Identification of Patient Zero in Static and Temporal Networks: Robustness and Limitations

Nino Antulov-Fantulin,[1] Alen Lančić,[2] Tomislav Šmuc,[1] Hrvoje Štefančić,[3,4] and Mile Šikić[5,6,*]
[1]Computational Biology and Bioinformatics Group, Division of Electronics, Rudjer Bošković Institute, Zagreb 10000, Croatia
[2]Department of Mathematics, Faculty of Science, University of Zagreb, Zagreb 10000, Croatia
[3]Theoretical Physics Division, Rudjer Bošković Institute, Zagreb 10000, Croatia
[4]Catholic University of Croatia, Zagreb 10000, Croatia
[5]Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb 10000, Croatia
[6]Bioinformatics Institute, A*STAR, Singapore 138671, Republic of Singapore
(Received 31 October 2014; revised manuscript received 6 February 2015; published 16 June 2015)

Detection of patient zero can give new insights to epidemiologists about the nature of first transmissions into a population. In this Letter, we study the statistical inference problem of detecting the source of epidemics from a snapshot of spreading on an arbitrary network structure. By using exact analytic calculations and Monte Carlo estimators, we demonstrate the detectability limits for the susceptible-infected-recovered model, which primarily depend on the spreading process characteristics. Finally, we demonstrate the applicability of the approach in a case of a simulated sexually transmitted infection spreading over an empirical temporal network of sexual interactions.

DOI: 10.1103/PhysRevLett.114.248701 PACS numbers: 89.75.Hc, 05.10.Ln, 87.19.X-

Introduction.—One of the most prevalent types of dynamic processes of public interest characteristic for the real-life complex networks are contagion processes [1–7]. Epidemiologists detect the epidemic source or the patient zero either by analyzing the temporal genetic evolution of virus strains [8–10], which can be time demanding, or trying to do a contact backtracking [11] from the available observed data. However, in cases where the information on the times of contact is unknown or incomplete or the infection is asymptomatic or subclinical the backtracking method is no longer adequate. Because of its practical aspects and theoretical importance, the epidemic source detection problem on contact networks has recently gained a lot of attention in the complex network science community. This has led to the development of many different source detection estimators for static networks, which vary in their assumptions on the network structure (locally treelike) or on the spreading process compartmental models (SI, SIR) [12–21], or both.

In the case of the SIR (susceptible-infected-recovered) model there are two different approaches. Zhu *et al.* proposed a sample path counting approach [15], where they proved that the source node minimizes the maximum distance (Jordan centrality) to the infected nodes on infinite trees. Lokhov *et al.* used a dynamic message-passing algorithm (DMP) for the SIR model to estimate the probability that a given node produces the observed snapshot. They use a mean-field-like approximation (node independence approximation) and an assumption of a treelike contact network to compute the source likelihoods [17]. Altarelli *et al.* remove the independence assumption and use the message passing method with an assumption of a treelike contact network to estimate the source [18]. In our study, we drop all the network structure and node independence assumptions and analyze the source probability estimators for general compartmental models. The main contributions of our Letter are the following: (i) We developed the analytic combinatoric, as well as the Monte Carlo methods (direct and Soft Margin) for determining exact and approximate source probability distribution, and have also produced the benchmark solutions on the 4-connected regular lattice structure. (ii) We measured the source detectability by using the normalized Shannon entropy of the estimated source probability distribution for each of the source detection problems, and have observed the existence of some highly detectable, as well as some highly undetectable regimes for the SIR and other spreading models. We notice that the detectability primarily depends on the spreading process characteristics. (iii) Using the simulations of the sexually transmitted infection (STI) on a realistic time interval of 200 days on an empirical temporal network of sexual contacts we demonstrate the robustness of the Soft Margin source estimator.

