Operations Research Letters 43 (2015) 59–64

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Optimization of stochastic virus detection in contact networks

Jinho Lee [a], John J. Hasenbein [b,*], David P. Morton [c]

[a] *Department of Management Science, Korea Naval Academy, Anggok-dong, Jinhae-gu, Changwon-si, Kyungsangnam-do, 645-797, South Korea*

[b] *Graduate Program in Operations Research and Industrial Engineering, Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78712-1591, USA*

[c] *Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3109, USA*

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form
16 November 2014
Accepted 16 November 2014
Available online 22 November 2014

Keywords:
Stochastic spread model
Submodular optimization
Stochastic integer programming

ABSTRACT

We develop network models to represent the dynamics of a virus spreading in a contact network. Based on the resulting dynamics governing the spread, we present optimization models to rapidly detect the virus. We consider two goals, maximizing the probability of detecting a virus by a time threshold and minimizing the expected time to detection. We establish submodularity results for these objective functions and, using data from a mobile service provider, we show that a greedy heuristic performs surprisingly well.

1. Introduction

We examine the problem of detecting a virus that spreads throughout a contact network in a stochastic manner. In the basic model we have an undirected graph representing a contact network (e.g., a group of friends on Facebook, or a cell phone network). At time 0, a virus begins at some node of the network and spreads in a manner that may be stochastic. Before time 0, the system manager can install a limited number of detectors at some subset of the nodes in the graph. The manager's goal is some metric related to rapid detection of the virus. The two goals we consider are to: (i) maximize the probability of detection before some fixed time t_0 and (ii) minimize the expected time until detection. At its root, this is a combinatorial stochastic optimization problem, which is difficult to attack directly. In this paper, we present a general class of models, obtain submodularity results, and investigate the efficiency of greedy heuristics using real data.

This model was motivated by our work with SK-Telecom (SKT), one of the largest mobile service providers in South Korea. In that work we examined a contact network arising from calls or texts between cell phone users. The threats in such a network are MMS viruses that may spread stochastically by, for example, picking random contacts from a user's address book. Messages can be scanned at MMS gateways for suspicious signatures. However, a well-written virus may evade detection at gateways. We focus on detecting viruses by monitoring cell phones for anomalous behavior, much in the same way a system administrator monitors computers. However, monitoring requires bandwidth, and so we are limited in the number of phones we can monitor. We use SKT data in the computational results we describe in Section 5. More extensive motivation and computational work can be found in [9,10]. Our model has potential application to any type of contact network subject to unwanted invasions, which move randomly – as opposed to adversarially – on that network.

There is related work in the literature. Berman et al. [2] and Gutfraind et al. [4] consider the spread model that we call TN1*C in the next section. While their applications differ from ours, in our terminology they install detectors to maximize the probability of detecting the virus before the virus reaches a "bad" node, and they establish a submodularity result for that objective function. Krause et al. [8] install sensors in a municipal water system to detect the malicious introduction of contaminants. They consider criteria including the expected time to detect the contaminants, the expected population affected by the contaminants, the expected amount of contaminated water consumed prior to detection, and the probability of detecting the contaminants. They formulate optimization models based on maximizing the reduction in a penalty function, and establish a submodularity result for this penalty-reduction metric. Kempe et al. [6] consider the problem of maximizing influence in a social network. Here, information is inserted at selected nodes in a social network after which the information diffuses across that network. The goal is to select the nodes of insertion to maximize influence. Kempe et al. establish submodularity properties for influence functions under a class of diffusion models.

[*] Correspondence to: 204 E. Dean Keeton St. Stop C2200 Austin, TX 78712-1591, USA.
E-mail address: jhas@mail.utexas.edu (J.J. Hasenbein).

http://dx.doi.org/10.1016/j.orl.2014.11.006
0167-6377/© 2014 Elsevier B.V. All rights reserved.

Table 1
Nomenclature for models of virus spread.

Characteristics	Models
Replication	Virus transits from node to node (T), Virus replicates itself and sends copies (R).
Persistence	Only newly infected nodes distribute the virus (N), All infected nodes distribute the virus (A).
Propagation	Virus propagates to one randomly selected neighbor (1), Virus propagates to every neighbor (E).
Susceptibility	Nodes are susceptible with probability one (1), Nodes are susceptible with probability $p < 1$ (P).
Transmission time	Transmission in constant unit time steps (C), Transmission according to an exponential distribution (M), Transmission according to a general distribution (G).

2. Modeling

2.1. The stochastic virus model

We define a contact network to be a graph $G = (V, E)$ on a set V of nodes and a set E of undirected edges, where $|V|$ and $|E|$ denote the number of nodes and edges, respectively. We call two nodes *neighbors* if an edge connects them. We assume a known probability distribution governs the single initial node from which a virus begins to spread at time 0, and we let w_i, $i \in V$, denote that probability mass function.

