On the Failure of Centrality Measures in Influence Maximization

Xiaolong CHEN
The Hong Kong University of Science and Technology (Guangzhou)
Nansha, Guangzhou, China
xchen738@connect.hkust-gz.edu.cn

ABSTRACT

The influence maximization (IM) problem, aiming to select a set of k nodes in a social network to maximize the expected number of influenced nodes, is an algorithmic problem in social influence analysis, which can be used in various applications such as viral marketing. Centrality measures are widely discussed in Network Science, which are also considered related to the social influence of nodes. However, it is a problem how their performance is, compared to other methods. If they fail to do a good job in Influence Maximization, what structure features of the network has an effect on such failure and how they affect it? Such questions are waiting for answers. In order to respond to these questions, this report implements these methods on some real world datasets and observe the results. We find that centrality ranking-based methods do fail to give a good results most of the time and the network density has positive correlation with the performance of centrality ranking-based methods. The code is publicly available in the github repository: https://github.com/chenxlong3/iota5001_proj.

CCS CONCEPTS

• **Mathematics of computing** → **Graph algorithms**; • **Information systems** → **Social networks**.

KEYWORDS

Social Networks, Influence Spread, Centrality

ACM Reference Format:
Xiaolong CHEN. 2022. On the Failure of Centrality Measures in Influence Maximization. In *Proceedings of ACM Conference (Conference'22).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The prevalence of online social networks (OSNs) during the last decades has prompted much attention on information diffusion, which is powerfull in many applications [10], such as political campaign and viral marketing [6], etc. A key problem related to information diffusion study is the influence maximization (IM) problem, which aims to select k users in an OSN with the maximum influence spread, i.e., the expected number of influenced users after the propagation process. Kempe et al. [11] formulated IM problem as a combinatorial optimization problem and showed that it is NP-hard. After that, a lot of approximation algorithms have been proposed [9, 12, 17, 18]. Such algorithms come with theoretical guarantees. On the other hand, researchers have put forward some centrality measures that are related to the social influences. A heuristic idea is to simply select top-k nodes using centrality ranking. Such methods are not theoretically guaranteed. So a bunch of questions appear, from which we formulate three research questions (RQs). They are listed below.

- RQ1: How do common centrality measures perform in Influence Maximization problem compared to other state-of-the-art methods?
- RQ2: If these centrality measures fail to return a good result, what are the reasons accounting for the failure of using centrality ranking for Influence Maximization?
- RQ3: What structure features of the graph affect the performance of those centrality ranking-based methods? measures?

The algorithms used in this project are Monte Carlo-greedy (MC-greedy)[11], CELF [12], Reverse Influence Sampling (RIS) [3]. Three centrality measures used for ranking are Degree centrality, Betweenness centrality [7] and PageRank centrality [15]. The structure features that are investigated are density, diameter, average length of shortest paths and clustering coefficient.

The main contributions of this work is summarized as follows. First, this project runs several IM algorithms on six real-world datasets and shows their effectiveness and efficiency. Second, we propose a metric to measure the performance of the centrality ranking-based methods. Last but not least, we investigate the relation between their performance and four fundamental network features.

2 PRELIMINARIES

To simplify the problem, we will give the definition under Independent Cascade (IC) model.

2.1 Problem Definition of Influence Maximization

Use G to denote a social network with a vertex (node) set V and an edge set E. Each edge is associated with a probability $p(e) \in [0, 1]$. The influence propagation process can be described as follows.

1. Nodes in the selected seed set S will be activated in timestamp 1 and other nodes are considered inactive.
2. If a node u is activated for the first time in timestamp i, then for every directed edge e from u to an inactive node v, u has $p(e)$ probability to activate v at timestamp $i + 1$. u cannot activate any node after timestamp $i + 1$.
3. Once a node is activated, it remains active from then on.

Conference'22, December 2022,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Given the seed set S, we use $I(S)$ to denote the number of nodes that are active in the end of the above process, which is called the spread of S. Given G and a constant $k \leq n$, the influence maximization problem under the IC model asks for a seed set of size k with the maximum expected spread $\mathbb{E}[I(S)]$. That is, we want to find k nodes that can influence the largest number of nodes in expectation.

