On Identifying the Causative Network of an Epidemic

Chris Milling, Constantine Caramanis, Shie Mannor and Sanjay Shakkottai

***Abstract*— The history of infections and epidemics holds famous examples where understanding, containing and ultimately treating an outbreak began with understanding its mode of spread. The key question then, is: which network of interactions is the main cause of the spread? And can we determine the *causative network* without any knowledge of the epidemic, other than the identify of a minuscule subsample of infected nodes? This comes down to understanding the *diagnostic power of network information*. Specifically, in this paper we consider an epidemic that spreads on one of two networks. At some point in time, we see a small random subsample (perhaps a vanishingly small fraction) of those infected. We derive sufficient conditions two networks must have for this problem to be identifiable. We provide an efficient algorithm that solves the hypothesis testing problem on such graphs, and we characterize a regime in which our algorithm succeeds. Finally, we show that the condition we need for this identifiability property is fairly mild, and in particular, is satisfied by two common graph topologies: the grid, and the Erdös-Renyi graphs.**

I. Introduction

We are fast moving toward a setting where people and devices interact through multiple networks. With such multi-network interactions comes both costs and benefits. On one hand, viruses and sickness can spread on any of these networks, thus rendering the task of pinpointing the causative agent harder. On the other hand, opinions can now propagate over any of these networks, thus providing more effective mechanisms for influencing society. Concrete examples include smartphone viruses spreading over cellular networks through SMS/MMS messaging or by short-range Bluetooth communications, influence and product marketing occurring through online social networks or through mass media communications, and human disease spreading via multiple possible relationship networks (professional vs. personal).

This paper focuses on determining the causative network for the spread of an epidemic (e.g. virus, sickness, or opinion) from limited samples of the network state. In other words, given that only a small fraction of the nodes can be tested or measured, can we determine the network over which the epidemic is spreading? Inferring the network can be beneficial: For instance, at the very early stages of the AIDS pandemic in the 1980s, there was much misinformation on the types of human relations that could lead to infection transfer. In fact, at one point it was called the "4H disease" where 4H referred to "Haitians, Homosexuals, Hemophiliacs, and Heroin users" [1], [2], by simply typecasting the individuals who were known to be infected. In retrospect, we speculate that if finer-grained network knowledge (professional, personal, etc.) as is known today was then available, such crude typecasting (which was very detrimental to certain communities [2]) might have been avoided.

Another example lies in the domain of cellphone viruses. It is known that cellphone viruses can spread either by Bluetooth connections based on proximity [3], or by (randomly) sampling the address book (contact network) and sending infected MMS messages [4]. Here again, it would clearly be beneficial to determine the causative network, as the countermeasure strategies would be very different depending on the network of spread. These examples motivate our model and results, where we build on our prior work in [5] that distinguishes between random sickness and epidemics.

A. Setting and Results

We consider a collection of n nodes, and the nodes are interconnected by two graphs G_1 or G_2. In other words, the two graphs share the same nodes but have different edges. We assume that the two graphs are independent, i.e., the node indices on the pair of graphs are randomly permuted with respect to each other (see Section II for details). An epidemic propagates along the edges of one of these graphs using the standard SI model [6]. Given a sub-sample of the infected nodes on the graph, our objective is to determine the network over which the epidemic is spreading. In this paper, we address the "single-snapshot" problem, where we are not given the time at which the node was infected, but only the identity of the (sub-sampled) infected nodes. In addition to the number of samples, we are also interested in the interval of time over which we can distinguish between the two networks[1]. Our main contributions are as follows:

(i) **Algorithm:** We develop the Comparative Ball Algorithm for inferring the causative network. This algorithm builds on the intuition that infected nodes are clustered more strongly on the true causative network. The algorithm outputs the causative network as the one with a smaller diameter-normalized ball that best fits the collection of infected node samples.

C. Milling, C. Caramanis and S. Shakkottai are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, USA, Emails: cmilling@utexas.edu, caramanis@mail.utexas.edu, shakkott@austin.utexas.edu. S. Mannor is with the Department of Electrical Engineering, Technion, Israel, Email: shie@ee.technion.ac.il. This work was partially supported by NSF Grants CNS-1017525, CNS-0721380, EFRI-0735905, EECS-1056028, DTRA grant HDTRA 1-08-0029 and Army Research Office Grant W911NF-11-1-0265.

