Influence Maximisation beyond Organisational Boundaries

Sebastian Stein*, Soheil Eshghi†, Setareh Maghsudi†, Leandros Tassiulas†,
Rachel K. E. Bellamy‡, Nicholas R. Jennings§
* *University of Southampton, Southampton, UK. Email: ss2@ecs.soton.ac.uk*
† *Yale University, New Haven, USA. E-mail: {soheil.eshghi, setareh.maghsudi, leandros.tassiulas} @yale.edu*
‡ *IBM T.J. Watson Research, Yorktown Heights, USA. E-mail: rachel@us.ibm.com*
§ *Imperial College, London, UK. E-mail: n.jennings@imperial.ac.uk*

***Abstract*—We consider the problem of choosing influential members within a social network, in order to disseminate a message as widely as possible. While this so-called problem of *influence maximisation* has been widely studied, little work considers partially-observable networks, where only part of a network is visible to the decision maker. Yet, this is critical in many applications, where an organisation needs to distribute its message far beyond its boundaries and beyond its usual sphere of influence. In this paper, we show that existing algorithms are not sufficient to handle such scenarios. To address this, we propose a set of novel adaptive algorithms that perform well in partially observable settings, achieving an up to 18% improvement on the non-adaptive state of the art.**

I. Introduction

Online social networks provide many connections across geographical, cultural and organisational boundaries that can be leveraged by multi-organisation federations for humanitarian mobilisation [1], participatory sensing [2] and crowdsourcing [3]. Unlocking this potential requires efficient message dissemination far beyond an organisation's usual domain of direct influence. In these circumstances, organisations can ask key individuals to spread the message within their respective social networks, ideally creating far-reaching cascades of repeated messages. Thus, identifying key individuals to target with their initial messaging is critical to aid those who wish to have their messages reach the most people, e.g., advertisers, marketers or public health professionals.

Efficiently choosing such influential individuals, or *influence maximisation* (IM), is a well-studied problem in the computer science literature [4]–[7]. This line of work proposes algorithms that identify influential individuals and predict network spread with some operational guarantees given a fully known network.

However, in practical settings, it is rare to have the comprehensive knowledge about a network required by most work in the IM tradition. In many applications, organisations and federations may know network links between their members, but have little information about the wider network. Some recent work has considered such partially observable networks, where only part of a network is visible to the decision maker (see Section II). Yet, this usually assumes that some information about the unobservable networks is known, e.g., the nodes or the edges between them (but not the strength of connections). Furthermore, these approaches typically consider settings where more information about the unseen network is revealed as messages are passed on. Thus, existing algorithms typically focus on exploring or probing the network and balancing this with the exploitation of influential nodes.

In contrast, we consider settings where no knowledge about the unobservable network is assumed and no information about the message spread beyond the visible part is revealed. This occurs in many real-life settings, e.g., where messages on online social networks are only visible to direct neighbours, or where members of other organisations wish to keep their communications secret.

In this paper, we take preliminary steps towards addressing these settings. We survey the state of the art and propose a new model for IM in partially observable networks. We propose two heuristic techniques for modifying existing IM algorithms to deal with these settings in an adaptive manner, and we show that the new algorithms we derive in this way outperform the non-adaptive state of the art by up to 18%.

II. Related Work

In a basic influence maximisation problem, a network is specified along with an *activation function* that describes the spread of influence. A message is passed through the network, and any individual influenced by the message is considered as being *activated*. Initially, no individual is influenced by the message (activated). The decision maker selects a fixed number of individuals as seeds, which become activated in order to spread the message. In essence, they activate their neighbours, according to the specified activation function. The decision maker seeks to select seeds in a way that maximises the total number of activated individuals [6]. While solving this problem is NP-hard in general, there are several computationally efficient approximate solutions with performance guarantees [6]–[10]. One particularly promising

set of algorithms use the concept of reverse reachable (RR) sets to select seed nodes. These include the TIM, TIM+ [11] and IMM [12] algorithms, which are currently considered to be among the top-performing influence maximisation approaches on realistic benchmark problems [13].

