A Subjective Evidence Model for Influence Maximization in Social Networks

Mohammadreza Samadi[a], Alexander Nikolaev[a], Rakesh Nagi[b]

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

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

msamadi@buffalo.edu, anikolae@buffalo.edu, nagi@illinois.edu

June 27, 2015

Abstract

This paper introduces the notion of subjective evidence, which fuels a new parallel cascade influence propagation model. The model sheds light on the phenomena of belief reinforcement and viral spread of innovations, rumors, opinions, etc., in social networks. Network actors are assumed to be testing a Bayesian hypothesis, e.g., for making judgment about the superiority of some product(s) or service(s) over others, or (dis)utility of a given program/policy. The model-based influence maximization solutions inform the strategies for market niche selection and protection, and identification of susceptible groups in political campaigning. The NP-Hard problem of influential seed selection is first solved as a mixed-integer program. Second, an efficient Lagrangian Relaxation heuristic with guaranteed bounds is presented. In small, medium and large-scale computational investigations, we analyze: (1) how the success of an influence cascade triggered in a (sub)community, long exposed to an opposite belief, depends on the structural properties of the underlying social network, (2) to what extent growing (increasing the density of) a consumer network within a market niche helps a company protect the niche, (3) given a competitor's strength, when a company should counter the competitor on "their turf", and when and how it should look for limited-time opportunities to maximally profit before eventually surrendering the market.

Keywords: influence maximization, social networks, Bayesian inference, evidence, seed selection.

1 Introduction and Motivation

People tend to view product recommendations received from friends or through friends more favorably compared to advertisements offered by commercial mass media channels (Chen *et al.,* 2010; Hinz *et al.,* 2011). Social connections enable the propagation of ideas, judgments and opinions; the phenomenon where knowledge transfer between individuals significantly affects their decisions about purchasing a product is known as social influence/contagion (Van den Bulte and Lilien 2001; Tang *et al.,* 2009; Chen *et al.,* 2013). Social influence and diffusion of innovations in social networks are mainly explored in managerial and sociological studies (Wejnert 2002; Angst *et al.,* 2010; Aral *et al.,* 2011). However, the need for simulating information diffusion/peer influence in social networks and solving optimization problems to algorithmically find potent success of cascade initiation strategies led to the introduction of the Influence Maximization (IM) problem. The objective of the IM problem is to find such early

starters, termed seeds, for influence spread in a social network that will direct information transfer so as to achieve a desired impact on the expected product adoption, or people's decisions/judgments/opinions with respect to a query of interest (Kempe *et al.,* 2003; Chen *et al.,* 2009).

Early mathematical formulations of the IM problem in social networks view social ties as indicators of dyadic dependence, where the random graph or Markov random field-based approach is a natural choice for model design (Domingos and Richardson 2001; Richardson and Domingos 2002). More recent literature on the algorithmic analysis of influence spread has been dominated by diffusion-based models (Kempe *et al.,* 2003), in which ties are viewed as information flow channels. The Independent Cascade (IC) and Linear Threshold (LT) models are most notable ones, both allowing for elegant discrete optimization problem statements; these models also provided the basis for a streak of subsequent studies (Kempe *et al.,* 2005; Goyal *et al.,* 2011; Wang *et al.,* 2012; Dinh *et al.,* 2014).

Application-wise, diffusion models have been found suitable for research studies in marketing (Arthur *et al.,* 2009; Chen *et al.,* 2010) and health care (Sangachin *et al.,* 2014). However, algorithmic investigations up to date failed to culminate in significant managerial insights and strategies. This is in part due to the fact that existing models do not specify the medium and nature of influence flow through a network, i.e., fail to explain the diffusion of *what* leads to social influence, and *how* it does so.

This paper takes a previously unexplored approach to modeling the spread of competitive influence in social networks, rooted in Bayesian Inference theory and focused on propagation of *evidence*. Bayesian inference logic helps quantify social influence under the premise that people treat new information as evidence and update their beliefs in support of or against the null hypothesis. In this approach, network nodes represent intelligent agents (actors) who seek to form judgments about a product/query by testing a relevant hypothesis (e.g., that a particular claim is true), based on their prior beliefs as well as the knowledge acquired through friends. A node's decision to significantly favor the null hypothesis signals the node's "positive activation"; significantly favoring the alternative implies "negative activation"; finally, whenever the collected evidence is inconclusive, the node is labeled "inactive".

This paper presents a Parallel Cascade (PC) diffusion framework for modeling evidence spread through social networks. The flow of information in this PC model is classified as parallel duplication in the typology of flow processes on social networks, introduced by Borgatti (2005), which supports the idea of belief reinforcement through subjective evidence duplication in social communication. The paper reports insightful observations, e.g., pertinent to the identification of penetrable market niches and convenient points of initial influence for conquering new market segments, obtained from solving basic instances of the PC model-based IM (PCIM) problem. The paper develops problem-specific optimization schemes for handling medium and large-scale instances of PCIM problem formulated as a Mixed-Integer program.

The rest of this paper is organized as follows. Section 2 reviews the literature on diffusion models for IM. Section 3 formally introduces the PC diffusion model, formulates PCIM problem and discusses its application to two empirical case studies. Offering a more computationally efficient approach to the problem, Section 4 presents a Lagrangian Relaxation heuristic tool suit, with solution quality guarantees achieved via two problem-specific heuristics for finding lower bounds for PCIM problem optima. Section 5 reports on the conducted experimental studies. Section 6 summarizes the findings, discusses the potential applications of the proposed methods and outlines future research directions. The paper contains two appendices: Appendix A presents the NP-hardness proof for the PCIM problem; Appendix B details the Subgradient Search algorithm for finding an upper bound for PCIM problem optima.

2 The Landscape of the Social Influence Research Domain

The concept of word-of-mouth has received attention in the 1940's as an effective way for diffusion of information (e.g., about new products) and soon became a coined term in the experimental marketing research (Merton 1947; Whyte Jr 1954; Katz 1957). Models of information diffusion over networks, also first introduced in the marketing field, were developed more recently and found use in health care (Newman 2002), sociology (Macy 1991; Valente *et al.,* 1994) and politics (Deroıan 2002). From an experimental point of view, the phenomenon of social contagion is known to be a significant factor affecting the strength of diffusion processes in social networks (Manchanda *et al.,* 2008; Angst *et al.,* 2010; Peres *et al.,* 2010; Aral and Walker 2011; Susarla *et al.,* 2012; Aral and Walker 2014).

The investigations into the impact of influential people, or opinion leaders, on cascade formation comprise a large part of the literature. Opinion leaders are defined as the individuals that have the ability to strongly affect the opinions or decisions of their network peers (Yoganarasimhan 2012). While some studies degrade the value/power of opinion leaders for social cascade progression (Becker 1970; Watts and Dodds 2007), most authors see opinion leader presence as a critical facilitator for cascade emergence (Lu *et al.,* 2013; Tucker 2008; Ghose and Ipeirotis 2011; Iyengar *et al.,* 2011; Van den Bulte and Joshi 2007). Hinz *et al.,* (2011) experimentally showed that a wisely selected group of opinion leaders can increase the influence spread rate in a cascade up to eight times. Yet, two questions remain unanswered: *How can one select the appropriate opinion leaders for maximizing the spread of influence in a social network?* and *How does this selection depend the social network structure?* While the literature reviewed above is more concerned with exploring the mechanisms of successful cascade propagation, it does not provide a readily available method/solution/strategy (for a company or a political party) to artificially create a cascade in support of a product or opinion by recruiting the "best" opinion leaders. The latter objective, however, may be highly sought-after by research-aware practitioners.

The first organized efforts for identifying influential nodes in social networks relied on centrality-based heuristics (Borgatti 2006). The degree centrality heuristic assumes that any node with a large number of direct connections (called a hub) must be highly influential in a social network. The distance centrality heuristic, on the other hand, considers a node influential if it has short paths to other nodes in the network (called a bridge) (Wasserman 1994; Hinz *et al.,* 2011). The centrality-based heuristics, however, provide no quality guarantee for the solution of the IM problems with multiple required seeds. To formulate an algorithmic approach to finding influential node sets, the term "influence maximization" was coined by Domingos and Richardson (2001). While the first attempts to address the IM problem employed a Markov random field approach (Domingos and Richardson 2001; Richardson and Domingos 2002), Kempe *et al.,* (2003) were first to re-frame it as a discrete optimization problem.

The Independent Cascade (IC) model and the Linear Threshold (LT) model, proposed by Kempe *et al.,* (2003), are the most well-known diffusion models for IM; the optimization problems based on these models are NP-hard (Wang *et al.,* 2012; Chen *et al.,* 2010). Kempe *et al.,* (2003) discovered a submodularity property of the IM objective function and presented a greedy seed selection algorithm with guaranteed, albeit loose, optimality bounds. The problem of assuming submodularity lies in the fact that under it, the marginal gain of adding new seeds should be decreasing, which does not support the idea of fixed threshold effects and reveals a manifest shortcoming in the respective diffusion models (Granovetter 1978; Macy 1991). Furthermore, the original greedy algorithm and even its extensions were found overly demanding computationally (Leskovec *et al.,* 2007; Goyal *et al.,* 2011; Chen *et al.,* 2009; Chen *et al.,* 2010; Wang *et al.,* 2012).

A separate branch of literature has explored social influence from the empirical data mining perspective. As discovered by Aral *et al.,* (2011), a financially viable cascade initiation requires the selection (buy-in) of no more then 0.2% of the nodes in a network. This finding underlines the value of precise seed selection algorithms that can ensure a desirable cost/returns ratio in cascade seeding. However, one observes a gap between the literature based on data-driven studies and algorithmic research. The latter efforts, unfortunately, have often focused on computational investigations in impractical settings and failed to produce managerial insights. The present paper serves as a bridge between these two research thrusts. It designs a realistic diffusion model, strongly supported by mathematical sociology findings, and solves the seed selection problem optimally over real social network datasets, and hence, paves a way to rigorously explore strategic decision-making in social networks.

Note that the original IM problem formulation was concerned with maximizing the expected number of activated nodes *at the end of the diffusion process*, when the activation status of all the nodes becomes fixed, irrespective of the sequence and timing of node activation. However, in many practical IM applications, an influence campaign has a predefined time window, over which it has to achieve the

maximized possible effect. Only recently, Goyal *et al.,* (2012) addressed the issue of unconstrained time horizon for IM and introduced MINTIME problem where the objective is to minimize the time until the activation of a predefined number of nodes.

Note also that in most IC and LT model-based seed selection problems have ignored the aspect of activation timing; furthermore, they assumed no competition. Meanwhile, the existing literature confirms the co-existence of (competing) opinions in real-world decision-making settings (Berger *et al.,* 2010; Bhagat *et al.,* 2012). Chen *et al.,* (2011) were first to recognize this issue by introducing an IC model that allows negative influence to impede the spread of positive influence.

In summary, the diffusion-based IM problems have received attention from the research community for finding influential nodes in social networks and have been confirmed to be useful for treating real-world problems. However, the current diffusion models for IM problems have not been able to produce many managerial insights, in part due to the underspecification of influence flow mechanisms. The present work proposes a mathematical framework for finding exact solutions to seed selection problems, which allows one to more rigorously explore the structure of such optimal solutions.

3 Bayesian Inference Logic in Influence Maximization

This section formalizes the Parallel Cascade (PC) model for diffusion of subjective evidence through social networks, provides the mathematical model for solving the problem of identifying the best positive seeds, and lastly, illustrates the use of the presented PC model with two case studies. The PC model views the node's adoption of a product or opinion, called activation, as a decision-making process based on continuously collected evidence. It has been established that human decision-making can be modeled as Bayesian inference with a high precision whenever the alternatives are given as mutually exclusive hypotheses (Tenenbaum *et al.,* 2006). These human subject experiments confirm that the collection rate and the impact of evidence on human decision-making process can be studied empirically, and hence, the models presented in this paper can be specified based on real-world data. Naturally, the current paper posits that people use their initial preferences and beliefs, as well as the incoming information they receive from peers, to decide whether a given null hypothesis (H_0) or the opposite alternative hypothesis (H_1) is more likely to be true. It is assumed that, based on the evidence accumulated and processed through Bayesian updates, a decision-maker may turn from an observer into a supporter of the hypothesis that convincingly appears more likely to them at a particular point in time; the described transition is defined by thresholds (on the evidence scale) as done in a great deal of sociology literature (Macy 1991; Valente *et al.,* 1994; Young 2001).

