Models and Algorithms for Network Immunization

George Giakkoupis
University of Toronto
ggiak@cs.toronto.edu

Aristides Gionis
University of Helsinki
gionis@cs.helsinki.fi

Evimaria Terzi
University of Helsinki
terzi@cs.helsinki.fi

Panayiotis Tsaparas
University of Helsinki
tsaparas@cs.helsinki.fi

ABSTRACT

Recently, there has been significant research activity in the algorithmic analysis of complex networks, such as social networks, or information networks. A problem of great practical importance is that of network immunization against virus spread. Given a network, a virus-propagation model, and an immunization cost function, we are interested in containing the spread of the virus while minimizing the immunization cost. In this paper, we consider two virus-propagation models and we propose immunization algorithms for each model. The first is a gossip model that has been used by Kempe et al. [12] in the context of designing algorithms for influence maximization in a social network. The second model is motivated by the work of Wang et al. [18]. Here we refine the analysis provided in [18], using a more accurate continuous time model, and we provide necessary and sufficient conditions for immunization in the model. The experimental evaluation of our algorithms show that they perform well on both synthetic and real graphs. Furthermore, our analysis and experiments reveal the following interesting facts (a) the simple heuristic of immunizing the nodes with the highest degree is not optimal, and (b) our algorithms perform significantly better in small-world networks.

1. INTRODUCTION

It is often the case that natural or man-made systems are organized in networks. Examples include the Internet, the Web, social networks, or networks of proteins. The normal operation of such networks is threatened by the diffusion of harmful information that is propagated through the links. Examples include cascading failures in an electrical grid, sexual transmitted diseases in a sexual network, computer viruses on the Internet, harmful gossip or panic in a social network. In all these cases the propagation of the information has a destructive effect on the network, therefore, we are interested in containing its impact, by taking either preventive, or curing action. We will collectively refer to the harmful information that is propagated as a *virus*, and we will refer to the process of impeding the spread of the virus as the immunization of the network. This is a problem of obvious practical importance. We want to prevent disease spreads, protect computer networks from viruses, and control the leakage of sensitive information and unpleasant gossip. At the same time our resources (vaccination, anti-virus software, advertisement targeting) are costly and limited, so we are interested in achieving the best possible effect, while allocating the minimum possible resources.

The problem of immunization is defined informally as follows. Given a network, and a virus-propagation model, assume that an adversary places a number of viruses in the network. We are interested in immunizing the minimum set of nodes, such that in the resulting network the spread of the virus is contained. Alternatively, given a fixed immunization budget, we are interested in containing the spread of the virus as much as possible without exceeding our budget.

The immunization process depends obviously on the virus propagation model. Models for diffusion of information in a network have been studied extensively in various disciplines, including computer science, sociology, physics, and epidemiology. In this paper, we consider two different types of models: the *independent-cascade model*, considered by Kempe et al. [12] for the propagation of gossip, and *dynamic-propagation models* similar to those considered by Wang et al. [18] for epidemic spread. Each of these models, defines a different objective for the immunization algorithm, and thus requires different immunization strategies.

Our contributions in this paper can be summarized as follows.

- We study analytically the dynamic propagation model. Under different formulations we provide necessary and sufficient conditions for an epidemic outbreak. Our analysis complements and extends that of Wang et al. [18].

- We define specific immunization problems for each of the virus-propagation models, and we propose immunization algorithms. For the independent-cascade model the proposed algorithm is a greedy method designed to minimize the spread of the virus in the network. For the dynamic-propagation models the virus spread depends on the eigenvalue of an appropriately defined matrix of the network. Thus, our algorithm attempts to reduce this eigenvalue with the least immunization cost.

- We study the algorithms experimentally on both real and synthetic networks. We observe that contrary to popular belief the greedy algorithm that removes the node with the highest degree does not yield the best results. The failure of this highest-degree heuristic is especially pronounced in graphs with high clustering coefficient, a case that is encountered in many social networks.

