Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

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

Sneha Mondal, Swapnil Dhamal, Y. Narahari
{sneha.mondal, swapnil.dhamal, hari}@csa.iisc.ernet.in
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, India

Abstract

The problem of maximizing the spread of influence with a limited budget is central to social networks research. Most solution approaches available in the existing literature devote the entire budget towards triggering diffusion at seed nodes. This paper investigates the effect of splitting the budget across two different, sequential phases. In phase 1, we adopt the classical approach of initiating diffusion at a selected seed-set. In phase 2, we use the remaining budget to offer *referral incentives*. We formulate this problem and explore suitable ways to split the budget 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, phase 2 yields a decisive improvement in influence spread.

Introduction

With the advent of online social networks, companies are increasingly giving importance to viral marketing through word-of-mouth. The added availability of platforms for mobile applications has empowered organizations to implement referral programs which motivate existing customers to promote a product amongst friends. It has been observed that referred customers are more valuable than regular ones; and that referral incentives are cost effective (Schmitt, Skiera, and Van den Bulte 2011), and there have been efforts to determine optimal pricing policies for referral reward programs (Hartline, Mirrokni, and Sundararajan 2008). On the other hand, it has been noted that referrals may adversely affect agents' responses since they cause referred friends to infer ulterior motives for the referral (Verlegh et al. 2013). More importantly though, 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 recommending the product or service to their friends and (b) *friend offers* which are given to the referred friends who buy the product or sign up for the service; thus incentivizing both parties involved in a successful referral.

Considerable work has been done on *influence maximization* using seed nodes, where the marketing company determines a small subset of users (seeds) who could be offered to adopt free samples of the product, hoping to trigger a cascade of subsequent product purchases owing to social influence (Kempe, Kleinberg, and Tardos 2003). There have also been studies on multi-phase approach for influence maximization, wherein the total budget (number of seed nodes) is split across multiple phases separated by a certain delay (Dhamal, Prabuchandran, and Narahari 2016). We depart from these studies in two key ways. First, we explicitly consider a reward scheme with individual incentives for successful referrals in phase 2. Second, in previous works, the seed-selection algorithm is run 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 decision involved. This introduces a unique constrained optimization problem to ensure that we do not exceed the budget subsequently.

The dynamics of word-of-mouth campaigns using seeding and referral programs are well understood in isolation, however, to the best of our knowledge, ours is the first work that analyzes the combined effect of word-of-mouth marketing using seeding coupled with referral incentives.

Problem Formulation

A social network is represented as a directed graph $G = (V, E)$, with pairwise influence probabilities p_{uv}. 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 α, expressed as a fraction of product price, denote the incentive (offered as discount or cashback) rewarded to both the referring and referred agents for a successful referral. Let $h(\alpha)$ be the resulting fractional increase in edge probability

i.e., under a referral incentive of α, the influence probability of edge (u, v) increases from its original value p_{uv} to $p^{\alpha}_{uv} = \min\{1, (1+h(\alpha))p_{uv}\}$. We assume $h(\cdot)$ to be non-negative, non-decreasing, continuous in [0,1] and satisfies $h(0) = 0$.

We employ the Independent Cascade (IC) model (Kempe, Kleinberg, and Tardos 2003) to study the stochastic process underlying diffusion. At time $t = 0$, a subset S^k of k initial adopters is selected, thus triggering phase 1, which terminates when no further nodes can be reached by the diffusion cascade. 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 nodes among $\bar{A}_{\text{diff}} = V \setminus A_{\text{diff}}$, the set of currently inactive nodes. For each $v \in \bar{A}_{\text{diff}}$, let $N(v) = \{u|(u, v) \in E; u \in A_{\text{diff}}\}$. Each $u \in N(v)$ gets (another) single opportunity to influence v. Note that u had already made a failed attempt at activating v in phase 1; and the probability of a successful attempt in phase 2, given failure in phase 1, is $(\frac{p^{\alpha}_{uv} - p_{uv}}{1-p_{uv}})$. If v is influenced, both u and v are rewarded the amount α. Once activated, node v can refer the product to each of its inactive neighbors, say w. This being v's first attempt at activating w, it would succeed with probability p^{α}_{vw}, with both v and w getting α reward for a successful referral.

If A_{ref} is the set of nodes that become active due to the referral program thus defined, the amount spent by the company on referrals is $2\alpha * |A_{\text{ref}}|$. Note that after phase 1, the budget remaining for referral incentives is $K-k$, so no more than $\frac{K-k}{2\alpha}$ nodes can be activated in phase 2. Since it is not feasible to ensure a bounded activation in every instance of the diffusion process, we aim to bound it in expectation.

