, in the initial time period) were either about the limit on the number of seeds to be selected [8, 25, 38] or the limit on the total budget to be spent on the seeds [37]. Prior to this work, no IM problem with a time-based profile of the available resources has been formulated and solved, to the authors' awareness. This research gap, addressed by this paper with the introduction of SASP, is likely due to the fact that all the original IM formulations were time independent.

SASP assumes that the activation of the same seed in multiple consecutive time periods leads to the reduction of its influenceability. This is because if the users in the second level perceive the "same message" in some consecutive posts by a blogger, they might suspect that the posts are paid and gradually lose the trust in the provided information. Thus, when a seed (a hired sponsored blogger) in the SASP maintains its activation over multiple time periods, its influenceability is multiplied by the factor $\gamma \leq 1$. To mitigate this effect, the decision-maker might want to deactivate seeds from time to time, to help them regain their influenceability. Note, however, that while this makes frequent deactivations and reactivations of the same seed appealing over the span of a campaign, this practice might not be cheap if each new (re)activation (think a new contract) incurs a fixed cost. The PPC diffusion model, presented in Section 3.2, models how the positive evidence (initiated in the first level network), as well as the negative evidence, spread through the second level network over time.

Continuing with the introduction of the SASP, it assumes that the cost C_j is incurred for having/keeping node j in the first level network active at each time period. Further, a fixed cost f is incurred with each new activation: if f is large, the preferred scheduling policies will keep the seeds activated for longer time periods, and vice versa. Also, the SASP allows one to account for the Net Present Value (NPV) of money. The corresponding multiplication factor π_t is introduced for time period t, where $0 \le \pi_t \le 1$ for $0 \le t \le T$ (T is the campaign horizon). It is natural to have π_t decrease with time to reflect a positive interest rate. Note that the seed activation spending is restricted at each time period $0 \le t \le T$ by the allocated budget R_t, and over the whole time horizon of the problem by the total campaign budget B.

In the objective function of the SASP, the positive activation of each node in the second level at each time period results in the gain of G_1, while the negative activation of each node similarly incurs the loss of G_2. Naturally, when G_1 is much greater than G_2, the SASP solutions will ignore the presence of a competitor, i.e., the seeds will be selected in a way to spread the positive evidence most efficiently. Meanwhile, With the large G_2, the solutions will focus on blocking the spread of the negative evidence.

3.2 Partial Parallel Cascade (PPC) model

Samadi *et al.*, [54] were first to model the spread of influence in a social network as a transfer of Bayesian evidence: they assume that each node is testing the null hypothesis that the opinion/claim preferred by the decision-maker is true, based on the evidence in support of or against this opinion/claim received from/through their directly connected peers. This paper presents a more general evidence-based diffusion model that allows for partial activation of nodes. This model can track not only opinion/product "adoption" but also "awareness", which is in line with the concepts in the experimental sociology and marketing literature [31, 44, 51].

The PPC model is now described in detail, in application to the blogger-centric marketing setting presented in Section 3.1. Consider a finite directed graph $G(N_2, A_2)$, where N_2 and A_2 denote the set of nodes and arcs in the second level network, respectively. Let N_1 denote the set of potential seeds in the first level network and A_1 be defined as the set of arcs directed from the first level network to the second level network. The diffusion process begins with the activation of seeds in the first level; the binary variable S^+_{jt} indicates whether node $j \in N_1$ is activated at time t. At each discrete time period $0 \le t \le T$, each node $i \in N_2$ receives the positive evidence e' from each of the activated seeds that it is following in first level (note that this amount is subject to the discount based on the activation history, and hence, the current influenceability of each seed), and a piece of negative evidence e'' from the competing evidence source. The nodes do not just count on their own observations in making judgments about the hypotheses; they communicate and transfer their impressions/opinions to each other: the opinion transferred to a node through a friend is viewed as a piece of evidence supporting

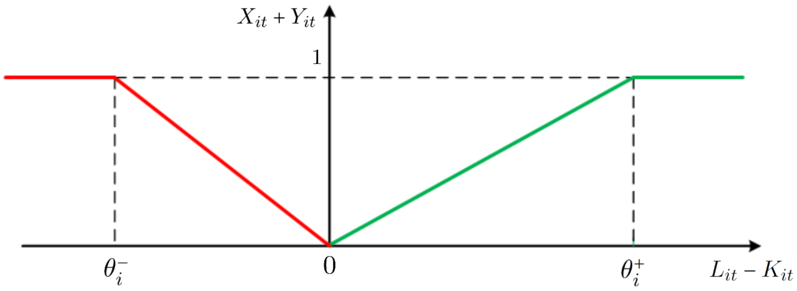

Figure 2: Partial activation of node i as a function its net evidence value at time t.

the corresponding hypothesis. At each time period $0 \le t \le T$, each node $i \in N_2$ delivers e^+X_{it} (e^-Y_{it}) of positive (negative) evidence to its out-neighbors, where $0 \le X_{it} \le 1$ and $0 \le Y_{it} \le 1$ denote the partial positive and negative activation levels of node i at time t, respectively. The amount of positive (negative) evidence resulting from a single new observation, supporting the null (alternative) hypothesis, defines the evidence increment e^+ (e^-); the numerical value of e^+ (e^-) can be calculated using the Bayesian inference logic [54]. Note that the PPC is a non-progressive diffusion model: the nodes may lose/change their activation states over time [7].

At the end of each time period $0 \le t \le T$, each node $i \in N_2$ updates its cumulative positive evidence tally ($L_{it} \ge 0$), and separately, negative evidence tally ($K_{it} \ge 0$), based upon the evidence it receives from its activated neighbors (in N_2) and/or connected seeds (in N_1). While testing the null hypothesis (which is supported by positive evidence) versus the alternative hypothesis (which is supported by negative evidence) at time t, node i aggregates the tallies L_{it} and K_{it}, which determines whether, in this time period, it adopts the positive view towards the marketed opinion/product (favoring the null hypothesis), or adopts the negative view (favoring the alternative hypothesis), or stays indifferent. The node's view is reflected in its activation state: either (partial) positive or (partial) negative activation, or no activation. Node i has its own positive (negative) threshold value $\theta_i^+ \ge 0$ ($\theta_i^- \ge 0$), representing the minimum net positive (negative) evidence level $L_{it} - K_{it}$ ($K_{it} - L_{it}$) it needs to match or exceed to get fully positively (negatively) activated. Once a node gets fully activated, it can be assumed ready to take a stance/action supporting the corresponding hypothesis, e.g., purchase the marketed product (if positively activated) or a competing product (if negatively activated). Any evidence that strengthens a node's stance beyond the threshold levels θ^+ and θ^-, does not lead to any further action in support of or against the marketed product [1]. Figure 2 has the partial positive and negative activation levels for a node plotted as a function of its net evidence value. The positive threshold (θ_i^+) and negative threshold (θ_i^-) levels reflect the node's sensitivity towards perceived positive and negative evidence, respectively, and are not necessarily equal. On the vertical axis in Figure 2, one has the sum of X_{it} and Y_{it}; at each time period t, at least one of the quantities X_{it} and Y_{it} is zero.

Further, in line with the mathematical sociology literature, at the end of each time period, each node is assumed to forget a part of the evidence it has aggregated. The presence of this *forgetfulness effect*, the strength of which is quantified by parameters β_1 and β_2, for positive and negative evidence, respectively, motivates the SASP as a further development of influence maximization problem ideas and formulations.

3.3 A Mathematical Model for the SASP under the PPC Diffusion Model

A mixed-integer program is designed for the SASP under the PPC diffusion model, so as to find such a schedule for activating the potential seeds in the first level network that maximizes (minimizes) the spread of positive (negative) evidence in the second level network. A summary of notation used hereafter is given in Table 1. The objective of the SASP is to maximize the cumulative gain resulting from the partial positive and negative activations in the second level network within the problem time window,

$$(P) \qquad \max Z = \sum_{i \in N_2} \sum_{t=0}^{T} \Big(G_1(X_{it}) - G_2(Y_{it}) \Big). \tag{1}$$

The SASP is a maximization problem subject to the constraints presented in groups as follows.

(I) *Setting the partial positive and negative evidence processing rules* so that (1) the positive (negative) activation level of a node linearly increases until its net evidence value reaches its positive (negative) threshold, and (2) after the net evidence value of a node passes the positive (negative) threshold, the node gets full positive (negative) activation. Also, these constrains ensure that no node can be both positively and negatively activated at the same time,

$$X_{it} \leq \frac{L_{it} - K_{it}}{\theta_i^+} + M_{it}(1 - U_{it}) \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{2}$$

$$X_{it} \leq U_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{3}$$

$$Y_{it} \geq \frac{K_{it} - L_{it}}{\theta_i^-} - M_{it}(Z_{it}) \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{4}$$

$$Y_{it} \geq Z_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{5}$$

$$U_{it} \leq 1 + \frac{L_{it} - K_{it}}{M'_{it}} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{6}$$

$$Z_{it} \geq \frac{K_{it} - L_{it} - \theta_i^-}{M''_{it}} \qquad i \in N_2, \quad t = 0, 1, ..., T. \tag{7}$$

(II) *Updating cumulative positive and negative evidence levels* for the nodes in the second level network,

$$L_{it} = \beta_1 L_{it-1} + \sum_{\substack{k \in N_2 \\ (k,i) \in A_2}} E2^+_{kt-1} + \sum_{\substack{j \in N_1 \\ (j,i) \in A_1}} E1^+_{jt-1} \qquad i \in N_2, \quad t = 1, 2, ..., T, \tag{8}$$

$$K_{it} = \beta_2 K_{it-1} + \sum_{(k,i) \in A_2} E2^-_{kt-1} + e'' \qquad i \in N_2, \quad t = 1, 2, ..., T, \tag{9}$$

Table 1: Definitions of the indices, input parameters and decision variables in mathematical problem

Indices	
j	Index used for the nodes in the first level network
i, k	Indices used for the nodes in the second level network
t	Time period index
Inputs	
A_1	Set of arcs connecting the nodes in the first level network to those in the second level network
A_2	Set of arcs connecting the nodes within the second level network
N_1	Set of nodes in the first level network
N_2	Set of nodes in the second level network
T	Total number of time periods in the time horizon
θ_i^+	Positive threshold for node i in the second level network
θ_i^-	Negative threshold for node i in the second level network
e^+	Positive evidence delivered in one transfer among the second level network nodes
e^-	Negative evidence delivered in one transfer among the second level network nodes
e'	Positive evidence value delivered in one transfer by each positive seed (before any discount)
e''	Negative evidence value delivered in one transfer by the external competitor
β_1	Rate at which the second level network nodes forget the previously received positive evidence
β_2	Rate at which the second level network nodes forget the previously received negative evidence
γ	Efficiency reduction discount factor for the seeds being active in consecutive time periods
B	Maximum budget available for seed activations over the time horizon of the problem
f	Fixed cost of activating a new seed for any number of consecutive time periods
G_1	Unit gain of a positive activation per time period
G_2	Unit cost of a negative activation per time period
C_j	Cost of node $j \in N_1$ to serve as a seed at each time period
π_t	Discount factor applied to the cost of seed activation at time period t
R_t	Budget allocation for seed activation at time period t
Decision Variables	
X_{it}	Partial positive activation level of node $i \in N_2$ at time t
Y_{it}	Partial negative activation level of node $i \in N_2$ at time t
U_{it}	$\begin{cases} 1, & \text{if node } i \in N_2 \text{ is fully positively activated at time } t \ (L_{it} - K_{it} \geq 0) \\ 0, & \text{otherwise} \end{cases}$
Z_{it}	$\begin{cases} 1, & \text{if node } i \in N_2 \text{ is fully negatively activated at time } t \ (K_{it} - L_{it} \geq \theta_i^-) \\ 0, & \text{otherwise} \end{cases}$
L_{it}	Cumulative level of positive evidence for node $i \in N_2$ at time t
K_{it}	Cumulative level of negative evidence for node $i \in N_2$ at time t
$E2_{it}^+$	Value of positive evidence that node $i \in N_2$ provides to each of its out-neighbors at time t
$E2_{it}^-$	Value of negative evidence that node $i \in N_2$ provides to each of its out-neighbors at time t
$E1_{jt}^+$	Value of positive evidence that node $j \in N_1$ provides to each of its followers in the second level at time t
S_{jt}^+	$\begin{cases} 1, & \text{if node } j \in N_1 \text{ is selected by the decision-maker to serve as a positive seed at time } t, \\ 0, & \text{otherwise} \end{cases}$
F_{jt}	Fixed cost paid for having seed $j \in N_1$ newly activated or reactivated at time t

$$L_{i0} = 0 \qquad i \in N_2, \tag{10}$$

$$K_{i0} = 0 \qquad i \in N_2. \tag{11}$$

(III) *Updating the evidence* delivered by each node in the second level network to its out-neighbors at each time period,

$$E2_{it}^+ \leq e^+ X_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{12}$$

$$E2_{it}^- \geq e^- Y_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T. \tag{13}$$

(IV) *Budget constraints* for activating seeds (taking into account the net present value of the fixed and per-period costs),

$$\sum_{t=0}^{T}\sum_{j\in N_1}\left(\pi_t\cdot C_j(S^+_{jt})+\pi_t(F_{jt})\right)\leq B, \tag{14}$$

$$\sum_{j\in N_1}\left(\pi_t\cdot C_j(S^+_{jt})+\pi_t(F_{jt})\right)\leq R_t \qquad t=0,1,...,T, \tag{15}$$

$$F_{jt}\geq f(S^+_{jt}-S^+_{jt-1}), \qquad j\in N_1, \quad t=1,2,...,T, \tag{16}$$

$$F_{j0}=f\cdot S^+_{j0} \qquad j\in N_1. \tag{17}$$

(V) *Updating the evidence (positive)* delivered by the activated seeds in the first level network to the nodes in the second level network,

$$E1^+_{jt}\leq(\gamma E1^+_{jt-1})+(1-S^+_{jt-1})e' \qquad j\in N_1, \quad t=1,2,...,T, \tag{18}$$

$$E1^+_{jt}\leq e'(S^+_{jt}) \qquad j\in N_1, \quad t=0,1,...,T. \tag{19}$$

(VI) *Sign constraints*,

$$0\leq X_{it},Y_{it},E2^+_{it},E2^-_{it} \qquad i\in N_2, \quad t=0,1,...,T, \tag{20}$$

$$0\leq E1^+_{jt},F_{jt} \qquad j\in N_1, \quad t=0,1,...,T, \tag{21}$$

$$S^+_{jt}\in\{0,1\} \qquad j\in N_1, \quad t=0,1,...,T, \tag{22}$$

$$U_{it},Z_{it}\in\{0,1\} \qquad i\in N_2, \quad t=0,1,...,T. \tag{23}$$

Constraints (6)-(7) ensure that the values of the binary variables U_{it} and Z_{it} are correctly assigned based on L_{it} and K_{it} for each node $i\in N_2$ at each time period t. Constraints (2)-(5) ensure that each node gets its positive or negative activation status properly updated in each time period. Note that the objective function of (P) favors higher values of X_{it} and lower values of Y_{it}.