Although most work in this space considers fully observable networks, where the structure and the probability of a message spreading from one node to another is known by the decision maker, some work investigates the influence maximisation problem in partially observable settings, where the structure of the network and/or the activation probabilities are (partially) unknown. If the seed selection is performed in one shot, the selection is either simply at random or based on the characteristics of the observable part of the network. In repeated seed selection, however, there is more room to develop sophisticated influence maximisation algorithms. For instance, by using an online decision making formulation, a balance is found between activating nodes (immediate influence spread) and probing the rest of the network to find best seeds (future influence spread). In particular, multi-armed bandit model and solutions are widely-used. For instance, [14] studies the influence maximisation in the absence of complete information on activation probability. The seed selection phase consists of two parts: (i) Exploration: any conventional influence maximisation algorithm is used to select seeds; (ii) Exploitation: a bandit algorithm, for instance UCB, selects the seeds. In the activation phase, the selected seeds are activated in the real world and the feedback is observed. Based on the feedback, the social graph is updated.

In [15], influence maximisation for unknown social networks is also investigated. Therein, a heuristic algorithm is proposed which works as follows: At every round, first a set of nodes are probed. The nodes are selected if they offer a large expected degree with high probability. Afterwards, all of the (so far) probed nodes are ranked based on their degree, and the nodes with the highest ranks are selected as seeds.

Moreover, some work, such as [16] or [17], uses robust optimisation techniques. In [16], the authors address the uncertainty in the edge influence probability. More precisely, in the model, instead of the exact probability of influence, every edge of the social graph is associated with some interval in which the true probability may lie. Then the goal is defined so as to maximise the worst-case ratio between the influence spread of the chosen seeds and the optimal set of seeds.

In contrast to all these approaches, we explore settings where parts of the network are known (e.g., individuals that belong to the decision maker's organisation), while other parts are completely unobservable and no further information is revealed about those.

III. Model and Problem

We consider a social network, $G = (V, E, w)$. Here, $V = \{1, 2, \ldots, N\}$ is a set of nodes and $E \subseteq V \times V$ is a set of directed edges. Nodes represent individuals and edges indicate who influences whom. We assume these edges are weighted through a function $w : E \to [0, 1]$. Initially, a seed set $S \subset V$ becomes activated, while the remaining nodes are inactivated. An activated node causes their neighbours to become (and stay) activated according to one of two types of activation functions commonly used in the literature:

1) **Independent Cascade [18]:** When a node i is first activated, it is given one chance to activate each neighbour j with probability $w(i, j)$.
2) **Weighted Cascade [6]:** When a node i is first activated, it is given one chance to activate each neighbour with probability $w(i, j) = \frac{1}{|\{(k,j) \in E\}|}$, where $|\{(k, j) \in E\}|$ is the in-degree of j.

The seed set S can be picked all at once, or one at a time. The benefit of the second, less studied, option is that it can incorporate feedback from the activations caused by prior choices, which are realisations of a stochastic process. We call the latter approach, which is a more realistic model for some network settings, *adaptive influence maximisation* to distinguish it from the former, more well studied case. Note that any adaptive process can replicate the all-at-once (i.e., traditional IM) solution. Unlike the state-of-the-art, in this paper we confine our attention to the adaptivec case. Further departing from existing models in the literature, we assume that the network is only partially observable. Specifically, we assume that the decision maker is aware of only a subset of nodes (the *observable* nodes) $V' \subseteq V$, a subset of edges (the *observable* edges) $E' \subseteq (E \cap V' \times V')$. Similarly, there is an *observable* weight function $w' \subseteq w$. We let $G' = (V', E', w')$ denote the observable part of the network.

