# Efficient Influence Maximization Under Network Uncertainty

Soheil Eshghi[1], Setareh Maghsudi[2], Valerio Restocchi[3,4], Sebastian Stein[3], and Leandros Tassiulas[1]

[1]School of Engineering and Applied Science, Yale University, USA
[2]Department of Electrical Engineering and Computer Science, Technical University of Berlin, Germany
[3]Electronics and Computer Science, University of Southampton, UK
[4]School of Informatics, The University of Edinburgh, UK

***Abstract*—We consider the influence maximization (IM) problem in a partially visible social network. The goal is to design a decision-making framework for an autonomous agent to select a limited set of influential seed nodes to spread a message as widely as possible across the network. We consider the realistic case where only a partial section of the network is visible to the agent, while the rest is one of a finite set of known structures, each with a given realization probability. We show that solving the IM problem in this setting is NP-hard, and we provide analytical guarantees for the performance of a novel computationally-efficient seed-selection approximation algorithm for the agent. In empirical experiments on real-world social networks, we demonstrate the efficiency of our scheme and show that it outperforms state-of-the-art approaches that do not model the uncertainty.**

*Keywords*: Influence Maximization, Social Networks, Partial Visibility, Uncertainty.

## I. Introduction

Harnessing the power of word-of-mouth communication to maximize the spread of information across a social network has shown great promise in diverse fields [1], [2]. To do this effectively, existing work has extensively studied how to design autonomous agents to solve the problem of influence maximization (IM) [3], [4]. Specifically, given a social network and an influence model, an agent seeks to select a set of influential seed nodes that maximizes the expected spread of information. The existing studies typically assume that the agent is aware of the complete network; however, in practice, such a scenario is unlikely. In fact, due to the noisy and imperfect data, the agent has only partial estimates of the network [5], [6]. When taking uncertainty into account, prior work has typically focused only on link uncertainty [7], has assumed the entire network is initially unknown [8], [9] or has assumed that the network has well-defined community structure [10].

To address this shortcoming, in this paper we assume that only a part of the network is visible, with the rest of the network taking one of several given structures with given probabilities. The influence maximizing agent seeks to select seeds within the visible part of the network, in order to maximize the expected spread of the message in the network. We also consider the case where the agent seeks to provide reasonable guarantees on the spread of a seed set regardless of which of the structures is the correct one.

Given the aforementioned setting, we extend the state of the art as follows: Based on the well-known algorithm *influence maximization via martingales (IMM)* [11], we develop a method for influence maximization under the aforementioned type of uncertainty. We show that the problem of seed selection for influence maximization under partial visibility is NP-hard, and the best approximation algorithm for seed selection can only provide a $(1-\frac{1}{e})$-optimality guarantee [12]. We also show how to efficiently compute a $(1-\frac{1}{e}-\epsilon)$-optimal seed set with high probability in the case where the expected influence of seed sets is not available and has to be estimated computationally. We then discuss performance degradation compared to the full-information setting. Finally, we run simulations on a real-world network to show the robustness of our algorithm. Our results suggest that the proposed algorithm performs well under different types of uncertainty, and also outperforms the state of the art (IMM) algorithm.

## II. Related Work

### A. Classical Setting

In the classical influence maximization problem, it is assumed that the network is completely *visible*, i.e., all possible links between network entities are known, and all network entities are available to the influence maximizer as potential seeds. In addition to the structure of the network, it is assumed that the activation function is also known, which describes the spread of influence among nodes. The IM problem is to select a specific number of most influential nodes, so as to maximize the influence spread in expectation. In their seminal work, Kempe et al. [3] formulate the influence maximization problem as a combinatorial optimization problem and show its NP-hardness. Under a variety of diffusion models, they developed a greedy approach, which achieves $1-1/e-\epsilon$ bound compared to the optimal solution. However, this approach was dependent on estimating the expected influence of each subset of network entities, which is by itself $\#P$-Hard [13], and thus a brute-force approach to computing its values will negate the gains from efficient seed selection. As a result, many near-optimal approximate solutions have been developed which are more computationally efficient in the expectation computation phase. Among these, the most important contribution was by Borgs et al. [13], who proposed the use of random Reverse Reachable (RR) sets for more efficient computation of an unbiased estimator for the influence function.

The key remaining question for the Borgs et al. [13] approach is how many RR sets (samples of network realizations) are needed to provide performance guarantees for the end-to-end IM algorithm. The most efficient algorithms to date are TIM/TIM+ [14], and IMM [11]. These two algorithms use tail-bounds for the estimation of the probability that the influence estimator is markedly different from the actual function, with the difference that [11] provides stronger bounds (and thus needs fewer RR sets/samples to guarantee the same level of performance). In both cases, these performance bounds depend on the (unknown) expected influence of the optimal size-$k$ seed set. In TIM/TIM+ [14], this value is lower-bounded from observations to provide a looser bound, while it is estimated more efficiently in IMM [11], providing even more computational efficiency. The technical approach of our work broadly follows IMM [11], but diverges in its domain of application and sought insights.

### B. *Partial Information Setting*

As mentioned previously, to date, a great majority of the research investigating the influence maximization problem has focused on the full-information case, where all the necessary information is available prior to seed selection. However, in many real-world applications, such information is extremely costly to acquire, if possible at all. Moreover, often the widely-used assumptions on the network structure do not match real networks, thereby limiting the generality of results. To address such shortcoming, influence maximization under uncertainty is an emerging topic of study. This problem corresponds to selecting a set of nodes to maximize influence spread when the network (or its characteristics) are only partially visible. For instance, the probability of influence among each pair of nodes might be unknown. Below we provide a brief review of the most important literature in this area.