Consider a triplet of actors who are testing the same hypothesis, e.g., that a new phone service is reliable. Suppose node 1 observes a fact supporting the hypothesis, e.g., using the new phone service for a month and experiencing few dropped calls, and presents their impression to nodes 2 and 3. Both nodes 2 and 3 update their beliefs about the hypothesis, and then, node 2 shares the absorbed information to node 3 without providing the source of information. Node 3 captures the information (in fact, the rumor) from node 2, treats it as if it provides new evidence supporting the hypothesis and updates its belief again. This process shows how a person's belief about a hypothesis can be reinforced multiple times as a result of a single external test/fact. In social networks, edges serve as channels that permit evidence duplication, and hence, can enable (unfounded) belief reinforcement.

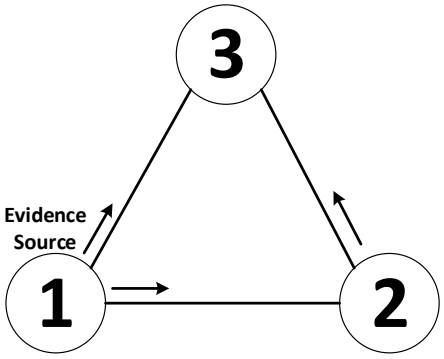


Figure 1: Belief reinforcement through subjective evidence spread in a social network.

Bayesian inference logic uses Bayes' rule to update beliefs in such hypotheses testing (i.e., update the probability that a particular hypothesis is true) as new evidence is received and processed; here, evidence is an objective quantity, that valuates the *new* information regarding a hypothesis, e.g., as a result of observing a *new* fact. However, in reality, beliefs are not necessarily updated based on such facts. When the source of in-coming information is not given (not traceable or forgotten), people still treat the information (supposedly new to them) as evidence, which we term *subjective*, and update their beliefs (Choi *et al.,* 2005; Golub and Jackson 2010). Figure 1 demonstrates the effect of subjective evidence spread on updating beliefs in social networks.

The PC model views positive activation as the event when an actor begins to significantly favor one hypothesis over the other. Once a network node becomes active, it begins to deliver the messages in support of their favored hypothesis to their connected peers. The evidence accumulation can be mathematically expressed by using the "Odds" function (O), defined as the probability that "H_0 is true" divided by the probability that "H_0 is false". Taking the logarithm of the Odds leads to an additive *evidence function*. The evidence function for H_0 is given as

$$e(H_0 \mid Rd) = 10\log_{10}\big(O(H_0 \mid Rd)\big) = e(H_0 \mid R) + 10\log_{10}\left[\frac{P(d \mid H_0R)}{P(d \mid H_1R)}\right], \tag{1}$$

where R is the prior knowledge of the null hypothesis (before the evidence diffusion begins) and d is one signal (a piece of new information) that supports the null hypothesis (Jaynes 2003). Thus, the evidence function combines the prior evidence and observed evidence. With no prior information (data) available, equal probabilities are typically assigned to the null and alternative hypotheses. When a sequence of multiple signals (data) D is received and processed, the updated evidence is given as

$$e(H_0 \mid RD) = 10\log_{10}\big(O(H_0 \mid RD)\big) = e(H_0 \mid R) + 10\sum_i \log_{10}\left[\frac{P(d_i \mid H_0R)}{P(d_i \mid H_1R)}\right]. \tag{2}$$

The increment of the positive evidence (e^+) resulting from a single observation supporting the null hypothesis (d), and the increment of the negative evidence (e^-) resulting from a single observation

supporting the alternative hypothesis (d') are respectively given by

$$e^+ = e(H_0 \mid d) = 10 * \Big(log_{10}P(d \mid H_0R) - log_{10}P(d \mid H_1R)\Big), \tag{3}$$

$$e^- = e(H_1 \mid d') = 10 * \Big(log_{10}P(d' \mid H_1R) - log_{10}P(d' \mid H_0R)\Big). \tag{4}$$

Therefore, upon collecting and processing multiple observations D with n^+ positive and n^- negative signals, the evidence supporting the H_0 becomes

$$e(H_0 \mid DR) = e(H_0 \mid R) + e^+ \cdot (n^+) - e^- \cdot (n^-). \tag{5}$$

The evidence increment values (e^+ and e^-) are used as parameters in the PC model.

3.1 The Parallel Cascade Diffusion Model

Define an Influence Graph as a directed graph $G = (N, A)$, with a set of nodes N and a set of arcs A. Let the sets of positive and negative seeds, i.e., the initial sets of evidence propagators, be denoted by S^+ and S^-, respectively. Note that the notion of "positivity" of evidence is arbitrary: the hypothesis that postulates a claim preferred by the grand policy-maker will hereafter be viewed as positive, hence the distinction between positive and negative evidence. For each node $i \in N$, let $\theta_i^+ \geq 0$ ($\theta_i^- \geq 0$) denote a positive (negative) threshold for the evidence that a node must accumulate, in support of (against) the null hypothesis, to become positively (negatively) activated. In a given problem, θ^+ and θ^- can be set using Bayesian logic: these values should reflect the desired levels of assurance for a node not to make a mistake (about the product/query) when it gets positively or negatively activated (Jaynes 2003).

At each discrete time period $t = 0, 1, 2, ..., T$, let $L_{it} \geq 0$ and $K_{it} \geq 0$ denote the cumulative levels of positive and negative evidence for node i, respectively. Node i is said to be positively (negatively) activated at period t if and only if $L_{it} - K_{it} \geq \theta_i^+$ ($K_{it} - L_{it} \geq \theta_i^-$); otherwise, it maintains the inactive status. Let $L_{i0} = \theta_i^+$ and $K_{i0} = 0$ denote the initial evidence levels of node $i \in S^+$, at time $t = 0$. Equivalently, $L_{j0} = 0$ and $K_{j0} = \theta_j^-$ denote the initial levels of evidence for node $j \in S^-$, at time period $t = 0$. Each node is assumed to accumulate evidence incoming from its activated neighbors, regardless of its own activation status. For each node $i \in N$, the nodes that have arcs toward (coming from) i are termed in-neighbors (out-neighbors) of i; $N_{in}(i)$ and $N_{out}(i)$ are the sets of in-neighbors and out-neighbors of i, respectively.

At time period $t = 0, 1, ..., T$, let $N_t^+(i) \subseteq N_{in}(i)$ $\big(N_t^-(i) \subseteq N_{in}(i)\big)$ denote the set of positively (negatively) activated in-neighbors of i. A node $p \in N_t^+(i)$ sends positive feedback (positive evidence) toward i and $n \in N_t^-(i)$ provides negative feedback (negative evidence) for i. The numerical values of the positive and negative evidence provided by node $i \in N$ at time $t > 0$ to its out-neighbors are denoted by $E_{it}^+ \geq 0$ and $E_{it}^- \geq 0$, respectively. If node i is positively activated at time t, E_{it}^- is zero; if

it is negatively activated at time t, E_{it}^{+} is zero; finally, when i is inactive at time t, both E_{it}^{-} and E_{it}^{+} are zero. The evidence value provided by a node to its out-neighbors in the time period immediately following positive (negative) activation is given by e^{+} (e^{-}). Note that e^{-} is defined as the *absolute* value of the negative evidence calculated using Bayesian logic (i.e., $e^{-} > 0$). At the end of each time period, each node updates its cumulative evidence levels (positive and negative) by adding the newly received evidence to the current evidence levels, and possibly, updates its activation status (to be used for the next time period). Once an activated node loses its activation (becomes inactive), its ability to propagate evidence is immediately revoked. Note that node activation does not have to be followed by an action (e.g., product purchase): the specific application of the model will dictate a desirable assumption in this regard (Bhagat *et al.,* 2012).

In order to realistically capture the effects of information transfer and evidence accumulation in social networks, two decay factors are incorporated in the presented evidence propagation model, one pertaining to evidence provision and the other pertaining to evidence collection and processing. The value of positive (negative) evidence provided by activated nodes decreases by α^{+} (α^{-}) as time passes from the last positive (negative) activation. As a result, the effect of the transferred information in updating the evidence level of out-neighbors is expected to diminish. Furthermore, "forgetfulness" rate β^{+} (β^{-}) is introduced into the PC model to allow nodes to forget (discount as old) a part of the positive (negative) evidence they previously collected. Forgetfulness rate, that has been well studied in marketing literature (Mahajan *et al.,* 1984; Bimpikis *et al.,* 2013), causes the recently observed evidence to make a greater contribution to the decision making process. Also, with time, the nodes will become indifferent to the query, as it often occurs in practice. Figure 2 illustrates the dynamics of PC model-driven evidence propagation over a small network.

Sets S^{+} and S^{-} include the Influence graph nodes that are positively and negatively activated, respectively, at $t = 0$; set S^{-} is given; the nodes in S^{+} are to be selected by the decision-maker solving the IM problem. The diffusion process is terminated after a pre-set (practically relevant) number of time periods (T). Following the traditional setup, the PC model-based IM (PCIM) problem is concerned with populating S^{+*} so as to maximize some measure of the evidence spread in the network. The measure taken in this paper accounts for both the earliness and sustainment of node activation: PCIM amounts to maximizing the count of time periods with positive activation ($\Gamma_G(S^{+}, S^{-})$) while minimizing the count of time periods with negative activation ($\Delta_G(S^{+}, S^{-})$) over all the nodes,

$$S^{+*} \in arg\,max_{(S^{+}\subseteq N | S^{-}\subseteq N, S^{+}\cap S^{-}=\emptyset)} \big(\Gamma_G(S^{+}, S^{-}) - \Delta_G(S^{+}, S^{-})\big).$$

It thus makes the model applicable for such marketing, political and military problems where the timing and duration of activation matter. For example, when activation stands for subscription for a service,

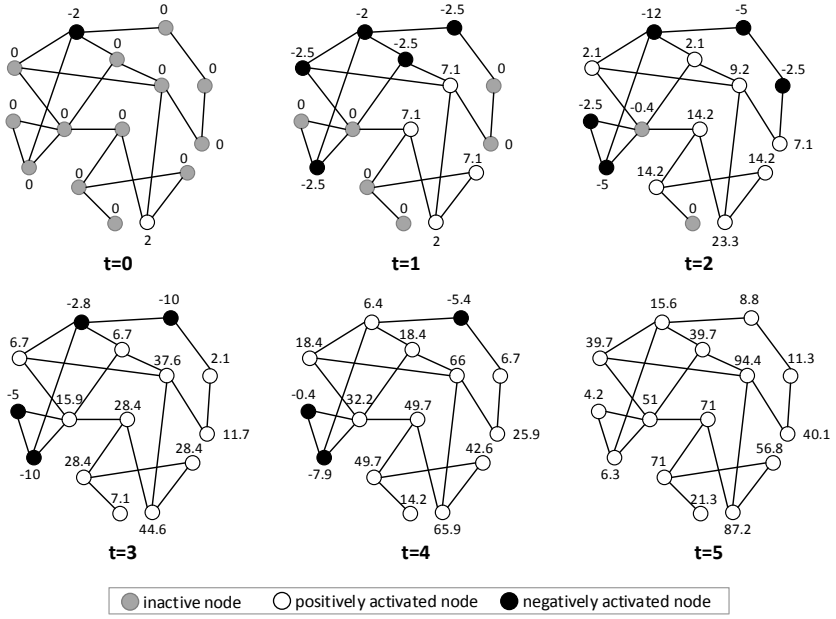


Figure 2: The spread of positive and negative evidence through a network with $|N|$ = 15, T = 5, e^+ = 7.1, e^- = 2.5, $\alpha^+ = \alpha^- = 1.0$, $\beta^+ = \beta^- = 1.0$, θ^+ = 2 and θ^- = −2. The net activation value ($L_{ti} - K_{it}$) for each node is found beside each node. Each graph reports the activation status of each node at a singe time period t.

each node generates profit in each time period that it's positively activated. As a result, the total duration of a node's positive activation determines its contribution to the objective function. Note that a positively activated node still observes both positive and negative evidence: as such, a positively activated node can become negatively activated after receiving enough negative evidence, and vice versa.

