Simultaneous Influencing and Mapping Social Networks

(Extended Abstract)

Leandro Soriano Marcolino[1], Aravind Lakshminarayanan[2], Amulya Yadav[1], Milind Tambe[1]
[1]University of Southern California, Los Angeles, CA, 90089, USA
{sorianom,amulyaya,tambe}@usc.edu
[2]Indian Institute of Technology, Madras, Tamil Nadu, 600036, India
aravindsrinivas@gmail.com

1. INTRODUCTION

Influencing a social network is an important technique, with potential to positively impact society, as we can modify the behavior of a community. For example, we can increase the overall health of a population; Yadav et al. (2015) [4], for instance, spread information about HIV prevention in homeless populations. However, although influence maximization has been extensively studied [2, 1], their main motivation is viral marketing, and hence they assume that the social network graph is fully known, generally taken from some social media network. However, the graphs recorded in social media do not really represent all the people and all the connections of a population. Most critically, when performing interventions in real life, we deal with large degrees of lack of knowledge. Normally the social agencies have to perform several interviews in order to learn the social network graph [3]. These highly unknown networks, however, are exactly the ones we need to influence in order to have a positive impact in the real world, beyond product advertisement.

Additionally, learning a social network graph is very valuable *per se*. Agencies need data about a population, in order to perform future actions to enhance their well-being, and better actuate in their practices [3]. As mentioned, however, the works in influence maximization are currently ignoring this problem. Each person in a social network actually knows other people, including the ones she cannot directly influence. When we select someone for an intervention (to spread influence), we also have an opportunity to obtain knowledge. Therefore, in this work we present for the first time the problem of simultaneously influencing and mapping a social network. We study the performance of the classical influence maximization algorithm in this context, and show that it can be arbitrarily low. Hence, we study a class of algorithms for this problem, performing an experimentation using four real life networks of homeless populations. We show that our algorithm is competitive with previous approaches in terms of influence, and is significantly better in terms of mapping.

2. INFLUENCING AND MAPPING

We consider the problem of maximizing the influence in a social network. However, we start by knowing only a subgraph. Each time we pick a node to influence, it may teach us about subgraphs. Our objective is to spread influence, at the same time learning the network. We call this problem "Simultaneous Influencing and Mapping" (SIAM). We consider a version of SIAM where we only need to map the nodes that compose the network. We assume that we always know all the edges between the nodes of the known subgraph. Formally, let $G := (V, E)$ be a graph with a set of nodes V and edges E. We pick one node at each one of η interventions. The selected node is used to spread influence and map the network. We do not know the graph G, we only know a subgraph $G_k = (V_k, E_k) \subset G$, where k is the current intervention number. G_k starts as $G_k := G_0 \subset G$. For each node v_i, there is a subset of nodes $V^i \subset V$, which will be called "teaching list". Each time we pick a node v_i, the known subgraph changes to $G_k := (V_{k-1} \cup V^i, E_k)$, where E_k contains all edges between the set of nodes $V_{k-1} \cup V^i$ in G. Our objective is to maximize $|V_k|$, given η interventions.

For each v_i, we assume we can observe a number γ_i, which indicates the size of its teaching list. We study two versions: in one γ_i is the number of nodes in V^i that are not yet in G_k (hence, the number of new nodes that will be learned when picking v_i). We refer to this version as "perfect knowledge". In the other, $\gamma_i := |V^i|$, and thus we cannot know how many nodes in V^i are going to be new or intersect with already known nodes in V_k. We refer to this version as "partial knowledge". Note that we may also have nodes with empty teaching lists ($\gamma_i = 0$). The teaching list of a node v_i is the set of nodes that v_i will teach us about once picked, and is not necessarily as complete as the true set of all nodes known by v_i. Some nodes could simply refuse to provide any information. We assume the teaching list and the neighbor list to be independent. That is, a node may teach us about nodes that it is not able to directly influence. For instance, it is common to know people that we do not have direct contact with, or we are not "close" enough to be able to influence. Similarly, a person may not tell us about all her close friends, due to limitations of an interview process, or even "shame" to describe some connections. We consider a probability φ that a node will have a non-empty teaching list.