KDD 2005 Chicago, Illinois, USA

The rest of the paper is structured as follows. In Section 2 we put our work in perspective with the related literature, and in Section 3 we define the general framework of the immunization problem we consider. In Section 4 we provide precise definitions for immunization and epidemic spread, and we discuss in detail the virus-propagation models we consider. Our immunization algorithms are presented in Section 5, and our experiments in Section 6. Section 7 is a short conclusion.

2. RELATED WORK

The study of mathematical models for epidemic spread has a long history in biological epidemiology due to the need of understanding the mechanism that infectious diseases spread among populations. The pioneering work of Kermack and McKendrick [15] establishes the first stochastic theory for epidemic spread and proves the existence of an *epidemic threshold*, a quantity that determines whether the epidemic will spread or it will die out. Similar threshold phenomena prevail in a many other epidemic models, and a lot of recent work in mathematical epidemiology focuses on providing analytic expressions for epidemic thresholds in different propagation models and different families of networks [3–6, 17].

Cohen [7] and Murray [16] pointed out the connection between the spread of computer viruses and epidemiology, and Kephart and White [13, 14] were among the first to apply epidemiology-based models for analyzing the propagation of computer-viruses.

In homogeneous networks, immunizing random nodes in the network is an effective mechanism for preventing the epidemic spread [1, 17]. However, the method of uniform immunization breaks down for scale-free networks due to the existence of highly connected nodes. Thus, for scale-free networks, Pastor-Satorras et al. [17] and Dezso et al. [9] suggest immunizing highly connected nodes.

The case that the topology of underlying network is unknown is studied by Cohen et al. [8], who show that under a percolation propagation model, the method of immunizing random acquaintances of random nodes is more effective than immunizing random nodes.

Finally, Aspnes et al. [2] study inoculation strategies from a game-theoretic point of view, where individual nodes can decide if they want to be immunized or run the chance to be infected and their goal is to minimize their individual expected cost. They also consider centralized versions of the problem, and they introduce the sum-of-squares partition problem, for which they obtain a polynomial-time $O(\log^2 n)$-approximation algorithm.

3. THE GENERAL FRAMEWORK

The immunization problem has the following components.

- The network, over which the virus propagates. This is modeled as a graph $G = (V, E)$. We will consider only *undirected* graphs, although most of our results go through for directed graphs as well. It is possible to assume that the graphs are drawn from a specific family. In the definitions and algorithms we consider all possible graphs. In our experiments we investigate various popular families of graphs.

- The virus propagation model, that determines how the virus spreads in the network.

- The immunization algorithm, which has the power to immunize a set of nodes in the network in order to minimize the spread of the virus. An immunized node cannot receive or transfer the virus. Conceptually, for the purpose of the virus propagation we can think of the immunized nodes as being removed from the network. The cost of the immunization algorithm is the number of nodes that are immunized.

- The adversary, who has knowledge of the virus propagation model, and she plants r copies of the virus in the network so as to maximize the spread of the virus. We will use A_r to denote such an adversary. The adversary may also have knowledge of the choices made by the immunization algorithm. We call such an adversary an *adaptive* adversary. We also consider a randomized adversary that places the virus copies uniformly at random. We call such an adversary *blind*.

4. VIRUS PROPAGATION MODELS AND EPIDEMIC SPREAD

Let $G = (V, E)$ denote the underlying graph of our network, and let n denote the number of nodes in the network. A common thread underlying all of the propagation models we consider is the *propagation probability* matrix P. For every edge (i, j) in the graph G, the entry p_{ij} of the matrix holds the probability that a copy of the virus is propagated across edge (i, j). We have $p_{ij} = 0$ for all $(i, j) \notin E$. The models that we consider differ in the way they use matrix P. Furthermore, in the dynamic propagation model, we also assume that at any time step if node i is infected, it gets cured with probability p_i.