We describe now a framework for organizing the assumptions of the manner in which the virus spreads on the network starting at time 0. The model classification is organized by considering the following questions:

- How does the virus disseminate through the network?
- How long are the nodes "contagious"?
- How does the virus select nodes for replication?
- How susceptible are nodes to the virus?
- What model of time dynamics represents virus transmission?

We label the model characteristics implied by answers to these questions as follows: replication, persistence, propagation, susceptibility, and transmission time. Varying these characteristics generates a diverse array of spread models, some of which have appeared in the literature. Table 1 summarizes the corresponding set of models.

Replication. A virus may either simply transit (T) the network without creating copies or replicate (R) itself by sending copies throughout the network. The former mode is unusual for viral models, but we include it because it is related to other models appearing in the literature.

Persistence. A virus may only attempt to infect other nodes one time or it may persistently attempt to infect other nodes. In discrete-time models this corresponds to only newly (N) infected nodes spreading the virus, or all (A) infected nodes spreading the virus.

Propagation. An infected node may attempt to propagate by choosing one randomly selected neighboring node (1) or it may attempt infection of every (E) neighboring node. Other variations can be taken into account via the next characteristic.

Susceptibility. Potential transmission of the virus to a node may not succeed. Some nodes may be deemed immune, due to biological reasons, or technical reasons, depending on the application. So, we distinguish the case in which every node that receives a virus becomes infected (1) from the case in which each attempted infection succeeds independently with probability p (P).

Transmission time. The transmission time indicates how much time passes before an infected node takes action to infect another node, or nodes. Under constant transmission times (C), we have a discrete-time model. Under exponential (M) transmission times, the model is a continuous-time Markov chain. Generally distributed (G) transmission times, in most cases, require a more complex stochastic model (such as a semi-Markov process).

To denote a specific spread model, we use a string of five letters and numbers. For example, RNE1C represents the spread model in which the virus replicates and sends copies from newly infected nodes to all neighbors, with probability one, in discrete time. Our computational results focus on discrete-time models. However, we present this framework for two reasons: (i) some of our theoretical results hold for the entire set of models, and (ii) some model variations relate to models in the literature. We make a few observations below on various model classes.

First, consider models whose designation begins with T. In such models, the virus hops from one node to another. In particular, the virus exists only at one node at a time. In this mode of replication, the sets of designations TA*** and T*E** (here * is a wildcard) do not make sense. So, we can restrict attention to the models in the class TN1**. The models TN11C and TN1PC can be represented by discrete-time Markov chains (DTMCs) with a state space of size $|V|$. In the former model, the virus hops from node to node by randomly selecting a neighboring node. The latter model could be interpreted similarly, but with the additional possibility that the virus can remain at the same node at subsequent time steps. The models TN11M and TN1PM are just continuous-time analogs of the first two models mentioned and can thus be represented by continuous-time Markov chains (CTMCs). Finally, the related models TN11G and TN1PG can be represented by semi-Markov processes.

Now, consider spread models whose designation begins with R. The distinction between the classes RN*** and RA*** involves whether the infected nodes persist in attempting infection. Furthermore, we have equivalence between the classes RAE1* and RNE1*, since the set of infected nodes is the same in either class. The spread model RAE1C is a special case in which an infected node replicates and sends copies to every neighboring node in discrete time, and this is the only model under consideration with completely deterministic dynamics. That said, even with a deterministic model of spread, the model has stochastic elements in that we allow the initial location of the virus to be random.

Finally, consider spread models in which transmission occurs in constant unit time steps, C. Each of our models of virus spread with constant-time transmission is a time-homogeneous DTMC with a state space of size at most $2^{|V|}$, because the state is given by the current set of infected nodes. If transmission times are instead exponentially distributed then each of our models is time-homogeneous CTMC with a finite state space.

2.2. Metrics and optimization

For purposes of exposition, we focus on two metrics to capture the goal of rapidly detecting a virus. Other metrics representing, for example, more complex risk attitudes, could be of interest also. For each metric, some fixed set of nodes, $S \subseteq V$, called *detectors* must be chosen in advance, and we detect the virus when it first infects a node in S. Thus, any realization of the virus propagation process yields a detection time. We assume that the detectors are perfectly reliable (i.e., the virus is always detected when it infects a detector), although it is straightforward to incorporate false negatives.

In the optimization model under our first metric, we choose S to maximize the probability of detecting the virus by a given time threshold (MPT). In our second optimization model, we choose S to minimize the expected time to detection (MET). In each case, the requisite probability distribution governing propagation derives from the probability mass function governing the initial location of the virus and the model of spread that we assume.