2.2 Notations

The frequently used notations are listed in Table 1.

Notation	Description
G	Network
V	The set of nodes
E	The set of edges
n	Number of nodes, $n = \|V\|$
m	Number of edges, $m = \|E\|$
k	Size of seed set
S	Seed set
$I(S)$	The number of influenced nodes from S
$\sigma(S)$	The estimator of $\mathbb{E}[I(S)]$
$\Delta(u\|S)$	Influence increment, $\sigma(S \cup \{u\}) - \sigma(S)$
ε	Error term
N_u	Neighbors of node u
d_u	Degree of node u
σ'_{st}	Number of shortest paths from s to t
$\sigma'_{st}(u)$	Number of shortest paths through u from s to t
δ	Deviation of the solution returned by centrality-based methods

Table 1: Notations

3 RELATED WORK

3.1 Influence Maximization

In 2003, Kempe et al. [11] firstly presented the algorithmic study on the influence maximization problem. They showed that this problem is NP-hard in general and proposed to use a greedy algorithm to approximate the solution with a factor of $1 - 1/e - \varepsilon$. The key idea is to use Monte Carlo simulation [14] to estimate the expected influence spread (i.e. $\mathbb{E}[I(S)]$). With this approach, Kempe's algorithm greedily selects the node in G that gives the largest increment of influence, until k nodes are chosen. However, the time complexity is too high for large graphs. There are a lot of work [2, 4, 5, 9, 12] trying to improve the efficiency of this algorithm. Among these techniques, the most famous ones are CELF [12] and CELF++ [9], which make use of the monotonicity and submodularity of the influence function. Although they have the same worst time complexity as simple greedy algorithm, it turns out that they can provide higher empirical efficiency.

In 2014, Borgs et al. [3] made a great breakthrough by presenting a near-linear time algorithm under the IC model. They proposed the idea of reverse influence sampling (RIS), which significantly improve the time complexity for giving a $(1-1/e-\varepsilon)$-approximate solution with high probability. After that, Tang et al. [18] pointed out the deficiency of Borgs's algorithm and proposed Tow-phase Influence Maximization algorithm (i.e., TIM), which makes it more efficient in practice. In 2015, they took a martingale approach [17] and further improve the algorithm (i.e., IMM).

3.2 Centrality Measures

Bavelas [1] first defined measure of centrality for connected graphs. After that, a lot of centrality measures have been proposed to indicate the importance of nodes. Freeman et al. developed three types of measures of centrality [7, 8], which are degree, closeness and betweenness centrality. In 1999, Page et al. [15] proposed PageRank, which brings order to web pages. This report will focus on Degree centrality, Betweenness centrality and PageRank centrality.

4 OVERVIEW OF ALGORITHMS

4.1 Sketch-based Algorithms

The sketch-based algorithms basically sample a sketch from the network to evaluate influence spread. The algorithms that will be introduced in this part are MC-greedy and RIS.

The basic greedy framework is given by Algorithm 1. For MC-greedy, the influence estimator uses Monte-Carlo simulation to measure the influence spread of a set of nodes. To further explain, the idea is to run the propagation process (described in section 2.1) for a large number of times and use the mean value of activated nodes to estimate $\mathbb{E}[I(S)]$. That is to say, the influence estimator can be expressed in the form of

$$\sigma(S) = \frac{1}{r}\sum_{i=1}^{r}(\#\text{ of activated nodes in } i\text{-th simulation}). \quad (1)$$

It has been proved that MC-greedy can return a $(1 - 1/e - \varepsilon)$-approximate solution if the number of MC simulations r is set to $\Theta(\varepsilon^{-2}k^2 n \log(n^2 k))$.

