Localized Epidemic Detection in Networks with Overwhelming Noise

Eli A. Meirom†, Chris Milling†, Constantine Caramanis♯, Shie Mannor†, Sanjay Shakkottai♯, Ariel Orda†
† Dept. Of Electrical and Computer Engineering, University of Texas, Austin, Texas, USA
♯ Dept. Of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

ABSTRACT

We consider the problem of detecting an epidemic in a population where individual diagnoses are extremely noisy. We show that exclusively local, approximate knowledge of the contact network suffices to accurately detect the epidemic. The motivation for this problem is the plethora of examples (influenza strains in humans, or computer viruses in smartphones, etc.) where reliable diagnoses are scarce, but noisy data plentiful. In flu or phone-viruses, exceedingly few infected people/phones are professionally diagnosed (only a small fraction go to a doctor) but less reliable secondary signatures (e.g., people staying home, or greater-than-typical upload activity) are more readily available.

Our algorithm requires only local-neighbor knowledge of this graph, and in a broad array of settings that we describe, succeeds even when false negatives and false positives make up an overwhelming majority of the data available. Our results show it succeeds in the presence of partial information about the contact network, and also when are many (hundreds, in our examples) of initial patients-zero.

1. INTRODUCTION

Any study, analysis or curbing of an epidemic begins with the fundamental question: is the phenomenon that we experience indeed an epidemic? The identification of such processes, be they malicious malware spreading on computer networks, memes on social networks, or trends in public opinion, is of major interest across several disciplines.

A common characteristic is a stark absence of reliable information: patients with flu-like symptoms rarely visit a medical professional. More typical, however, is the availability of significant amounts of information indicating a secondary signature of the outbreak: flu patients may stay home from work, infected computers may exhibit somewhat encumbered performance, adopters of a new technological trend (e.g., smart-watch) may show new usage patterns, or may simply “self-report.” While more plentiful, such data may be riddled with false positives and negatives. Further complicating the problem is the fact that the underlying network is rarely fully known and in practice it may be possible to recover only a local subgraph near any infected node. In an epidemic setting, there may be multiple epidemic sources rather than a single “patient zero.”

How can we obtain accurate population-diagnoses from unreliable single-node information, and only local (and possibly approximate) knowledge of the contact network? We consider the basic problem in the most dire setting: at some instant in time we are informed that a given subset of nodes exhibiting unusual behavior (in the computer setting) or are sick. Exact reporting times are inaccessible. Moreover, we assume we are unable to observe the time evolution of the sickness reporting process. Even if there is an epidemic, we may have a large (possibly an overwhelming) fraction of false positives and false negatives among the reporting subset and its complement, respectively. Given this single (and very noisy) snapshot in time, the statistical inference problem is to determine if there is an epidemic which propagates through the network from one or possibly many initial nodes, or rather nodes have become infected via an independent, external mechanism.

We do not wish to require each node to know more than its local neighborhood (and in fact, our simulations show that this too need only be approximate). We provide the only algorithm, to the best of our knowledge, that has these key properties. Our theoretical results show that our algorithm succeeds in successfully solving the hypothesis testing problem even when the fractions of false negatives and positives go to 1. Our simulation results corroborate this finding.

2. PROBLEM DEFINITION

Consider a graph $G = (V, E)$, where the number of nodes is $N = |V|$. The decision making task is distinguishing between the two following alternative scenarios.

An epidemic: At time $t = 0$, an arbitrary subset of nodes is infected. Infected nodes infect their neighbors according to an exponential clock set on each edge. At this time, each infected node reports it is infected with probability $q(N)$. The number of reporting nodes is $|S_r|$. The non-reporting nodes are false negatives. In addition, there are $f(N)|S_r|$ nodes, picked uniformly over the network, that also report infected. Each node that is not infected yet reports infection constitutes a false positive.

Uniform reporting: Each node, independently of all others, reports an infection with some probability. The problem is most challenging when the expected numbers of reporting
nodes in each scenario match, as otherwise simple counting would solve the problem.

When the reporting probability increases there exists a large connected component with a ball-like shape about the set of “patient-zero”s, the initial sources of the epidemic, and the problem is easier. Likewise, the problem is more difficult when the ratio of the number of falsely reporting to the number of truly reporting nodes, \( f(N) \), tends to infinity, as the truly reporting nodes are washed out by the sea of falsely reporting nodes. We describe an algorithm that is shown to converge correctly even in settings where the reporting probability tends to zero and \( f(N) \to \infty \).

