On the Robustness of Cascade Diffusion under Node Attacks

Alvis Logins
Aarhus University

Yuchen Li
Singapore Management University

Panagiotis Karras
Aarhus University

ABSTRACT

How can we assess a network's ability to maintain its functionality under attacks? *Network robustness* has been studied extensively in the case of deterministic networks. However, applications such as online information diffusion and the behavior of networked public raise a question of robustness in *probabilistic* networks. We propose three novel robustness measures for networks hosting a diffusion under the Independent Cascade (IC) model, susceptible to node attacks. The outcome of such a process depends on the selection of its initiators, or seeds, by the *seeder*, as well as on two factors outside the seeder's discretion: the attack strategy and the probabilistic diffusion outcome. We consider three levels of seeder awareness regarding these two *uncontrolled* factors, and evaluate the network's viability aggregated over all possible extents of node attacks. We introduce novel algorithms from building blocks found in previous works to evaluate the proposed measures. A thorough experimental study with synthetic and real, scale-free and homogeneous networks establishes that these algorithms are effective and efficient, while the proposed measures highlight differences among networks in terms of robustness and the surprise they furnish when attacked. Last, we devise a new measure of diffusion entropy that can inform the design of probabilistically robust networks.

ACM Reference Format:
Alvis Logins, Yuchen Li, and Panagiotis Karras. 2020. On the Robustness of Cascade Diffusion under Node Attacks. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366423.3380028

1 INTRODUCTION

Networks are ubiquitous in the modelling of infrastructures [39, 44], social interactions [25, 33, 34], physical and life-science phenomena [9, 53]. Yet such networks are subjects to failures or *attacks* [17], whereby some of their elements may be disabled or removed.

Network robustness expresses the degree in which a network retains essential features of its functionality despite attacks [52].

Deterministic robustness. Some measures of network robustness gauge the change of a *deterministic* graph property, such as diameter, average path length [36], or inverse shortest path length [46], after *random* failures. Another measure, grounded on *percolation theory* [55], aggregates the size of the largest connected component (LCC) [36, 52] over all possible sets of blocked nodes [53].

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380028

Stochastic robustness. In applications such as information diffusion and epidemiology there is uncertainty regarding the *connections* in the network, i.e., the network is *stochastic*. We study the operation of such a stochastic network under attacks on nodes, expressed as the expected number of activated nodes under a diffusion process. We refer to this type of robustness as *probabilistic* network robustness. Despite the extensive study of deterministic network robustness [29], its probabilistic counterpart has been scantily studied. There are studies on how to engineer a robust diffusion in an adversarial environment [12, 21], but an investigation on how to measure robustness in such environments is missing.

In this paper, we study the robustness of probabilistic networks expressed by means of the capacity to carry out a successful independent cascade diffusion under node attacks. We introduce three robustness measures built around two sources of uncertainty: attacks on nodes and probabilistic diffusion outcomes on edges.

2 BACKGROUND

Processes in large networks, such as the flow of electricity in a power grid, package routing and delivery in the internet, and protein delivery in a cell, are vulnerable to attacks. Such events may cause electricity blackouts [15] or epidemics [58].

2.1 Deterministic Robustness of Integrity

Network robustness reflects a network's ability to maintain its connectivity under attacks [40]. The connectivity of an *undirected* network is measured by the expected size of its largest connected component (LCC) after an attack [40]. This expected LCC size is also defined on probabilistic undirected networks [24].

Scale-free networks are highly robust to *random* node failures but vulnerable to *targeted* node attacks [3]; increasing their robustness against attacks is in conflict with maintaining their natural robustness against random failures [5]. Some robustness measures take into consideration both random and target failures [47]. Such an *inclusive* measure of robustness, targeted by a local-search heuristic in [53], is the sum of *worst-case* LCC sizes over all cardinalities of sets of blocked nodes:

$$R(G) = \frac{1}{n^2} \sum_{Q=1}^{n} s(Q) \tag{1}$$

where n is the number of nodes in the network and $s(Q)$ is the size of the LCC after removing Q nodes; the normalization by n^2 ensures values are comparable across networks, being in the range $[\frac{1}{n}, \frac{n-1}{2n}]$. The heuristic in [53] leads to an onion-like graph structure, with nodes of similar degree tending to be connected to each other [19]. The closely related *network reliability* problem [18] secures connectivity between two predefined node sets under edge failures. We are interested in the robustness of stochastic diffusion processes under node attacks, which resembles the robustness of deterministic networks under node attacks and random edge failures, yet has received limited attention [10].

2.2 Stochastic Robustness of Diffusion

Network robustness also refers to a network's capacity to host a diffusion process despite the exclusion of some network elements [5, 7, 11, 21]. The mathematical modelling of diffusion is independent of semantics: it may be a diffusion of *information*, of *cascading failures*, or a viral *infection epidemic* [15]. Similarly, a node attack is mathematically equivalent to a node immunization or failure. As the effect of node attacks is evaluated by a stochastic process, we reach the concept of *stochastic robustness*.

A diffusion may be *epidemic*, *threshold*, or *cascading* [61]. There are two popular epidemic models [49]: By the **SIS** model, nodes are either ***s**usceptible* or ***i**nfected*; a node may get infected from its neighbors and become susceptible again after some time. By the **SIR** model, it may ***r**ecover* and becomes immune. The *expected* size of an SIR epidemic starting at u is equal to the expected size of the connected component that contains u [16]. Epidemic models typically consider a homogeneous *infection rate*, yet two models study *information diffusion* with heterogeneous rates [35]: the *Independent Cascade* (IC) model (a special case of SIR [62]) and the *Linear Threshold* (LT) model [28, 34]. Under these models, the *Influence Maximization* (IM) problem [28] seeks a set of initially active nodes, or *seeds*, that maximizes the expected number of activated nodes.

