Two-phase Influence Maximization in Social Networks with Seed Nodes and Referral Incentives

Paper 267

Abstract

The problem of maximizing the spread of influence with a limited budget is central to social networks research. Most problems addressed in existing literature spend the entire budget towards triggering diffusion at seed nodes. In this paper, we investigate the effect of splitting the budget across two sequential phases; each phase corresponds to a different way of spreading influence. In phase 1, we adopt the classical approach of triggering diffusion at a selected set of seed nodes to spread the influence. In phase 2, we use the remaining budget to offer referral incentives. Assuming the independent cascade model, we formulate an objective function for the above two-phase influence maximization problem, and investigate its properties. We determine an effective budget-split between the two phases with detailed experiments on synthetic and real-world datasets. The principal findings from our study are: (a) when the budget is low, it is prudent to use the entire budget for phase 1; (b) when the budget is moderate to high, it is preferable to use much of the budget for phase 1, while allocating the remaining budget to phase 2; (c) in the presence of moderate to strict temporal constraints, phase 2 is not warranted; (d) if the temporal constraints are low or absent, referral incentives (phase 2) yield a decisive improvement in influence spread.

Introduction

With the advent of online social networks, companies have increasingly started giving importance to viral marketing through word-of-mouth. The added availability of platforms for mobile applications has now empowered companies to implement referral programs with the hope of boosting the conversion rate of product purchases using referrals.

It has been empirically observed that referred customers are more valuable than regular ones (Schmitt, Skiera, and Van den Bulte, 2011); and that referral incentives are cost effective (Schmitt, Skiera, and Van den Bulte, 2011; Xiao, Tang, and Wirtz, 2011), despite being prone to the opportunistic behavior of customers (Schmitt, Skiera, and Van den Bulte, 2011; Wirtz and Chew, 2002). Furthermore, it has been noted that referrals may, at times, be counterproductive and adversely affect agents' responses because they cause referred friends to infer ulterior motives for the referral (Verlegh et al., 2013). More importantly though, in the same study, the authors demonstrate that rewarding both the referring and the referred agent can eliminate such a negative effect.

In view of the above findings, referral incentives are generally implemented as a two-way scheme: (a) *referral rewards* given to existing customers for successfully referring their friends and (b) *friend offers* which are given to the referred friends who buy the product or sign up for the service. Referral rewards incentivize existing customers to put in additional effort to recommend the product to their friends, while friend offers increase the willingness of the referred friends to become customers. Figure 1 below shows a typical implementation of the referral scheme.

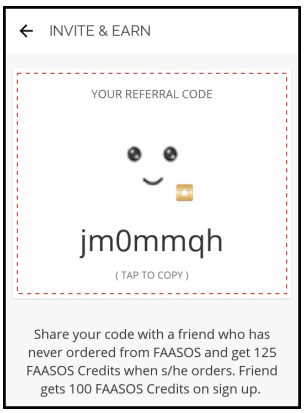


Figure 1: Referral scheme employed by FAASOS, a popular Indian food-on-demand company that takes customer orders via its mobile app

Another often employed marketing scheme is "word-of-mouth" propagation, where the marketing company wishes to find and convince a small subset of users (seeds) to adopt the product, hoping to trigger a cascade of product purchases owing to social influence. The problem of selecting the best seed nodes to maximize diffusion, called *influence maximization*, has been extensively studied in (Easley and Kleinberg, 2010; Guille et al., 2013).

The dynamics and design of word-of-mouth campaigns, as well as referral programs is well understood in isolation, however, it is unclear how to effectively employ both

schemes to achieve greater influence under budget/time constraints. In this work, we attempt to combine the above two methods for spreading influence, and make the following specific contributions:

- With a fixed initial budget in mind, we formulate a discrete optimization problem that captures *budget-splitting* between two sequential phases, where phase 1 involves triggering the spread of influence through seed nodes, and phase 2 employs referral incentives to help spread the influence further. We show that the objective function is non-negative, monotone increasing, and submodular.
- We evaluate this approach over both synthetic and real datasets. Although our proposed algorithm is not guaranteed to achieve constant-factor approximation to the optimal solution, we show that an optimal budget-split would outperform single phase diffusion. Finally, we study the temporal spread of the cascade, and provide an insightful cost-benefit analysis of the two-phase scheme.

Preliminaries

We represent a social network as a graph $G = (V, E)$, where V is the set of n nodes and E is the set of m weighted, directed edges. In this paper, we study information diffusion under the independent cascade model.

The Independent Cascade (IC) Model In this model, an *influence probability* p_{uv} is associated with each directed edge $(u, v) \in E$, where p_{uv} specifies with what probability the source node u would influence the target node v. Diffusion starts at time step 0 with simultaneous triggering of a set of initially activated or influenced seed nodes; following this, diffusion proceeds in discrete time steps. In each step, nodes which got influenced in the previous time step attempt to influence their (inactive) neighbors, and succeed in doing so with the influence probabilities associated with the corresponding edges. Nodes, once activated, remain active for the rest of the diffusion process. The process terminates when no further nodes can be activated.

Live Graph The notion of a *live graph* is crucial to analyzing diffusion under the IC model. A live graph $\mathcal{X}$ of a graph G is an instance of graph G, obtained by sampling the edges; an edge (u, v) is present in the live graph with probability p_{uv} and absent with probability $1 - p_{uv}$, independent of the presence of other edges in the live graph (thus, a live graph is a directed graph with no edge probabilities). The probability $p(\mathcal{X})$ of occurrence of any live graph $\mathcal{X}$, can be obtained as $\prod_{(u,v)\in\mathcal{X}}(p_{uv}) \prod_{(u,v)\notin\mathcal{X}}(1 - p_{uv})$.

Relevant Work

Domingos and Richardson (2001) are the first to study influence maximization as an algorithmic problem in a probabilistic setting. Kempe, Kleinberg, and Tardos (2003) are the first to formulate it as a discrete optimization problem. They show that maximizing the objective function under the IC model is NP-hard, and present an approximate greedy hill-climbing algorithm. Chen, Wang, and Yang (2009) propose fast heuristics for influence maximization under the IC model since the greedy algorithm is computationally intensive. However, all the above approaches select seeds only once and exhaust them all at the same time.

