Proceedings of the 3rd International Workshop on Social Influence Analysis (SocInf 2017)
August 19th, 2017 - Melbourne, Australia

Influence Maximization with an Unknown Network by Exploiting Community Structure

Bryan Wilder[1], Nicole Immorlica[2] Eric Rice[1], and Milind Tambe[1]

[1] University of Southern California,
Center for Artificial Intelligence in Society
Los Angeles, CA
{bwilder, erice, tambe}@usc.edu,
[2] Microsoft Research, New England,
Cambridge, MA
nicimm@gmail.com

Abstract. In many real world applications of influence maximization, practitioners intervene in a population whose social structure is initially unknown. We formalize this problem by introducing *exploratory influence maximization*, in which an algorithm queries individual network nodes to learn their links. The goal is to locate a seed set nearly as influential as the global optimum using very few queries. We show that this problem is intractable for general graphs. However, real world networks typically have community structure, in which nodes are arranged in densely connected subgroups. We present the ARISEN algorithm, which leverages community structure to find an influential seed set by querying only a fraction of the network. Experiments on real world networks of homeless youth, village populations in India, and others validate ARISEN's performance.

1 Introduction

In contexts ranging from health, to international development, to education, practitioners have used the social network of their target population to rapidly spread information and to change behavior in socially desirable ways. The challenge is to identify the influential members of the population. While previous work has delivered many computationally efficient algorithms for this *influence maximization* problem [7, 21, 12], this work assumes that the social network is given explicitly as input. However, in many real-world domains, the network is not initially known and must be gathered via laborious field observations. For example, collecting network data from vulnerable populations such as homeless youth, while crucial for health interventions, requires significant time spent gathering field observations [19]. Social media data is often unavailable when access

to technology is limited, for instance in developing countries or with vulnerable populations. Even when such data is available, it often includes many weak links which are not effective at spreading influence [2]. For instance, a person may have hundreds of Facebook friends who they barely know. In principle, the entire network could be reconstructed via surveys, and then existing influence maximization algorithms applied. However, exhaustively reconstructing the network is very labor-intensive and considered impractical in many situations [22]. For influence maximization to be relevant to many real-world problems, it must contend with limited *information* about the network, not just limited *computation*.

The major informational restriction is the number of nodes which may be surveyed to explore the network. Thus, a key question is: *how can we find influential nodes with a small number of queries?* Existing field work uses heuristics, such as sampling some percentage of the nodes and asking them to nominate influencers [22]. We formalize this problem as *exploratory influence maximization* and seek a principled algorithmic solution, i.e., an algorithm which makes a small number of queries and returns a set of seed nodes which are approximately as influential as the the globally optimal seed set. To the best of our knowledge, no previous work directly addresses this question from an algorithmic perspective (we survey the closest work in Section 3).

We show that for general graphs, any algorithm for exploratory influence maximization may perform arbitrarily badly unless it examines almost the entire network. However, real world networks have useful structure. In particular, social networks often have strong *community* structure, where nodes are arranged into groups which are connected tightly internally, but only weakly to the rest of the network [10, 16]. Consequently, influence mostly propagates in a local fashion. Community structure has been used to develop more computationally efficient influence maximization algorithms [23, 8]. Here, we use it to design a highly information-efficient algorithm. We make three main contributions. *First*, we introduce exploratory influence maximization and show that it is intractable for general graphs. *Second*, we present the ARISEN algorithm, which exploits community structure to find an influential seed set. *Third*, we show experimental results on a variety of networks(both synthetic and real) that verify ARISEN's performance. Our focus here is on introducing the algorithm and showing experimental results; theoretical analysis of ARISEN's performance will be presented in future work. In this paper, we focus on the description of the problem and survey related work. We then briefly present the high-level idea of our algorithm and give an example of experimental results.

2 Exploratory influence maximization

As a motivating example, consider a homeless youth shelter which wishes to spread HIV prevention information [19]. It would try to harness the youths' social network and select the most influential peer leaders to spread information, but this network is not initially known. Constructing the network requires a

laborious survey [19]. Our motivation is to mitigate this effort by querying only a few youth. Such queries require much less time than the day-long training peer leaders receive. We now formalize this problem.

Influence maximization: The influence maximization problem [13], starts with a graph $G = (V, E)$, where $|V| = n$ and $|E| = m$. We assume throughout that G is undirected; social links are typically reciprocal [20]. An influencer selects K seed nodes with the aim of maximizing the expected size of the resulting influence cascade. We assume that influence propagates according to the independent cascade model (ICM), which is the most prevalent model in the literature. Initially, all nodes are inactive except for the seed set. When a node becomes active, it makes one attempt to activate each neighbor. Each attempt succeeds independently with probability q, where q is typically assumed to be the same for all edges [7, 13, 25]. Let $f(S)$ denote the expected number of activated nodes with seed set $S \subseteq V$. The objective is to compute $\arg\max_{|S|\leq K} f(S)$.