For the activation of node i at time t, four cases (namely, (a)-(d)) are possible: (a) node i is positively activated but its net evidence level is less than its positive threshold ($0\leq L_{it}-K_{it}<\theta^+_i$); then, constraints (6) and (7) result in $U_{it}=1$ and $Z_{it}=0$; this setting makes (4) and (5) redundant and leads to $Y_{it}=0$; on the other hand, (2) and (3) assign the partial positive activation of node i per the PPC diffusion model: $X_{it}=\frac{L_{it}-K_{it}}{\theta^+_i}$; (b) node i has full positive activation ($L_{it}-K_{it}\geq\theta^+_i$); then, the values of binary variables turn out the same as in case (a) ($U_{it}=1$ and $Z_{it}=0$) as ensured by constraints (6) and (7); the value of Z_{it} makes (4) and (5) redundant and ensures $Y_{it}=0$; for the positive activation, (2) becomes redundant and (3) results in $X_{it}=1$; (c) node i has a partial negative activation ($0\leq K_{it}-L_{it}<\theta^-_i$); then, both binary variables take zero value as ensured by (6) and (7); as such, (2) and (3) result in $X_{it}=0$; meanwhile, (4) and (5), with the help of the objective function, ensure the correct partial negative activation of node i ($Y_{it}=\frac{K_{it}-L_{it}}{\theta^-_i}$); (d) node i has a full negative activation ($K_{it}-L_{it}\geq\theta^-_i$);

then, by definition, $U_{it} = 0$ and $Z_{it} = 1$, ensured by (6) and (7), and also, (2) and (3) ensure that $X_{it} = 0$, (4) becomes redundant and (5) forces $Y_{it} = 1$. The minimum values of the "big M" variables that make constraints (2), (3), (6) and (7) in set (I) work are calculated in Appendix C.

Each node $i \in N_2$ starts the diffusion process in a neutral state ($L_{i0} = K_{i0} = 0$), which is guaranteed by (10) and (11). Constraint (8) updates the cumulative positive evidence level of node $i \in N_2$ at time period t as a sum of its positive evidence accumulated by period $t-1$ (discounted with β_1 for forgetfulness), the positive evidence received from all the connected nodes $k \in N_2$ at the second level ($E2^+_{kt-1}$) and the evidence received from all the seeds $j \in N_1$ that i follows in the first level ($E1^+_{jt-1}$). Similarly, (9) updates the cumulative negative evidence level for each node in the second level network, where e'' is the evidence amount that each node receives from the external source of negative evidence.

Constraints (12) and (13) ensure that each node in the second level delivers the right level of positive or negative evidence (depending on its activation) to its out-neighbors. Constraint (14) ensures that the total seed activation cost and the fixed cost of seed activations over the time horizon of the problem, after applying the discount factors, is less than or equal to the total budget. Constraint (15) guarantees that the total seed activation cost at each time period does not exceed the allocated budget. Constraints (16) and (17) ensure that the fixed cost is incurred whenever an inactive seed gets activated.

Constraints (18) and (19) determine the evidence that a seed $j \in N_1$ delivers to its followers at time t: whenever an inactive seed j just becomes active, it can deliver the e' amount of evidence to its followers, but with every subsequent time period that seed j maintains its activation, the amount of evidence that it delivers to its followers drops by the factor of γ.

The mathematical program (P) includes two binary variable sets, U_{it} and Z_{it}, that track the positive and negative activations: these variables are the key drivers of the problem's complexity.

THEOREM 1. *The SASP under PPC diffusion model is NP-hard.*

PROOF: See Appendix B.

On a final note, it is easy to show that the option to activate some seeds late can never make a solution to an IM problem worse: i.e., the optimal value of the SASP provides an upper bound for the OIMP (where seeds can only be selected/activated in the initial time period).

PROPOSITION 1. *The optimal value of the SASP is a valid upper bound for the OIMP.*

PROOF: See Appendix B.

3.4 Case Studies

This section presents two case studies that help explore the properties of optimal SASP solutions. Case Study 1 demonstrates the value of seed activation scheduling as opposed to relying exclusively on "early starters" for IM, in application to two real-world social networks. A detailed description of these data, along with an explanation of how they informed the construction of the networks used in the presented experiments, is provided in Section 5. Case Study 2 is conducted with a small artificial social network to capture the effect of the problem parameters and objective function on the optimal seed activation strategies. The SASP is considered with two objective functions: first, counting the weighted activation (positive minus negative) over time, i.e., $\sum_{i\in N_2}\sum_{t=0}^{T}\big(G_1(X_{it}) - G_2(Y_{it})\big)$, named Obj.I, and second, counting the weighted activation (positive minus negative) at the final time period, i.e., $\sum_{i\in N_2}\big(G_1(X_{iT}) - G_2(Y_{iT})\big)$, named Obj.II. While Obj.I is suitable for such problems where the positive activation of each user benefits the decision-maker in each period, Obj.II works better for a time constrained campaign, e.g., an election, where the decision-maker only cares about the activation status/awareness level of the nodes at a pre-set time deadline.

Case Study 1. In this study, the optimal values for the SASP and OIMP are reported for problem instances with the forgetfulness rate (β_1) values ranging from 0 to 1, with no budget constraints.

Figure 3 compares the optimal values of the SASP and OIMP for Obj.I and Obj.II on the real-world social network of people contributing in "Multiple Sclerosis" forum, named Network 1. Network 1 includes 709 bloggers, of which six superusers are designated as potential seeds and placed in the first level network, with all the other nodes forming the second level network. Note that in online health forum research, most well-connected users and voluminous contributors are conventionally referred to as "superusers."

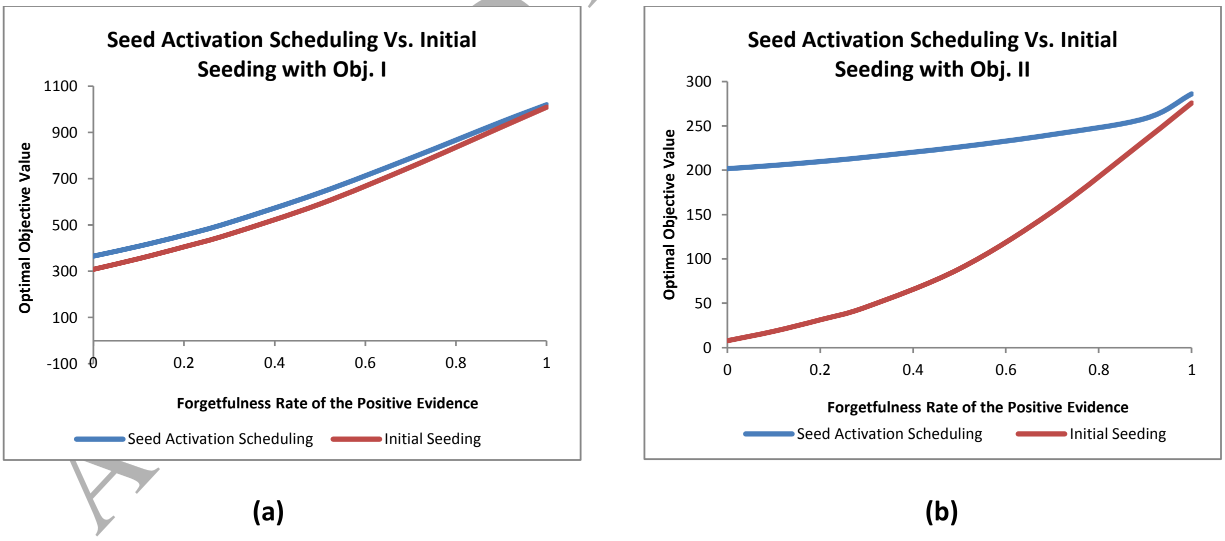

Figure 3: SASP optimal values vs. those for strategies relying purely on initial seeding, over the range of feasible values for forgetfulness, with $N_1 = 6$, $N_2 = 703$, $T = 4$, $\beta_2 = \gamma = 0.7$, $e^+ = 0.45$, $e^- = 0.2$, $e' = 1.1$, $e'' = 0.1$, $G_1 = G_2 = 1$, $B = 900$, $f = 25$, $I_t = 0.02$ and $\theta^+ = \theta^- = 2$: (a) with Obj.I, (b) with Obj.II.

The results in Figure 3 show that, even with a small time horizon, seed activation scheduling has an advantage over the original IM, particularly if the forgetfulness rate is high, i.e., when β_1 is small. As β_1 grows, the forgetfulness has a smaller impact on the spread of evidence, and the objective values for the SASP and the OIMP get closer. The advantage of using the SASP solution is greater with Obj.II. When the seeds are all selected in the initial time period, and the results are evaluated at a certain time horizon, the forgetfulness effect reduces the influence produced by the initial seeds. On the other hand, scheduling seed activations over time allows one to continuously fuel evidence diffusion, i.e., stopping the diffusion of negative evidence early and adding seeds at carefully selected time periods during the diffusion process to combat the forgetfulness effect. Note that the optimal solutions to the SASP and OIMP problems are still different when forgetfulness is removed from the problem ($\beta_1 = 1$), because the SASP permits the decision-maker to activate a single blogger multiple times. If superusers' out-degrees significantly differ, and/or the magnitude of the positive evidence delivered by positive seeds is small (so that multiple evidence updates are required to cause activations in the second level network), then the OIMP optima turn out far worse when plugged into an SASP, compared to the SASP true optima.

Another set of experiments is performed with Network 2 – a social network of users of "Herpes" forum (see Section 5.1 for more details on this dataset). In 2015, 784 users have contributed to it; four superusers with the average degree of 153 are placed into the first level of Network 2, with the second level formed by 780 other forum users. The time horizon of the problem is slightly increased ($T = 6$) to make the forgetfulness effect more pronounced. The results (see Figure 4) reveal a trend similar to the one observed in the experiments with Network 1. However, the difference between the optimal objective values of the SASP and those of the IM problem in the absence of forgetfulness is now more visible. In summary, Case Study 1 experimentally validates the claim of Section 1: under the PPC diffusion model, the idea of seed activation scheduling has merit.

Case Study 2. The impact of the SASP input parameters on the optimal seed activation schedules is analyzed next. Consider a network in Figure 5(a), consisting of ten regular bloggers (second level) that can be influenced by three high-profile bloggers (first level) and an external negative evidence source (not shown in the figure).