Methods.—In a general case, the contact network during an epidemic process can be temporal and weighted, but we first concentrate our analysis on a static undirected and nonweighted network $G = (V, E)$, where V denotes a set of nodes and E denotes a set of edges. The random binary vector $\vec{R}$ indicates which nodes got infected up to a certain time T. For the contagion model, we use the SIR model with the simultaneous updates in time described by the

probability p that an infected node infects a susceptible neighbor node in one discrete step and the probability q that an infected node recovers in one discrete step. We observe one epidemic realization $\vec{r}_*$ of $\vec{R}$ at a time T of the SIR process (p, q, T) on a network G and want to calculate the source posterior probabilities $P(\Theta = \theta_i|\vec{R} = \vec{r}_*)$. We have developed two complementary approaches that can provide exact posterior probability distributions over nodes in the spreading realization $\vec{r}_*$ via the Bayesian approach: the direct Monte Carlo approach and analytical combinatoric approach.

Using the *direct Monte Carlo approach*, for each potential source node i (infected node in the realization $\vec{r}_*$), a large number n of epidemic spreading simulations with maximum duration T is performed with i as an epidemic source. The number of simulations n_i which coincides with the realization $\vec{r}_*$ is recorded. To cut down on the extensive calculation required for the Monte Carlo simulations, we employ a pruning mechanism (no errors introduced), stopping the simulations at $t < T$ if the current simulation realization has infected a node which is not infected in $\vec{r}_*$. The probability of the node i being the source of the epidemic is then calculated as $P(\Theta = \theta_i|\vec{R} = \vec{r}_*) = n_i/\sum_j n_j$. The statistical significance of the direct Monte Carlo results are controlled with the convergence conditions. For more information, see the Supplemental Material [22], Sec. 2.

An alternative approach, the *analytical combinatoric approach* assigns to each node of degree n a generating function which is maximally $(n + 1)$ dimensional, which captures the events of node first infection and infection spreading through its edges at specific times. Then, by multiplication of the generating functions of all the infected nodes from a realization, we are able to merge all contributions together and get the source probability distribution. In the Supplemental Material [22], Sec. 1, along with the detailed description of the analytical combinatoric method, we demonstrate the correspondence between the direct Monte Carlo and analytical combinatorics. A serious disadvantage of the analytical method is that the calculations become prohibitively intricate in the case of nontreelike configurations.

We have generated a series of benchmark cases on a 4-connected lattice $(N = 30 \times 30)$, for which we have calculated the probability distributions over the potential source candidates using the direct Monte Carlo estimator (see Supplemental Material [22], Sec. 4). The source detectability $D(\vec{r}_*) = 1 - H(\vec{r}_*)$, is characterized via the normalized Shannon entropy H (normalization by entropy of uniform distribution) of the calculated probability distribution $P(\Theta = \theta_i|\vec{R} = \vec{r}_*)$.

Results depicting distributions of H for different parts of the SIR parameter space for the regular lattice are given in Fig. 1, plots (a), (b), and (c). The figures show qualitatively the same detectability behavior across the p parameter, for

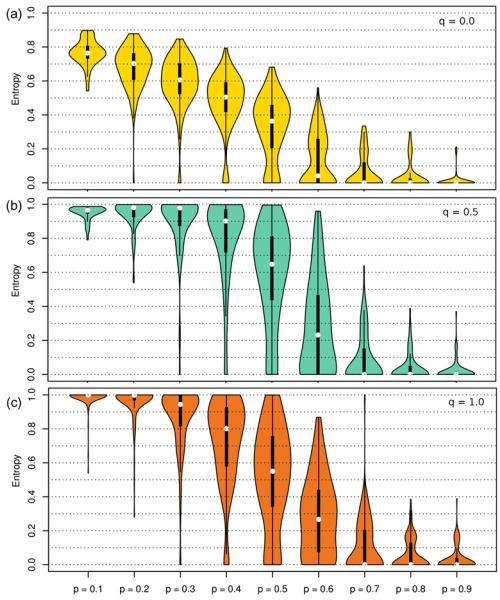

FIG. 1 (color online). Plots (a), (b), and (c): Box plots depicting distribution of entropy values (H) of source probability distributions for a number of randomly generated spreading realizations across with different (p, q) parameters on the 4-connected lattice: $N = 30 \times 30$ nodes with $T = 5$, calculated by the direct Monte Carlo method with 10^6–10^8 simulations per source.