Algorithm 1 Greedy

Require: k: A number, σ: Influence estimator
1: $S \leftarrow \emptyset$
2: **for** $i = 1 \ldots, k$ **do**
3: $\quad u^* \leftarrow \arg\max_{u \in V \setminus S}(\sigma(S \cup \{u\}) - \sigma(S))$
4: $\quad S \leftarrow S \cup \{u^*\}$
5: **end for**
6: **return** S

Following the MC-greedy algorithm, several methods have been proposed to reduce the number of MC simulations. CELF [12] exploits the submodularity of the influence function. Using $\Delta(u|S_i)$ to denote $\sigma(S_i \cup \{u\}) - \sigma(S_i)$, the submodularity of the influence function indicates that $\Delta(u|S_j) \leq \Delta(u|S_{j-1})$, which means the marginal gain of influence spread will decrease with the increasing number of seed set size. This property allow us to derive an upper bound for $\Delta(u|S_j)$, which makes it easy for us to stop early so that we can prune the search space. To describe it more detailedly, the key idea is the following inequation.

$$\Delta(u|S_j) \geq \Delta(v|S_{j-1}) \geq \Delta(v|S_j) \quad (2)$$

This chain of inequations show that if the influence increment of u at timestamp j is larger than that of v at timestamp $j-1$, then

we do not need to compute $\Delta(v|S_j)$ since we know that it certainly will not be chosen at the current timestamp. Although CELF has the same worst-case time complexity as MC-greedy, it significantly improves the practical efficiency.

In 2014, Borgs et al. [3] proposed the idea of reverse sampling for the IM problem, which is called Reverse Influence Sampling (RIS) in other following works. To illustrate their method, two important concepts will be introduced first:

Definition 4.1 (Reverse Reachable Set). *Let v be a node in G, and g be the graph obtained by removing each edge in G with probability $1 - p(e)$. The reverse reachable set (RR set) for v is the set of nodes in g that can reach v.*

Definition 4.2 (Random RR Set). *Suppose $\mathcal{G}$ is the distribution of g induced by the randomness in edge removals from G. A random RR set is defined as an RR set generated by a random sample from $\mathcal{G}$, for a node selected uniformly at random from g.*

The key observation of Borgs et al. is that if an RR set generated for v has a probability of ρ to overlap with a node set S, then if we use S as the seed set to start influence propagation, we have a probability of ρ to activate v. With this observation, RIS algorithm runs in the following steps:

1. Generate a certain number of random RR sets from G.
2. Formulate the problem as a maximum coverage problem, which aims to select k nodes to cover the maximum number of RR sets.
3. Use the greedy algorithm to return a $(1 - 1/e)$-approximate solution S_k^*.

4.2 Centrality Ranking-based Algorithms

In this report, we focus on three popular centrality measures, degree centrality, betweenness centrality [7] and pagerank centrality [15]. The computation of these measures are given in equation 3, 4, 5.

$$D_C(u) = deg(u) \tag{3}$$

$$B_C(u) = \sum_{s,t \in V, s \neq t \neq u} \frac{\sigma'_{st}(u)}{\sigma'_{st}} \tag{4}$$

$$PR(u) = \sum_{v \in N_u^{(in)}} \frac{PR(v)}{d_v^{(out)}} \tag{5}$$

Using these metrics to select the seeds is also called simple ranking proxy methods [13]. To use them in the IM problem, we can simply compute these centrality measures for all nodes and select top-k nodes as the seed set. Such methods, however, are not theoretically guaranteed.

5 METHODOLOGY

This report will use three kinds of centrality measures to compare with greedy algorithm and RIS algorithm, which are outdegree centrality, betweenness centrality and pagerank centrality. Since it needs factorial time complexity to compute the exact solution to the IM problem, we simply use the highest spread of all five methods as reference and use the difference between the highest influence spread and the spread by centrality-based methods to indicate the deviation of the solution returned by centrality-based methods, which is denoted by δ. Equation 6 gives the formulation of δ, where S_{best} denotes the seed set with the highest influence spread in the corresponding experiment and S_{cen} denotes the seed set returned by the centrality ranking method (Degree, Betweenness or PageRank). The network features that we investigate are listed in Table 2.

$$\delta = \frac{\sigma(S_{best}) - \sigma(S_{cen})}{\sigma(S_{best})} \tag{6}$$

Notations	Description
Density	Number of edges / Number of all possible edges
Diameter	The maximum distance of two nodes in the largest connected component
Avg. dist.	The average length of all shortest paths
Cluster. Coef.	The clustering coefficient of the network

Table 2: Investigated features

First we run the aforementioned methods in every dataset and document the influence spread and time cost. Then for each dataset, we compute δ with the seed set size being 5. Draw the scatter plots to investigate the correlation between δ and the structured features for every centrality measures.

