Source Detection for Large-Scale Epidemics

Brunella M. Spinelli

LCA3, I&C, EPFL

***Abstract*—Epidemic modeling is a well-studied problem. We review the two main approaches: random mixing models and contact networks. We explain why a realistic model for pandemics should encompass both.**

A far less studied problem is that of detecting the source of an epidemic with only partial observations of the diffusion process. We present the first contribution to the field [3] and describe how we plan to improve the existing models by employing data collected by sparse observers and by establishing an estimator for the position of the source of a large-scale epidemic.

***Index Terms*—Source detection, epidemics, sensor placement**

I. Introduction

EPIDEMICS have been the object of many modeling efforts for a long time and many interesting results have been reached in the field. The two classical approaches, reviewed in detail in [1], are random mixing models and network-based models. While in the first the population is modeled as a group of individuals in which everyone has the same probability of being infected by any of the other individuals, in the second the diffusion happens through a network of contacts. Both approaches come with their advantages and shortcomings. Random mixing processes are described through well-known differential equations but the key underlying assumption is not realistic for human everyday life, even within a city. On the other hand, network-based models are more lifelike but require contacts among people to be known or, at least, reliably estimated. Furthermore this approach is sometimes too restrictive, not accounting for the possibility for individuals to meet someone who is not among their usual contacts. Some efforts have already been done to build a model which binds together these two approaches, obtaining what we could refer to as a *network of random mixing communities*: random mixing is assumed only at a microscopic level, while on a larger scale there are only some predetermined connections. With such a model, interesting results in the study of world-scale epidemics have been reached [2].

Once a meaningful epidemic model is established, the inverse problem arises: is it possible to localize the source of a diffusion process? This is a far less studied question, first tackled in 2011 by Shah and Zaman's work about rumors in a network [3]. In the era of social networks, estimating the source of a rumor spreading on a large network is an intriguing challenge which, since this first paper came out, has received a significant amount of attention. Beyond the practical application, the authors also develop seminal theoretical results. After defining a maximum-likelihood (ML) estimator for the source of diffusion on a tree, they state and prove several results about the correctness of their estimator for different tree topologies, showing that the ability of tracing back the source from the configuration of infected nodes at a given moment depends heavily on the structural properties of the network on which the estimation is performed. To the best of our knowledge, none of the subsequent works in source localization thoroughly investigated this link between network topology and probability of success of the different estimators proposed. In contrast, we believe that such analyses, even if not directly applicable to real topologies, would lead to new and inspiring insights on the problem.

Moreover, we think that [3], as well as many other works on source detection, suffers from the unfeasibility of collecting the large amounts of data needed to perform the estimation. To overcome this, the authors of [5] proposed a source estimator based on sparse observations. Our current work is oriented towards developing the latter model, going further in the exploration of the use of sparse observers. More specifically, we would like to answer questions such as *which are the best nodes to monitor for data collection* and *which kind of information is more relevant to the disambiguation of the source*.

In this proposal we give a critical review of [1], [2] and [3] and propose a way of advancing the state of the art for the problem of source localization according to a twofold plan: on

Proposal submitted to committee: June 27th, 2013; Candidacy exam date: July 4th, 2013; Candidacy exam committee: Matthias Grossglauser, Patrick Thiran, Martin Vetterli.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(name and signature)

Thesis director: _____
(name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(B. Falsafi) (signature)

one side, continuing the present work on how to exploit sparse observations for epidemics containment and source detection, and on the other side generalizing source detection to a two-scale model for large epidemics.

II. Survey of the Selected Papers

A. Modeling Epidemics

The biology and mathematics communities have long been interested in epidemics and many models for the spreading of diseases have been proposed. Early models were mainly based on a random mixing assumption: any two individuals in the population can meet and possibly infect each other.

