Short Paper

An Experimental Study on Network Immunization

Alvis Logins
Aarhus University

Panagiotis Karras
Aarhus University

ABSTRACT

Given a network in which a undesirable rumor, disease, or contamination spreads, which set of network nodes should we *block* so as to contain that spread? Past research has proposed several methods to address this *network immunization* (NI) problem, which is to find a set of k nodes, such that the undesirable dissemination is minimized in expectation when they are blocked. As the problem is NP-hard, some algorithms utilize solely features of the network structure in a *preemptive* manner, to others that take into account the specific source of a contamination in a *data-aware* fashion. This paper presents an experimental study on NI algorithms and baselines under the independent cascade (IC) diffusion model. We employ a variety of synthetic and real-world networks with diverse graph density, degree distribution, and clustering coefficients, under realistically calculated influence probabilities. We conclude that data-aware approaches based on the construct of *dominator trees* usually perform best; however, in networks with a power-law degree distribution, preemptive approaches utilizing spectral network properties shine out by virtue of their efficiency in identifying central nodes.

1 INTRODUCTION

Real-world networks facilitate the spread of ideas, behaviors, inclinations, or diseases via *diffusion* processes [10]. Oftentimes a diffusion of malicious nature needs to be contained via countermeasures [11]. One such countermeasure is the *blocking* of a subset of network nodes. *Network Immunization* (NI) calls to find an optimal set of nodes to block so as to arrest a diffusion.

Early works on NI were motivated by epidemiology [3, 12], categorizing individuals as Susceptible $\mathbb{S}$, Infected $\mathbb{I}$, or Recovered $\mathbb{R}$. Those who are infected infect their susceptible neighbors with a transition rate β, and become recovered (hence immune) with transition rate γ. In the context of social networks [14], the *Independent Cascade* (IC) model [4] generalizes the SIR model, assigning an independent transition rate β to each edge. Kempe et al. [6] formulated the Influence Maximization (IM) problem under the IC model, where the goal is to select k seed nodes that maximize the expected diffusion spread; since then, the problem has been studied extensively [10, 14].

The NI problem is complementary to the IM problem. Certain notions are useful in both. For example, eigenvalue centrality [12] has been used to guide seed selection in IM. Similarly, Chen et al. [2] employ the first eigenvalue λ as a proxy to the objective of NI problem, scoring nodes by the eigen-drop $\Delta\lambda$ that their removal causes, leading to a succession of techniques aiming to to maximize the eigen-drop of immunized nodes [20].

We distinguish two variants on network immunization: *preemptive* immunization finds a solution before the epidemic starts; by contrast, *data-aware* immunization tailors the solution to a particular diffusion seed [19]. The state-of-the-art data-aware solution, *Data-Aware Vaccination Algorithm* (DAVA) [21] employs structures called *dominator trees*. Still, the experimental study in [21] is limited to four datasets with synthetic propagation probabilities; it is not clear how the topology of the network influences the algorithm's performance. At the same time, recent preemtive immunization methods [11, 13] significantly outperform the baselines used in [21], yet have not been compared to DAVA itself. Thus, to the best of our knowledge, no previous work has studied how data-aware and and preemptive immunization strategies fare under different graph topologies.

In this paper, we investigate the performance of state-of-the-art data-aware and preemptive NI solutions on a variety of real-world and synthetic network structures with diverse characteristics, and under realistic influence probabilities with the IC model. Our study features the *first*, to our knowledge, application of the most recent algorithm for eigen-drop maximization and a generic spectral method of activity shaping, to NI under the IC model. We demonstrate that data-aware approaches are leading in a majority of configurations, yet preemtive ones stand out under particular settings of graph density, influence probabilities, degree distribution, and clustering coefficients.