Now, let G be a graph of size n, and let N_r be a subset of r nodes of the graph where r copies of the virus are placed. Assume now that the propagation process is completed. In the case of the independent-cascade model this is after a finite number of steps. In the case of the dynamic propagation model, we consider the state of the system as the time t approaches infinity. Let $S(N_r, G)$ denote the expected number of infected nodes in G. The expectation is taken over all random choices made by the propagation model. Also, let

$$S_r(G) = \max_{N_r} S(N_r, G),$$

denote the maximum expected number of infected nodes, where the maximum is taken over all possible virus placements. The subset $A_r = \arg\max_{N_r} S(N_r, G)$ corresponds to choices of the adaptive adversary. We call $S_r(G)$ the *epidemic spread* in G. We now give the definitions of an *immunized* graph and an *epidemic outbreak*.

DEFINITION 1. *Let G be a graph with n nodes, r be a number of initial virus copies, and γ be a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$. We say that the graph G is* $\gamma(n, r)$-immunized *against an adaptive adversary A_r, if*

$$S_r(G) \leq \gamma(n, r).$$

Alternatively, we say that there is a $\gamma(n, r)$-epidemic outbreak *in the graph, if the graph is not $\gamma(n, r)$-immunized.*

We can give similar definitions for immunization and epidemic outbreak against a blind adversary. In this case we define

$$\widehat{S}_v(G) = E_{A_r}[S(A_r, G)],$$

to be the expected epidemic spread, where now the expectation is taken over all possible positions of the r viruses.

We now define the following basic immunization problem

PROBLEM 1 (IMMUNIZATION). *Given a graph G with n nodes, a number r of initial viruses, and a function $\gamma : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, find the minimum number k of nodes in G that need to be immunized, such that the resulting graph G' is $\gamma(n, r)$-immunized.*

4.1 The Independent Cascade Model

The characteristic of the independent cascade model is that each node has only one opportunity to infect its neighbors. Time in the model proceeds in discrete time steps. At time $t = 0$ the adversary plants r viruses to some nodes of the graph. Then, if a node i becomes infected for first time at time t it is given a single chance to infect each of its neighbors j that is currently uninfected. The probability that node i succeeds in infecting node j is p_{ij}, the (i, j) entry of the propagation probability matrix P. If node i succeeds in infecting j, then j becomes infected at time $t + 1$; otherwise node i never attempts to infect node j again in the future (eventhough j might eventually get infected by some of its other neighbors). The virus-propagation process continues until no more infections are possible — clearly the process stops after at most n steps.

We now define the immunization problems we will consider in this paper, with respect to the independent cascade model.

PROBLEM 2 (EPIDEMIC SPREAD MINIMIZATION). *Given a graph G, a number r of initial viruses, and a number k, immunize k nodes in G, such that the epidemic spread $S_r(G)$ in the resulting graph is minimized.*

PROBLEM 3 (EXPECTED EPIDEMIC SPREAD MINIMIZATION). *Given a graph G, a number r of initial viruses, and a number k, immunize k nodes in G, such that the expected epidemic spread $\widehat{S}_r(G)$ in the resulting graph is minimized.*

Both problems 2 and 3 are NP-Hard.

4.2 Dynamic Propagation Models

In this section we view virus propagation as a dynamical birth-death process that evolves over time. Viruses are continuously propagated in the network, but they may also die. In a discrete time model, a virus dies at node i in a single time step with probability p_i. We refer to p_i as the *healing probability* of node i. A virus is propagated from node i to node j with probability p_{ij}. In the continuous time approach, the healing and propagation probabilities are replaced by the corresponding rates of healing and propagation, respectively.

The discrete time version of this model was considered by Wang et al. [18] They simplify the model by setting all propagation probabilities to $p_{ij} = \beta$, and all healing probabilities to $p_i = \delta$. Then they consider a discrete time process that proceeds roughly as follows. At time $t = 0$ the adversary A_r plants r viruses in the network. At time step t, every node i that has the virus infects neighboring node j that does not have the virous with probability p_{ij}. At the same step, node i may become cured with probability p_i. The propagation and healing events may happen at any order with equal probability. We will refer to this model as the WCWF model.