different values of parameter q. It is important to observe the existence of three different regions: the low detectability–high entropy region $(p < 0.2)$, the intermediate detectability–intermediate entropy region $(0.2 < p < 0.7)$, and the high detectability–low entropy region $(p > 0.7)$. We observe that the detectability transition is still present even for different spreading models (SI, ISS, IC) and we observe the interplay of the network size and stopping time T on the detectability (see Supplemental Material [22], Sec. 10 and Fig. 2). In Fig. 2, plot (a), we observe that in a regime, when the network size restricts the epidemic spreading but not the epidemic itself via it's natural evolution characterized by the parameters (p, q) or stopping time T, the entropy is high as the realizations from different sources are almost identical.

As the application of direct Monte Carlo and analytical combinatoric approaches becomes prohibitively expensive for realistic network sizes, we formulate an estimator which is much more efficient in approximating the true underlying source probability distribution for the particular epidemic spread. We continue with the definition of the *Soft-Margin* estimator, a generalization of the Monte Carlo inference method, in which the direct Monte Carlo method represents a limiting case. In order to proceed we first need to introduce some useful definitions. The random binary vector $\vec{R}_\theta$ describes the outcome of the epidemic process and sample vectors: $\{\vec{r}_{\theta,1}, \ldots, \vec{r}_{\theta,n}\}$ describe n independent outcomes of that process. Each sample vector $\vec{r}_{\theta,i}$ is obtained using the Monte Carlo simulation of the contagion

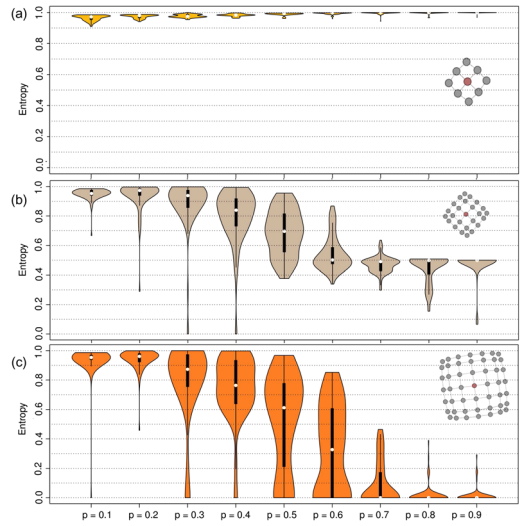

FIG. 2 (color online). Plots (a), (b), and (c): Box plots depicting distribution of entropy values (H) of source probability distributions for a number of realizations starting from the central node denoted with red color on the 4-connected lattice with different sizes (3×3, 5×5, and 7×7) with the SIR model for $q = 0.5$, $T = 5$, and different p values, calculated with the Soft Margin method with the (10^4–10^6) simulations per source and adaptive a chosen from the convergence condition.

process with the θ as the source. We measure the similarity between vectors $\vec{r}_1$ and $\vec{r}_2$ by the Jaccard similarity function $\varphi\colon (R^N \times R^N) \to [0, 1]$ calculated as the ratio of the size of the interaction of set of infected nodes in $\vec{r}_1$, $\vec{r}_2$ and the size of their union. The random variable $\varphi(\vec{r}_*, \vec{R}_\theta)$ measures the similarity between a fixed realization $\vec{r}_*$ and a random vector realization that comes from the SIR process with the source θ. The empirical cumulative distribution function of the n samples from the random variable $\varphi(\vec{r}_*, \vec{R}_\theta)$ is denoted $\hat{F}_\theta(x)$, where x is the value of the similarity variable. By taking the derivative of $\hat{F}_\theta(x)$, we get the probability density function (PDF) estimate:

$$\hat{f}_\theta(x) = \frac{d}{dx}\hat{F}_\theta(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - \varphi(\vec{r}_*, \vec{r}_{\theta,i})), \qquad (1)$$

where $\delta(x)$ denotes the Dirac delta distribution. Having defined the PDF for the observed similarities $\hat{f}_\theta(x)$, we can now define the main Soft-Margin inference expression as

$$\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta) = \int_0^1 w_a(x)\hat{f}_\theta(x)dx, \qquad (2)$$

where $w_a(x)$ is a weighting function. We use the following Gaussian weighting form: $w_a(x) = \exp[-(x-1)^2/a^2]$. In the limit where the parameter $a \to 0$, we obtain the direct Monte Carlo likelihood estimation. For cases when the parameter $a > 0$, we obtain an estimator which estimates the likelihood by using the weighting function $w_a(x)$ to accept contributions from realizations whose similarity to observed realization is less than 1. Using the property of delta distribution, we simplify the expression for the Soft Margin estimator to (for more details see Supplemental Material [22], Sec. 5):

$$\hat{P}(\vec{R} = \vec{r}_*|\Theta = \theta) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{-[\varphi(\vec{r}_*, \vec{r}_{\theta,i}) - 1]^2}{a^2}\right). \qquad (3)$$

Note, that alternative view on the Soft Margin estimator is the nonparametric density estimation with the Gaussian kernels [24]. Finally, we do not need to set the Soft Margin width parameter a in advance. After we calculate the estimated PDF for every potential source $\hat{F}_\theta(x)$, we can choose the parameter a as the infimum of the set of parameters for which the PDFs have converged. The implementation details, time complexity analysis, and pruning mechanism for the Soft Margin estimator can be found in the Supplemental Material [22], Secs. 5, 6, and 7.

Results.—We now demonstrate the applicability of our inference framework to detect the source of the simulated STI epidemic spreading in an empirical temporal network of sexual contacts in Brazil (see Fig. 3, plot (a)). This publicly available data set [25] was obtained from the Brazilian Internet community and is used as an

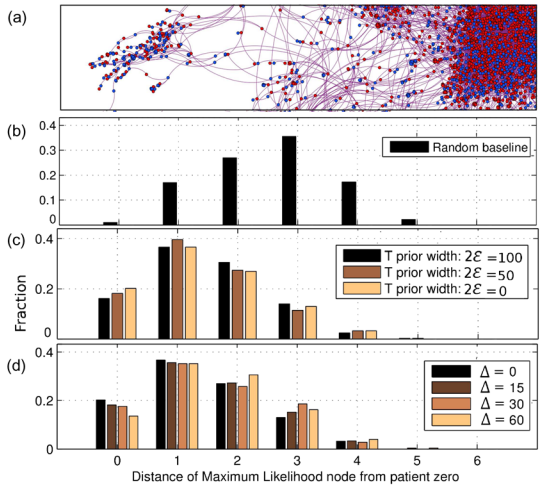

FIG. 3 (color online). Plot (a): Visualization of a part of the aggregated empirical temporal network of sexual contacts in Brazil [25]. In plots (b), (c), and (d) the performance is measured as the fraction of 500 experiments with a specific graph distance of the maximum likelihood candidate to the true source. The average execution time of a single experiment to calculate source probability distribution over all potential candidates was around 12 s (on 50 cpu cores) with 20 000 STI simulations per node. Plot (b): The baseline performance of a random estimator, which uniformly assigns likelihood to potential nodes. Plot (c): The influence of prior knowledge about initial outbreak moment $[t_0 - \epsilon, t_0 + \epsilon]$ of the outbreak on performance. Plot (d): The influence of randomized temporal ordering of interactions within Δ days, with $\epsilon = 0$ (we know the starting time t_0) on performance.

