Supplemental Case Studies

Pedro C. Pinto,[1] Patrick Thiran,[1] Martin Vetterli[1]

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

We present two supplemental case studies using the framework introduced in the paper *Locating the Source of Diffusion in Large-Scale Networks*, by the same authors, published in Physical Review Letters, vol. 109, issue 6, no. 068702, pp. 1–5, Aug. 2012.

I. Localizing the Source of Contamination in a Subway

Networks are often vulnerable to accidental or intentional contamination. In 1998, Sydney's main water supply was contaminated by two microscopic parasites, and health warnings were issued to three million residents. Similar events occurred in water distribution systems worldwide, such as the E. Coli bacteria contamination in Walkerton, Canada in 2000, and the chemical solvent contamination in Crestwood, USA in 2009. In 1995, terrorists released sarin gas on several lines of the Tokyo metro, one of the world's busiest commuter transport systems, killing thirteen people and injuring thousands. In these scenarios, it is important to determine the location of the contamination source, since it can help infrastructure personnel stop or limit the contamination.

To evaluate the effectiveness of the SparseInf algorithm, we consider the scenario of airborne contamination in the New York City subway. Fig. 1(a) depicts the Red Line of the NYC subway (composed of Routes 1, 2, and 3). There are $N = 91$ nodes representing subway stops with associated GPS locations, obtained from [1]. Each stop is a potential access point for injection of an air contaminant into the subway tunnels. The propagation delay over edge i is modelled as $\mathcal{N}(\mu_i, \sigma_i^2)$, where both μ_i and σ_i are proportional to the physical length of the tunnel, so that the *propagation ratio* $\eta \triangleq \frac{\mu_i}{\sigma_i}$ is a constant which can be interpreted as a signal-to-noise ratio. The propagation delays are independent for different edges i.[1] To determine the source of contamination, we deploy K sensors uniformly at random over the network, which is appropriate in the absence of a-priori information about the source location.

Figure 1(b) shows the probability of correct localization $P_{\text{loc}} = \mathbf{P}(\hat{s} = s^*)$ versus the observer density $\frac{K}{N}$, for various values of the propagation ratio η. The probability P_{loc} is averaged over the random source location, the random observer locations, and the random propagation. Note

[1]More elaborate aerosol transport models can be used, accounting for factors such as turbulence, temperature, and contaminant properties.

that all the curves in Fig. 1(b) are lower-bounded by the *unit-slope* line corresponding to $P_{\text{loc}} = \frac{K}{N}$. This line represents the performance when all arrival information is ignored, so that each observer can only identify the source if they are in the same location. As expected, the localization probability P_{loc} increases with the η, reaching its maximum value when propagation is deterministic ($\eta = \infty$). Even in this limit, we have that $P_{\text{loc}} < 1$ in general, since there is a nonzero probability that the source s^* falls inside an unresolvable set, in which case we can only guess the location of s^* within the appropriate subset.

Figure 1(c) plots the average distance between $\hat{s}$ and s^*, as a function of the observer density $\frac{K}{N}$. We conclude that even with a *sparse* arrangement of observers with 20% density, we can achieve an average error of *less than one hop* for a relatively unfavorable signal-to-noise ratio of $\eta = 1$. This *small distance error* also implies a proportionately *small delay* between the contaminant release and the first detection.

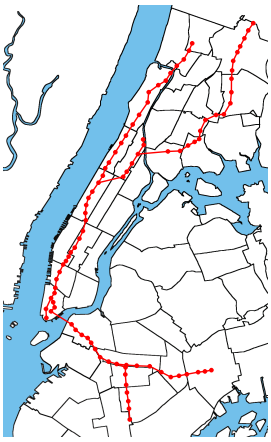
Overall, the results in this section suggest that for the purpose of localizing the source of contamination, a sparse deployment of sensors may provide an effective alternative to the individual monitoring (either human or automatic) of every station in the subway network.

II. Localizing the Leader of a Terrorist Organization

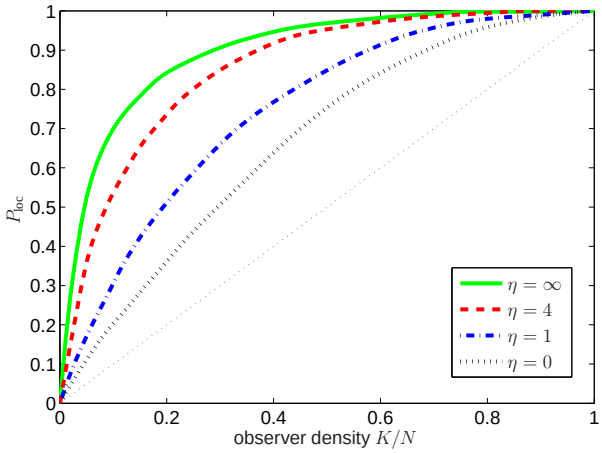
An organizational chart of a company rarely captures how information and influence circulate among its employees. Although the hierarchical chart is usually a tree with the leader at the root, the day-to-day operations of the organization are often conducted through a complex network of informal interactions. Furthermore, the actual sources of influence are not clearly identified in many cases. The identification of thought leaders or opinion leaders is an important challenge in network science. In traditional approaches, a node with high centrality (e.g., large degree) is often associated with the leader. However, the two do not necessarily coincide—for example, it is possible that the thought leader is a weakly-connected node that diffuses his ideas to strongly-connected nodes, which further spread it throughout the network. We propose an alternative approach: rather than simply consider the network structure, we monitor the actual information that is circulated in the network at a few specific nodes (the observers), and use it to estimate the original source of the information.

