Minimizing the Spread of Contamination by Blocking Links in a Network

Masahiro Kimura
Deptartment of Electronics and Informatics
Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp

Kazumi Saito
School of Administration and Informatics
University of Shizuoka
Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp

Hiroshi Motoda
Institute of Scientific and Industrial Research
Osaka University
Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract

We address the problem of minimizing the propagation of undesirable things, such as computer viruses or malicious rumors, by blocking a limited number of links in a network, a dual problem to the influence maximization problem of finding the most influential nodes in a social network for information diffusion. This minimization problem is another approach to the problem of preventing the spread of contamination by removing nodes in a network. We propose a method for efficiently finding a good approximate solution to this problem based on a naturally greedy strategy. Using large real networks, we demonstrate experimentally that the proposed method significantly outperforms conventional link-removal methods. We also show that unlike the strategy of removing nodes, blocking links between nodes with high out-degrees is not necessarily effective.

Introduction

Considerable attention has recently been devoted to investigating the structure and function of various networks including computer networks, social networks and the World Wide Web (Newman 2003). From a functional point of view, networks can mediate diffusion of various things such as innovation and topics. However, undesirable things can also spread through networks. For example, computer viruses can spread through computer networks and email networks, and malicious rumors can spread through social networks among individuals. Thus, developing effective strategies for preventing the spread of undesirable things through a network is an important research issue. Previous work studied strategies for reducing the spread size by removing nodes from a network. It has been shown in particular that the strategy of removing nodes in decreasing order of out-degree can often be effective (Albert, Jeong, and Barabási 2000; Broder et al. 2000; Callaway et al. 2000; Newman, Forrest, and Balthrop 2002). Here notice that removal of nodes by necessity involves removal of links. Namely, the task of removing links is more fundamental than that of removing nodes. Therefore, preventing the spread of undesirable things by removing links from the underlying network is an important problem.

In contrast, finding influential nodes that are effective for the spread of information through a social network is also an important research issue in terms of sociology and "viral marketing" (Domingos and Richardson 2001; Richardson and Domingos 2002; Gruhl et al. 2004). Thus, researchers (Kempe, Kleinberg, and Tardos 2003; Kimura, Saito, and Nakano 2007) have recently studied a combinatorial optimization problem called the *influence maximization problem* on a network under the *independent cascade (IC) model*, a widely-used fundamental probabilistic model of information diffusion. Here, the influence maximization problem is the problem of extracting a set of k nodes to target for initial activation such that it yields the largest expected spread of information, where k is a given positive integer. Note also that the IC model can be identified with the so-called *susceptible/infective/recoverd (SIR) model* for the spread of disease in a network (Gruhl et al. 2004).

The problem we address in this paper is a dual problem to the influence maximization problem. The problem is to minimize the spread of undesirable things by blocking a limited number of links in a network. More specifically, when some undesirable thing starts with any node and diffuses through the network under the IC model, we consider finding a set of k links such that the resulting network by blocking those links minimizes the expected contamination area of the undesirable thing, where k is a given positive integer. We refer to this combinatorial optimization problem as the *contamination minimization problem*. For this problem, we propose a novel method for efficiently finding a good approximate solution on the basis of a naturally greedy strategy. Using large real networks including a blog network, we experimentally demonstrate that the proposed method significantly outperforms link-removal heuristics that rely on the well-studied notions of betweenness and out-degree. In particular, we show that unlike the case of removing nodes, blocking links between nodes with high out-degrees is not necessarily effective for our problem.

Problem Formulation

In this paper, we address the problem of minimizing the spread of undesirable things such as computer viruses and malicious rumors in a network represented by a directed graph $G = (V, E)$. Here, V and E ($\subset V \times V$) are the sets of all the nodes and links in the network, respectively. We

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

assume the IC model to be a mathematical model for the diffusion process of some undesirable thing in the network, and investigate the contamination minimization problem on G. We call nodes *active* if they have been contaminated by the undesirable thing.

Independent Cascade Model

We define the IC model on graph G according to the work of Kempe, Kleinberg, and Tardos (2003).