For one fixed parameter setting, Figure 5(b) presents the optimal seed activation schedules for the SASP instances with Obj.I and Obj.II, respectively. The schedule corresponding to each solution is represented by a vertical ribbon, where the time periods go from $t = 0$ to $t = T$, separated from each other by black lines. In each time period, a black circle in a square indicates seed activation; the lower square is for node A, the middle one for B, and the upper one for C. Under Obj.I, all the activations are equally profitable over the time horizon of the problem; thus, the corresponding optimal seeding strategy focuses more on the initial time periods, in order to initiate the cascade, and also, holds a part of budget for some late activations to keep the positive cascade alive. The optimal seeding strategy

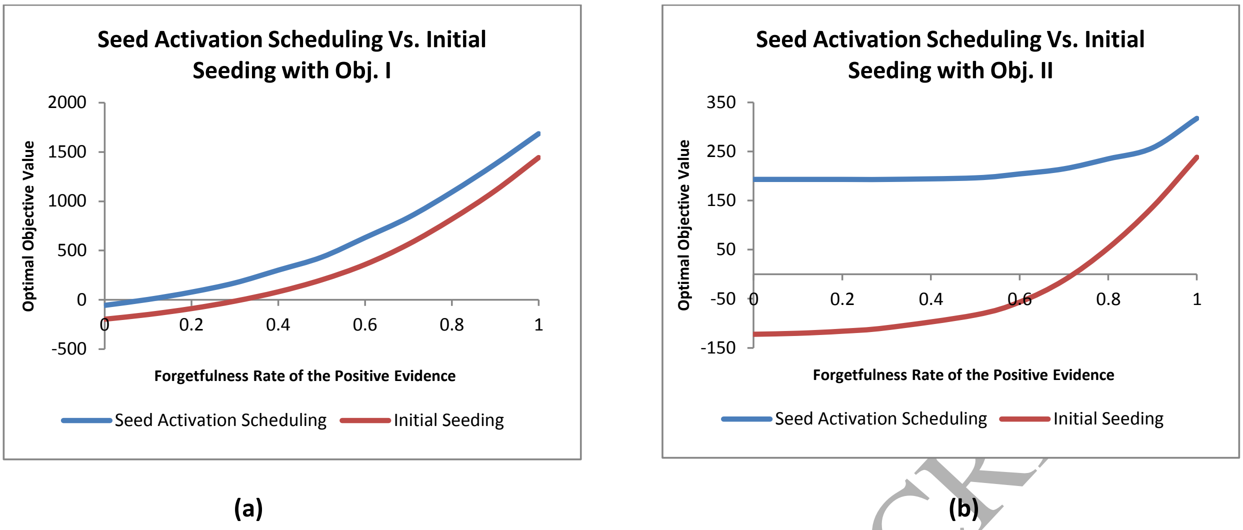

Figure 4: SASP optimal values vs. those for strategies relying purely on initial seeding, over the range of feasible values for forgetfulness, with $N_1 = 780$, $N_2 = 4$, $T = 6$, $\beta_2 = \gamma = 0.7$, $e^+ = 0.4$, $e^- = 0.2$, $e'' = 0.1$, $G_1 = G_2 = 1$, $B = 1100$, $f = 25$, $I_t = 0.02$ and $\theta^+ = \theta^- = 2$: (a) with Obj.I, (b) with Obj.II.

under Obj.II, on the other hand, focuses more on the few last time periods, to secure the largest number of active nodes at the very end of the campaign. It spends a small portion of the budget early, to hamper the negative evidence spread, and then, uses most of the budget late to achieve a favorable network state by the preset deadline T.

Unrestricted Budget: A campaign with an unlimited budget represents a situation where the decision-maker is not worried about the seeding costs; both the budget (B) and allocated budget at time t (R_t) are assumed infinite. In this case, nothing can stop the decision-maker from activating all the seeds at all the time periods ($0 \leq t \leq T$). The γ parameter (the efficiency reduction parameter of the activated seeds) is the only driver behind the decision-maker's willingness to have seeds deactivated at times.

Figure 6 shows how the optimal objective value and optimal seed activation schedule change under the varied values of γ. According to Figure 6(a), the same trend emerges with both Obj.I and Obj.II: the higher γ values allow the influence of each seed to have a lasting effect, propelling the positive cascades on. Under the optimal seed activation strategy (see Figure 6(b)), when γ is large (so that the efficiency of activated seeds does not decrease over time), the decision-maker activates all the seeds in each time period. As γ decreases, a trade-off emerges between keeping the less efficient seeds activated for a longer time and alternating the seed activations while making each activation short. When the γ value is either too small or too large, an optimal seeding strategy is easy to find, but finding an optimum under the medium values of γ becomes challenging. In the latter case, the optimal strategy activates at

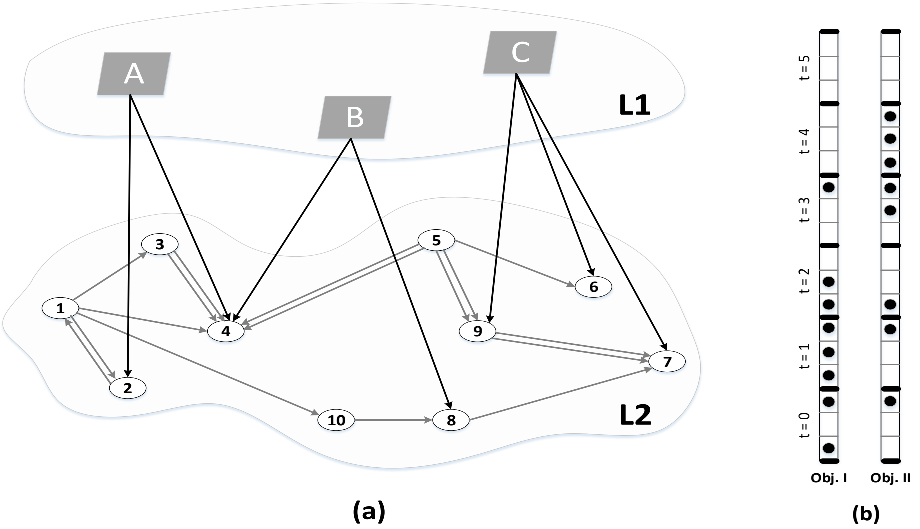

Figure 5: Analysis of the optimal seed activation scheduling over a small blogger network: (a) two-level network with potential seeds in the first level and the campaign targets in the second level, (b) optimal seed activation schedules for Obj.I and II, with $N_1 = 3$, $N_2 = 10$, $T = 5$, $\beta_1 = \beta_2 = \gamma = 0.7$, $e^+ = e^- = 0.4$, $e^- = 0.2$, $e'' = 0.1$, $G_1 = G_2 = 1$, $B = 480$, $f = 25$, $I_t = 0.04$, $\theta^+ = \theta^- = 2$.

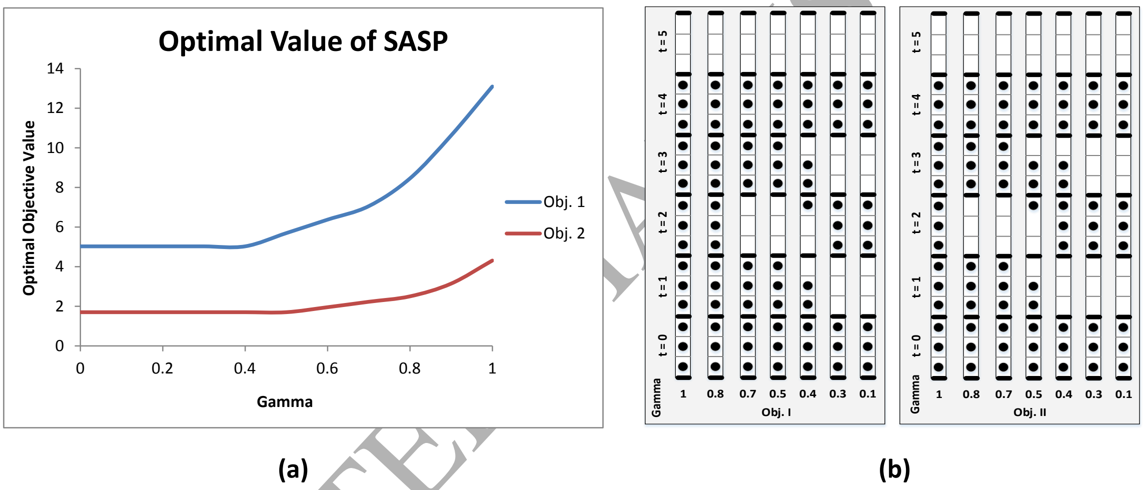

Figure 6: Analysis of optimal seed activation schedules under unrestricted budget: (a) the optima for Obj.I and Obj.II, with γ parameter, (b) the optimal seed activation schedules with γ parameter.

least one seed at each time period to fuel the diffusion in the second level network while coping with the seed efficiency reduction (over consecutive time periods). One insight of this case study is that spending all the available resources does not always work out for the best. Indeed, consider a billionaire running an election campaign in a community whose members do not like to be fed the same/repeating content by the media. In such a situation, activating all the possible bloggers, although financially feasible, can make the electorate unresponsive to all-out campaigns. Finding an optimal campaign strategy, considering the population characteristics on the one hand and the competitor influence on the other hand, is a hard problem that can be addressed as a carefully formulated instance of SASP.

Time Horizon Analysis: This case study explores the effect of SASP time horizon on optimal seed activation scheduling. Under a limited campaign budget, a decision-maker in search of a schedule that optimizes the positive evidence spread must be mindful of the campaign length and deadline(s).

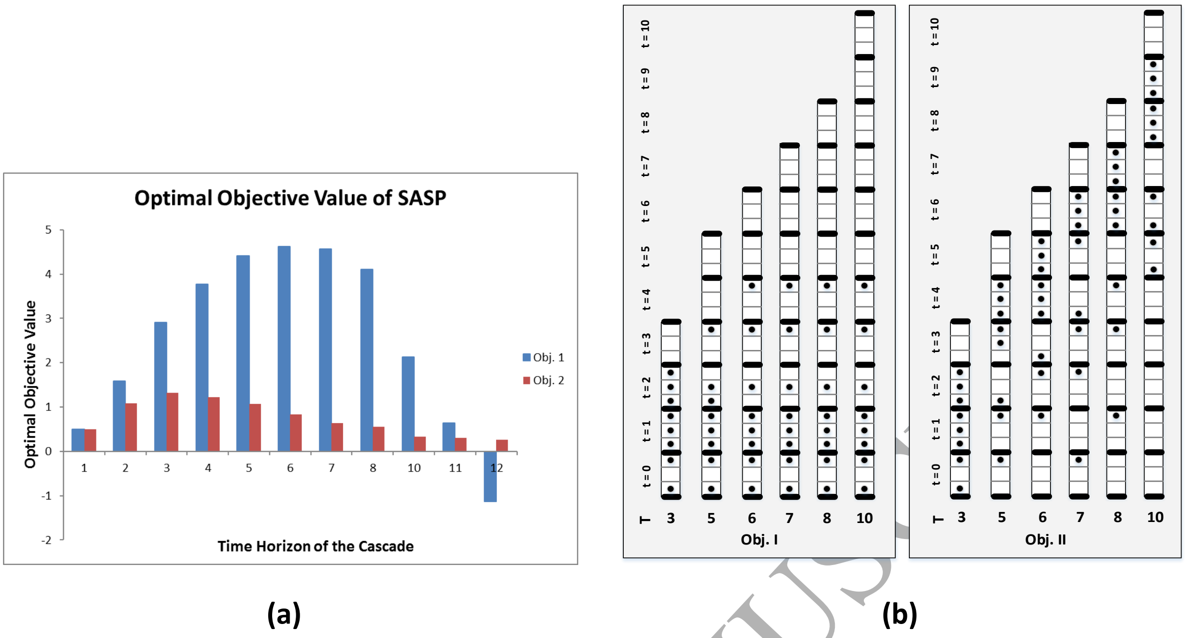

Figure 7: Analysis of the variation in the optimal seed activation schedules under a varied time horizon: (a) a trend in the optimal value of the objective function (Obj.I and Obj.II) with time horizon, (b) the optimal seed activation schedules with time horizons.

The results of a series of experiments with SASPs with competition are reported in Figure 7. As expected, for $T = 1$, the optimal solutions for the problem instances with Obj.I or Obj.II turn out the same. Indeed, Obj.I returns the count of the activated nodes in the second level network at times $t = 0$ and $t = 1$, and since no node is activated at $t = 0$, then Obj.I turns out equal to the count at the cascade deadline ($t = T = 1$), defined as Obj.II (see Figure 7(a)). The time-based dynamics of evidence cascade propagation and decline have been studied in another work of the authors [55]. The non-zero forgetfulness factor β_2, and the presence of a competitor whose influence intensity does not decrease over time, are the key factors responsible for the unimodal form of the curves in Figure 7(a). Note that it is important for a decision-maker to define the problem time horizon in a way most fitting their needs, before solving and interpreting the results of the SASP. Observe that for each pre-set time horizon value $T > 0$, when the same SASP is solved with either Obj.I or Obj.II, the range of the possible objective function values (distance between its upper and lower bounds) for Obj.I is greater than or equal to that of Obj.II. As such, when the time horizon is defined so that the positive evidence is able to spread, Obj.I is always greater than Obj.II. On the other hand, when the time horizon is large enough so that the predefined fixed budget cannot keep the positive cascade alive, Obj.I turns negative earlier than Obj.II.

The results of this experiment suggest that under a limited budget, the SASP optimum is necessarily a unimodal function of its time horizon. This, in general, is not true. For example, with a large interest rate, a large time horizon provides more opportunities for the decision-maker to activate more late seeds under the same budget. The results of the same experiment, repeated with a higher interest rate value ($I_t = 0.14$), are shown in Figure 8: the objective value-based bar chart now has multiple peaks. Finally, with no NPV effect, i.e., when $I_t = 0$, an optimum of the SASP with a fixed budget does not monotonically increase with the growing time horizon.