6 EXPERIMENT

6.1 Settings

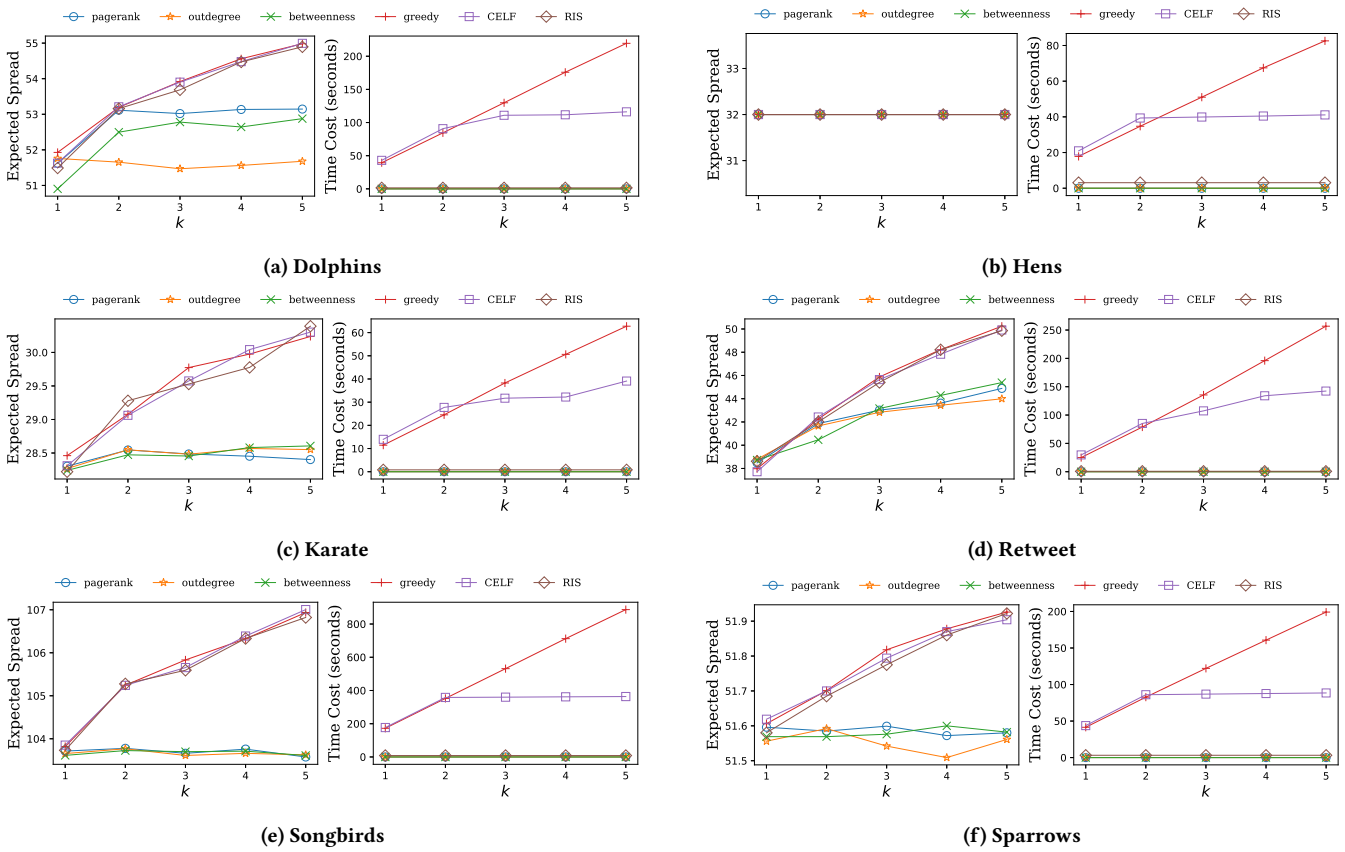
Due to the high time complexity of greedy algorithm, we only carry out the experiments on six small real world datasets (with at most 110 nodes), which are downloaded from https://networkrepository.com [16]. Since the propagation is directed, we will change all undirected networks to directed networks. The number of Monte Carlo simulation times for measuring $I(S)$ is set as 1000. The maximum k is set as 5. The probability attached with every edge is fixed as 0.5. The information of datasets is given by Table 3. The experiments are run on a linux machine with an Intel(R) Xeon(R) CPU @ 2.60GHz and 80GB memory. The programming language used here is Python.