In the IC model, the diffusion process unfolds in discrete time-steps $t \geq 0$, and it is assumed that nodes can switch their states only from inactive to active, but not from active to inactive. Given an initial active node v, we assume that the node v has first become active at time-step 0, and all the other nodes are inactive at time-step 0. We specify a real value p with $0 < p < 1$ in advance. Here, p is referred to as the *propagation probability* through a link. The diffusion process proceeds from the initial active node v in the following way. When a node u first becomes active at time-step t, it is given a single chance to activate each currently inactive child node w, and succeeds with probability p. If u succeeds, then w will become active at time-step $t+1$. If multiple parent nodes of w first become active at time-step t, then their activation attempts are sequenced in an arbitrary order, but all performed at time-step t. Whether or not u succeeds, it cannot make any further attempts to activate w in subsequent rounds. The process terminates if no more activations are possible.

For an initial active node v, let $\sigma(v; G)$ denote the expected number of active nodes at the end of the random process of the IC model on G. We call $\sigma(v; G)$ the *influence degree* of node v in graph G.

Contamination Minimization Problem

Now, we give a mathematical definition of the contamination minimization problem on graph $G = (V, E)$. For preventing the undesirable thing from spreading through the network under the IC model, we aim to minimize the expected contamination area (that is, the expected number of active nodes) by appropriately removing a fixed number of links.

First, we define the *contamination degree* $c(G)$ of graph G as the average of influence degrees of all the nodes in G, that is,

$$c(G) = \frac{1}{|V|} \sum_{v \in V} \sigma(v; G). \tag{1}$$

Here, $|A|$ stands for the number of elements of a set A. For any link $e \in E$, let $G(e)$ denote the graph $(V, E \setminus \{e\})$. We refer to $G(e)$ as the graph constructed by *blocking* e in G. Similarly, for any $D \subset E$, let $G(D)$ denote the graph $(V, E \setminus D)$. We refer to $G(D)$ as the graph constructed by *blocking* D in G. We define the *contamination minimization problem* on graph G as follows: Given a positive integer k with $k < |E|$, find a subset D^* of E with $|D^*| = k$ such that $c(G(D^*)) \leq c(G(D))$ for any $D \subset E$ with $|D| = k$.

For a large network, any straightforward method for exactly solving the contamination minimization problem suffers from combinatorial explosion. Therefore, we consider approximately solving the problem.

Proposed Method

We propose a method for efficiently finding a good approximate solution to the contamination minimization problem on graph $G = (V, E)$. Let k be the number of links to be blocked in this problem.

Geedy Algorithm

We approximately solve the contamination minimization problem on $G = (V, E)$ by the following greedy algorithm:

```
1. Set $D_0 \leftarrow \emptyset$.
2. Set $E_0 \leftarrow E$.
3. Set $G_0 \leftarrow G$.
4. **for** $i = 0$ to $k - 1$ **do**
5.    Choose a link $e_* \in E_i$ minimizing $c(G_i(e))$, $(e \in E_i)$.
6.    Set $D_{i+1} \leftarrow D_i \cup \{e_*\}$.
7.    Set $E_{i+1} \leftarrow E_i \setminus \{e_*\}$.
8.    Set $G_{i+1} \leftarrow (V, E_{i+1})$.
9. **end for**
```

Here, D_k is the set of links blocked, and represents the approximate solution obtained by this algorithm. G_k is the graph constructed by blocking D_k in graph G, that is, $G_k = G(D_k)$.

To implement this greedy algorithm, we need a method for calculating $\{c(G_i(e));\ e \in E_i\}$ in Step 5 of the algorithm. However, the IC model is a stochastic process model, and it is an open question to exactly calcluate influence degrees by an efficient method (Kempe, Kleinberg, and Tardos 2003). Therefore, we develop a method for estimating $\{c(G_i(e)); e \in E_i\}$.

Kimura, Saito, and Nakano (2007) presented the bond percolation method that efficiently estimates the influence degrees $\{\sigma(v; \tilde{G});\ v \in V\}$ for any directed graph $\tilde{G} = (V, \tilde{E})$. Thus, we can estimate $c(G_i(e))$ for each $e \in E_i$ by straightforwardly applying the bond percolation method. However, $|E_i|$ becomes very large for a large network unless i is very large. Therefore, we propose a method that can estimate $\{c(G_i(e));\ e \in E_i\}$ in a more efficient manner on the basis of the bond percolation method.