Note also that in the absence of negative seeds, when communication can only reinforce the nodes' beliefs, the PC model with the decay factors set to $\alpha^+ = 1$ and $\beta^+ = 0$, reduces to a special case equivalent to the original LT model introduced by Kempe *et al.,* (2003), with the fixed threshold values.

By accommodating conflicting evidence and thanks to its objective function, the PCIM problem can inform decisions even in situations where the decision-maker stands to eventually loose its market position(s). Via the threshold values and forgetfulness rates, the PC model also easily accommodates the non-symmetry of positive and negative influence effects in social networks, i.e., the phenomenon known as the "Negativity Bias", which, e.g., reflects the fact that only a few negative product feedback comments can turn a potential buyer away (Nam *et al.,* 2010; Baumeister *et al.,* 2001; Taylor 1991).

3.2 Optimization Model Specification and Solution Methodology

In this section, a Mixed-Integer Program is constructed for finding exact optimal solutions to the PCIM problem. It is first noted that the PCIM problem is NP-hard.

THEOREM 1. *The PCIM problem is NP-hard.*

PROOF: By a polynomial Turing reduction from the Maximum Coverage Problem (see Appendix A).

The PCIM problem is now formally stated, with the notation summarized in Table 1.

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

Indices	
i, j	node indices
t	time period index
Inputs	
$G(N, A)$	the Influence Graph; a directed graph with a set of nodes N and a set of arcs A
$\|N\|$	total number of nodes in the network
T	total number of time periods in the time horizon considered in the problem
$\|S^+\|$	total number of positive seeds
θ_i^+	the value of positive threshold for i^{th} node
θ_i^-	the value of negative threshold for i^{th} node
e^+	maximum value of positive evidence a node can send in a single time period
e^-	maximum value of negative evidence a node can send in a single time period
α^+	discount rate for the value of positive evidence sent by a positively activated node
α^-	discount rate for the value of negative evidence sent by a negatively activated node
β^+	the rate that each node forgets the previously received positive evidence
β^-	the rate that each node forgets the previously received negative evidence
S_i^-	$\begin{cases} 1, & \text{if } i^{th} \text{ node is a negative seed,} \\ 0, & \text{otherwise,} \end{cases}$
Decision Variables	
X_{it}	$\begin{cases} 1, & \text{if node } i \text{ is positively activated at time } t \\ 0, & \text{otherwise} \end{cases}$
Y_{it}	$\begin{cases} 1, & \text{if node } i \text{ is negatively activated at time } t \\ 0, & \text{otherwise} \end{cases}$
L_{it}	cumulative level of positive evidence for i^{th} node at time t
K_{it}	cumulative level of negative evidence for i^{th} node at time t
E_{it}^+	the value of positive evidence that i^{th} node provides for its neighbors at time t
E_{it}^-	the value of negative evidence that i^{th} node provides for its neighbors at time t

As stated earlier in the paper, at every time period, each node is either positively activated, negatively activated or inactive. At the end of each time period, every node collects all the incoming evidence and updates its cumulative evidence level to determine its activation status for the next time period. The Mixed-Integer Programming model (P) for the PCIM problem is given,

$$(P) \qquad \max Z = \sum_{i=1}^{|N|} \sum_{t=0}^{T} (X_{it} - Y_{it}) \tag{6}$$

Subject to:

$$Y_{it} \geq ((K_{it} - L_{it}) - \theta_i^-)/M \qquad i = 1, 2, ...|N|, \quad t = 0, 1, ..., T, \tag{7}$$

$$1 - X_{it} \geq (\theta_i^+ - (L_{it} - K_{it}))/M \qquad i = 1, 2, ...|N|, \quad t = 0, 1, ..., T, \tag{8}$$

$$X_{it} + Y_{it} \leq 1 \qquad i = 1, 2, ...|N|, \quad t = 0, 1, ..., T, \tag{9}$$

$$L_{it} = \beta^+ L_{it-1} + \sum_{(j,i) \in A} E_{jt-1}^+ \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{10}$$

$$K_{it} = \beta^- K_{it-1} + \sum_{(j,i) \in A} E_{jt-1}^- \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{11}$$

$$L_{i0} = X_{i0}(\theta_i^+) \qquad i = 1, 2, ...|N|, \tag{12}$$

$$K_{i0} = Y_{i0}(\theta_i^- + \epsilon) \qquad i = 1, 2, ...|N|, \tag{13}$$

$$E^+_{it} \leq (\alpha^+ E^+_{it-1}) + (1 - X_{it-1})e^+ \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{14}$$

$$E^+_{it} \leq e^+(X_{it}) \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{15}$$

$$E^-_{it} \geq (\alpha^- E^-_{it-1}) + (Y_{it} - Y_{it-1})e^- \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{16}$$

$$E^-_{it} \leq e^-(Y_{it}) \qquad i = 1, 2, ...|N|, \quad t = 1, 2, ..., T, \tag{17}$$

$$Y_{i0} = S^-_i \qquad i = 1, 2, ...|N|, \tag{18}$$

$$E^+_{i0} = X_{i0}e^+ \qquad i = 1, 2, ...|N|, \tag{19}$$

$$E^-_{i0} = Y_{i0}e^- \qquad i = 1, 2, ...|N|, \tag{20}$$

$$\sum_{i=1}^{|N|} X_{i0} \leq |S^+|, \tag{21}$$

$$0 \leq L_{it}, K_{it}, E^+_{it}, E^-_{it} \qquad i = 1, 2, ...|N|, \quad t = 0, 1, ..., T, \tag{22}$$

$$Y_{it}, X_{it} \in \{0, 1\} \qquad i = 1, 2, ...|N|, \quad t = 0, 1, ..., T. \tag{23}$$

The objective function in (6) takes into account the timing of node activation through the counts of positively and negatively activated nodes in each time period. Note that removing the timing of activation from the objective function in (6) would generate the problem of maximizing the number of positively activated nodes and minimizing the number of negatively activated nodes at the end of the diffusion process, i.e., in period T, which can be solved as a special case of (P).

The constraints (7) and (8) ensure that each node gets positively activated when its net evidence level (the difference between cumulative positive evidence and cumulative negative evidence) is greater than or equal to the positive threshold (θ^+), and gets negatively activated when the net evidence level is less than or equal to the negative threshold (θ^-). In constraints (7) and (8), M is a large positive number greater than or equal to $[\max_{i \epsilon N}(\theta^+_i + \theta^-_i)] + \epsilon + (|N|-1)(T+1)e^-$. Constraint (9) guarantees that, at each time period, each node is either positively or negatively activated, or otherwise inactive. Constraints (10) and (11) ensure the correct updates of the level of cumulative evidence for each node at each time period. The diffusion process starts with the cumulative level of positive and negative evidence set to zero for all the nodes except the seeds. Constraints (12) and (13) ensure that the cumulative level of positive evidence for each positive seed is greater than the positive threshold (θ^+), and the cumulative level of negative evidence for each negative seed is greater than the negative threshold (θ^-). This is required to ensure that the seeds do not lose their ability for propagating influence immediately following the initial time period. As the objective function favors reducing the number of negative activations (deactivates a negatively activated node in case that its negative level of evidence is exactly equal to its negative threshold), a very small positive parameter ϵ, as small as 0.0001, is needed to force the

model to keep the negative seeds negatively activated at the end of the initial time period. Such a parameter is not needed in constraint (12) because the objective function favors keeping the positive seeds positively activated when the level of positive evidence and the positive threshold are the same. Note that assigning a large value to ϵ and adding it to both (12) and (13) would increase the time over which positive and negative seeds can sustain their respective activation. Constraints (14) and (15) set the value of the positive evidence that any node can propagate over a single time period $t \geq 0$ (E^+_{it}); they guarantee that: (a) E^+_{it} is zero when node i is not positively activated at time t ($X_{it} = 0$), (b) E^+_{it} is equal to e^+ when i has become positively activated at time t ($X_{it} - X_{it-1} = 1$), and (c) E^+_{it} is equal to $\alpha^+ E^+_{it-1}$, otherwise. Constraints (16) and (17) set the value of the negative evidence that any node can propagate over a single time period $t \geq 0$ (E^-_{it}). Constraint (18) ensures the initial activation of the negative seeds. Constraints (19) and (20) set the initial value of the evidence propagated by each node in the network. Constraint (21) ensures that the total number of positive seeds at time $t = 0$ does not exceed the pre-defined number of seeds in the problem. The non-negativity and binary constraints for the decision variables in the problem are defined in (22) and (23).

To the best of our knowledge, this paper presents a first mixed-integer program for solving IM problems. The experimental results with (P) are reported in Section 5.

3.3 Case Studies with the PC Model

The most meaningful and valuable IM modeling efforts, reported in the literature, allow for the characterization of the properties of optimal solutions, derived from the analyses of distinct small- and medium-sized problem instances; such findings can then be extrapolated to more general, large problem instances. This section presents two examples using data from real-world networks that illustrate the power of the PC model in explaining the consequences of seed selection decisions when positive and negative influences clash in social networks with specific structure. The PC model reveals how and why the optimal strategy for positive influence spread depends on the selected seeds' positions, on the time length of the window of opportunity the decision-maker has, on the network structure, on the locations of the opponent's seeds, and on the specifications of the evidence accumulation mechanism.

Case Study 1: This example studies the flow of information over the Zachary's karate club network, a well-known network in the literature of social network analysis. The dataset contains 34 members of a karate club who were observed for two years and the friendship links were extracted based on the interactions among members outside of the club-related activities (Zachary 1977). During the data collection course, a disagreement grew between the club's administrator and instructor, which led to the club's break-up into two clubs. Figure 3 shows the Zachary's karate club social network, named Network 1, where node 1 denotes the instructor who is the central node in the first cluster (C_1) and

node 34 denotes the administrator who is the central node of the second cluster (C_2). The clusters depict the eventual student memberships in the two separated clubs (Girvan and Newman 2002).

In order to define a PCIM problem on Network 1, consider it as a new market for a vitamin supplement product. Through personal connections, the students can share information with each other and observe each other using the product: consequently, they can process such observations as evidence in support of the hypothesis that the new product is good.

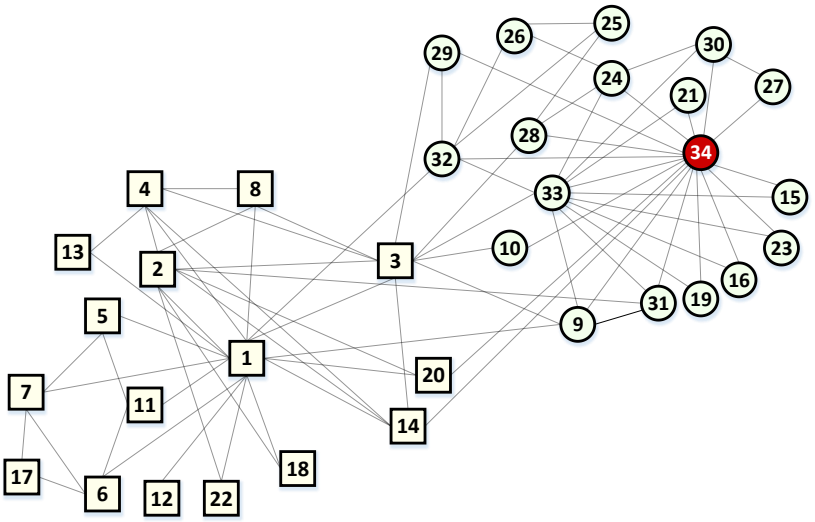


Figure 3: The IM problem on Network 1 with $|N| = 34$, $T = 5$, $\alpha^+ = \alpha^- = 0.8$, $\beta^+ = \beta^- = 1.0$, $\theta^+ = 2$ and $\theta^- = -2$. Square nodes represent members of the first cluster (C_1) and circle nodes represent members of the second cluster (C_2). Node 34 is the club administrator, who serves as the negative seed.