In [15], for instance, the uncertainty lies in the edge influence probability. More precisely, in the considered model, instead of the exact probability of influence, every edge of the graph is associated with some interval, to which the true probability belongs. The problem is solved by sampling the edges uniformly or adaptively, followed by estimating the true influence probability using the sample set. In the former, the edges are sampled uniformly at random, whereas in the latter, critical edges (such as those with possibly-higher influence probability) are sampled more often. Clearly, the adaptation is performed according to the historical results of sampling. Uncertainty with respect to the diffusion probability is also considered in [7] and [16]. In the former, the problem is formulated as a maximum likelihood decision, whereas in the latter, bandit theory is used to address it. Generally-speaking, a bandit problem can be categorized as an online optimization problem. In the most seminal setting of this problem, an agent selects one of the available options sequentially, without having any prior information. The goal is to maximize the average (discounted) reward or to minimize the average regret. A similar work is [17], where a combinatorial bandit model is used. In [8], the authors define the uncertainty in terms of partial observability of the network structure. Influence maximization for unknown social networks is also investigated in [6], [9], [18].

## III. System Model

In this section, we first formalize the information that is available to the agent (the *influence maximizer*), then describe the influence model we assume and finally summarize the IM problem under partial visibility.

### A. *Information Structure*

Assume that a directed graph $G = (V, E)$ exists such that $|V| = n$. However, only a subset of its nodes, as well as the edges between such nodes, are *visible* to the agent. That is, we have $G_v = (V_v, E_v)$ such that $V_v \subseteq V$ and $E_v \subseteq V_v \times V_v$, with the condition that if $e = (b, c) \in E$ and $b, c \in V_v$, then $e \in E_v$. $G_v$ is the only part of the network that is known to the influence maximizer, corresponding to the real-world scenario where accurate information is only known about a part of the network. We assume some information is available about the graph:

- The number of nodes $n$ is known.
- The structure of the unobserved part of the network $G = (V, E')$ is one of $M$ different known graphs $G_1, G_2, \ldots, G_M$, each with probability $q_1, q_2, \ldots, q_M$ (such that $\sum_{i=1}^{M} q_i = 1$). Note that in this case, if $e = (b, c) \in E'$, then either $b \notin V_v$ or $c \notin V_v$.

Despite being relatively strong, this latter assumption makes our approach applicable to the settings where a generative model can produce a finite set of samples of the unseen part of the network [19] or where multiple real-world sources of information (surveys, OSNs) are used without knowing which represents the real method of influence spread. In these settings, priors for $q_i$'s can be obtained based on the past success of the various methods/sources.

We consider a discrete-time model, where decisions by the influence maximizer are made at time 0. The system evolves from then on according to the influence model that is described in the following section.

### B. *Influence Model*

We consider the Independent Cascade (IC) and the Weighted Cascade (WC) models of influence propagation. In the IC influence model, each edge $e = (b, c)$ of a graph structure is associated with a probability $p(e)$ that determines the chance that $u$ successfully influences $v$ at time $t + 1$, provided that it has been itself influenced at time $t$. In the WC model, this probability is modulated by the in-degree of $v$ (i.e., the total probability of being influenced by neighbors does not increase with its number of neighbors). Following the classical assumptions of [3], the attempt for influence only occurs once at time $t+1$, and if unsuccessful, that link will not play a role in the future spread of influence.

At time 0, the agent activates a node set $S \subseteq V_v$, with all other nodes being inactive. These newly activated nodes each sample from among their outgoing edges to find neighbors according to the relevant probabilities $p(e)$ and, if successful, activate them at the next time instance. Activated nodes remain active henceforth, while only *newly activated nodes* attempt to influence their neighbors (for one time-step). Under this model, we define the function $I(\cdot) : 2^{|V_v|} \rightarrow [0, n]$ that maps each subset of visible nodes (i.e., possible seed nodes) to the number of influenced nodes once the aforementioned process has concluded. Note that $I(S)$ (such that $S \subseteq V_v$) is a random variable, as its value depends on the end result of the activation pattern (the so-called live-edge graph $X$ [3]). The goal of the influence maximizer is to maximize the expected spread $\mathbb{E}\{I(S)\}$ for a seed $S \subseteq V_v$ of size $k$.

### C. Problem Statement

**Problem 1** (PV-IM). *Given visible graph $G_v$, possible realizations of the unobserved graph $G_1, G_2, \ldots, G_M$ (with probabilities $q_1$, $q_2$,..., $q_M$) a diffusion model $D \in \{IC, WC\}$ and influence functions $p(e)$ for each edge $e$ in the realizations, and a positive integer $k$, find a node-set $S_k^o$ of size $k$ that maximizes $\mathbb{E}\{I(S_k^\circ)\}$.*

## IV. Influence Maximization under Partial Visibility

We show (in §IV-A) that Problem 1 is NP-hard and that its objective is sub-modular and monotone, and thus it only admits a $(1 - \frac{1}{e})$-optimal approximate solution. The rest of this section deals with how to approach this approximate solution given that the influence of each set of seeds has to be computationally evaluated before a solution to Problem 1 can be attempted.