Bond Percolation Method

First, we revisit the bond percolation method (Kimura, Saito, and Nakano 2007). Here, we consider estimating the influence degrees $\{\sigma(v; G_i); v \in V\}$ for the IC model with propagation probability p in graph $G_i = (V, E_i)$.

It is known that the IC model is equivalent to the bond percolation process that independently declares every link of G_i to be "occupied" with probability p (Newman 2003). Let M be a sufficiently large positive integer. We perform the bond percolation process M times, and sample a set of M graphs constructed by the occupied links,

$$\{{G_i}^m = (V, {E_i}^m);\ m = 1, \cdots, M\}.$$

Then, we can approximate the influence degree $\sigma(v; G_i)$ by

$$\sigma(v; G_i) \simeq \frac{1}{M} \sum_{m=1}^{M} |\mathcal{F}(v; {G_i}^m)|.$$

Here, for any directed graph $\tilde{G} = (V, \tilde{E})$, $\mathcal{F}(v; \tilde{G})$ denotes the set of all the nodes that are *reachable* from node v in the graph. We say that node u is reachable from node v if there is a path from u to v along the links in the graph. Let

$$V = \bigcup_{u \in \mathcal{U}(G_i{}^m)} \mathcal{S}(u; G_i{}^m)$$

be the strongly connected component (SCC) decomposition of graph $G_i{}^m$, where $\mathcal{S}(u; G_i{}^m)$ denotes the SCC of $G_i{}^m$ that contains node u, and $\mathcal{U}(G_i{}^m)$ stands for a set of all the representative nodes for the SCCs of $G_i{}^m$. The bond percolation method performs the SCC decomposition of each $G_i{}^m$, and estimates all the influence degrees $\{\sigma(v; G_i);\ v \in V\}$ in G_i as follows:

$$\sigma(v; G_i) = \frac{1}{M} \sum_{m=1}^{M} |\mathcal{F}(u; G_i{}^m)|\,, \quad (v \in \mathcal{S}(u; G_i{}^m))\,, \quad (2)$$

where $u \in \mathcal{U}(G_i{}^m)$.

Estimation Method

We are now in a position to give a method for efficiently estimating $\{c(G_i(e));\ e \in E_i\}$ in Step 5 of the greedy algorithm. We develop such an estimation method on the basis of the bond percolation method.

For any directed graph $\tilde{G} = (V, \tilde{E})$, we define $\varphi(\tilde{G})$ by

$$\varphi(\tilde{G}) = \frac{1}{|V|} \sum_{v \in V} \left| \mathcal{F}(v; \tilde{G}) \right|. \quad (3)$$

Using the bond percolation method, we consider estimating the contamination degree $c(G_i)$ of the graph $G_i = (V, E_i)$. Then, by Equations (1), (2) and (3), we can estimate $c(G_i)$ as

$$c(G_i) = \frac{1}{M} \sum_{m=1}^{M} \varphi(G_i{}^m). \quad (4)$$

Here, note that $\varphi(G_i{}^m)$ is calculated by

$$\varphi(G_i{}^m) = \frac{1}{|V|} \sum_{u \in \mathcal{U}(G_i{}^m)} |\mathcal{F}(u; G_i{}^m)|\ |\mathcal{S}(u; G_i{}^m)|\,. \quad (5)$$

We assume that M is sufficiently large. Then, by the independence of the bond percolation process, we can estimate $c(G_i(e))$ for every $e \in E_i$ as

$$c(G_i(e)) = \frac{1}{|\mathcal{M}_i(e)|} \sum_{m \in \mathcal{M}_i(e)} \varphi(G_i{}^m), \quad (6)$$

where $G_i{}^m = (V, E_i{}^m)$, and

$$\mathcal{M}_i(e) = \{m \in \{1, \cdots, M\};\ e \notin E_i{}^m\}\,.$$

We efficiently estimate $\{c(G_i(e));\ e \in E_i\}$ by Equations (5) and (6) without applying the bond percolation method for every $e \in E_i$. Namely, the proposed method can achieve a great deal of reduction in computational cost compared with the coventional bond percolation method.

Experimental Evaluation

Using two large real networks, we experimentally evaluated the performance of the proposed method.

