Detecting Epidemics Using Highly Noisy Data

Chris Milling
UT Austin
1 University Station
Austin, TX
cmilling@utexas.edu

Constantine Caramanis
UT Austin
1 University Station
Austin, TX
constantine@utexas.edu

Shie Mannor
The Technion
Haifa, Israel
shie@ee.technion.ac.il

Sanjay Shakkottai
UT Austin
1 University Station
Austin, TX
shakkott@austin.utexas.edu

ABSTRACT

From Cholera, AIDS/HIV, and Malaria, to rumors and viral video, understanding the *causative network* behind an epidemic's spread has repeatedly proven critical for managing the spread (controlling or encouraging, as the case may be). Our current approaches to understand and predict epidemics rely on the scarce, but exact/reliable, expert diagnoses. This paper proposes a different way forward: use more readily available but also more noisy data with *many false negatives and false positives*, to determine the *causative network* of an epidemic. Specifically, we consider an epidemic that spreads according to one of two networks. At some point in time we see a small random subsample (perhaps a vanishingly small fraction) of those infected, along with an *order-wise similar number* of false positives. We derive sufficient conditions for this problem to be detectable, and provide an efficient algorithm that solves the hypothesis testing problem. We apply this model to two settings. In the first setting, we simply want to distinguish between random illness (a complete graph) and an epidemic (spread along a structured graph). In the second, we have a superposition of both of these, and we wish to detect which is the strongest component.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Stochastic Processes

Keywords

epidemic process, network inference

1. INTRODUCTION

The study of epidemic spread over social, communication, and human contact networks, be it a contagion of a human or computer virus, or a rumor, opinion or trend, begins with two basic questions: do we indeed have a spreading epidemic, and if so, what is the causative network spreading it? Numerous famous examples from the history of epidemiology ([22, 3]) have illustrated the importance and difficulty of determining the causative network. With accurate data collected over time, for example, from high accuracy medical diagnoses of a known illness, the causative network essentially reveals itself. Yet such data are rarely available. More to the point, highly incomplete and noisy data *often are available*. Indeed, the challenge arises in particular, when time lapse data of "true" illness is not available, and when the data we do have is highly noisy with many false positives and negatives. For example, online records (flu-related keywords in social networks [4], or Internet searches such as in Google Flu Trends [11]) provide large but noisy data sources for detecting flu epidemics, but potentially containing many false positives.

A similar fundamental difficulty arises with other epidemics. Consider the adoption of the latest tablet or smartphone. The spread is likely driven *both* by "word-of-mouth" (online social networks via tweets or Facebook posts) advertising, as well as explicit advertising campaigns over television, Internet ads, etc. While both modes likely play a part in driving sales, is it the broad-spectrum advertising (which is in effect a star network that connects the advertiser to all the television viewers) that serves as the dominant driver of spread, or is it the word-of-mouth viral marketing that is dominant? Surveys may reveal (noisy) data on who owns the new tablet, but pinpointing time-of-acquisition and the causative network is much more difficult.

In a communications setting, carriers need to worry about similar problems. On observing abnormal interactions from some smartphones on their network, they need to decide if this is due to a buggy firmware update, or something more malicious (such as malware/virus spread). Unfortunately however, these carriers rarely have access to these user devices themselves, thus, need to recourse to inferring from the limited and noisy samples (e.g., phone-to-network interactions). While currently small, malwares [15] and viruses that spread via user contact networks [8, 23] are receiving increasing attention.

The key idea in this work, is that different spreading mechanisms have different statistical signatures, in terms of the subset of people infected. This is certainly the case when the

MobiHoc'13, July 29–August 1, 2013, Bangalore, India.

causative graphs are very different, and the subset of nodes (people, machines, etc.) the epidemic has reached ("infected nodes") are completely and accurately revealed. As discussed, however, the data available are typically noisy, with many false positives and negatives. Moreover, the larger the fraction of the network the contagion has reached, the more this "network signature" is washed out. This paper explores these tradeoffs. We consider a broad class of graphs: graphs with bounded degree. The degree controls the infection's speed. We give sufficient conditions on when the causative network can be determined when only a vanishing number of infected nodes report, and moreover when a constant fraction of those reporting are false positives. Then, we consider the case most relevant in spread of rumors, technology and ideas: the superposition of two spreading mechanisms. Indeed, in the age of mass advertising and mass media, trends spread friend-to-friend, but also through television, Internet ads, and similar advertising efforts that exhibit a "star-like" contagion network. We provide sufficient conditions for determining which is the dominant effect, again when only a vanishing fraction of infected nodes report, and when no time-lapse data are available.

1.1 Related Work

Analyzing the spread of epidemics under the susceptible-infected (SI) model [9] has been considered in depth for a variety of graphs and circumstances [1, 12]. While there has been much work on what we call the *forward problem*, i.e., predicting what an infection may look like or do, the present work falls under the heading of *backward problems*. Thus, the work in this paper is related to the task of inferring various characteristics of the infection given the infected nodes. We briefly mention several related works. Demiris and O'Neill estimate the transmission rates of the epidemic [5, 6]. Shah and Zaman provide an algorithm to estimate the node most likely to be the source of the infection [20, 21]. Luo and Tai consider a similar problem with multiple infection sources [14]. Myers *et al.* estimate the proportion of infections that occur randomly (from unknown sources external to the network) given the full sequence of infected nodes under a similar mixed infection setting [18].

In even more closely related work, Netrapalli and Sanghavi analyze the problem of estimating the structure of the contagion network given the times each node is infected for several epidemics on the same network [19]. Gomez-Rodriguez *et al.* provide an algorithm to solve a similar problem with a somewhat different infection model [10]. These works use time-lapse information of multiple epidemics, and from those are able to detect the edges of the causative network. This is more general in the sense that very little is known about the underlying spreading mechanics, compared to what amounts to a hypothesis testing problem in our setting. On the other hand, the data-availability regime under which we operate, is much more harsh. We have very limited information, viewing only a very partial and also noisy, subset of the infected nodes of a single contagion *at a single point in time*. We have no time-related information. Under these limited-information conditions, inferring the network structure, as in [19] or [10], would not be possible. The work in [16] and [17] on hypothesis testing for determining the graph corresponding to the epidemic spread, by Milling *et al.*, is also very relevant (we elaborate on the connection in Section 3). However, these papers do not consider false positives, or superposed spreading (spreading simultaneously via multiple networks), whereas these are precisely the main contributions of this current paper.