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 underlying the diffusion process in 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 underlying the diffusion process in phase 2 with α referral incentive. Note that $\mathcal{X} \subseteq \mathcal{Y}$, that is, $\mathcal{Y}$ contains all the edges of $\mathcal{X}$, along with edges absent in $\mathcal{X}$ resampled based on the scaled edge probabilities. Hence $p((u, v) \in \mathcal{Y}|(u, v) \in \mathcal{X}) = 1$. Sampling edges under the IC model can be understood as follows: for each edge, we independently sample 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 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 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 $A^{\mathcal{X}}_{\text{diff}}$ 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,

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

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

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

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,

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

and we get the following constrained optimization problem,

Select (S^k, α) to maximize

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

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

where the constraint bounds the expected number of nodes activated in the referral phase. When $k = K$ and $\alpha = 0$, $f(S^k, \alpha)$ reduces to single-phase influence function $\sigma(S^K)$, maximizing which is NP-hard (Kempe, Kleinberg, and Tardos 2003). Hence the above problem is NP-hard.

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 edge probability p_e and enhanced probability p^{α}_e. Let $\mathcal{X}$ and $\mathcal{Y}$ be live graphs for phase 1 and 2 respectively. Then, $\mathcal{X} \subseteq \mathcal{Y}$ and $P(e \in \mathcal{Y}) = P(e \in \mathcal{X}) + P(e \notin \mathcal{X}).P(e \in \mathcal{Y}|e \notin \mathcal{X}) = p_e + (1 - p_e).\left(\frac{p^{\alpha}_e - p_e}{1-p_e}\right) = p^{\alpha}_e$.

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.*

It is important to note that, despite these properties, owing to the additional constraint on the number of nodes that can be activated in the referral phase, the greedy hill-climbing algorithm is not guaranteed to give a constant factor approximation of $(1 - \frac{1}{e})$ unlike in the unconstrained case (Nemhauser, Wolsey, and Fisher 1978).

Experimental Evaluation

For experimental evaluation, we require a concrete function $h(\cdot)$ satisfying the basic properties mentioned previously. To model most practical scenarios, we desire $h(\cdot)$ to obey the law of diminishing returns. In our context, this means that as the 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, which also accounts for risk-aversion (Kahneman and Tversky 1979; Bernoulli 1954), another well-observed attribute of rational agents. Hence we consider a simple log function $h(\alpha) = \ln(1 + \alpha)$, which satisfies all aforementioned properties.

We conduct simulations on synthetic datasets with commonly observed degree distributions in complex networks (power-law, stretched exponential, and log-normal), as well as on real-world datasets. We employ *weighted cascade (WC)* and *trivalency (TV)* models to transform an undirected, unweighted network into a directed, weighted one. The WC model assigns a weight to every directed edge (u, v) equal to the reciprocal of v's degree in the undirected network, while the TV model assigns a weight to every edge by uniformly sampling from the set of values $\{0.001, 0.01, 0.1\}$. For computing objective function value, we run 10^4 Monte-Carlo iterations.

Similar quantitative and qualitative results are observed on all network data, and we report only representative observations for our experiments on the following datasets - (a) Les Miserables (LM), consisting of 77 nodes and 508 directed edges (Knuth 1993), used for computationally intensive experiments, and (b) a co-authorship network in the "High Energy Physics - Theory" papers (NetHEPT) consisting of 15,233 nodes and 62,774 directed edges, for making deductions on general social networks (Kempe, Kleinberg, and Tardos 2003; Chen, Wang, and Wang 2010).

Key Implementation Details We have shown 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. For each 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 in phase 1 with the corresponding α-reward in phase 2 violates the budget constraint. In such a case, 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 successively adds a node with the maximum marginal influence until k nodes are reached. In the 2P-SRI model, including this node in the seed set may cause too many 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.

Simulation Results

Farsighted versus Myopic Seed Selection The farsighted approach takes both phase 1 and 2 into consideration to determine the optimal (k, α) pair, as well as the seed set that maximizes the two-phase objective $f(S^k, \alpha)$. In contrast, the myopic method does not account for the presence of referral phase; that is, the selected seed set S^k aims to maximize only the single phase objective function $\sigma(S^k)$. 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. On LM dataset, the farsighted and myopic versions of the greedy algorithm perform nearly equally well, hence we implement only the myopic algorithm for a computationally feasible running time. Furthermore, in terms of expected influence using seeding, the PMIA heuristic (Chen, Wang, and Wang 2010) performs close to the greedy algorithm for our optimization problem.

Effect of k and α Figure 1(a) presents the expected spread as a function of varying k and α, for a fixed total budget K. A clear trade-off emerges between (a) the extent of diffusion owing to the initial seed-set and (b) the percentage of referral incentive offered. For the 2P-SRI model to be effective, it is crucial that there be 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 and hence a rather limited spread; this explains the high values of optimal k. Also, an improved spread is never attained at very high values of α (beyond 20%) since for a fixed K and k, a higher α lowers the maximum permissible number of nodes which can be influenced in phase 2.