Network Datasets

First, we employed a trackback network of blogs because a piece of information can propagate from one blog author to another blog author through a trackback. Since bloggers discuss various topics and establish mutual communications by putting trackbacks on each other's blogs, we regarded a link created by a trackback as a biderectional link for simplicity. By tracing up to ten steps back in the trackbacks from the blog of the theme "JR Fukuchiyama Line Derailment Collision" in the site "goo" (`http://blog.goo.ne.jp/usertheme/`), we collected a large connected trackback network in May, 2005. This network was a directed graph of $12,047$ nodes and $79,920$ links. We refer to this network data as the blog network.

Next, we employed a network of people that was derived from the "list of people" within Japanese Wikipedia. Specifically, we extracted the maximal connected component of the undirected graph obtained by linking two people in the "list of people" if they co-occur in six or more Wikipedia pages, and constructed a directed graph regarding those undirected links as bidirectional ones. We refer to this network data as the Wikipedia network. Here, the total numbers of nodes and directed links were $9,481$ and $245,044$, respectively.

Note here that these two networks are strongly connected.

Experimental Settings

The IC model has the propagation probability p as a parameter. So we determine the typical values of p for the blog and Wikipedia networks, and use them in the experiments. Let us consider the bond percolation process corresponding to the IC model with propagation probability p in graph $G = (V, E)$. Let S be the expected fraction of the maximal SCC in the network constructed by occupied links. S is a function of p, and as the value of p decreases, the value of S decreases. In other words, as the value of p decreases, the original graph G gradually fragments into small clusters under the corresponding bond percolation process. Figures 1 and 2 show the network fragmentation curves for the blog and Wikipedia networks, respectively. Here, we estimated S as follows:

$$S = \frac{1}{M} \sum_{m=1}^{M} \max_{u \in \mathcal{U}(G_i{}^m)} |\mathcal{S}(u; G_i{}^m)|,$$

where $M = 10000$. We focus on the point p_* at which the average rate dS/dp of change of S attains the maximum, and regard it as the typical value of p for the network. Note that p_* is a critical point of dS/dp, and defines one of the features intrinsic to the network. From Figures 1 and 2, we estimated p_* to be $p_* = 0.2$ for the blog network and $p_* = 0.03$ for the Wikipedia network.

For the proposed method, we need to specify the number M of performing the bond percolation process. In the experiments, we used $M = 10000$.

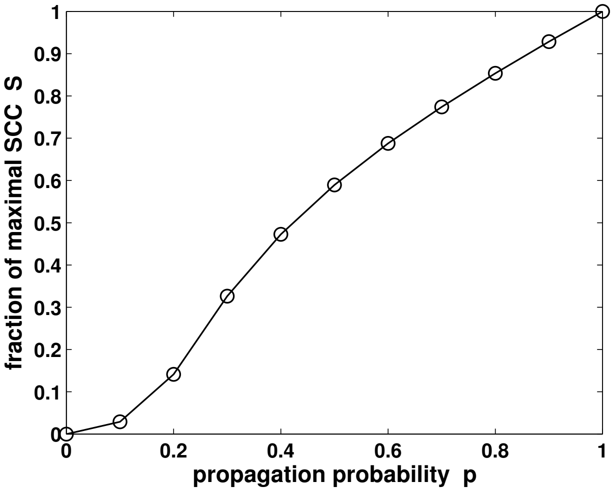

Figure 1: Fragmentation of the blog network for the IC model. The fraction S of the maximal SCC as a function of the propagation probablity p.

Comparison Methods

We compared the proposed method with two heuristics based on the well-studied notions of betweenness and out-degree in the field of complex network theory, as well as the crude baseline of blocking links at random. We refer to the method of blocking links uniformly at random as the *random method.*

The *betweenness score* $b_{\tilde{G}}(e)$ of a link e in a directed graph $\tilde{G} = (V, \tilde{E})$ is defined as follows:

$$b_{\tilde{G}}(e) = \sum_{u,v \in V} \frac{n_{\tilde{G}}(e; u, v)}{N_{\tilde{G}}(u, v)},$$

where $N_{\tilde{G}}(u, v)$ denotes the number of the shortest paths from node u to node v in $\tilde{G}$, and $n_{\tilde{G}}(e; u, v)$ denotes the number of those paths that pass e. Here, we set $n_{\tilde{G}}(e; u, v)/N_{\tilde{G}}(u, v) = 0$ if $N_{\tilde{G}}(u, v) = 0$. Newman and Girvan (2004) successfully extracted community structure in a network using the following link-removal algorithm based on betweeness:

1. Calculate betweenness scores for all links in the network.
2. Find the link with the highest score and remove it from the network.
3. Recalculate betweenness scores for all remaining links.
4. Repeat from Step 2.