2 BACKGROUND

The classic approach to preemptive NI is the NetShield algorithm [2]. NetShield greedily selects a set of nodes S, aiming to maximize its *Shield value*:

$$Sv(S) = \sum_{i \in S} 2\lambda \mathbf{u}(i)^2 - \sum_{i,j \in S} \mathbf{A}(i,j)\mathbf{u}(i)\mathbf{u}(j)$$

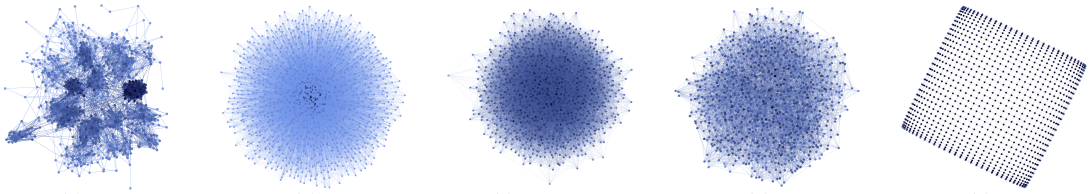
where λ and $\mathbf{u}$ are the largest eigenvalue and the corresponding eigenvector of the network's adjacency matrix $\mathbf{A}$. A set S has high Sv if its elements have high eigenscore $\mathbf{u}(i)$ and are not connected to each other (zero $\mathbf{A}(i,j)$). A high eigenscore implies that their removal leads to a significant eigen-drop $\Delta\lambda$. The algorithm has a $O(n|S|^2)$ complexity, where n is the size of a network.

NetShield defines an *epidemic threshold* β' such that any edge transition probability $\beta > \beta'$ would result in a significant portion of the network being contaminated. The algorithm utilizes the fact that the *epidemic threshold* is related to the first eigenvalue of the network adjacency matrix as $\beta' = 1/\lambda$ [16]. Thus, λ expresses the *vulnerability* of the network to an epidemic. Tariq et al. [13] improved upon NetShield by approximating the eigen-drop, relying on the fact that λ can be expressed as the limit trace of the p-exponential adjacency matrix $\mathbf{A}$, which equals the number of p-sized closed walks in the graph, cw_p:

$$\lim_{p \to \inf, p \text{ even}} (trace(\mathbf{A}^p))^{1/p} = \lambda$$

$$trace(\mathbf{A}^p) = cw_p(G)$$

The proposed method greedily selects a set of nodes to block based on their contribution to closed walks, hence to network vulnerability, approximating cw_p by a submodular score function, calculated by partitioning vertices into α equal-size groups by means of a set of hash functions $i = \{1..\gamma\}$. In our experiments, we use $\alpha = 200$ and $i \in \{1..3\}$. The published version suggested using $p = 6$, yet in communication with the authors we confirmed that $p = 8$ feasibly leads to improved results; we refer to this algorithm as *Walk8*; its complexity is $O(n^2 + \gamma(n + \alpha^3) + nk^2)$.



(a) GRP (b) BA (c) Binomial (d) WS (e) Grid

Figure 1: Generated Graph Examples. Darker color indicates higher degree, normalized per graph.

Graph Type	$\|V\|$ [$\cdot 10^3$]	$\|E\|$ [$\cdot 10^3$]	$deg_{\text{min/avg/med/max}}$	clust. coeff. cl	infl. prob. W	seed fract. sf	k fract. kf	Other Parameters
Binomial	1.0	14.7	4/15/15/30	0.012	0.2	0.05	0.05	edge exist. $p = 0.015$
GRP	1.0	14.6	2/30/26/90	0.325	0.1	0.01	0.05	shape param. $s = 20$, $v = 0.9$, intra-group prob. $p_{in} = 0.4$, inter-group $p_{out} = 0.001$
WS	1.0	7.0	18/28/28/44	0.237	0.2	0.05	0.05	neighbors in a ring $l = 15$, rewiring prob. $p = 0.3$
BA	1.0	29.4	28/59/40/498	0.103	0.1	0.05	0.05	prob. of triangle $p = 0.2$, density $m = 15$
Grid	1.0	39.7	4/8/8/8	0.000	0.7	0.05	0.05	–
Stanford	9.9	36.9	0/7/5/555	0.392	0.2	0.2	0.2	
Gnutella	62.6	147.9	1/4/2/95	0.007	0.2	0.2	0.2	–
VK	2.8	40.9	1/29/14/288	0.235	–	0.2	0.2	