Wang et al. [18] argue the following:

THEOREM 1 (WANG ET AL. [18]). *Given a graph G with adjacency matrix M, propagation probability β, and healing probability δ, the network G is 0-immunized in the* WCWF *model against any adaptive adversary A_r, $1 \leq r \leq n$, if $\lambda_1(M) < \delta/\beta$.*

We can refine the proof and prove the even stronger result:

THEOREM 2. *Let G be a graph, and let Δ be the system matrix of the model. The network G is 0-immunized, against any adaptive adversary A_r, $1 \leq r \leq n$, if and only if $\lambda_1(\Delta) < 1$.*

5. IMMUNIZATION STRATEGIES

5.1 The Independent Cascade Model

Minimizing the epidemic spread: In this section we discuss the proposed algorithms for network immunization and epidemic-spread containment. We start by describing the algorithm for Problem 2, EPIDEMIC SPREAD MINIMIZATION.

For the rest of this section we consider the case that k nodes are immunized against an adversary who places only one virus ($r = 1$). The basic ingredient of the algorithm, based on the observation of Kempe et al. [12], is to view the probabilistic process of virus propagation as sampling over all $2^{|E|}$ possible graphs according to the distribution defined by the probabilities p_{ij} and then run the deterministic-cascade model on the sampled graph. For completeness, we repeat the argument here: when a node i becomes infected, each currently uninfected neighbor j of i becomes also infected with probability p_{ij}. This process is equivalent with a process where each edge (i, j) is *live* in the graph with probability p_{ij}, and the virus is propagating only over the live edges. In turn, this is equivalent to sampling a graph X from the set of all subgraphs of G, where each edge (i, j) of G is present in X with probability p_{ij}, and then propagating the virus *deterministically* on the graph X.

Consider a graph X sampled from G as described in the previous paragraph. Assume that the adversary decides to place the single copy of the virus at the node u of the graph, and define $s(\{u\}, X)$ to be the number of infected nodes (the epidemic spread in X) when the virus starts at node u. Clearly, $s(\{u\}, X)$ is the size of the connected component of X that u belongs to; since we established that the virus spreads deterministically in X, all nodes that are reachable from u will be infected. Then, the expected epidemic spread $S(\{u\}, G)$ in G with initial placement of the virus at node u can be expressed as

$$S(\{u\}, G) = \sum_X Pr[X] s(\{u\}, X), \tag{1}$$

where $Pr[X]$ is the probability of obtaining graph X from G when sampling according to edge probabilities p_{ij}. Therefore, given a graph G, we can estimate $S(\{u\}, G)$ using Equation (1): sample graphs X from G according to the distribution defined by the probabilities p_{ij} and compute the expected size of the connected component that u belongs to.

Let $G|_{w_1,w_2,...}$ be the graph resulting from G after immunizing the nodes $w_1, w_2, \ldots$. To describe our algorithm, assume first that we want to immunize only one node ($k = 1$). For all candidate nodes w_1 to be immunized, we can compute the value of $S(G|_{w_1}) = \max_u S(\{u\}, G|_{w_1})$, which is the worst-case (over all possible initial virus placements) expected epidemic spread if the node w_1 is chosen to be immunized. Then we choose to immunize the node w that minimizes the epidemic spread $S(G|_w)$. In the case that we want to immunize $k > 1$ nodes we proceed in a *greedy* fashion: we first immunize the node w_1 that minimizes $S(G|_{w_1})$. Then we find the best node w_2 to be immunized in the graph $G|_{w_1}$, that is, we find the node w_2 that minimizes $S(G|_{w_1,w_2})$ given the choice of w_1 in the previous step, and we continue until we select k nodes. We call this algorithm GREEDY.

Running time: Overall, the running time of GREEDY in order to immunize k nodes with Q samples per iteration is $O(Q(n^2 + nm)k)$.