1.2 Main Contributions

The fundamental problem we consider is diagnosing the causative network of an epidemic or contagion, *using noisy and highly incomplete* data, and in particular, data without temporal information. The focus and main contribution of this paper is in tempering the effect of this noise in the data, and hence greatly expanding the available data sets that can be used. To the best of our knowledge, this is the first paper that has considered epidemic forensics in this regime.

The first part of the paper attempts to diagnose a contagion as arising from node-to-node contact via a specific contact network, or through a random infection process. The latter can be modeled as a contagion spreading from the center node of a star graph, to the leaves. As discussed above, we assume there is an overwhelming fraction of false negatives (that is, only a vanishing fraction of infected nodes report). Of this vanishing fraction, we assume that a constant fraction are false positives. We note that there is no way to identify false positives or false negatives, but only to provide a system-level diagnosis. Indeed, our goal is to diagnose the spreading network of the contagion.

Next, we consider the superposition of these two contagion processes, and attempt to determine the stronger of the two components. That is, a contagion spreads through a star network as described above, *and also simultaneously through subsequent node-to-node contact.* This is the case with much advertising: television and other mass-media advertisements provide initial "seeds," but subsequent spread may occur through word-of-mouth. For different products, the relative effects of these two mechanisms may differ. Under what circumstances can we determine the stronger component?

We note that robustness is again at the crux: we are attempting to determine system-wide effects; at the local level, it is impossible to say if the contagion reached an infected node through what we have called the star model, or through local interactions (e.g., word of mouth).

Specifically, our contributions are as follows.

- Algorithm Development: We provide an algorithm we call the *Median Ball Algorithm.* This algorithm is simple, efficient to run, and terminates quickly even for very large graphs. In the first case of diagnosing a contagion as a spreading epidemic or a random illness (the star graph), it outputs what it believes is the most likely candidate. Similarly, for the case of a superposition of those two processes, it outputs what it believes is the stronger of the two components.

- Arbitrary (Adversarial) False Positives: We give sufficient conditions in terms of number of total sick nodes, number of sick nodes reporting (i.e., fraction of false negatives) and fraction of false positives, under which the algorithm above correctly diagnoses the causative network, i.e., with the probability of Type I and Type II error going to zero in the size of the graph. In particular, our results show that our algorithm correctly identifies the causative network even when only a vanishingly small fraction of sick nodes report, and moreover, when up to 50% of the reporting nodes are false positives, *even when those false positives are adversarially selected.*

- Random False Positives: We give sufficient conditions for the same problem, when the false positives are randomly (independently and uniformly) distributed in the graph. Here, we show that our algorithm correctly identifies the causative network *for any fraction of false positives up to* 100%.

- Superposed Spreading (Mixed Infection Types): Finally, we consider the setting where the spreading occurs through both random infection (the star spreading graph) representing, for instance, television or Internet advertisement, and subsequently through node-to-node contact (word-of-mouth) and give sufficient conditions for when our algorithm correctly determines which of the two components is the dominant factor in the contagion spread. Again, we require only a vanishingly small fraction of sick nodes to report.

2. MODEL AND ALGORITHM

We have described intuitively the contagion spreading models we attempt to distinguish. In this section, we describe them more precisely and build on the models in [16, 17]. We consider two distinct infection regimes: a contagion spreading through node-to-node contact, versus random spread of the contagion, when each node becomes sick independently of its location in the graph or the status of its neighbors. We can model this, too, as an epidemic spreading over a star graph where initially only the center node is infected. For ease of discussion, we call the node-to-node mode of spreading an *epidemic*, and the star-mode of spreading a *random sickness*. In both cases, we start with a single infected node at time 0.

The Infection Process

Let $G = (V, E)$ denote the graph along which the infection spreads. As discussed above, in the case of an epidemic spreading node-to-node, G is a structured graph (e.g., d-dimensional grid). For the case of a random illness, the graph G is a star graph, with every node connected to a central node assumed to be infected at time zero. We let $n = \text{card(V)}$, the size of the graph. The diameter of the graph is denoted $\text{diam}(G)$.

Given a graph G, the contagion spreads as follows. At time zero, an initial node is selected and called "infected." For the structured graph case, we assume this initial infected node is selected uniformly at random. For the star graph, it is the central node. The infection spreads from that node to its neighbors, across the edges of the graph. The spreading occurs according to a standard susceptible-infected (SI) model [9, 7, 13] for an epidemic. The spreading rate is parameterized by a single number, or rate. To make clear the distinction between the rate for a structured graph or for a star graph, we use η to represent the rate of the structured graph, and γ/n the rate of the star graph. We divide by n in the case of the star graph so that new infections appear at rate γ (ignoring the shrinking number of susceptible nodes). This means the following: for each infected node and for each edge incident to that node, we start an exponential clock, i.e., a clock that expires after an exponentially distributed length of time, of expectation $1/\eta$, i.e., of rate η for a structured graph, and n/γ, i.e., of rate γ/n, for the star graph. The expiration of a clock indicates that the adjacent node becomes infected (if it is not already infected) and new clocks are started for each edge from this newly infected node. In this way, the infection spreads along the edges of the graph in a node-to-node fashion.

Let S denote the set of infected nodes at a given time. The rate of new infections is (roughly) proportional to the number of uninfected nodes ($V \setminus S$) incident to an infected node. Thus, for a random sickness (G a star graph) new nodes become infected at a rate $(\text{card}(V \setminus S))\gamma$, and hence the rate of new infections in fact decreases as more nodes become infected. For most graphs, the rate of infections initially increases, before decreasing as more and more nodes become infected. The most challenging regime to pose the problem of diagnosing the causative network of the epidemic, is where the expected number of infected nodes is the same under both models. Thus, for the remainder of this paper, all results are stated under precisely this assumption.

The second half of this paper considers mixtures of both of these types of infections: the star graph infects nodes at rate γ/n, and then these infected nodes infect their neighbors on the structured graph (e.g., the grid) at rate η. Thus, in this superposed process, nodes become infected at random at some rate γ, and the infection then spreads from these nodes as an epidemic at the (different) rate η. In this setting, we consider two different processes: one where the dominant factor is the random infection (the spread from the star graph) and the other where it is the spread along the structured graph that dominates. Thus, in the first setting we have $\gamma \gg \eta$, and the random infection dominates the epidemic, and in the second setting, $\eta \gg \gamma$, and the epidemic spread dominates the infection process.