[1] As we discussed in [5], once sufficient time has elapsed such that most or all the nodes are infected, it is impossible to determine the causative network. Thus, we are interested in determining the largest possible time at which the snapshot is taken, when we can still distinguish the networks. On the other end for small values of time, our interest in determining how few samples are needed to reliably find the causative network.

(ii) **General Graphs:** For any pair of graphs G_1, G_2 that both satisfy: *(a) Speed condition* – the epidemic ball radius increases linearly in time, and *(b) Spread condition* – a randomly selected collection of nodes are sufficiently spread apart, we derive conditions on the number of samples and the time interval over which we can determine the causative network (as $n \to \infty$ and with high probability).

(iii) **Grids and the Erdös-Renyi Random Graphs:** For d-dimensional grids, and the giant component of the Erdös-Renyi random graph (with constant asymptotic average degree), we derive bounds on the parameters associated with the speed and spread conditions, thus, providing sufficient conditions on the regime where we can determine the causative network.

B. Related Work

This paper builds on the approach and results in our prior work [5], where we derived sufficient conditions on time interval and number of samples to distinguish between a random sickness (which does not depend on network structure) and an epidemic (e.g., allergy vs. flu). As in [5], our methods strongly rely on first-passage percolation on discrete structures, specifically, shape and speed theorems for grids and random graphs [7], [8].

Epidemic spread and inference has been well studied in various contexts [6], [9], [10], [11], [12]; however, we are unaware of prior analytical work that attempts to determine the causative network as in this paper.

C. Outline of the Paper

The paper is organized as follows. In the following section, Section II, we precisely define the problem and the infection model. In addition, we specify the Comparative Ball Algorithm that we analyze in the rest of the paper. Section III provides success criteria for generic graphs. That is, we show that if these criteria are satisfied, then our algorithm determines the true spreading mechanism with high probability. In Section IV, we analyze two standard graph topologies, grids and Erdös-Renyi graphs, and show that they satisfy the aforementioned criteria. Finally, Section V contains simulation information and illustrates empirically the performance of our algorithm on these graphs.

II. Model and Algorithm

We consider a collection of n nodes (vertices V) which are members of two different networks (graphs). These graphs are denoted by $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$; they share the same vertex set but have different edge sets. For example, G_1 could represent the n vertices arranged on a $d-$dimensional grid, and G_2 could be an Erdös-Renyi graph. Note that G_2 does *not* need to have qualitatively different structure from G_1: Indeed G_2 could also be a $d-$dimensional grid, but with a different node-to-edge mapping.

Spread Model: We study the situation where an epidemic is propagating on one of these two graphs, and the objective is to determine on which network it is spreading. From our previous discussion, this 'epidemic' could model many situations, including the spread of a cellphone virus, physical sickness of humans, and opinions or influence about products or ideas.

Given that the epidemic is on graph G_i, the spread occurs as follows (the standard SI dynamics [6]). A node is randomly selected to be the epidemic seed, and a random variable associated with it is set to '1' (all other nodes are set to '0'). Nodes with value '1' are referred to as infected nodes. Associated with each edge on the graph G_i are independent exponential random variables, each with mean '1'. These represent the transit time of the infection across that edge – a random variable. An infected node proceeds to infect its neighbors, with each non-infected neighbor getting infected after the random transit time associated with the edge between the infected node and this neighbor. This process proceeds until the entire graph G_i is infected.

Reporting Model: At some fixed time t, a sub-sample of the infected nodes report their infection state independently, with some probability $q < 1$. We denote $S_r(n)$ to be the set of reporting infected nodes. We henceforth suppress the dependence on n, i.e., we use the notation S_r unless otherwise specifically needed for clarity.

Graph Structure: Clearly, we require that the graphs G_1 and G_2 be "different enough" if we would want to reliably determine the causative network. As an extreme example, if the graphs were identical, clearly there is no hope of distinguishing between the two.

In this paper, we require that corresponding nodes on the two graphs have independent neighborhoods.[2] This is a condition that approximately holds in typical settings. Consider for instance the several hundred "nodes" (people, or devices) that come within blue-tooth range during a walk through the mall. This list likely has extremely small overlap (possibly only the few friends accompanying us on the mall excursion) with the set of nodes that send us e-mail or SMS on a regular basis.