The impact of recommendations and word-of-mouth marketing on product sales revenue is also well-studied in marketing literature; see for example, (Aral and Walker, 2011, 2012; Reichheld, 2003). There have been efforts to study the effectiveness of referral incentive programs and also determining an optimal reward program (Reingen and Kernan, 1986).

Most relevant to our study is the methodology of (Dhamal, Prabuchandran, and Narahari, 2015), who adopt a multi-phase approach for influence maximization, wherein the total budget (number of seed nodes) is split across multiple phases separated by a certain delay. A similar approach is adopted in (Mochalova, 2015), where the initial seed set is split across multiple stages and the degree centrality score is used to select the most influential seeds at every stage. Nevertheless, we depart from this earlier work in two key ways. First, unlike both aforementioned studies, we explicitly consider a referral scheme with individual incentives for successful referrals. Second, in both these works, the seed-selection algorithm is run repeatedly at *each* stage, thus implicitly never exceeding the total budget. In contrast, we decide on the seed-set as well as referral amount only once at the start of phase 1; thereafter there is no selection involved. This introduces a unique constrained optimization problem to ensure that we do not exceed the budget subsequently.

To the best of our knowledge, ours is the first work that analyses the combined effect of traditional 'word-of-mouth' marketing coupled with referral incentives.

Motivation and Agenda

Edge influence probabilities are crucial to diffusion in IC model and in some sense determine the conversion rate of edges during diffusion in social networks. Low edge probabilities cause the diffusion to fade off within a few hops, leading to a concentration of influenced nodes in the vicinity of seed nodes, and very few distant nodes getting influenced. A referral scheme increases the edge influence probabilities and boosts up the chances of the diffusion reaching distant nodes, because it relies on incentivizing every referring agent, instead of only seed nodes. This comes at a cost; owing to the budget split, the number of initial seed nodes that trigger the diffusion is reduced. Moreover, a delay in the diffusion process may be undesirable when the value of the product or information decreases with time, or when there is a competing diffusion and people are influenced by the product or information which reaches them first.

Thus there is a natural trade-off between (a) increased edge probabilities that improve the conversion rate and (b) reduction in the number of seed nodes plus introduction of an additional delay.

A Motivating Example

We now illustrate the usage of referral incentive scheme with a simple stylized example.

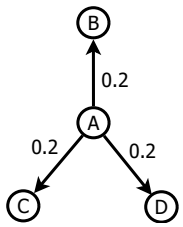


Figure 2: Two phase diffusion using seed set and referral incentives: a motivating example

Let us study single phase diffusion on the graph presented in Figure 2, where node A can influence each of its neighbors with probability 0.2. Consider a total budget of $K = 2$ seeds. Let A and B be the two seed nodes selected by an influence maximizing algorithm at time step 0. The expected number of activated nodes (including seed nodes) at the end of this single phase diffusion process is 2 + 0.2 + 0.2 = 2.4.

Next consider a proposed budget split of 1 seed node for initial diffusion, and the remaining budget (equivalent to 1 seed) to be used as referral incentives. As already mentioned, both agents involved in a successful referral receive this benefit. Suppose under a referral offer of 15% cashback/discount, the edge probabilities increase from 0.2 to 0.6. Now let us assume that the algorithm selects A as the seed node for triggering diffusion in the first phase. The referral scheme operates under the enhanced edge probabilities. A careful consideration of edge probabilities shows that conditioned on the event that A is unable to influence B in the first phase (the probability of this event is 0.8), it should be able to influence B with a probability $\frac{0.6-0.2}{0.8} = 0.5$ under the referral scheme which guarantees both of them a 15% incentive.

The outcomes can be classified into four possible cases. (a) In the best possible scenario, A is able to influence B, C, and D in the first phase itself, thereby negating the need to employ the second phase. Note that in this case, we achieve the target spread while exhausting only a part of the budget. (b) A is able to influence exactly two of its neighbors in the first phase, and fails to influence the third (say, D). In the subsequent referral phase, it is able to influence D with probability 0.5, taking the total expected spread to $1 + 2 + 1(0.5) = 3.5$. For a successful referral, both A and D are awarded a 15% bonus each. We thus exhaust 1 + 2(0.15) = 1.3 out of a total budget of 2, while achieving a higher spread than single phase diffusion. (c) A similar analysis follows for the case when A activates exactly one neighbor in the first phase. (d) Even in the worst case when A is not able to activate any of its neighbors in the first phase, the expected spread after the referral phase is $1 + 3(0.5) = 2.5$, which is higher than the expected spread for single phase diffusion. Nodes B, C and D get a referral award each, and A is awarded 3 bonuses for as many successful referrals. The budget exhausted is 1 + 3(0.15) + 3(0.15) = 1.9, and the spread achieved is higher than in the single phase diffusion.

This stylized example highlights that an increase in edge influence probabilities with the aid of referral incentives can lead to improved diffusion *with the same total budget*.

Problem Formulation

Without loss of generality, we assume that each product has unit price, and K is the total available budget. Under the unit-price assumption, a budget of K corresponds to K free samples that can be given to initial adopters; we use the terms 'budget' and 'seed nodes' interchangeably throughout the paper. Our model is henceforth referred to as *2P-SRI* (**2 P**hase diffusion with **S**eed nodes and **R**eferral **I**ncentives)

Proposed Model for Referral Incentives

Let $h(\alpha)$ be the fractional increase in edge probability owing to the referral scheme, that is, under a referral incentive of α, the influence probability of an edge (u, v) increases from its original value p_{uv} to $p^{\alpha}_{uv} = \min\{1, (1 + h(\alpha))p_{uv}\}$. Thus, $h(\alpha)$ captures the effect of additional effort on part of the referring agent, as well as increased willingness to buy on part of the currently inactive neighbor. For the remainder of this paper, we maintain the natural assumption that $h(\cdot)$ is non-negative, non-decreasing, continuous in [0,1] and satisfies $h(0) = 0$.

At time step $t = 0$, a subset S^k of k initial adopters is selected, and phase 1 is triggered by distributing k free samples among them. As per the IC model, the first phase terminates when no further nodes can be reached by the diffusion process. Let A_{diff} denote the set of active nodes at the end of phase 1. Phase 2 is now initiated by offering a referral incentive of α to this set of customers, hoping to further influence $\bar{A}_{diff} = V \setminus A_{diff}$, the set of currently inactive nodes.