Name	n	m
Dolphins	62	159
Hens	32	496
Karate	34	78
Retweet	96	117
Songbirds	110	1027
Sparrows	52	454

Table 3: Datasets

6.2 Results

The influence spread of the seed set returned by the aforementioned methods and the time cost of them are shown in Figure. 1. From the results we can see that although it takes far less time to use centrality measures to select the seed set, such methods may fail



(a) Dolphins (b) Hens

(c) Karate (d) Retweet

(e) Songbirds (f) Sparrows

Figure 1: Expected influence spread and time cost of each method for the six datasets that we look into.

most of the time. In the experiment, apparently RIS is the method that performs well in both efficiency and effectiveness.

Given the experiment results, we are able to answer the research questions raised before.

6.2.1 RQ1. How do common centrality measures perform in IM problem? The centrality ranking-based methods are not stable. Sometimes they can return the seed set with influence spread close to other methods (e.g. Hens and Sparrows). But overall, the sketch-based algorithms consistently outperform the centrality ranking-based methods.

6.2.2 RQ2. What are the reasons accounting for the failure. By thinking about how the diffusion process goes on and how we compute the centrality measures, we can find that none of the ranking methods accounts for the influence overlaps between different seeds. This is because they simply consider the influence spread as a linear combination of the influence spread of every individual node in the seed set. Thus, such methods will overestimate the influence spread.

6.2.3 RQ3. What kinds of structure features affect the performance of those centrality measures? From Figure 2 we can see that the relation between the centrality ranking-based methods and the structure features presents similar patterns for all three centrality measures we investigate here. Qualitatively speaking, the difference δ has negative correlation with graph density. But the relations of δ and other features are not obvious, due to the small number of data points.