### A. Complexity of PV-IM

**Theorem 1.** *The PV-IM problem is NP-hard and has a monotone and submodular objective. A greedy selection of seeds gives a $(1 - \frac{1}{e})$-approximation of the optimal result.*

*Proof.* Define $\sigma_X^m(S)$ to be the function that maps each subset of visible nodes (seed set $S$) to the spread resulting from them in live-edge realization $X$ of graph $G_m \cup G_v$. In [3], it was shown that for both the $IC$ and $LT$ influence models, $\sigma_X^m(S)$ is monotone and sub-modular, and therefore $\sigma^m(S)$, the expected spread under graph $G_m \cup G_v$ is also monotone and sub-modular (as it is a weighted linear sum of the $\sigma_X^m(S)$ terms). It is also shown that both problems are NP-hard (by a mapping to Set-Cover and Vertex-Cover, respectively). A more general result applying for a wider variety of triggering functions (including WC) was obtained by [20].

First, note that if $M = 1$ (and $q_1 = 1$), our problem maps to the original IM problem, which is NP-hard. Therefore, partially visible IM is also NP-hard. Furthermore, if we define $\sigma(S)$ to be the *a priori* expected spread from $S$ among all graph realizations, then $\sigma(S) = \sum_{m=1}^{M} q_m \sigma^m(S)$, which is a linear, positively-weighted sum of monotone, sub-modular functions. Therefore, $\sigma(S)$ is also still monotone and sub-modular, and given the function $\sigma(\cdot)$ the greedy seed selection algorithm will give a $(1 - \frac{1}{e})$-approximation to the optimal solution [12]. □

This proof, however, assumed that $\sigma(\cdot)$ is available to the greedy algorithm, which is not a reasonable assumption as the computation of this function is itself $\#P$-hard [13]. Thus, in the next section, we develop computationally-tractable ways to approximate $\sigma(\cdot)$, and to provide guarantees for the performance of a partially-visible IM algorithm considering the two sources of approximation errors.

### B. Performance Guarantees for PV-IM

We now characterize ways of providing reasonable approximations for $\sigma(\cdot)$ given the uncertainty in the realization of the invisible part of the network. We first adapt the notion of Reverse Reachable sets [13] to the partial visibility setting, and then show how many RR sets are necessary for adequate performance guarantees on the greedy seed selection algorithm.

**Definition 1.** *A Reverse Reachable (RR) Set $P^i(q)$ for a node $q$ in graph realization $G_i \cup G_v$ is the set of nodes that can reach $q$ in a sampled live-edge graph $g_i$ created from $G_v \cup G_i$. A Random Reverse Reachable Set $P^i$ is an RR set generated for a random node $q$ in $G_v \cup G_i$.*

Let $S$ be any set of nodes in $V_v$. Define $x_j^i(S) \in \{0, 1\}$ to be the indicator variable that determines whether the $j-$th random RR set $P^i$ generated for graph realization $G_v \cup G_i$ overlaps with any of the nodes in $S$.

Borgs et al. [13] showed that the probability that $S$ influences a node $v$ when $G_i$ is the structure of the unobserved graph (the generic setting with an observable graph) is equal to the probability that $S$ overlaps with an RR set for graph $G_v \cup G_i$ generated for $v$. Therefore, $\mathbb{E}\{I(S)|G_i\} = n.\mathbb{E}\{x^i(S)|G_i\}$, where $x^i(S)$ is a *random* RR set (generated uniformly at random for some $v \in V$) and the expectations are conditional on the structure of the invisible part of the network being $G_i$. Thus, if we have $\theta_i$ random RR sets generated for structure $G_i$ respectively for all $i$, by the linearity of expectation we will have:

$$\mathbb{E}\{I(S)|G_i\} = \frac{n}{\theta_i}.\mathbb{E}\{\sum_{j=1}^{\theta_i} x_j^i(S)|G_i\}. \quad (1)$$

Thus, using the law of total expectation:

$$\mathbb{E}\{I(S)\} = \sum_{i=1}^{M} q_i \frac{n}{\theta_i}.\mathbb{E}\{\sum_{j=1}^{\theta_i} x_j^i(S)|G_i\}. \quad (2)$$

Define $F_R^i(S) := [\sum_{j=1}^{\theta_i} x_j^i(S)]/\theta_i$, the fraction of RR sets generated for $G_v \cup G_i$ that overlap seed-set $S$, with $F_R(S) := \sum_{i=1}^{M} q_i F_R^i(S)$ being the weighted average of these fractions. By (1), $n.F_R(S)$ is an unbiased estimator of $\mathbb{E}\{I(S)\}$.

For the greedy seed selection algorithm to be able to use this approximation to deliver the sought-after $(1 - \frac{1}{e} - \epsilon)$- approximate solution with high probability, we need this estimator of the influence function to be close enough to its expected value that the greedy algorithm can only choose a seed-set within the acceptable performance range. In TIM/TIM+ [14] and IMM [11], this is accomplished in two steps: 1) by choosing enough RR sets to ensure that the estimator of the influence of the optimal size-$k$ seed set is close to its expected value with high probability, and 2) by (if necessary) choosing more RR sets to ensure that the influence estimators of all "extremely sub-optimal" (i.e., worse than $(1-\frac{1}{e}-\epsilon)$-approximation) seed sets will also be close to their expected values. Looking at the union of these two conditions, we can then say that the size-$k$ seed set chosen by the greedy algorithm, $S_k^g$, will only *not* be within the required $(1-\frac{1}{e}-\epsilon)$ ratio if either condition fails, the probability of which can be upper-bounded by the sum of each condition's failure probabilities. We formalize these intuitions following the outline of [11]:

**Definition 2.** *Let $p^\circ$ be the expected influenced fraction of nodes for the optimal size-k seed set $S_k^\circ$, $p^\circ := \mathbb{E}\left[I(S_k^\circ)\right]/n$.*

**Definition 3.** *Let $p_i^\circ$ be the expected influenced fraction of nodes for the optimal seed set selection when the invisible part of the network has structure $G_i$: $p_i^\circ := \mathbb{E}\left[I(S_k^\circ)|G_i\right]/n$.*

Therefore,

$$p^\circ = \mathbb{E}\left[I(S_k^\circ)\right]/n = \sum_{i=1}^{M} q_i p_i^\circ. \tag{3}$$

We now specify how many sample RR sets, denoted $\theta$, are needed to guarantee that $F_R(S_k^\circ)$ is close to $p^\circ$ with high probability, and how they should be divided between the different possible realizations of the unobserved network:

**Lemma 1.** *For $\delta_1 \in (0,1)$, and $\epsilon_1 > 0$, if*

$$\theta^* := \frac{\sum_{i=1}^{M} q_i^2 p_i^\circ(1-p_i^\circ)}{\sum_{i=1}^{M} q_i^2 (p_i^\circ)^2} \frac{\log(1/\delta_1)}{\epsilon_1^2}, \tag{4}$$

*then for $\theta \geq \theta^*$, we have $n(F_R(S_k^\circ)) \geq (1-\epsilon_1).np^\circ$ with probability at least $1-\delta_1$, each realization being sampled*

$$\theta_i^* := \theta^* \frac{q_i \sqrt[2]{p_i^\circ(1-p_i^\circ)}}{\sum_{j=1}^{M} q_j \sqrt[2]{p_j^\circ(1-p_j^\circ)}} \tag{5}$$

*times.*

*Proof.* **Step 1**: We prove that:

$$\Pr\left[n(F_R(S_k^\circ)) \leq (1-\epsilon_1)np^\circ\right] = \Pr[\sum_{i=1}^{M}\sum_{k=1}^{\theta_i}(q_i \prod_{j\neq i}\theta_j)\cdot$$
$$[x_k^i(S_k^\circ) - p_i^\circ] \leq -\epsilon_1(\prod_{j=i}^{M}\theta_j)\sum_{i=1}^{M}(q_i p_i^\circ)], \tag{6}$$

by using the definition of $F_R^i(S_k^\circ)$, multiplying both sides inside the probability by $\prod_{j=1}^{M}\theta_j$, and grouping of terms.

**Step 2**: For $r \leq \theta := \sum_{j=1}^{M}\theta_j$, define: $m_r := \min\{l \in \{1,\ldots,M\} | \sum_{i=1}^{l}\theta_i \geq r\}$, and

$$Y_r := \sum_{i=1}^{m_r} \sum_{k=1}^{r-\sum_{d=1}^{m_r-1}\theta_d} (q_i \prod_{j\neq i}\theta_j)[x_k^i(S_k^\circ) - p_i^\circ]. \tag{7}$$

With this definition, (6) becomes

$$\Pr\left[n(F_R(S_k^\circ)) \leq (1-\epsilon_1).np^\circ\right] =$$
$$\Pr[Y_\theta \leq -\epsilon_1(\prod_{j=i}^{M}\theta_j)\sum_{i=1}^{M}(q_i p_i^\circ)]. \tag{8}$$

**Step 3**: We show that that the sequence $Y_1, Y_2, \ldots$ is a Martingale, and appeal to a Martingale tail-bound for $Y_\theta$ to upper-bound the right-hand side of (8):

**Lemma 2.** *[21, Theorem 18]: Let the sequence $Z_1, Z_2, \ldots$ be a Martingale and let there exist $a, b$ such that for some $i$ and for all $1 < j \leq i$, $|Z_1| \leq a$, $|Z_j - Z_{j-1}| \leq a$, as well as $Var[Z_1] + \sum_{k=2}^{i} Var[Z_k|Z_1,\ldots,Z_{k-1}] \leq b$. Then, for any $\gamma > 0$, we have $\Pr\left[Z_i - \mathbb{E}\{Z_i\} \geq \gamma\right] \leq \exp\left(\frac{-\gamma^2}{\frac{2}{3}\gamma a + 2b}\right)$.*

Appealing to Lemma 2 for $-Y_\theta$ and combining the resulting bound with (8) and simplifying, we obtain:

$$\Pr\left[n(F_R(S_k^\circ)) \leq (1-\epsilon_1)np^\circ\right]$$
$$\leq \exp\left(-\frac{\epsilon_1^2[\sum_{i=1}^{M}(q_i p_i^\circ)]^2}{2\sum_{i=1}^{M}\frac{q_i^2 p_i^\circ(1-p_i^\circ)}{\theta_i}}\right) \tag{9}$$

**Step 4:** We choose $\{\theta_1,\ldots,\theta_M\}$ to minimize such an error probability given a total number of possible RR sets $\theta$, leading to the following optimization problem:

$$\min \quad \sum_{i=1}^{M}\frac{q_i^2 p_i^\circ(1-p_i^\circ)}{\theta_i} \qquad \text{s.t.} \quad \sum_{i=1}^{M}\theta_i = \theta, \qquad \theta_i \geq 0$$

Solving the above using the KKT conditions, we will have

$$\theta_i^* = \theta^* \frac{q_i \sqrt[2]{p_i^\circ(1-p_i^\circ)}}{\sum_{j=1}^{M} q_j \sqrt[2]{p_j^\circ(1-p_j^\circ)}} \tag{10}$$

Replacing these values into an upper-bound of $\delta_1$ for (9) gives us the desired value of $\theta^*$ (4).