2.3 Robustness under the IC model

We focus on the IC model, widely used to study word-of-mouth effects in social networks [34], by which a diffusion proceeds in discrete time steps. At time $t = 0$, a set of *seed* nodes $S \in V$ are activated. Any node v activated at time t tries to activate its out-neighbours at time $t + 1$, and succeeds with an independent probability $p_e = p_{uv}$ for each neighbor u. In case of success, the edge e is active. This cascading process terminates when there are no more trials for activation. The set of active nodes and edges forms a deterministic *live-edge* graph g [28]. The *spread*, or expected number of activated nodes, is the expected number of nodes reachable from S in G, while each edge may fail independently with probability $1 - p_e$. Hence, diffusion robustness under the IC model corresponds to the deterministic robustness under targeted node attacks and random edge failures with respect to seeds.

Related problems are sensitivity to *edge perturbations* [2, 20, 57] and *robust influence maximization* (RIM) under edge perturbation [12] or any adversarial source of uncertainty [21]. Given a finite set of adversarial strategies Θ, the objective in [21] is:

$$\max_{S,|S|\leq k} \min_{\theta\in\Theta} \frac{\sigma_\theta(S)}{\sigma_\theta(S^*_\theta)} \tag{2}$$

where $\sigma_\theta(S)$ is the spread achieved by seed set S under strategy θ, S^*_θ is the optimal seed set for θ, and k is a budget constraint; the normalization by $\sigma_\theta(S^*_\theta)$ measures the fraction over optimal influence; an absolute measure is used with continuous θ in [26].

The Saturate Greedy (SatGreedy) algorithm [21] solves the RIM problem by targeting the cumulative effect of all strategies, which is a submodular objective. This algorithm, applicable on any monotonic and submodular parameterization of the spread function, provides a bi-criteria approximation guarantee: violating the budget constraint k by an $O(k \ln |\Theta|)$ factor leads to an $(1 - \frac{1}{e})$ approximation of the optimal solution. We adopt the RIM objective as a component in one of the measures we introduce.

Stochastic graph with nodes V and edges E	$G = (V, E)$
Number of nodes and edges of G	n, m
Deterministic graph sampled from G	$g \sim G$
Edge probability parameter	W
Set of attack strategies	$\Theta = \{\theta_i\}$
Degree of a node v	$d(v \in V)$
Number of blocked (removed) nodes	ℓ
Reachability indicator function	$I(v, S)$
EMR-RNI	D
Expected number of activated nodes	σ
A seed set S and size of the set k	S, $k = \|S\|$

Table 1: Notations

3 DIFFUSION ROBUSTNESS MEASURES

We propose three robustness measures, anchored on the awareness of a *seeder*, who selects seed nodes, regarding node attacks and probabilistic diffusion outcomes. Table 1 lists our notations.

3.1 Attack Strategies

We measure robustness against an attacker who disables nodes. A consideration of all possible attack strategies amounts to the NP-hard problem of *node immunization* [11, 22, 38, 62]; instead, we demarcate a strategic set of structure-aware attack strategies on a directed stochastic network G, $\Theta_G = \{\theta^i_G\}$ [21]; $\theta^i_G(\ell)$ is a set of ℓ nodes in G chosen by strategy θ; g_θ denotes the graph obtained by removing nodes from a deterministic instance g of G according to $\theta(\ell)$. We opt for strategies that are also node ranking functions. A recent study assigns attack strategies of four *types* to three or four *clusters* by applying several distance measures on their outputs [4]. We select six strategies that represent each type and cluster in [4], plus a spectral-based baseline, NetShield [11, 37, 50, 62]:

(1) *Degree* picks nodes with the largest degree;
(2) *Random* picks seed nodes uniformly at random;
(3) *Acquaintance* [14] picks a random node's neighbor;
(4) *PageRank* ranks nodes by PageRank values [45];
(5) *Katz* centrality [27] equals $x_i = \alpha \sum_j \mathbf{A_{ij}} x_j + \beta$, where $\alpha = 0.1$, $\beta = 1$, and $\mathbf{A}$ the network's adjacency matrix.
(6) *Betweenness* centrality is the sum of the fraction of all-pairs shortest paths that pass through a node.
(7) *NetShield* [11] greedily selects a set of nodes S, aiming to maximize a spectrally defined *Shield value*.

3.2 Awareness-based Robustness Measures

We define three robustness notions based on the abstraction of *seeder awareness* of attacks and diffusion events, aggregating outcomes over all possible attack sizes, and a notion of *diffusion entropy* that shows how much difference seeder awareness can make.

3.2.1 EMR. Assume an *omniscient* seeder with access to an oracle that predicts the outcome g of a diffusion on G and of an attack on g that produces g_θ. As discussed in Section 2.1, the robustness of a deterministic undirected network G can be expressed by its largest connected component (LCC) [40]. When G is a *directed* network, the LCC substructure is generalized to either of the largest strongly or weakly connected component [54]. Here, we define the expected maximum number of nodes such an omniscient seeder can reach by diffusion from a seed set S of size k in G under a worst-case attack strategy $\theta \in \Theta(\ell)$ as the *Expected Maximum Reach* (EMR):

$$EMR_G(\ell) = \min_{\theta\in\Theta(\ell)} \mathbb{E}_{g_\theta\sim G}\left[\max_{S:|S|\leq k} \sum_{v\in g_\theta} I(v, S)\right] \tag{3}$$

where $I(v, S)$ indicates whether there exists a path from S to node v in a live-edge instance of a directed network, g_θ; $\sum_{v \in g_\theta} I(v, S)$ is the size of a maximum forest with at most k roots. Our first measure aggregates $EMR_G(\ell)$ over all values of ℓ, normalized by network size. We call this measure *sum of EMR* or *SEMR*:

$$SEMR_G = \frac{1}{n} \sum_{\ell=1}^{n} EMR_G(\ell) \tag{4}$$

We introduce an algorithm for SEMR computation in Section 3.3.