For each $v \in \bar{A}_{diff}$, let $N(v) = \{u | (u, v) \in E; u \in A_{diff}\}$. Each $u \in N(v)$ gets a single opportunity to influence v by sending a direct product recommendation. If v has multiple active neighbors, their activation attempts are sequenced in some arbitrary order. Note that node u must have already made a (failed) attempt at activating node v in phase 1 itself; and given this information, u succeeds in activating v in phase 2 by a certain probability which we arrive at in Section on 'Objective Function'. If v is influenced, the company rewards an amount α to u as well as to v. Once activated, node v can recommend the product to each of its inactive neighbors, say w. Since the edge (v, w) has not been visited before, the probability of v influencing w under α-incentive is just p^{α}_{vw}, and a successful referral gets both v and w the same α fractional cashback (or discount). If A_{ref} is the set of nodes that become active due to the referral program thus defined, then amount spent by the company on referrals is $2\alpha * |A_{ref}|$.

Note that after phase 1, the budget left for referral incentives is $K - k$, so the number of nodes that can be activated in the subsequent referral phase 2 should be no more than $\frac{K-k}{2\alpha}$. Since it is not feasible to ensure a bounded activation in every instance of diffusion (live graph), we make a practical assumption that a company would want to bound this number in expectation.

In addition to the basic properties outlined above, we would like the function $h(\cdot)$ to obey the law of diminishing returns. In our context, this means that as the amount of referral incentive increases, additional incentive has lower

perceived value. We thus model $h(\alpha)$ as a *concave* function. A common choice for concave utilities is the logarithmic function, and in (Kahneman and Tversky, 1979; Bernoulli, 1954) it is shown that logarithmic concave utilities account for risk-aversion, another well-observed attribute of rational agents. Hence for the rest of our work, we consider a simple log function $h(\alpha) = \ln(1+\alpha)$, where the constant of 1 corresponds to the baseline when no referral incentive is received. So $h(\alpha)$ is 0 when $\alpha = 0$, 0.22 when $\alpha = 0.25$ or 25%, 0.4 when $\alpha = 0.5$, and 0.7 when $\alpha = 1$.

Objective Function

Let k be the seed budget reserved for the first phase, and S^k be the corresponding seed set of size k. Assume that $\mathcal{X}$ is the live graph destined to occur at the end of phase 1. While $\mathcal{X}$ is not visible during the diffusion process, we can calculate $p(\mathcal{X})$ from edge probabilities in G as follows:

$$p(\mathcal{X}) = \prod_{(u,v)\in\mathcal{X}} p_{uv} \prod_{(u,v)\notin\mathcal{X}} (1-p_{uv})$$

Similarly, let $\mathcal{Y}$ be the live graph destined to occur at the end of phase 2 with α referral incentive. Note that $\mathcal{X} \subseteq \mathcal{Y}$, that is, the live graph $\mathcal{Y}$ contains all the edges of $\mathcal{X}$, along with edges absent in $\mathcal{X}$ sampled based on the scaled edge probabilities. Hence $p((u,v) \in \mathcal{Y} | (u,v) \in \mathcal{X}) = 1$. If an edge (u,v) is absent in $\mathcal{X}$, it will be present in $\mathcal{Y}$ with probability $(\frac{p^\alpha_{uv} - p_{uv}}{1-p_{uv}})$ and absent with probability $(\frac{1-p^\alpha_{uv}}{1-p_{uv}})$. We now provide an explanation for this scaled value.

Sampling edges under the traditional IC model can be understood as follows: for each edge, we independently sample a number z uniformly at random in $[0,1]$. An edge (u,v) becomes active if and only if $z \leq p_{uv}$. Similarly, the sampling of an edge in $\mathcal{Y}$ given its absence in $\mathcal{X}$ can be thus explained; the probability that it will be present is $p(z \leq p^\alpha_{uv} | z > p_{uv}) = \frac{p^\alpha_{uv} - p_{uv}}{1-p_{uv}}$; and the probability that it will be absent is $p(z > p^\alpha_{uv} | z > p_{uv}) = \frac{1-p^\alpha_{uv}}{1-p_{uv}}$.

So given the occurrence of $\mathcal{X}$, we have that $\mathcal{Y}$ occurs with the following probability:

$$p(\mathcal{Y}|\mathcal{X};\alpha) = \prod_{(u,v)\in\mathcal{Y}\setminus\mathcal{X}} \left(\frac{p^\alpha_{uv} - p_{uv}}{1-p_{uv}}\right) \prod_{(u,v)\notin\mathcal{Y}} \left(\frac{1-p^\alpha_{uv}}{1-p_{uv}}\right)$$

where $p^\alpha_{uv} = \min\{1, (1+h(\alpha))p_{uv}\}$, as defined earlier.

From $\mathcal{X}$, the set of nodes activated in the first phase can be determined. Let $A^{\mathcal{X}}_{diff}$ be the set of nodes active at the end of the influence process that starts at S^k when the resulting live graph is $\mathcal{X}$, that is,

$$A^{\mathcal{X}}_{diff} = \{v |\ v \text{ is reachable from } S^k \text{ in } \mathcal{X}\}$$

The nodes activated thus act as effective seed nodes for the next phase. As above, we define $A^{\mathcal{Y}}_{ref}$ to be the set of additional nodes influenced in the referral phase, that is,

$$A^{\mathcal{Y}}_{ref} = \{v |\ v \text{ is reachable from } A^{\mathcal{X}}_{diff} \text{ in } \mathcal{Y}\} \setminus A^{\mathcal{X}}_{diff}$$

Now as both $\mathcal{X}$ and $\mathcal{Y}$ are unknown at the beginning of the first phase, the influence function $f(S^k, \alpha)$ is in expectation over all such $\mathcal{X}$'s and $\mathcal{Y}$'s. Thus,

$$f(S^k,\alpha) = \sum_{\mathcal{X}} p(\mathcal{X}) \left\{ |A^{\mathcal{X}}_{diff}| + \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X};\alpha) |A^{\mathcal{Y}}_{ref}| \right\}$$

and we have the following constrained optimization problem,

Select (S^k, α) to maximize

$$f(S^k,\alpha) = \sum_{\mathcal{X}} p(\mathcal{X}) \left\{ |A^{\mathcal{X}}_{diff}| + \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X};\alpha) |A^{\mathcal{Y}}_{ref}| \right\}$$

subject to $\sum_{\mathcal{X}} p(\mathcal{X}) \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X};\alpha) |A^{\mathcal{Y}}_{ref}| \leq \frac{K-k}{2\alpha}$

where the constraint bounds the expected number of nodes activated in the referral phase 2, as explained in Section on 'Proposed Model for Referral Incentives'.