One construction that has this property, and the one we assume for our results in the sequel, is as follows. We start with two graph topologies G_1 and G_2 of the same size n, and *whose nodes are unlabeled*. Then we randomly label the nodes of graph G_1 from '1' to 'n' uniformly. Likewise, and independently, we label the nodes of graph G_2. Now nodes of the same label are considered the same entity (person, device), i.e., if a node on one graph is infected, the corresponding node on the other graph is also infected. With this labeling, the graphs can be considered independent. A key property that results from this model is that clustered nodes in one graph are likely to be separated apart on the other.

Objective: Given the two graph topologies, G_1 and G_2, and a sample of infected nodes S_r, our objective is to design an algorithm that (asymptotically) correctly determines which

[2]We note that we can envision other conditions based on clustering of epidemics on the two graphs, which also serves as alternate sufficient conditions. For simplicity, we restrict ourselves to the 'random node index' condition in this paper.

graph the epidemic is spreading on.

By 'asymptotically correct' we mean the following. Observe that in this setting, there are two types of error events: the epidemic is spreading on G_1 but the algorithm outputs G_2, and vice versa. The algorithm is said to be asymptotically correct if the probabilities of each of these error events go to zero as $n \to \infty$.

A. The Comparative Ball Algorithm

Given a graph G_i, a node v, and a radius r, we denote by $Ball_{v,r}(G_i)$ the collection of all nodes on the Graph G_i that are at most a distance r from node v (graph distance measured by hop-count). Further, the diameter of the graph is denoted by $diam(G_i)$. Given any collection of nodes S, we now denote by $Ball(G_i, S)$ the smallest-radius ball that contains all the nodes in S, and we let $RadiusBall(G_i, S)$ denote its corresponding radius.

We term our algorithm the 'Comparative Ball Algorithm'. For each graph, we find the smallest ball on that graph that contains all the reporting infected nodes. We take the ratio of the radius of this ball to that of its diameter. These ratios (called the *score* of each graph) serve as a topology independent measure of clustering on each graph. The Comparative Ball Algorithm returns the graph with the smallest normalized clustering ratio. This is formally described below.

Algorithm 1 Comparative Ball Algorithm

Input: Two graphs, G_1 and G_2; Set of reporting infected nodes S;
Output: G_1 or G_2

$a_1 \leftarrow RadiusBall(G_1, S)$
$b_1 \leftarrow diam(G_1)$
$x_1 \leftarrow a_1/b_1$
$a_2 \leftarrow RadiusBall(G_2, S)$
$b_2 \leftarrow diam(G_2)$
$x_2 \leftarrow a_2/b_2$
if $x_1 \leq x_2$ **then**
  return G_1
else
  return G_2
end if

III. Main Result: General Graphs

Now we classify the types of graphs for which this algorithm can determine which graph an infection has spread on. The key criteria for these graphs are as follows. First, an infection spreading over a graph must be localized and clustered for sufficiently small times t. Second, when a sufficient number of random nodes are infected, then they must be spread out over a relatively large area of the graph with high probability. That is, the smallest ball containing a set of infected nodes should be small, and the smallest ball containing a set of random nodes should be large.

We call a graph G with diameter $diam(G)$ *detectable* if it satisfies the following conditions with probability approaching 1 as n increases, for some positive constants s_G, b_G, and β_G:

Speed Condition: There is a speed s_G such that for an infection S_r at time t, $RadiusBall(G, S_r) < s_G t$.
Spread Condition: A random set S of nodes on G, with $card(S) > \beta_G \log n$, satisfies $RadiusBall(G, S) > b_G diam(G)$.

Now we prove that these conditions are sufficient to distinguish infections on any two such graphs. These conditions are fairly mild. In Section IV we show that two commonly encountered, standard types of graphs satisfy these properties: $d-$dimensional grids and Erdös-Renyi graphs.

Theorem 1: Consider detectable graphs G_1 and G_2 and infection times t such that the number of reporting infected nodes scales at least as $\max(\beta_{G_1}, \beta_{G_2}) \log n$. Then if $t < b_{G_2} diam(G_1)/s_{G_1}$, the Comparative Ball Algorithm correctly identifies an infection on G_1 with probability approaching 1. In addition, if $t < b_{G_1} diam(G_2)/s_{G_2}$, then the Comparative Ball Algorithm correctly identifies an infection on G_2 with probability approaching 1.