The review by Keeling and Eames [1] illustrates how networks came into play and started being used to describe the contagion process in terms of a more realistic *mixing network*, where individuals have only a finite set of contacts that they can possibly meet and infect or be infected by. Let us start by having a glance at how random mixing models for epidemics look like. A *random mixing model* is defined by a set of differential equations that describe how the partition of the population in different status classes (e.g. infectious, susceptible,...) changes over time. Many models can be derived as variations of the well-known SIR (Susceptible-Infectious-Recovered) and SIS (Susceptible-Infectious-Susceptible) models which are defined as follows.

SIR model:

$$\begin{cases} \frac{dS}{dt} &= bN - \lambda S - dS, \\ \frac{dI}{dt} &= \lambda S - gI - dI, \\ \frac{dR}{dt} &= gI - dR, \end{cases} \tag{1}$$

SIS model:

$$\begin{cases} \frac{dS}{dt} &= gI - \lambda S, \\ \frac{dI}{dt} &= \lambda S - gI, \end{cases} \tag{2}$$

Here S, I, R are the number of susceptible, infectious and recovered individual in a population of size N, while b, λ, g, d are, respectively, birth, contagion, recovery and death rates. These models have been surprisingly successful, mainly because they provide an efficient description of microscopic dynamics and they are adaptable to specific diseases and population features. As mentioned before, the main inconvenient they have is the unrealistic assumption that each single individual gets in contact with a random sample of the whole population.

In contrast, a network structure can capture the permanent contacts of each individual, whose number is usually notably smaller than the population size, and the precise structure of these contacts.

Here a question arises: how to ascertain a population contact network or simulate it in a reliable way? In [1] it is argued that the quest for representative networks comes across many problems such as quantifying the intensity of contact needed for an infection to spread, taking into account disease-specific features, collecting data from many different individuals and linking them, and handling incomplete data. On the other hand, simulating networks in order to incorporate and complete available data has the shortcoming of producing models that do not necessarily accurately capture the precise characteristics of the real network.

An interesting contribution to the understanding of the spread of epidemics comes from the study of *idealized networks*. These networks are designed to exhibit some specific features. They are defined in terms of the spatial positioning of individuals or in terms of the distribution or the number of individuals' contacts. Some examples are random networks (e.g. the Erdős-Rényi model), lattices, small-world networks, spatial networks and scale-free networks (e.g. the Barabási-Albert model). A very interesting fact about these models, with respect to epidemic modeling, is that, when it comes to disease diffusion, they display very diverse properties. The isotropic structure of lattices, for example, leads to a wave-like spread of infection (i.e. the diffusion happens in a roughly circular manner). A small-world network (e.g. the Watts-Strogatz model), is likely to display independent epidemics in distant parts of the network; a scale-free network, whose node degree distribution is heavy-tailed, may witness some super-spreading events. These are just a few examples that highlight that a key point in epidemics modeling is understanding how the relevant network looks like, Section II-B goes deeper in this direction.

The review by Keeling and Eames also discusses alternative models and possible directions for future research in the field. Two ideas interest us in particular:

- *Dynamic networks*. A static network is often quite far from reality. Contacts among people change over time and, more interestingly, the outbreak of an epidemic can have a considerable effect on the propensity to travel or to hang out in public places. [1] does not propose an answer to this challenge. We think that a first interesting step would be to employ a two-phase model where the structure of the mixing network changes after the percentage of infectious population reaches a given threshold.
- *Integration of a random mixing component in a network*. A purely contact-network-based approach can sometimes yield results that are too restrictive. Intuitively, individuals are likely to have some random encounters in an unpredictable way. In [1] it is mentioned that a model where any two individuals can potentially interact, but predominant partnerships are imposed, has already been studied for the case of STDs (sexually transmitted diseases). A somehow similar idea is that of considering short-distance random mixing combined with some long distance contacts which are activated with a certain probability. In the next subsection we will see how the work of Colizza, Barrat, Barthélemy, and Vespignani [2] showed that this is indeed a very effective model for describing large-scale epidemics.

B. A Model for World Epidemics

The article published in 2006 by Colizza et al. [2] is a seminal work in modeling world epidemics. Their contribution takes advantage of the publicly-available world airport network (WAN) data set to build a model that is the superposition of city-scale random mixing SIR models and network-based city-to-city movements of individuals.