The Reporting Process

At a given point in time, a subset of the infected nodes is revealed – these nodes are discovered to be infected, or they self-report as infected. Given this snapshot, the task is to determine the spreading process. If we had access to the entire set of infected nodes, then a simple test of the connectivity of the infected nodes would easily distinguish the infection mechanism with overwhelming probability. In most contagion processes, however, only a small – perhaps a vanishingly small – fraction of infected nodes are detected, or self-report. Indeed, most people suffering from flu symptoms do not visit a doctor; no survey or poll reveals more than a minuscule fraction of adopters of a new technology; and likewise, only few virus-infected computers are reported/detected. In short, there may be a large – possibly an overwhelming – fraction of false negatives, i.e., of infected nodes that do not report (or are not detected). As a consequence, reporting infected nodes are likely to be disconnected, possibly with relatively large distance between them. This may be particularly challenging in small-world graphs.

As the theorem statements below make clear, we indeed assume that the fraction of false negatives is overwhelming. Our theorems show that the algorithm we provide can recover the true infection mechanism when only a logarithmic number of infected nodes reports, i.e., when the fraction of reporting infected nodes is *exponentially small.*

Further confounding the task of diagnosing an epidemic versus random illness may be the presence of false positives among the reporting infected nodes. This again is the case with many available data sets. Obtaining accurate diagnoses (i.e., with very low false-positive rate) is difficult. In human illnesses, in many cases (in particular, easily treatable sexually transmitted diseases) public policy prescriptions have focused on tests that are inexpensive, yield results quickly, and have a low false-negative rate. Moreover, data on self-reported illness (i.e., without medical diagnosis) are increasingly available, and should be exploited. Answers to surveys and polls suffer from precisely the same possibility of false positives. Moreover, depending on the setting, it is important to consider the case of correlated (clustered) or even worse, manipulative false positives, that may collude to obscure the infection propagation mechanism, or, simply, may not have an easily describable distribution. On the other hand, one would expect that if false positives are randomly (uniformly, and independently) distributed across the graph, that their effect would be less pernicious. We consider both settings, and indeed, show that this is the case. In the case of adversarially distributed false positives, our algorithm succeeds when up to 50% of the reporting nodes are false positives. When the false positives are randomly distributed, our algorithm can tolerate nearly 100% false positives.

The notation we use is as follows. As above, we let S denote the set of actually infected nodes (revealed or not). We assume that each of these reports its infection with some probability. We note that we do not require the reporting process to be independent, i.e., the set of infected nodes that report may well be correlated. We denote by q the probability that infected nodes report, and we denote the resulting subset by $S_r \subseteq S$ (and hence, $\mathrm{card(S_r)}$ has expected value $q\mathrm{card(S)}$). We further assume that some fraction of the uninfected nodes, $V \setminus S$, may (falsely) report infection as well. As discussed, we consider both cases where this fraction of falsely reporting nodes is chosen by an adversary and chosen randomly. In the random case, each false positive node is chosen uniformly at random from the entire graph, where repeats are allowed. We allow the "falsely reporting" nodes to be in S_r (so they are not truly false positives) to reduce dependence on S_r. This also ensures that the density of reporting nodes is highest in S_r. We thus denote the set of reporting sick nodes (including both truly infected, and false positives) by $\bar{S}_r \supseteq S_r$. We parameterize the number of false reporting nodes by a constant $f \geq 0$. Let the number of false positives be given by $\lfloor f \cdot \mathrm{card(S_r)} \rfloor$. Thus, $f/(1+f)$ is approximately the fraction of all reporting nodes that are false positives. When $f \to 1$, then fully half the reporting nodes are false positives. As f continues to increase, this fraction approaches 100%.

We consider false positives in the setting where we must determine if the infection spreads as an epidemic (a structured graph) or a random illness (a star graph). We show that our algorithm succeeds against adversarially selected false positives even as $f \to 1$. In the case of randomly selected false positives, our algorithm succeeds for any value of f. In the second half of the paper, we consider the superposition of the two processes. Here we focus only on false negatives. As the proof makes clear, incorporating false positives is a straightforward extension.

2.1 Graphs

Our results apply to a broad family of graphs, with different topologies. The key property we require is that the graph should have bounded degree (where this bound is a constant). From this property, we show that the infection can spread at only limited asymptotic speed, and that the neighborhood sizes are sufficiently small.

That is, there exists a constant $\bar{d}$ greater than or equal to the degree of each node for sufficiently large n. As a result of this property, the infection can only travel at a certain maximum rate through the graph. Define the random variable W as the maximum distance an epidemic has spread from its source. We define the condition *limited epidemic speed* as follows:

Definition 1. A graph has *limited epidemic speed* if there exist finite, positive constants s, λ_1 such that for sufficiently large n and an epidemic starting at any node a and duration t,

$$P(W > st) < e^{-\lambda_1 t}.$$

The speed s mentioned in the above definition is in fact an upper bound on the speed, in that there is no matching lower bound. Nevertheless, we refer to it as the speed for brevity. In addition, we also need a constraint on the neighborhood size.

Definition 2. A graph G has *limited neighborhood size* if $\mathrm{diam}(G)$ scales as $\Omega(\log n)$ and there exists a increasing concave function $b(x)$ such that for all $0 < x < 1$, $b(x) > 0$ and all balls of radius no more than $b(x)$ contains less than xn nodes for sufficiently large n with probability tending to 1.

In fact, both of these previous conditions follow from a bounded degree distribution, as stated formally below.

THEOREM 1. *Let G be a graph with maximum degree $\bar{d}$. Then G has both limited epidemic speed and limited neighborhood size.*

PROOF. First, the spread of the epidemic on G can be upper bounded by a tree of degree $\bar{d}$ where nodes are repeated for each path to them. See [17] for details on this bound. Then using a speed upper bound for trees, we find that G has *limited epidemic speed*, where the exponential probability of error follows from a Chernoff bound [2]. Next, using the maximum degree condition, the number of nodes within distance r from an arbitrary node a of G is at most $\bar{d}^{r+1}$. Therefore, for any x, $0 < x < 1$, no ball of radius $\log_{\bar{d}} xn - 1$ contains more than xn nodes. From this, we see that $\mathrm{diam}(G) \geq \log_{\bar{d}} n - 1$. Letting $b(x) = \log_{\bar{d}} xn - 1$, we see this satisfies the desired condition for *limited neighborhood size.* This completes the proof. □

Our bounds in the ensuing Theorems depend explicitly on the parameters that define the limited epidemic speed and neighborhood size conditions. We comment that tighter bounds can be derived with additional graph structure. For instance in a grid (lattice), first passage percolation results [13] provide sharper estimates, which in turn, can lead to stronger sufficient conditions in the ensuing Theorems. We refer to [17, 16] for an analogous discussion on graph-specific conditions (however, without false positives or superposed spreading).