Local information: The edge set E is not initially known. Instead, the algorithm explores portions of the graph using local operations. We use the popular "Jump-Crawl" model [5], where the algorithm may either jump to a uniformly random node, or crawl along an edge from an already surveyed node to one of its neighbors. When visited, a node reveals all of its edges. We say that the *query cost* of an algorithm is the total number of nodes visited using either operation. Our goal is to find influential nodes with a query cost that is much less than n, the total number of nodes.

Stochastic Block Model: In our formal analysis, we assume that the graph is drawn from the SBM. The SBM originated in sociology [9] and lately has been intensively studied in computer science and statistics (see e.g. [1, 15, 18]). In the SBM, the network is partitioned into disjoint communities $C_1....C_L$. Each within-community edge is present independently with probability p_w and each between-community edge is present independently with probability p_b. Notice that each community is an Erdős-Rényi random graph with additional random edges to other communities. We assume that $p_w \geq \frac{\log |C_i|}{|C_i|}$ for all C_i, since this is necessary for C_i to be internally connected [11]. While the SBM is a simplified model, our experimental results show that ARISEN performs well on real-world graphs. ARISEN takes as input the parameters n, p_w, and p_b, but is not given any prior information about the realized draw of the network. It is reasonable to assume that the model parameters are known since they can be estimated using existing network data from a similar population (in our experiments, we show that this approach works well).

Objective: We compare to the globally optimal solution, i.e, the best performance if the entire network structure were known in advance. Let $f_E(S)$ give the expected number of nodes influenced by seed set S when the set of realized edges are E. Let $\mathcal{A}(E)$ be the (possibly random) seed set containing our algorithm's selections given edge set E. Let OPT be the expected value of the globally optimal solution which seeds K nodes. We measure the algorithm's performance by the ratio $OPT/\,\mathbb{E}[f_E(\mathcal{A}(E))]$, where the expectation is over both the randomness in the graph and the algorithm's choices.

3 Related work

First, Yadav et al. [25] and Wilder et al. [24], studied dynamic influence maximization over a series of rounds. Some edges are "uncertain" and are only present with some probability; the algorithm can gain information about these edges in each round. However, the majority of potential edges are known in advance. By contrast, our work does not require *any* known edges. Mihara et al. [17] also consider influence maximization over a series of rounds, but in their work the network is initially unknown. In each round, the algorithm makes some queries, selects some seed nodes, and observes all of the nodes which are activated by its chosen seeds. The ability to observe activated nodes makes our problem incomparable with theirs because these activations can reveal a great deal about the network and give the algorithm information that even the global optimizer does not have (in their work, the benchmark does not use the activations). Thus, we emphasize that our algorithm uses strictly less information. Further, in many applications, activations correspond to private behaviors (e.g. getting tested for HIV) and cannot be observed for practical or legal reasons.

Another line of work concerns local graph algorithms, where a local algorithm only uses the neighborhoods around individual nodes. Borgs et al. [3] study local algorithms for finding the root node in a preferential attachment graph and for constructing a minimum dominating set. Other work, including Bressen et al. [6] and Borgs et al. [4], aims to find nodes with high PageRank using local queries. These algorithms are not suitable for our problem since a great deal of previous work has observed that picking high PageRank nodes as seeds can prove highly suboptimal for influence maximization [14, 7, 12]. Essentially, PageRank identifies a set of nodes that are *individually* central, while influence maximization aims to find a set of nodes which are *collectively* best at diffusing information. We also emphasize that our technical approach is entirely distinct from work on PageRank.

4 Proposed algorithm and results

We now provide a brief overview of our algorithm for exploratory influence maximization. The main idea is to sample a set of T random nodes $\{v_1 ... v_T\}$ from G and explore a small subgraph H_i around each v_i by taking R steps of a random walk. We discard the first B steps of each walk as burn-in. R, T and B are inputs set by the user according to the size of the network and the number of seeds they wish to select. Intuitively, T should be greater than K so we can be sure of sampling each of the largest K communities. The subgraphs H_i are used to construct a weight vector $\boldsymbol{w}$ where w_i gives the weight associated with v_i. The algorithm then independently samples each seed from $\{v_1 ... v_T\}$ with probability proportional to $\boldsymbol{w}$. Further details will appear in an extended version of the paper.

Figure 1 shows experimental results on a network of households in a village in rural india. We compare our algorithm (in blue with diagonal hatches) to a series

of baselines. From left to right, the baselines are (1) selecting K random node and seeding their highest degree neighbor (2) starting at a random node and iteratively seeding the highest degree neighbor of the previous node (3) selecting K seeds at random. The y axis plots the fraction of the optimal value (assuming that the true network were known) attained by each algorithm. We see that the proposed algorithm outperforms all baselines.