The network considered consists of a set V of $M = 3100$ vertices representing main city airports and 18,810 weighted edges representing passenger flows between each airport pair. Denote by N_j, $j \in \{1, .., M\}$, the population size of the metropolitan area served by airport j, then for each j it holds that

$$N_j = S_j(t) + I_j(t) + R_j(t),$$

where S_j, I_j and R_j are the number of susceptible, infectious and recovered individuals at time t. More generally, we could write

$$N_j = \sum_m X^{[m]}(t),$$

letting m indicate one of the possible states.

Within each city, the epidemic evolution is governed by a stochastic version of the SIR equations which associates a noise term to each transition process[1]. Moreover, equations are coupled among them by a *transport operator* Ω encoding the population movements from one city to another, and are solved numerically. Let us now look more closely at the definition of Ω. The basic idea is to make use of the WAN data set to model the number $\xi_{jl}(X_j^{[m]})$ of people in a certain state traveling from city j to city l during each time interval. Denote the average passenger flux per day on connection (j, l) as w_{jl}. Then, for each city j, the variables $\{\xi_{jl}(X_j^{[m]})\}_l$ follow a multinomial distribution with parameters $\{w_{jl}\Delta t/X_j^{[m]}\}$, Δt being the discrete time interval considered. Hence the transport operator can be written as

$$\Omega(X_j^{[m]}) = \sum_l (\xi_{lj}(X_l^{[m]}) - \xi_{jl}(X_j^{[m]})).$$

Besides defining this original model, the authors tackle the challenging tasks of demonstrating the importance of the network topology and of assessing forecast reliability of the model. As for the first task, they compare the *entropy* (see Def.1 below) of epidemics simulated on the actual network with the entropy observed for epidemics simulated on two other models that we will describe shortly.

Definition 1. *Given a fixed time t, the prevalence of the epidemic in city j is* $i_j(t) \doteq I_j(t)/N_j$. *Let*

$$\vec{\rho}(t) = [\rho_j(t)] \doteq \Big[\frac{i_j(t)}{\sum_l i_l(t)}\Big],$$

then the entropy of the epidemic at time t is the entropy of the vector $\vec{\rho}$*:*

$$H(t) = -\frac{1}{\log(M)} \sum_j \rho_j(t) \log \rho_j(t).$$

Note that H measures the homogeneity of the infection in the network: $H = 1$ if the epidemic is homogeneously affecting all the cities, $H = 0$ if only one city is infected.

In both models used for comparison, passenger fluxes and population sizes are assumed to be uniform and equal to the averages of the real data. Moreover, in one case (A) the network is identical to that of the proposed model, in the other (B) they use a synthetic Erdős-Rényi network with the same number of vertices and where each possible edge is drawn with probability $\langle k\rangle/|V|$, $\langle k\rangle$ being the average node degree in the real network. Then, the two models have the same average node degree and the same weights on the edges but different edge sets. In fact, here the goal is to show that the WAN data set leads to a good choice for the connections to include in the model.

The simulations show that the entropy values obtained with the proposed model are similar to the values obtained under the comparison model A but quite different from those obtained under B. This leads the authors to conclude that there is striking evidence of a direct relation between the epidemic pattern and the network topology, a finding that confirms that, as it was also claimed by [1], an essential task in building an epidemic model is understanding how the network which is relevant for the disease spread looks like.

To assess the reliability of the forecast, the measure used is a product of Hellinger affinities.