We also want to maximize the influence over the network. We consider the traditional independent cascade model, with observation, as in Golovin and Krause (2010) [1]. That is, a node may be either influenced or uninfluenced. An uninfluenced node may change to influenced, but an influenced node will never change back to uninfluenced. Each time we pick a node for an intervention, it will change to influenced. When a node changes from uninfluenced to influenced, it will "spread" the influence to its neighbors with some probability. That is, at each edge e there is a probability p_e. When a node v_1 changes to influenced, if there is an edge $e = (v_1, v_2)$, the node v_2 will also change to influenced with probability p_e. Similarly, if v_2 changes to influenced, it will spread the influence to its neighbors by the same process. Influence only spreads in the moment a node changes from uninfluenced to influenced. As in [1], we consider that we have knowledge about whether a node is influenced or not (but in our case, we can only know about nodes in the current known subgraph G_k). Let I_k be the number of influenced

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016),*
J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.),
May 9–13, 2016, Singapore.

nodes after k interventions. Our objective is to maximize I_k given η interventions. Influence may spread beyond G_k. We consider I_k as the number of influenced nodes in the full graph G. We denote as σ_i the expected number of nodes that will be influenced when picking v_i (usually calculated by simulations).

We must maximize both $|V_k|$ and I_k. Similarly to previous influence maximization works [2, 1], we study greedy solutions. The fundamental problem of SIAM is whether to focus on influencing or mapping the network. Hence, we propose as a general framework to select the node v_i such that: $v_i = \text{argmax}(c_1 \times \sigma_i + c_2 \times \gamma_i)$. Constants c_1 and c_2 control the balance between influencing or mapping. $c_1 = 1, c_2 = 0$ is the classical influence maximization algorithm (*"influence-greedy"*); $c_1 = 0, c_2 = 1$, on the other hand, only maximizes the knowledge-gain at each intervention (*"knowledge-greedy"*). $c_1 = c_2 = 1$ is an algorithm where both objectives are equally balanced (*"balanced"*). In order to better handle *partial knowledge*, we also propose the *"balanced-decreasing"* algorithm, where c_2 constantly decreases until reaching 0. Hence, we define c_2 as: $c_2 := \begin{cases} c_2' - \frac{1}{d} \times c_2' \times k & \text{if } k \leq d \\ 0 & \text{otherwise} \end{cases}$, where c_2' is the value for c_2 at the very first iteration, and d controls how fast it decays.

We begin by studying *influence-greedy*. It was shown that when picking the node v which argmax(σ_v) at each intervention, we achieve a solution that is a $(1 - 1/e - \epsilon)$ approximation of the optimal, as long as our estimation of σ_v (by running simulations) is "good enough" [2]. However, even though the actual influence spread may go beyond the known graph G_k, we can only run simulations to estimate σ_v in the current G_k. Hence, the previous results are no longer valid. We show with an example that we can obtain arbitrarily low-performing solutions by using *influence-greedy*.

Consider the graph in Figure 1, and assume we will run 2 interventions (i.e., pick 2 nodes). There is a probability 1 to spread influence in any edge. Our initial knowledge is $V_0 = \{A, A', B, B', C\}$. A and B can influence A' and B', respectively. However, C cannot influence any node. A, B, A' and B' have empty teaching lists. C, on the other hand, can teach us about a connected graph of z nodes. *Influence-greedy*, by running simulations on the known graph, picks nodes A and B, since each can influence one more node. The optimal solution, however, is to pick node C, which will teach us about the connected graph of z nodes. Then, we can pick one node in that graph, and influence $z + 1$ nodes in total. Hence, the *influence-greedy* solution is only $\frac{4}{z+1}$ of the optimal. As z grows, *influence-greedy* will be arbitrarily far from the optimal solution.