While we present the epidemic dynamics using the concrete example of an SI (susceptible-infected) model, our algorithm may be applied to other infection models as well. Indeed, the algorithm and proofs are diffusion-model-agnostic — they assume no knowledge of the diffusion mechanics.

The problem we consider here was first introduced in [1]. However, the algorithm and analysis used in this paper is fundamentally different. In [1], the authors rely on tools from first passage percolation, which in some settings (for some graph topologies, for initial infection conditions) may be fragile. For example, it is not clear if the algorithms given there are able to handle large numbers of initial seeds. Additionally, our algorithm is inherently local and easily parallelizable. Most significantly, our algorithm requires only local knowledge of the graph: each infected node is required to know only information about its local neighborhood.

3. ALGORITHM: HOTSPOT AGGREGATOR

The intuition is simple: neighbors of infected nodes are more likely to be infected in an epidemic, when compared to the baseline of random illness. Let \( H^{NN}(K, s) \) be an indicator equal to one if there are more than \( s \) reporting nodes within the the \( K \) nearest neighbors of node \( i \).

If the indicator evaluates to 1, we call node \( i \) a hotspot, and the sum of these indicators gives the number of hotspots. Note that the evaluation of these, and hence the counting of hotspots, can be done in a distributed fashion, with each reporting infected node requiring only local network information. Our Hotspot Aggregator Algorithm counts the number of hotspots, and then performs a threshold test: if there are more than \( T \) hotspots, it declares an epidemic.

The algorithm depends on the local neighborhood size \( K \), and the threshold \( T \). The theorems that make up our main results [2], demand a choice for \( K \) and \( T \) that requires only approximate knowledge of the false positive rate and the graph topology. Computing the correct values for \( K \) and \( T \) is straightforward for essentially all topologies, and in fact, for some topologies the values for \( K \) and \( T \) can be computed analytically [2]. Our experimental results suggest that the algorithm’s performance is quite stable. For instance, choosing \( K \) corresponding to the \( r \)-hop neighborhood, for \( r = 2, 3 \) yields uniformly good computational results.

Algorithm Information and Complexity. The complete graph information (its global structure) requires \( O(N^2) \) to encode. In contrast, our algorithm requires only the number of reporting nodes within each local environment. For many cases of interest, \( N_{\text{reporting}} \) can be as small as \( O(\log(N)) \) or even lower. The algorithm’s complexity is \( O(N_{\text{reporting}}) \) and it scales with the number of reporting nodes (which, again, is typically vanishingly small compared to the number of nodes in the network).

Algorithm Performance. The conditions on the size of the local neighborhoods our algorithm must consider depend on level of noise. We show that our algorithm succeeds even in the extreme setting where the number of falsely reporting nodes is \( \Theta(N) \) and the number of truly reporting nodes is \( \omega(1) \). In this case, we require \( \Theta(\log N) \)-size local neighborhoods in order to guarantee convergence. If there are only \( o(N) \) falsely reporting nodes, we show we can do much better; a local environment containing only a constant number of nodes is sufficient.

Statistical Performance: We provide sufficient conditions w.r.t. the topology, the number of infected nodes, and the number of false negatives and positives, for our algorithm to correctly identify the infection cause (epidemic or random). Among our results, we show that our algorithm succeeds even when the number of infected nodes is \( O(N) \), i.e., a constant fraction of all the nodes, and also in the setting where the false positives grossly outnumber the true positives. We also consider the small-infection regime. We show here that our algorithm succeeds even if each node only knows the graph up to its 1-hop neighbors.

Experiments: we provide experiments on random graphs, as well as real-world graphs. We show that our algorithm’s performance is at least as good as the theory predicts. Moreover, its efficiency allows it to scale easily to very large networks. We compare to state-of-the-art and demonstrate that our algorithm is both faster and more accurate.

Acknowledgments

This research was supported in part by the European Union through the CONGAS project in the 7th Framework Programme, NSF Grants CNS-1017525, CNS-1320175, EECS-1056028, DTRA grant HDTRA 1-08-0029, ARO Grants W911NF-11-1-0265 and W911NF-14-1-0387, and the US DOT supported D-STOP Tier 1 UTC.

4. REFERENCES

[1] C. Milling, C. Caramanis, S. Mannor, and S. Shakkottai, “Network forensics,” in SIGMETRICS ’12, vol. 40, no. 1, Jun. 2012, p. 223.

[2] E. A. Meiron, C. Milling, C. Caramanis, S. Mannor, A. Orda, and S. Shakkottai, “Localized epidemic detection in networks with overwhelming noise.” [Online]. Available: http://arxiv.org/abs/1402.1263