Typically, we observe that for TV model, an optimal k is nearly 85-90% of K with moderately high α (10-15%) (see Table 1), whereas for WC model, the optimal k is about 95% of K, with comparatively lower values of α (1.5-2.5%). Further, the edge probability enhancement $h(\alpha)p_{uv}$ for an edge (u, v) depends on the initial probability p_{uv}. The larger this probability, lower is the α required to convert an edge from being non-live to live. For NetHEPT dataset, edge probabilities are in a much higher range under WC model than under TV model; so it suffices to have a very low α for the WC model, and a relatively higher α for the TV 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**
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**
200	496.45	527.06	188	0.15	**6.17**

Table 1: Results of simulations on NetHEPT (TV model)

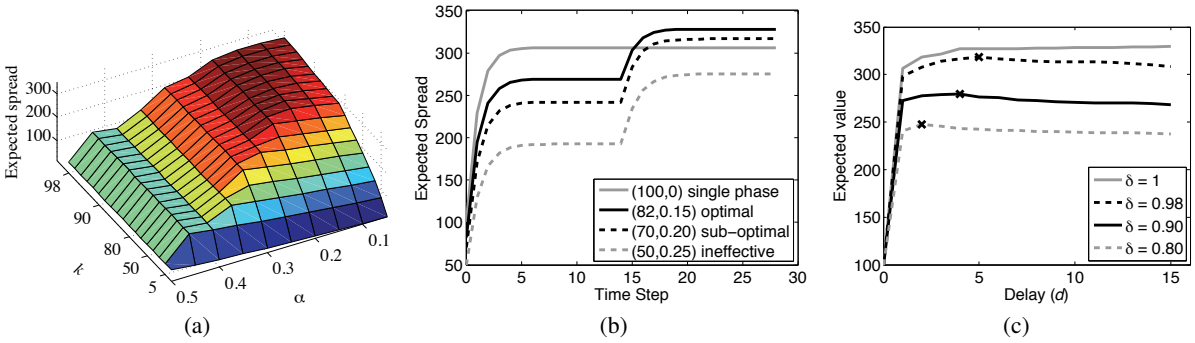

Figure 1: Performance of 2P-SRI on NetHEPT (TV model) for $K = 100$: (a) as a function of k and α, (b) with respect to the progression in time steps, (c) as a function of the delay after which the referral phase is initiated

Effect of Total Budget In our experiments on synthetic data with log-normal and power-law degree distributions, we observe that a budget-split is detrimental for low values of total budget (see Figure 2). If the initial seed set is not of a reasonable size, a very limited number of nodes is activated in the regular diffusion phase. This adversely affects the final spread despite referral incentives; hence if the total budget is low to begin with, splitting it further may not be warranted.

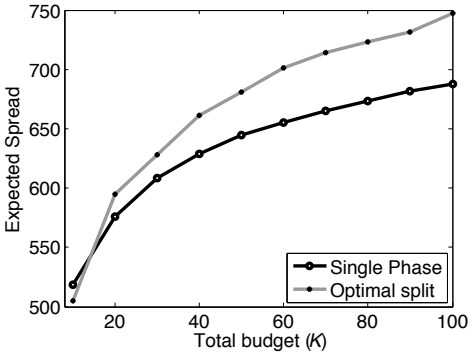

Figure 2: Performance of 2P-SRI as a function of total seed budget on a power-law network of 1000 nodes (TV model)

Scheduling the Referral Phase Figure 1(b) depicts a typical temporal progression of diffusion for different (k, α) splits when we wait long enough for the diffusion in phase 1 to terminate before initiating phase 2. However, this wait may not be advisable in the presence of temporal constraints, e.g., product value decaying over time, where the *rate of diffusion* is critical. This can be captured in a time-discounted objective function $\nu(S) = \sum_{t=0}^{\infty} \delta^t \cdot \sigma_t(S)$, where $\sigma_t(S)$ is the expected number of newly activated nodes at time t, and $\delta \in [0, 1]$ (lower δ means faster decay). Figure 1(c) presents the performance of 2P-SRI as a function of delay d after which the referral phase is initiated, for different values of δ with the corresponding optimal (k, α) pairs. The optimal delay (marked on the plots) decreases as the value of δ lowers. In Figure 1(c), 2P-SRI gives an improvement when δ is relatively high (≥ 0.9), while it under-performs for low δ.

If phase 1 is cut-off after a few time steps (lower delay), the number of active nodes is not large enough to trigger a successful phase 2. Besides, we might end up expending referrals on nodes that could be activated anyway without an incentive, i.e. in phase 1 itself. On the other hand, a higher delay leads to considerable decaying of the product value, which is reflected in the value of expected spread $\nu(S)$. This induces a trade-off in determining the optimal delay.

Acknowledgment

We acknowledge the continued support of Adobe Research Labs, Bangalore in carrying out this project.

References

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

Chen, W.; Wang, C.; and Wang, Y. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *SIGKDD*, 1029–1038. ACM.

Dhamal, S.; Prabuchandran, K. J.; and Narahari, Y. 2016. Information diffusion in social networks in two phases. *IEEE Transactions on Network Science and Engineering* 3(4):197–210.

Hartline, J.; Mirrokni, V.; and Sundararajan, M. 2008. Optimal marketing strategies over social networks. In *Proceedings of the 17th international conference on World Wide Web*, 189–198. ACM.

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

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *SIGKDD*, 137–146. ACM.

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

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

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

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