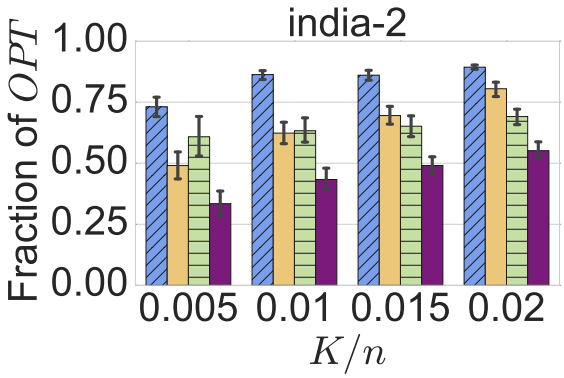

Fig. 1. Influence spread compared to OPT as K varies with $q = 0.15$.

5 Conclusion

We introduced exploratory influence maximization to study influence maximization when the network is initially unknown. We presented a novel algorithm, which exploits the community structure present in many real world social networks. Experimental results on a real world network show that our algorithm is competitive with the global optimum and outperforms several baselines.

References

1. Abbe, E., Sandon, C.: Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In: FOCS. pp. 670–688. IEEE (2015)
2. Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E., Fowler, J.H.: A 61-million-person experiment in social influence and political mobilization. Nature 489(7415), 295–298 (2012)
3. Borgs, C., Brautbar, M., Chayes, J., Khanna, S., Lucier, B.: The power of local information in social networks. In: WINE. pp. 406–419. Springer (2012)
4. Borgs, C., Brautbar, M., Chayes, J., Teng, S.H.: Multiscale matrix sampling and sublinear-time pagerank computation. Internet Mathematics 10(1-2), 20–48 (2014)

5. Brautbar, M., Kearns, M.J.: Local algorithms for finding interesting individuals in large networks. In: Innovations in Theoretical Computer Science. pp. 188–199 (2010)
6. Bressan, M., Peserico, E., Pretto, L.: The power of local information in pagerank. In: WWW. pp. 179–180. ACM (2013)
7. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: KDD. pp. 1029–1038. ACM (2010)
8. Chen, Y.C., Zhu, W.Y., Peng, W.C., Lee, W.C., Lee, S.Y.: Cim: community-based influence maximization in social networks. ACM Transactions on Intelligent Systems and Technology (TIST) 5(2), 25 (2014)
9. Fienberg, S.E., Wasserman, S.S.: Categorical data analysis of single sociometric relations. Sociological methodology 12, 156–192 (1981)
10. Girvan, M., Newman, M.E.: Community structure in social and biological networks. PNAS 99(12), 7821–7826 (2002)
11. Janson, S., Luczak, T., Rucinski, A.: Random graphs, vol. 45. John Wiley & Sons (2011)
12. Jung, K., Heo, W., Chen, W.: Irie: Scalable and robust influence maximization in social networks. In: ICDM. pp. 918–923. IEEE (2012)
13. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD. pp. 137–146. ACM (2003)
14. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Finding influential nodes in a social network from information diffusion data. In: Social Computing and Behavioral Modeling, pp. 1–8. Springer (2009)
15. Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., Zhang, P.: Spectral redemption in clustering sparse networks. PNAS 110(52), 20935–20940 (2013)
16. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics 6(1), 29–123 (2009)
17. Mihara, S., Tsugawa, S., Ohsaki, H.: Influence maximization problem for unknown social networks. In: ASONAM. pp. 1539–1546. ACM (2015)
18. Mossel, E., Neeman, J., Sly, A.: Reconstruction and estimation in the planted partition model. Probability Theory and Related Fields 162(3-4), 431–461 (2015)
19. Rice, E., Tulbert, E., Cederbaum, J., Adhikari, A.B., Milburn, N.G.: Mobilizing homeless youth for HIV prevention. Health education research 27(2), 226–236 (2012)
20. Squartini, T., Picciolo, F., Ruzzenenti, F., Garlaschelli, D.: Reciprocity of weighted networks. Scientific Reports (2012)
21. Tang, Y., Xiao, X., Shi, Y.: Influence maximization: Near-optimal time complexity meets practical efficiency. In: KDD. ACM (2014)
22. Valente, T.W., Pumpuang, P.: Identifying opinion leaders to promote behavior change. Health Education & Behavior (2007)
23. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: KDD. pp. 1039–1048. ACM (2010)
24. Wilder, B., Yadav, A., Immorlica, N., Rice, E., Tambe, M.: Uncharted but not uninfluenced: Influence maximization with an uncertain network. In: AAMAS. pp. 740–748 (2017)
25. Yadav, A., Chan, H., Xin Jiang, A., Xu, H., Rice, E., Tambe, M.: Using social networks to aid homeless shelters: Dynamic influence maximization under uncertainty. In: AAMAS. pp. 740–748 (2016)