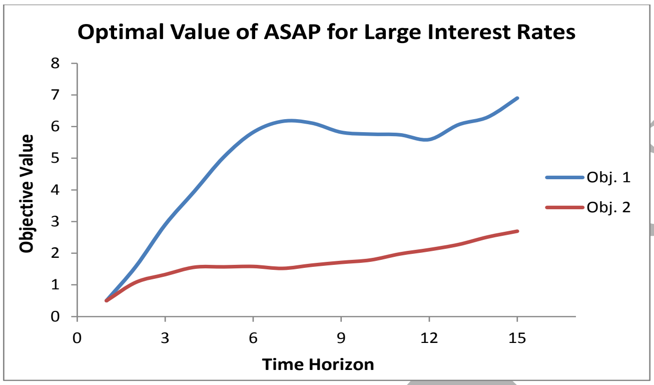

Figure 8: Analysis of the optima for SASP under varied time horizons, with high interest rates.

Pure positive cascades: In an SASP with no competition, e.g., for a decision-maker promoting a new product in a new market, increasing the time horizon never reduces the objective function value; however, it decreases the marginal gain in the objective value per period, due to forgetfulness.

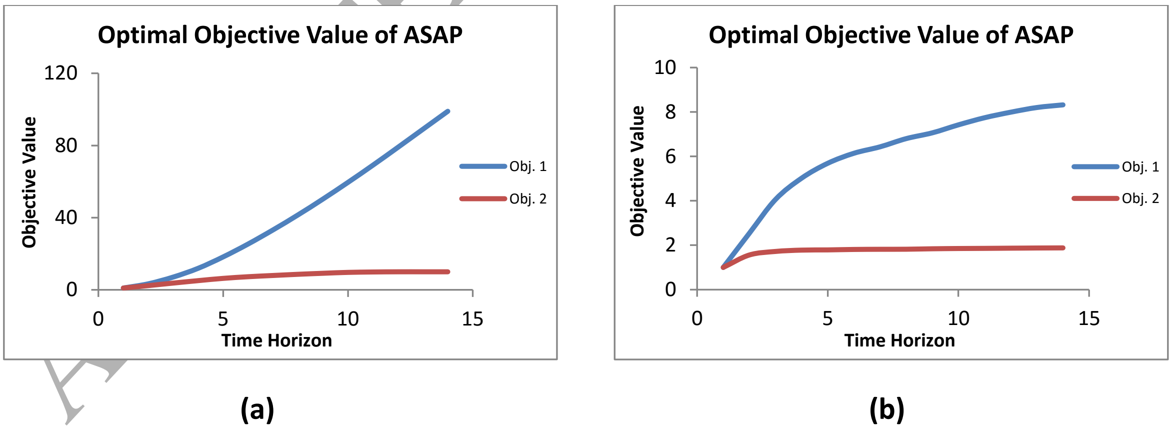

Figure 9: Analysis of the behavior of optimal value of the objective function (Obj.I and Obj.II) with time horizon in pure positive cascade: (a): no forgetfulness effect, (b): high forgetfulness effect.

The results of an experiment with a pure positive cascade are shown in Figure 9. When the forgetfulness effect is negligible, increasing the time horizon helps the decision-maker to secure a larger gain, under a fixed budget (Figure 9(a)). Note that Obj.II is bounded from above by a total number of nodes in the second level network. When this bound is reached, a further increase of the time horizon has no effect. Meanwhile, with Obj.I, increasing the time horizon always allows the decision-maker to achieve a higher possible gain, to the point where all the resources have been optimally exploited.

With a non-zero forgetfulness effect, the marginal gain in the optimal objective value reduces over time, even for purely positive cascades. This is because the amount of fresh evidence injected into the network by the seeds is fixed, and the intensity of communication between the second level nodes drops due to forgetfulness. Figure 9(b) shows how the addition of forgetfulness affects theT SASP optima.

Cascades with intermediate and final deadlines: In this case study, Obj.II is revised to Obj.III, to quantify the activation gain (the number of positively activated nodes minus the number of negatively activated ones) at multiple deadlines: e.g., an SASP with Obj.III would assist a political party candidate in maximizing the electorate support both at the time of a primary election (intermediate deadline) and general election (final deadline). Obj.III is given as $\sum_{i \in N_2} \big(G_1(X_{iT}) - G_2(Y_{iT})\big) + \sum_{i \in N_2} \big(G_3(X_{iT'}) - G_4(Y_{iT'})\big)$, where G_3 and G_4 are the weights applied to the counts of the positively and negatively activated nodes at the intermediate deadline $q * T$, with $0 \le q \le 1$, and at final deadline T', respectively. Note that Obj.I is a special case of Obj.III, with T intermediate deadlines uniformly distributed over the time horizon.

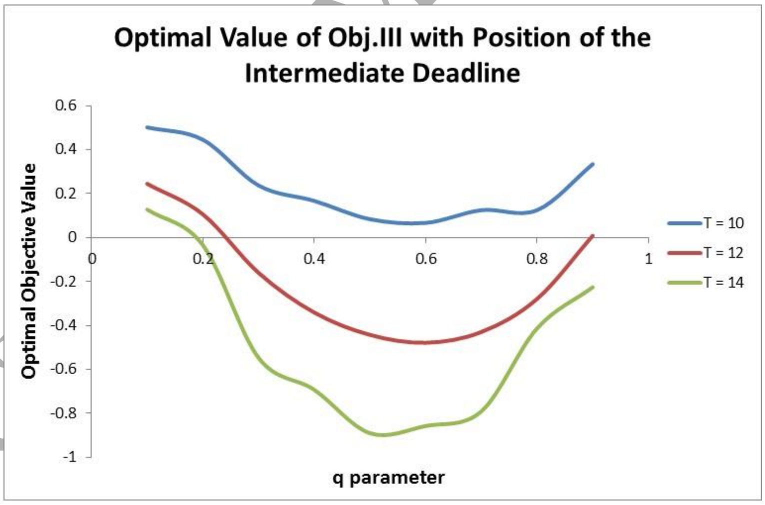

Figure 10: Optima for Obj.III with an intermediate deadline, for varied time horizons.

An experiment is performed with the SASP instances with Obj.III with the varied parameter q. The results in Figure 10 show that the intermediate deadline in the middle of the time horizon presents the biggest challenge for the decision-maker. In this case, the decision-maker needs to ignite the cascade early to achieve a high gain, while trying to save resource in order to finish strong.

Note that an SASP with Obj.III, where the resource availability is stochastically dependent on the intermediate campaign results, presents an interesting extension opportunity to the present work.

4 A Heuristic Solution Methodology using Column Generation

Most real-world SASP instances in the blogger-centric marketing domain cannot be expected to be solved to optimality using exact optimization tools. This section presents a decomposition method, using column generation, to solve the SASP, somewhat similarly to solving the well-studied vehicle routing problem [14]. A solution to the relaxed master problem of the column generation procedure provides an upper bound for an optimal solution to the mixed-integer program (P). On the other hand, the structure of the problem allows for extracting a tight integer solution within the column generation procedure, offering a lower bound.

4.1 A Relaxed Master Problem for the SASP

The relaxed master problem keeps a pool of solutions (feasible schedules) generated by solving the sub-problem and works to find an optimal combination of the schedules to maximize the SASP objective. The binary decision variable representing a schedule is relaxed so that the dual information can be extracted and exported to the sub-problem to generate new, improving feasible schedules. Table 2 provides a list of notation used hereafter, in addition to those in Table 1: the new notation are used in the formulation of the column generation heuristic. To clarify the column generation procedure, the inputs and outputs for the SASP relaxed master problem and sub-problem are presented in Table 3.

The relaxed master problem (RMP) for SASP is formulated as follows:

Table 2: Notation used in the Sub-Problem and the Relaxed Master Problem

Notation	
H	Set of solutions of the Relaxed Master Problem
$\hat{S}^h_{jt}$	Activation of seed j at time t in solution $h \in H$
λ	Dual cost of constraint (25)
δ_{jt}	Dual cost of constraint (38) for seed $j \in N_1$ at time t
σ_{jt}	Dual cost of constraint (39) for seed $j \in N_1$ at time t
W^h	The portion of solution $h \in H$ included in the solution of the Relaxed Master Problem

Table 3: List of inputs and outputs of the Sub-Problem and the Relaxed Master Problem

	Relaxed Master Problem	Sub-Problem
Inputs	H $\hat{S}^h_{jt}$ $j \in N_1,\ t = 0, 1, ..., T,\ h \in H$	λ δ_{jt} $j \in N_1,\ t = 0, 1, ..., T$ σ_{jt} $j \in N_1,\ t = 0, 1, ..., T$
Outputs	λ δ_{jt} $j \in N_1,\ t = 0, 1, ..., T$ σ_{jt} $j \in N_1,\ t = 0, 1, ..., T$ W^h $h \in H$	S^+_{jt} $j \in N_1,\ t = 0, 1, ..., T$

$$(RMP) \qquad \max Z = \sum_{i \in N_2} \sum_{t=0}^{T} \Bigg(G_1(X_{it}) - G_2(Y_{it}) \Bigg) \tag{24}$$

Subject to:

$$\sum_{h \in H} W^h = 1 \qquad \text{Dual cost: } \lambda, \tag{25}$$

$$X_{it} \le \frac{L_{it} - K_{it}}{\theta_i^+} + M_{it}(1 - U_{it}) \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{26}$$

$$X_{it} \le U_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{27}$$

$$Y_{it} \ge \frac{K_{it} - L_{it}}{\theta_i^-} - M_{it}(Z_{it}) \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{28}$$

$$Y_{it} \ge Z_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{29}$$

$$U_{it} \le 1 + \frac{L_{it} - K_{it}}{M'_{it}} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{30}$$

$$Z_{it} \ge \frac{K_{it} - L_{it} - \theta_i^-}{M''_{it}} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{31}$$

$$L_{it} = \beta_1 L_{it-1} + \sum_{\substack{k \in N_2 \\ (k,i) \in A_2}} E2^+_{kt-1} + \sum_{\substack{j \in N_1 \\ (j,i) \in A_1}} E1^+_{jt-1} \qquad i \in N_2, \quad t = 1, 2, ..., T, \tag{32}$$

$$K_{it} = \beta_2 K_{it-1} + \sum_{(k,i) \in A_2} E2^-_{kt-1} + e'' \qquad i \in N_2, \quad t = 1, 2, ..., T, \tag{33}$$

$$L_{i0} = 0 \qquad i \in N_2, \tag{34}$$

$$K_{i0} = 0 \qquad i \in N_2, \tag{35}$$

$$E2^+_{it} \le e^+ X_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{36}$$

$$E2^-_{it} \ge e^- Y_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{37}$$

$$E1^+_{jt} - \gamma E1^+_{jt-1} + e' \sum_{h \in H} W^h \hat{S}^h_{jt-1} \le e' \qquad j \in N_1, \quad h \in H, \ \ t = 1, 2, ..., T, \qquad \text{Dual cost: } \delta_{jt-1}, \tag{38}$$

$$E1^+_{jt} - e' \sum_{h \in H} W^h \hat{S}^h_{jt} \le 0 \qquad j \in N_1, \quad h \in H, \ \ t = 0, 1, ..., T, \qquad \text{Dual cost: } \sigma_{jt}, \tag{39}$$

$$0 \le X_{it}, Y_{it}, U_{it}, Z_{it}, L_{it}, K_{it}, E2^+_{it}, E2^-_{it} \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{40}$$

$$0 \le E1^+_{jt} \qquad j \in N_1, \quad t = 0, 1, ..., T, \tag{41}$$

$$0 \le W^h \qquad h \in H, \tag{42}$$

$$U_{it}, Z_{it} \le 1 \qquad i \in N_2, \quad t = 0, 1, ..., T, \tag{43}$$

$$W^h \le 1 \qquad h \in H. \tag{44}$$

Problem (RMP) includes all the constraints of (P) except for constraint sets (IV) and (VI). Index h is the index of a schedule in the solution pool of (RMP), and the decision variable W^h denotes the portion of solution h included in the optimal solution of the (RMP). The summation of the selected portions of the complete available feasible solutions is equal to one (constraint set (25)), since (P) is designed to find a single schedule. Note that the value of the dual variable is extracted only for constraint sets (25), (38) and (39) – those that include the decision variable W^h; the dual values of the other constraints do not help the sub-problem in introducing new improving feasible schedules into the (RMP).

4.2 A Sub-Problem for the Partial Activation Model

The sub-problem receives the dual variable values from the relaxed master problem and finds a new feasible schedule which, when added to the solution pool in the (RMP), could enter the basis and improve the problem's objective value. As far as the optimal objective value of the sub-problem is negative, adding the corresponding schedule to the (RMP) is beneficial. As such, the column generation routine stops when the sub-problem returns a non-negative objective value. The constraints of the sub-problem (SP) ensure the feasibility of the introduced schedules:

$$(SP) \qquad \min Z' = \lambda + e' \sum_{j \in N_1} \left(\sum_{t=0}^{T-1} \big(\delta_{jt} - \sigma_{jt}\big) S^+_{jt} - \sigma_{jT} S^+_{jT} \right) \tag{45}$$

Subject to:

$$\sum_{t=0}^{T} \sum_{j \in N_1} \big(\pi_t \cdot C_j(S^+_{jt}) + \pi_t(F_{jt})\big) \le B, \tag{46}$$

$$F_{jt} \ge f(S^+_{jt} - S^+_{jt-1}), \qquad j \in N_1, \quad t = 1, 2, ..., T, \tag{47}$$

$$F_{j0} = f \cdot S^+_{j0} \qquad j \in N_1, \tag{48}$$

$$\sum_{j \in N_1} \big(\pi_t \cdot C_j(S^+_{jt}) + \pi_t(F_{jt})\big) \le R_t \qquad t = 0, 1, ..., T, \tag{49}$$

$$0 \le F_{jt} \qquad j \in N_1, \quad t = 0, 1, ..., T, \tag{50}$$

$$S^+_{jt} \in \{0, 1\} \qquad j \in N_1, \quad t = 0, 1, ..., T. \tag{51}$$

This sub-problem is a generalized knapsack problem that selects as many items (seeds) as it can afford, so that the objective function is optimized. However, this problem has multiple time periods (namely, $T + 1$ periods) and the decisions are made sequentially; each item can be selected more than once, and the fixed cost of item selection depends on what period(s) an item is being selected for. The computational results with the SASP instances on real social networks in Section 5 showcase that in practice, (SP) can be solved to optimality rather fast even for large networks.