One specific case of partially observable networks that we are particularly interested in are *organisation-partitioned* networks. In these networks, a subset of nodes reveal their neighbours to the decision makers, e.g., because they are members of the decision maker's organisation (or members of a multi-organisation federation). More formally, the set V is partitioned into two sets O and $\bar{O}$ to mark members of the organisation and those outside it. Here, V' is the union of O and those nodes in $\bar{O}$ where an edge exists between that node and a node in O. We denote the latter set as *boundary* nodes (sometimes known as frontier nodes), while the nodes in O are denoted as *fully observable* nodes. Similarly, E' comprises all edges $(i, j) \in E$, where $i \in O$ or $j \in O$, and w' is defined for all edges in E'. In such networks, we assume that only members of the decision maker's

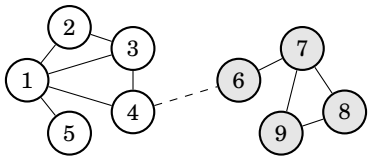


Figure 1: Illustrative example of an organisation-partitioned network.

organisation can be chosen as seeds (i.e., those in O).

Concluding this section, we define the general problem of *adaptive influence maximisation in partially observable networks* as follows: Given a specific activation function, a seed set size M and the observable part of a network G', adaptively choose a seed set $S \subseteq V'$ (with $|S| \leq M$) that maximises the total number of activated nodes in the *underlying* network G.

IV. Motivating Example

Figure 1 provides an illustrative example of an organisation-partitioned network. Here, the unshaded nodes represent members of the decision maker's organisation (i.e., O), while the shaded nodes represent those outside it (i.e., $\bar{O}$). Now, it is possible to use standard IM algorithms on such a partially observable network (with $V' = \{1,2,3,4,5,6\}$), assuming that no nodes exist beyond the observable boundary of the network. However, this will result in highly connected nodes within the organisation to be preferred (e.g., node 1 if $M = 1$). This is not necessarily optimal, because it ignores the possible existence of unobserved nodes. In Figure 1, node 4 acts as a bridge to another cluster of nodes. Such *weak* ties (shown as a dashed line) are in fact widely recognised as being instrumental in the spread of information [19], and this needs to be considered by IM techniques in partially observable networks. In Section V-A, we will present a simple technique for attaching a higher importance to such nodes.

This example also helps illustrate the importance of adaptive methods for influence maximisation in practical settings. Note that if a highly connected node within an organisation is picked as a seed, nodes that are close by have a high likelihood of becoming activated. On the other hand, activation is a stochastic process, and if a highly connected neighbour of the original seed is observed to have not activated, it would be very beneficial to target this node in the second round of seed selection. For example, in the example above, if node 4 is the first seed chosen and is observed not to have activated node 1, then it would be reasonable to pick node 1 as the next seed. On the other hand, if node 4 activates node 1, then targeting another node, such as node 3, as the second seed would be more reasonable. Traditional influence maximisation algorithms cannot differentiate between these two scenarios.

V. Techniques for Influence Maximisation in Organisation-Partitioned Networks

In order to address settings with partially observable organisation-partitioned networks, we propose two techniques. These are heuristic approaches that can be used to improve the performance of existing influence maximisation algorithms. We describe these in the following two sections.

A. Boundary Node Bias

One possible modification to existing algorithms is to add a bias for nodes connected to boundary nodes in organisation-partitioned networks. This is achieved by adding a weight w to boundary nodes that dictates how much more important a boundary node is compared to regular nodes that are part of the decision maker's organisation. For example, a weight $w = 2$ indicates that potentially activating one boundary node should be treated as activating two regular nodes by the algorithm. How to implement this depends on the particular influence maximisation algorithm that is being modified. We briefly outline how to do this for two common algorithms below (which we will use as key benchmarks throughout the remainder of the paper):

- **Degree Centrality**: This algorithm selects seed nodes with the highest degree centrality first, a common measure for identifying influential nodes within a social network [20]. Specifically, we here consider the out-degree of a node i, i.e., the number of neighbours that i may potentially influence. It is straightforward to adapt this to a weighted version — each regular neighbour in O contributes 1 to the weighted out-degree, while each boundary neighbour in $\bar{O}$ contributes w.
- **Reverse Reachable (RR) Sets**: Algorithms based on RR sets, like TIM, TIM+ and IMM [11], [12] can be similarly modified to associate a greater weight w to boundary nodes. Specifically, these algorithms rely on maximising the number of RR sets that a set of seed nodes covers. Instead of treating these sets equally, we can attach a different weight w to RR sets for boundary nodes.