Definition 2. *Let* $\vec{p}, \vec{q} \in \mathbb{R}^N$ *be two points in the N-simplex. Their Hellinger affinity is given by*

$$\mathrm{sim}(\vec{p}, \vec{q}) = \sum_{i=1}^{N} \sqrt{p_i q_i}.$$

Note that the definition encodes a notion of similarity between two probability vectors: $\mathrm{sim}(\vec{p}, \vec{q}) = 1$ if $\vec{p} = \vec{q}$, while $\mathrm{sim}(\vec{p}, \vec{q}) = 0$ if they are non-zero on sets that do not intersect each other. Now call $\vec{\pi}(t)$ the vector s.t. $\pi_j(t) = \frac{I_j(t)}{\sum_l I_l(t)}$, $j \in \{1, .., n\}$. The authors measure the similarity of two epidemics I and II through the values over time of

$$\Theta(t) = \mathrm{sim}(\vec{\pi}^I(t), \vec{\pi}^{II}(t)) \times \mathrm{sim}(\vec{i}^I(t), \vec{i}^{II}(t))$$

where $\vec{i}(t) = [i_j(t)]_j$ is the infection prevalence vector defined above. Simulations are run under the actual model and under the two comparison models A and B. Results show that the overlap between different realizations of the epidemic process is fairly high for model B. On the other hand the predictability is much smaller for model A. This is due to the fact that in B the connectivity pattern is quite homogeneous, while in A the heterogeneity of the network leads to more diversified pathways for the epidemic evolution. In the actual model the effect of the heterogeneity of the network is mediated by the non-uniform weights, resulting in a higher predictability. However the proposed model is less predictable than model B, especially at early stages of the epidemics. Moreover the authors remark that the predictability increases when the epidemic starts in a city with few connections, while it is very low when the first infected city is a large hub in the WAN data set, due to the fact that the epidemic can spread through many different pathways. These results about predictability are very interesting because they go in the direction of answering the question whether the epidemic evolution contains, even after a long time interval, some information about the city or the area where it all started. This is, in a few words, the idea on

[1]For example, the number of individuals who get infected in city j during the time interval $[t, t + \Delta t]$ is given by $\lambda N_j^{-1} X_j^{[S]}(t) X_j^{[I]}(t)\Delta t + \eta\sqrt{\lambda N_j^{-1} X_j^{[S]}(t) X_j^{[I]}(t)\Delta t}$, where $\eta \sim N(0, 1)$.

which the source localization problem that we will introduce in the next subsection is founded.

Finally, it is worth mentioning that, in 2007, the same authors published the results obtained testing the model on the data collected during the SARS pandemic in 2002-2003 [4]. The model is used to predict for which countries the risk of being infected is higher than a fixed threshold. It is shown that the predictions match with the empirical data with a maximal error of 7% during the whole evolution of the epidemic. Quite surprisingly, larger accuracy is achieved in European countries or in the US that are far away from the estimated epidemic center (Hong Kong, China). The precision is much lower for nearby countries such as Singapore, Taiwan or Vietnam. On one hand this shows that the use of the WAN dataset is an appropriate instrument to study world epidemics; on the other, it indicates that the (relatively) short-distance spread cannot be captured in the same way and needs a more detailed analysis. However, since all these observations are based on a single realization of the process, it is difficult to determine whether the imperfections in the match are due to outlying features of the realization or to structural deficiencies in the model. In this specific case, the large amount of connections within South-East Asia is likely to lead to a considerable variability in the epidemic evolution and, from the point of view of containment policies, a very interesting extension would be to investigate the reasons for this variability: the social diversity of traveling individuals, the travel frequency and the scope of visiting a foreign country could be some starting points.

C. Where Did the Epidemic Start?

Now we come to the core of our discussion and introduce the problem of detecting the diffusion source of an epidemic. In the most general possible terms we can formulate the problem in this way: given partial observations of a diffusion phenomenon on a network, and given a particular model for this diffusion, can we infer where the process started? At least two expressions ask for further precision: *partially observed* and *infer*. How much of the process we can assume to be able to observe? How can we build an estimator for the position of the diffusion source and assess its reliability?