Proof: Suppose we have graphs G_1, G_2, and infection time t as given in the theorem statement. By symmetry, it is sufficient to prove that an infection is detected on G_1. Then suppose we have such an infection on G_1 with reporting nodes S_r, where $card(S_r) > \beta_{G_2} \log n$. Note that by the independence assumption, this set of nodes is randomly distributed over G_2. Using the properties for a graph to be detectable, we see that with probability approaching 1, $RadiusBall(G_1, S_r) < s_{G_1} t$ and $RadiusBall(G_2, S) > b_{G_2} diam(G_2)$. Then the score for the first graph satisfies $x_1 < s_{G_1} t/diam(G_1) < b_{G_2}$ by hypothesis. Similarly, $x_2 > b_{G_2} diam(G_2)/diam(G_2) = b_{G_2}$. Therefore, the algorithm correctly identifies an infection. ■

IV. Speed and Spread Conditions: Grids and the Erdös-Renyi Graph

In this paper, we consider two types of graphs: the d-dimensional grid, and the Erdös-Renyi graph. The d-dimensional grid graph represents a contact graph where the infection spreads between nodes in spatial proximity (e.g., the Bluetooth virus, human sickness). The second topology is an Erdös-Renyi graph, a random graph forming a network with low diameter. This topology models an infection spreading over long distance networks, such as the Internet or over social networks. We show that both of these networks are detectable, and hence that the Comparative Ball Algorithm works on these graphs. The calculations in this section directly follow from the results in [5]. We however elaborate on the proofs in this section for clarity and provide simpler alternate proofs.

A. $d-$Dimensional Grids

First we consider a grid graph, modeling infections that spread geographically through proximal connections. Let the graph $G = Grid(n, d)$ be such a grid network with n

nodes and dimension d, so the side length is $n^{1/d}$. We avoid edge effects by wrapping around the grid (a torus). This avoids dealing with distracting complexities resulting from the choice of the initial source of the infection.

Now we establish two results on the behavior of infections on this topology, which thus shows that this type of graph is detectable. First, we establish limits on the speed of the infection after time t has passed. Next, we show lower bounds on the spread, i.e., the ball size needed to cover a random selection of nodes of sufficient size.

Proposition 1: Let $G = Grid(n, d)$ and time t scaling without bound as n increases. Then there exists a constant μ such that

$$RadiusBall(G, S_r) < 1.1d\mu t,$$

with probability converging to 1.

Proof: The proof follows from results in first passage percolation [8], and is available in Theorem 1, [5]. ∎

The following theorem provides a lower bound on the radius of the ball needed to cover a collection of random nodes uniformly selected from the grid. We require that the number of random nodes grows at least as $\log n$.

Proposition 2: Let $G = Grid(n, d)$. Let S_r be a collection of random nodes in G, such that $card(S_r) > \log n$ for sufficiently high n. Then

$$RadiusBall(G, S_r) > n^{1/d}/4,$$

with probability converging to 1.

Proof: We present a similar result in previous work [5], but the more relaxed conditions allow us to demonstrate this proposition with a simpler and clearer proof. By assumption, we have a set S_r of $c = card(S_r)$ random nodes, and seek to show the probability they are all within some ball of radius $n^{1/d}/4$ decays to 0 with n. Then consider one of the n such balls. There are less than $L = (n^{1/d}/2)^d$ nodes in that region (the number of nodes in a 'box' of side $n^{1/d}/2$). Then we see within this ball, there are at most $\binom{c}{L}$ arrangements of the sick nodes out of $\binom{c}{n}$ total possible arrangements. Therefore, the probability all the sick nodes are within the region is no more than

$$\begin{aligned}\binom{L}{c}\Big/\binom{n}{c} &= \frac{L!(n-c)!}{(L-c)!n!} \\ &\leq (L/n)^c. \end{aligned} \tag{1}$$

Using a union bound, we find that the probability there is a ball of that size containing all nodes in S_r is at most $n(L/n)^c$. Then

$$\begin{aligned} n(L/n)^c &< n\left(\frac{1}{2^d}\right)^{\log n} \\ &= n^{1-d\log 2} \\ &\to 0. \end{aligned} \tag{2}$$

Therefore, $RadiusBall(G, S_r) > n^{1/d}/4$ with probability converging to 1. ∎

Since the diameter of a grid is (nearly) $d/2n^{1/d}$, we see that a grid satisfies both the speed condition (Theorem 1) and the spread condition (Theorem 2), and hence is detectable. Therefore, the Comparative Ball Algorithm performs well on grid graphs.