B. Adaptive Influence Maximisation

Any existing influence maximisation algorithm can be run in an adaptive manner, as explained in Section III. Specifically, the algorithm can be used to select a single seed in the original network. This is then activated and the spread from the chosen seed is observed. Then, when the spread is observed, any activated nodes (and their adjacent edges) are removed from the network and the

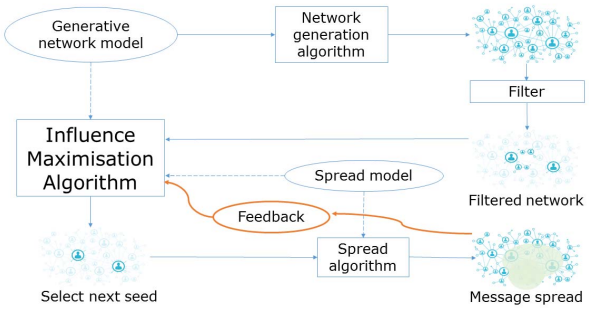


Figure 2: Benchmarking framework

algorithm is run again on the new network to select the next seed. This continues until M seeds are chosen.

In practical terms, such an approach is only feasible when the decision maker has sufficient time to observe the results before choosing the next seed. Furthermore, running certain algorithms (like those based on RR sets) can be computationally expensive, even if only one seed is selected. Thus, an adaptive execution of an existing algorithm may substantially increase the computational burden. However, we leave these issues to future work.

In the following section, we investigate whether boundary node bias and adaptive influence maximisation can improve performance in settings with partial observability.

VI. Experiments

We developed a comprehensive benchmarking framework for evaluating influence maximisation algorithms on partially observable networks (see Figure 2). In this figure, dashed lines indicate the use of a probabilistic model, and solid lines indicate data flow. Briefly, the figure shows how a full network is generated, then filtered to a partially observable network. The IM algorithm then iteratively selects one seed node and observes the generated spread on the visible portion of the network ("feedback") until it has chosen M seeds.

A. Experimental Setup

We use the NetHept[1] dataset, a network of 15k nodes and 31k edges (representing citations within the high energy physics theory community). We choose this due to the following reasons: (i) It constitutes a real-world dataset of reasonable size; (ii) It has been widely used to benchmark influence maximisation algorithms [21].

Throughout our experiments, we will generate partially observable networks with varying numbers of *fully observable* nodes (i.e., nodes that are observable and whose neighbours are also observable — see Section III). We do this by initialising the set of fully observable nodes (O) to contain a number of seed nodes, chosen uniformly at random. In our experiment, we initially choose 4 seed nodes, but our results are similar for other settings. Then, we iteratively add one additional node to O at a time, either by picking a node from O and then adding one of its neighbours that is currently not in O to O (both uniformly at random), or, when no such neighbours exist, by adding a random node to O. This continues until O contains a target number of nodes.

Again, due to space reasons, we focus on the weighted cascade model, although we have observed broadly similar trends more generally in the independent cascade model (for various edge weight distributions). We also focus on two benchmark algorithms: *IMM* as one of the current state-of-the-art influence maximisation algorithms, and *Degree Centrality (DC)* as a fast heuristic that is widely used. We also show the performance of *Random* as a baseline (which selects nodes uniformly at random). All results are averaged over 1,000 trials.

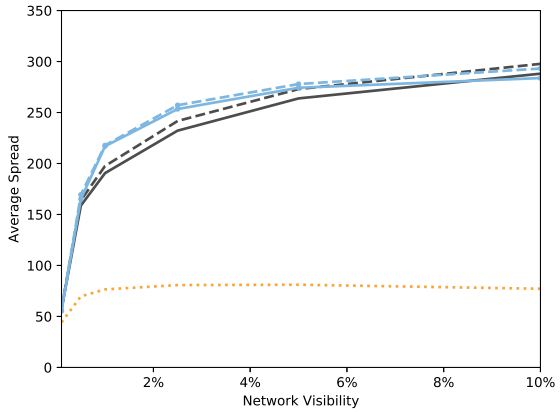
B. Results

In our experiments, we separately consider the effect of boundary node bias and adaptive influence maximisation.