Firm F_1, the producer of a particular model (variation) of the new product, plans to offer the product at a discounted price to two people in the network, as seeds, to encourage other people to adopt the product, which they were reluctant to adopt otherwise. Meanwhile, a competing producer F_2, that produces an alternative product's model, has also identified Network 1 as a niche and has already incentivized node 34, the administrator who has the highest degree in Network 1, to serve as their seed. It is assumed that each person, when exposed to both F_1's and F_2's products, tests the null hypothesis that "F_1's model is better than F_2's" versus the alternative hypothesis that "F_2's model is better than F_1's". The acceptance (rejection) of the null hypothesis by any node corresponds to the adoption of F_1's (F_2's) model, while staying undecided signals that the node has not yet adopted the product. The set position of the negative seed serves as a constraint in the problem that F_1 formulates, with the objective of locating its own seeds more efficiently.

In competing against each other, each company (F_1 and F_2) not only tries to maximize its own profit, but also tries to minimize the competitor's profit. Without any further assumptions, if F_2 has a significantly stronger brand image than F_1, the best intuitive strategy for F_1 is to locate its seeds far away from the negative seed to influence a group of people and reap some profit in the limited time window before all the people adopt the F_2's product. A more challenging problem arises when F_1's brand image is as strong as F_2's. In this situation, F_1 can assign both seeds to cluster C_1 to influence

all people in this cluster, while a more reasonable strategy is to assign one seed to cluster C_2, in the neighborhood of the negative seed to cancel it out, and assign the other seed to cluster C_1.

Turning the described intuition into exact PCIM solutions, however, is not trivial. To this end, one can use program (P); see the results in Table 2. In the solved PCIM instances, all the nodes process evidence in the same manner (i.e., the community is homogeneous), and the evidence threshold values are set to $\theta_i^+ = 2$, $\theta_i^- = -2$, for every node $i \in N$. The positive seeds are gradually strengthened over several problem instances, by increasing the value of positive evidence increment (e^+), which leads to the changing optimal seed locations. The time horizon in every problem instance is set to $T = 5$ (note that, since the network is small, with the diameter of five links, then any two positive nodes can potentially reach the whole network within five time periods).

Table 2: Optimal positive seeds competing the negative seed in node 34 for $T = 5$ (Network 1).

Exp. Index	e^+	e^-	Opt. Seeds	Remarks
1	0.5	3.5	(6,7)	both seeds are as far away from the neg. seed as possible
2	0.8	3.5	(5,6)	both seeds are far away from the neg. seed
3	1.1	3.5	(1,2)	both seeds inch closer to the neg. seed, still in C_1
4	1.7	3.5	(1,33)	one seed stays in C_1 and the other one moves to counter the neg. seed in C_2
5	2.1	3.5	(3,33)	one seed moves to the bridge of C_1 and C_2 and the other one is still in C_2
6	2.6	3.5	(32,33)	both seeds move to C_2 to block the neg. seed in its own cluster
7	3.5	3.5	(3,33)	one seed stays close to the neg. seed and the other one begins to move away
8	4.4	3.5	(1,33)	the seeds spread out over the network, without regard to the neg. seed
9	6	3.5	(1,33)	the seeds spread out over the network, without regard to the neg. seed

In Table 2, the first column shows the experiment index, the second and third columns show the evidence increment values, reflective of the relative quality levels of the F_1's (e^+) and F_2's (e^-) products, and the last column reports the optimal seed set for each instance. The analysis of the optimal seed sets showcases the transition in the optimal seed allocations, as the problem parameters are varied. When the positive evidence increment value is too small, the optimal positive seeds find themselves in the locations most distant from the negative seed. As the positive evidence strength grows in the subsequent instances, the optimal positive seed locations first gradually move toward the negative seed and then spread out evenly over the network. These results are well in line with the intuition.

In order to study the effect of the different time horizon settings on the optimal solution for F_1, the experiments are repeated with various time horizons and the results are reported in Table 3. Tracking the changes in the optimal positive seed locations with the varied T reveals that the decision-maker should become more conservative as the time horizon for the problem increases. In order to further study the patterns in the optimal solution formation with the growing T, assume that the decision-maker (F_1) earns (loses) one dollar per positive (negative) activation per time period. Then, the PCIM objective can be interpreted as the amount of money that the decision-maker earns by the end of the marketing campaign. Taking any action other than the optimal one leads to a regret compared to the

objective value that would be obtained under the optimal seed selection. As such, a decision-maker that relies on centrality-based heuristics, will always select the nodes (1,33) as the positive seeds, as they have the highest degree and betweenness centrality values (except for node 34, which cannot be selected), irrespective of the evidence values and T.

Table 3: The effect of time horizon on the Optimal position of positive seeds (Network 1).

Exp. Index	e^+	e^-	Opt. Seeds $T=2$	Heu. Reg.	Opt. Seeds $T=4$	Heu. Reg.	Opt. Seeds $T=7$	Heu. Reg.	Opt. Seeds $T=9$	Heu. Reg.	Opt. Seeds $T=15$	Heu. Reg.
1	0.5	3.5	(7,14)	3	(7,11)	6	(6,7)	6	(6,7)	6	(6,7)	7
2	0.8	3.5	(1,2)	5	(5,6)	10	(5,6)	13	(7,11)	12	(7,11)	13
3	1.1	3.5	(1,2)	10	(1,2)	35	(1,2)	48	(1,2)	48	(5,6)	55
4	1.7	3.5	(1,33)	0	(1,33)	0	(1,33)	0	(1,3)	2	(1,3)	9
5	2.1	3.5	(1,33)	0	(3,33)	9	(3,33)	24	(3,33)	21	(3,33)	104
6	2.6	3.5	(1,33)	0	(32,33)	31	(32,33)	116	(32,33)	184	(32,33)	393
7	3.5	3.5	(1,33)	0	(3,33)	20	(3,33)	29	(3,33)	31	(3,33)	31
8	4.4	3.5	(1,33)	0	(1,33)	0	(1,33)	0	(1,33)	0	(1,33)	0
9	6	3.5	(1,33)	0	(1,33)	0	(1,33)	0	(1,33)	0	(1,33)	0

Table 3 shows the regret of the heuristic solution; the regret increases with T, which in part explains why the decision-maker becomes more conservative as T increases. Note that the regret values should be standardized to allow for proper comparison across that problem instances with different time horizons. As the maximum amount of money that F_1 can theoretically make in each instance is $N(T+1)$, termed the maximum theoretical revenue (MTR), the heuristic regret of each problem is divided by MTR and the standardized regrets are plotted in Figure 4.

When the positive seeds are weak, the negative evidence conquers the whole network, and vice versa. The peak in Figure 4 corresponds to the case where the groups of positive seeds and the negative seed are almost equally strong - this is when calculated seed selection can have a big impact. The calculations of the area under the standardized regret curve on Figure 4 reveal that the regret value of the heuristic-based seed selection increases with T. Overall, these results emphasize the importance of optimal seed selection in (a) the problems with a large time horizon, and (b) the problems where neither positive nor negative evidence is overly dominant.

Note that when the positive evidence increment (e^+) becomes much larger than the negative evidence increment (e^-), such that any strategy eventually leads to full positive activation in the network, then the optimal strategy is indifferent to both the location of the negative seed and the time horizon, and places the positive seeds so as to minimize the time of reaching all the nodes. Interestingly, this observation brings up the idea of minimizing the maximum distance (or the average distance) of nodes to positive seed(s) as a heuristic method for locating positive seeds in social networks, when positive evidence strongly dominates the negative evidence. This finding connects the problem of locating positive seed(s) for maximizing the spread of evidence in a non-competitive social network to the p-center and $p-$median

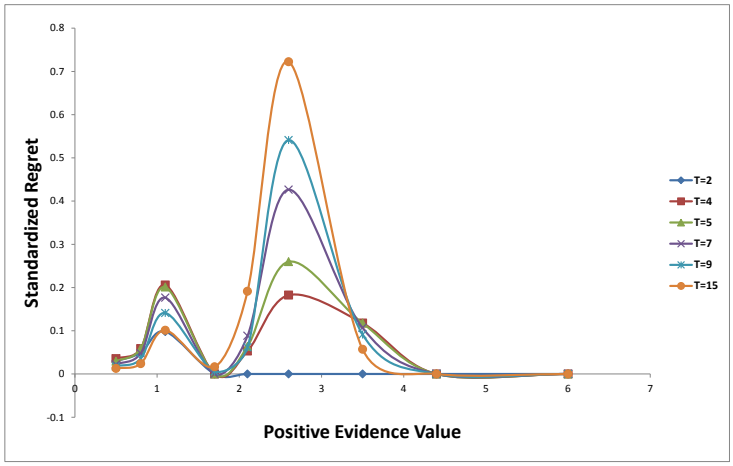


Figure 4: The standardized regret value of the centrality-based heuristics in Table 3.

problems in facility location literature, that try to locate facilities so that it minimizes the maximum or average distance of facilities to the points of demand (Hakimi 1964; Tansel *et al.,* 1983).

Case Study 2: This example studies the stability of judgments in a network in the presence of external influence. To this end, a new concept reflecting the consequences of subjective evidence reinforcement is introduced, and its utility is illustrated in application to the Florentine families' marriage network (Padgett and Ansell 1993). Define a network cluster's "defendability" as the number of its nodes that withstand the pressure of an external judgement, i.e., do not change their opinions/decisions (e.g., related to product purchasing, political party support, etc.). This case study showcases how the defendability of a cluster depends on its interconnectedness and the timing of an external "attack".

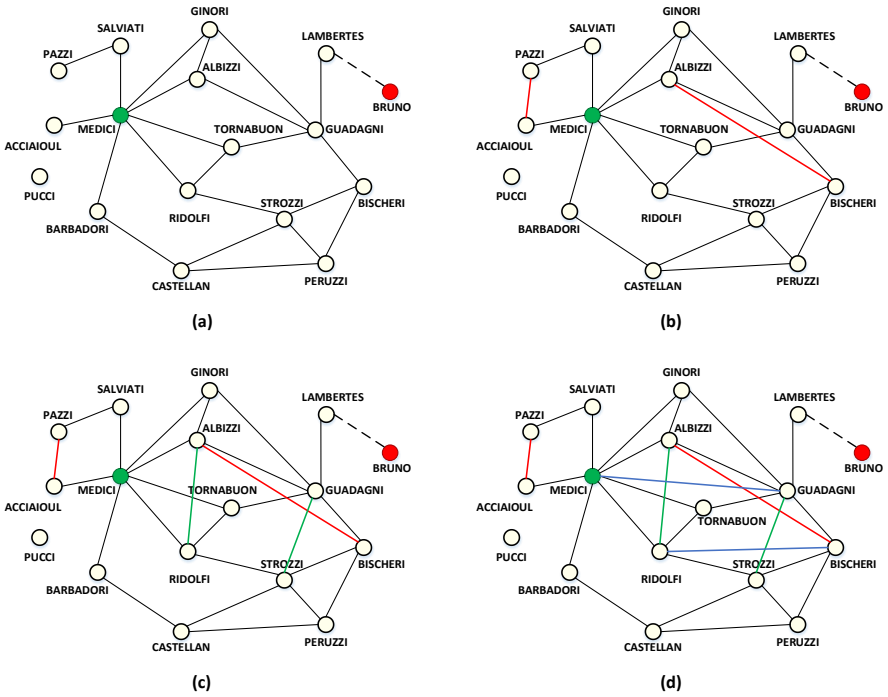


Figure 5: The IM problem on Network 2 with $|N| = 16$, $T = 50$, $\alpha^+ = \alpha^- = 0.7$, $\beta^+ = \beta^- = 1.0$, $\theta^+ = 2$ and $\theta^- = -2$: (a) the initial setup, (b) the red edges are added to the network, (c) the green edges are added to the network, (d) the blue edges are added to the network. Adding edges to the cluster (increasing the density of the cluster) increases its defendability and makes it more difficult for the Bruno family (negative seed) to penetrate the network cluster.

The Florentine families' marriage network, named Network 2, contains 16 elite families in Florence in which the links represent the inter-family marriages in the time period 1394-1434. Padgett and Ansell (1993) illustrated how Medici family took power through creating strategic marriage links in this network. It is of interest to explore how the growing number of within-cluster marriage links would help Medici remain in power, if a new family were to emerge from the outside and attempt to impose its own influence on the cluster (see Figures 5(b) - 5(d)).