3.2.2 RNI. Let us now consider a seeder lacking knowledge of diffusion outcomes, but having access to an oracle that predicts node attacks. We define the *maximum* number of nodes such a seeder can expect to reach in G under a worst-case attack strategy $\theta \in \Theta(\ell)$ as the *Robust Network Immunization* (RNI):

$$RNI_G(\ell) = \min_{\theta \in \Theta(\ell)} \max_{S:|S| \leq k} \mathbb{E}_{g_\theta \sim G} \left[\sum_{v \in g_\theta} I(v, S) \right] \tag{5}$$

where $\mathbb{E}_{g_\theta \sim G}[\sum_{v \in g_\theta} I(v, S)]$ is the *expected* size of the number of nodes $v \in g_\theta$ to which a path exists from S. Our second robustness measure aggregates $RNI_G(\ell)$ over all values of ℓ, normalized by network size. We call this measure *SRNI*:

$$SRNI_G = \frac{1}{n} \sum_{\ell=1}^{n} RNI_G(\ell) \tag{6}$$

The computation of RNI requires solving an influence maximization (IM) problem on a graph with $\theta(\ell)$ nodes removed for each attack strategy $\theta \in \Theta$ and each value of ℓ. We do so while building sampled networks g_θ incrementally, using the dynamic IM algorithm (DIM) [43], which extends IMM [56].

3.2.3 RIM. Last, we consider a seeder who has information neither about diffusion outcomes, nor about node attacks. We define the *maximum* number of nodes such a seeder can expect to reach under a *worst-case* $\theta \in \Theta(\ell)$ as *Robust Influence Maximization* (RIM) [21]:

$$RIM_G(\ell) = \max_{S:|S| \leq k} \min_{\theta \in \Theta(\ell)} \mathbb{E}_{g \sim G} \left[\sum_{v \in g_\theta} I(v, S) \right] \tag{7}$$

where $\mathbb{E}_{g_\theta \sim G}[\sum_{v \in g_\theta} I(v, S)]$ is the *expected* number of nodes $v \in g_\theta$ to which a path exists from S. Our third measure aggregates $RIM_G(\ell)$ over ℓ, normalized by network size; we call it SRIM:

$$SRIM_G = \frac{1}{n} \sum_{\ell=1}^{n} RIM_G(\ell) \tag{8}$$

To calculate SRIM we apply SatGreedy [21] with the objective in Equation 2 modified to account for node removals rather than edge perturbation and normalizing spread by network size $|V|$ rather than by the optimal spread under strategy θ, since we are interested in robustness in the absolute sense:

$$\max_S \rho'(S) = \max_S \min_\theta \frac{\sigma_\theta(S)}{|V|}$$

Further, we enhance the runtime of SatGreedy using the same *dynamic* approach as for SRNI [43] to estimate spread. We also consider the baselines proposed in [21]: *SingleGreedy* selects k seeds sequentially, choosing a seed that maximizes the objective in each step. *AllGreedy* finds the best seed set for each adversary, and selects the one of these that maximizes the objective.

3.3 SEMR Computation

To compute the SEMR measure for a single seed, we need to calculate expected maximum tree sizes over randomly sampled attacked networks g, under each attack strategy. We consider attack strategies $\theta \in \Theta$ under which the set of blocked nodes for $\ell + 1$ is a superset of that for ℓ: $\theta(\ell) \subset \theta(\ell + 1)$. To obtain a sequence of attack sets $\theta_g(\ell)$ for different ℓ on g, it suffices to sequentially remove nodes from g, or, equivalently, sequentially add nodes to g. We compute maximum tree sizes over several random samples g from G, with edges pre-sampled and nodes incrementally added according to each strategy θ, and average values per ℓ to get $EMR(\ell)$. For the sake of efficiency, we employ a dynamic reachability index that returns nodes reachable from any node and also supports node insertions, building upon DAGGER [60]. Given g, the index maintains a directed acyclic graph (DAG), where each node represents a strongly connected component (SCC) in g, called *graph condensation*. A node's insertion incurs the insertion of its incident edges. Assume a new edge $e = (u, v)$ is inserted, and s and t being the SCCs u and v belong to, respectively. DAGGER checks whether there is a path from t to s, using its reachability index. If there is, then DAGGER merges all SCCs on all paths from t to s.

Algorithm 1 SEMR Computation

```
1:  function INSERT(w, H)
2:      DAGGER.INSERT(w)
3:      w' ← SCC(w)                          ▷ w' corresponds to an SCC in g and has label r
4:      R ← set of nodes removed from g'
5:      for all v'|(w', v') ∈ g' do
6:          w'.r ← w'.r ∪ v'.r
7:      Q ← {u'|∃ path w' ⇝ u' in (g')^T}
8:      for all u' ∈ Q do
9:          u'.r ← u'.r ∪ w'.r \ R
10:         if ∄v'|(v', u') ∈ E' then
11:             H.insert(< u', |{v ∈ g|SCC(v) ∈ u'.r}| >)
12: function SEMR
13:     for all θ ∈ Θ do
14:         s_θ ← empty list
15:         Initialize DAGGER with empty graph
16:         H ← a descending heap of < key, value >
17:         for all w ∈ θ.reverse() do
18:             INSERT(w, H)
19:             v', s ← H.top()              ▷ Apply CELF here for k > 1
20:             s_θ[ℓ] = s
21:     s_min ← empty list
22:     s_min[ℓ] ← min_θ s_θ[ℓ] ∀ℓ
23:     return ∑ s_min
```

We extend DAGGER with a query that computes SEMR for a single seed (Algorithm 1). Let $g' = (V', E')$ be the DAG that corresponds to g. For each node $v' \in V'$, we maintain a label $v'.r$ as the set of nodes $u' \in V'$ reachable from v': $v'.r = \{u' \in V' | \exists \text{path } v' \rightarrow u'\} \cup \{v'\}$, and a heap H organizing tree root nodes (i.e., nodes with zero in-degree) by the sum of reachable SCC sizes. Upon the insertion of a new node w to g, we collect the ids of w's SCC (Line 3) and invalidated SCCs R (Line 4), calculate the reach $w'.r$ of the SCC w belongs to, $w' \in g'$, based on its out-neighbours (Lines 5-6), and update the labels of all ascendant nodes of w', u' reachable from w' in the reverse DAG $(g')^T$, accordingly (Lines 8-11). Upon reaching an ascendant root node u', we update H (Lines 10-11). To compute SEMR for a single seed, we obtain maximum tree sizes from H (Line 19). For k seeds, we pick k nodes from H, prioritized by marginal gain in terms of reachable nodes in *lazy greedy* fashion [42], as the objective function is submodular. The performance of SEMR computation depends on set union and subtraction operations (Lines 6 and 9).

