Toward Optimal Vaccination Strategies for Probabilistic Models

Zeinab Abbassi
Columbia University
zeinab@cs.columbia.edu

Hoda Heidari
Sharif University of Technology
heidari@ce.sharif.edu

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Applications—*Data Mining*

General Terms

Algorithms, Economics, Performance, Human Factors

Keywords

Vaccination, Social Networks, Probabilistic Model, Virus Propagation, SIR

1. INTRODUCTION

Epidemic outbreaks such as the recent H1N1 influenza show how susceptible large communities are toward the spread of such outbreaks. The occurrence of a widespread disease transmission raises the question of vaccination strategies that are appropriate and close to optimal. The seemingly different problem of viruses disseminating through email networks, shares a common structure with disease epidemics. While it is not possible to vaccinate every individual during a virus outbreak, due to economic and logistical constraints, fortunately, we can leverage the structure and properties of face-to-face social networks to identify individuals whose vaccination would result in a lower number of infected people.

The models that have been studied so far [3, 4] assume that once an individual is infected all its adjacent individuals would be infected with probability 1. However, this assumption is not realistic. In reality, if an individual is infected by a virus, the neighboring individuals would get infected with some probability (depending on the type of the disease and the contact). This modification to the model makes the problem more challenging as the simple version is already NP-complete [3].

Here we consider the following epidemiological model computationally: A number of individuals in the community get vaccinated which makes them immune to the disease. The disease then outbreaks and a number of nodes that are not vaccinated get infected at random. These nodes can transmit the infection to their friends with some probability. In this work we consider the optimization problem in which the number of nodes that get vaccinated is limited to k and our objective is to minimize the number of infected people overall. We design various algorithms that take into account the properties of social networks to select k nodes for vaccination in order to achieve the goal. We perform experiments on a real dataset of 34,546 vertices and 421,578 edges and assess their effectiveness and scalability.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. MODEL

We represent the social contact network by a graph $G = (V, E)$, where each individual is denoted by a node in the graph. There is an edge between two nodes, if the probability of virus transmission between its endpoints is non-zero. For example, in the case of flu there is an edge between two co-workers.

In this graph, let q_i be the probability that node i gets infected initially (when this probability is assumed to be the same for all nodes, we denote it by q). Also, let's assume that for all $i, j \in V$, p_{ij} is the probability that j would get infected, given i is infected.

The problem that we consider in this paper is defined as follows: Given a social contact network which is represented by $G(V, E)$, the probabilities q_i and p_{ij}, and a parameter k; our objective is to find k nodes to vaccinate such that the total expected number of infected nodes is minimized. In other words, let T be the set of vaccinated nodes and $f(T)$ be the expected number of nodes that get infected after set T is vaccinated. The goal is to find a set T ($|T| = k$) of nodes to vaccinate in order to minimize $f(T)$.

Given the set T of vaccinated nodes and set S of initial infected nodes, let $f_S(T)$ be the set of nodes that get infected through the propagation of infection according to the random process. It can be seen that $f(T) = \sum_{S \subset V(G)} q(S) f_S(T)$ where $q(S)$ is the probability that the set S of nodes gets initially infected and is equal to $\Pi_{i \in S} q_i \Pi_{i \notin S} (1 - q_i)$

Evaluation Function. It can be shown that for any $\epsilon > 0$, function $f(T)$ can be computed within a factor $1 - \epsilon$ in time polynomial in the size of the input and $\frac{1}{\epsilon}$ [1].

Remark: $f(T)$ is a monotone function and is neither submodular nor supermodular. For proofs please refer to [1].

3. ALGORITHMS

Since the problem of vaccination is shown to be NP-hard [3], we design several heuristic algorithms to tackle it. Our algorithms have a common basic structure: for all of them, we define a measure (M) by means of which we calculate the vaccination priority of the nodes. More precisely, we iteratively calculate M for all remaining nodes of the input network and vaccinate the node whose M is extremum, until sufficient number of nodes get vaccinated. We call this template Iterative Candidate Selection Algorithm (ICSA).

In order to improve the running time of ICSA, we need to decrease the number of times it updates M. So we modify it in a way that it only updates M when $c^n (n = 0, 1, 2, ...)$ percentage of nodes have already been vaccinated. The reason that we update the measure more frequently at the beginning is that deleting a node when the graph is dense has probably more impact on the expected number of infected nodes than when the graph has become sparse.

The rest of this section is devoted to the measures we have used.

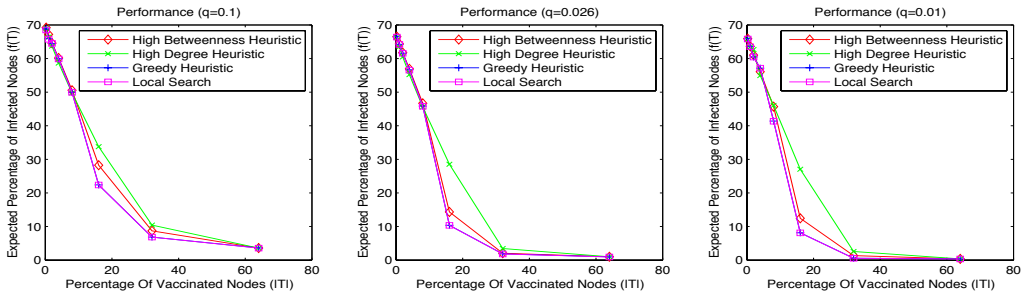

Figure 1: Performance of different heuristics on HEP-PH with $p = 0.05$.