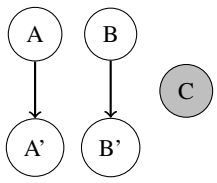
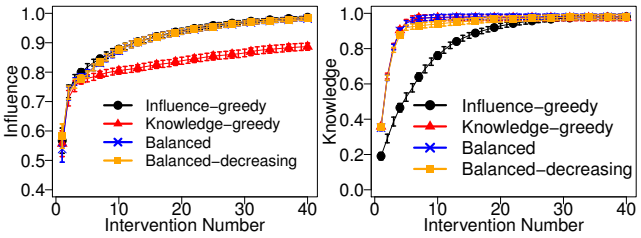


Figure 1: Greedy algorithm has low performance.

3. RESULTS

We run experiments using four real life social networks of the homeless population of Los Angeles, provided by Eric Rice, from the School of Social Work of the University of Southern California. All the networks are friendship-based social networks of homeless youth who visit a social agency. We run 100 executions per network. At the beginning of each execution, 4 nodes are randomly chosen to compose our initial subgraph (G_0). We noticed similar tendencies in the results across all four social networks. Due to space constraints, we plot here the results considering all networks simultaneously (that is, we average over all the 400 executions). In all graphs, the error bars show the confidence interval, with $\rho = 0.01$.

We measure the percentage of influence in the network ("Influence") and percentage of known nodes ("Knowledge"). We consider



(a) Influence (b) Knowledge

Figure 2: Results in real world networks across many interventions.

balanced-decreasing with $c_2' = 1.0$, and $d = 5$. In order to estimate the expected influence spread (σ_v) of each node, we run 1000 simulations before each intervention. Simulations are run in the current known subgraph G_k, although the actual influence may go beyond G_k (which will be considered when we measure Influence). Concerning γ_v, we consider it to hold the number of *new* nodes that would be learned if v is selected, for *balanced* and *knowledge-greedy* (i.e., perfect knowledge). For *balanced-decreasing*, we consider γ_v to hold the full teaching list size, including nodes that are already known (i.e., partial knowledge). Hence, we evaluate if *balanced-decreasing* approximates well *balanced*.

We simulate the teaching lists, since there are no real world data available yet (we only have data about the connections in the four networks). If a node has a teaching list, we fix its size according to a uniform distribution from 0 to $0.5 \times |V|$. Each node in the graph is also equally likely to be in the teaching list of a node v_i.

Figure 2 shows the result at each intervention for $\varphi = 0.5$ and $p = 0.5$. The Influence obtained by *influence-greedy*, *balanced*, and *balanced-decreasing* are similar. Out of all 40 interventions, their result is not significantly different in any of them (and they are significantly better than *knowledge-greedy* in around 75% of the interventions). This shows that *balanced* is able to successfully spread influence in the network, while at the same time mapping the graph. We also notice that perfect knowledge about the number of new nodes in the teaching lists is not necessary, as *balanced-decreasing* obtained close results to *balanced*. In terms of Knowledge, all algorithms clearly outperform *influence-greedy* with statistical significance. Moreover, the result for *knowledge-greedy*, *balanced* and *balanced-decreasing* are not significantly different in any of the interventions. This shows that we are able to successfully map the network (as well as *knowledge-greedy*), but at the same time spreading influence successfully over the network (as well as *influence-greedy*), even in the *partial knowledge* case.

Acknowledgments: This research was supported by MURI grant W911NF-11-1-0332, and by IUSSTF.

REFERENCES

[1] D. Golovin and A. Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *COLT*, 2010.

[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[3] P. V. Marsden. Recent developments in network measurement. In *Models and methods in social network analysis*. Cambridge University Press, 2005.

[4] A. Yadav, L. S. Marcolino, E. Rice, R. Petering, H. Winetrobe, H. Rhoades, M. Tambe, and H. Carmichael. Preventing HIV spread in homeless populations using PSINET. In *IAAI*, 2015.