4 EXPERIMENTAL STUDY

We investigate the nature of all three measures and study their interrelationships. Experiments ran on a 378G RAM Intel Xeon CPU @ 3.10GHz running Ubuntu 18.04. All algorithms are implemented[1] in C++ and compiled with gcc 7.4 with -O3 optimization. We set timeout 10h per one measure computation. Runtime and timeout do not include time for the strategy set Θ computation, which is the same for all measures. We assign edge probabilities either randomly, or uniformly. For random assignment, we pick a value for each edge uniformly from 0 to W, where W is a parameter. For uniform assignment, we assign a certain W value to each edge. We refer to these two types of assignment as *Random* and *Uniform*.

Synthetic Networks. We study *power-law networks*, represented by the **Barabási-Albert** (BA) model, and *homogeneous networks*, represented by the **Gaussian Random Partition** (GRP) [8] and **Watts Strogatz** (WS) models. For **BA**, we use the algorithm of Holme and Kim [23], which extends the original Barabási-Albert model, yet use the BA label as its basis. The algorithm randomly creates μ 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. **GRP** groups nodes so that group sizes follow a Gaussian distribution with expected size s and variance of size equal to s/v, where v is a shape parameter. It uses a probability value p_{in} for edges across nodes in the same group, and p_{out} otherwise. **WS** models self-organizing small-world systems [59], with two parameters: l indicates how many neighbors each node is joined with in a ring; p is a probability of edge rewiring, inducing disorder.

Network	$\lvert V\rvert\,[\cdot 10^3]$	$\lvert E\rvert\,[\cdot 10^3]$	$d_{max}, \overline{d}$	cl
Blogs	1.2	19.0	467, 31	0.336
Minnesota	2.6	3.3	5, 2	0.024
VK	2.8	40.8	288, 29	0.247
Advogato	6.6	47.3	947, 14	0.211
DBLP	12.6	49.7	710, 8	0.117
Brightkite	56.7	212.9	1134, 8	0.117
Gnutella	62.6	147.9	95, 5	0.007
Stanford	281.9	2312.5	38626, 16	0.597

Table 2: Real-world datasets. $d_{max}, \overline{d}$ is maximum and average degree, cl is average clustering coefficient [51].

Real-world networks. We use real-world datasets of various sizes and degree distributions: Blogs contains front-page hyperlinks between blogs during the 2004 US election [1, 30]. DBLP is a citation network of scientific papers [30, 32]. Advogato is a network of trust relationships in an online community platform for free-software developers [30, 41]. Minnesota is a road network [48]. VK is a social network with influence probabilities derived from the content of posts published by users [37]. Brightkite is a location-based social network [13]. Gnutella is snapshots of the Gnutella peer-to-peer file sharing network [31]. Table 2 lists our real-world datasets.

4.1 Choice of Algorithm for RIM Computation

As a preliminary experimental choice, we study the performance of methods for RIM calculation, including algorithms and baselines proposed in [21]. We use the IMM algorithm for influence maximization [56] as a non-robust baseline. We include SingleGreedy *with* the CELF (i.e., lazy greedy) optimization, proposed in [21], and also its variant *without* it, given that, on this non-submodular problem objective, the CELF optimization affects quality.

[1] The code is available at https://github.com/allogn/robustness.

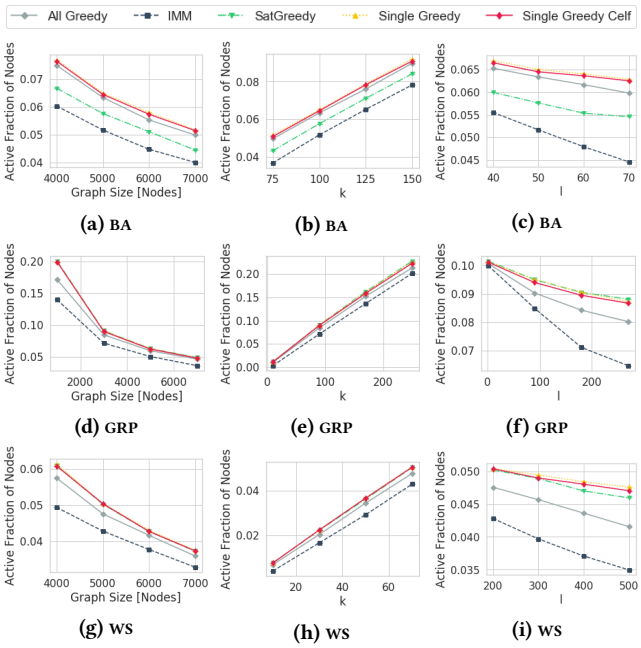


Figure 1: RIM under node attacks. BA ($n = 5 \cdot 10^3$, $\ell = 50$, $k = 100$, $\mu = 2$), WS ($n = 5 \cdot 10^3$, $\ell = 200$, $k = 70$), GRP ($n = 3 \cdot 10^3$, $\ell = 180$, $k = 90$). $W = 0.1$, SatGreedy: $\gamma = 10^{-4}$.

We compare the performance of algorithms in the computation of the unaggregated RIM objective, with BA, GRP, and WS networks. Figure 1 illustrates the results vs. graph size n, seed set size k, and number of attacked nodes ℓ. We observe that SingleGreedy with and without CELF matches or outperforms SatGreedy, while IMM has a disadvantage that grows with ℓ, imprinting the significance of using robust algorithms.