In what follows, we use the SparseInf algorithm to identify the leader in the 9/11 terrorist network. Figure 2(a) depicts the network of $N = 62$ hijackers and associates who were allegedly involved in the September 11th 2001 attack. The data was collected in [2] from publicly released information in major newspapers. Shortly after 9/11, Mohamed Atta was identified as the ringleader of the attack. Our goal is to determine whether observing the information flow at a few nodes of the network would lead to Atta as the possible source, with high probability. In such scenario, the *information* that is propagated across the network corresponds to messages sent by Atta (e.g., intelligence, plans, or instructions), while the *observers* correspond to terrorists whose communication is wiretapped. Furthermore, we would like to optimally place the wiretaps (usually in limited number), in order to maximize the localization probability.

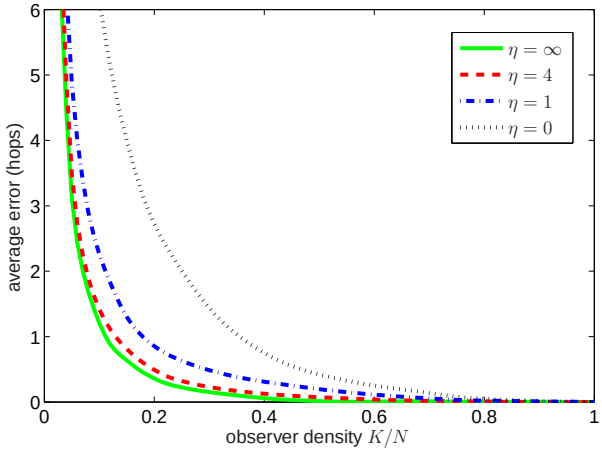
The successful localization of a terrorist leader could enable either prevention or prosecution, depending on whether the source localization can be performed in real-time or not. However, this task presents tremendous challenges. First, it is extremely difficult to obtain an accurate representation of a terrorist network, due to its covert nature. The best we can do is to estimate



(a) The Red Line of the New York City subway network, composed of Routes 1, 2, and 3. Each of the $N = 91$ subway stops is a potential access point for injection of an air contaminant into the subway tunnels. We randomly deploy K contamination sensors, which estimate the source of contamination based on the timing and direction of arrival of the contaminant.



(b) Probability of correct localization P_{loc} versus the observer density K/N, for various values of the propagation ratio $\eta \triangleq \mu_i/\sigma_i$.



(c) Average distance between $\hat{s}$ and s^* (in hops) versus the observer density K/N, for various values of the propagation ratio η.

Figure 1. Localizing the source of an airborne contamination in the New York City subway.

the network based on all available intelligence, being wary that incomplete data might greatly affect the results. Second, it is hard to accurately estimate the statistics of the propagation delays, even just their mean and variance. Since messages can be delivered through various mediums or infrastructure networks, the propagation statistics are not as easy to compute as in the example of Section I. Lacking better information, we assume that all propagation delays are IID $\mathcal{N}(\mu, \sigma^2)$, where the *propagation ratio* $\eta \triangleq \frac{\mu}{\sigma}$ is the only parameter that needs to be estimated. Lastly, it is extremely difficult to intercept the communication between terrorists. As a result, we focus on the placement of only two wiretaps on the network.