Next, the following simple lemma (using a balls-in-bins argument) proves useful in the sequel, so we give it here.

LEMMA 1. *Suppose graph G has limited neighborhood size. Let $0 < x < 1$, $\epsilon > 0$ and R be a collection of nodes with*

$\text{card}(R) < (1-\epsilon)xn$. *Let* S *be a collection of uniformly random nodes with* $\text{card}(S) = \omega(\log n)$. *Then the probability that* R *contains at least* x *fraction of the random nodes in* S *decays to* 0 *as* n *increases. In particular, there exists a constant* $\lambda_2 > 0$ *such that*

$$P(\text{card}(R \cap S) \geq x\text{card}(S)) < e^{-\lambda_2 \text{card}(S)}.$$

The main way we use this lemma is to show that the probability that a large fraction of randomly selected nodes fall in a ball around a given node, goes to zero.

2.2 Algorithm

We develop a single algorithm we call the Median Ball Algorithm to solve the hypothesis testing problems in this paper – both for the case of detection of epidemic versus random illness (the first part of the paper), as well as the case of determining the dominant factor in the spread of the contagion (the second part of the paper). The Median Ball Algorithm is simple to describe: it searches for the smallest ball that covers a fraction of the reporting infected nodes. Of course, it has no way to tell if a reporting sick node is truly infected or a false positive, and as emphasized above, this is not the goal of this paper. If the resulting radius of this ball is small enough, it declares that there is an epidemic; otherwise, it concludes that the infection process is in fact a random illness. This algorithm is efficient, as even the brute-force implementation runs in time at most $O(|V| \cdot \text{diam}(G))$.

The algorithm takes two parameters α, and r. These parameters are tailored to the problem at hand, including, in the case of r, the size of the graph. As input, it takes a graph G and a set of reporting infected nodes S_r. If the algorithm can cover an α-fraction of the infected nodes in a ball of radius at most r, it declares the infection to be an epidemic; otherwise, it labels the infection a random illness.

Define $Ball_G(a, r)$ as all nodes in G within distance r from node a.

Algorithm 1 Median Ball Algorithm

```
Input: Graph $G$; Set of reporting infected nodes $S_r$;
Output: Epidemic or Random

  $c \leftarrow \alpha\,[\text{card}(S_r)]$
  for all $d \in V$ do
    $B \leftarrow Ball_G(d, r)$
    if $\text{card}(B \cap S_r) \geq c$ then
      return Epidemic
    end if
  end for
  return Random
```

3. FALSE POSITIVES

We consider the problem of determining if the spreading mechanism of a contagion is what we have termed an epidemic, or a random illness. We consider the superposition of these two in Section 4.

When all reporting nodes are truly infected (i.e., no false positives) then a special case of our algorithm can solve this problem with asymptotically (as the graph size scales) zero error: taking $\alpha = 1$, our algorithm reduces to the special case considered in [16] where one seeks a small ball containing *all the reporting sick nodes*. However, the algorithm in [16] fails even with a vanishing fraction of false positives – indeed, even with one single false positive. In contrast, the algorithm we give here succeeds with up to 50% *adversarially placed false positives*, and up to any fixed fraction less than 100% randomly placed outliers. Both of these results are the best possible under our formulation, where the number of false positives is proportional to the number of true infected nodes.

We show that by looking at the α-quantile ball, our algorithm becomes effectively immune to outliers. This happens for the following reason: suppose the true spreading process is an epidemic. We show that there is in fact a small-radius ball that covers all truly infected nodes. Now, the false positives either fall near or within this ball, or far outside it. In the first case, they do not require the ball to be larger and hence do not lead the algorithm to incorrectly pronounce the epidemic a random illness. In the latter case, they are ignored by the quantile ball algorithm, and thus again do not cause the algorithm to produce an error. If the true mechanism is random infection (i.e., the star graph) then the algorithm produces an error if it declares the infection mechanism to be an epidemic. For this to happen, either the true infections must appear clustered, even though they are independently distributed uniformly on the graph, or the false positives must exhibit a sufficient clustering to fool the algorithm. As we see, this is possible under adversarial placement of enough false positives, but probabilistically extremely unlikely otherwise.

Adversarial False Positives: First, consider the case where the false positives are chosen by an adversary with full knowledge of our algorithm and the true type of infection. The adversary can act non-randomly to attempt to confound the algorithm. In particular, the adversary can spread the false positives to look randomly placed when the infection is in fact an epidemic, and when the infection is random, can cluster the false positives to make the random sickness appear like an epidemic. We show that in both of these cases, if $f < 1$, i.e., the reporting nodes are less than 50% false positives, it is possible (for appropriate infection parameters) to distinguish the type of infection with probability of error tending to 0 as the number of nodes, n, tends to infinity.

We consider a graph G, with limited epidemic speed and limited neighborhood size. Let s denote the speed of an epidemic on the graph G (in fact, an upper bound on this, as discussed in Definition 1). Let $b(x)$ denote the limited neighborhood size function as defined previously in Definition 2. Note that from Theorem 1, all graphs with bounded degree have limited speed and neighborhood size.

THEOREM 2. ***[Adversarial False Positives]*** *Suppose* G *is as described. Suppose further that* $f < 1$ *and set* $f' = (1-f)/(1+f) > 0$. *Suppose* t *scales such that the number of reporting nodes is* $\omega(\log n)$ *and* $t < b(f'/2)/s$. *Then the Median Ball Algorithm with* $\alpha = 1/(1+f)$ *and* $r = st$ *correctly determines the type of infection with probability tending to* 1 *with the number of nodes,* n.

PROOF. First we show that the Type II error probability decays to 0. To this end, suppose the infection is in fact an epidemic. Consider only the true reporting nodes S_r. Note that $\text{card}(S_r) \geq \alpha\text{card}(\bar{S}_r)$. By the definition of speed s, the probability the epidemic spreads outside a ball of radius $r = st$ decays to 0, so this ball covers S_r and hence at least

α fraction of the reporting nodes. Therefore it is correctly labeled an epidemic.

Now we show that the Type I error probability also decays to 0. We need to show no ball of radius r can cover $\alpha = 1/(1+f)$ fraction of the nodes. Since only $f/(1+f)$ of the nodes are false positives, the ball must contain at least $(1-f)/(1+f) = f' > 0$ true reporting nodes. Then it is sufficient that the probability there exists a ball of radius r covering $f'\text{card}(S_r)$ true reporting nodes (which are located randomly) decays to 0.