B. Erdös-Renyi Graphs

Now we consider Erdös-Renyi graphs, representing infections that spread over low diameter networks (the diameter grows logarithmically with network size). An Erdös-Renyi graph is a random graph with n nodes, where there is an edge between any pair of nodes, independently with probability p. We study the Erdös-Renyi graph in the regime where $p = c/n$, for some positive constant $c > 1$. It is known that in this regime, the graph is disconnected, but there exists a giant component with $\Theta(n)$ nodes with high probability in the large n regime. In this paper, we restrict our attention to epidemics on the giant component (otherwise the problem is trivial). Thus as in [5], we limit both the infection and the random set of reporting nodes (due to the labeling when the infection occurs on the alternative graph) to occur exclusively on the giant connected component. If the infection on the other graph contains too many nodes for the giant component, we simply ignore the excess, but this point is already outside the regime of interest.

Like before, we establish two results in this section. The first theorem proves an upper bound on the ball size for an infection up to a limited time. Next, we demonstrate a lower bound on the ball size for a random collection of nodes.

Proposition 3: Let $G = G(n, p)$ and time t scaling with n. Then there exists a constant C_2 such that

$$RadiusBall(G, S_r) < C_2 t,$$

with probability converging to 1.

Proof: Roughly speaking, this theorem states that there is a maximum speed at which the infection can travel on an Erdös-Renyi graph. The statement follows from a similar maximum speed result for trees [13], and the proof is provided in our previous work, [5], Theorem 4. However, we now provide additional detail on exactly how one can transition from the result for trees to one for Erdös-Renyi graphs.

To do this, we upper bound an infection on an Erdös-Renyi graph by a tree that represents the routes on which an infection can travel. Since an Erdös-Renyi graph is locally tree-like [14], we expect this approximation to be fairly accurate for low times, though this is not necessary for the proof. Then consider the tree formed as follows. The root of the tree is the initial infected node. The next level contains copies of all nodes adjacent to the original node in the Erdös-Renyi graph. Each of these is connected to copies of their neighbors, and so on. Note all nodes can (and likely do) have multiple copies.

Now consider the induced set of infected nodes, $\tilde{S}_r$, as the set of nodes which have copies that are infected. Since the distance of a copy from the root is no less than the distance from the original node to the original root, we see that the distance the infection has traveled on the tree is no less than the distance the infection traveled on $\tilde{S}_r$. Now we show that

the $\tilde{S}_r$ stochastically dominates the true infected set S. That is, for all sets T, $P(T \subset \tilde{S}_r) \geq P(T \subset S_r)$. This follows from the fact that the transition rate from a set of infected nodes T to the set $T \cup \{x\}$ is higher for the induced set for all sets T and nodes $x \notin T$. On the real graph, the transition rate is simply the number of nodes in T adjacent to x. For the induced set, the rate depends on the exact set of infected copies, and is equal to the number of copies of nodes in T adjacent to x. Since each node in T has at least one infected copy, we see the number of copies is always no less than the number of nodes.

Thus, we find that as stated, the transition rates are universally equal or higher for the induced set, from which the stochastic dominance result follows. Hence, $RadiusBall(G, S_r)$ is also stochastically dominated by $RadiusBall(G, \tilde{S}_r)$, and the latter is upper bounded by the depth of the tree, which using the speed result, is bounded by $C_2 t$ for some speed C_2. That is,

$$RadiusBall(G, S_r) < C_2 t.$$

■

Next, we use the neighborhood sizes on this graph to provide a lower bound to the ball size needed to cover a random infection.

Proposition 4: Let $G = G(n, p)$. Let S_r be a collection of random nodes in G, such that $card(S_r)$ scales at least with $\log n$. Then

$$RadiusBall(G, S_r) > \frac{\log n}{3 \log c},$$

with probability converging to 1.

Proof: This statement is also proved in our prior work [5], but we now provide much greater detail. We proceed by bounding the probability that all the random nodes are within a ball of radius m. This is possible only if all nodes in S_r are within distance $2m$ from any given node in S_r. Now, the number of nodes within a distance $2m$ from a given node is no more than $16m^3c^{2m} \log n$ with probability $1 - o(n^{-1})$ [15]. Then the probability of all nodes fitting inside one such ball is at most

$$\left(\frac{16m^3c^{2m} \log n}{n}\right)^{card(S_r)-1} < \left(\frac{16m^3c^{2m} \log n}{n}\right)^{\log n - 1}.$$

