

# Scalable Methods for Adaptively Seeding a Social Network\*

Thibaut Horel  
Harvard University  
thorel@seas.harvard.edu

Yaron Singer  
Harvard University  
yaron@seas.harvard.edu

## ABSTRACT

In many applications of influence maximization, one is restricted to select influencers from a set of users who engaged with the topic being promoted, and due to the structure of social networks, these users often rank low in terms of their influence potential. To alleviate this issue, one can consider an adaptive method which selects users in a manner which targets their influential neighbors. The advantage of such an approach is that it leverages the friendship paradox in social networks: while users are often not influential, they often know someone who is.

Despite the various complexities in such optimization problems, we show that scalable adaptive seeding is achievable. To show the effectiveness of our methods we collected data from various verticals social network users follow, and applied our methods on it. Our experiments show that adaptive seeding is scalable, and that it obtains dramatic improvements over standard approaches of information dissemination.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—Data Mining; F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems

## 1. INTRODUCTION

Influence Maximization [1] is the algorithmic challenge of selecting a fixed number of individuals who can serve as early adopters of a new idea, product, or technology in a manner that will trigger a large cascade in the social network. In many cases where influence maximization methods are applied one cannot select any user in the network but is limited to some subset of users. In general, we will call the *core set* the set of users an influence maximization campaign can access. When the goal is to select influential users from the core set, the laws governing social networks can lead to poor outcomes: due to the heavy-tailed degree distribution

\*The full version of this work is available as [4].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).  
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.  
ACM 978-1-4503-3473-0/15/05.  
<http://dx.doi.org/10.1145/2740908.2744108>.

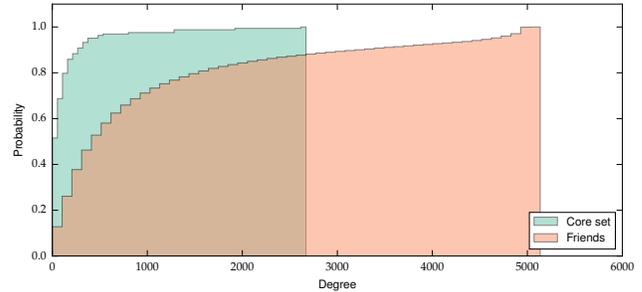


Figure 1: CDF of the degree distribution of users who liked a post by Kiva on Facebook and that of their friends.

of social networks, high degree nodes are rare, and since influence maximization techniques often depend on the ability to select high degree nodes, a naive application of influence maximization techniques to the core set is ineffective.

An alternative method recently introduced in [5] is a two-stage approach called adaptive seeding. In the first stage, one can spend a fraction of the budget on the core users so that they invite their friends to participate in the campaign, then in the second stage spend the rest of the budget on the influential friends who hopefully have arrived. The idea behind this approach is to leverage a structural phenomenon in social networks known as the friendship paradox [2]: even though individuals are not likely to have many friends, they likely have a friend that does (“your friends have more friends than you”). Figure 1 illustrates this effect on Facebook.

In this work, we present efficient algorithms for adaptive seeding achieving an optimal approximation ratio of  $(1 - 1/e)$ . The guarantees of our algorithms hold for linear models of influence. While this class does not include models such as the independent cascade and the linear threshold model, it includes the well-studied *voter model* [3]. We then use these algorithms to conduct a series of experiments to show the potential of adaptive approaches for influence maximization both on synthetic and real social networks.

## 2. FRAMEWORK

**Model.** Given a graph  $G = (V, E)$ , for  $S \subseteq V$  we denote by  $\mathcal{N}(S)$  the neighborhood of  $S$ . The notion of influence in the graph is captured by a function  $f : 2^V \rightarrow \mathbb{R}_+$  mapping a subset of nodes to a non-negative influence value. In this work, we focus on linear influence functions:  $f(S) = \sum_{u \in S} w_u$  where  $(w_u)_{u \in V}$  are non-negative weights capturing the influence of individual vertices. The input of the *adaptive seeding* problem is a *core set* of nodes  $X \subseteq V$  and for any node  $u \in \mathcal{N}(X)$  a probability  $p_u$  that  $u$  realizes if one of its

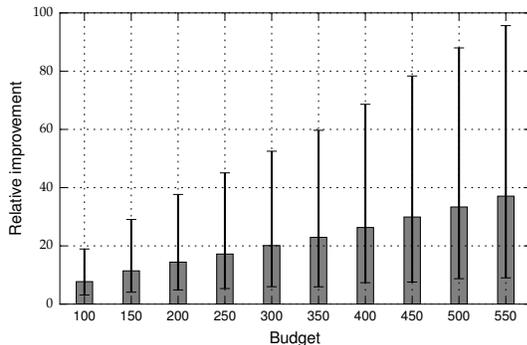


Figure 2: Ratio of the performance of adaptive seeding to IM. Bars represents the mean improvement across all verticals, and the “error bars” represents the range of improvement across verticals.

neighbor in  $X$  is seeded. The goal is to solve:

$$\begin{aligned} \max_{S \subseteq X} \sum_{R \subseteq \mathcal{N}(S)} p_R \max_{T \subseteq R} f(T) \\ \text{s.t. } |S| \leq k \end{aligned} \quad (1)$$

where  $p_R$  is the probability that the set  $R$  realizes,  $p_R \equiv \prod_{u \in R} p_u \prod_{u \in \mathcal{N}(S) \setminus R} (1 - p_u)$ . Intuitively, we want to select at most  $k$  nodes in the core set  $X$  such that the expected maximum influence which can be derived from the set  $R$  of neighbors realizing using the remaining budget is maximal.