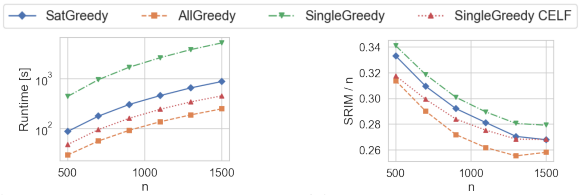


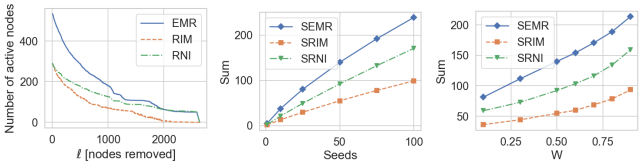
Figure 2: SRIM on BA, $p = 0.4$, $\mu = 10$, $W = 0.3$, $k = 50$.

Now we drop the non-robust IMM algorithm out of the comparison, and study the performance of robust algorithms, with the DIM algorithm embedded, on the runtime for computing, and value of, the *aggregate* SRIM robustness measure on the BA network. Figure 2 shows our results for $k = 50$ seeds. As in Figure 1, SingleGreedy stands out in terms of objective, at the cost of higher runtime. The difference in objective is more prominent now, as we aggregate the measure over all values from 1 to ℓ. The runtime for computing Θ is negligible, reaching 4s for the largest network.

These results indicate that SingleGreedy (without CELF) offers the best effectiveness, but significantly worse efficiency. SingleGreedy with CELF matches the performance of SingleGreedy, matches or outperforms that of SatGreedy, is more efficient, and does not require any accuracy parameter γ, as SatGreedy does. Ergo, we opt for SingleGreedy *with* CELF in the following.

4.2 Measure relationships

We now study the relation between measures and their sensitivity to the set of attack strategies, using two homogeneous networks (Minnesota and GRP) and two power-law networks (Blogs and VK).



(a) $k = 50, W = 0.8$ **(b)** $W = 0.5$ **(c)** $k = 50$

Figure 3: Measures on Minnesota road network.

Figure 3a plots plain EMR, RNI, and RIM values, without aggregation, vs. ℓ on Minnesota. Values decrease gradually, revealing some irregularities of graph structure in the middle range of ℓ. EMR and RNI follow a similar pattern, while RIM differs. For instance, from $\ell = 1000$ to 2000 EMR and RNI present two abrupt drops at the same value of ℓ. RIM has more and smaller irregularities. Figures 3b and 3c present the summed measures (SEMR, SRNI, and SRIM) vs. seed set size and influence probabilities W, respectively. The difference between them grows especially with seed set size.

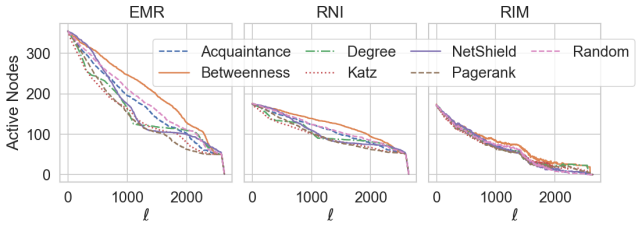
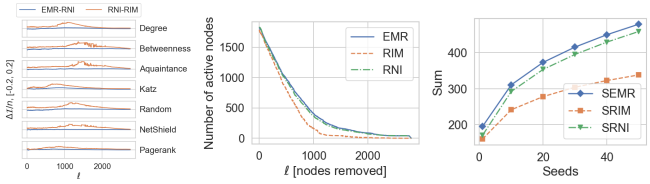


Figure 4: Effect of 7 attack strategies. Minnesota road network, $W = 0.5$**,** $k = 50$**,** *Random*.

Figure 4 presents a decomposition: instead of a minimum over all strategies, we plot the expected influence per strategy, with the seed set selected by each algorithm. We observe that EMR and RNI follow the same trend *also* for each strategy separately. This is especially conspicuous with NetShield, which shows poor performance in its immunization objective for small values of ℓ, but swiftly improves in the middle range; it then becomes the most effective strategy for a short ℓ range, but looses that position to PageRank. Remarkably, results for RNI presents the same outline, but scaled to a smaller values of active nodes. On the other hand, RIM exhibits a different behaviour, as all strategies mostly produce the same response to the selected seeds. This result illustrates the difference of RIM from the other two measures: RIM is based on the worst case among the complete set of strategies by nature, hence can afford to let the selected seeds perform almost equally well on any attack.

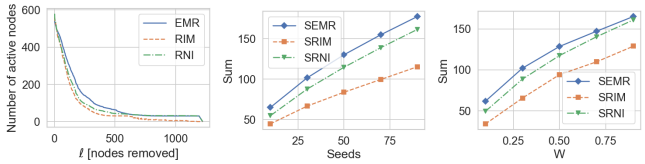
Figure 5a plots the differences EMR-RNI and RNI-RIM vs. ℓ on the VK network. RNI-RIM has a convex shape with a maximum in the middle-range ℓ, while EMR-RNI is almost zero in the whole range. This behavior differs from the one we observed with the BA and DBLP networks, where there is a peak on EMR-RNI. Figure 5b plots non-aggregate measure values for $k = 40$. RNI is very close to EMR along the whole range of ℓ; on the other hand, RNI-RIM also peaks close to the maximum curvature of lines. Figure 5c shows that the effect becomes stronger with larger k, *aggregating* over all ℓ values: SRNI remains close to SEMR, while SRIM diverges from the others; this divergence implies that, on power-law networks, knowledge about the attack, gained when moving from RIM to RNI, is more valuable than knowledge about the stochastic edge outcome, gained when moving from RNI to EMR.



(a) Deltas **(b) Measures vs.** ℓ **(c) Sums vs.** k

Figure 5: Dependency of measures on VK social network.