Without loss of generality, the Medici family is taken as a positive seed in Figure 5: it is assumed to begin a political campaign at time period $t = 0$. After d time periods, a new family Bruno, taken as a negative seed, creates a marriage link to Lambertes family, a peripheral node in the original network, hoping to initiate an oppositional campaign. The negative influence is assumed stronger than positive ($e^+ = 1$ and $e^- = 3.5$) ensuring that the negative influencer has the potential to penetrate the cluster. Intuitively, one expects the network cluster to reinforce a particular view as it is exposed to it for a long time. Also, the number of within-cluster connections should accelerate the information exchange, and thereby, make the cluster more defendable.

With both positive and negative seed locations given, the PC diffusion model is employed for evaluating the spread of evidence through the network over time (up to $T = 50$) until the optimal solution no longer changes with the growing T. In order to gauge the impact of the cluster density on defendability, two marriage links first are added to the original network: Acciaioul to Pazzi, and Albizzi to Bischeri (Figure 5(b)); then, two additional links are added: Ridolfi to Albizzi, and Strozzi to Guadagni (Figure 5(c)); and finally, two more links: Medici to Guadagni and Ridolfi to Bischeri (Figure 5(d)).

Table 4: The counts of positively and negatively activated nodes in Network 2 at time $T = 50$.

Network	2 (a)		2 (b)		2 (c)		2 (d)	
(Density)	(0.167)		(0.183)		(0.2)		(0.217)	
Delay	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
0	-	15	-	15	-	15	-	15
1	-	15	-	15	-	15	14	1
2	-	15	-	15	-	15	14	1
3	-	15	3	9	14	1	14	1
4	-	15	8	4	14	1	14	1
5	1	11	9	4	14	1	14	1
6	6	9	9	3	14	1	14	1
7	13	1	13	1	14	1	14	1

Table 4 summarizes the results for the four clusters (2(a)-2(d) shown in Figures 5(a)-5(d)): the first row gives the cluster labels; the second row reports the clusters' densities. The first column of Table 4 reports the delay (in the number of time periods) after which the cluster gets exposed to the negative influence. For each cluster in Table 4, the first (second) column reports the total number of nodes (families) that adopt the positive (negative) political opinion by the end of the diffusion process. The

Table 5: Attacking low-degree families.

Network (Density)	Attacking Pazzi (0.167)		Attacking Acciaioul (0.167)	
Delay	(+)	(-)	(+)	(-)
0	13	2	-	15
1	13	2	14	1
2	13	2	14	1
3	13	2	14	1
4	13	2	14	1
5	13	2	14	1
6	13	2	14	1
7	13	2	14	1

Table 6: Attacking Lambertes family

Network (Density)	Attacking Lambertes (0.175)	
Delay	(+)	(-)
0	-	15
1	-	15
2	13	1
3	13	1
4	13	1
5	13	1
6	13	1
7	13	1

results reported in Table 4 quantifies how the cluster defendability increases with the growing density, confirming the claim of Easley and Kleinberg (2010) about the association of *pluralistic ignorance* and the number of direct contacts in the network. The clusters are also observed to become more defendable after a certain delay period, termed *critical delay threshold*, which depends on the evidence strengths, cluster connectivity and proximity of the point of attack to the positive seed in the cluster.

In the real-world scenario, Bruno family would be unlikely to be able to marry into any family in the core of Network 2. A link to Pucci, an isolated node, would hardly be useful. The results of the diffusion process with the same settings as in Table 4, but with Bruno targeting Acciaiiuol and Pazzi (low-degree families), are reported in Table 5. The results showcase the fact that attacking a network through a point far from the positive seed provides a better opportunity for the external evidence to succeed. In order to see how the distance of the point of attack from the positive seed affects the success of the external influence to spread in the cluster, the same experiments are repeated when a link is added to the network to connect the point of attack (Lambertes) to the positive seed (Medici) - see Table 6. The comparison of Tables 4 and 6 reveals that decreasing the distance between the positive seed and the point of attack hurts the prospects of the negative seed. More generally, reinforcing a network with more links makes it more defendable against an opposing influence.

Case Study 2 highlights the fact that investments into influencing a well-connected community must be carefully calculated. Both the community structure and intervention timing are important to such ventures. Note that in a marketing problem of occupying and protecting a market niche, the delay considered in this section can be viewed as that of introducing a competing product. In this context, the PC diffusion model can help valuate long-term marketing strategies, i.e., assess the trade-off between an earlier yet more expensive or a delayed but less expensive product introduction.

In summary, Section 3.3 showcases the value of the PC diffusion model for expressing the spread of evidence in practical IM problems. Furthermore, the provided case studies exemplify how sensitive the

optimal solutions to IM problems may be to the numerical values of problem parameters. Notably, the present section connects the seed positioning problem in social networks to the facility location problem, a well-studied problem in the literature of Operation Research, that opens a door to applying location theory models for IM problems in social networks. Section 4.2.1 explains how the methods from the location theory literature can inform new IM heuristics.

4 A Set of Lagrangian Heuristic Tools for PCIM

As mentioned in Section 3 (and proved in Appendix A), the PCIM problem is NP-hard. The sources of complexity of the PCIM problem include the number of nodes in the Influence Graph, the total number of time periods for spreading the evidence and the number of positive seeds. An efficient approach is required for solving PCIM problem instances with large Influence Graphs.

This section presents a guaranteed-performance heuristic method for PCIM using Lagrangian Relaxation. It works by relaxing a preselected subset of constraints in (P) and including weighted penalty terms for violating the relaxed constraints into the objective function. Lagrangian Relaxation has been applied for solving optimization problems in various areas, including supply chain network design (Pan and Nagi 2013), scheduling (Diaby *et al.,* 1992), network planning (Siomina *et al.,* 2007) and data clustering (Ding *et al.,* 2005). A Lagrangian Relaxation heuristic for PCIM is designed in Section 4.1: it identifies good feasible solutions in reasonable time, while returning a tight upper bound for the optima. To achieve the latter, a Subgradient algorithm is presented in Appendix B as a method for finding the lowest upper bound for PCIM. Finally, two heuristic methods are presented in Section 4.2 for finding near-optimal feasible PCIM solutions and obtaining tight lower bounds for the optima.

4.1 Lagrangian Relaxation for Finding an Upper Bound for PCIM Solutions

By definition, incorporating the removed PCIM constraints into its objective function results in a valid "relaxation" of the original formulation (Held and Karp 1970; Fisher 1981).

A Lagrangian Relaxation problem (LR_u) for (P) is given,

$$(LR_u) \qquad \max Z^{LR_u}(u) = \sum_{i=1}^{|N|}\sum_{t=0}^{T}(X_{it} - Y_{it}) + u(|S^+| - \sum_{i=1}^{|N|} X_{i0}), \tag{24}$$

Subject to:

$$u \geq 0, \tag{25}$$

$$(7)\text{-}(20),$$

$$(22)\text{-}(23).$$

Each feasible solution of (P) is feasible for the corresponding (LR_u), since (LR_u) is at most as constrained as (P). In order to make (LR_u) a valid relaxation for (P), a non-negativity constraint is required for the Lagrangian multiplier (u). As a result of defining (LR_u) as a relaxation for (P), each feasible solution for (LR_u) provides an upper bound for (P). In an effort to obtain tight upper bounds for PCIM, a Lagrangian dual problem (LD_u) is formulated,

$$(LD_u) \qquad Z^{LD_u} = Min_{u'} Z^{*LR_u}(u'), \tag{26}$$

where $Z^{*LR_u}(u')$ is the optimal solution of (LR_u), for a given u'. The Lagrangian dual problem (LD_u) can be iteratively solved for finding the dual multipliers that minimize the optimal solution of (LR_u) to obtain the best (lowest) upper bound for (P).

To make the iterative search procedure of solving LD_u more efficient, a loose relaxation of (LR_u) is preferable. Tighter relaxations, however, are expected to provide better bounds for (P); thus, a trade-off arises between executing fewer iterations of the search procedure for solving (LD_u) with a tighter relaxation and executing more iterations of the search procedure for solving (LD_u) with a less tight relaxation. Such relaxations that keep X_{it} $(i = 1, 2, ..., |N|, t = 0, 1, ..., T)$ binary and keep the negative seeds fixed are computationally easier because adding (dropping) positive seeds to (from) their optimal solutions can provide valid feasible solutions for (P). With this idea in mind, constraint set (21) is relaxed with dual multiplier u. Although the selected relaxed problem removes the constraint for the exact number of positive seeds in (P), the maximum number of positive seeds in (P) is still constrained by the sets (9) and (18), which do not allow a node to be a negative seed and a positive seed at the same time. In this paper, a Subgradient search procedure, a famous hill climbing algorithm (Fisher 2004), is applied to solve the Lagrangian dual problem (see Appendix B). There are other methods including simplex-based methods and multiplier adjustment methods proposed in the literature for solving Lagrangian dual problems, but Subgradient-based procedures, in general, achieve better computational performance (Fisher 1981; Fisher 2004). Two heuristic methods are proposed next for finding near-optimal feasible solutions for (P) to provide the lower bound for calculating the heuristic gap and updating the step size for the Subgradient algorithm.

4.2 Obtaining the Lower Bounds for Optimal PCIM Solutions

Each feasible solution for (P) presents a valid lower bound for the optimal solution for (P). The presented PCIM problem always has at least one feasible solution if the total number of positive seeds and negative seeds is less than or equal to the total number of nodes in the Influence Graph. The simplest method for finding a feasible solution for (P) is to trivially select any $|S^+|$ nodes, which are not negative seeds, as positive seeds. Although such solutions satisfy the stopping criterion in the

Subgradient algorithm, the resulting lower bound is not necessarily tight. In this section, two heuristic methods are presented for finding near-optimal feasible solutions.

4.2.1 The Iterative Seed Removal (ISR) Algorithm

The PCIM problem has three properties, discovered through experimental studies with the mathematical model (P) over Network 1, Netowrk 2 and real Facebook datasets from SNAP collection (Leskovec and Krevl 2014), and presented in this section as observations. The ISR algorithm to be presented utilizes these properties to efficiently find near-optimal feasible solutions for PCIM problem.

Observation 1. In the PCIM problem, when the positive seed locations are given at time $t = 0$, the calculation of the resulting objective function value in (P) takes $O(T|N|^2)$ time.

Observation 2. For the PCIM problems with the varying number of positive seeds to be selected ($|S^+|$), the solution time is a concave function of $|S^+|$.

Observation 3. The intersection of the sets of optimal seeds for two instances of the PCIM problem with a different number of positive seeds is generally non-empty; moreover, in a vast majority of cases, the optimal seed set for a PCIM problem instance is a subset of the optimal seed set for the PCIM instance with the same parameter specification but more positive seeds.

The analysis of PCIM run-times in Section 5.2 experimentally confirms Observations 1 and 2. As a piece of evidence for Observation 3, Table 7 gives the results for three PCIM problems with the growing number of positive seeds. The first problem uses the Mexican Political Elite (MPE) network that contains the significant friendship, kinship, political and business connections within a powerful political group in Mexico (De Nooy *et al.,* 2011). The next two problems use Facebook data subsets, FB_1 and FB_2, found in the SNAP collection (Leskovec and Krevl 2014). In each case, the number of positive seeds in the original PCIM problem BP (Base Problem) is equal to four, and new PCIM problems are generated by iteratively incrementing the number of positive seeds by one, and then, solved to see if the optimal solution of a problem with more positive seeds includes the seeds from the optimal solution for BP. The results confirm that the optimal solution for all the problems with more than four positive seeds contain the solution for BP. Note that Observation 3 experimentally authenticates the utility of the Greedy algorithm of Kempe *et al.,* (2003) and the facility location Stingy algorithm (also known as the Greedy-Drop algorithm) of Feldman *et al.,* (1966), for solving PCIM problems.

The ISR algorithm employs the PCIM problem properties captured in the three presented Observations to efficiently obtain good and tight solutions to practical problem instances. Consider an instance of the PCIM problem with $|S^+|$ positive seeds to be identified (henceforth referred to as the original problem). The ISR algorithm increases the number of positive seeds and defines a Dummy problem with

Table 7: Computational results for small- and medium-sized PCIM problem instances.