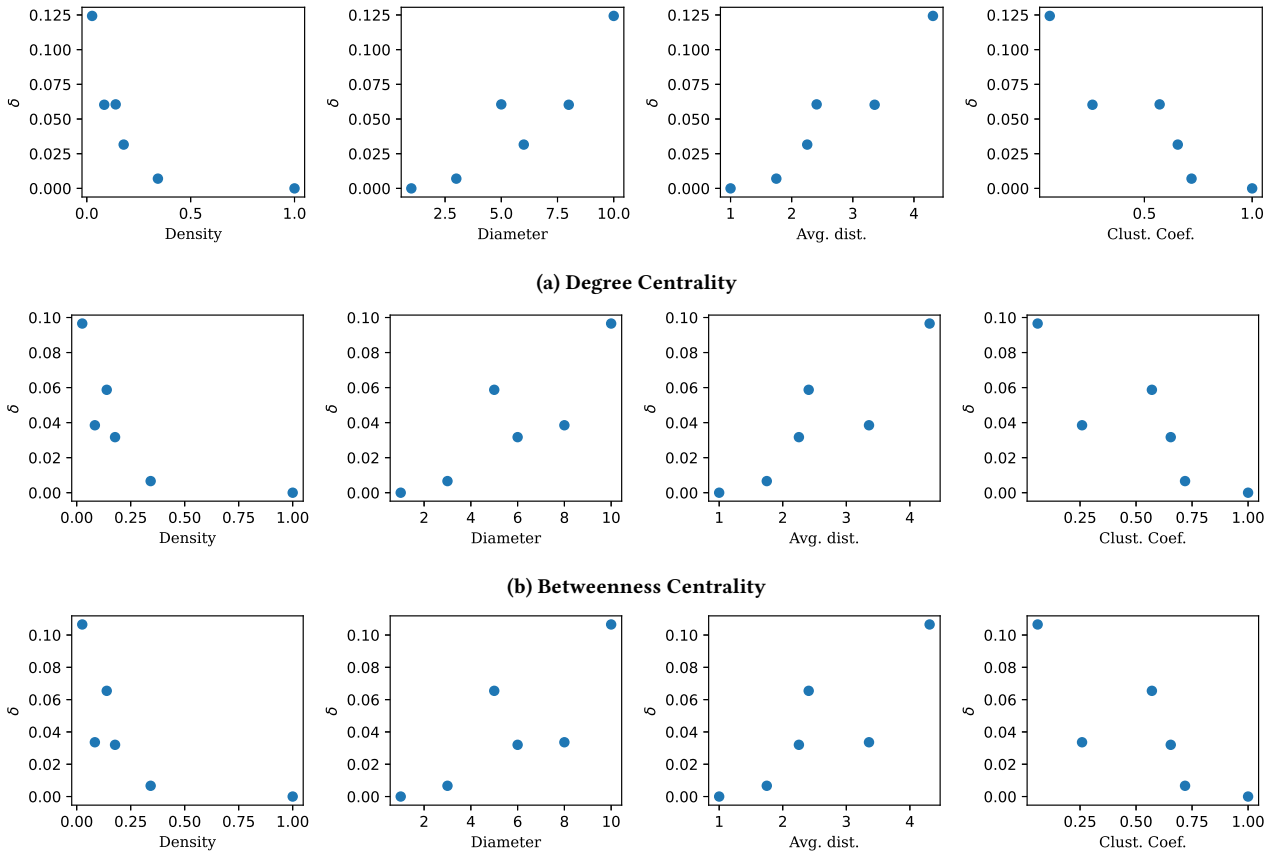
7 LIMITATIONS

A number of limitations of this work are listed here, which will be considered in future study.

- Due to the limitation of the computation source, this project only uses small datasets.
- Only three network features are considered here. Hopefully if we consider more features, some interesting pattern can appear.
- The number of datasets is not enough for us to carry out meaningful quantitative analysis. We can consider using man-made random graphs in further study.

8 CONCLUSION

This work investigates the effectiveness and efficiency of using centrality measures to select the seed set for Influence Maximization problem. It turns out that centrality ranking-based methods fail to give us stable and good result most of the time. To find out whether such failure is relevant to the structure features of the network,



(a) Degree Centrality

(b) Betweenness Centrality

(c) PageRank Centrality

Figure 2: The relation between δ and the investigated network features.

several experiments have been conducted. The result shows that higher density can lead to better performance of centrality ranking-based methods. However, the effects of other structure features are not so clear due to the limited number of data points.

REFERENCES

[1] Alex Bavelas. 1950. Communication patterns in task-oriented groups. *The journal of the acoustical society of America* 22, 6 (1950), 725–730.

[2] Shishir Bharathi, David Kempe, and Mahyar Salek. 2007. Competitive influence maximization in social networks. In *International workshop on web and internet economics*. Springer, 306–311.

[3] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. 2014. Maximizing social influence in nearly optimal time. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 946–957.

[4] Wei Chen, Chi Wang, and Yajun Wang. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1029–1038.

[5] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 199–208.

[6] Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 57–66.

[7] Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry* (1977), 35–41.

[8] Linton C Freeman. 1978. Centrality in social networks conceptual clarification. *Social networks* 1, 3 (1978), 215–239.

[9] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Celf++ optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*. 47–48.

[10] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record* 42, 2 (2013), 17–28.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[12] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 420–429.

[13] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. 2018. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1852–1872.

[14] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[16] Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*. https://networkrepository.com

[17] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 1539–1554.

[18] Youze Tang, Xiaokui Xiao, and Yanchen Shi. 2014. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 75–86.