Figures 6 and 7 show the proximity among the three aggregate measures on the Blogs and GRP networks. On the *power-law* Blogs network, the trend is similar to VK, with RNI close to EMR. However, on the *homogeneous* GRP network, RNI is close to RIM for the whole spectrum of network shape parameters. We conclude that network topology determines what gain of knowledge matters most; on a homogeneous network, knowledge about the stochastic edge outcome is more valuable than knowledge about the attack.



(a) $k = 50, W = 0.8$ **(b)** $W = 0.5$ **(c)** $k = 50$

Figure 6: Dependency of measures, Blogs network. *Random*.

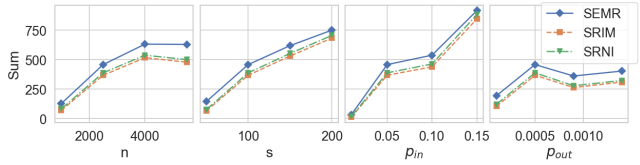
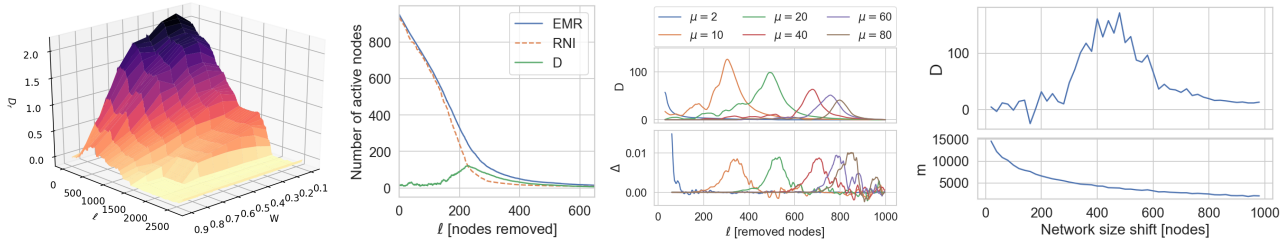


Figure 7: Dependency on network parameters. GRP. $n = 2500$**,** $s = 100$**,** $p_{in} = 0.05$**,** $p_{out} = 5 \cdot 10^{-4}$**,** $k = 10$**,** $W = 0.3$**,** *Random*.

Another interesting feature is the shape of the tail of distributions (Figures 4, 5b and 6a). There exists a value of $\ell = \ell'$, such that all three measures converge to the value of k as ℓ grows towards ℓ', but for $\ell > \ell'$ RIM drops to 0, while others remain at k. The drop of RIM is concave, with a gap of first derivative. The region $\ell > \ell'$ corresponds to the case where the attacker blocks all nodes by at least one strategy for any seed set. That strategy determines RIM. However, for EMR and RNI, seeds are selected after the attack, therefore there are at least k non-blocked nodes.

4.3 EMR vs RNI: the diffusion entropy

The EMR and RNI measures both represent cases in which the attacker has to prepare for the worst-case, i.e., the case in which the seeder is aware of the attacker's actions. In other words, both these measures correspond to *robust immunization* problems. Their



(a) D_r**, Minnesota,** $k = 10$ **(b) Peak Location, BA** **(c)** D **vs.** Δ**, BA** **(d) Fixed-size network, BA**

Figure 8: Local maximum of D**; BA parameters:** $n = 1000$, $p = 0.4$, $\mu = 10$**; (b):** $k = 5$**; (d):** $k = 1$, $\Theta = \{\text{Degree}\}$.

difference lies in the fact that, under EMR, the seeder is also aware of the probabilistic network outcome. Thus, the difference between these two probabilistic network robustness measures expresses the surprise effect or, so to speak, *negative entropy* that a probabilistic diffusion outcome can present to the attacker; it shows how much worse the spread can be in the case of a seeder aware of probabilistic outcomes in comparison to the best guess of a seeder unaware of such outcomes. We study the impact of this difference in more detail, using uniform probability assignment so as to focus on structural effects. We consider the absolute difference D among the two measures; and also the relative difference with respect to RNI, D_r.

$$D = \text{EMR} - \text{RNI}, \quad D_r = \frac{\text{EMR} - \text{RNI}}{\text{RNI}} \tag{9}$$

Figure 8a shows the surface of D_r for different values of ℓ and W on the Minnesota network. D_r is larger for smaller number of removed nodes ℓ, and drops with larger edge probabilities. Still, it is not monotonic vs. W; it obtains a maximum value around $W = 1.5$, and the peak is more explicit with smaller ℓ. Figure 8b shows that this non-monotonic behavior of D_r also appears with respect to ℓ on a BA network, and indicates exactly where the peak is located. Compared to Figure 8c, where peaks are presented only for a single seed, we see that on Figure 8b the peak has larger width.

D also relates to the relative marginal gain seeds addition by the seeder. We define $\delta_{\theta_i}(\ell)$ as the relative marginal gain of the second seed for any strategy $\theta_i \in \Theta$ under ℓ attacked nodes:

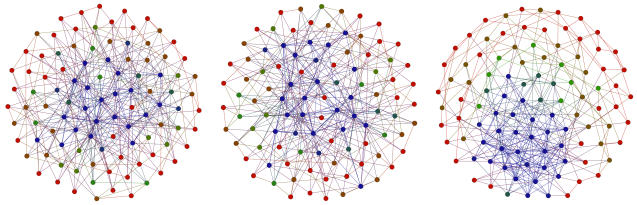
$$\delta_{\theta_i}(\ell) = \frac{\max_{S:|S|=1} \sigma_{\theta_i(\ell)}(S) - \max_{S:|S|=2} \sigma_{\theta_i(\ell)}(S)}{\max_{S:|S|=1} \sigma_{\theta_i(\ell)}(S)} \tag{10}$$

We then calculate a new quantity $\Delta(\ell)$ as the maximum differential quotient of δ over all strategies for each ℓ:

$$\Delta(\ell) = \max_{\theta_i \in \Theta} \left\{ \delta_{\theta_i}(\ell) - \delta_{\theta_i}(\ell - 1) \right\} \tag{11}$$

Figure 8c juxtaposes D and Δ, plotted with moving average smoothing. Their two peaks align, with a slight shift to the right for Δ. This finding implies that, on BA networks, the values of ℓ for which the network ceases to be strongly centralized, hence Δ flattens out, would also cause the highest surprise to an attacker.