It is easy to prove that the problem of maximizing $f(S^k)$ is NP-hard. In particular when $k = K$ and $\alpha = 0$, the problem reduces to the objective function for single-phase diffusion $\sigma(\cdot)$, and the problem of maximizing $\sigma(\cdot)$ is NP-hard (Kempe, Kleinberg, and Tardos, 2003).

We present an illustrative example to demonstrate the computation of the objective function.

Example 1. *Consider the network given in Figure 3. We study two-phase referral scheme with $K = 2$, $k = 1$, and $S^1 = \{A\}$. The initial edge probability is p on both edges, and under some fixed referral bonus α, the probability scales to q. Table 1 lists all possible realizations of $\mathcal{X}$ and $\mathcal{Y}$ along with their respective probabilities of occurrence. As a concrete example, let $p = 0.5$, $\alpha = 0.5$, $q = (1+\ln(1+\alpha))p \approx 0.7$. Hence it can be computed that $f(\{A\}, 0.5) \approx 2.19$.*

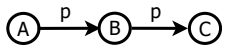


Figure 3: Example

$K=2, k=1, S^1=\{A\}$					
$\mathcal{X}$	$p(\mathcal{X})$	$\lvert A^{\mathcal{X}}_{diff}\rvert$	$\mathcal{Y}$	$p(\mathcal{Y}\lvert\mathcal{X};\alpha)$	$\lvert A^{\mathcal{Y}}_{ref}\rvert$
$\{AB, BC\}$	p^2	3	$\{AB, BC\}$	1	0
$\{AB\}$	$p(1-p)$	2	$\{AB\}$	$\frac{1-q}{1-p}$	0
			$\{AB, BC\}$	$\frac{q-p}{1-p}$	1
$\{BC\}$	$p(1-p)$	1	$\{BC\}$	$\frac{1-q}{1-p}$	0
			$\{AB, BC\}$	$\frac{q-p}{1-p}$	2
$\{\}$	$(1-p)^2$	1	$\{AB\}$	$\frac{(q-p)(1-q)}{(1-p)^2}$	1
			$\{BC\}$	$\frac{(q-p)(1-q)}{(1-p)^2}$	0
			$\{AB, BC\}$	$\frac{(q-p)^2}{(1-p)^2}$	2
			$\{\}$	$\frac{(1-q)^2}{(1-p)^2}$	0

Table 1: Table shows computation of values in Example 1

Properties of the Objective Function

Lemma 1. *The objective function $f(S, \alpha)$ is equivalent to $\sum_{\mathcal{Y}} P(\mathcal{Y}; \alpha)|A^{\mathcal{Y}}_{diff}|$.*

Proof. Consider an edge e with original activation probability p_e and enhanced probability p_e^α. Let $\mathcal{X}$ and $\mathcal{Y}$ be live graphs observed at the end of phase 1 and 2 respectively; $\mathcal{X} \subseteq \mathcal{Y}$. Then,
$P(e \in \mathcal{Y}) = P(e \in \mathcal{X}) + P(e \notin \mathcal{X}).P(e \in \mathcal{Y}|e \notin \mathcal{X}) = p_e + (1 - p_e).\left(\frac{p_e^\alpha - p_e}{1 - p_e}\right) = p_e^\alpha$.
Note that the set of nodes reachable from S^k in $\mathcal{Y}$ is precisely the set of influenced nodes at the end of both phases. Thus, the (unconstrained) two-phase objective is equivalent to the single-phase objective that operates on a graph with enhanced probabilities. □

Owing to the above equivalence and the single phase objective function being non-negative, monotone, and submodular, the following result follows.

Proposition 1. *For a fixed α, $f(S, \alpha)$ is non-negative, monotone, and submodular with respect to S.*

On Appropriateness of the Greedy Hill-climbing Algorithm We have shown that the objective function underlying the considered problem is non-negative, monotone, and submodular. However, owing to the additional constraint on the number of nodes that can be activated in the referral phase 2, the greedy hill-climbing algorithm is not guaranteed to give a $(1 - \frac{1}{e})$-approximate solution. If the greedily chosen optimal seed set does not belong in the constraint set, the algorithm is forced to pick sub-optimal nodes to satisfy the constraint, and the constant factor approximation guarantee (Nemhauser, Wolsey, and Fisher, 1978) may no longer be valid. In this paper, we use greedy algorithm as a first heuristic targeting to maximize the objective function $f(\cdot, \cdot)$, since it can take this farsighted function into consideration unlike other heuristics such as PMIA (Chen, Wang, and Wang, 2010).

Experimental Evaluation

Simulation Setup

Whenever there is a need for transforming an undirected, unweighted network into a directed and weighted network, we employ two well-accepted special cases of the IC model, namely, the *weighted cascade (WC) model* and the *trivalency (TV) model*. The WC model does this transformation by making all edges bidirectional and assigning a weight to every directed edge (u, v) equal to the reciprocal of v's degree in the undirected network (Kempe, Kleinberg, and Tardos, 2003). The TV model makes all edges bidirectional and assigns a weight to every edge by uniformly sampling from the set of values $\{0.001, 0.01, 0.1\}$. For computing the objective function value accurately, we run 10^4 Monte-Carlo iterations.

Key Implementation Details In Section on 'Properties of the Objective Function', we show that for a fixed α, the influence function is monotone and submodular with respect to S. This leads to the following natural approach for finding the best split : We perform grid search over a discrete range for potential values of α; and for each fixed choice of α, we use a suitable algorithm to find the influence maximizing seed set S^k, while respecting the budget constraint.

We call a (k, α) pair *infeasible* if selecting exactly k seed nodes with the corresponding α value violates the upper bound on the number of permissible referred nodes. When such a case arises, we reject this value of k, find the largest $k' \leq k$ for which (k', α) is a feasible pair, and replace $f(k, \alpha)$ with $f(k', \alpha)$.