Table 1: Default parameters for graph types

DAVA [21] accepts the seed set of a network diffusion as input and builds its NI solution around *dominator trees*. A node u *dominates* a node w w.r.t. a seed node s if all paths from s to w pass through u. A *dominator tree* is a tree where each node is dominated by its ancestors. The *benefit* of removing a node is calculated as $\gamma(v) = 1 + \sum_{u \in \text{children of } v} \gamma(u) \cdot p_{vu}$, where p_{vu} stands for the probability that influence propagates along *any* path, approximated via the *most probable* path, from v to u.

DAVA iteratively removes the node of highest benefit and reconstructs the tree. In a DAVA variant, DAVA-fast, the tree is built only once and top-k nodes are selected based on their benefit in one go. A dominator tree is built in $O(E \log N)$ [8].

NetShape [11] immunizes a network via a convex relaxation approach, maximizing the eigen-drop of the network's integrated and symmetrized *Hazard matrix*, a matrix of a continuous integrable transition rate functions $\{\beta(u, v, t)\}_{u,v \in V}$, which indicate the probability that v is influenced by u at time t after u gets infected. The first eigenvalue bounds the expected spread of an infection. We apply NetShape as a heuristic for the IC model, setting the integral of the transition rate $\beta(u, v, t)$ as equal to the influence probability between u and v, and minimize the first eigenvalue by the projected subgradient descent method in the space of possible Hazard matrices *after* immunization, while setting the effect of immunizing u on the integrated hazard matrix element as 0, if u is a seed. The complexity is $O(\frac{1}{\epsilon^2} p_{max}^2 E \ln E)$, where p_{max} is the maximum propagation probability, and ϵ is a parameter affecting a step of subgradient descent.

3 METHODOLOGY

Consider a directed graph $G = (V, E)$ with set of nodes V and set of edges E. Each edge is associated with a probability of propagation. By the *independent cascade* model, a diffusion occurs in discrete time steps. In step t_0, a *seed set* $S \subset V$ becomes *activated*. Any node v activated in step t_i attempts to activate each of its inactive neighbors in step t_{i+1}, and succeeds by the probability associated with the edge from v to that neighbor. The process terminates when there are no more newly activated nodes. The *Network Immunization* (NI) problem calls to *block* a select set of k nodes $R \subseteq V \setminus S$ so as to minimize the expected *spread* of activated nodes, by a given seed set S in a graph G.

3.1 Algorithms

We compare six solutions to the NI problem in three categories:

- **Naïve**: *Degree* selects the top-k nodes with highest degree; *Random* selects k nodes uniformly at random.
- **Preemptive**: *NetShield* [2] and *Walk8* [13],
- **Data-Aware**: *NetShape* [11] and *DAVA* [21].

On NetShape, we use the default $\epsilon = 0.2$. As exact spread computation is #P-hard, we estimate spread with any solution via 1000 Monte-Carlo IC simulations. We use the original Matlab code of Walk8. As seeds cannot be blocked, we fetch $k + |S|$ nodes to be blocked with Walk8, ensuring that at least k nodes are blocked. We implemented all other algorithms in Python[1].

3.2 Data

We use both synthetic and real data obtained as follows.

3.2.1 Synthetic Data. We generated graphs of different properties using five models. By the *Erdős-Rényi model*, each edge is present with probability p; generated graphs have a low clustering coefficient and a binomial degree distribution. We refer to this generator as **Binomial**. We render the graph directed by selecting a random direction for each edge with 50% probability.

A **Gaussian Random Partition** (GRP) [1] selects edges as with Erdős-Rényi, but with a prior grouping, where group size follows a Gaussian distribution; it uses a probability value p_{in} for edges across nodes in the same group, and p_{out} otherwise, hence varying intra-group and inter-group density.