High Degree. An intuitive vaccination strategy is to vaccinate the individuals who have more contact with others. In order to consider the likelihood of transmission of disease as well, we define the expected degree of a node to be the sum of the probabilities (p) of its incident edges. The High Degree heuristic is the (modified) ICSA where its priority measure is set to expected degree.

High Betweenness. Another centrality measure that we consider is node betweenness. Vaccinating nodes of high betweenness can prevent an epidemic to be transmitted to a large group of nodes.

Again here, we take into account the likelihood that virus reaches other nodes from a node and define expected betweenness of a node to be the overall expected amount of flow that it receives. In the High Betweenness Algorithm we use expected betweenness as the measure M for the ICSA.

Greedy. A natural strategy for selecting good candidate nodes is the greedy algorithm: in each step, vaccinate the node which decreases f by the most value. The greedy algorithm would be time-consuming for large networks, since at each step, one has to try all nodes $u \notin T$, and compute $f(T \cup \{u\})$ via sampling. One way to improve the running time is to compute $f(T \cup \{u\})$ only for nodes in a priority queue that contains important nodes, like nodes with high expected degree. By manipulating the size of this queue, we can adjust accuracy and running time of the algorithm.

Local Search. To improve the results of the above algorithms, one can try a natural local search, a.k.a Swap algorithm. An improving swap here means to find $u \in T$ and $v \notin T$ so as $f(T \cup \{v\} \backslash \{u\}) < f(T)$ and substitute u with v, i.e. $T = T \cup \{v\} \backslash \{u\}$. In the local search algorithm, we perform the swap that decreases f by most value, as far as an improving swap exists. To make the local search scalable, we examine only the possible swaps between high expected degree nodes of $V - T$ and T.

4. EXPERIMENTAL EVALUATION

We conduct experiments on our algorithms with the objective to asses and compare their performance and running time. For this work we have selected the HEP-PH citation dataset which was initially used in the 2003 KDD cup challenge and covers all the citations within a dataset of 34,546 papers with 421,578 edges. Citation/collaboration networks are shown to be good representations of social networks [5, 2]. If a paper p cites paper q, the graph contains a directed edge from p to q. For the purposes of our application we treat the graph as undirected [6].

The experiments are run with the following parameters: $c = 2$, $\epsilon = 0.001$, $p = 0.05$, and 0.005 and $q = 0.1, 0.01$, and $\frac{5}{\sqrt{|V|}}$. The results for the experiments with $p = 0.05$ are shown in Figure 1. The graphs are concave confirming that the nodes that nodes that get vaccinated at the beginning have more marginal impact. The experiments show that greedy is more effective, decreasing the percentage of expected infected nodes to less than a third by only vaccinating %16 of the nodes.

The local search algorithm depicted in Figure 1 is performed on the results of the greedy algorithm, therefore it is expected to perform at least the same. The results show slight improvements, which implies that our greedy algorithm works well leaving minimal space for improvements while taking much less time than the local search operations. The above trends hold for the trial in which $p = 0.005$ which is not shown here due to space constraints.

We also created an equivalent Erdos-Renyi graph and performed the same experiments on them. All the algorithms are less effective on random graphs as expected, but their relevant performance is the same. The running times of the algorithms are shown in Table 1. For figures and more details of the experiments please see [1].

Table 1: Running times in seconds.

Degree	Betweenness	Greedy	LocalSearch
149	22069	64676	178703

5. CONCLUSION AND FUTURE DIRECTIONS

In this work we consider the problem of vaccination (over a contact network or an email network), where constraints only allow k individuals to be vaccinated. We propose a few heuristics and compare their performance and running time on a real citation network with $34,546$ vertices and $421,578$ edges. In general, we observe that by leveraging the properties of the network to select the best candidates for vaccination, we get very good results. We show that by vaccinating only a small fraction of the nodes we can decrease the spread of viruses by a lot. We also conclude that greedy is the best algorithm, both in terms of performance and running time. Interesting future directions are devising improved heuristics and possibly approximation algorithms, and also looking at the problem of minimizing overall cost when k is not given.

6. REFERENCES

[1] www.cs.columbia.edu/~zeinab/poster.pdf
[2] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the Spread of Influence through a Social Network, KDD'03.
[3] J. Aspnes, K. Chang, A. Yampolskiy, Inoculation Strategies for Victims of Viruses and the Sum-of-Squares Partition Problem, Journal of Computer and System Sciences, 2006.
[4] P. Chen, M. David, D. Kempe, Better Vaccination Strategies for Better People, EC'10.
[5] M. Newman. The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. 98(2001).
[6] J. Gehrke, P. Ginsparg, J. M. Kleinberg. Overview of the 2003 KDD Cup. SIGKDD Explorations 2003.