In particular, the notion of betweenness can be interpreted in terms of signals traveling through a network. If signals travel from source nodes to destination nodes along the shortest paths in a network, and all nodes send signals at the same constant rate to all others, then the betweenness score of a link is a measure of the rate at which signals pass along the link. Thus, we naively expect that blocking the links with the highest betweenness score can be effective for preventing the spread of contamination in the network. Therefore, we apply the method of Newman and Girvan (2004) to the contamination minimization problem. We refer to this method as the *betweenness method.*

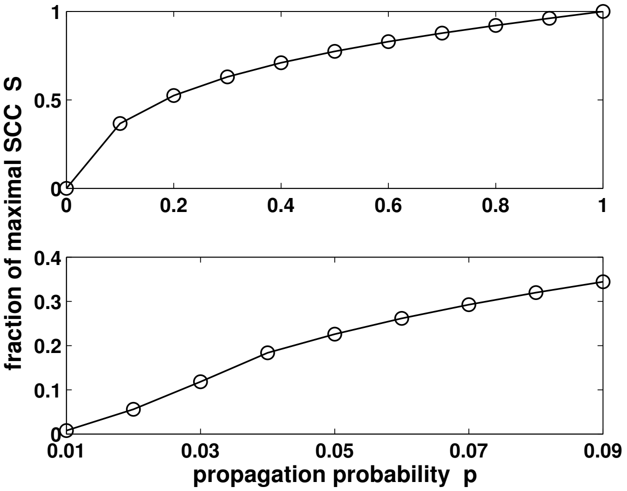

Figure 2: Fragmentation of the Wikipedia network for the IC model. The fraction S of the maximal SCC as a function of the propagation probablity p. The upper and lower frames show the network fragmentation curves for the whole range of p and the range of $0.01 \leq p \leq 0.09$, respectively.

On the other hand, previous work has shown that simply removing nodes in order of decreasing *out-degrees* works well for preventing the spread of contamination in most real networks (Albert, Jeong, and Barabási 2000; Broder et al. 2000; Callaway et al. 2000; Newman, Forrest, and Balthrop 2002). Here, the out-degree $d(v)$ of a node v means the number of outgoing links from the node v. Thus, blocking links between nodes with high out-degrees looks promising for the contamination minimization problem. Therefore, as a comparsion method, we employ the method of recursively blocking links $e = [u, v]$ from u to v in decreasing order of their scores $\overline{d}(e)$,

$$\overline{d}(e) \;=\; d(u)\, d(v).$$

We refer to this method as the *out-degree method.*

Experimental Results

We evaluated the performance of the proposed method and compare it with that of the betweenness, out-degree and random methods. Clearly, the performance of a method for solving the contamination minimization problem can be evaluated in terms of contamination degree c. We used the value of c (see Equations (4) and (5)) that is estimated by the bond percolation method with $M = 10000$.

Figures 3 and 4 show the contamination degree c of the resulting network as a function of the number k of links blocked for the blog network, where the circles, triangles, diamonds and squares indicate the results for the proposed, betweenness, out-degree and random methods, respectively.

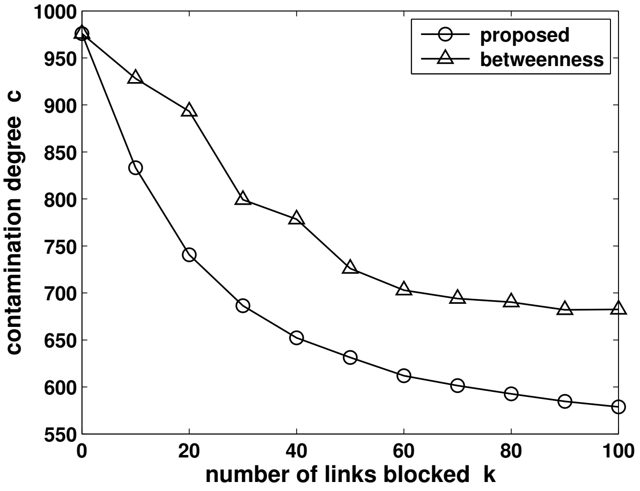

Figure 3: Performance comparison between the proposed and betweenness methods in the blog network for the IC model with $p = 0.2$.