We exploit this observation to generate networks of *enhanced robustness*: we fix size to 1000 nodes, yet first generate a network of larger size and then remove superfluous nodes by the Degree strategy. We call the amount of nodes first added and then removed *shift*. Figure 8d plots D vs. shift. Shifting improves network robustness in terms of D; we create networks in which a seeder has the potential to perform surprisingly well against an attacker. The lower subfigure plots the number of edges in the obtained network; as there is no correlation between the peak of D and number of edges, the peak must be attributed to the network's structure.



(a) Original **(b) SRIM-based** **(c) SEMR-based**

Figure 9: Robust BA networks

4.4 Case Studies

We provide examples of robust networks using the local search heuristic of [53], which randomly samples pairs of edges, e.g., pair $\{(v_1, u_1), (v_2, u_2)\}$, and rewires them to $\{(v_1, u_2), (v_2, u_1)\}$ if that leads to a higher robustness measure. We experiment with SRIM and SEMR, since SRNI exhibits similar behavior to SEMR (see Section 4.2). The sampling proceeds until $|E|$ iterations bring no change.

We experiment with a random BA network of 100 nodes, uniform edge probability of 0.5, and 2 seeds. Figure 9 shows the original network (non-robust), and two networks obtained by the aforementioned procedure for SRIM and SEMR, respectively. Colors indicate similar node degrees, blue for larger, green for medium, and red for smaller. We plot the networks using the Fruchterman Reingold algorithm [6]. We note that the network targeting SEMR has a layered onion-like structure, similar to robust static networks [19], while the other two networks do not show evident patterns.

5 CONCLUSIONS

We introduced three aggregate measures that evaluate the diffusion robustness of probabilistic networks, anchored on a seeder who orchestrates an Independent Cascade diffusion under node attacks. Each measure is based on a notion of worst-case maximum expected spread. We introduced efficient algorithms to calculate these measures and sample-based versions thereof that enable their computation on realistic networks of up to 10^5 nodes. Our experimental study determined that, on scale-free networks, measures sharing the same notion of seeder awareness regarding the adversarial attack are closer, while those sharing the same notion of awareness regarding the network instance are closer on homogeneous networks. Our results provide tools for assessing the robustness of real-world probabilistic networks, and offer guidelines on how to achieve and enhance network robustness.

REFERENCES

[1] Lada A. Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proc. 3rd Intl Workshop on Link Discovery*. 36–43.

[2] Abhijin Adiga, Chris J. Kuhlman, Henning S. Mortveit, and Anil Vullikanti. 2014. Sensitivity of Diffusion Dynamics to Network Uncertainty. *Journal of Artificial Intelligence Research* 51 (2014), 207–226.

[3] Réka Albert, Hawoong Jeong, and Albert-László Barabási. 2000. Error and attack tolerance of complex networks. *Nature* 406, 6794 (2000), 378.

[4] Mirza Basim Baig and Leman Akoglu. 2015. Correlation of Node Importance Measures: An Empirical Study through Graph Robustness. In *WWW Conference Companion*. 275–281.

[5] Albert-László Barabási. 2016. *Network science*. Cambridge university press.

[6] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009).

[7] Ilija Bogunovic. 2012. *Robust protection of networks against cascading phenomena*. Master's thesis. Department of Computer Science, ETH Zürich.

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

[9] Nicholas E Brunk, Lye Siang Lee, James A Glazier, William Butske, and Adam Zlotnick. 2018. Molecular jenga: the percolation phase transition (collapse) in virus capsids. *Physical Biology* 15, 5 (2018), 056005.

[10] Lincoln Chayes, Roberto H Schonmann, et al. 2000. Mixed percolation as a bridge between site and bond percolation. *The Annals of Applied Probability* 10, 4 (2000), 1182–1196.

[11] 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 TKDE* 28, 1 (2016), 113–126.

[12] Wei Chen, Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou. 2016. Robust Influence Maximization. In *KDD*. 795–804.

[13] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *KDD*. 1082–1090.

[14] Reuven Cohen, Shlomo Havlin, and Daniel Ben-Avraham. 2003. Efficient immunization strategies for computer networks and populations. *Physical review letters* 91, 24 (2003), 247901.

[15] Paolo Crucitti, Vito Latora, and Massimo Marchiori. 2004. Model for cascading failures in complex networks. *Physical Review E* 69, 4 (2004), 045104.

[16] Rick Durrett. 2010. Some features of the spread of epidemics and information on a random graph. *Proc. National Academy of Sciences* 107, 10 (2010), 4491–4498.

[17] Wendy Ellens and Robert E. Kooij. 2013. Graph measures and network robustness. *CoRR* abs/1311.5064 (2013). arXiv:1311.5064

[18] Christian Frey, Andreas Züfle, Tobias Emrich, and Matthias Renz. 2018. Efficient Information Flow Maximization in Probabilistic Graphs. *IEEE TKDE* 30, 5 (2018), 880–894.

[19] Yukio Hayashi and Naoya Uchiyama. 2018. Onion-like networks are both robust and resilient. *Scientific Reports* 8, 1 (2018), 11241.

[20] Xinran He and David Kempe. 2014. Stability of influence maximization. In *KDD*. 1256–1265.

[21] Xinran He and David Kempe. 2016. Robust Influence Maximization. In *KDD*. 885–894.

[22] Xinran He, Guojie Song, Wei Chen, and Qingye Jiang. 2012. Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model. In *SDM*. 463–474.