**Step 4$^*$:** If the different graphs $G_i \cup G_v$ have similar expected normalized influence spreads $p_i^\circ = p^\circ$, *the optimal number of RR sets to draw from each possible realization is directly proportional to its probability of occurrence ($\theta_i^* = q_i\theta^*$).* Furthermore replacing this result into (9) gives us the following:

$$\Pr\left[n(F_R(S_k^\circ)) \leq (1-\epsilon_1)np^\circ\right] \leq \exp\left(-\frac{\epsilon_1^2\theta^* p^\circ}{2}\right),$$

which recovers the same bound as in [11, Lemma 3] for small $p_i^\circ$. Therefore, we can see that uncertainty about the invisible part of the graph *does not, in many cases, degrade performance, a surprising result.* □

**Corollary 1.** *If the expected influence of the optimal set is equal in all graph realizations, it is optimal to sample each realization in accordance with its probability of occurrence.*

Replacing $p_i^\circ$ from Definition 3 in (4), we also see that:

**Corollary 2.** *$\theta^*$ increases at most linearly in network size $n$.*

This mirrors a result in [11]. It is important to note that these results are only possible through intelligently choosing the number of samples $\theta_i^*$ to allocate to each of the realizations, which is a contribution of this paper.

Now assume that these $\theta^*$ RR-sets are used to select a size-k seed $S_k^g$ set via a greedy approach. As the objective of this selection problem is also monotone and submodular, we will have the following guarantee with probability $1-\delta_1$:

$$nF_R(S_k^g) \geq (1-\frac{1}{e})nF_R(S_k^\circ) \geq (1-\frac{1}{e})(1-\epsilon_1).np^\circ \\ = (1-\frac{1}{e})np^\circ - \epsilon_1(1-\frac{1}{e}).np^\circ \tag{11}$$

Thus, the result of a greedy optimization on $\theta^*$ RR-sets can fall a factor of $(1-\frac{1}{e})(1-\epsilon_1)$ below the optimal value in the worst case with probability $1-\delta_1$. We now guarantee that no extremely sub-optimal seed set (much worse than a $(1-\frac{1}{e})$-approximation of the optimal seed-set) will be picked by a greedy algorithm: we show how many RR sets are needed so that that the estimators of all extremely sub-optimal seed sets will be close to their expected values (as in [11]).

Using the same methodology as in the proof of Lemma 1, we can bound the probability that the unbiased estimator of the sub-optimal set is very far (expressed as a multiple of the optimal value for computational reasons) from its expected value. First, we define what we mean by an extremely-sub-optimal seed set at level $\epsilon > \epsilon_1(1-\frac{1}{e}) > 0$:

**Definition 4.** *A size-$k$ seed set $S_k$ is* extremely sub-optimal *at level $\epsilon > 0$ ($ESO_\epsilon$) if $\mathbb{E}[I(S_k)] \leq (1-\frac{1}{e}-\epsilon)np^\circ = (1-\frac{1}{e})np^\circ - \epsilon np^\circ$.*

The number of such seed-sets is upper-bounded by $\binom{n}{k}$, the total number of size-$k$ seed sets. Therefore, if we bound the probability that each $ESO_\epsilon$ seed set is selected with probability $\delta_2/\binom{n}{k}$ for some $\delta_2 > 0$, then the total probability that an $ESO_\epsilon$ seed set is selected will be at most $\delta_2$. Alternatively, this means that with probability $1-\delta_2$, the selected node set will not be $ESO_\epsilon$. We now clarify how many RR-sets are needed so that each $ESO_\epsilon$ seed set is selected with probability $\delta_2/\binom{n}{k}$:

**Lemma 3.** *For $\delta_2 \in (0,1)$, $\epsilon > \epsilon_1(1-\frac{1}{e}) > 0$, if Lemma 1 holds,*

$$\theta' := \frac{2\log(\frac{\binom{n}{k}}{\delta_2})[(1-\epsilon_1)(1-\frac{1}{e})-\frac{2}{3}\epsilon]}{(\epsilon-\epsilon_1(1-\frac{1}{e}))^2p^\circ}, \tag{12}$$

*and $\theta'_i := q_i\theta'$ (sampling according to realization probabilities), then for $\theta \geq \theta'$ and $ESO_\epsilon$ seed set $k$, we will have $\mathbb{E}\{I(S_k^g)\} \geq (1-\frac{1}{e}-\epsilon).\mathbb{E}\{I(S_k^\circ)\}$ with probability at least $1-\delta_2$.*

*Proof.* **Step 1:** As $\theta \geq \theta^*$, from (11) $nF_R(S_k^g) \geq (1-\frac{1}{e})np^\circ - \epsilon_1(1-\frac{1}{e})np^\circ$. We bound the probability that $nF_R(S_k) \geq (1-\frac{1}{e})np^\circ - \epsilon_1(1-\frac{1}{e})np^\circ$ for $S_k \in ESO_\epsilon$ (i.e., that such a set could be chosen by the greedy algorithm), which is the complement of the statement in the lemma. As $\mathbb{E}\{I(S_k)\} \leq (1-\frac{1}{e}-\epsilon)np^\circ$, we will have

$$nF_R(S_k) - \mathbb{E}\{I(S_k)\} \geq np^\circ\epsilon_2, \tag{13}$$

where $\epsilon_2 := (\epsilon - \epsilon_1(1-\frac{1}{e})) > 0$.