Then this decays to 0 at least as fast as n^{-1} if

$$\frac{16m^3c^{2m} \log n}{n} < n^{-1/\log n}.$$

Finally we set $m = \frac{\log n}{3 \log c}$ as desired. Hence $c^{2m} = n^{2/3}$. Using this substitution, the above term reduces to

$$\begin{aligned} \frac{16m^3c^{2m} \log n}{n} &= \frac{16m^3 n^{2/3} \log n}{n} \\ &= \frac{16(\log n)^4}{27(\log c)^3 n^{1/3}} \\ &< (\log n)^4 n^{-1/3} < n^{-1/\log n} \end{aligned} \quad (3)$$

for sufficiently large n. Therefore, $RadiusBall(G, S_r) > \frac{\log n}{3 \log c}$ with probability converging to 1. ■

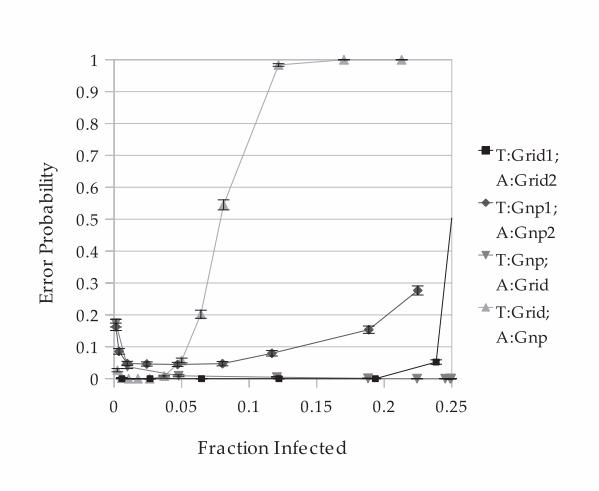

Fig. 1. This figure shows the error probability for the algorithm on pairs of standard graphs. Various (conditional) error probabilities are illustrated – 'T:' corresponds to the true network, and 'A:' corresponds to the algorithm output.

The diameter of the giant component of an Erdös-Renyi graph is $\Theta(\log n / \log c)$ [14]. Thus, Theorem 3 and Theorem 4 establish that an Erdös-Renyi graph satisfies both the speed and spread conditions, and hence is detectable.

V. Simulations

We have demonstrated that as the graph size increases, eventually the probability of mistaking which graph an infection spreads on, decays toward zero. We simulated the performance of the Comparative Ball Algorithm to evaluate the performance empirically. We determined the error rate over a range of t for several pairs of graphs. We evaluated the two different standard graph topologies considered earlier, grids and Erdös-Renyi graphs.

We simulated the infections on various pairs of the graphs over a range of times. In order to portray the results in a comparable way, we plotted the error rate versus the average infection size instead of time. This is necessary because different times result in very different infection sizes for the different graphs. That is, the infection is large even at low t on an Erdös-Renyi graph, and vice versa for a grid graph. This would introduce a misleading effect in the results.

Each node in the graphs received a random label to ensure independence. We use $n = 4000$ for each graph with $q = 0.25$. For the Erdös-Renyi graphs, we use $p = 2/4000$. The probability of error was computed over 1000 trials. There are two possible types of errors in each simulation, when the infection spreads on the first graph, and when it spreads on the second. We label the error event 'T:G_1; A:G_2' for the error where the infection in fact travels on graph G_1 (True event), but the algorithm incorrectly labels it as occurring on graph G_2 (Algorithm output).

The results of these simulations are shown in Figure 1. Note that up to about 5% of the network reporting an infection, the error rates are low in all cases. The error rates are consistently low for the 'T:Grid1;A:Grid2' comparison

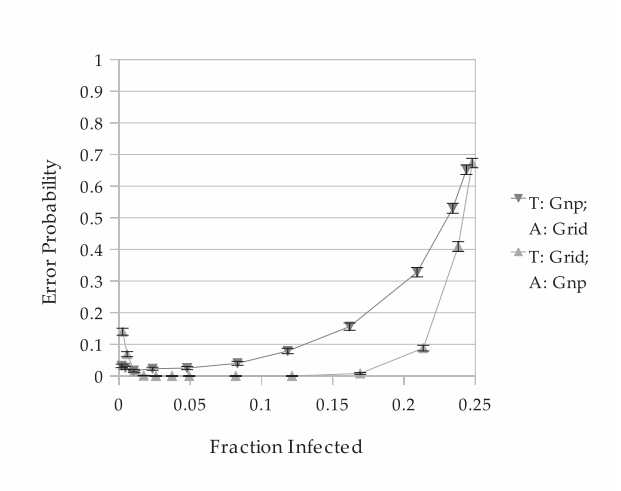

Fig. 2. This figure shows the error probability for the G(n,p) vs. Grid graphs for the scaled diameter setting (diameter of G(n,p) graph is scaled by 1.7).