4.3 Initialization of the Solution Pool for Column Generation

The presented column generation algorithm needs a pool of initial solutions to start with, to which it iteratively adds more solutions (columns). The quality of the initial solutions may affect the number of iterations till convergence. However, an organized selection of these solutions guarantees neither a faster convergence to an optimum of the (RMP), nor a higher quality of the best integer solution (schedule) obtained at the end of the column generation procedure [49]. Random selection of initial solutions has been previously explored in column generation research [20]. Appendix A presents four initialization algorithms that exploit the properties of the SASP; each has a specific preference in scheduling seed activation, and together, they generate a diverse set of good initial SASP solutions.

Each algorithm provides one initial solution; thus, the column generation process begins with four initial columns. The results in Section 5 confirm the viability of this initialization method.

4.4 A Heuristic Toolbox for Finding Computational Bounds

The proposed column generation heuristic, when converges, provides an optimal solution to the relaxed master problem. At that point, there is no new column (schedule), that, if added to the solution pool, can improve the objective function of the relaxed master problem. As (P) is a mixed-integer program, the optimal solution to (RMP) only provides an upper bound for (P). On the other hand, solving an integer program over all the columns in the solution pool of (RMP) provides a feasible solution and a valid lower bound for (P). Only one schedule (column) is selected by the integer program from all the generated columns – the column with the maximum objective value. Due to the special structure of the problem, finding a lower bound can be done more efficiently than running the integer version of (RMP), i.e., with variables $U, Z \in \{0,1\}$. Given a valid schedule (column), the objective value of SASP is calculated using a deterministic simulation process that spreads evidence over the network and records the outcome. When the objective value of each available column is evaluated, the algorithm returns the schedule with a maximum outcome. Running the evaluation module for each schedule takes $O(T|A|)$ time, where $A = A_1 \cup A_2$; thus, running the evaluation module takes less time than a solver.

At each iteration of the column generation procedure, when the sub-problem introduces a new column into the (RMP), the heuristic toolbox first sends a new schedule to the evaluation module to calculate the corresponding objective value, and if this value turns out greater than that for the best integer solution found so far, then the solution is stored as the new best integer solution. At the end of the column generation procedure, the objective value of the best integer solution is returned as the lower bound for (P). An optimal solution to (P) is contained in the range between this lower bound and the upper bound, obtained for the solution to (RMP) at the end of the column generation procedure. In practice, the column generation upper bound turns out to be loose. In fact, it turns out to be not far from the linear programming relaxation bound, and hence, using the Branch and Price approach, initiated at the corresponding solution, is also not advisable. In order to tackle the challenge of improving the upper bound, the structural properties of (P) must be exploited.

The basic idea of strengthening the upper bound is to develop a simpler problem by relaxing some constraints in (P) and then solving the remaining problem to optimality (using Cplex); indeed, this upper bound remains valid. To obtain the tightest such upper bound, one can relax as far as possible the constraints that are inactive, so that the remaining problem can be solved easily without affecting the upper bound quality too much. The lower bound solution is examined to see which constraints of set (I) can be relaxed. The diffusion process is simulated from the lower bound solution S^*_{jt} $(\forall j \in N_1, 0 \le t \le T)$, and thereby, the values of the binary variables U_{it} and Z_{it} and of the activation variables X_{it} and Y_{it} $(\forall i \in N_2, 0 \le t \le T)$ are obtained. When the value of X_{it} is 1, all constraints in set (I) can be relaxed for node i at time t. On the other hand, when the value of Z_{it} is 1, then constraints (3) and (5) can be relaxed. The results in Section 5 confirm that this developed upper bound performs well in practice. To recap, Figure 11 summarizes the interaction among modules in the presented heuristic toolbox.

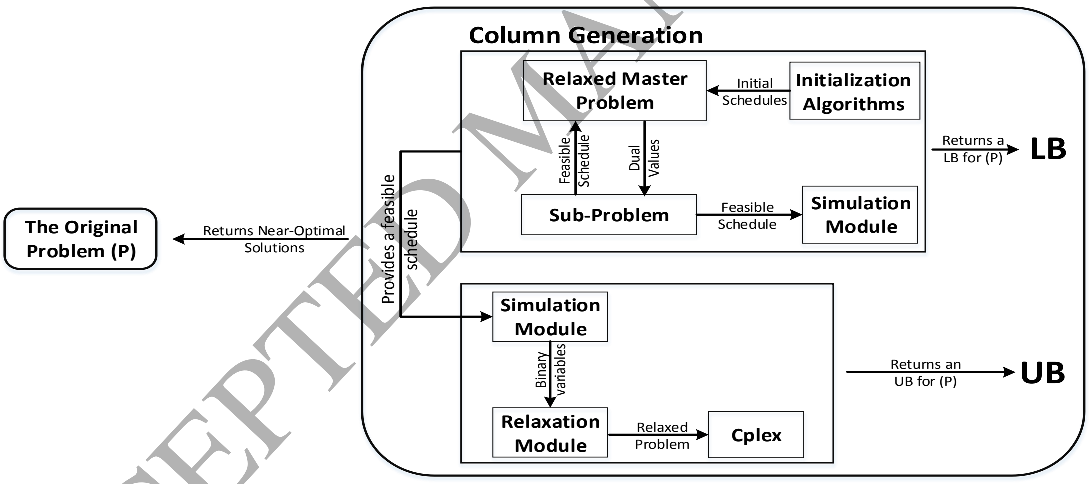

Figure 11: The heuristic toolbox.

5 Computational Results

This section presents computational results for SASP instances with real-world data. Section 5.1 describes the data. Section 5.2 compares the performance of the SASP heuristics of Section 4 against Cplex. Section 5.3 performs sensitivity analyses of the SASP parameters.

5.1 Data Description

For the purposes of this study, the pro-health discussion forum data were retrieved from a website popular among English speakers[1]. The website users maintain the personal pages and online "friend" lists, and discuss health-related topics within medical issue-specific "communities". (See Kaplan and Haenlein [32] to learn more about the characteristics typical of modern online content communities.) The website, with its ontology being similar to other forum-hosting platforms [59], was crawled in 2015 [58]. The user personal profile data were not saved and the usernames were anonymized. The website contains more than a hundred communities; overall, more than a Million users contributed to about two Million threads, consisting of ten Million messages. Figure 12(a) plots the number of messages grouped by day. Figure 12(b) gives a log-log plot of the message counts across different topics (forum threads); the horizontal axis shows the number of topics each of which has the number of messages shown on the vertical axis. The largest number of messages observed in a topic exceeded 10,000.

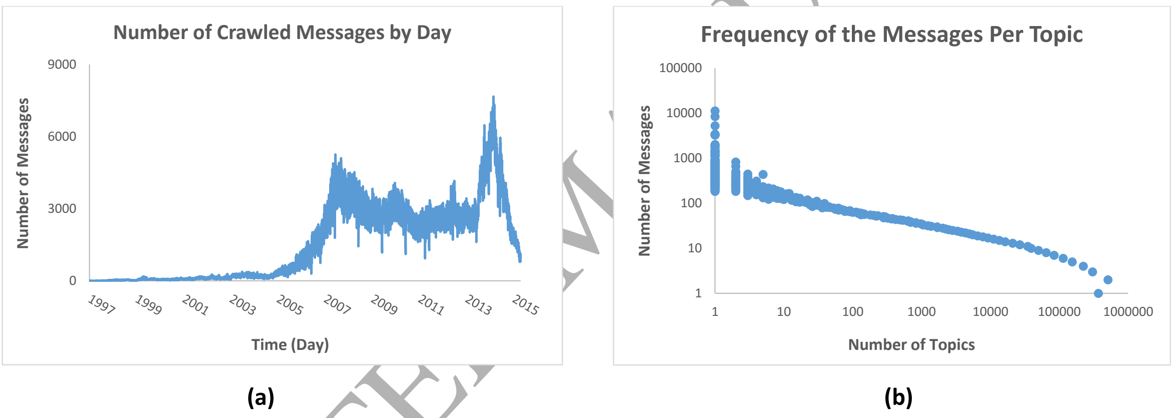

Figure 12: The aggregate patterns based on the crawled website data: (a) the number of messages grouped by day, (b) the number of contributed messages per topic on the logarithmic scales.

The collected data were used to build the social networks of the forum contributors; Appendix D details the algorithm used for the network generation. Note that the built networks may not necessarily represent friendship connections; however, they properly reflect the likely channels of information spread, as appropriate for the SASP. Multiple SASP instances were created with the built networks, with the identified (most active) superusers designated as first-level nodes: indeed, these individuals are most likely to be approached by decision-makers looking to advertise their products, or promote the spread of health behavior-related information.

The density of the edges in the second level network is a factor that affects the SASP solution time

[1] For user privacy purposes, the name of the website is not disclosed.

very much. Recall that simulating the evidence exchange resulting from the activation of a given set of positive seeds takes $O(T|A|)$ time. In order to experimentally explore this dependence, an experiment is run over the network based on the "Stroke" forum: while the SASP parameters remain fixed, the network density is iteratively increased by random edge addition. Figure 13 exhibits an exponential growth of the solution time, as a function of the density of the second level network.

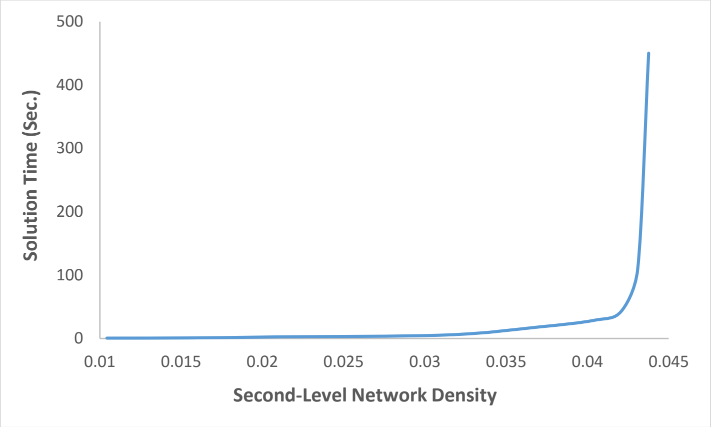

Figure 13: Analyzing impact of second level network density on the SASP solution time.

Note that when the density of the second level network for problem (P) is reduced, the resulting problem (P') is significantly different from (P). However, while the optimal objective value of (P') does not provide any bound for the optimal objective value of (P), the optimal solution (seed activation schedule) of (P') provides a valid feasible solution for (P). This observation is used in Section 5.2 along with the column generation-based algorithm to obtain a tighter lower bound for the SASP.

5.2 Column Generation-Based Heuristic Performance

The presented heuristic toolbox provides both upper and lower bounds for the SASP optimum, and informs us of the quality of the best heuristically obtained solution. However, with the availability of the true optimal value (attainable by Cplex for small- and medium-sized problems), one can evaluate the performance of the algorithms that return the lower and upper bounds separately from each other.

Table 4 presents the results of a computational study with a set of small and medium-sized SASP instances. In each instance, the nodes with out-degree above $\mu + 3\sigma$ (where μ and σ are the mean and standard deviation of out-degree over the instance network) were designated as potential seeds. The Mixed-Integer Program and the Column Generation-based heuristic were implemented using Concert Technology in JAVA and the commercial solver CPLEX 12.6. All the experiments have been performed on a desktop with an Intel(R)Core(TM)i5 3.2GHz processor, 64-bit operating system and 12GB RAM.

Table 4: Computational results with small- and medium-sized SASP problem instances.

Dataset	$\lvert N_1\rvert$	$\lvert N_2\rvert$	T	B	Cplex Sol. (Opt.)	Cplex Time (sec.)	Heu. Sol. LB	UB	Heu. Time (sec.)	Opt. Gap (%)	Heu. Gap (%)
EBV	1	51	7	450	18.25	0.8	18.25	18.25	2.6	0.00	0.00
Alcoholism	2	98	6	350	-29.14	5.61	-29.14	-29.14	7.22	0.00	0.00
Alcoholism	4	131	7	620	51.22	107.38	51.22	53.11	142.56	0.00	3.55
Relationships	6	405	6	1250	624.38	175.24	618.92	642.21	57.9	0.87	3.62
Pregnancy	21	925	6	2450	2486.27	16168.69	2440.24	2507.31	306.28	1.86	2.67

The first column of Table 4 contains (working) dataset name, columns 2 through 5 specify the SASP parameters, and the other columns contain the computational results. For the instances solved by Cplex, the optimality gap is reported, along with the heuristic gap. The results indicate that for small instances of SASP, Cplex outperforms the presented heuristic algorithm, which is expected. Running solution initialization algorithms as well as lower bound and upper bound routines for small SASP problems is a more expensive computation than running Cplex. As the problem size increases, however, the Cplex runtime grows fast while the heuristic runtime grows marginally.