Since $r < b(f'/2)$, no ball of radius r contains over $f'n/2$ nodes. Consider one of the n balls of radius r (one ball for each possible center node), call it R. Then by Lemma 1, there exists a strictly positive λ_2 such that

$$P(\text{card}(R \cap S_r) \geq f'\text{card}(S_r)) < e^{-\lambda_2 \text{card}(S_r)}.$$

Since $\text{card}(S_r) = \omega(\log n)$, $e^{-\lambda_2 \text{card}(S_r)} = o(1/n^2)$. Therefore, from a union bound, there is some ball of radius r containing over f' fraction of the true reporting nodes with probability at most $o(1/n)$. Hence, no such ball covers α fraction of the nodes in $\bar{S}_r$ with probability tending to 1 so the Type I error probability goes to 0. □

Therefore, as long as the number of false positives is (a fraction) less than the number of true reporting nodes, it is possible to determine whether an infection is due to an epidemic or a random sickness. Given the unlimited adversarial model, it is clear that this is tight. That is, if $f = 1$, it is impossible to distinguish the types of infection in the adversarial setting by any algorithm of any complexity. We state this simple converse result as a theorem.

THEOREM 3. *Suppose $f = 1$ and the random sickness is normalized so that the infection size distribution is equal for both infection processes. Then with adversarial false positives, the probability of error for any algorithm is at least* 0.5.

PROOF. There is a simple adversarial algorithm that guarantees a probability of error of 0.5. Recall the *a priori* probability for each infection process is equal. When the infection is from an epidemic, the adversary chooses nodes randomly exactly as in the random sickness. When the infection is from a random sickness, the adversary chooses nodes exactly as in an epidemic. Therefore, in all cases, exactly half the nodes are due to an epidemic, and half are due to a random sickness. Since the infection size is normalized, each collection of infected nodes is equally likely to be an epidemic as a random sickness. Then the probability of error for every set $\hat{S}_r$ is 0.5 (no matter the algorithm), and hence the overall probability is 0.5. □

Random False Positives: When an adversary places the false positives, the worst case scenario is generally when it places them in a cluster when the infection is in fact a random sickness. Therefore, when the false positives are located randomly over the graph, one would expect that the infection process is distinguishable for a larger range of f. We show that this is in fact the case. It is possible to distinguish an epidemic from a random sickness *for all values of f*. We note, though, that as one would expect, the larger the f, the tighter the constraint on the time of detection, i.e., than total number of infected nodes.

THEOREM 4. ***[Random False Positives]*** *Let $f > 0$. Suppose t scales such that number of reporting nodes is $\omega(\log n)$ and $t < b\left(\frac{1}{2(1+f)}\right)/s$. Then the Median Ball Algorithm with $\alpha = 1/(1+f)$ and $r = st$ correctly determines the infection type with prob. tending to* 1.

PROOF. The proof proceeds in a very similar way to Theorem 2. First suppose the infection is an epidemic. We can cover all true reporting nodes with probability scaling to 1 using the speed definition. Since at least an α fraction of the reporting nodes are truly infected, our algorithm correctly reports the infection is an epidemic. Therefore the Type II error probability decays to 0.

Now suppose the infection is a random sickness. Since the false positives are also random, the reporting nodes with the false positives are simply are larger set of random nodes. Note $r = b\left(\frac{1}{2(1+f)}\right)$. Using Lemma 1 in the same way as in Theorem 2, we see that no ball of radius r contains over a $\alpha = 1/(1+f)$ fraction of the random nodes with probability approaching 1. In this case, our algorithm returns random sickness. Thus the Type I error probability also tends to 0. □

4. MIXED INFECTION TYPES

Now we turn to the problem of mixed infection types. In this case, we deal with infections where the infection is spreading both as an epidemic and a random sickness. We term the nodes that become infected randomly as *seeds*, from which the infection starts spreading as an epidemic. We consider two distinct infection processes. In Process 0, the infection spreads mostly randomly. Let γ_0, η_0 be the infection rates for the random sickness and epidemic respectively and t_0 be the infection time for Process 0. For clarity, we also call Process 0 "Process SR-WE" (Strong random, weak epidemic). In Process 1, the infection is dominated by the epidemic, and let γ_1, η_1, and t_1 be the corresponding parameters as before. We label Process 1 "Process WR-SE" (Weak random, strong epidemic). Note that the infection is the same if the rates are scaled up by the same factor that time is scaled down. Then we can say that the epidemic dominates in Process 1 relative to Process 0 if $\eta_1/\gamma_1 \gg \eta_0/\gamma_0$. Unlike in the previous section, we apply no explicit normalization. Rather, we provide sufficient conditions on the range of the parameters for which the Median Ball Algorithm succeeds.

First we consider Type I errors. Assume the infection spreads by Process SR-WE [Process 0]. We use the Median Ball Algorithm with parameters α and r. Then the following theorem characterizes when the probability of error decays to 0. Let s and $b(x)$ be the speed and neighborhood size function as defined previously.

THEOREM 5. *Consider an infection spreading as in Process 0. Suppose $q\gamma_0 t_0 = \omega(\log n)$. Suppose there exists a constant integer $c_1 \geq 1$ where $\eta_0 t_0 = o\left((\gamma_0 t_0)^{-1/(1+c_1)}\right)$ and for some $\epsilon > 0$, suppose that $r + c_1 < b\left(\frac{\alpha}{\bar{d}^{c_1+1}(1+\epsilon)}\right)$. Then the Type I error probability decays to 0 as n increases.*

PROOF. First we show that no infection (from a single seed) spreads farther than a distance c_1, so each infection contains at most a constant $\bar{d}^{c_1+1}$ nodes (where, recall, $\bar{d}$ is a bound on the maximum degree of the graph). Consider an arbitrary seed a and all paths of length $c_1 + 1$ beginning at

a. There are at most $\bar{d}^{c_1+1}$ such paths. An infection from a must spread over one such path in time t_0 to spread farther than distance c_1. Since the traversal time of an edge has distribution $\text{Exp}(\eta_0)$, the probability the infection can spread over the edge in time t_0 is $1 - e^{-\eta_0 t_0} < \eta_0 t_0$. Then using a union bound, the probability that the infection spreads more than a distance c_1 is less than $(\bar{d}\eta_0 t_0)^{c_1+1}$. Let ϵ_2 satisfy $0 < \epsilon_2 < 1$. By hypothesis, the expected number of seeds is $\omega(\log n)$, so from standard concentration results, the number of seeds is at most $1 + (1+\epsilon)\gamma_0 t_0$ with probability tending to 1. Let P be the probability the infection spreads farther than distance c_1. Then from a final union bound,

$$\begin{aligned} P &< (1 + (1+\epsilon)\gamma_0 t_0)\,(\bar{d}\eta_0 t_0)^{c_1+1} \\ &= o\left(2\gamma_0 t_0 \bar{d}^{c_1+1}(\gamma_0 t_0)^{-1}\right) \qquad (1) \\ &= o(2\bar{d}^{c_1+1}). \end{aligned}$$

Eq. (1) follows from our hypothesis $\eta_0 t_0 = o\left((\gamma_0 t_0)^{-1/(1+c_1)}\right)$. Therefore, $P \to 0$ so the infection travels no more than a distance c_1 with probability tending to 1.