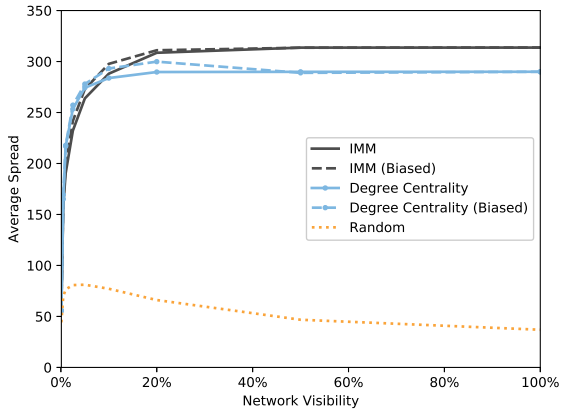
Boundary Node Bias: Figure 3 shows a setting with 10 seeds (selected non-adaptively at the same time). Here, both benchmark algorithms, standard *IMM* and *DC* are shown (solid lines), as well as their respective version with boundary node bias (we set $w = 2$ throughout this section, which works well across a wide range of settings). Several interesting trends emerge here. First, *DC* works better than *IMM* for settings with low visibility. This is likely because the immediate neighbours of a node are a better indicator of its importance rather than its influence over the very limited visible part of the network. Second, adding a bias can lead to a small (but significant) improvement in average spread (in this particular example between 1–4%). Finally, this gain is lowest in settings with either very low or high visibility. This is reasonable, because nodes are mostly connected to boundary nodes when visibility is very low, and when visibility is high, this actually causes the algorithm to favour nodes that are on the boundary of the overall network.

Adaptive Influence Maximisation: Figure 4 shows the results when we allow the algorithms to select the 10 seeds adaptively. Here, performance is significantly and consistently improved for both benchmark algorithms. This is not surprising, because the adaptive component allows the algorithms to respond to the realisation of the epidemic spread. Interestingly, the gain for *DC* is significantly higher than for *IMM* (up to 11% for *DC*

[1]This can be downloaded at https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/weic-graphdata.zip.

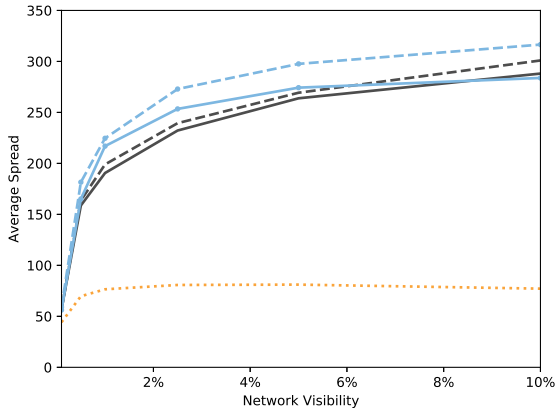


(a) Proportion of fully observable nodes up to 10%.

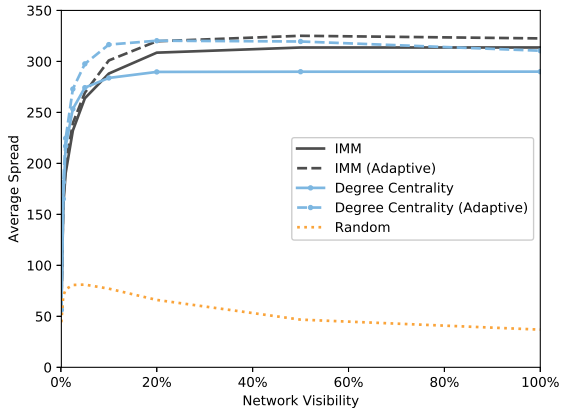


(b) Proportion of fully observable nodes up to 100%.

Figure 3: Effect of boundary node bias on average spread.



(a) Proportion of fully observable nodes up to 10%.