Minimizing the expected epidemic spread: We now discuss the modifications needed in the above GREEDY algorithm in order to address the problem of minimizing the expected epidemic spread in the network. Recall that this problem corresponds to the case of a blind adversary who plants viruses uniformly at random in the graph. The GREEDY algorithm for EXPECTED EPIDEMIC SPREAD

MINIMIZATION is similar to the one we described before. Nodes are immunized one at a time until a total of k nodes are selected. To select the i-th node w_i we sample graphs X from the current graph $G' = G|_{w_1,\ldots,w_{i-1}}$ and for each sample X we compute $\widehat{s}(X|_w)$ for all possible candidate nodes w. Then, using the equation $\widehat{S}(G'|_w) = \sum Pr[X]\widehat{s}(X|_w)$, which is analogous to Equation (1), we can estimate $\widehat{S}(G'|_w)$ for all nodes w and select to immunize the one that yields the smallest value. For one sample graph X, we can compute $\widehat{s}(X|_w)$ for all nodes w in time $O(n + m)$ using DFS. Thus, the overall running time of the algorithm is $O(Q(n + m)k)$, where Q is the number of samples at each iteration of the GREEDY.

5.2 Dynamic Propagation Models

Following the discussion in Section 4 the spread of an epidemic depends on the largest eigenvalue of the adjacency matrix M. Therefore, the objective of the immunization algorithm is to decrease this eigenvalue, while incurring the minimum possible damage to the network.

The algorithm we propose, named EIG is the following. Proceed in iterations, where each iteration takes as input a matrix B. For the first iteration, $B = \Delta$ the system matrix (where $\Delta = M$ the adjacency matrix when all propagation, and healing probabilities are equal). Compute the largest eigenvalue λ_1 and the corresponding eigenvector $\boldsymbol{w}_1$ of matrix B. If $\lambda_1 < 1$ the algorithm stops. Otherwise, find the node i with the maximum value in the eigenvector $\boldsymbol{w}_1$, and remove it from the graph, that is, remove the corresponding row and column from matrix B. The resulting matrix will be given as input to the next iteration. The running time of the algorithm is $O(kT)$, where k is the number of nodes removed, and T is the time to compute the first eigenvalue and eigenvector matrix. If the graph is sparse this can usually be done in time proportional to the edges of the graph.

6. EXPERIMENTS

In this section we compare the performance of the proposed immunization strategies against other natural heuristic strategies. We also explore how the underlying graph structure affects the relative performance of the various algorithms. For the experimental evaluation we use both synthetic and real datasets. We will demonstrate that our algorithms out-perform the other heuristics. We will also demonstrate that our algorithms perform better for graphs with high clustering coefficient, thus being more appropriate for "small world" networks. Here, we present a subset of the experiments we did due to lack of space.

6.1 Datasets

We next describe the main characteristics of the graph types we generated and the model that we use for the graph generation.

Random graphs: Random graphs are graphs with nodes randomly connected to each other with probability p. Given the number of nodes n and the parameter p, a random graph is generated by creating an edge between each pair of nodes u and v with probability p.

"Small-world" graphs: A graph is said to be a small-world network [19] if it has small *characteristic path length* L, and large *clustering coefficient* C. The characteristic path length L is defined as the average shortest path between any two pairs of vertices in the graph. The clustering coefficient C is a measure of the local graph structure, and it is defined as the average fraction of pairs of neighbors of a node that are also connected to each other. We generate "small-world" graphs using the generating model proposed in [19]. We use $\mathcal{G}_W$ to denote the family of graphs generated by this model. The generation process of these graphs is governed by a parameter α. Intuitively, α determines the probability that two nodes will be connected, given the number of their common neighbors, thus, it controls to which extent the graph will contain communities. For small values of α the graph has small and densely connected components. As α approaches infinity, the generated graphs become random graphs. There is a range of values of α for which the generated graphs exhibit the "small-world" property. Figure 1 shows the characteristic path length and the clustering coefficient of a graph $G \in \mathcal{G}_W$ with 4000 nodes as a function of the parameter α. For values of α between 5 and 8 the generated graph exhibits small average path lengths and high clustering coefficient, and thus it is a small world graph.

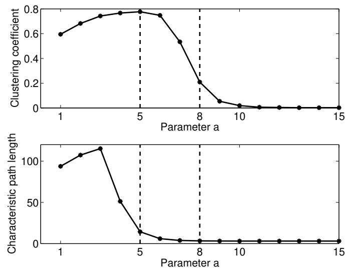


Figure 1: C and L for a graph in $\mathcal{G}_W$ as a function of α

In addition to the synthetic datasets we also experiment on real graphs: the *co-authors* graph.