**Step 2:** We first make to definitions:

**Definition 5.** *Define $p$ to be the expected influenced fraction of nodes resulting by the size-$k$ seed set $S_k \in ESO_\epsilon$: $p := \mathbb{E}[I(S_k)]/n \leq (1-\frac{1}{e})p^\circ - \epsilon p^\circ$.*

**Definition 6.** *Define $p_i$ to be the expected influenced fraction of nodes resulting from seed set $S_k$ when the invisible part of the network has structure $G_i$:*

$$p_i := \mathbb{E}[I(S_k)|G_i]/n. \tag{14}$$

We then show that for $\gamma := \frac{\epsilon_2 p^\circ}{p}$:

$$\Pr[nF_R(S_k) - \mathbb{E}\{I(S_k)\} \geq np^\circ\epsilon_2] = \tag{15}$$
$$\Pr\left[\sum_{i=1}^{M}\sum_{k=1}^{\theta_i}(q_i\prod_{j\neq i}\theta_j)[x_k^i(S_k)-p_i] \geq \gamma(\prod_{j=i}^{M}\theta_j)\sum_{i=1}^{M}q_ip_i\right]$$

Note that $\gamma \geq \frac{\epsilon_2}{(1-\frac{1}{e}-\epsilon)}$ as $S_k \in ESO_\epsilon$.

**Step 3:** We define $m_r$ and $Y_r$ as in (7) (but with $S_k$ replacing $S_k^\circ$ and $p$ replacing $p^\circ$) and appeal to Lemma 2 again to bound the right-hand side of (15):

$$\Pr[nF_R(S_k) - \mathbb{E}\{I(S_k)\} \geq \epsilon_2 np^\circ] \\ \leq \exp(\frac{-\gamma^2p^2}{2\sum_{i=1}^{M}\frac{q_i^2p_i(1-p_i)}{\theta_i} + \frac{2}{3}\gamma p[\max_l \frac{q_l}{\theta_l}]}) \tag{16}$$

**Step 4:** We again choose $\{\theta_1,\ldots,\theta_M\}$ to minimize the right-hand side of (16), equivalent to minimizing the denominator of the exponent. Obtaining a closed form in this case is difficult, but we can see that the first term is minimized by (10) (with $\theta'$ replacing $\theta^*$) and the latter by $\theta'_i = q_i\theta'$ for all $i$, which may be similar if $p_i$ is the same over all realizations.

**Step 5:** We obtain an upper-bound to the right-hand side of (16), using $\theta'_i = q_i\theta'$:

$$\Pr[nF_R(S_k) - \mathbb{E}\{I(S_k)\} \geq \epsilon_2 np^\circ] \leq \exp(\frac{-\gamma^2p\theta'}{2+\frac{2}{3}\gamma}).$$

Setting the right-hand side to less than $\frac{\delta_2}{\binom{n}{k}}$ leads to (12) after suitable replacements, as $\gamma p = \epsilon_2 p^\circ$, $\epsilon_2 = (\epsilon - \epsilon_1(1-\frac{1}{e})) > 0$, and $\gamma \geq \frac{\epsilon_2}{(1-\frac{1}{e}-\epsilon)}$.

Therefore, in this setting too *the optimal number of RR sets to draw from each possible realization of the invisible part of the graph is closely proportional to its probability of occurrence ($\theta_i^* = q_i\theta$ is close to optimal).*

Thus, among the unbiased estimators we consider, the most efficient (in terms of guarantees per number of RR set samples) samples each possible $G_i$ in proportion to its probability of occurrence $q_i$. □

Putting Lemmas 1 and 3 together, we state the following theorem:

**Theorem 2.** *For $\epsilon > \epsilon_1(1-\frac{1}{e}) > 0$, $\delta_1, \delta_2 \in (0,1)$ and $\delta_1 + \delta_2 = \delta$, if $\theta_i > \max\{\theta_i^*, \theta'_i\}$ for all $i$, then our greedy algorithm will return a $(1-\frac{1}{e}-\epsilon)$ solution to Problem 1 with probability at least $1-\delta$.*

Algorithm 1 describes the steps needed to solve Problem 1. First, the requisite number of RR sets are created for each realization to match both Lemma 1 and Lemma 3, allowing the algorithm to achieve the bound in Theorem 2. We then describe the greedy approximation algorithm that once applied to the empirically generate RR coverage fraction functions (with suitable adjustments) will give us the $(1-\frac{1}{e}-\epsilon)-$approximation performance with high probability.