We go into the problem of source detection by presenting and discussing the article by Shah and Zaman that is considered to have opened the field in 2011 [3]. The model used in this work is a simplified version of the SIS model that we could denote by SI: once a susceptible node gets infected it stays infected forever. Infection flows on edges according to independent exponentially distributed random times. The authors propose to estimate the source in a ML fashion, based on the knowledge of the network status at time t. Let N be the number of nodes which are infected at time t and call G_N the connected component of the graph formed by all the infected nodes. Then the ML estimator is given by

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(G_N|v),$$

where $\mathbf{P}(G_N|v)$ is the probability of observing the configuration G_N of infected nodes under the hypothesis of v being the source. Denote by $\Omega(v, G_N)$ the set of all possible[2] permutations starting with v and resulting in the infected component G_N: the key quantity considered in the paper is $R(v, G_N) = |\Omega(v, G_N)|$, called *rumor centrality*. The name is based on the fact that the article addresses in particular the diffusion of rumors in social networks. What is interesting about this quantity is that it allows a concise expression for the ML estimator in the case of *regular trees*[3]: for such trees, one can check that every permutation σ of given length has the same probability and the ML estimator can be expressed as

$$\begin{aligned} \hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(G_N|v) &= \arg\max_{v \in G_N} \sum_{\sigma \in \Omega(G_N, v)} \mathbf{P}(\sigma) \\ &= \arg\max_{v \in G_N} R(G_N, v). \end{aligned} \tag{3}$$

The estimator is then generalized to the case of non-regular trees and general graphs via the BFS (Breadth-First-Search) approximation, i.e. assuming that the diffusion happens only through a BFS tree[4]. Of course, the resulting estimator is no longer the ML estimator. Moreover, $\mathbf{P}(\sigma)$ is no more constant and appears in the expression:

$$\hat{v}_{\text{BFS}} \in \arg\max_{v \in G_N} \mathbf{P}(\sigma_v^*) R(G_N, v),$$

where σ_v^* is a BFS permutation of $\Omega(G_N, v)$.

Here, the whole idea underlying source estimation is fairly simple: the node that is most likely to be the source is the one which has the largest (weighted) number of permutations that lead to the observed configuration G_N. Finally, it is worth mentioning that the quantity $R(G_N, v)$ is computationally very cheap: it is proved that it can be computed in $O(N)$ time.

However, the most substantial parts of the work from Shah and Zaman are the results on the estimator's performance for different tree topologies. Let $\mathcal{C}_t$ be the event of detecting correctly the source after time t. Looking at regular trees of degree d, it can be shown that:

- if $d = 2$, i.e., the graph is a line, $\mathbf{P}(\mathcal{C}_t)$ goes to 0 as $t \to \infty$. In particular,
$$\mathbf{P}(\mathcal{C}_t) = O\Big(\frac{1}{\sqrt{t}}\Big)$$
.
- if $d > 2$, there exists a constant α_d such that
$$0 < \alpha_d \leq \liminf_{t \to \infty} \mathbf{P}(\mathcal{C}_t) \leq \limsup_{t \to \infty} \mathbf{P}(\mathcal{C}_t) \leq \frac{1}{2}.$$
- if $d = 3$, $\lim_{t \to \infty} \mathbf{P}(\mathcal{C}_t) = \frac{1}{4}$.

Moreover, they manage to identify a class of trees for which the detection probability goes to 0 for $t \to \infty$: this is the class of *geometric trees*.

Definition 3. *Let $\alpha > 0$ and $0 < b \leq c$. A tree is called geometric if for every fixed node v^* of degree d the following*

[2]Possible with respect to the diffusion model: each newly infected node has at least one of the previously infected nodes in his neighborhood.

[3]A *regular tree* of degree d is a tree in which each node has degree d.

[4]A BFS tree is formed through the following process: start with a root node and repeatedly add all the neighbors of the already picked nodes that have not been picked yet together with the edges connecting them to previously reached nodes.

happens. Consider the sub-trees rooted at v^*, say $T_1, .., T_d$. For any node $v \in T_i, r \in \mathbb{N}$, let $n_i(v, r)$ be the number of nodes in T_i at distance exactly r from the node v. Then we require that, for all $1 < i < d$ and $v \in T_i$,

$$b \cdot r^\alpha < n_i(v, r) < c \cdot r^\alpha.$$

The condition states that each sub-tree rooted at a given node should satisfy polynomial growth with exponent α. Note that for regular trees the growth is exponential while here it is sub-exponential. The theorem proved is the following.