Watts Strogatz (WS) networks model self-organizing small-world systems [17], which have small average shortest path length, and are highly clustered, hence susceptible to infectious spread. The generator employs two parameters: l indicates how many nearest neighbors each node is joined with in a ring; p is a probability of edge rewiring, which induces disorder.

Barabási-Albert (BA) networks have both high clustering coefficients (as GRP and WS graphs) and power-law degree distribution, hence are better imitations of real-world social networks. We use the algorithm of Holme and Kim [5], which extends the original Barabási-Albert model, yet use the BA label as its basis; this algorithm randomly creates m edges for each node in a graph, and for created edge with a probability p adds an edge to one of its neighbors, thus creating a triangle.

Grid graphs have each node connected to four neighbors on a lattice. With this graph type, we explore the applicability of solutions on spatial graphs such as geosocial contact networks [22].

[1] Available at https://github.com/allogn/Network-Immunization

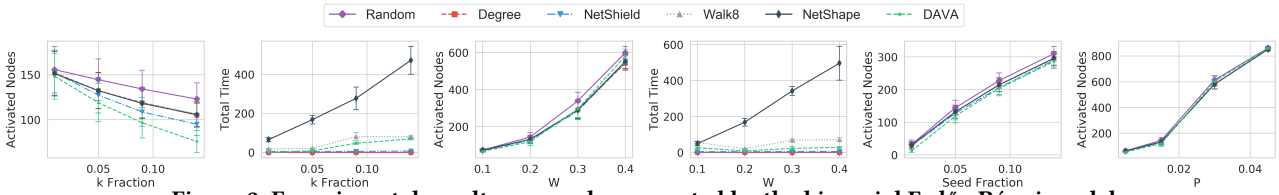


Figure 2: Experimental results on graphs generated by the binomial Erdős-Rényi model

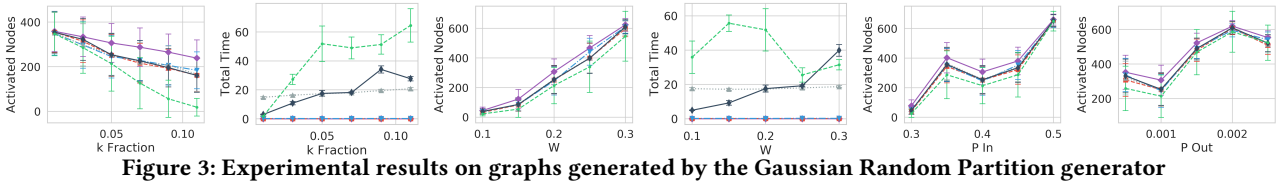


Figure 3: Experimental results on graphs generated by the Gaussian Random Partition generator

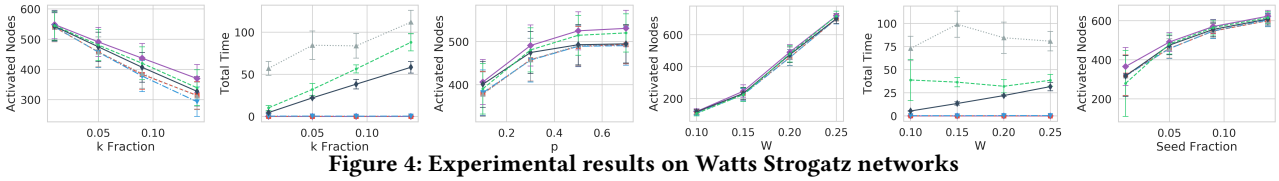


Figure 4: Experimental results on Watts Strogatz networks

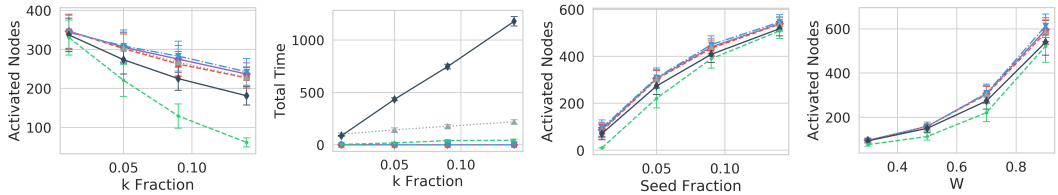


Figure 5: Experimental results on regular grids

Table 1 lists the default parameters for all models, where fractions $sf = |S|/|V|$ and $kf = k/|V|$. Figure 1 shows example graphs. All synthetic graphs have 1000 nodes, as in [16].