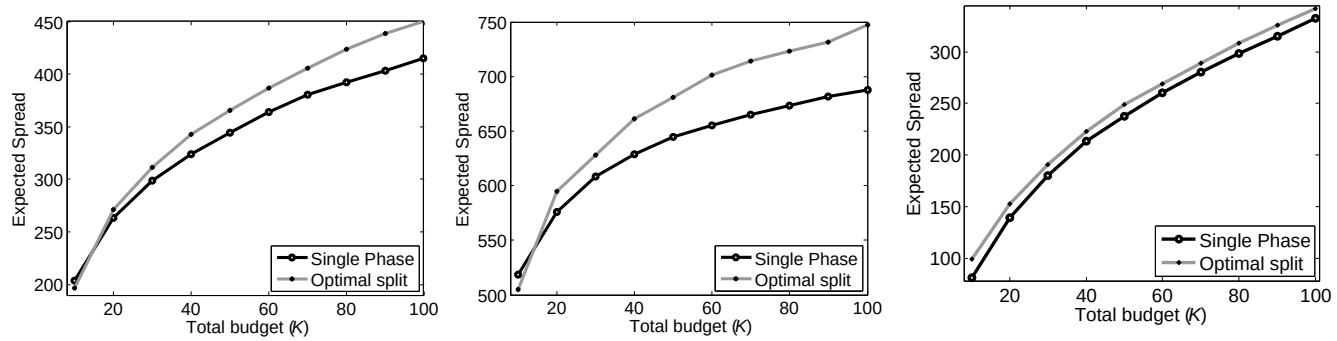
The conventional greedy algorithm starts with an empty set and then successively adds a node t with the maximum marginal influence until k nodes are reached. In the 2P-SRI model, it is possible that including t into the seed set may cause too many additional nodes to become active in the referral phase, thereby violating the referral budget constraint. This behavior is typical at higher values of α, where the referral budget is low but the probability of nodes getting activated in the referral phase is high. In such cases, we modify the greedy algorithm to forego the 'best' seed, and pick instead a node that yields the highest spread while respecting the budget constraint.

Evaluation on Synthetic Data

In our synthetic experiments, we simulate real-word network data by studying three commonly observed degree distributions in complex networks, namely, power-law (Ahn et al., 2007), stret- ched exponential (Newman, Forrest, and Balthrop, 2002), and log-normal (Lerman and Ghosh, 2010).

Our synthetic data experiments follow the setup adopted in (Liu et al., 2016). We first distribute 1000 nodes according to a homogenous Poisson point process over a 2-D space of unit area. Next, we randomly assign the out-degree of each node according to the power law, stretched exponential, or log-normal distribution with appropriate parameters. For each node with out-degree l, we add l directed edges into the network with this node as the starting point. The ending point of an added edge is chosen according to the WPR model (Wong, Pattison, and Robins, 2006), in which connection probability between two nodes is a step-function of the distance between them. Following this, we randomly assign weights to edges using the TV or WC model, as mentioned previously. As the results obtained on both models were very similar, we present results only for the TV model. Since there is a probabilistic element in the generation of synthetic graphs, our reported results are averaged over 20 instantiations of each graph model.

Key Observations and Insights Figure 4 depicts the performance of 2P-SRI compared to the baseline of single-stage (batch) seed selection for graphs with power law, stretched exponential, and log-normal degree distributions. The total available budget, K varies from 1% of total network nodes (10 seeds) to 10% (100 seeds). 2P-SRI clearly outperforms the baseline for all three synthetic graphs; Figure 4(a) and

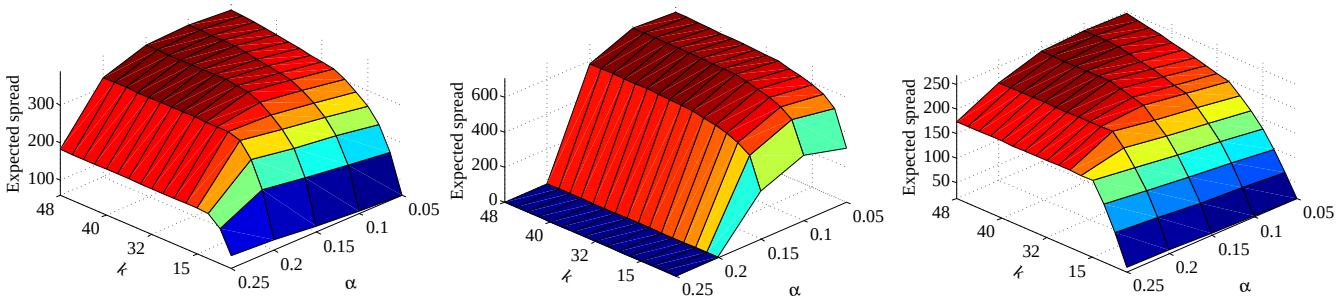


(a) Performance on log-normal network

(b) Performance on power-law network

(c) Performance on stretched exponential network

Figure 4: Performance comparison between 2P-SRI at optimal budget-split, and single-phase (batch) selection that does not involve budget-splitting or referral incentives, on networks with different degree distributions (1000 nodes)



(a) Log-normal network, optimal at (41, 0.10)

(b) Power-law network, optimal at (38, 0.12)

(c) Stretched exponential network, optimal at (45, 0.075)

Figure 5: Performance of 2P-SRI as a function of (k, α) for fixed $K = 50$

(b) show that the performance-gain is more pronounced with increase in K. Further, note that for a very small value of initial budget, $K = 10$, 2P-SRI may be detrimental. Similar trends are observed in our experiments on real-world data; please refer to Section on 'Effect of Total Budget' for detailed explanations.

Figure 5 shows the influence spread as a function of k and α for a fixed value of $K = 50$. Observe that for a fixed α, the influence spread is monotone increasing with k. Also note that the maximum spread is obtained at relatively high values of k (in the range 38 to 45) and low values of α (typically 7.5% to 10%). This corroborates with our observations on real-world data, please see Section on 'Effect of k and α' for a detailed analysis of this observation.

