Heuristic Algorithms for Influence Maximization with Partial Information*

Soheil Eshghi[2], Setareh Magshudi[2,3], Valerio Restocchi[1], Leandros Tassulias[2], Rachel K. E. Bellamy[4], Nicholas R. Jennings[5], Sebastian Stein[1]

[1] University of Southampton, Southampton, UK (corresponding author, email: ss2@soton.ac.uk)
[2] Yale University, New Haven, USA
[3] Technische Universität Berlin
[4] IBM T.J. Watson Research, Yorktown Heights, NY, USA
[5] Imperial College London, UK

1 Summary

Models to study the propagation of opinions and influence in social networks have been extensively studied and, in particular, the computer science literature has focused on developing algorithms to maximise the spread of influence. However, little work considers the common real-world scenario in which only portions of the full network are visible or only a subset of nodes can be chosen to spread influence from. In particular, in this paper we explore influence maximisation under a type of uncertainty which has not been investigated so far. In our setting, a part (or some parts) of a network is known (e.g., individuals that belong to the decision maker's organisation), while the rest is completely unobservable.

We propose a set of heuristic algorithms designed to maximise the spread of influence in such a setting, by preferentially targeting boundary nodes. We consider the case of organisation-partitioned networks, i.e., networks in which a subset of nodes (a community) and all links among them are fully visible, but the rest are unknown. We show that, in such a setting, the proposed algorithms outperform the state of the art by up to 38%.

2 Method

We propose three heuristic algorithms to select seed nodes:

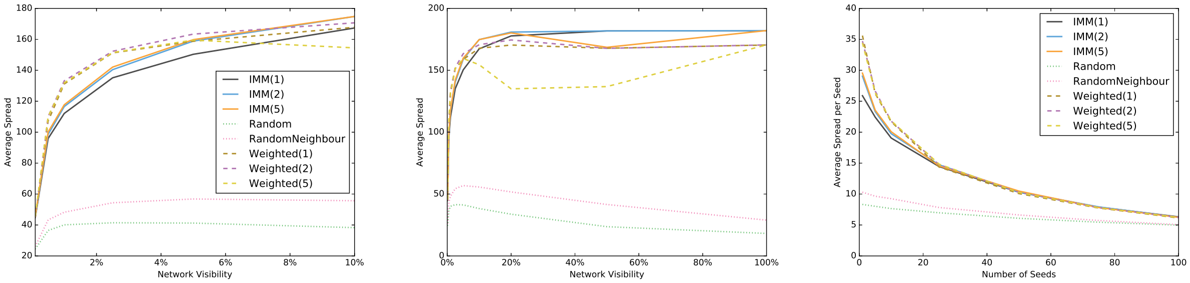
- **Random Selection:** In this case, we select the seeds simply at random.
- **Random with Neighbour Activation:** Here, we first select the seeds at random, and then for each seed, we activate one of its neighbours. This approach is based on the friendship paradox [2]. It states that on an average basis, most people have fewer friends than their friends have. Thus, if we select some seed at random, it is beneficial to activate one of its neighbours, instead of the original node itself, due to possible larger number of connections.
- **Selection based on (Weighted) Degree:** In this heuristic, we first rank the nodes in the known part of the network based on their degree, so that nodes with larger number of neighbours have higher ranks. Then we select node with highest rank as seeds. Here we use the intuition that nodes with larger number of connections are very likely to be highly influential. In the weighted version of this approach, we still rank the nodes based on their degree; however, in order to improve the influence probability in the unknown part of the network, we attach a higher weight for neighbours that are in the boundary set.

To compute the propagation of influence, we use the NetHept dataset, a network of 15k nodes and 31k edges (representing citations within the high energy physics theory community). We choose this because it constitutes a real-world dataset of reasonable size and because it has been widely used to benchmark influence maximisation algorithms [1]. We compare the performance of the proposed heuristics (measured as the average number of nodes influenced per seed) with that of the most most successful influence maximisation algorithm with theoretical performance guarantees, IMM [1].

3 Results

Figures 1a and 1b show the average spread of all algorithms in a setting with five seed nodes, as we vary the network visibility, i.e., the proportion of fully observable nodes. As expected, the state-of-the-art IMM algorithm performs

[*] A full version of this paper is available: S. Stein, S. Eshghi, S. Maghsudi, L. Tassiulas, R. Bellamy, N. Jennings. Heuristic Algorithms for Influence Maximization in Partially Observable Social Networks. SOCINF 2017



(a) (b) (c)

Fig. 1: Figures a and b show the average spread in partially observable networks for 5 seeds. Figure c shows the average spread for varying numbers of seed with visibility 1%.

will throughout all settings. However, looking more closely at cases with low observability (Figure 1a), some of the heuristic approaches outperform it. Specifically, the degree-based heuristics perform consistently well, sometimes achieving an up to 19% higher average spread than IMM. As visibility rises (also continuing in Figure 1b), this difference becomes less pronounced, and by 10% visibility, they achieve a similar performance. As visibility rises further, IMM(1) eventually achieves the highest performance (from 50% visibility onwards).

Looking specifically at the effect of adding higher weights for boundary nodes to the heuristics (denoted as WD(*w*) and IMM(*w*), where *w* is the weight attached to boundary nodes), this can lead to a significant increase in the average spread. However, the performance is sensitive to the exact parameter value, and for networks with higher visibility, a high weight can indeed lead to a decrease in performance, and is most pronounced for Weighted(5) in settings with 20-50% visibility. This is because the boundary nodes actually decrease in importance in the network as more of it is known to the algorithm.

Finally, Figure 1c shows the average spread per initial seed chosen as the number of initial seeds is increased (in a setting with 1% visibility). This highlights that our heuristic techniques achieve the highest performance gains over the state of the art in settings with fewer initial seeds (specifically, a gain of up to 38% when there is just a single initial seed).

Overall, these are promising results, showing that in settings where large parts of the network are not observable and where only few seeds can be chosen, the state-of-the-art algorithm does not necessarily perform best. Instead simple heuristics perform well, and both those heuristics and the current state of the art benefit from explicitly favouring nodes at the boundary of the known network. It should also be noted that the heuristics are several order of magnitude faster than IMM —a typical run of IMM took about 0.1-0.2 seconds, while the degree-based heuristics typically completed within 0.2-0.3ms.

References

1. Tang, Y. and Xiao, X. and Shi, Y.: Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (2014)
2. Zuckerman, E. and Jost, J.T.: What makes you think you're so popular? Self-evaluation maintenance and the subjective side of the "friendship paradox". Social Psychology Quarterly (2001)