**Algorithm 1** Pseudocode of Partial Visibility IM Agent.

```
**Data:** $G_v$, $G_1, G_2, \ldots, G_M$ ($q_1$, $q_2$,..., $q_M$), $D$, $p(e)$, $k$, $\delta$, $\epsilon$.
**Result:** Algorithm for picking $S_k^g$ seed set, which is $(1-\frac{1}{e}-\epsilon)$-optimal with probability $1-\delta$.
Set $F_R(S)=0$, $F_R^i(S)=0$ for all sets, all $i$, $S_k^g=\emptyset$;
for $i=1,\ldots M$ do
    for $j=1,\ldots \max\{\theta_i^*, \theta_i'\}$ do
        Generate a random RR set for $G_v \cup G_i$;
        Update $F_R^i(S)$ for all nodes;
    end for
    Update $F_R(S)$ for all nodes;
end for
for $i=1,\ldots k$ do
    Add node $q$ with maximum $F_R(q)$ to seed set $S_k^g$;
    for $j=1,\ldots M$ do
        for $l=1,\ldots \max\{\theta_i^*, \theta_i'\}$ do
            if $x_l^j(q)=1)$ then
                Decrease $F_R^i(S)$;
            end if
        end for
        Update $F_R(S)$ for all nodes;
    end for
end for
Return $S_k^g$;
```

**Corollary 3.** *Note that $\theta^*$ & $\theta'$ in Lemmas 1 and 3 do not, in general, scale with the number of candidate realizations $M$.*

However, when $M$ is exponentially large, it is inescapable that the number of total RR-set samples may increase exponentially. This latter case is less amenable to the PV-IM approach, as stated in the introduction.

## V. Simulations

We showed analytically that in certain cases, it is optimal to sample RR sets from realizations in accordance with the realization's probability of occurrence given a fixed total number of RR sets. Here, we compare the empirical performance of such an approach with more naive variants. Note that as our agent was designed with performance guarantees in expectation in mind, it is not guaranteed to be optimal in any particular case. However, we have seen experimentally that in many real data-sets, it performs just as well as other widely-used heuristics.

### A. Experimental Setting

To illustrate the impact of sampling proportions on the spread of influence, we run simulations on two different networks, $G_1$ and $G_2$. $G_1$ has a probability $q_1$ to be the true network, and probability $q_1^* = q_1 + a$ to be sampled when computing an RR set. Also, to model uncertainty in $G_2$, we rewire $G_1$ by removing existing links and reassigning them to other nodes.

We present only results obtained by using a network with negative assortativity to model $G_1$, random rewiring to generate $G_2$, and 10% partial visibility for both networks. Specifically, we choose to use NetHept[1], a network that consists of 15k nodes and 31k edges, representing citations in the high energy physics community, as it has been widely used to show the performance of influence maximization algorithms [14], [22]. Several other combinations of data sources, rewiring methods, and visibility parameters were tested and yielded qualitatively similar results.[2]

Each simulation consists of four steps. First, NetHept is assigned to represent $G_1$, and $G_2$ is generated according to the random rewiring algorithm. Second, $\theta$ RR sets are created, with each being created from $G_1$ (resp. $G_2$) with probability $q_1^*$ (resp. $1-q_1^*$). Third, the seeds are chosen greedily by the algorithm. Fourth, the *realization* of the uncertain network is chosen between $G_1$ and $G_2$ with probability $q_1$ and $1-q_1$, respectively, and the influence process is simulated. This process is repeated T=1000 times for each value of $q_1$ and $q_1^*$, to ensure the robustness of the results.

### B. Benchmarks

To evaluate our algorithm, we compare its performance with those of three benchmarks:

- IMM: this algorithm, which represents the state of the art, samples $\theta$ RR sets on the network with the highest realization probability, and chooses those nodes with the highest expected spread [11].
- AverageDegree: the $k$ nodes with highest degrees $d_i$, computed as $d_i = d_{i,1}q_1 + d_{i,2}q_2$ when node i has degree $d_{i,1}$ in $G_1$ and $d_{i,2}$ in $G_2$, are selected as seed nodes
- Random Choice: all seed nodes are selected randomly.

### C. Results

Interestingly, although our results apply for worst-case guarantees and are approximate, we find that sampling RR sets with probabilities $q_m$ in this setting leads to a marginal improvement of the influence spread than under- and over-sampling $G_1$ (Figure 1). In other simulations on smaller networks, we found that undersampling sometimes leads to a slightly larger influence propagation (up to 10%). These results suggest that our algorithm is broadly robust to uncertainty in realization probabilities $q_m$ and that a good seed set will be provided even when $q_m$ are not estimated accurately.

Our algorithm outperforms IMM by up to 14.9%, when uncertainty on the network is highest (i.e., $q_1 = q_2 = 0.5$). Importantly, the influence spread becomes similar only under low uncertainty ($q_1 > 0.8$). Our algorithm significantly outperforms the proposed heuristics at all levels of uncertainty, with a performance increment between 38.5% and 45.1%, and has the advantage of providing theoretical guarantees.

[1] The data set is downloadable at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/weic-graphdata.zip.

[2] We conducted a number of experiments with NetHept and two other real-world networks using two rewiring algorithms for the unknown part of the network: one removes and reassigns links randomly, shifting $G_2$ towards a random network, whereas the other reassigns links such that the degree distribution of $G_2$ is the same as that of $G_1$. We also varied the proportion of observable nodes.