Evaluation on Real Data

We first conduct simulations on the Les Miserables (LM) dataset (Knuth, 1993) consisting of 77 nodes and 508 directed edges in order to study the performance of computationally intensive greedy algorithms. To study two-phase diffusion on a larger dataset, we consider an academic collaboration network obtained from co-authorships in the "High Energy Physics - Theory" papers published on the e-print arXiv from 1991 to 2003. It contains 15,233 nodes and 62,774 directed edges (with WC or TV transformation), and is popularly denoted as NetHEPT. This network exhibits many structural features of large-scale social networks and is widely used for experimental justifications, for instance, in (Kempe, Kleinberg, and Tardos, 2003; Chen, Wang, and Yang, 2009; Chen, Wang, and Wang, 2010). In subsequent sections, the following aspects are empirically examined:

- Selecting seed nodes based on two methods, namely farsighted and myopic algorithms (Section on 'Farsighted versus Myopic Seed Selection')
- Variation in performance of 2P-SRI over a wide range of initial seed budget, and role of k and α in the extent of influence spread (Sections on 'Effect of k and α', 'Effect of Total Budget')
- Rate of diffusion, and the dynamics of influence spread under temporal constraints, wherein we also seek to find an optimal delay after which the referral phase should be initiated. (Section on 'Scheduling the Referral Phase')

Simulation Results

Farsighted versus Myopic Seed Selection The farsighted method corresponds to determining the optimal (k, α) pair, as well as the seed set that maximizes $f(S^k, \alpha)$, the two-phase influence function. In other words, the farsighted method determines the above parameters while fully taking into account the second phase (referral phase) for prediction purposes. In contrast, the myopic method does not account for the presence of the referral phase; that is, the selected seed set S^k aims to maximize only the single phase objective function $\sigma(S^k)$.

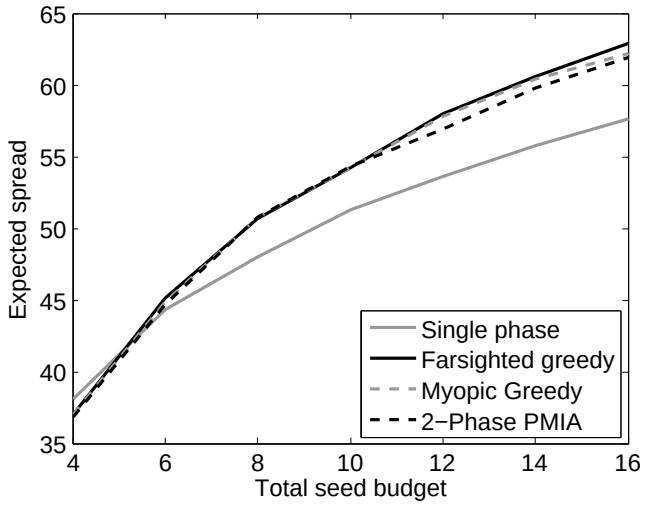


Figure 6: Performance of algorithms as a function of total seed budget on LM dataset (WC model)

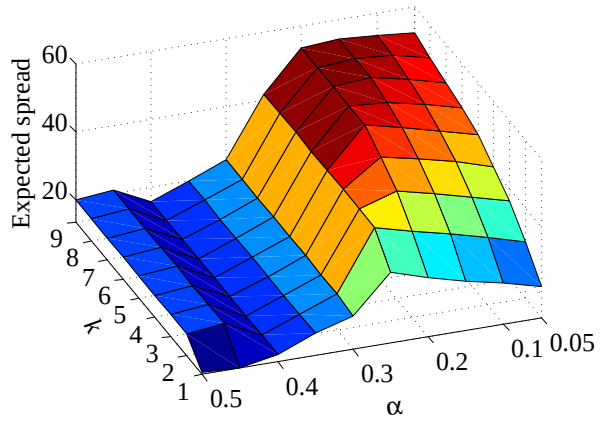


Figure 7: Performance of 2P-SRI as a function of k and α on LM dataset (WC model) for $K = 10$

The farsighted greedy algorithm involves two levels of Monte-Carlo iterations (thus squaring the effective number of iterations) and is not suitable to be run on a large dataset. Figure 6 draws a comparison between farsighted greedy hill-climbing and myopic methods (greedy and heuristic) on the LM dataset. The difference in performance of the farsighted and myopic versions of the greedy algorithm is practically negligible. So it is reasonable to implement only the myopic algorithm in the interest of computationally feasible running time. Furthermore, the PMIA heuristic which performs close to the greedy algorithm (in terms of expected influence of the seed set) for single phase diffusion (Chen, Wang, and Wang, 2010), performs close to optimal in our optimization problem as well.

Effect of k and α Figures 7, 8, and 9 present the expected spread as a function of the initial seed budget k (in phase 1) and referral incentive α (in phase 2), for a fixed total budget K. It is observed that the highest expected spreads are

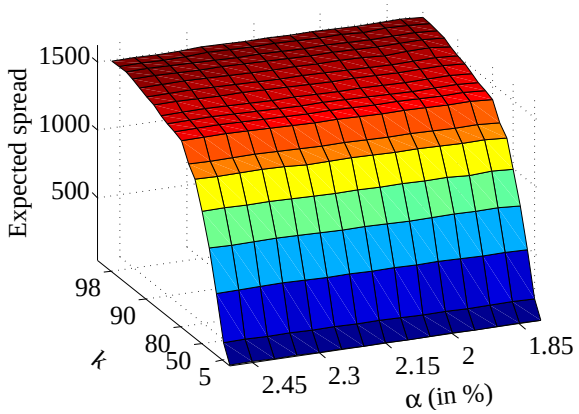


Figure 8: Performance of 2P-SRI as a function of k and α on NetHEPT (WC model) for $K = 100$

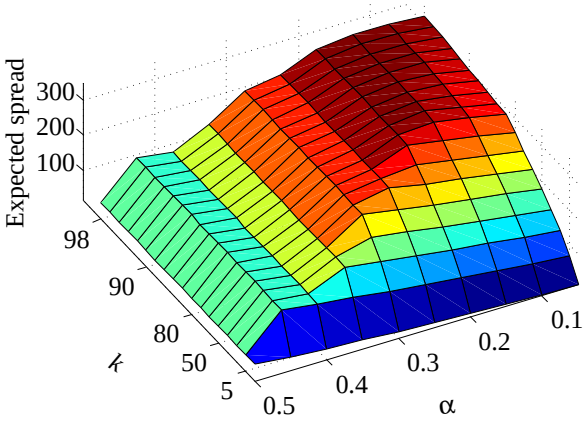


Figure 9: Performance of 2P-SRI as a function of k and α on NetHEPT (TV model) for $K = 100$

obtained at relatively high values of k and low values of α. In fact, for a reasonably high total budget, we consistently observe that for the TV model, it is optimal to set k equal to 85-90% of the total budget K with moderately high α (10-15%), whereas for the WC model the optimal k is typically about 95% of K, with the value of α on the lower side (1.5-2.5%). Tables 2 and 3 present detailed observations.