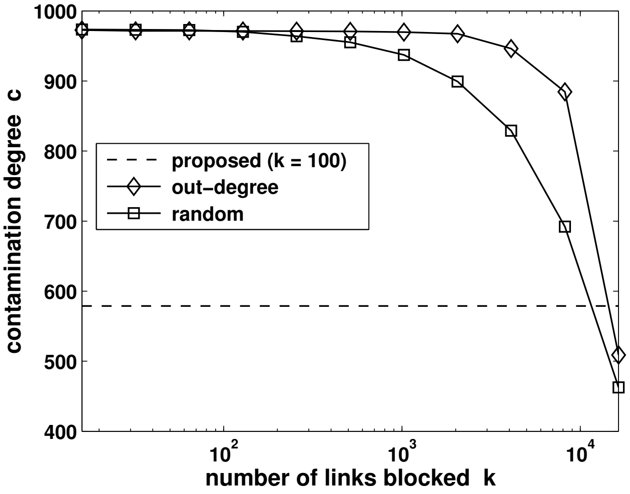

Figure 4: Performance comparison of the proposed method for $k = 100$ with the out-degree and random methods in the blog network for the IC model with $p = 0.2$.

In Figure 4, the dashed line indicates the contamination degree of the network obtained by the proposed method for $k = 100$. From Figures 3 and 4, we first see that the proposed method outperformed the betweenness, out-degree and random methods for the blog network. By taking into account the definition (1) of contamination degree, we can mention from Figure 3 that the proposed method decreased the expected number of nodes contaminated from about 980 nodes to about 580 nodes by blocking appropriate 100 links for the blog network. Here note that blocking 100 links means blocking about 0.13% of the links in the blog network. Thus, we find from Figure 3 that by appropriately blocking about 0.13% of the links in the blog network, the proposed and betweenness methods decreased contamination degree by about 41% and 30%, respectively. Hence, we can deduce that the proposed method was effective, and also outperformed the betweennes method by over 10% at $k = 100$ for the blog network. Moreover, we find from Figure 4 that blocking 100 links by using the proposed method was the same as blocking over 10000 links by using the out-degree and random methods for the blog network in effect. Namely, we can deduce that the proposed method was 100 times more effective than the out-degree and random methods at $k = 100$ for the blog network.

Figures 5 and 6 display the contamination degree c of the resulting network as a function of the number k of links blocked for the Wikipedia network. Here, as in Figures 3 and 4, the circles, triangles, diamonds and squares indicate the results for the proposed, betweenness, out-degree and random methods, respectively. In Figure 6, the dashed line indicates the contamination degree of the network obtained by the proposed method for $k = 300$. We also see from Figures 5 and 6 that the proposed method outperformed the betweenness, out-degree and random methods for the Wikipedia network. In particualr, we observe from Figure 5 that as the value of k increased, the performance difference between the proposed and betweenness methods gradually increased. Note here that blocking 300 links means blocking about 0.12% of the links in the Wikipedia network. Thus, we find from Figure 5 that by appropriately blocking about 0.12% of the links in the Wikipedia network, the proposed and betweenness methods decreased contamination degree by about 26% and 16%, respectively. Hence, we can deduce that the proposed method was effective, and also outperformed the betweennes method by about 10% at $k = 300$ for the Wikipedia network. Moreover, we find from Figure 6 that blocking 300 links by using the proposed method was the same as blocking about 30000 links by using the out-degree and random methods for the Wikipedia network. Namely, we can deduce that the proposed method was effective about 100 times as much as the out-degree and random methods at $k = 300$ for the Wikipedia network.

These results imply that the proposed method works effectively as expected, and significantly outperforms the conventional link-removal heuristics, that is, the betweenness, out-degree and random methods. This shows that a significantly better link-blocking strategy for reducing the spread size of contamination can be obtained by explicitly incorporating the diffusion dynamics of contamination in a network, rather than relying solely on structual properties of the graph.