Theorem 1. *Let T a geometric tree and $v \in T$ with degree $d \geq 3$ such that $d > \frac{c}{b} + 1$. If the rumor starts spreading from v then $\lim_t \mathbf{P}(\mathcal{C}_t) = 1$.*

The intuition behind these results is that very regular topologies do not allow to deduce the source from the infection pattern. The extreme but very intuitive case is that of the regular tree with $d = 2$: given a sequence of adjacent infected nodes the probability of detecting the true source decreases dramatically with the length of the sequence. For $d > 2$ the exponential symmetric expansion of regular trees leads to high variance in the evolution and hence to low detection rates.

In view of these interesting results on the correctness of the estimator, the question arises whether it is possible to derive similar results for more general diffusion models and related estimators. More specifically, we think that the model of [3] is quite restrictive both for the kind of diffusion in itself and for the amount of information needed for source estimation: for many real situations, it is not realistic to model the epidemic in a SI fashion and it is often not feasible to know the status of the entire network.

Some significant steps towards a more applicable, yet mathematically tractable model are made in the work by Pinto, Thiran, and Vetterli [5]. In the next section we will describe the model and estimator of this latter work, whose analysis and development is at the core of our present research.

III. Current Work & Research Proposal

A. Localizing the Source Through Sparse Observations

Our current work is focused on the search for optimal observers placements for the model of [5], where an ML estimator that performs source detection through sparse observations is proposed. Let us describe which model for the diffusion is employed, and define the related estimator for the source.

Consider the graph representing the network where the information/infection diffuses. The process is started at an unknown time t^* by a source s^* which is randomly distributed among the nodes. Each node u that gets infected at time t_u will re-transmit the information to each non-infected neighbor v at time $t_u + X_{uv}$, where $\{X_{uv}\}_{u,v \in V}$ are random variables with a known, arbitrary joint distribution. Out of the n nodes of the graph, K are chosen to form the observers set O and record from which neighbor and at what time they receive the infection. Hence, the observation set is $\mathcal{O} = \{(o, v, t_o)\}$ where o is an observer, v is the neighbor who transmitted the infection and t_o is the time at which the infection was received.