	Dataset	$\|S^+\|$ = 4 (BP)	$\|S^+\|$ = 5	$\|S^+\|$ = 6	$\|S^+\|$ = 7
Opt. Sol. ($\|N\|$=35, T=10)	MPE	(2,10,12,20)	(2,10,12,18,20)	(2,10,12,14,20,31)	(2,10,12,14,20,29,31)
Inclusion of BP Solution (%)		-	100	100	100
Opt. Sol. ($\|N\|$=40, T=30)	FB_1	(6,8,17,19)	(6,8,17,19,25)	(6,8,17,19,25,34)	(5,6,8,17,19,25,34)
Inclusion of BP Solution (%)		-	100	100	100
Opt. Sol. ($\|N\|$=50, T=25)	FB_2	(10,19,25,43)	(10,16,19,25,43)	(10,16,19,25,43,48)	(10,16,19,25,43,47,48)
Inclusion of BP Solution (%)		-	100	100	100

$|S_d^+| > |S^+|$ positive seeds. The Dummy problem is exactly the same as the original PCIM problem in the Influence Graph and input parameters, but it seeks for a greater number of positive seeds. An optimal solution for the Dummy problem is necessarily infeasible for the original PCIM; the ISR algorithm works to iteratively obtain the best combination of the seeds to be removed from the optimal solution for the Dummy problem and obtain a good feasible solution for the original problem. The number of positive seeds in the Dummy problem is chosen to be large, but not too large, so that it can be solved fast, and also, the seed removal procedure can be efficient.

According to Observation 2, it is always possible to find a simple Dummy problem with $|S^+| + |S^-| \leq |N|$. According to Observation 3, an optimal solution for the Dummy problem is expected to include the positive seeds present in the optimal solution for the original problem. Hence, the ISR algorithm executes the greedy algorithm of Kempe *et al.,* (2003) backwards. At each iteration of the ISR algorithm, the problem with more positive seeds is called a *superior problem* because its objective function value is necessarily greater than or equal to that of a subproblem achieved by removing one seed, hence called an *inferior problem*. To begin with, let the first superior problem have $|S_d^+|$ positive seeds and define an inferior problem as a maximization problem for finding the best set of $|S_d^+| - 1$ positive seeds. Instead of solving the inferior problem, the ISR algorithm traverses all the distinct combinations of $|S_d^+| - 1$ positive seeds in the solution of the superior problem and selects the combination that maximizes the objective function of the inferior problem. According to Observation 1, computing the objective function value of the inferior problem for each possible combination of $|S_d^+| - 1$ positive seeds in the optimal solution of the superior problem takes $O(T|N|^2)$ time. To proceed, the ISR algorithm keeps removing the positive seeds until it obtains a set of $|S^+|$ positive seeds, and reports it as a feasible solution for the original PCIM problem and a lower bound for the optimal solution. The total number of inferior problems that the ISR algorithm solves to obtain the feasible solution for the original PCIM problem with $|S^+|$ positive seeds using a Dummy problem with $|S_d^+|$ positive seeds is

$$\binom{|S_d^+|}{|S_d^+|-1} + \binom{|S_d^+|-1}{|S_d^+|-2} + \ldots + \binom{|S^+|+1}{|S^+|} = |S_d^+| + |S_d^+ - 1| + \ldots + |S^+| + 1 = \frac{|S_d^+|^2 + |S_d^+| - |S^+| - |S^+|^2}{2}. \quad (27)$$

The ISR algorithm elegantly employs the PCIM problem properties. However, its main drawback is its independence from the Subgradient algorithm: the upper bound for the PCIM problem, found by

the Subgradient algorithm, does not feed into the ISR algorithm. Furthermore, the ISR algorithm may not work well if all the available Dummy problems are hard to solve.

Algorithm 1 - The ISR Algorithm for PCIM

```
Initialize |S+| and |S+_d| in a Dummy problem, with |S+_d| > |S+|;
Initialize bestSolutionValue with −M and currentSolutionValue with 0;   /* M is a large positive number */
Solve the Dummy Problem with |S+_d| positive seeds;
Store the solution in S*_sup;
for t ← 0 to (|S+_d| − |S+| − 1) do
    for i ← 1 to (|S+_d| − t) do
        createNewSolution(i);     /* This function removes i^th seed from S*_sup to obtain a new solution for the inferior problem*/
        evaluateNewSolution(i);     /* This function evaluates the objective function for the new solution*/
        storeCurrentSolution(i);     /* This function stores the objective value of the new solution in currentSolutionValue*/
        if currentSolutionValue ≥ bestSolutionValue then
            recordBestSolutionIndex();     /* This function records i as the index of the best solution for PCIM*/
            updateBestSolutionValue();     /* This function updates the best solution for PCIM*/
        end if
    end for
    removeOneSeed();     /* This function removes the seed with best solution index from seed set and updates S*_sup*/
    updateBestSolutionValue();     /* This function initializes bestSolutionValue with −M*/
end for
```

4.2.2 The Adaptive Subgradient-Based (ASB) Algorithm

The ASB algorithm is designed to utilize the information obtained in executing the Subgradient algorithm to iteratively improve the lower bound for PCIM. The optimal solution of (LR_u) does not necessarily provide a feasible solution for (P), since (LR_u) is not constrained by the number of positive seeds. Let S_L^+ be the set of positive seeds in the optimal solution of (LR_u). At each iteration of the Subgradient algorithm, if $|S_L^+| \geq |S^+|$, the ASB algorithm selects the first $|S^+|$ positive seeds (with respect to a fixed random ordering) from S_L^+ to obtain a feasible solution for (P) with $|S^+|$ positive seeds. On the other hand, when $|S_L^+| < |S^+|$, the ASB algorithm selects all the positive seeds in S_L^+ and randomly selects $|S^+| - |S_L^+|$ positive seeds from the nodes in the network that are neither in S_L^+ nor in S^-. In the early iterations of the Subgradient algorithm, the ASB algorithm blindly selects $|S^+|$ positive seeds, however, the selection process becomes more precise as the Subgradient algorithm runs further.

Algorithm 2 - The ASB Algorithm for PCIM

```
Initialize |S+|;
Initialize bestSolutionValue with solution of ISR algorithm and currentSolutionValue with 0;
while gapValue ≤ acceptableGap do
    storeLagrangianSolution();     /* This function stores the solution of LR_u to be used for finding the lower bound*/
    createFeasibleSolution();     /* This function creates a feasible folution for (P) using S+_L*/
    evaluateNewSolution();     /* This function calculates the objective value of PCIM for the stored solution*/
    storeCurrentSolution();     /* This function stores the objective value of the new solution in currentSolutionValue*/
    if currentSolutionValue ≥ bestSolutionValue then
        recordBestSolution();     /* This function updates the best feasible solution for PCIM*/
    end if
end while
```

The ISR and ASB algorithms are very efficient when used together in practice. The ASB algorithm first stores the best feasible solution obtained by the ISR algorithm as the best feasible solution of (P)

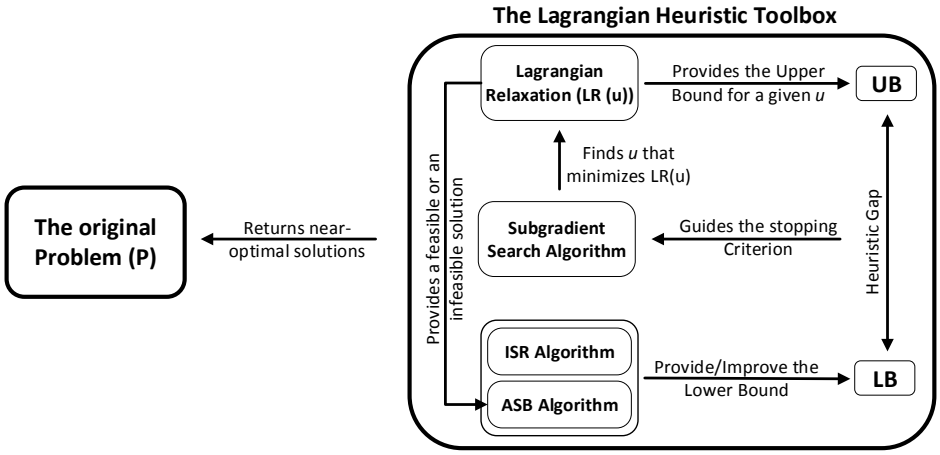


Figure 6: The Lagrangian heuristic toolbox: an overview of the components.

found so far. At each iteration of the Subgradient algorithm, the ASB algorithm extracts a feasible solution for (P), using Algorithm 2, and quickly finds the corresponding objective value (see Observation 1). At each iteration of the Subgradient algorithm, the feasible solution for (P), obtained by the ASB algorithm, is accepted only if it provides a greater objective function value than the current best feasible solution.

To summarize, the presented algorithms form a Lagrangian heuristic toolbox for obtaining near-optimal PCIM problem solutions with rigorously evaluated bounds; the utility of and relationships between the algorithms are explained in Figure 6.

5 Computational Results

This section presents the computational results with the PCIM instances on some real social networks. Subsection 5.1 studies the performance of the Lagrangian Relaxation toolbox for PCIM. Subsection 5.2 focusing on run-time and discusses the sources of complexity in the PCIM problem.

5.1 Lagrangian Relaxation Performance

In order to analyze the performance of the presented Lagrangian Relaxation heuristic, this section solves the PCIM problem instances formulated on four Facebook networks found in SNAP collection (Leskovec and Krevl 2014). The network size- and structure-dependent statistics of these undirected datasets, indexed F1, F2, F3 and F4, are reported in Table 8. The nodes in these networks are labeled; in order to evaluate the performance of the heuristic method, each experiment takes a sub-network of the main dataset with $|N|$ nodes.

In this work, the Mixed-Integer Program and the Lagrangian Relaxation heuristic are implemented using Concert Technology in JAVA and the commercial solver CPLEX 12.5. All the experiments have

Table 8: Dataset Statistics.

Dataset	Nodes	Edges	Directed	Density
F1	150	1693	No	0.151
F2	747	30025	No	0.108
F3	534	4813	No	0.034
F4	1034	26749	No	0.05

been performed on a desktop with Intel(R)Core(TM)i3 3.3GHz processor, with 8GB RAM and 64 bit operating system. Table 9 shows the computational results for small and medium-sized problems, all solved to optimality using CPLEX. The availability of the optimal solutions for these problems permits calculating both the optimality gap and heuristic gap. For the small problems, CPLEX outperforms the Lagrangian Relaxation heuristic in terms of solution time. As the problem size increases the PCIM solution time with CPLEX increases rapidly (see Section 5.2 for the sources of the PCIM problem complexity), while the Lagrangian Relaxation heuristic remains fast. Note that in the majority of the PCIM problem instances reported in Table 9, the ISR and ASB algorithms have found the optimal solution (the optimality gap is equal to zero).

Table 9: Computational results with small- and medium-sized PCIM problem instances.

Dataset	$\|N\|$	T	$\|S^+\|$	LR Time (sec.)	LR LB	LR UB	Cplex Time (sec.)	Cplex Sol. (Opt.)	Opt. Gap (%)	Heu. Gap (%)	Iter. #
F1	30	40	6	11.53	1167	1186	0.69	1167	0	1.6	20
F1	40	100	9	72.01	4020	4021	7.89	4020	0	0.02	20
F1	60	50	7	81.32	2849	2931	74.32	2853	0.1	2.7	20
F2	45	64	7	26.59	2795	2870	7.71	2795	0	2.6	20
F2	60	50	9	32.06	2887	2929	45.84	2887	0	1.4	20
F2	85	75	14	49.66	6233	6289	3425.23	6233	0	0.8	30

The results of the computational study with large-sized problems are given in Table 10. For these problems, CPLEX runs out of computer memory and fails to return optimal solutions. In such cases, the Lagrangian Relaxation heuristic runs in a reasonable computational time and provides an acceptable heuristic gap. For large problems, the optimality gap is unknown, due to unknown optimal solution, and the heuristic gap remains the only criterion for the evaluation of the heuristic's performance. The run-time for the Lagrangian Relaxation heuristic smoothly increases with the dimensions of PCIM problem instances and it illustrates the supreme contribution provided by the heuristic approach.

In order to assess the scalability of the Lagrangian Relaxation heuristic for solving practical PCIM problems, it is executed with large Facebook networks, where CPLEX cannot even create a feasible solution in the computer memory. It is observed that the Lagrangian Relaxation heuristic still provides acceptable bounds for optimal PCIM solutions. Table 11 reports the results of a computational study with five large problems where the only concerns are the heuristic gap and solution time of the Lagrangian Relaxation heuristic. The results of computational studies in Table 11 show that the proposed

Table 10: Computational results with large-sized PCIM problem instances.