Co-author graph: The co-authors dataset consists of 8000 authors of papers in VLDB, PODS and SIGMOD available at the Collection of Computer Science Bibliographies[1]. The co-author graph is constructed by creating undirected edges between authors that have been co-authors in the same paper. We think of the co-authors dataset, as representative of a social network.

6.2 Experiments with the Independent Cascade Model

In this section we evaluate the proposed immunization strategies for the independent cascade model. We compare the GREEDY algorithms with some natural and straightforward heuristics. Two of them, MAXDEGREE and MAXDEGREEIT, select the nodes to immunize based on their degrees. The MAXDEGREE algorithm immunizes the nodes of the graph in decreasing order of their degree in the input graph. The MAXDEGREEIT immunizes nodes also based on their degree, but at every step it considers immunizing the node with the highest degree in the graph induced by the changes completed in the previous immunization steps. Therefore, the two algorithms do the same choice in their first step. We also compare with the SORT algorithm which is defined as follows. Initially, the algorithm computes for each node in the network the gain achieved by removing only this node from the network. It then sorts the nodes in decreasing order of that gain, and proceeds by removing the nodes in that order. Finally, we compare with the algorithm RANDOM, which at every step selects a random node to immunize. Throughout the experimental section if the results for RANDOM are omitted is because the algorithm performs incomparably bad.

For realizing the independent cascade model we assign a uniform probability p to each edge of the graph. We experiment with different values for the parameter p and report the results. In all

[1] http://liinwww.ira.uka.de/bibliography

cases we consider as a measure for the performance of the algorithms the epidemic spread S, or the expected epidemic spread $\widehat{S}$, depending on whether we consider the spread minimization or the expected spread minimization problem.

We first experiment with real datasets, and we consider the co-authors graph.Figures **??** shows the performance of the different algorithms as a function of the number of nodes being immunized, for both the spread minimization, and the expected spread minimization problem. The parameter p is set to value 1 and 0.6, however, experiments with other values of the parameter exhibit similar trends. In all the experiments we conducted, including the ones shown in Figure **??** the GREEDY algorithm consistently outperformed the other heuristics. We also noticed that the SORT algorithm tends to make good choices, although it does not take into account the changes in the underlying graph structure when deciding upon the next node to immunize.

An important observation at this point is that the relative performance of the algorithms varies in the two datasets. In the co-authors graph the GREEDY algorithm performs clearly better than the other methods. In the AS graph however, this difference is not that striking. The clustering coefficient of the co-authors graph is $C_{CO} = 0.64$. This raises the question of how the structural properties of the graph to be immunized, affect the performance of the different algorithms.

In order to further investigate this question we consider graphs from the family $\mathcal{G}_W$. We fix the number of nodes ($n = 4000$), and we consider three different values for the parameter $\alpha = 3, 6, 14$. Notice that the choice of the parameter values is not arbitrary. We set the value of α so that we obtain three different types of graphs, with different structural properties. As it was already illustrated in Figure 1, for $\alpha = 3$ the generated graphs have high clustering coefficient and large average path length. For $\alpha = 6$ they exhibit small-world properties and for $\alpha = 14$ they are similar to random graphs. The plots in Figure 3 show the expected epidemic spread as a function of the number of immunized nodes for the different values of α. From those plots it is apparent that the superiority of the GREEDY algorithm is much more strongly pronounced for the small values of α. Thus, the GREEDY algorithm is significantly more efficient than the other heuristics when the graph has high clustering coefficient. This includes also the case when the characteristic path is small, that is, the graph has the small world property, although the relative performance of the algorithm does not seem to depend on the characteristic path length. An intuitive explanation for this is that for small values of α, the $\mathcal{G}_W$ graphs have lots of locally dense components, sparsely interconnected, and thus our algorithm can easily identify then interconnecting edges and disconnect them. For large values of α, where the graph is expected to be almost random, the performance of the GREEDY algorithm is only a slight improvement over the rest of the algorithms. In this case, all other algorithms, including the RANDOM algorithm, perform almost the same. This is expected, since there is no clear structure in the graph that the immunization algorithm can take advantage of. Similar results are obtained when we consider epidemic spread instead of expected epidemic spread.