3.2.2 Real-World Datasets. We use 3 real-world graphs. Stanford and Gnutella, have been employed in related literature; a third, VK, provides a case of real-world propagation probabilities.

The **Stanford** data consists of pages and hyperlinks in the Stanford University website[2] [21]. The **Gnutella** peer-to-peer file sharing directed network is part of the SNAP dataset [9]. We use the biggest snapshot of 62586 nodes, with a diameter of 11 nodes and a clustering coeficient of 0.0055. It has been used in [11, 19, 21]. **vKontakte**[3] (VK) is a Russia-based social network of more than 500 million users[4]. Its public API allows to download information about public profiles, subscriptions, and posts. We fetch public posts of users to train the IC model.

3.3 Parameters

We consider **blocked node set size** k as a fraction of network size [15]. We employ *random* **seed selection** [3, 18, 21]; we pick 10 random seed sets, and show the mean and standard deviation of activated nodes. We choose **influence probabilities** uniformly at random from 0 to a maximum value W. We learn influence probabilities on the VK data using user posts as actions. We download 100 latest posts at the moment of publishing per user, resulting in 21M posts. Most posts are short, hence we can apply the same Natural Language Processing methods as for short messages. After preprocessing, we collected 536,073 non-empty messages belonging to non-isolated nodes in the VK graph, with median length of 11 words, std 187 and max 2977, leaving us with 3% of the original dataset. We define the closeness of actions by comparing the content of text messages, as in [7], to learn vector embeddings of short messages. We define term proximity as $p(w_2|w_1) = \frac{1}{|M|}\frac{c(w_1,w_2)}{c(w_1)}$, where M is a set of all posts with non-zero text content, $c(w)_m$ is the number of messages with w, and $c(w_1, w_2)$ is the number of messages with w_1 and w_2 present together. We learn stemmed term proximities and enrich the term frequency–inverse document frequency vectors of messages by increasing the probability of any words similar to words present in the message. We consider all message pairs with similarity above the median as similar. Scanning the action log to calculate the influence probability from a node u to any nove v as the ratio of successful reposts of similar messages. Filtering zero-probability edges, we select the largest component of 2.8K nodes and 40.9K edges as our VK network.

[2] https://www.cise.ufl.edu/research/sparse/matrices/Gleich/ [3] http://vk.com/
[4] https://en.wikipedia.org/wiki/VKontakte

4 EXPERIMENTAL RESULTS

Here we present the results of our study. We set a timeout of 1h for all experiments for a single solver instance.

4.1 Synthetic Data

Figure 2 shows results with **Binomial** graphs. As the number of blocked nodes grows, DAVA's advantage of knowing the seeds becomes evident. Surprisingly, NetShield achieves better results than NetShape and Walk8 in this graph type. As the graph has a uniform structure, spectral-based algorithms do not perform well. This uniformity results in performance of algorithms not being dependent on the number of seeds and influence probability W. Still, as Figure 2c shows, with large W DAVA is slightly worse than preemptive approaches. DAVA assumes that the influence probability between two successive dominators in the dominator tree is equal to the probability along the shortest path. When there are many paths between two dominators, this assumption fails, hence the accuracy of the algorithm drops. We observe that NetShape is the least scalable algorithm.