Table 5 reports the results for large SASP instances, which could not be solved by Cplex in the set time of four hours; after this time, a significant optimality gap (> 120%) was still observed, meaning that Cplex was not able to find even one good feasible solution. Meanwhile, the heuristic performs well.

Table 5: Computational results with large-sized SASP problem instances.

Dataset	$\lvert N_1\rvert$	$\lvert N_2\rvert$	T	B	Cplex Sol. (Opt.)	Cplex Time (sec.)	Heu. Sol. LB	UB	Heu. Time (sec.)	Heu. Gap (%)
HIV Prev.	14	3683	6	3250	–	>4 hrs	3631.85	3740.14	1130.35	2.91
Anxiety	8	1295	7	1300	–	>4 hrs	1436.45	1459.74	226.29	1.59
Women's Health	7	3401	7	1400	–	>4 hrs	1073.63	1132.77	480.45	5.21

The presented column generation-based algorithm provides a feasible SASP solution and a tight lower bound on the optimum. Given such a bound, one can exploit the idea of *warm start*, i.e., using the solver's "mipstart" feature, to start with a tight lower bound and make the solver's convergence faster. In fact, the warm start strategy runs Cplex as a heuristic that works on the mixed-integer program until a pre-defined gap is obtained. The performance of this method, compared to that of the heuristic algorithm for the SASP instances, is also reported in Table 6.

Table 6: Comparing the heuristic algorithm and the warm start strategy for finding the upper bound.

Dataset	$\lvert N_1\rvert$	$\lvert N_2\rvert$	T	B	Time Limit (sec.)	Heu. Gap (%)	Warm Start Gap (%)
Relationships	6	405	6	1250	57.9	3.62	1.06
Pregnancy	21	925	6	2450	306.28	2.67	0.60
HIV Prev.	14	3683	6	3250	1130.35	2.91	9.16
Anxiety	8	1295	7	1300	226.29	1.59	8.74
Women's Health	7	3401	7	1400	480.45	5.21	16.93

For a fair comparison between the warm start method and the heuristic algorithm, the time limits for both were set to be equal to the heuristic runtimes as in Tables 4 and 5. The results in Table 6 show that for the medium-sized instances of SASP, the warm start method provides a tighter bound on the optimum; however, it is not the case with the larger problem instances.

5.3 Network Density and Evidence Spread

Another study is performed with the "Stroke" forum network – a small, sparse network of 58 nodes, three of which were designated as superusers. Across multiple SASP instances with the low-density second level network with 31 edges, the positive evidence increment value e^+, $e^+ < e'$, is iteratively increased. In order to investigate the effect of the network density on the optimal SASP values, edges are added randomly to the second level network. The statistics of the original network of the "Stroke" forum users and the revised networks (obtained via edge additions) are presented in Table 7.

Table 7: Statistics of the networks used for density analysis in SASP.

Network	# of edges	Density
Original Net.	31	0.0104
Net. 2	46	0.0154
Net. 3	93	0.0313
Net. 4	155	0.0522

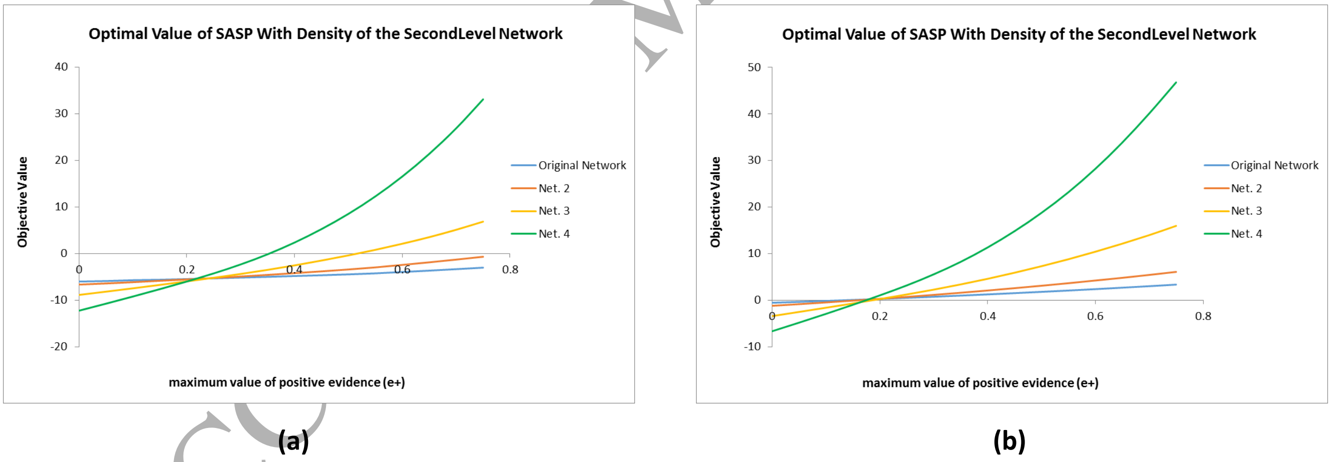

Figure 14: Analyzing impact of second level network density on the evidence spread: (a) low budget (580), (b) high budget (800).

The results of the experiments for two different budget limits, low and high, are presented in Figures 14(a) and 14(b), respectively. The optimal SASP values grow consistently as e^+ is increased: a higher evidence increment allows the positive evidence to spread more effectively. A comparison of the results across the instances with the networks of varied density suggests that the higher the density, the greater the effect of the network structure on the campaign outcome. When the positive party is stronger (weaker) than the opponent, a denser network generates a greater positive (negative) outcome. The objective function dynamics is similar in the experiments, irrespective of the budget limit.

The presented results suggest that a dense social network benefits decision-makers promoting highly appealing products. Otherwise, it is beneficial to the decision-maker to limit the consumer communication on the subject. Note that this is in line with the findings of Easley and Kleinberg [16].

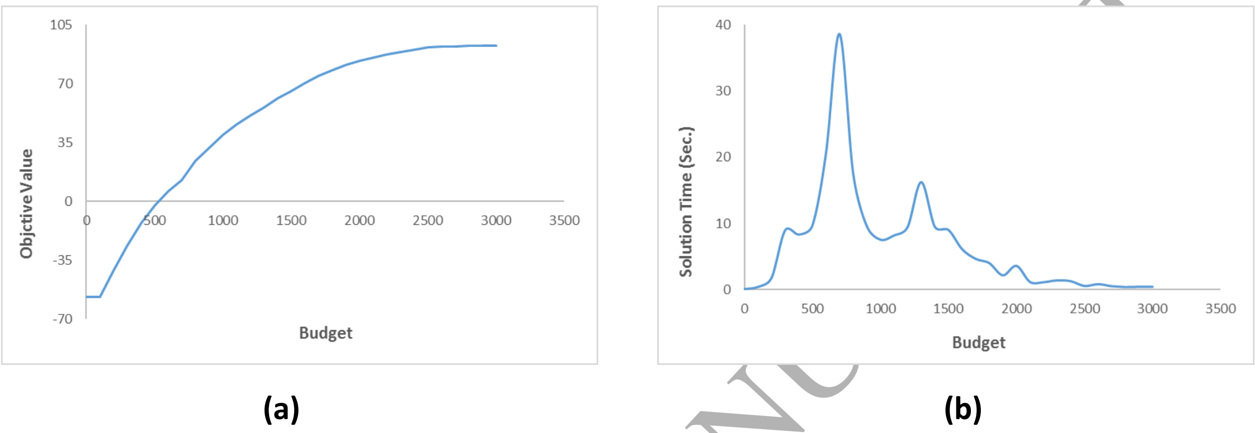

Figure 15: The impact of the second level network density on the evidence spread.

The final computational study is performed over the "Hepatitis B" forum network, with five designated superusers and 98 nodes in the second level network, to investigate how the budget limit impacts the SASP solution value and the runtime to find it. The results are reported in Figures 15(a) and 15(b). As the budget limit increases, the objective value grows up to a certain point. As for the runtime, when the SASP budget limit is low, the problem becomes easy; if the budget is large, any schedule becomes feasible, and the runtime is also low. The highest runtimes are observed with medium budget limits. This observation is consistent with the results in [54].

6 Discussion

This paper solves for an optimal seed activation schedule in IM. In the presented Seed Activation Scheduling Problem (SASP), depending on the budget limitations, the problem time horizon, and/or under positive NPV, the activation of late seeds is shown to help keep the influence cascades alive. The new evidence-based diffusion model, the Partial Parallel Cascade (PPC) model, describes the spread of evidence through networks where nodes can be partially activated, in line with the modern product adoption theories in the experimental marketing literature. The SASP is formulated as a mixed-integer program and solved using Cplex. As the SASP problem under the PPC diffusion model is NP-hard, a heuristic decomposition algorithm is designed using a Column Generation technique that provides both the upper bounds and lower bounds for the optima.

In order to set up realistic SASP experiments, the networks based on the public communication (questions, answers and discussions) in a large healthcare forum website were collected. The presented experiments, motivated by the marketing practices in the blogger-centric domain, reveal how the optimal SASP solutions depend on the type of the objective function. It is shown that the decision-maker has to be particularly careful when pursuing time-dependent goals, e.g., in order to build a momentum before each targeted campaign deadline.

The presented Column Generation-based heuristic is proposed as a means for solving the IM problem efficiently. While the current paper focuses on the SASP under the PPC diffusion model, the column generation method can be applied to any other IM problem, with the sub-problem introducing feasible seed sets, with respect to budget and other problem instance-specific constraints, and the objective function being improved in the relaxed master problem. Furthermore, the idea of column generation bodes well for the situations where a decision-maker works to compose a portfolio of seed activation strategies, to minimize the investment risk through assigning a part of the budget to each strategy.

This paper opens up a new area for modeling two-level diffusion-based optimization problems. An example of another such problem in transportation planning is accepting/rejecting shipment orders considering the impact of the decisions on the future orders and on the diffusion process within the business network of the order owners. Such problems require the advances in both the two-level IM and the dynamic and stochastic knapsack problems [35, 48]. Future research can investigate the stochastic seed activation for IM. Moreover, further research can establish a connection between the presented two-level seed activation scheduling problem and reward-based scheduling problems.

Acknowledgments

This work was supported in part by the National Science Foundation Award 1635611, “Operational Decision-Making for Reach Maximization of Incentive Programs that Influence Consumer Energy-Saving Behavior”. Additionally, while contributing to this work, the last author was supported in part by the 2016 U.S. Air Force Summer Faculty Fellowship Program, sponsored by the Air Force Office of Scientific Research.

References

[1] Beal, G. M., Bohlen, J. M., *et al.,* (1957). *The diffusion process.* Agricultural Experiment Station, Iowa State College.

[2] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature 489*(7415), 295–298.

[3] Bone, P. F. (1995). Word-of-mouth effects on short-term and long-term product judgments. *Journal of business research 32*(3), 213–223.

[4] Brown, J., Broderick, A. J., and Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing 21*(3), 2–20.

[5] Budak, C., Agrawal, D., and El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pp. 665–674. ACM.

[6] Cao, Z., Chen, X., and Wang, C. (2013). How to schedule the marketing of products with negative externalities. In *Computing and Combinatorics*, pp. 122–133. Springer.

[7] Chen, W., Lakshmanan, L. V., and Castillo, C. (2013). Information and influence propagation in social networks. *Synthesis Lectures on Data Management 5*(4), 1–177.

[8] Chen, W., Lu, W., and Zhang, N. (2012). Time-critical influence maximization in social networks with time-delayed diffusion process. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI11)*, pp. 592–598.

[9] Chen, W., Wang, C., and Wang, Y. (2010). 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.

[10] Chen, W., Wang, Y., and 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.

[11] Chen, Y. and Krause, A. (2013). Near-optimal batch mode active learning and adaptive submodular optimization. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 160–168.

[12] Chierichetti, F., Kleinberg, J., and Panconesi, A. (2014). How to schedule a cascade in an arbitrary graph. *SIAM Journal on Computing 43*(6), 1906–1920.

[13] Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine 357*(4), 370–379.

[14] Desaulniers, G., Desrosiers, J., and Solomon, M. M. (2006). *Column generation*, Volume 5. Springer Science & Business Media.

[15] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 57–66. ACM.

[16] Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world.* Cambridge University Press.

[17] Fan, L., Wu, W., Zhai, X., Xing, K., Lee, W., and Du, D.-Z. (2014). Maximizing rumor containment in social networks with constrained time. *Social Network Analysis and Mining 4*(1), 1–10.

[18] Fan, T.-K. and Chang, C.-H. (2011). Blogger-centric contextual advertising. *Expert Systems with Applications 38*(3), 1777–1788.

[19] Feige, U. (1998). A threshold of ln n for approximating set cover. *Journal of the ACM (JACM) 45*(4), 634–652.

[20] Galvão, R. D. (1981). A note on garfinkel, neebe and rao's lp decomposition for the p-median problem. *Transportation Science 15*(3), 175–182.

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

[22] Golovin, D. and Krause, A. (2010). Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *arXiv preprint arXiv:1003.3967*.