(b) Proportion of fully observable nodes up to 100%.

Figure 4: Effect of adaptive influence maximisation on average spread.

and up to 5% for *IMM*). This difference is likely because *DC* does not consider the overlap between neighbouring nodes (which is handled by *IMM*). During adaptive execution, this is not an issue, because activated nodes are removed from consideration. Finally, we note that the adaptive *DC* beats the state-of-the-art *IMM* by up to 18% for networks with low visibility (up to 20%) and remains competitive even when visibility is high. This makes it a promising candidate in settings where computation time is important (adaptive *IMM* took 1 minute on average to solve the fully visible network on a standard workstation, while adaptive *DC* took 181 ms — an over 300-fold improvement).

We have also experimented with combining boundary node bias and adaptive influence maximisation. Due to space reasons, we do not show these results here, but we note briefly that the bias had a slightly negative effect on *DC*. This could be because this centrality measure is already a good heuristic in partially visible network, especially when the issue with overlapping neighbours is resolved through adaptive execution. Adaptive *IMM*, on the other hand, benefits slightly from the bias.

VII. Conclusions and Future Work

In this paper, we proposed two techniques for influence maximisation in partially observable networks: introducing a boundary node bias and adaptively executing the algorithm. Both show promise, but the adaptive approach performed particularly well, improving performance by up to 18% in some settings. We also showed that a simple degree-based heuristic outperforms a state-of-the-art algorithm in settings where the network is only partially visible. Combining this heuristic with an

adaptive execution makes it competitive even when large parts of the network are visible. As the true size of a network may often be unknown in realistic settings, this approach may thus be a robust choice in such settings.

In future work, we plan to test our approaches on more varied real-world data with partial observability, both from existing datasets (e.g., Slack or Twitter) and from real deployments. The latter will be conducted in situations where information needs to spread across sub-organisations. Possibilities include advertising academic talks to university departments, and dissemination of information about a food-waste reduction programme across teams within an industrial research organisation.

Acknowledgments

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorised to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon. The work of S. M. was supported by a post-doctoral fellowship from the German Research Foundation (DFG) under Grant Number MA 7111/1-1.

References

[1] O. Okolloh, "Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information," *Participatory learning and action*, vol. 59, no. 1, pp. 65–70, 2009.

[2] A. Zenonos, S. Stein, and N. R. Jennings, "Coordinating measurements for air pollution monitoring in participatory sensing settings," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 2015, pp. 493–501.

[3] L. Von Ahn, "Human computation," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008, pp. 1–2.

[4] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 57–66.

[5] P. D. M. Richardson, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the Eighth Intl. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 61–70.

[6] D. Kempe, J. M. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *KDD*, 2003, pp. 137–146.

[7] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2014, pp. 946–957.

[8] E. Mossel and S. Roch, "On the submodularity of influence in social networks," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 128–134.

[9] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 88–97.

[10] C. Wang, W. Chen, and Y. Wang, "Scalable influence maximization for independent cascade model in large-scale social networks," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, p. 545, 2012.

[11] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *2014 ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 75–86.

[12] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1539–1554.

[13] A. Arora, S. Galhotra, and S. Ranu, "Debunking the myths of influence maximization: An in-depth benchmarking study," in *2017 ACM International Conference on Management of Data (SIGMOD)*. ACM, 2017.

[14] S. Lei, S. Maniu, L. Mo, R. Cheng, and P. Senellart, "Online influence maximization," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. ACM, 2015, pp. 645–654.

[15] S. Mihara, S. Tsugawa, and H. Ohsaki, "Influence maximization problem for unknown social networks," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '15. ACM, 2015, pp. 1539–1546.

[16] W. Chen, T. Lin, Z. Tan, M. Zhao, and X. Zhou, "Robust influence maximization," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 795–804.

[17] B. Wilder, A. Yadav, N. Immorlica, E. Rice, and M. Tambe, "Uncharted but not uninfluenced: Influence maximization with an uncertain network," in *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, 2017.

[18] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[19] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[20] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[21] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. ACM, 2014, pp. 75–86.