Now we need to show no ball of radius r contains over an α fraction of the reporting nodes. We first consider all infected nodes. Let $\epsilon > 0$ be a constant as specified in the theorem statement. For convenience, let $c_2 = \bar{d}^{c_1+1}$, the maximum number of nodes in a ball of radius c_1. Consider an arbitrary node a, and let $B_{\text{inner}} = Ball(a, r)$, $B_{\text{outer}} = \text{Ball}(a, r + c_1)$. Then from the previous result, any seed that has an infection that spreads to a node in B_{inner} must be inside B_{outer} (since it can only travel a distance c_1). By the hypothesis that $r + c_1 < b\left(\frac{\alpha}{c_2(1+\epsilon)}\right)$, $\text{card}(\text{B}_{\text{outer}}) < \frac{\alpha n}{c_2(1+\epsilon)}$. Let u be the number of seeds, so $u = \omega(\log n)$, again by hypothesis. Then from Lemma 1, the number of seeds within B_{outer} is less than $\frac{\alpha u}{c_2(1+\epsilon/2)}$ with probability greater than $1 - 1/n^2$. Each of these seeds infects less than c_2 nodes, so the total number of infected nodes within B_{inner} (which must all be from seeds in B_{outer}) is less than $\frac{\alpha u}{1+\epsilon/2}$. Hence, this ball contains less than a $\frac{\alpha}{1+\epsilon/2}$ fraction of the infected nodes.

Finally, we need to show the reporting process does not significantly impact the fraction of infected nodes seen in that ball. We consider an equivalent method of choosing the reporting nodes: first the number of reporting nodes is chosen (with the appropriate distribution), and then these are distributed uniformly over the infected nodes. Let Q be the number of reporting nodes. Then we need to find the probability that αQ reporting nodes are within B_{inner}. Let X be the number of reporting nodes in this region. As we just showed, the probability that any particular reporting node is within that region is at most $\frac{\alpha}{1+\epsilon/2}$. From a standard balls-in-bins arguement like in Lemma 1, since $\alpha Q = \omega(\log n)$, $P(X > \alpha Q) < 1/n^2$. That is, the probability that at least αQ of the reporting nodes are in that region is at most $1/n^2$.

Since each ball contains over an α fraction of the reporting nodes with probability no more than $1/n^2$, from a union bound, we find the probability that any of the n possible balls exceeds this bound is at most $1/n$. In this case, our algorithm correctly labels it 'Random'. Therefore, the Type I error probability decays to 0 as desired. □

Next consider the infection spreading by Process WR-SE [Process 1]. Define each of the parameters as before. Then we can characterize the range for which the Type II error goes to 0 as follows.

THEOREM 6. *Consider an infection spreading as in Process 1. Suppose $r > s\eta_1 t_1$, where s is the speed of the infection when it spreads at rate 1, and $\eta_1 t_1$ scales to infinity. Suppose $\alpha = o((1+\gamma_1 t_1)^{-1})$, and $\log(1/\alpha) = o(\eta_1 t_1)$. Then the Type II error probability decays to 0 as n increases.*

PROOF. First we show an upper bound on the number of seeds (recall seeds are the nodes randomly infected). The number of seeds is equal to one (the initially infected node) plus a Poisson random variable with mean $\gamma_1 t_1$. Let U be the set of seeds, and let $u = \frac{1}{\alpha}$. Since $\frac{1}{\alpha} = \omega(1 + \gamma_1 t_1)$, from the distribution, $u > \text{card(U)}$ with probability scaling to 1.

From the speed definition, there exists a constant λ_1 such that for each seed a,

$$P(W > s\eta_1 t_1) < e^{-\lambda_1 \eta_1 t_1},$$

where W is the radius of the infection starting at a. Now we apply a union bound to see that,

$$\begin{aligned} P(\exists a \in U s.t. W > s\eta_1 t_1) &< \frac{1}{\alpha} e^{-\lambda_1 \eta_1 t_1} \\ &< e^{\lambda_1 \eta_1 t_1/2} e^{-\lambda_1 \eta_1 t_1} \qquad (2) \\ &= e^{-\lambda_1 \eta_1 t_1/2} \to 0, \end{aligned}$$

where Equation 2 follows from the fact that $\log(1/\alpha) = o(\eta_1 t_1)$. Therefore, each seed spreads no farther than a distance r with probability tending to 1.

We now show that our algorithm returns 'Epidemic' in this case. Cover the seed with the largest (reporting) infection using a ball of radius r, which we showed covers the entire infection for that seed. Since there are at most $1/\alpha$ seeds total, the fraction of reporting infected nodes covered is at least $1/\frac{1}{\alpha} = \alpha$. Therefore, an α fraction of the reporting infected nodes has been covered a ball of radius r, so the Median Ball Algorithm returns 'Epidemic' as desired. □

The previous theorems establish the set of conditions sufficient for the algorithm to succeed. As the conditions are a little opaque, we summarize them here: *(i)* The total number of nodes that can be covered by a ball of radius $2r$ (where r increases with n) must scale a constant factor less than the total number of nodes times α. *(ii)* In Process SR-WE [Process 0], the expected number of reporting seeds must be order-wise more than $\log n$. *(iii)* In Process SR-WE, the infection spreads no more than a constant distance. *(iv)* For Process WR-SE [Process 1], the threshold r must be set large enough that a ball of radius r covers the largest infection (using the epidemic speed). *(v)* For Process WR-SE, the expected number of seeds must be order-wise less than α^{-1}. *(vi)* For Process WR-SE, α^{-1} must be order-wise less than exponentials in $\eta_1 t_1$.

Finally, recall we can choose the algorithm parameters α and r. Then the question is, when can we choose appropriate algorithm parameters so that the probability of error goes to 0? This is answered by the following theorem.