**Non-adaptive Optimization.** We say that a policy is *non-adaptive* if it selects a set of nodes  $S \subseteq X$  to be seeded in the first stage and a vector of probabilities  $\mathbf{q} \in [0, 1]^n$ , such that each neighbor  $u$  of  $S$  which realizes is included in the solution independently with probability  $q_u$ . The constraint will now be that the budget is only respected in expectation, *i.e.*  $|S| + \mathbf{p}^T \mathbf{q} \leq k$ . Formally the optimization problem for non-adaptive policies can be written as:

$$\begin{aligned} \max_{\substack{S \subseteq X \\ \mathbf{q} \in [0, 1]^n}} \sum_{R \subseteq \mathcal{N}(X)} \left( \prod_{u \in R} p_u q_u \prod_{u \in \mathcal{N}(X) \setminus R} (1 - p_u q_u) \right) f(R) \\ \text{s.t. } |S| + \mathbf{p}^T \mathbf{q} \leq k, q_u \leq \mathbf{1}\{u \in \mathcal{N}(S)\} \end{aligned} \quad (2)$$

where  $\mathbf{1}\{E\}$  is the indicator variable of the event  $E$ . Non-adaptive policies are related to adaptive policies:

**PROPOSITION 1.** *Let  $(S, \mathbf{q})$  be an  $\alpha$ -approximate solution to (2), then  $S$  is an  $\alpha$ -approximate solution to (1).*

### 3. ALGORITHMS

Proposition 1 allows us to focus on designing non-adaptive policies for (2) which is easier to solve than (1).

Our first algorithm is obtained by considering a relaxation of (2) where the binary choices of including vertices in  $S$  are relaxed to fractional values. The solution must then be rounded using the Pipage Rounding framework.

The second algorithm is combinatorial: first, we note that for additive influence functions and for fixed  $S$ , the maximization over  $\mathbf{q}$  in (2) is a simple fractional knapsack problem which can be solved efficiently. Furthermore, the optimal value of this problem is a monotone submodular function of  $S$ . Our algorithm can thus be obtained by applying the celebrated greedy algorithm for monotone submodular maximization where we repeatedly solve fractional knapsack problems when greedily constructing the solution.

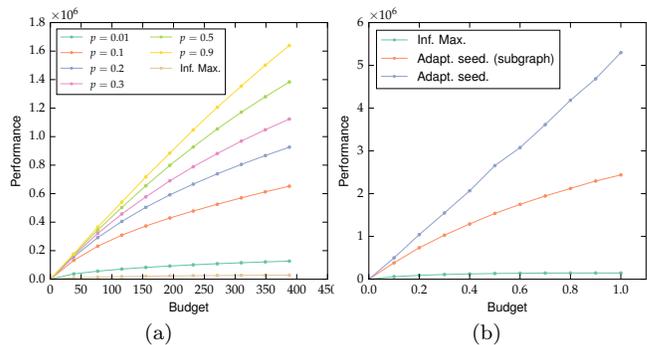


Figure 3: (a) Performance of adaptive seeding for various propagation probabilities. (b) Performance of *adaptive seeding* when restricted to the subgraph of users who *liked* HBO (red line).

Both algorithms achieve an optimal  $(1 - 1/e)$  approximation ratio. The first algorithm is extremely efficient over instances where there is a large budget. The second algorithm can be easily parallelized and implemented in MapReduce, has good theoretical guarantees on its running time and does well on instances with smaller budgets.

### 4. EXPERIMENTS

The main component of our experiments involved collecting publicly available data from Facebook. Despite the extreme difficulty of collecting such data, we were able to collect large networks. For 10 several Facebook Pages, each associated with a commercial entity that uses the Facebook page to communicate with its followers, we selected a post and then collected data about the users who expressed interest (“liked”) the post and their friends. The advantage of this data set is that it is highly representative of the scenario we study here. We focused on posts which were liked by about 1,000 users, which when we include their friends, leads to networks of about 100,000 users.

Figure 2 compares the performance of our approach to running the standard influence maximization (IM) approach to the core set. Figure 3a shows the impact of the probability of neighbors realizing, while Figure 3b shows the performance of adaptive seeding when restricted to users who previously expressed interest in the vertical and for which we could expect the probability of realizing to be close to one. These results suggest that adaptive seeding can have dramatic improvements over standard IM. [4] contains additional experiments to analyze the impact of various parameters as well as evaluations on synthetic data.

**Acknowledgement.** This research is supported in part by a Google Research Grant and NSF grant CCF-1301976.

### 5. REFERENCES

- [1] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.
- [2] S. L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 1991.
- [3] R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975.
- [4] T. Horel and Y. Singer. Scalable methods for adaptively seeding a social network. *CoRR*, abs/1503.01438, 2015.
- [5] L. Seeman and Y. Singer. Adaptive seeding in social networks. In *FOCS*, 2013.