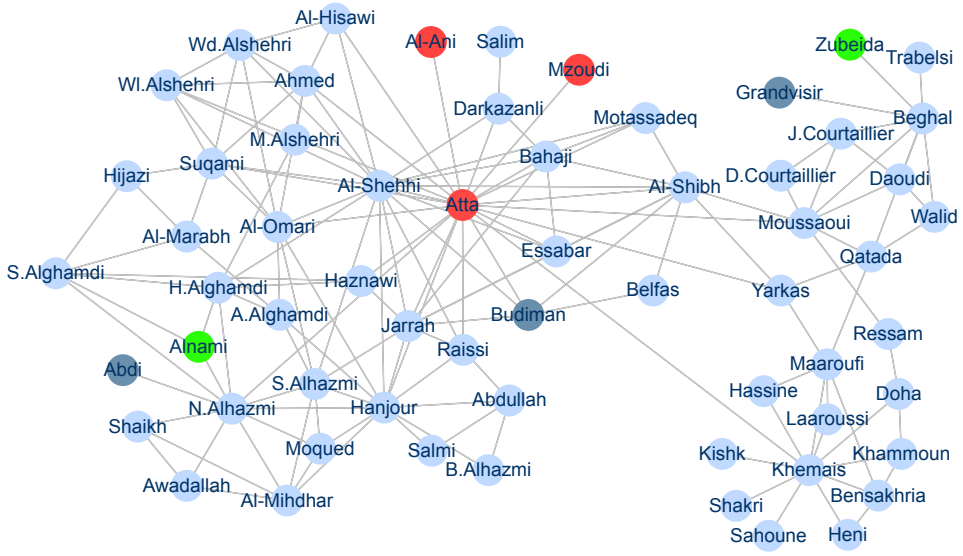
Figure 2(a) shows a particular outcome of the SPARSEINF algorithm, with $\eta = 4$ and $C = 20$ cascades. The two observers—Alnami and Zubeida, in green—gather all the incoming cascades, which lead them to identify three possible sources, marked in red: Atta, Al-Ani, and Mzoudi. Note that for the chosen observers, this is the absolute best it can be done. This is because Atta, Al-Ani, and Mzoudi form an *unresolvable set* of size 3: they all exhibit the same deterministic delay $\boldsymbol{\mu}_s$, and therefore are indistinguishable in terms of source estimation. The size of such unresolvable sets is shown in Table 2(b) for various choices of observers. Note that the placement of the observers has an effect the resolvability of the source. In general, observers that are well-separated from each other and surround the source (e.g., Alnani and Zubeida) ensure better resolvability than observers that are close to each other, and both far away from the source (e.g., Grandvisir and Zubeida). Of course, since the source location in unknown a priori, the challenge relies in carefully choosing the observer locations so that the accuracy is high for any possible source. Optimal observer placement is still an open problem and is currently being investigated.

The outcome of the SPARSEINF algorithm is itself random, due to the random propagation delays on the network. Figure 2(c) shows the normalized localization performance $P_{\text{loc}}/P_{\max}$ versus the number C of information cascades, for two distinct choices of observers. Here, the normalization factor is $P_{\max} = \frac{1}{3}$ and represents the maximum attainable probability of localization for this scenario, since the minimum unresolvable set has size 3. As the figure suggests, by collecting information from successive cascades, the observers can average out the variance associated with random propagation, thus achieving higher accuracy of localization.

REFERENCES

[1] Metropolitan Transportation Authority, "Subway entrance and exit GIS data," http://mta.info, 2010.

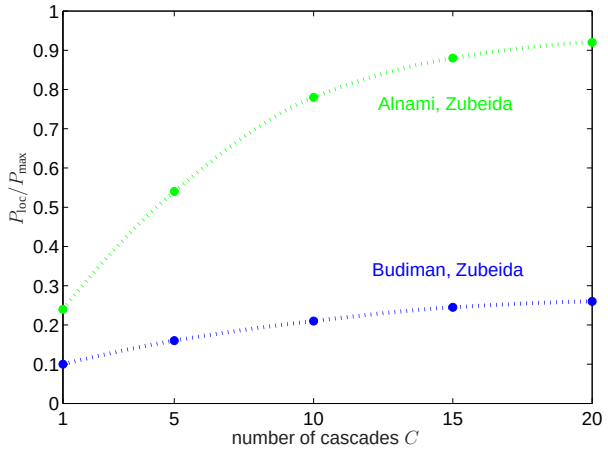
[2] V. E. Krebs, "Uncloaking terrorist networks," *First Monday*, vol. 7, no. 4, 2002.



(a) The network of hijackers involved in the September 11th 2001 attack, and their associates. Shortly after the incident, Mohamed Atta was identified by the authorities as the ringleader of the attack. Here, we use two observers (Alnami and Zubeida, in green), who monitor all the incoming messages. Based only on the timing and direction of arrival of the messages received by the two observers, the SPARSEINF algorithm identified three possible sources (in red): Atta, Al-Ani, and Mzoudi. The dark grey nodes represent alternative observer locations, used in the analysis of Fig. 2(c) and Table 2(b).

Observers	Size of unresolvable set
Alnani, Zubeida	3
Budiman, Zubeida	7
Abdi, Zubeida	12
Grandvisir, Zubeida	60

(b) Size of the unresolvable set for various choices of observers.



(c) Probability of correct localization $P_{\rm loc}$ versus the number C of cascades, for two distinct choices of observers ($\eta = 4$).

Figure 2. Localizing the leader of the 9/11 terrorist organization.