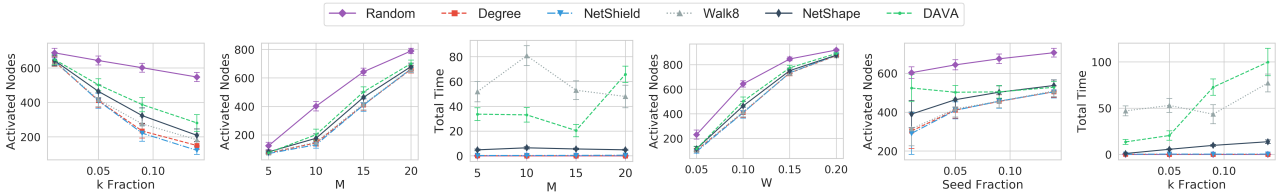


Figure 6: Experimental results on graphs generated by the Barabasi-Albert growth model

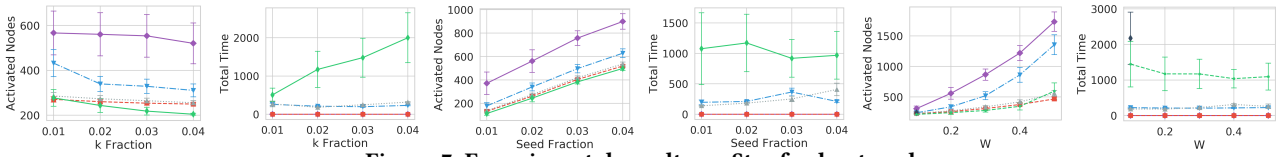


Figure 7: Experimental results on Stanford network

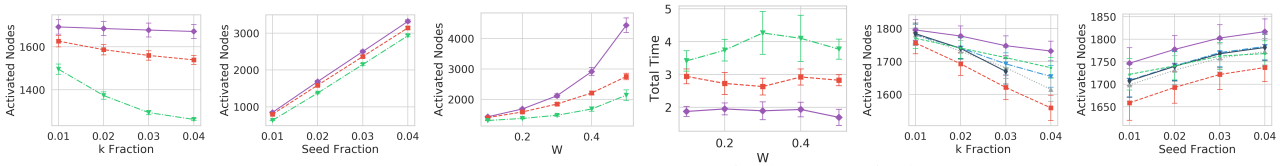


Figure 8: Experimental results on Gnutella (a-d) and VK (e-f) networks

Figure 3 shows results with **GRP** graphs. Again, the gap increases as k grows. DAVA achieves the best results on all parameters, except for the largest p_{out}. We observe that, as the inter-group probability p_{out} grows, DAVA shows slightly worse performance; in other words, as the graph forfeits its clustered structure, DAVA provides less accurate probability estimates.

Figure 4 shows results with **WS** graphs. Here, preemptive algorithms perform significantly better than DAVA, while the difference is accentuated as the number of blocked nodes k grows.

Figure 5 shows results with **Grid** graphs. All algorithms except DAVA fail to isolate seeds. NetShape outperforms other spectral approaches thanks to its data-awareness. Runtimes are similar to those in the Binomial case, with DAVA being sufficiently scalable.

Last, Figure 6 shows results with **BA** graphs; the degree heuristic and NetShield perform best. This result indicates that there are limits to the versatility of DAVA.

4.2 Real Data

Figure 7 shows results on the **Stanford** network. We employ the fast DAVA that builds a dominator tree only once so as to scale. Exploring a larger range of parameters than [21] reveals that DAVA performs similarly to the Degree heuristic, and slightly worse as W grows, due to the scale-free data topology. NetShape and Walk8 could not scale to such size. **Gnutella** has a more random topology than the Stanford network. Running on the 62K-node Gnutella snapshot, only fast-DAVA and baselines terminated within the time limit. Figure 8 shows the results, with DAVA reasserting its advantage. Our **VK** graph has high clustering coefficient and power-law degree distribution. Figure 8 shows that, on this data, DAVA is outperformed by preemptive methods. We deduce that, in real-world social networks, isolating diffusion sources is less critical than immunizing influence hubs.

5 CONCLUSIONS

We conducted an exhaustive experimental study of network immunization methods. We conclude that, while data-aware approaches stand out on networks with uniform topologies, spectral structure-based approaches are competitive on networks with power-law topologies. This result calls for further research.