A clear tradeoff emerges between (a) the extent of diffusion owing to the initial set of seed nodes and (b) the amount of referral incentives to be offered. For a referral incentive scheme to work effectively, it is crucial that there is a significant population of activated nodes that act as referring agents in phase 2. A small sized initial seed set limits the number of active nodes at the end of phase 1, leading to a dearth of referring agents for the next phase. So a low value of k combined with any value of α will likely result in a rather limited spread; this explains the high values of k in all optimal (k, α) pairs.

Furthermore, an improved spread is never attained at very high values of α (beyond 20%) because it is evident that for a fixed K and k, higher values of α lower the maximum permissible number of nodes which can be influenced in phase

K	Expected spread		k	α	% gain
	Single phase	With referral			
10	266.58	275.43	9	0.021	**3.32**
15	365.33	375.51	13	0.017	**2.79**
20	447.26	459.22	18	0.027	**2.76**
30	628.14	643.49	28	0.016	**2.44**
50	944.55	971.53	47	0.021	**2.86**
80	1332.05	1372.413	77	0.017	**3.03**
100	1554.96	1605.844	96	0.022	**3.28**
150	2002.26	2061.85	144	0.025	**2.98**
200	2389.21	2458.55	194	0.021	**2.91**

Table 2: Detailed results of simulations on NetHEPT (WC model)

K	Expected spread		k	α	% gain
	Single phase	With referral			
10	60.93	63.94	9	0.05	**4.93**
15	82.57	87.72	13	0.05	**6.24**
20	103.32	109.26	15	0.10	**5.75**
30	128.0	134.8	25	0.10	**5.31**
50	192.24	204.21	46	0.15	**6.23**
80	263.44	284.89	72	0.15	**8.14**
100	307.29	327.88	82	0.15	**6.69**
150	404.68	433.07	140	0.15	**7.02**
200	496.45	527.06	188	0.15	**6.17**

Table 3: Detailed results of simulations on NetHEPT (TV model)

2. In our model, the edge probability enhancement $h(\alpha)p_{uv}$ for an edge (u, v) depends on the initial probability p_{uv} itself. The larger this probability, lower is the α required to convert an edge (u, v) from a non-live edge to a live edge. For the NetHEPT dataset, edge probabilities are in a much higher range in WC model compared to TV model; so it suffices to have a very low α for the WC model, and a relatively higher α for the TV model.

Effect of Total Budget Figures 6 and 10 depict the performance of 2P-SRI in comparison with single-phase diffusion for LM (WC model) and NetHEPT dataset (TV model), respectively. It is observed that when the total budget K is low, splitting it further is detrimental, while a split is advantageous when the budget is moderate or high. The results are along similar lines for both models. A detailed set of observations capturing the optimal values of k and α as well as the %-superiority of 2P-SRI over the single phase approach, for a wide range of total budget on NetHEPT, is provided in Tables 2 and 3. The gain in performance varies between 3-8% for a reasonably high total budget. Note that this gain could have significant implications when the company is concerned with monetary profits or a long-term customer base.

As mentioned earlier, if the initial seed set is not of a reasonable size, the number of nodes activated in the regular diffusion phase is likely to be very limited. This adversely affects the final spread despite the referral incentive scheme, hence if the total budget is low to begin with, splitting it is not warranted.

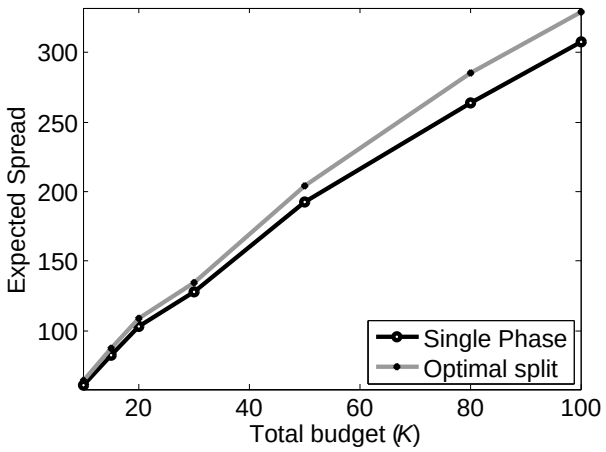


Figure 10: Performance of 2P-SRI as a function of total seed budget on NetHEPT dataset (TV model)

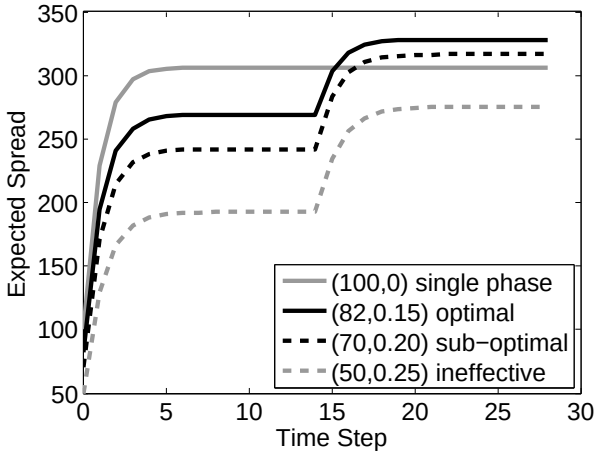


Figure 11: Discrete time performance of 2P-SRI for different (k, α) pairs on NetHEPT (TV model) for $K = 100$

Scheduling the Referral Phase In the model considered, it is presumed that one should wait long enough for regular diffusion in phase 1 to terminate before initiating referral phase 2. However, this wait may not be advisable in the presence of temporal constraints where the rate of diffusion is critical; or in presence of a competing campaign where the aim is to reach as many nodes as early as possible.