THEOREM 7. *Suppose there exists c_1 such that $\eta_0 t_0 = o\left((\gamma_0 t_0)^{-1/(c_1+1)}\right)$ and $q\gamma_0 t = \omega(\log n)$. Suppose $\eta_1 t_1 = \omega(\log(\gamma_1 t_1))$, $\gamma_1 t_1 = \omega(1)$, and $s\eta_1 t_1 = o\left(b(\frac{1}{\gamma_1 t_1})\right)$. Then the algorithm parameters can be chosen so that the probability of error tends to 0.*

Proof. We need to choose r and α so that $s\eta_1 t_1 < r < b\left(\frac{\alpha}{c_2(1+\epsilon)}\right) - c_1$ and $\alpha = o((\gamma_1 t_1)^{-1})$, $\log(1/\alpha) = o(\eta_1 t_1)$, where $c_2 = \bar{d}^{c_1+1}$. First we consider the conditions on α. Define an arbitrary slowly increasing function $g(n) = \theta(1)$, $g(n) = o(\gamma_1 t_1)$. This is possible since $\eta_1 t_1 = \omega(1)$. Choose $\alpha = (\gamma_1 t_1 g(n))^{-1}$. Then we have

$$\begin{aligned}\log(1/\alpha) &= \log(\gamma_1 t_1 g(n))\\ &< 2\log(\gamma_1 t_1)\\ &= o(\eta_1 t_1).\end{aligned}$$

Thus α satisfies the desired conditions. Now we will show it is possible to choose an appropriate r. By hypothesis, $s\eta_1 t_1 = o(b(\frac{1}{\gamma_1 t_1}))$. From our choice of α, for sufficiently large n, $\frac{\alpha}{c_2(1+\epsilon)} < \frac{1}{\gamma_1 t_1}$. Using the concavity of $b(x)$,

$$\begin{aligned}b\left(\frac{1}{\gamma_1 t_1}\right) &< \frac{\gamma_1 t_1}{\alpha/(c_2(1+\epsilon))} b\left(\frac{\alpha}{c_2(1+\epsilon)}\right)\\ &= o\left(b\left(\frac{\alpha}{c_2(1+\epsilon)}\right)\right). \qquad (3)\end{aligned}$$

Therefore, $s\eta_1 t_1 = o(b(\frac{\alpha}{c_2(1+\epsilon)}))$, with $s\eta_1 t_1 = \omega(1)$ by hypothesis. Thus it is clear r can be chosen with $s\eta_1 t_1 < r < b\left(\frac{\alpha}{c_2(1+\epsilon)}\right) - c_1$, for example by averaging each side. With this choice of parameters, the conditions of Theorem 5 and Theorem 6 are satisfied. Hence, both the Type I and Type II error probabilities will tend to 0. □

5. SIMULATIONS

In the previous sections, we show that the Median Ball Algorithm can distinguish epidemics and random sicknesses in both the cases of false positives, and when the infection process is mixed. In this section we illustrate via simulation the probability of error for reasonable graph sizes and parameters, and how it changes as the parameters are adjusted. While the theory developed so far applies to many types of graphs, here we specifically consider only one structure – grid graphs with wrapping edges – to explore various aspects (false positves, mixed infection) and for different parameter choices, within the page-length space constraints. Naturally, these graphs have bounded degree, and hence have the necessary properties to detect an infection. We use this graph to explore the non-asymptotic behavior of the Median Ball algorithm.

We performed these simulations with false positives and for mixed infections. We evaluate our algorithm by the empirical error probability, the average error probability for both Type I and Type II errors, weighting both equally. The results give insight in how the error probability is affected by graph topology, algorithm parameters, and infection time.

Each simulation was performed as follows. We used a grid graph with $n = 4900$, and infection time $t = 10$. The reporting probability was fixed at $q = 0.25$. The infection was simulated for 10000 trials for each infection processes (a random sickness and an infection), running the Median Ball Algorithm for each set of reporting nodes. We set the ball size parameter (r) to the optimal value as determined empirically. The other parameters were set as described in each caption. The probability of error is mostly plotted against the empirical expected fraction of infected nodes. That is, for each set of parameters, we estimated the expected number of infected nodes from the simulations, which was divided by n to determine the fraction infected. This expected fraction of infected nodes conveys the size of the infection, and hence the difficulty of the problem (since the task is more difficult the larger the infection is). Note that since $q = 0.25$, the expected fraction of reporting nodes is approximately 0.25 times as large. Finally the probability of error was estimated from the frequency at which the Median Ball Algorithm mischaracterized the type of infection.

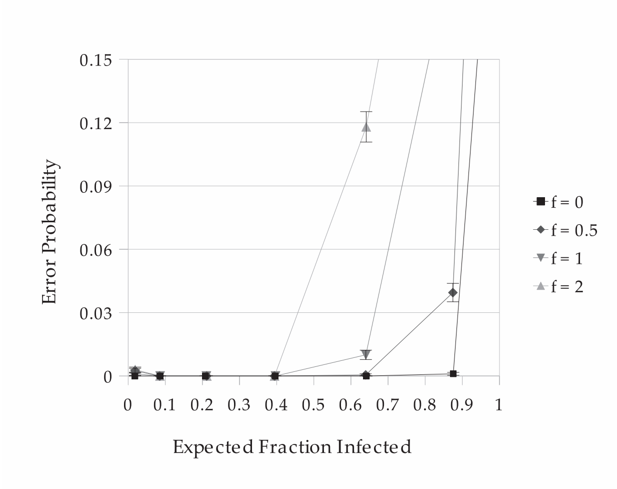

Figure 1: [False Positive Model] This figure shows the overall error probability, the sum of equally weighted Type I and Type II error rates, for a grid graph. The false positives were located randomly on the graph. The x-axis measures the expected fraction of nodes truly infected. As in our results, $\alpha = 1/(1 + f)$. The ball radius r was set to the optimal value empirically.

False Positives: Our first simulation results are on the probability of error for grid graphs for a variety of false positive frequencies. As in our analytical setting, the random sickness infection size was normalized to the same distribution as the epidemic as determined empirically. The results are shown in Figure 1.

The error probability is very low up to a very large number of truly infected nodes. It climbs fairly slowly as the number of false positives increases. Even when two-thirds of the reporting nodes are false positives, the error probability is low even up to an expected 40% of the network infected. Therefore our algorithm works very well in this setting.

Mixed Infection: Next, we present the simulation results for infections with mixed spreading regimes. Unlike for false positives, there is no direct normalization of the infection sizes. Rather, we adjusted the rates so that the infection sizes for both infection processes would be similar. This was done by first choosing the epidemic rate, and then empirically finding the random rate to three significant digits so that expected number of infected nodes hit a target value. This was done so that all the infections (for the various parameters) would be fairly comparable. Process SR-WE [Process 0] used an infection rate of 0.2.