REFERENCES

[1] Ulrik Brandes, Marco Gaertler, and Dorothea Wagner. 2003. Experiments on Graph Clustering Algorithms. In *ESA*. 568–579.

[2] Chen Chen, Hanghang Tong, B. Aditya Prakash, Charalampos E. Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos, and Duen Horng Chau. 2016. Node Immunization on Large Graphs: Theory and Algorithms. *IEEE Trans. Knowl. Data Eng.* 28, 1 (2016), 113–126.

[3] Wen Cui, Xiaoqing Gong, Chen Liu, Dan Xu, Xiaojiang Chen, Dingyi Fang, Shaojie Tang, Fan Wu, and Guihai Chen. 2016. Node Immunization with Time-Sensitive Restrictions. *Sensors* 16, 12 (2016), 2141.

[4] Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters* 12, 3 (2001), 211–223.

[5] Petter Holme and Beom Jun Kim. 2002. Growing scale-free networks with tunable clustering. *Physical review E* 65, 2 (2002), 026107.

[6] David Kempe, Jon M. Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. 137–146.

[7] Ricardo Lage, Peter Dolog, and Martin Leginus. 2013. Vector Space Models for the Classification of Short Messages on SN Services. *WEBIST* (2013), 209–224.

[8] Thomas Lengauer and Robert Endre Tarjan. 1979. A fast algorithm for finding dominators in a flowgraph. *ACM TPLS* 1, 1 (1979), 121–141.

[9] Jure Leskovec and Rok Sosič. 2016. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM TIST* 8, 1 (2016), 1.

[10] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. 2018. Influence Maximization on Social Graphs: A Survey. *IEEE TKDE* 30, 10 (2018), 1852–1872.

[11] Kevin Scaman, Argyris Kalogeratos, Luca Corinzia, and Nicolas Vayatis. 2017. A Spectral Method for Activity Shaping in Continuous-Time Information Cascades. *CoRR* abs/1709.05231 (2017).

[12] Paulo Shakarian, Abhinav Bhatnagar, Ashkan Aleali, Elham Shaabani, and Ruocheng Guo. 2015. The independent cascade and linear threshold models. In *Diffusion in Social Networks*. Springer, 35–48.

[13] Juvaria Tariq, Muhammad Ahmad, Imdadullah Khan, and Mudassir Shabbir. 2017. Scalable Approximation Algorithm for Network Immunization. In *PACIS*.

[14] V. Tejaswi, P. V. Bindu, and P. Santhi Thilagam. 2016. Diffusion models and approaches for influence maximization in social networks. *ICACCI* (2016).

[15] Biao Wang, Ge Chen, Luoyi Fu, Li Song, and Xinbing Wang. 2017. DRIMUX: Dynamic rumor influence minimization with user experience in social networks. *IEEE Trans. Knowl. Data Eng.* 29, 10 (2017), 2168–2181.

[16] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. 2003. Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint. In *22nd Symposium on Reliable Distributed Systems*. 25–34.

[17] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of "small-world" networks. *Nature* 393, 6684 (1998), 440.

[18] Arie Wahyu Wijayanto and Tsuyoshi Murata. 2018. Pre-emptive spectral graph protection strategies on multiplex social networks. *Applied Network Science* 3, 1 (2018), 5.

[19] Dingda Yang, Xiangwen Liao, Huawei Shen, Xueqi Cheng, and Guolong Chen. 2018. Dynamic node immunization for restraint of harmful information diffusion in social networks. *Physica A* 503 (2018), 640–649.

[20] Yao Zhang. 2017. *Optimizing and Understanding Network Structure for Diffusion*. Ph.D. Dissertation. Virginia Tech.

[21] Yao Zhang and B. Aditya Prakash. 2015. Data-Aware Vaccine Allocation Over Large Networks. *TKDD* 10, 2, 20:1–20:32.

[22] Yao Zhang, Arvind Ramanathan, Anil Vullikanti, Laura L. Pullum, and B. Aditya Prakash. 2017. Data-Driven Immunization. In *ICDM*. 615–624.