Dataset	$\|N\|$	T	$\|S^+\|$	LR Time (sec.)	LR LB	LR UB	Cplex Time	Cplex Gap	Heu. Gap (%)	Iter. #
F1	80	50	9	70.91	3839	3923	> 4 hr	> 195%	2.1	60
F1	100	60	10	112.99	5764	5981	> 4 hr	> 190%	2.4	60
F2	120	70	10	259.73	8021	8280	> 4 hr	> 198%	3.1	60
F2	120	70	10	259.73	8021	8280	> 4 hr	> 198%	3.1	60
F3	80	50	11	78.65	3691	3836	> 4 hr	> 192%	3.7	60
F3	120	70	10	259.73	8021	8280	> 4 hr	> 198%	3.1	60

heuristic method provides encouraging results for large-scale problems, establishing its practical value.

Table 11: Computational results with large-sized PCIM problem instances.

Dataset	$\|N\|$	T	$\|S^+\|$	LR Time (sec.)	LR LB	LR UB	Heu. Gap (%)	Iter. #
F2	550	50	40	594.96	26940	27494	2.0	60
F2	720	35	60	831.72	24467	25070	2.4	60
F3	480	30	84	522.73	13977	14208	1.6	60
F4	1034	30	100	1589.34	28991	29822	2.8	60

5.2 A Sensitivity Analysis of the PCIM Problem Run-time Dynamics

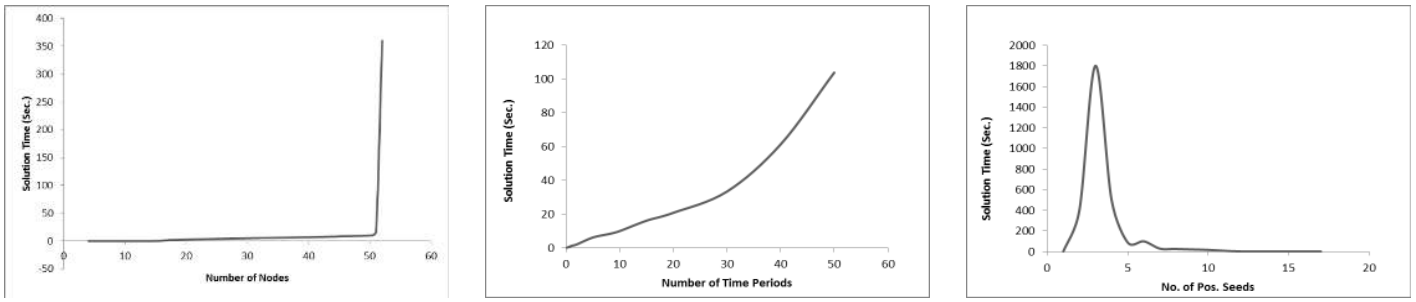
Three elements affect the solution time of program (P) for PCIM: the number of nodes in the Influence Graph, number of time periods for evidence spread and number of positive seeds to be selected in the problem. In this section, the PCIM input parameters are varied selectively, and three sets of experiments are performed with F3 dataset; in each case, only one of the three aforementioned factors is changed to see how it affects the solution time. The results in Figure 7(a) show that the solution time increases rapidly with the growing number of nodes in the Influence Graph, e.g., solving a problem with 90 nodes takes about 50 times more time over a problem with 80 nodes.

Figure 7(b) shows the effect of the total number of time periods, T, on the run-time of (P) revealing a linear trend. Problems with a small number of nodes appear to remain tractable even with large T.

The results of the third set of experiments (see Figure 7(c)) show how the solution time of (P) is affected by the number of positive seeds in the PCIM problem. These results authenticate the second observation given in Section 4.2.1. As shown in Figure 7(c), the solution time resembles a concave function in the number of positive seeds, which motivates the ISR Algorithm: the number of positive seeds in a hard PCIM problem instance needs to be just slightly increased to find a Dummy problem with a significantly lower solution time.

6 Discussion

This section discusses the limitations of the presented models and methods, and concludes the paper.



(a) (b) (c)

Figure 7: Sensitivity analyses of (P) with $\alpha^+ = \alpha^- = 0.7$, $\beta^+ = \beta^- = 0.9$, $e^+ = 1.2$, $e^- = 1.25$: (a) run-time dynamics as a function of the total number of nodes in the network ($T = 10$, $|S^+| = 4$, $|S^-| = 3$), (b) run-time dynamics as a function of the number of time periods ($|N| = 50$, $|S^+| = 4$, $|S^-| = 3$), and (c) run-time dynamics as a function of the number of positive seeds ($|N| = 50$, $T = 10$, $|S^-| = 3$).

6.1 Study Limitations

This paper provides insightful findings and develops a framework for modeling the spread of influence in a social network. However, this study has limitations worth mentioning.

First, while the PC model relies on the theory and findings established in the sociology literature for human decision-making (Tenenbaum *et al.,* 2006), it treats stochasticity only implicitly (through Bayesian updates) and does not emphasize the differences between network actors and the uncertainty in capturing such differences. The deterministic diffusion process makes the PCIM problem mathematically tractable, i.e., allows one to solve it as a mixed-integer program, design efficient heuristics exploiting known and fixed network characteristics, and make insightful observations (after all, linear programs are often found useful in practice even though real-world problems are rarely truly deterministic). Future work, however, can involve stochastic optimization for PCIM.

Second, data-focused studies are needed to uncover and address potential challenges in specifying the model parameters, i.e., learning how people really process subjective, as opposed to objective, evidence. The investigations with the latter have been previously conducted (Xu and Tenenbaum 2007; Goodman *et al.,* 2011), which gives promise to the expansion of the presented research in this direction, too; such studies, however, should lie in the consumer psychology domain. On a positive note, from the modeling and algorithmic perspectives, the PC model can be used with user-defined parameters, and its ability to produce practical insights is confirmed through the reported case studies.

6.2 Concluding Remarks

This paper models social influence as a consequence of subjective evidence transfer, and quantitatively derives general insights about cascading behavior and belief reinforcement in social networks. The

presented Parallel Cascade (PC) diffusion model defines the rules of exchange and accumulation of subjective evidence, which feeds into node-level hypothesis testing, *en route* to making decisions, forming judgments, etc. The preference of a null hypothesis over an alternative hypothesis determines a node's participation status in regards to further evidence propagation. The value of evidence collected and accumulated, in support of or against the null hypothesis, is calculated using Bayesian update logic. The optimization problem of finding the set of influential nodes to initiate the evidence spread in support of the null hypothesis under the PC diffusion model (PCIM) is formulated as a mathematical program and solved using CPLEX. The PCIM problem is shown to be NP-hard, and next, an efficient, guaranteed-performance heuristic tool set is presented, exploiting Lagrangian Relaxation.

The studies of the spread of evidence in social networks using the PC diffusion model showcase that the ability of the decision-maker to trigger a successful cascade or keeping a cascade alive is sensitive to the density of network connections and the presence of the opposite opinions in a target cluster. This paper focuses on node-level IM solutions, utilizing the exact fine features of the network structure; however, it also opens a door to studying the problem on the network level, e.g., describing the general properties of the seeds' optimal locations based on metrics such as density and clusterization. Based on the presented PC diffusion model, one can potentially develop new centrality metrics for evaluating network ability to reinforce/preserve beliefs. Future research can also explore how PCIM instances with extremely large Influence Graphs can be reduced, e.g., via clustering, to become manageable.

The PC model quantifies belief reinforcement through social connections, and informs the changes in optimal seed allocation for creating successful cascades. As noted in Section 3, the model can incorporate actors' actions: based on the collected evidence, the actors may not only be active in spreading their opinions and judgments, but also, choose to buy a product, vote for a party, etc. Such actions will result in the acquisition of first-hand objective evidence by the actors, which can be processed differently in comparison with the processing of subjective evidence). The addition of actions in the model can lead to more insightful analyses, e.g., of low-quality but actively advertised goods where customers may get excited about a product but only until they buy one.

Also, this paper opens up a new area for modeling the defendability of cohesive clusters in social networks against strong external opinions and for the identification of "vulnerable" nodes in network clusters. Furthermore, the paper establishes connection between PCIM and location theory models. Further efforts will pursue the construction of a theoretical method for solving stochastic PCIM instances. Moreover, future studies can apply the proposed optimization scheme for modeling the spread of evidence in the social networks that are growing, and in situations where neither the structure of a social network nor the locations of the opponent's opinion leaders are precisely specified.

Acknowledgments

This work was supported in part by the National Science Foundation grant ICES-1216082, and a Multidisciplinary University Research Initiative (MURI) grant W911NF-09-1-0392. This support is gratefully acknowledged.

References

Angst, C. M., Agarwal, R., Sambamurthy, V., and Kelley, K. (2010). Social contagion and information technology diffusion: The adoption of electronic medical records in us hospitals. *Management Science 56*(8), 1219–1241.

Aral, S., Muchnik, L., and Sundararajan, A. (2011). Engineering social contagions: Optimal network seeding and incentive strategies. In *Winter Conference on Business Intelligence*.

Aral, S. and Walker, D. (2011). Identifying social influence in networks using randomized experiments. *Intelligent Systems, IEEE 26*(5), 91–96.

Aral, S. and Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science 60*(6), 1352–1370.

Arthur, D., Motwani, R., Sharma, A., and Xu, Y. (2009). Pricing strategies for viral marketing on social networks. In *Internet and Network Economics*, pp. 101–112. Springer.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology 5*(4), 323.

Becker, M. H. (1970). Sociometric location and innovativeness: Reformulation and extension of the diffusion model. *American Sociological Review*, 267–282.

Berger, J., Sorensen, A. T., and Rasmussen, S. J. (2010). Positive effects of negative publicity: when negative reviews increase sales. *Marketing Science 29*(5), 815–827.

Bhagat, S., Goyal, A., and Lakshmanan, L. V. (2012). Maximizing product adoption in social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 603–612. ACM.

Bimpikis, K., Ozdaglar, A., and Yildiz, E. (2013). Competing over networks. *submitted for publication. Available online at http://web.mit.edu/asuman/www/publications.htm*.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks 27*(1), 55–71.

Borgatti, S. P. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory 12*(1), 21–34.

Chen, W., Collins, A., Cummings, R., Ke, T., Liu, Z., Rincon, D., Sun, X., Wang, Y., Wei, W., and Yuan, Y. (2011). Influence maximization in social networks when negative opinions may emerge and propagate. In *Proceedings of SIAM International Conference on Data Mining*, pp. 379–390.

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.

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.

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.

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

Choi, S., Gale, D., and Kariv, S. (2005). Learning in networks: An experimental study. *Unpublished manuscript*.

De Nooy, W., Mrvar, A., and Batagelj, V. (2011). *Exploratory social network analysis with Pajek*, Volume 27. Cambridge University Press.

Deroıan, F. (2002). Formation of social networks and diffusion of innovations. *Research Policy 31*(5), 835–846.

Diaby, M., Bahl, H. C., Karwan, M. H., and Zionts, S. (1992). A Lagrangean relaxation approach for very-large-scale capacitated lot-sizing. *Management Science 38*(9), 1329–1340.

Ding, C., He, X., and Simon, H. D. (2005). Nonnegative Lagrangian relaxation of k-means and spectral clustering. In *Machine Learning: ECML 2005*, pp. 530–538. Springer.

Dinh, T. N., Zhang, H., Nguyen, D. T., and Thai, M. T. (2014). Cost-effective viral marketing for time-critical campaigns in large-scale social networks. *IEEE/ACM Transactions on Networking (TON) 22*(6), 2001–2011.

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.

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

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

Feldman, E., Lehrer, F., and Ray, T. (1966). Warehouse location under continuous economies of scale. *Management Science 12*(9), 670–684.

Fisher, M. L. (1981). The Lagrangian relaxation method for solving integer programming problems. *Management science 27*(1), 1–18.

Fisher, M. L. (2004). The Lagrangian relaxation method for solving integer programming problems. *Management science 50*(12 supplement), 1861–1871.