Figure 11 depicts a typical temporal progression of the influence spread for different values of k and α. Owing to the use of the entire budget in the beginning itself, single phase diffusion is the fastest and reaches its saturation point within a few time steps. The other plots present the progression of the optimal (k, α) pair, a sub-optimal one, as well as a pair whose final influence spread is less than that of the single

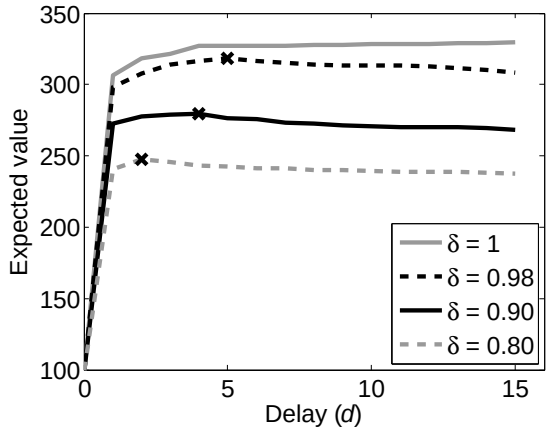


Figure 12: Performance of 2P-SRI as a function of the delay after which the referral phase is initiated on NetHEPT (TV model) for $K = 100$

phase. In the 2P-SRI model, influence stagnates at the end of phase 1, then shoots up and reaches its optimum a few time steps after initiating phase 2.

The standard IC model does not capture aforementioned temporal factors; a more practical objective function should capture not only the influence spread, but also its rate. One such function could be $\nu(S) = \sum_{t=0}^{\infty} \Gamma(t)\sigma_{(t)}(S)$, where $\Gamma(\cdot)$ is a non-increasing function such that $\Gamma(t) \leq 1$ for all values of t, and $\sigma_{(t)}(S)$ is the expected number of newly activated nodes at time step t. As an educated first guess for a decay function in several problems, we consider $\Gamma(t) = \delta^t$ where $\delta \in [0, 1]$ in our simulations.

Figure 12 captures the performance in terms of *expected value*, which refers to the value of $\nu(\cdot)$ with $\Gamma(t) = \delta^t$. In particular, it exhibits the expected value obtained in 2P-SRI as a function of the delay after which the referral phase is initiated, for different values of δ with the corresponding optimal (k, α) pairs. The corresponding expected values for single phase diffusion are 307 ($\delta = 1$), 301 ($\delta = 0.98$), 275 ($\delta = 0.9$), and 254 ($\delta = 0.8$). The optimal delays are marked on the plots; and for the NetHEPT dataset, we observe that 2P-SRI gives an improvement over single phase diffusion when δ is relatively high (≥ 0.9), while it should not be used when δ is in a lower range. Furthermore, it can be seen that the optimal delay decreases as the value of δ lowers, i.e. phase 2 is initiated earlier.

As mentioned previously, for 2P-SRI to be effective, it is critical that the number of active nodes at the beginning of phase 2 be large enough (which requires some diffusion time); and that referrals are not expended on nodes that anyway could been activated without referral incentives, i.e. in phase 1 itself. A lower value of delay d adversely affects the above conditions. On the other hand, a higher value of d leads to considerable decaying of the product value and hence the expected value. This induces a natural trade-off in determining the optimal value of d, the number of time steps for which phase 1 is executed.

Summary and Actionable Insights

Based on our experiments on synthetic and real-data, we have the following insights applicable to any typical marketing campaign:

- An optimal budget split with referrals is advantageous over single-phase seed selection, provided the total budget is moderate to high. For low initial budget, it is advisable to use the entire budget as seeds and forego the referral phase.
- An optimal split is typically observed when 85-90% of total budget is used for seed selection in phase 1. If the edge probabilities are high, a small referral incentive of 1.5-2.5% during phase 2 is sufficient, for low edge probabilities the incentive must be higher, around 10-15%
- Referrals are best avoided in the presence of moderate to high temporal constraints, but work well in practice when there is no time constraint.

References

Ahn, Y.-Y.; Han, S.; Kwak, H.; Moon, S.; and Jeong, H. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, 835–844. ACM.

Aral, S., and Walker, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* 57(9):1623–1639.

Aral, S., and Walker, D. 2012. Identifying influential and susceptible members of social networks. *Science* 337(6092):337–341.

Bernoulli, D. 1954. Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society* 23–36.

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 International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 1029–1038. ACM.

Chen, W.; Wang, Y.; and Yang, S. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 199–208. ACM.

Dhamal, S.; Prabuchandran, K. J.; and Narahari, Y. 2015. A multi-phase approach for improving information diffusion in social networks. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1787–1788. IFAAMAS.

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*, 57–66. ACM.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Guille, A.; Hacid, H.; Favre, C.; and Zighed, D. A. 2013. Information diffusion in online social networks: A survey. *ACM Special Interest Group on Management of Data (SIGMOD) Record* 42(1):17–28.

Kahneman, D., and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society* 263–291.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 137–146. ACM.

Knuth, D. E. 1993. *The Stanford GraphBase: A Platform for Combinatorial Computing*, volume 37. Addison-Wesley Reading.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. 90–97.

Liu, H.; Ioannidis, S.; Bhagat, S.; and Chuah, C.-N. 2016. Adding structure: Social network inference with graph priors. In *ACM Workshop on Mining and Learning with Graphs, co-located at ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Mochalova, A. 2015. Multi-stage application of seed selection process for viral marketing. *Available at SSRN 2630949*.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming* 14(1):265–294.

Newman, M. E.; Forrest, S.; and Balthrop, J. 2002. Email networks and the spread of computer viruses. *Physical Review E* 66(3):035101.

Reichheld, F. F. 2003. The one number you need to grow. *Harvard Business Review* 81(12):46–55.

Reingen, P. H., and Kernan, J. B. 1986. Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research* 370–378.

Schmitt, P.; Skiera, B.; and Van den Bulte, C. 2011. Referral programs and customer value. *Journal of Marketing* 75(1):46–59.

Verlegh, P. W.; Ryu, G.; Tuk, M. A.; and Feick, L. 2013. Receiver responses to rewarded referrals: the motive inferences framework. *Journal of the Academy of Marketing Science* 41(6):669–682.

Wirtz, J., and Chew, P. 2002. The effects of incentives, deal proneness, satisfaction and tie strength on word-of-mouth behaviour. *International Journal of Service Industry Management* 13(2):141–162.

Wong, L. H.; Pattison, P.; and Robins, G. 2006. A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications* 360(1):99–120.

Xiao, P.; Tang, C. S.; and Wirtz, J. 2011. Optimizing referral reward programs under impression management considerations. *European Journal of Operational Research* 215(3):730–739.