approximation of temporal sexual contacts. The data set (see Supplemental Material [22], Sec. 8) consists of the triplets (v_i, v_j, t), which represents the event that the nodes v_i and v_j had a sexual interaction at a time t. The first 1000 days in the original data set are discarded due to the transient period with sparse encounters [25] and therefore all temporal moments are measured relative to day 1000, as the authors have done in the original study [25]. For our temporal network, we use the SIR model $(p = 0.3, q = 0.01)$ for STI. The upper limit of the transmission probability for the STI that was previously used on this contact network is $p = 0.3$ [25]. The recovery parameter $q = 0.01$ represents a disease with the mean recovery of 100 days.

Note that here the calculation of exact source probability distributions is computationally too demanding for both the direct Monte Carlo and the analytical combinatoric method. Therefore, we use the Soft Margin estimator with the smallest width a for which the ML node probability estimate converged. Our experiments consist of two parts: (i) simulation of STI spreading through a temporal network of sexual contacts and (ii) detection of the patient zero from the observed process.

In order to demonstrate applicability of the approach in realistic conditions, we introduce uncertainty in the epidemic starting time t_0, and later on also with respect to node states in observed epidemic realization. Note that uncertainties in (p, q) parameters can also be relaxed by the marginalization procedure (see Supplemental Material [22], Sec. 5). The relaxation of knowing the starting point of the epidemic t_0 is done by using the marginalization over time, sampling over all possible starting points t_0 from a uniform probability distribution over $[t_0 - \epsilon, t_0 + \epsilon]$, $2\epsilon = \{0, 50, 100\}$ days. In Fig. 3, plot (c), we show the summary results from 500 independent experiments, when the starting time t_0 was chosen from the interval of [100–200] days, the end of the epidemic was set to the day $t = 300$ and using different uniform priors (ϵ) for the moment t_0. Using the uniform uncertainty of $\epsilon = 50$ days, we can still detect the source within its first neighborhood (distance 0 and 1 from the source) in approximately 60% of the experiments. These results are of great practical importance, since in reality we do not know the exact starting times, but rather only an upper and a lower bound on the starting point.

Next, we demonstrate how the uncertainty in the temporal orderings of interactions within a time window of the length Δ affects the performance of source detection. We use a randomization algorithm which permutes time stamps inside of a bin of Δ days from the start to the end of the contact interaction network in a nonoverlapping way. From Fig. 3, plot (d), we observe that higher uncertainty in orderings (higher Δ) reduces the detectability of the source of infection. However, the estimation framework is robust to small-scale interaction noise.

We have also shown that our Soft Margin algorithm estimates source probabilities with much higher precision than other estimators (Jordan and DMP estimator) on benchmark cases by comparing the results against the direct Monte Carlo source probability estimations on the regular lattice (see Supplemental Material [22], Sec. 3). Results for source detection for different values of (p, q) parameters and for the case when only a random subset of the node states is observed can be found in the Supplemental Material [22], Sec. 9.

Discussion.—The assumption about missing dynamic information about times of infection or recovery in our case study seems rather plausible for two realistic cases: STI infections and computer viruses. Many STIs generate silent epidemics since many of them are unrecognized, asymptomatic, or subclinical as the pathogens are being transmitted from patients with mild or totally absent symptoms. A large number of people with STIs: chlamydia [26], gonorrhea [26], human papillomavirus, and others show mild or no symptoms at all. The second motivation comes from silent spreading of a certain class of computer viruses and worms through computer networks which become active simultaneously on a specific date. Unlike other approaches [12–21], we identified different source detectability regimes and our methodology is applicable to arbitrary network structures, and is limited solely by the ability to computationally produce realizations of the particular contagion process.