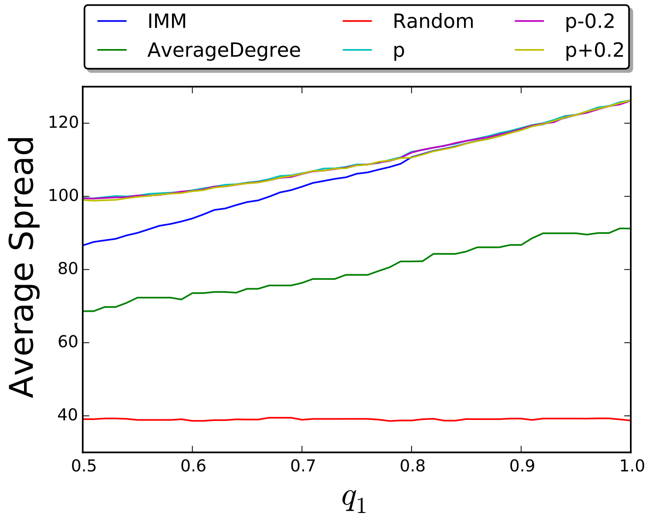

Fig. 1. Average spread given the realization probability $q$ of $G_1$, a choice of $k = 10$ seeds, and a corresponding sampling frequency of $q^* = q + a$ (with imposed boundaries $0 \leq q^* \leq 1$). The error bars are the same size or smaller than the line.

## VI. Conclusions and Future Work

We extended the results from influence maximization with full-information to the case where only a part of the network is visible to the decision-maker. We established that the problem remains NP-hard with a monotone and submodular objective. Moreover, we showed that the efficient algorithms devised to approximate the optimal solution in the full-information case have efficient analogs under our general setting. We also identified the effect of network uncertainty on the required number of samples from possible network realizations, establishing novel results on scaling behaviors for the expectation maximization framework, including showing that total run time of the algorithm does not, in general, scale with the number of possible realizations. We showed analytically that generating random RR sets for realizations according to their probability of occurrence is broadly optimal in terms of performance guarantees for the IMM algorithm, and also performs well in practice on real data-sets. In future work, we seek to compare the results of our approach to other robust models that optimize against the worst case realization of uncertainty (akin to playing a game with nature), while seeking to more efficiently exploit shared information in possible network realizations to further decrease the computational burden. We will also investigate a case where the $M$ graphs represent the unobservable network realization with probability $1 - \delta$ for $\delta \ll 1$.

## Acknowledgments

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. V.R is also partially supported by EPSRC under grant EP/P010164/1, and L.T. was supported in part by the ARO under grant W911NF1810378.

## References

[1] P. Domingos and M. Richardson, "Mining the network value of customers," in *SIGKDD'01*. 2001, pp. 57–66, ACM.

[2] A. Yadav, B. Wilder, E. Rice, R. Petering, J. Craddock, A. Yoshioka-Maxwell, M. Hemler, L. Onasch-Vera, M. Tambe, and D. Woo, "Influence maximization in the field: The arduous journey from emerging to deployed application," in *AAMAS'17*. 2017, pp. 150–158, AAAI.

[3] D. Kempe, J. M. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD'03*. 2003, pp. 137–146, ACM.

[4] J. Yuan and S. Tang, "No time to observe: Adaptive influence maximization with partial feedback," in *IJCAI'17*. 2017, pp. 3908–3914, AAAI.

[5] A. Yadav, H. Chan, A. Xin Jiang, H. Xu, E. Rice, and M. Tambe, "Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty," in *AAMAS'16*. 2016, pp. 740–748, AAAI.

[6] B. Wilder, A. Yadav, N. Immorlica, E. Rice, and M. Tambe, "Uncharted but not uninfluenced: Influence maximization with an uncertain network," in *AAMAS'17*. 2017, pp. 1305–1313, AAAI.

[7] K. Satio, R. Nakano, and M. Kimora, "Prediction of information diffusion probabilities for independent cascade model," in *KES'08*. 2008, pp. 67–75, Springer.

[8] A. Carpentier and M. Valko, "Revealing graph bandits for maximizing local influence," *AISTATS'16*, pp. 10–18, 2016.

[9] B. Wilder, L. Onasch-Vera, J. Hudson, J. Luna, N. Wilson, R. Petering, D. Woo, M. Tambe, and E. Rice, "End-to-end influence maximization in the field," in *AAMAS'18*. 2018, pp. 1414–1422, AAAI.

[10] B. Wilder, N. Immorlica, E. Rice, and M. Tambe, "Maximizing influence in an unknown social network," in *AAAI'18*. 2018, AAAI.

[11] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *SIGMOD'15*. 2015, pp. 1539–1554, ACM.

[12] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[13] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *SODA'14*. 2014, pp. 946–957, SIAM.

[14] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *SIGMOD'14*. 2014, pp. 75–86, ACM.

[15] W. Chen, T. Lin, Z. Tan, M. Zhao, and X. Zhou, "Robust influence maximization," in *SIGKDD'16*. 2016, pp. 795–804, ACM.

[16] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart, "Online influence maximization," in *SIGKDD'15*. 2015, pp. 645–654, ACM.

[17] S. Vaswani, L. Lakshmanan, and M. Schmidt, "Influence maximization with bandits," in *arXiv preprint arXiv:1503.00024*, 2015.

[18] S. Mihara, S. Tsugawa, and H. Ohsaki, "Influence maximization problem for unknown social networks," in *ASONAM'15*. 2015, pp. 1539–1546, ACM.

[19] D. R. Hunter, S. M. Goodreau, and M. S. Handcock, "Goodness of fit of social network models," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 248–258, 2008.

[20] E. Mossel and S. Roch, "On the submodularity of influence in social networks," in *STOC'07*. 2007, pp. 128–134, ACM.

[21] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

[22] S. Stein, S. Eshghi, S. Magshudi, L. Tassiulas, R. Bellamy, and N. Jennings, "Heuristic algorithms for influence maximization in partially observable social networks," in *SOCINF'17*. 2017, pp. 20–32, ACM.