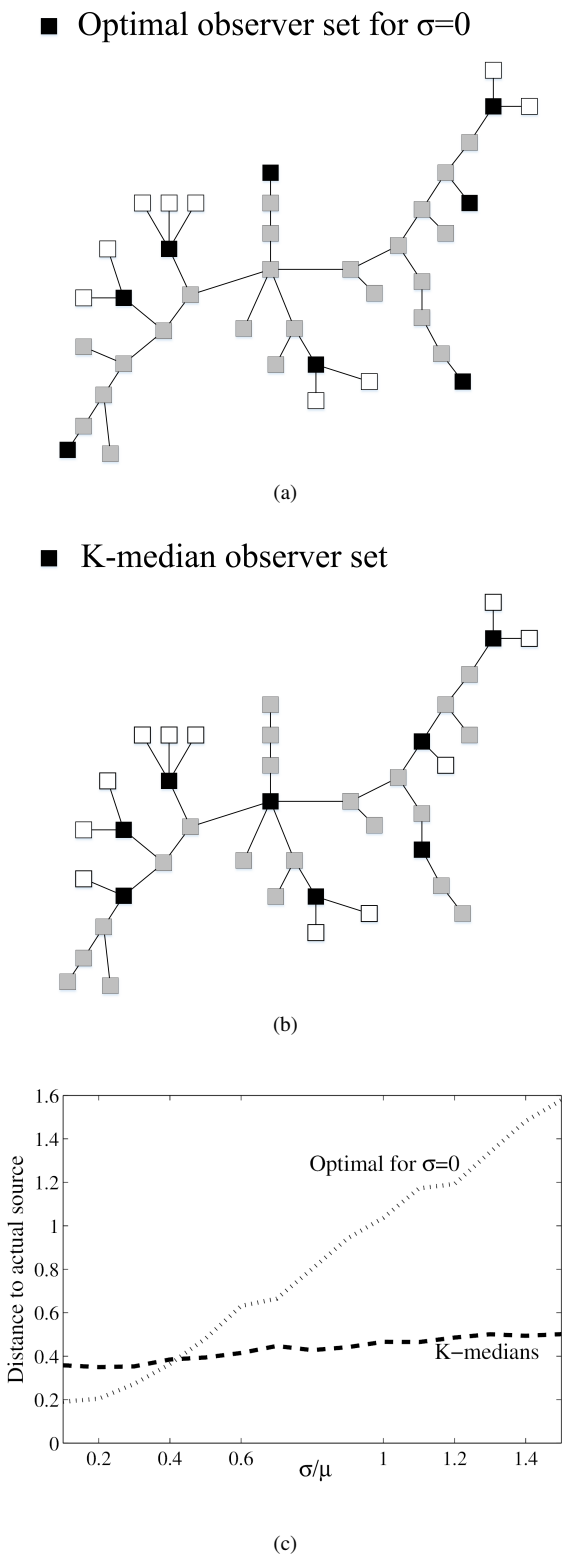

Fig. 1. Comparison between the optimal placement for $\sigma = 0$ (a) and the k-medians placement (b). Observers (black) divide the tree in sets of adjacent nodes (grey). If the real source is one of the white nodes it is immediately detected; if it is in one of the grey components, the estimator compare the likelihood for all the nodes in the component. Note that the large component of (a) is replaced in (b) by several smaller componenents. (c) shows the mean error (in hops) obtained from 3000 simulations.

The location of the source is recovered from the measurements $\mathcal{O}$ by the maximum likelihood criterion

$$\hat{s} = \arg\max_{s \in V} \mathbf{P}(s^* = s|\mathcal{O}).$$

If we consider the graph G to be a tree and the propagation delays associated with the edges of G to be $i.i.d.$ Gaussians with known mean μ and variance σ^2, the ML estimator can be written in a simple closed form (see [5] for details). These assumptions are not too restrictive because the generalization to general graphs can be obtained through the BFS heuristic (as in [3]). Furthermore, Gaussian distributed delays represent well many situations in which the transmission time can be modeled as the sum of a constant and a zero-mean Gaussian noise.

Clearly, the ML estimation proposed has an associated probability of error that depends on the number and the position of the observers in the network. We studied the probability of error of the estimator and found that the optimal observer placement depends on the signal-to-noise ratio (SNR) μ/σ. When $\sigma = 0$ (i.e. the transmission times are deterministic), placing observers in peripheral locations of the network allows to trace back the position of the source by taking differences between the observed infection times. For growing values of σ this strategy becomes less and less satisfactory as the measurements taken at extreme locations of the network tend to be very noisy. In this case we propose to model the observer placement as an instance of the *k-medians problem*[5]. This corresponds to the intuitive idea of placing observers in an uniform way, minimizing the average distance from a non-observer node to the closest observer. In this way we avoid an accumulation of noise along paths that connect observers and perform a faster and more precise source detection. In fact, we can see the k-medians as partitioning the network uniformly, hence enabling us to reduce the number of candidate source to a small set of nodes (see Figure 1).

B. Further Development & Research Plan

1) Dependence of the Optimal Observer Placement on the Network Topology: As explained in Section III-A, increasing values of the noise σ lead to a change in the optimal placement. In the next few months we will investigate this change thoroughly for different tree topologies. Our goal is to find out if there are some extreme topologies, i.e., some class of trees for which either one of the two placement is always optimal or there is a sudden phase transition at a particular value of the SNR. We think that this would be an enlightening first step towards the understanding of data collection for source localization on networks.

Another interesting question concerns optimal observer placement for general graphs. In fact, even if it seems that placing observers uniformly in the network according to the the k-medians criterion would still be an efficient choice, this may not be the case for networks that exhibit very small diameter. This is the case, for example, for many common families of idealized network such as Erdős-Rényi or Watts-Strogatz networks. In such situations, few observers would be enough to reach the k-medians optimum and we would not make use in the most efficient way of the available budget of observers. To this end, we think that a more refined notion of centrality should be employed in the observers' choice, e.g. taking into account the number of multiple paths connecting nodes in the network.