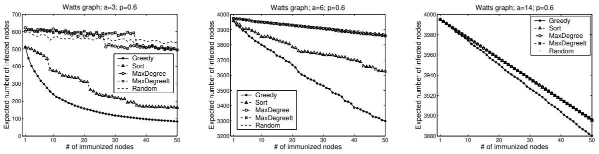


Figure 3: Expected epidemic spread for graphs in $\mathcal{G}_W$

6.3 Experiments with the Dynamic Propagation Model

In this section we evaluate the performance of our immunization strategies for the dynamic propagation model. We again compare our proposed algorithm EIG with the MAXDEGREE and MAXDEGREEIT heuristics, as well as, with RANDOM. All of them are defined in the same way as in the previous subsection. We also consider the BATCH algorithm, a faster variant of the EIG algorithm that processes nodes in batches. At every step, the BATCH algorithm removes the ℓ nodes with the highest value in the first eigenvector of matrix P. When $\ell = 1$, we have the EIG algorithm. In our experiments we set $\ell = 2$. This means that BATCH cuts the computational requirements of the EIG algorithm into half. Experimental evidence shows that using $\ell = 2$ does not degrade the performance of the algorithm. For values $\ell > 2$, we observed a significant drop in the performance.

Recall that in the dynamic propagation model, the objective of the algorithm is to drop the eigenvalue of the matrix below the epidemic threshold. In our experiments, we assume that the threshold is given as input to the algorithm, and it does not depend on the size of the graph. The performance of the algorithm is measured by the number of nodes it removes from the graph, in order to bring the eigenvalue of the matrix below the threshold.

We begin the experimental evaluation with the real datasets, where we observe the same trends as in the independent cascade model. That is, the EIG algorithm performs always better than the max-degree heuristics. Figure 4 shows the performance of the different algorithms for the co-authors and the AS graph[2], when varying the threshold values. The EIG algorithm achieves the required drop in the eigenvalue by removing the least number of nodes from the graph. The relative advantage of the EIG algorithm is more clearly observed in the co-authors graph. Coupling this observation with the already observed difference in the clustering coefficients of the two real graphs, we again proceed in exploring how the structural properties of the graph influence the performance of the different algorithms.

In order to explore this we design an experiment similar to the one we did for the independent cascade model. That is, we generate graphs from $\mathcal{G}_W$ family, each consisting of 4000 nodes, for 15 different integer values of the parameter α from 1 to 15. The values of the clustering coefficient and the characteristic path length follow the curves shown in Figure 1. Figure 5(a) shows the performance of the various algorithms as a function of the parameter α, for two different thresholds. The results at each point are averages over three independently generated graphs. The first observation is that for large values of α, that is, as the graphs become close to random graphs, all the algorithms (including RANDOM) tend to behave the same way. For small values of α however, the EIG algorithm performs significantly better than the other heuristics, independently of the size of the characteristic path length in the underlying graph. This is not the case for MAXDEGREE and MAXDEGREEIT algorithms that are apparently affected by both the clustering coefficient and the average path length.

We also experiment with a set graphs of different sizes from $\mathcal{G}_W$ families. The sizes of the graphs are set to be 1000, 2000, 4000, 8000 and 16000 nodes. For those graphs, we fix a threshold, and we report the number of nodes that need to be removed from each graph in order for the eigenvalue to drop below the threshold. Apparently, the threshold is chosen in such a way that it is lower than the lowest existing eigenvalue in the set of input graphs. In that

[2]Here we use the same AS graph as the one used for the independent cascade model experiments