[23] Petter Holme and Beom Jun Kim. 2002. Growing scale-free networks with tunable clustering. *Physical review E* 65, 2 (2002), 26107.

[24] Sergei Ivanov and Panagiotis Karras. 2016. Harvester: Influence Optimization in Symmetric Interaction Networks. In *IEEE Intl Conf. Data Science and Advanced Analytics*. 61–70.

[25] Sergei Ivanov, Konstantinos Theocharidis, Manolis Terrovitis, and Panagiotis Karras. 2017. Content Recommendation for Viral Social Influence. In *SIGIR*. 565–574.

[26] Dimitris Kalimeris, Gal Kaplun, and Yaron Singer. 2019. Robust Influence Maximization for Hyperparametric Models. In *ICML*. 3192–3200.

[27] Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 1 (1953), 39–43.

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

[29] Gunnar W. Klau and René Weiskircher. 2005. Robustness and Resilience. In *Network Analysis*. Springer Berlin Heidelberg, 417–437.

[30] Jérôme Kunegis. 2013. KONECT: the koblenz network collection. In *WWW Conference*. 1343–1350.

[31] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *ACM TKDD* 1, 1 (2007), 2.

[32] Michael Ley. 2002. The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *Proc. 9th Intl Symp. String Processing and Information Retrieval*. 1–10.

[33] Yuchen Li, Ju Fan, George V. Ovchinnikov, and Panagiotis Karras. 2019. Maximizing Multifaceted Network Influence. In *ICDE*. 446–457.

[34] 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.

[35] Sungsu Lim, Joongbo Shin, Namju Kwak, and Kyomin Jung. 2016. Phase transitions for information diffusion in random clustered networks. *The European Physical Journal B* 89, 9 (2016), 188.

[36] Jing Liu, Mingxing Zhou, Shuai Wang, and Penghui Liu. 2017. A comparative study of network robustness measures. *Frontiers of Computer Science* 11, 4 (2017), 568–584.

[37] Alvis Logins and Panagiotis Karras. 2019. Content-Based Network Influence Probabilities: Extraction and Application. In *ICDM Workshops*. 69–72.

[38] Alvis Logins and Panagiotis Karras. 2019. An Experimental Study on Network Immunization. In *EDBT*. 726–729.

[39] Alvis Logins, Panagiotis Karras, and Christian S. Jensen. 2019. Multicapacity Facility Selection in Networks. In *ICDE*. 794–805.

[40] Oriol Lordan and Maria Albareda-Sambola. 2019. Exact calculation of network robustness. *Reliability Engineering & System Safety* 183 (2019), 276–280.

[41] Paolo Massa, Martino Salvetti, and Danilo Tomasoni. 2009. Bowling Alone and Trust Decline in Social Network Sites. In *The 8th IEEE Intl Conf. Dependable, Autonomic and Secure Computing*. 658–663.

[42] Michel Minoux. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*. Springer, 234–243.

[43] Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, and Ken-ichi Kawarabayashi. 2016. Dynamic Influence Analysis in Evolving Networks. *PVLDB* 9, 12 (2016), 1077–1088.

[44] Alessio Pagani, Guillem Mosquera, Aseel Alturki, Samuel Johnson, Stephen Jarvis, Alan Wilson, Weisi Guo, and Liz Varga. 2019. Resilience or robustness: identifying topological vulnerabilities in rail networks. *Royal Society Open Science* 6, 2 (2019), 181301.

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

[46] Vitor Hugo Louzada Patricio, Fabio Daolio, Hans J. Herrmann, and Marco Tomassini. 2015. *Propagation Phenomena in Real World Networks. Generating Robust and Efficient Networks Under Targeted Attacks*. Springer, 215–224.

[47] G. Paul, T. Tanizawa, S. Havlin, and H. E. Stanley. 2005. Optimization of robustness of complex networks. *The European Physical Journal B* 48, 1 (2005), 149–149.

[48] Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*. 4292–4293. http://networkrepository.com

[49] Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. 2013. *Modern Epidemiology*. Lippincott Williams & Wilki.

[50] 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).

[51] Thomas Schank and Dorothea Wagner. 2005. Approximating Clustering Coefficient and Transitivity. *Journal of Graph Algorithms and Applications* 9, 2 (2005), 265–275.

[52] Tiago A. Schieber, Martín Gómez Ravetti, and Panos M. Pardalos. 2016. A Review on Network Robustness from an Information Theory Perspective. In *Proc. 9th Intl Conf. on Discrete Optimization and Operations Research*. 50–60.

[53] Christian M Schneider, André A Moreira, José S Andrade, Shlomo Havlin, and Hans J Herrmann. 2011. Mitigation of malicious attacks on networks. *Proc. National Academy of Sciences* 108, 10 (2011), 3838–3841.

[54] N. Schwartz, R. Cohen, D. ben Avraham, A.-L. Barabási, and S. Havlin. 2002. Percolation in directed scale-free networks. *Physical Review E* 66, 1 (2002), 015104.

[55] Dietrich Stauffer and Ammon Aharony. 2003. *Introduction to percolation theory*. Taylor & Francis.

[56] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *SIGMOD*. 1539–1554.

[57] Sho Tsugawa and Hiroyuki Ohsaki. 2017. On the Robustness of Influence Maximization Algorithms against Non-Adversarial Perturbations. In *ASONAM*. 91–94.

[58] Emilia Vynnycky and Richard White. 2010. *An introduction to infectious disease modelling*. Oxford University Press.

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

[60] Hilmi Yildirim, Vineet Chaoji, and Mohammed J. Zaki. 2013. DAGGER: A Scalable Index for Reachability Queries in Large Dynamic Graphs. *CoRR* abs/1301.0977 (2013).

[61] Huiyuan Zhang, Subhankar Mishra, My T Thai, J Wu, and Y Wang. 2014. Recent advances in information diffusion and influence maximization in complex social networks. *Opportunistic Mobile Social Networks* 37, 1.1 (2014), 37.

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