2) Generalizations of the Model Proposed in [5]: We think that the model of [5] is a good starting point for several generalizations and has the potential to meet many of the challenges that came to light in our discussion.

First of all, in a number of real situations it is not possible to assume that the observers reveal from whom they receive the information both for privacy issues or because the information is not available to the observer himself (e.g. in the case of airborne diseases). We already know that the hypothesis of observers knowing from whom they got infected can be removed with slight changes in the estimator. How does the probability of error increase if we suppose that this information is no longer available? We believe that the k-medians would still form an efficient observer set, at least for the tree case, but this needs to be studied further.

In the end of Section II-A we mentioned that a model accounting for time variability in the network would be desirable. This could be obtained by associating to each edge a probability of transmission and allowing these probabilities to vary at different stages of the epidemics. Would it be possible to obtain a closed-form expression for the ML estimator in this context? The case in which each edge e is traversed with a probability p_e constant over the time can be easily studied thanks to the independence between *whether* an edge is traversed and *when* it is traversed. Considering probabilities that change over time is more challenging. A promising approach could be that of estimating the source through two successive backward steps. Let us suppose that we have ascertained that at time $\tilde{t}$ the probabilities associated to the network edges change. We could start by focusing on what happened after $\tilde{t}$, identifying one or more nodes that are responsible for the last stage of the spreading. In the second phase of our estimation, we would then use the probabilistic topology relative to $t < \tilde{t}$ to estimate the node that is more likely to have caused the infection of the nodes selected in the first backward step. In order to develop this idea we would start off by studying diffusion models involving networks varying over time and their power to explain real data.

3) Localizing the Source of a Pandemic: Our ultimate goal is to perform source estimation on a two-scale model for disease diffusion similar to the one proposed in [2]. Recalling our observations at the conclusion of Section II-B, we think that here the main challenge is to find the appropriate data to collect and exploit in order to disambiguate between different sources: in view of the high variability of epidemic paths described by [2], the question about some common features of these paths, that could help in tracing back the source, becomes more intriguing. The substantial modification that we would like to propose with respect to the model of [2] is that of substituting the city mixing components with highly

[5]Given a budget k of nodes that can be selected, the problem is that of finding the set of nodes O^*, $|O^*| = k$, such that the average distance from a non-selected node to the nearest selected node is minimized.

connected graphs where the probability of connection between two individuals should ideally vary with the distance between them. This would give a more realistic model on which to perform source detection.

We think that the solution to this problem could come as a further step from the two-phase detection described in Section III-B2. In fact, the epidemic can still be seen as the outcome of two processes (local diffusion and long range spreading) even if they do not happen in subsequent times but simultaneously and hence interacting.

ACKNOWLEDGMENTS

The work presented in Section III-A was realized in collaboration with Pedro Pinto and Filip Pavetić.

The author would like to thank Lucas Maystre and Farid Movahedi Naini for helpful discussions and Holly Cogliati-Bauereis for useful suggestions concerning English writing.

REFERENCES

[1] M. J. Keeling, and K. T. D. Eames, *Networks and epidemic models*, J.R.Soc. Interface, Vol. **2**, September 2005.

[2] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, *The role of the airline transportation network in the prediction and predictability of global epidemics*, Proceedings of the National Academy of Sciences of the USA, Vol. **103**, No. 7, February 2006.

[3] D. Shah, and T. Zaman, *Rumors in a network: who's the culprit?*, IEEE Transactions on information theory, Vol. **57**, No. 8, August 2011.

[4] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, *Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study*, BMC Medicine, Vol. **5**, November 2007.

[5] P. Pinto, P. Thiran, and M. Vetterli, *Locating the Source of Diffusion in Large-Scale Networks*, Physical Review Letters, Vol. **109**, August 2012.

[6] R. Durrett, *Random graphs dynamics*, Cambridge University Press, 2007.