We note from Figures 4 and 6 that the out-degree method was almost the same as or worse than the random method in performance. In the task of removing nodes from a network, the out-degree heuristic has been effective since many links can be blocked at the same time by removing nodes with high out-degrees. However, we find that in the task of blocking a limited number of links, the strategy of blocking links between nodes with high out-degrees is not necessarily effective.

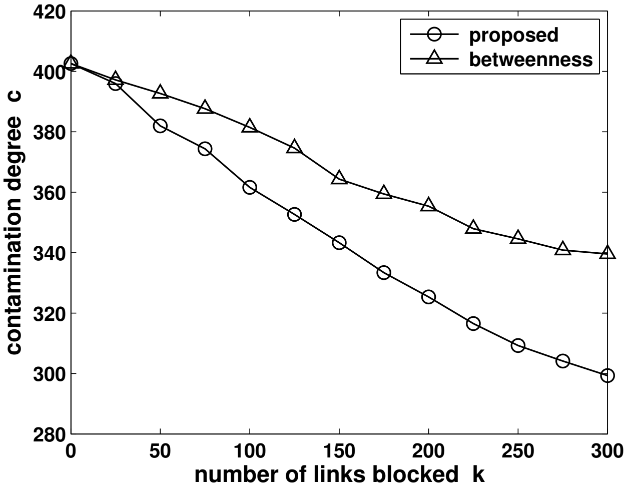

Figure 5: Performance comparison between the proposed and betweenness methods in the Wikipedia network for the IC model with $p = 0.03$.

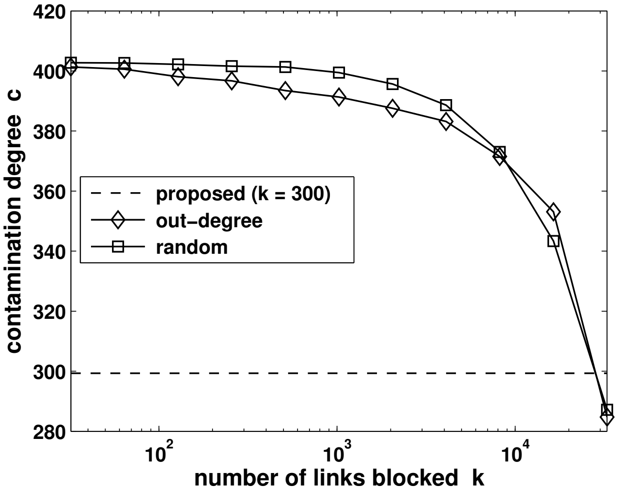

Figure 6: Performance comparison of the proposed method for $k = 300$ with the out-degree and random methods in the Wikipedia network for the IC model with $p = 0.03$.

Conclusion

In an attempt to minimize the spread of undesirable things by blocking links in a network, we have considered the contamination minimization problem, a dual problem to the influence maximization problem for social networks. This minimization problem is another approach to the problem of preventing the spread of contamination by removing nodes in a network, We have proposed a novel method for efficiently finding a good approximate solution to this problem on the basis of the greedy algorithm and the bond percolation method. Using large-scale blog and Wikipedia networks, we have experimentally demonstrated that the proposed method effectively works, and also significantly outperforms the conventional link-removal heuristics based on the betweenness and out-degree. Moreover, we have found that unlike the task of removing nodes, the strategy of blocking links between nodes with high out-degrees is not necessarily effective for our problem.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research, U.S. Air Force Research Laboratory under Grant No. AOARD-08-4027.

References

Albert, R.; Jeong, H.; and Barabási, A.-L. 2000. Error and attack tolerance of complex networks. *Nature* 406:378–382.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; and Wiener, J. 2000. Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference*, 309–320.

Callaway, D. S.; Newman, M. E. J.; Strogatz, S. H.; and Watts, D. J. 2000. Network robustness and fragility: Percolation on random graphs. *Physical Reveiw Letters* 85:5468–5471.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 57–66.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference*, 107–117.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Kimura, M.; Saito, K.; and Nakano, R. 2007. Extracting influential nodes for information diffusion on a social network. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 1371–1376.

Newman, M. E. J., and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69:026113.

Newman, M. E. J.; Forrest, S.; and Balthrop, J. 2002. Email networks and the spread of computer viruses. *Physical Review E* 66:035101.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review* 45:167–256.

Richardson, M., and Domingos, P. 2002. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 61–70.