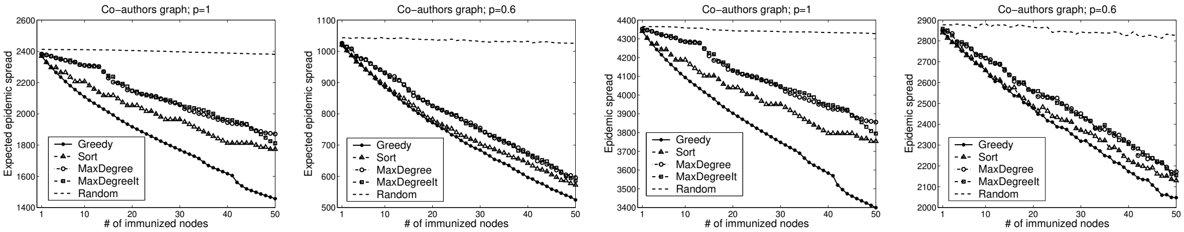


Figure 2: (Expected)Epidemic spread for real graphs

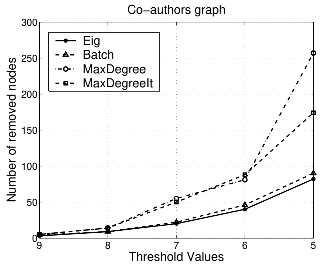
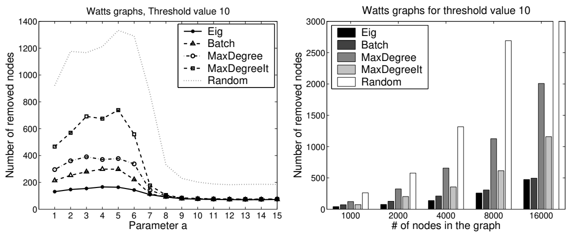


Figure 4: Experiments with real graphs and different thresholds



(a) (b)

Figure 5: Experiments with $\mathcal{G}_W$ graphs

case the threshold is set to 10, while the eigenvalues in the input graphs take values between 11.1 and 11.4. The results obtained for $\mathcal{G}_W$ are shown in Figure 5(b). Each bar in the plot is an average over three independent runs. It is again obvious that for $\mathcal{G}_W$ graphs the superiority of the EIG algorithm is striking.

7. CONCLUSIONS

In this paper, we consider the problem of network immunization against a virus spread. We study the immunization problem under two different models for virus propagation. For the independent cascade model we show that the problem of minimizing the spread both against an adaptive and a blind adversary is NP-hard. The greedy algorithm we propose performs better than the natural heuristic of immunizing the highest-degree nodes. For the case of dynamic propagation models, we give a rigorous and we provide necessary and sufficient conditions for an epidemic outbreak in the network. For this problem we propose a simple heuristic, which is based on intuition drawn from linear algebra. We experimentally show that the algorithm performs extremely well in practice and much better than degree-related heuristics.

8. REFERENCES

[1] R. M. Anderson and R. M. May. *Infectious diseases in humans*. 1992.

[2] J. Aspnes, K. Chang, and A. Yampolskiy. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. In *SODA*, 2005.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[4] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A.Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 2005.

[5] M. Boguna and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E 66*, 2002.

[6] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. *Statistical Mechanics of Complex Networks*, 2003.

[7] F. Cohen. Computer viruses, theory and experiments. *Computers & Security*, 1987.

[8] R. Cohen, S. Havlin, and D. ben Avraham. Efficient immunization strategies for computer networks and populations. *Phys Rev Lett.*, 2003.

[9] Z. Dezso and A.-L. Barabasi. Halting viruses in scale-free networks. *Phys. Rev. E 66*, 2002.

[10] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[11] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing*, 2001.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, 2003.

[13] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. *IEEE Computer Society Symposium on Research in Security and Privacy*, 1991.

[14] J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. *IEEE Computer Society Symposium on Research in Security and Privacy*, 1993.

[15] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. Roy. Soc. Lond.*, 1927.

[16] W. H. Murray. The application of epidemiology to computer viruses. *Computers & Security*, 1988.

[17] R. Pastor-Satorras and A. Vespignani. Epidemics and immunization in scale-free networks. *Handbook of Graphs and Networks: From the Genome to the Internet*, 2002.

[18] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *SRDS*, 2003.

[19] D. J. Watts. Networks, dynamics and the small world phenomenon. *American Journal of Sociology*, 105, 1999.