up to the point where the whole network is infected. When comparing a grid and an Erdös-Renyi graph, there is a bias to label it an Erdös-Renyi graph at higher times, causing the 'T:Grid;A:G(n,p)' error to be very high and conversely, the 'T:G(n,p);A:Grid' error to be very low. This suggests that by simply modifying the Comparative Ball Algorithm to normalize with respect to a scaled graph diameter (where the scaling parameter would be graph dependent), we could balance these two error probabilities, and thus result in improved performance. To illustrate, by choosing a diameter scaling value of 1.7 for the Grid graph, the plot in Figure 2 indicates that one could distinguish between G(n,p) and Grid graphs for a significantly larger range. We plan to study a systematic approach for such scalings as future work.

VI. Conclusions

When an infection/virus is seen spreading over a group of people/machines, one may have multiple possibile spreading regimes for the infection in mind, and want to know which the infection is most likely travelling on. We have shown that this is possible to do with high accuracy if the regimes are independent and satisfy two properties: 1) An infection spreading according the regime should be localized in the contact graph, and 2) A random set of nodes should be spaced far apart on the graph. When these conditions are satisfied (in the sense given in this paper), the correct spreading regime can be detected accurately with high probability by determining on which graph the infection appears to be more clustered. In addition, we have shown two standard types of graphs, grids and Erdös-Renyi graphs, satisfy these properties. Our simulations here demonstrate the efficacy of our algorithm.

We are currently extending this work in several directions. First, in this paper, we have assumed that the infection spreads at rate 1 between all connected nodes. However, in real circumstances, the infection may travel faster between some pairs of nodes than others. Then, we would need to consider a weighted contact graph. This requires understanding how the weights affect the shape of the 'ball' containing the infected nodes. Another necessary extension is improving the robustness of the algorithm. If a single node falsely reports being infected, the ball size used in the algorithm can be wildly distorted. This would require appropriately filtering the reporting nodes to remove outliers.

References

[1] Wikipedia, "Hiv/aids — Wikipedia, the free encyclopedia," 2012, [Accessed 30-Sept-2012]. [Online]. Available: http://en.wikipedia.org/wiki/HIV/AIDS

[2] J. Cohen, "Making headway under hellacious circumstances," *SCIENCE*, vol. 313, pp. 470–473, July 2006.

[3] F-Secure, "Bluetooth-worm:symbos/cabir," 2012, [Accessed 30-Sept-2012]. [Online]. Available: http://www.f-secure.com/v-descs/cabir.shtml

[4] Wikipedia, "Commwarrior-a — Wikipedia, the free encyclopedia," 2012, [Accessed 30-Sept-2012]. [Online]. Available: http://en.wikipedia.org/wiki/Commwarrior-A

[5] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai, "Network forensics: random infection vs spreading epidemic," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 223–234, June 2012.

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

[7] R. Lyons and R. Pemantle, "Random walk in a random environment and first-passage percolation on trees," *The Annals of Probability*, vol. 20, no. 1, pp. 125–136, 1992.

[8] H. Kesten, "On the speed of convergence in first-passage percolation," *The Annals of Applied Probability*, vol. 3, no. 2, pp. 296–338, Nov 1993.

[9] G. Streftaris and G. J. Gibson, "Statistical inference for stochatic epidemic models," in *Proc. 17th International Workshop on Statistical Modeling*, 2002, pp. 609–616.

[10] N. Demiris and P. D. O'Neill, "Bayesian inference for stochastic multitype epidemics in structured populations via random graphs," *Journal of the Royal Statistical Society Series B*, vol. 67, no. 5, pp. 731–745, 2005.

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

[12] ——, "Rumors in a network: Who's the culprit?" *IEEE Transactions on Information Theory*, vol. 57, August 2011.

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

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

[15] F. Chung and L. Lu, "The diameter of sparse random graphs," *Adv. in Appl. Math*, vol. 26, pp. 257–279, 2001.