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on 23*(10), 1498–1512.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Goffin, J. (1977). On convergence rates of subgradient optimization methods. *Mathematical Programming 13*(1), 329–347.

Golub, B. and Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 112–149.

Goodman, N. D., Ullman, T. D., and Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review 118*(1), 110.

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.

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.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology 83*(6), 1420.

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research 12*(3), 450–459.

Held, M. and Karp, R. M. (1970). The traveling-salesman problem and minimum spanning trees. *Operations Research 18*(6), 1138–1162.

Held, M., Wolfe, P., and Crowder, H. P. (1974). Validation of subgradient optimization. *Mathematical Programming 6*(1), 62–88.

Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. (2011). Seeding strategies for viral marketing: an empirical comparison. *Journal of Marketing 75*(6), 55–71.

Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science 30*(2), 195–212.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge university press.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly 21*(1), 61–78.

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.

Kempe, D., Kleinberg, J., and Tardos, É. (2005). Influential nodes in a diffusion model for social networks. In *Automata, Languages and Programming*, pp. 1127–1138. Springer.

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.

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

Lu, Y., Jerath, K., and Singh, P. V. (2013). The emergence of opinion leaders in a networked online community: A dyadic model with time dynamics and a heuristic for fast estimation. *Management Science 59*(8), 1783–1799.

Macy, M. W. (1991). Chains of cooperation: Threshold effects in collective action. *American Sociological Review*, 730–747.

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.

Manchanda, P., Xie, Y., and Youn, N. (2008). The role of targeted communication and contagion in product adoption. *Marketing Science 27*(6), 961–976.

Merton, R. K. (1947). Selected problems of field work in the planned community. *American Sociological Review*, 304–312.

Nam, S., Manchanda, P., and Chintagunta, P. K. (2010). The effect of signal quality and contiguous word of mouth on customer acquisition for a video-on-demand service. *Marketing Science 29*(4), 690–700.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical Review E 66*(1), 016128.

Padgett, J. F. and Ansell, C. K. (1993). Robust action and the rise of the medici, 1400-1434. *American Journal of Sociology*, 1259–1319.

Pan, F. and Nagi, R. (2013). Multi-echelon supply chain network design in agile manufacturing. *Omega 41*(6), 969–983.

Peres, R., Muller, E., and Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing 27*(2), 91–106.

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.

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.

Siomina, I., Värbrand, P., and Yuan, D. (2007). Pilot power optimization and coverage control in wcdma mobile networks. *Omega 35*(6), 683–696.

Susarla, A., Oh, J.-H., and Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research 23*(1), 23–41.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 807–816. ACM.

Tansel, B. C., Francis, R. L., and Lowe, T. J. (1983). State of the artlocation on networks: A survey. part i: The p-center and p-median problems. *Management Science 29*(4), 482–497.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological Bulletin 110*(1), 67.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences 10*(7), 309–318.

Trigeiro, W. W., Thomas, L. J., and McClain, J. O. (1989). Capacitated lot sizing with setup times. *Management Science 35*(3), 353–366.

Tucker, C. (2008). Identifying formal and informal influence in technology adoption with network externalities. *Management Science 54*(12), 2024–2038.

Valente, T. W., Frautschi, S., Lee, R., OKeefe, C., Schultz, L., Steketee, R., Chitsulo, L., Macheso, A., Nyasulu, Y., Ettling, M., *et al.,* (1994). Network models of the diffusion of innovations. *Nursing*

Times 90(35), 52–3.

Van den Bulte, C. and Joshi, Y. V. (2007). New product diffusion with influentials and imitators. *Marketing Science 26*(3), 400–421.

Van den Bulte, C. and Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort1. *American Journal of Sociology 106*(5), 1409–1435.

Wang, C., Chen, W., and Wang, Y. (2012). Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery 25*(3), 545–576.

Wasserman, S. (1994). *Social network analysis: Methods and applications*, Volume 8. Cambridge university press.

Watts, D. J. and Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research 34*(4), 441–458.

Wejnert, B. (2002). Integrating models of diffusion of innovations: a conceptual framework. *Sociology 28*(1), 297.

Whyte Jr, W. H. (1954). The web of word of mouth. *Fortune 50*(1954), 140–143.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review 114*(2), 245.

Xu, J. and Nagi, R. (2013). Solving assembly scheduling problems with tree-structure precedence constraints: A Lagrangian relaxation approach. *Automation Science and Engineering, IEEE Transactions on Automation Science and Engineering 10*(3), 757–771.

Yoganarasimhan, H. (2012). Impact of social network structure on content propagation: A study using youtube data. *Quantitative Marketing and Economics 10*(1), 111–150.

Young, H. P. (2001). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 452–473.

Appendix A. Proof of Theorem 1.

PCIM problem is shown to be NP-hard by a polynomial Turing reduction from the Maximum Coverage Problem (MCP), also referred to as the max k-cover problem or set k-cover problem in the literature (Feige 1998). The objective of MCP is to select a group of sets, where some sets have common elements, such that the total number of selected sets is less than the predefined limit and the total number of selected elements is maximized. MCP is first formally stated and then, the reduction from PCIM to MCP is presented.

Maximum Coverage Problem

INSTANCE: A number $k > 0$ and a a collection of sets $J = J_1, J_2, ..., J_m$.

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

Given an arbitrary instance of MCP, define a particular instance of PCIM as follows: Assume the Influence Graph $G(N, A)$ is given and let $T = 1$, $|N| = m$, $|S^+| = k$ and $|S^-| = 0$. Let $e^+ > \max \theta_i^+; i = 1, 2, ..., |N|$, $e^- = 0$, and lastly, set $\alpha^+ = \alpha^- = \beta^+ = \beta^- = 1$. Set J_i for $i = 1, 2, ..., m$ can be defined such that $j \in J_i$ iff $j = i$ or $(i, j) \in A, j = 1, 2, ..., N$ (all the nodes in the first hop of node i). This transformation can be performed in polynomial time in the size of the arbitrary instance of the MCP.

In order to show that an optimal solution to PCIM problem maps to an optimal solution of MCP, let X_{i0}^* for $i = 1, 2, ..., |N|$ ($X_{i0} \in \{0, 1\}$) to be an optimal solution to PCIM problem. Then, $\sum_{i=1}^{|N|} X_{i0} \leq |S^+|$, $X_{jT} \leq \sum_{(i,j)\in A} X_{i,0} + X_{j0}$ for $j = 1, 2, ..., |N|$, $Y_{it} = 0$ for $i = 1, 2, ..., |N|, t = 0, 1, ..., T$ and $\sum_{i=1}^{|N|} \sum_{t=0}^{T} (X_{it} - Y_{it})$ is maximized. The claim is that X_{i0}^* is an optimal solution for MCP. Note that X_{i0}^* for $i = 1, 2, .., |N|$ is a feasible solution for MCP because $\sum_{i=1}^{|N|} X_{i0} \leq |S^+| = k$.

Suppose there exists such solution to MCP $\bar{X}_{i0}$ for $i = 1, 2, .., |N|$ that $|\bigcup_{J_i \in \bar{J}'} J_i| > |\bigcup_{J_i \in J'^*} J_i|$. Solution $\bar{X}_{i0}$ for $i = 1, 2, .., |N|$ is a feasible solution for PCIM: $\sum_{i=1}^{|N|} \bar{X}_{i0} \leq |S^+| = k$. Therefore, the PCIM objective function for this solution is $\sum_{i=1}^{|N|} \sum_{t=0}^{T} (\bar{X}_{it} - \bar{Y}_{it}) = |\bigcup_{J_i \in \bar{J}'} J_i| + k > |\bigcup_{J_i \in J'^*} J_i| + k = \sum_{i=1}^{|N|} \sum_{t=0}^{T} (X_{it}^* - Y_{it}^*)$, which is a contradiction. Thus, X_{i0}^* for $i = 1, 2, .., |N|$ is an optimal solution for MCP. ■

Appendix B. Subgradient Search Algorithm for the Lagrangian Dual Problem

The Lagrangian dual problem (LD_u) is presented in Section 5.1 for finding the best (lowest) upper bound for the optimal solution of (P). Since $Z^{LD_u}(u)$ is non-differentiable, the subgradient of this function is employed in the implementation of a search algorithm for finding the improved multipliers.

DEFINITION 1: Vector s is a subgradient of $Z^{LD_u}(u)$ at point u' if:

$$Z^{LD_u}(u) \leq Z^{LD_u}(u') + s(u - u'), \qquad \forall u. \tag{28}$$

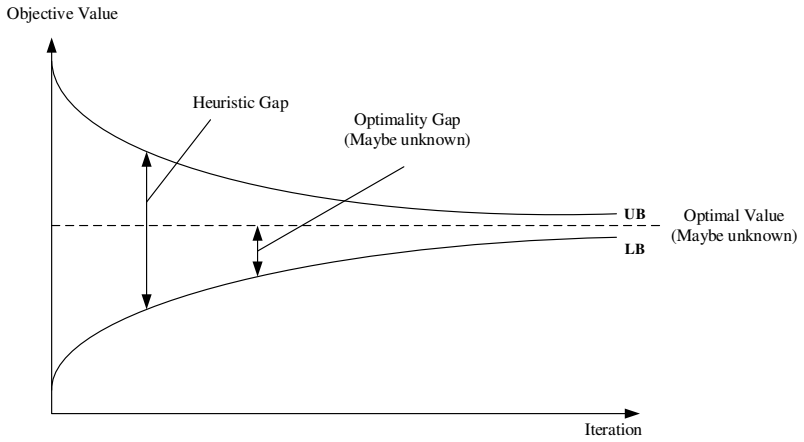


Figure 8: The Heuristic and optimality gaps achieved with the Subgradient algorithm.

A multiplier u^* is optimal for $(LD_u(u))$ iff subgradient of $Z^{LD_u}(u^*)$ is zero. At iteration k of the Subgradient search algorithm, the subgradient can be expressed as

$$s^k = \sum_i X_{i0}^k - |S^+|, \tag{29}$$

where $X_{i0}^k, i = 1, 2, ..., N$, are the optimizers of (LR_{u^k}). According to Fisher (2004), the iterative Subgradient search algorithm for generating the sequence of Lagrangian multipliers u^k, given an initial value u^0, is defined as

$$u^{k+1} = u^k - l_k(s^k), \tag{30}$$

where l_k denotes the step size and $Z^{LD_u}(u^k) \rightarrow Z^{*LD_u}$ if $l \rightarrow 0$ with $\sum_{i=0}^k l_i \rightarrow \infty$ (Goffin 1977). As l_k approaches zero, it is guaranteed that the Subgradient algorithm does not overstep u^*. Note that the summation of step size values approaches positive infinity, which, theoretically, guarantees the convergence to u^*. At the end of each iteration of the Subgradient algorithm, the step size value can be updated using the quality of the solution obtained for (LD_u) at the same iteration,

$$l_k = \frac{\lambda_k(Z^{LD_u}(u^k) - Z')}{\| s^k \|^2}, \tag{31}$$

where λ_k is a positive scalar for the step size and Z' is a lower bound for Z^{LD_u}. The appropriate range of values for λ_k can be defined experimentally; the range $0 < \lambda_k \leq 2$ has been found to work well in practice (Fisher 2004). The maximum value in the selected range for the step size is assigned to the initial value of the step size (λ_0), and it is split when Z^{LD_u} fails to decrease for a given number of consecutive iterations of the Subgradient algorithm (Held *et al.,* 1974).

There is no mathematical proof for the optimality in the Subgradient algorithm. As (P) is a maximization problem, the Subgradient algorithm stops when the gap between the lower bound, obtained by the ISR and ASB algorithms presented in Section 4.2, and the upper bound, obtained by the Subgradient

algorithm, becomes less than a preselected threshold value, which guarantees the quality of the solutions for Lagrangian Relaxation heuristic. Alternatively, the Subgradient algorithm can be terminated after a predetermined number of iterations have been executed or a predefined run-time limit has been reached (Trigeiro *et al.,* 1989; Xu and Nagi 2013). In this paper, the quality of the gradually improved solutions drives the stopping criteria for the Subgradient algorithm to obtain a guaranteed-performance heuristic method for solving (P). Figure 8 shows how the heuristic and optimality gaps change in the Lagrangian Relaxation method to reduce the gap around the optimal solution.