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

[24] Goyal, A., Bonchi, F., Lakshmanan, L. V., and Venkatasubramanian, S. (2012). On minimizing budget and time in influence propagation over social networks. *Social Network Analysis and Mining*, 1–14.

[25] Goyal, A., Bonchi, F., Lakshmanan, L. V., and Venkatasubramanian, S. (2013). On minimizing budget and time in influence propagation over social networks. *Social Network Analysis and Mining 3*(2), 179–192.

[26] Goyal, A., Lu, W., and 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.

[27] Hajiaghayi, M., Mahini, H., and Malec, D. (2014). The polarizing effect of network influences. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 131–148. ACM.

[28] Hajiaghayi, M., Mahini, H., and Sawant, A. (2013). Scheduling a cascade with opposing influences. In *Algorithmic Game Theory*, pp. 195–206. Springer.

[29] Huckfeldt, R. and Sprague, J. (1987). Networks in context: The social flow of political information. *American Political Science Review 81*(04), 1197–1216.

[30] Inkpen, A. C. and Tsang, E. W. (2005). Social capital, networks, and knowledge transfer. *Academy of management review 30*(1), 146–165.

[31] Kalish, S. (1985). A new product adoption model with price, advertising, and uncertainty. *Management science 31*(12), 1569–1585.

[32] Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons 53*(1), 59–68.

[33] Kempe, D., Kleinberg, J., and 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.

[34] Khuller, S., Moss, A., and Naor, J. S. (1999). The budgeted maximum coverage problem. *Information Processing Letters 70*(1), 39–45.

[35] Kleywegt, A. J. and Papastavrou, J. D. (1998). The dynamic and stochastic knapsack problem. *Operations Research 46*(1), 17–35.

[36] Kozinets, R., Wojnicki, A. C., Wilner, S. J., and De Valck, K. (2010). Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing, March*.

[37] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). 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.

[38] Li, Y., Chen, W., Wang, Y., and Zhang, Z.-L. (2013). Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 657–666. ACM.

[39] Liu, B., Cong, G., Xu, D., and Zeng, Y. (2012). Time constrained influence maximization in social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pp. 439–448. IEEE.

[40] Lu, L.-C., Chang, W.-P., and Chang, H.-H. (2014). Consumer attitudes toward bloggers sponsored recommendations and purchase intention: The effect of sponsorship type, product type, and brand awareness. *Computers in Human Behavior 34*, 258–266.

[41] Mahajan, V., Muller, E., and Sharma, S. (1984). An empirical comparison of awareness forecasting models of new product introduction. *Marketing Science 3*(3), 179–197.

[42] Martin, T., Schoenebeck, G., and Wellman, M. (2014). Characterizing strategic cascades on networks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 113–130. ACM.

[43] Mochalova, A. and Nanopoulos, A. (2015). Multi-stage seed selection for viral marketing. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pp. 1181–1183. ACM.

[44] Newton, J. D., Klein, R., Bauman, A., Newton, F. J., Mahal, A., Gilbert, K., Piterman, L., Ewing, M. T., Donovan, R. J., and Smith, B. J. (2015). The move study: a study protocol for a randomised controlled trial assessing interventions to maximise attendance at physical activity facilities. *BMC public health 15*(1), 1.

[45] Nieves, J. and Osorio, J. (2013). The role of social networks in knowledge creation. *Knowledge Management Research & Practice 11*(1), 62–77.

[46] Ozanne, U. B. and Churchill Jr, G. A. (1971). Five dimensions of the industrial adoption process. *Journal of Marketing Research*, 322–328.

[47] Pal, S. K., Kundu, S. K., and Murthy, C. (2014). Centrality measures, upper bound, and influence maximization in large scale directed social networks. *Fundam. Inform. 130*(3), 317–342.

[48] Papastavrou, J. D., Rajagopalan, S., and Kleywegt, A. J. (1996). The dynamic and stochastic knapsack problem with deadlines. *Management Science 42*(12), 1706–1718.

[49] Reese, J. (2006). Solution methods for the p-median problem: An annotated bibliography. *Networks 48*(3), 125–142.

[50] Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 61–70. ACM.

[51] Rogers, E. M. (2010). *Diffusion of innovations.* Simon and Schuster.

[52] Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., and Benevenuto, F. (2012). Finding trendsetters in information networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1014–1022. ACM.

[53] Samadi, M. (2016). *Optimal Strategies for Controlling Cascades in Social Networks: An Influence Maximization Approach.* Ph. D. thesis, State University of New York at Buffalo.

[54] Samadi, M., Nikolaev, A., and Nagi, R. (2016a). A subjective evidence model for influence maximization in social networks. *Omega 59*, 263–278.

[55] Samadi, M., Nikolaev, A., and Nagi, R. (2016b). The temporal aspects of the evidence-based influence maximization on social networks. *Optimization Methods and Software*, 1–22.

[56] Sangachin, M. G., Samadi, M., and Cavuoto, L. A. (2014). Modeling the spread of an obesity intervention through a social network. *Journal of Healthcare Engineering 5*(3), 293–312.

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

[58] Semenov, A. (2013). Principles of social media monitoring and analysis software.

[59] Semenov, A. and Veijalainen, J. (2013). A modelling framework for social media monitoring. *International Journal of Web Engineering and Technology 8*(3), 217–249.

[60] Singer, Y. (2012). How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 733–742. ACM.

[61] Trusov, M., Bucklin, R. E., and Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of marketing 73*(5), 90–102.

[62] Webster Jr, F. E. (1969). New product adoption in industrial markets: a framework for analysis. *The Journal of Marketing*, 35–39.

[63] Zhang, B., Teng, J., Bai, X., Yang, Z., and Xuan, D. (2011). P 3-coupon: A probabilistic system for prompt and privacy-preserving electronic coupon distribution. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*, pp. 93–101. IEEE.

Appendix A. Initialization Algorithms for the Presented Column Generation Method

The column generation algorithm presented for SASP in Section 4 needs a set of initial solutions (columns) to begin with. The quality of the initial solutions may impact the solution time for the column generation module of the heuristic toolbox. This appendix details four initialization algorithms that generate such initial solutions exploiting the properties of SASP.

Algorithm 1 - The Time Priority Initialization Algorithm

```
$S^+ = 0$;     /* Initializes the schedule with zero*/
$R'_t = R_t$;     /* Initializes the remaining budget at time $t$ with the allocated budget constraint at $t$ */
$R'' = B$;     /* Initializes the remaining budget with the maximum seed activation budget */
for  t ← 0 to $T$  do
    $C_{jt} = NPV(C_j, t)$ /* Updates the cost of seed activation considering the net present value of money*/
    Initialize $f_t$ /* Updates the fixed cost of seed activation considering the net present value of money*/
    if $t = 0$  then
        A = 1 /* Sets the coefficient of the fixed cost of seed activation */
    else do
        A = 0 /* Sets the coefficient of the fixed cost of seed activation */
    end if
    for  j ← 1 to $|N_1|$  do
        if $R'_t \geq C_{jt} + Af_{jt}$  then
            if $R'' \geq C_{jt} + Af_{jt}$  then
                $S^+_{jt} = 1$     /*schedules the activation of ode $j$ at time $t$*/
                $R'' - = C_{jt} + Af_{jt}$     /*updates the total remaining budget*/
                $R'_t - = C_{jt} + Af_{jt}$     /*updates the remaining budget at the current time period*/
            end if
        else do
            Break;     /*stops searching at the current time period and goes to $t + 1$, if possible*/
        end if
    end for
    if $R'' \leq \min_{j \in N_1} C_{jt}$  then
        Break;     /*stops searching for more activations*/
    end if
end for
Return $S^+$
```

The Time Priority Initialization algorithm, named Algorithm 1, starts with the initial time period and activates seeds in the order of the node index until the budget constraint is met, and then, moves to the next time period. The initialization algorithm needs to include the seed activation fixed cost only in the initial time period ($t = 0$); if node $j \in N_2$ is selected at time t, the algorithm definitely has selected it at time $t - 1$, which removes the need for paying the fixed cost of seed activation after the initial time period. After selecting each seed for activation, the algorithm compares the remaining budget against the minimum seed activation cost and decides whether to continue searching for more seeds or stop. To include NPV in the seed activation process, Algorithm 1 works with C_{jt} as the NPV of the cost for activating node $j \in N_1$ at time t.

The Knapsack initialization algorithm, named Algorithm 2, on the other hand, keeps the time priority and selects the seeds within each time period based on their values. The value function for each potential seed $j \in N_1$ is calculated as the unit cost of influencing each node on the second level network through seed j; the algorithm gives a priority to the seeds with lower unit cost. Algorithm 2 focuses on selecting the seeds with a greater number of followers and tries to spend the budget more strategically.

Algorithm 2 - The Knapsack Initialization Algorithm

```
$S^{+} = 0$;    /* Initializes the schedule with zero*/
$R'_t = R_t$;    /* Initializes the remaining budget at time $t$ with the allocated budget constraint at $t$*/
$R'' = B$;    /* Initializes the remaining budget with the maximum seed activation budget*/
for  j ← 1 to $|N_1|$  do
    $V_j = \frac{C_j}{\text{outDegree}_j}$;    /* Evaluates the unit cost of node $j$ to influence each regular node*/
end for
SortV();    /* This function runs an ascending sort based on V value of seeds*/
for  t ← 1 to $T$  do
    $C_{jt} = NPV(C_j, t)$ /* Updates the cost of seed activation considering the net present value of money*/
    Initialize $f_t$ /* Updates the fixed cost of seed activation considering the net present value of money*/
    if $t = 0$  then
        A = 1 /* Sets the coefficient of the fixed cost of seed activation */
    else do
        A = 0 /* Sets the coefficient of the fixed cost of seed activation */
    end if
    for  j ← 1 to $N_1$  do
        if $R'_t \geq C_{jt} + Af_{jt}$  then
            if $R'' \geq C_{jt} + Af_{jt}$  then
                $S^{+}_{jt} = 1$    /*schedules the activation of ode $j$ at time $t$*/
                $R'' -= C_{jt} + Af_{jt}$    /*updates the total remaining budget*/
                $R'_t -= C_{jt} + Af_{jt}$    /*updates the remaining budget at the current time period*/
            end if
        else do
            Break;    /*stops searching at the current time period and goes to $t + 1$, if possible*/
        end if
    end for
    if $R'' \leq \min_{j \in N_1} C_{jt}$  then
        Break;    /*stops searching for more activations*/
    end if
end for
Return $S^{+}$
```

Algorithms 1 and 2 are designed with the idea that seed activation timing is the key driver of the SASP complexity. The computational results of solving (P) on small social networks showcase a trade-off between the time and seed values in the optimal seed activation schedule. Time and Connectivity Trade-off Algorithm, called Algorithm 3, is designed to exploit this trade-off.

The algorithm assigns a value to each node $j \in N_1$ that can be selected as a seed at the initial time period, based on its out-degree and cost; larger out-degree and lower seed cost are preferred. In order to make the value calculation fair, the value of a node at each time period is discounted based on time horizon T and standardizing parameter δ so that each node has the greatest value at the initial time period. Considering time as an important element in the seed activation, the algorithm prefers earlier activation of a seed, because in this case, the seed has more time to spread evidence over the network.

All the nodes at the first level network are considered as viable candidates to become seeds. The algorithm compares the value of all these nodes at the initial time period and selects the one with a maximum value. If node $j \in N_1$ is selected at the initial time period, the earliest time that node j can become available as a seed again is time $t = 1$. As such, the algorithm updates the availability of node j and discounts its value. The algorithm also updates the remaining budget after each new seed selection. In the next iteration, the value of all the nodes at their earliest availability are compared.

The algorithm continues scheduling nodes' activation until the remaining budget becomes insufficient for activating the node with a maximum value. Algorithm 3 allows a potential seed with a high value (compared to other available candidates) to be selected multiple times before selecting other nodes; if the values of some nodes turn out to be close to each other, then the algorithm emphasizes earliness for the activation.