Figure 2 shows the probability of error for various infection sizes. The infection rate for Process WR-SE [Process 1] is given on the x-axis. As expected, the larger the infection, the more difficult it is to use clustering to determine whether an infection is mostly random or mostly an epidemic. When

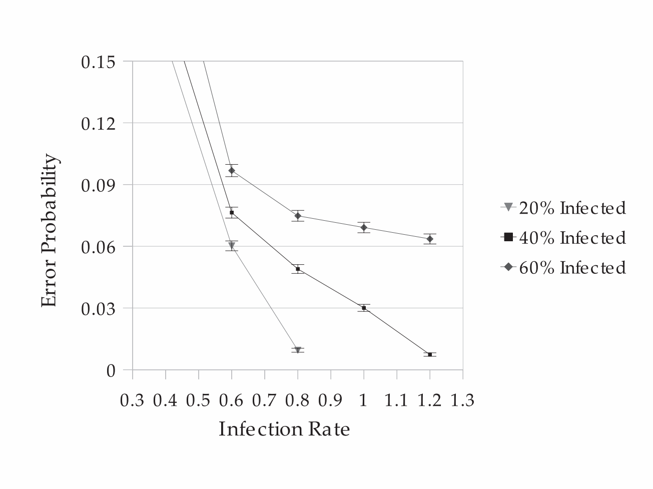

Figure 2: [Mixed Infection Model] This figure shows the overall error probability, the sum of equally weighted Type I and Type II error rates, for various expected fraction infected and Process WR-SE infection rates. The Process SR-WE infection rate is 0.2. The parameter $\alpha = 0.5$. The ball radius r was set to the optimal value empirically.

an expected 60% of the nodes in the network are infected, then the probability of error stays high, even for much larger infection rates. Note that there is a maximum infection rate before the target infection size is exceeded regardless of the random sickness rate. We used Process WR-SE infection rates close to that maximum.

Next we determine the effect of α on the probability of error. These results are shown in Figure 3. Surprisingly, changing α has a relatively small effect on the probability of error. The largest effect seen is using too large a value for larger Process WR-SE infection rates (when the probability of error is low). However, that is still relatively small. Then our algorithm seems fairly insensitive to the value of α.

6. ACKNOWLEDGMENTS

This work was partially supported by NSF Grants CNS-1017525, EECS-1056028, DTRA grant HDTRA 1-08-0029 and Army Research Office Grant W911NF-11-1-0265.

7. REFERENCES

[1] F. Ball and P. Neal. Poisson approximation for epidemics with two levels of mixing. *The Annals of Probability*, 32(1B):1168–1200, 2004.

[2] I. Benjamini and Y. Peres. Tree-indexed random walks on groups and first passage percolation. *Probability Theory and Related Fields*, 98:91–112, 1994.

[3] J. Cohen. Making headway under hellacious circumstances. *SCIENCE*, 313:470–473, July 2006.

[4] C. Corley, D. Cook, A. Mikler, and K. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7:596–615, 2010.

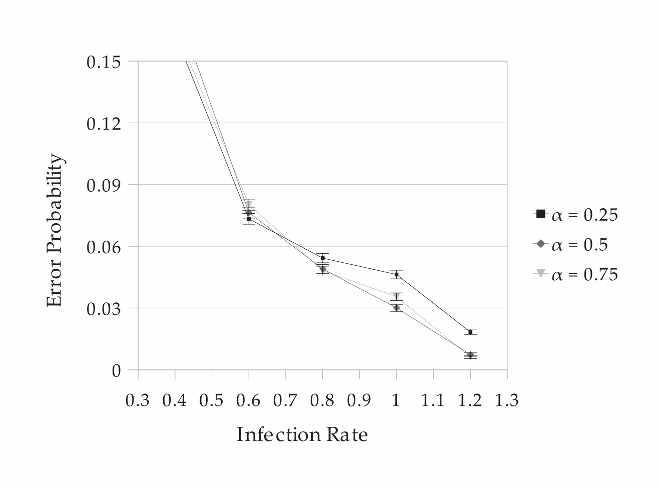

Figure 3: [Mixed Infection Model] This figure shows the overall error probability, the sum of equally weighted Type I and Type II error rates, for various values of α and Process WR-SE infection rates. The Process SR-WE infection rate is 0.2. The expected fraction infected was 40%. The ball radius r was set to the optimal value empirically.

[5] N. Demiris and P. D. O'Neill. Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Stat.*, 32:265–280, 2005.

[6] N. Demiris and P. D. O'Neill. Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Stat. Society Series B*, 67(5):731–745, 2005.

[7] R. Durrett. *Random Graph Dynamics*. Cambridge University Press, 2007.

[8] F-Secure. Bluetooth-worm:symbos/cabir, 2012. http://www.f-secure.com/v-descs/cabir.shtml.

[9] A. J. Ganesh, L. Massoulié, and D. F. Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM*, pages 1455–1466, 2005.

[10] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data*, 5(4):21:1–21:37, Feb. 2012.

[11] Google Flu Trends, http://www.google.org/flutrends/.

[12] A. Gopalan, S. Banerjee, A. Das, and S. Shakkottai. Random mobility and the spread of infection. In *Proc. IEEE Infocom*, 2011.

[13] H. Kesten. On the speed of convergence in first-passage percolation. *The Annals of Applied Probability*, 3(2):296–338, Nov 1993.

[14] W. Luo and W. P. Tay. Identifying infection sources in large tree networks. In *9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pages 281–289, June 2012.

[15] New York Times Bits Blog, http://bits.blogs.nytimes.com/2012/12/13/lookout-toll-fraud/.

[16] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai. Network forensics: random infection vs

spreading epidemic. *SIGMETRICS Perform. Eval. Rev.*, 40(1):223–234, June 2012.

[17] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai. On identifying the causative network of an epidemic. In *In Proceedings of 50th Annual Allerton Conference on Communication, Control, and Computing*, October 2012.

[18] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 33–41, New York, NY, USA, 2012. ACM.

[19] P. Netrapalli and S. Sanghavi. Learning the graph of epidemic cascades. *SIGMETRICS Perform. Eval. Rev.*, 40(1):211–222, June 2012.

[20] D. Shah and T. Zaman. Detecting sources of computer viruses in networks: Theory and experiment. *SIGMETRICS Perform. Eval. Rev.*, 86:203–214, 2010.

[21] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, 57, August 2011.

[22] J. Snow. *On the mode of communication of cholera.* John Churchill, 1855.

[23] Wikipedia. Commwarrior-a — Wikipedia, the free encyclopedia, 2012. [Accessed 30-Sept-2012].