The authors would like to thank Professor Dirk Brockmann for valuable discussions about the epidemic source detection problems during internship of NAF at the Robert Koch Institute in Berlin, Germany. For proofreading the manuscript, we would like to thank Vinko Zlatić, Sebastian Krause, and Ana Bulović. The work is financed in part by the Croatian Science Foundation under Project No. I-1701-2014, the EU-FET project MULTIPLEX under Grant No. 317532, and the FP7-REGPOT-2012-2013-1 InnoMol project under Grant No. 316289.

*Corresponding author.
mile.sikic@fer.hr

[1] A. Vespignani, Nat. Phys. **8**, 32 (2012).
[2] W. Broeck, C. Gioannini, B. Goncalves, M. Quaggiotto, V. Colizza, and A. Vespignani, BMC Infect. Dis. **11**, 37 (2011).
[3] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. GonzÃ¡lez, and V. Colizza, PLoS Comput. Biol. **10**, e1003716 (2014).
[4] V. Colizza, R. P. Satorras, and A. Vespignani, Nat. Phys. **3**, 276 (2007).
[5] C. Castellano and R. Pastor-Satorras, Phys. Rev. Lett. **105**, 218701 (2010).
[6] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[7] N. Antulov-Fantulin, A. Lancic, H. Stefancic, and M. Sikic, Information Sciences **239**, 226 (2013).

[8] M. Worobey, G.-Z. Han, and A. Rambaut, Proc. Natl. Acad. Sci. U.S.A. **111**, 8107 (2014).
[9] X. Du, L. Dong, Y. Lan, Y. Peng, A. Wu, Y. Zhang, W. Huang, D. Wang, M. Wang, Y. Guo *et al.*, Nat. Commun. **3**, 709 (2012).
[10] E. M. Volz and S. D. W. Frost, PLoS Comput. Biol. **9**, e1003397 (2013).
[11] D. M. Auerbach, W. W. Darrow, H. W. Jaffe, and J. W. Curran, Am. J. Med. **76**, 487 (1984).
[12] D. Shah and T. Zaman, in *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (ACM, New York, 2010), pp. 203–214.
[13] W. Dong, W. Zhang, and C. W. Tan, in *Proceedings of the IEEE International Symposium on Information Theory 2013* (IEEE, New York, 2013).
[14] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, Perform. Eval. Rev. **42**, 1 (2014).
[15] K. Zhu and L. Ying, in *IEEE/ACM Transactions on Networking* (2014), Vol. PP, p. 1.
[16] P. C. Pinto, P. Thiran, and M. Vetterli, Phys. Rev. Lett. **109**, 068702 (2012).
[17] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, Phys. Rev. E **90**, 012801 (2014).
[18] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, Phys. Rev. Lett. **112**, 118701 (2014).
[19] N. Antulov-Fantulin, A. Lancic, H. Stefancic, M. Sikic, and T. Smuc, in *Proceedings of the 2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems Workshops, 2014*, http://dx.doi.org/10.1109/SASOW.2014.35.
[20] C. H. Comin and L. da Fontoura Costa, Phys. Rev. E **84**, 056105 (2011).
[21] D. Brockmann and D. Helbing, Science **342**, 1337 (2013).
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.248701, which includes Refs. [23], for additional information about the used data set.
[23] L. E. C. Rocha, F. Liljeros, and P. Holme, Proc. Natl. Acad. Sci. U.S.A. **107**, 5706 (2010).
[24] J. S. Marron and D. Nolan, Stat. Probab. Lett. **7**, 195 (1988).
[25] L. E. C. Rocha, F. Liljeros, and P. Holme, PLoS Comput. Biol. **7**, e1001109 (2011).
[26] E. L. Korenromp, M. K. Sudaryo, S. J. de Vlas, R. H. Gray, N. K. Sewankambo, D. Serwadda, M. J. Wawer, and J. D. F. Habbema, Int. J. STD AIDS **13**, 91 (2002).