Algorithm 3 - Time and Connectivity Trade-off Algorithm

```
$S^+ = 0$;    /* Initializes the schedule with zero*/
$R'_t = R_t$;    /* Initializes the remaining budget at time $t$ with the allocated budget constraint at $t$*/
$R'' = B$;    /* Initializes the remaining budget with the maximum seed activation budget */
for  j ← 1 to $|N_1|$  do
    $V_{j1} = \frac{C_j}{\text{outDegree}_j}$;    /* Evaluates the unit cost of node $j$ to influence each regular node*/
    $V_{j2} = 0$;    /* Assigns the next available time period for each node */
    $V_{j3} = 1$;    /* Assigns the coefficient of fixed cost for each node */
end for
Sort(V,1);    /* This function sorts $V$ by values of the first column descendingly*/
while  $R'' > \min_{j \in N_1} C_{jt}$  do
    for  j ← 1 to $|N_1|$  do
        if $R'_t \geq C_{V_{j1}V_{j2}} + V_{j3} f_{V_{j1}V_{j2}}$  then
            if $R'' \geq C_{V_{j1}V_{j2}} + V_{j3} f_{V_{j1}V_{j2}}$  then
                $S^+_{V_{j1}V_{j2}} = 1$    /*schedules the activation of node $V_{j1}$ at time $V_{j2}$*/
                $R'' - = C_{V_{j1}V_{j2}} + V_{j3} f_{V_{j1}V_{j2}}$    /*updates the total remaining budget*/
                $R'_t - = C_{V_{j1}V_{j2}} + V_{j3} f_{V_{j1}V_{j2}}$    /*updates the remaining budget at the current time period*/
                if $V_{j1} < T$  then
                    $V_{j1} = V_{j1} * \frac{T-1}{T} * \delta$
                    $V_{j2} = V_{j2} + 1$
                    $V_{j3} = 0$
                else do
                    $V_{j1} = 0$
                end if
                Sort(V,1)
                Break;    /*Stops running the for loop and jumps to the while loop*/
            else do
                Continue;    /*Jumps to index $j+1$ in the for loop, if possible*/
            end if
        else do
            Continue;    /*Jumps to index $j+1$ in the for loop, if possible*/
        end if
    end for

end while
Return $S^+$
```

The idea of random seed activation is exploited in the fourth initialization algorithm. The Random Initialization Algorithm – Algorithm 4 – randomly selects a time period and the index of a node to be activated, and continues as far as the budget constraint is not violated. Such random selection, however, is not sensitive to the fixed cost: a seed selected at time t can be later selected for time $t-1$. To tackle this issue, when a seed, selected at time t, has already been selected at time periods $t-1$ and $t+1$, the algorithm adds the paid fixed cost at time $t+1$ to the remaining budget. Otherwise, if the seed selected at time t is also activated at time $t-1$, then, the fixed cost is ignored. On the other hand, if the node selected at time t has been already selected to be activated at time $t+1$, the fixed cost at time $t+1$ is added to the remaining budget and the fixed cost at time t is deducted from the remaining budget. Note that here, the fixed cost of seed activation is a fixed number, independent of the seed index, while the fixed cost of seed activation changes over time as a result of applying NPV. The random seed activation procedure continues till the remaining budget falls below the summation of the minimum seed activation cost and fixed costs. As such, it is possible for Algorithm 4 to stop when it is still possible to add more seeds; this possibility, however, is ignored because the algorithm is just providing a random initial solution for the column generation – this solution does not have to be the best possible one.

Algorithm 4 - The Random Initialization Algorithm

```
$S^+ = 0$;    /* Initializes the schedule with zero*/
$R'_t = R_t$;    /* Initializes the remaining budget at time $t$ with the allocated budget constraint at $t$*/
$R'' = B$;    /* Initializes the remaining budget with the maximum seed activation budget*/
for  t ← 1 to $T$  do
    Initialize $C_t$    /* Updates the cost of seed activation considering the net present value of money*/
    Initialize $f_t$    /* Updates the fixed cost of seed activation considering the net present value of money*/
end for
while  $R'' \geq \min_{j \in N_1}(C_{jt} + f_{jt})$ to $T$  do
    t = Rand(T);    /* Selects a random time period $0 \leq t \leq T$*/
    j = Rand($|N_1|$);    /* Selects a random seed $j \in N_1$*/
    if $R' \geq C_{jt} + f_{jt}$  then
        Continue();    /* Tries to find another combination of $t$ and $j$*/
    else do
        if $S^+_{jt}$ already selected  then
            Continue();    /* Tries to find another combination of $t$ and $j$*/
        else
            $S^+_{jt} = 1$;
            if $S^+_{jt-1} = S^+_{jt+1} = 1$  then
                $R'' - = C_{jt} - f_{jt+1}$;    /* Deducts the seed activation cost from and adds the fixed cost of time $t+1$ to the remaining budget*/
                $R'_t - = C_{jt} - f_{jt+1}$;    /* Deducts the seed activation cost from and adds the fixed cost of time $t+1$ to the remaining budget of time $t$*/
            else if $S^+_{jt-1} = 1$  then
                $R'' - = C_{jt}$;    /* Deducts the seed activation cost from the remaining budget*/
                $R'_t - = C_{jt}$;    /* Deducts the seed activation cost from the remaining budget of time $t$*/
            else if $S^+_{jt+1} = 1$  then
                $R'' - = C_{jt} + f_{jt} - f_{jt+1}$;    /* Deducts the seed activation and fixed cost from the remaining budget and adds the fixed cost of time $t+1$*/
                $R'_t - = C_{jt} + f_{jt} - f_{jt+1}$;    /* Deducts the seed activation and fixed cost from the remaining budget of time $t$ and adds the fixed cost of time $t+1$*/
            else
                $R'' - = C_{jt} + f_{jt}$;    /* Deducts the seed activation and fixed cost from the remaining budget*/
                $R'_t - = C_{jt} + f_{jt}$;    /* Deducts the seed activation and fixed cost from the remaining budget of time $t$*/
            end if
        end if
end while
Return $S^+$
```

Appendix B. SASP Complexity

Consider the case of SASP where a hire of any seed has a unit cost. As the SASP under the PPC diffusion model has a budget constraint for seed selection over time, and relaxes the unit cost constraint of the OIMP, it is natural to reduce the Budgeted Maximum Coverage Problem (BMCP), rather than the MCP, to it. The BMCP is known to be NP-hard [34]. From the budgeting point of view, one can see that the relationship between BMCP and MCP is similar to the relationship between the IM problem under the PC diffusion model [54] and that under the PPC diffusion model. However, the SASP under the PPC diffusion model is a multi-period sequential seed selection problem, which makes it much harder than the IM problem under the PC diffusion model and the reason for this increased complexity lies in the cost that the decision-maker pays to better control the cascade over time.

Proof of Theorem 1. SASP under PPC diffusion model is NP-hard by a polynomial Turing reduction from the Maximum Coverage Problem (MCP). The MCP selects a group of sets from a number of given sets that may have common elements to maximize a resulting total number of unique selected elements; MCP is NP-hard [19]. MCP is first formally stated, and then, the reduction from SASP to MCP is presented.

MCP INSTANCE: A number $k > 0$ and a collection of sets $V = \{V_1, V_2, ..., V_{m_1}\}$.

MCP OBJECTIVE: Find a subset $V' \subseteq J$ such that $|V'| \leq k$ and the number of covered elements $|\bigcup_{V_j \in V'} V_j|; j \in N_1$ is maximized.

Given an arbitrary MCP instance, define a particular instance of SASP as follows. Assume $T = 1$, $|N_1| = m_1$, $|N_2| = m_2$, $B = k$, $G_1 = G_2 = 1$ $e'' = 0$ and $C_j = 1$ for each node $j \in N_1$. Let $e^+ > \max \theta_i^+; i = 1, 2, ..., m_2$, $e^- = 0$, $f = 0$, and set $\gamma = \beta^+ = \beta^- = 1$. Define set V_j for $j = 1, 2, ..., m_1$ such that $v \in V_j$ iff $(j, i) \in A_1, i = 1, 2, ..., |N_2|$ (all the nodes one hop away from j). This transformation can be performed in polynomial time in the size of the arbitrary MCP instance.

In order to show that an optimal solution to SASP maps to an optimal solution to MCP, let X_{j0}^* for $j = 1, 2, ..., |N_1|$ ($X_{j0} \in \{0, 1\}$) be an optimal solution to SASP. Then, $\sum_{i=1}^{|N_1|} C_j X_{j0} \leq B$, $Y_{it} = 0$ for $i = 1, 2, ..., |N_2|, t = 0, 1, ..., T$ and $\sum_{i=1}^{|N_2|} \sum_{t=0}^{T} (X_{it} - Y_{it})$ is maximized. The claim is that X_{j0}^* is an optimal solution to MCP. Note that X_{j0}^* for $j = 1, 2, .., |N_1|$ is a feasible solution to MCP because $\sum_{j=1}^{|N_1|} C_j X_{j0} \leq B = k$.

Suppose there exists such a solution to MCP, $\bar{X}_{j0}$, for $j = 1, 2, .., |N_1|$ that $|\bigcup_{V_j \in \bar{V}'} V_j| > |\bigcup_{V_j \in V'^*} V_j|$. Solution $\bar{X}_{j0}$ for $j = 1, 2, .., |N_1|$ is a feasible solution to SASP: $\sum_{j=1}^{|N_1|} C_j \bar{X}_{j0} \leq B = k$. Therefore, the SASP objective function for this solution is $\sum_{i=1}^{|N_2|} \sum_{t=0}^{T} (\bar{X}_{it} - \bar{Y}_{it}) = |\bigcup_{V_j \in \bar{V}'} V_j| + k > |\bigcup_{V_j \in V'^*} V_j| + k = \sum_{i=1}^{|N_2|} \sum_{t=0}^{T} (X_{it}^* - Y_{it}^*)$, which is a contradiction. Thus, X_{j0}^* for $j = 1, 2, .., |N_1|$ is an optimal solution to MCP. ■

Proof of Proposition 1. Assume that for a given set of parameters, the optimal solution of the OIMP (S_o^+) dominates the optimal solution of the SASP (S_s^+) with the same set of parameters. The objective functions in both the SASP and original IM are identical; let the objective value corresponding to solution sets S_o^+ and S_s^+ be O_o^* and O_s^*, respectively. Each feasible solution to the OIMP is necessarily feasible for the SASP, i.e., satisfies the total budget and allocated budget constraints. If $O_o^* > O_s^*$, then there is a feasible solution for SASP – S_o^+ – that provides a greater objective value, and hence, S_s^+ is not the optimal solution of the SASP. This leads to a contradiction, which completes the proof. ■

Appendix C. Big M Values for the Mathematical Program (P)

Three sets of large positive numbers (M_{it}, M'_{it} and M''_{it}) are defined for constraint group (I) of the mathematical model (P). These numbers are chosen, and hence, labeled differently in the different constraints, to facilitate the presentation. In this appendix, the minimum acceptable big M values that guarantee the correctness of (P) are derived.

With I_i denoting the in-degree of node i in the second level network, these numbers are bounded from below,

$$M_{it} \geq \left(\frac{e^- I_i + e''}{Min\{\theta_i^+, \theta_i^-\}}\right)\left(\frac{1-\beta_2^{t-1}}{1-\beta_2}\right) \qquad i \in N_2, \quad t = 0,1,...,T, \tag{52}$$

$$M'_{it} \geq \left(e^- I_i + e''\right)\left(\frac{1-\beta_2^{t-1}}{1-\beta_2}\right) \qquad i \in N_2, \quad t = 0,1,...,T, \tag{53}$$

$$M''_{it} \geq \left(e^- I_i + e''\right)\left(\frac{1-\beta_2^{t-1}}{1-\beta_2}\right) - \theta_i^- \qquad i \in N_2, \quad t = 0,1,...,T. \tag{54}$$

In order to ensure that constraints (2), (4), (6) and (7) always correctly enforce the PPC diffusion process, the following conditions need to be met, respectively:

$$M_{it} \geq \frac{K_{it} - L_{it}}{\theta_i^+} \qquad i \in N_2, \quad t = 0,1,...,T, \tag{55}$$

$$M_{it} \geq \frac{K_{it} - L_{it}}{\theta_i^-} \qquad i \in N_2, \quad t = 0,1,...,T, \tag{56}$$

$$M'_{it} \geq K_{it} - L_{it} \qquad i \in N_2, \quad t = 0,1,...,T, \tag{57}$$

$$M''_{it} \geq K_{it} - L_{it} - \theta_{it}^- \qquad i \in N_2, \quad t = 0,1,...,T. \tag{58}$$

The right hand sides in (55)-(58) are maximized when K_{it} attains its maximum value and L_{it} is zero. The maximum value of K_{it} is achieved if all the in-neighbors of node i have been negatively activated over the time periods 0 through $t-1$. Node i starts off in a neutral state, with $K_{i0} = 0$; from the period $t = 1$ onward, the maximum amount of new negative evidence that node i receives at each time period is $I_i e^- + e''$. At the end of each time period, the newly received negative evidence is added to K_{it}, while the previously collected negative evidence is discounted by β_2. Thus, the maximum value of K_{it} can be calculated as the summation of a geometric series,

$$K_{it} \leq \beta_2^{t-2}(I_i e^- + e'') + \beta_2^{t-3}(I_i e^- + e'') + ... + \beta_2(I_i e^- + e'') + (I_i e^- + e'') = \left(\frac{1-\beta_2^{t-1}}{1-\beta_2}\right)(I_i e^- + e''). \tag{59}$$

From (59), one directly obtains (52)-(54).

Appendix D. Network Generation Algorithm for Thread Contributors

This appendix provides the algorithm that was used to generate networks from thread contribution data.

Algorithm 5 - Network Generation Algorithm for Thread Contributors

```
Crawl($i$);    /* Crawls all the messages in thread $i$ */
Sort($i$);    /* Sorts all the crawled messages in ascending order by posting date */
$root = i(0)$;    /* Assigns the first message to the root of the message tree */
for each  message in thread $i$  do
    $c_1$ = user(message)    /* Stores the user of the current message as the current user */
        if message is marked as “reply to”  then
            $c_2$ = user(reply)    /* Stores the user whose message has been replied */
            $c_2 \leftarrow c_1$    /* Creates a communication edge */
            Engagement($c_2 \leftarrow c_1$)    /* calculates the engagement score of the edge */
        else
            FindLeaves()    /* Finds all the unanswered leaves of the tree */
            Edge()    /* Creates edge from the unanswered leaves to the current node */
            Engagement()    /* calculates and assigns the engagement score to the extracted edges */
        end if
end for
Aggregate();    /* Aggregates the edge observations */
ExtractNetwork();    /* Applys the minimum threshold over the network and keep the strong edges */
Return network
```