SOURCE LOCALIZATION ON GRAPHS VIA ℓ_1 RECOVERY AND SPECTRAL GRAPH THEORY

Rodrigo Pena, Xavier Bresson, Pierre Vandergheynst

Ecole Polytechnique Federale de Lausanne (EPFL)
School of Engineering (STI)

ABSTRACT

We cast the problem of source localization on graphs as the simultaneous problem of sparse recovery and diffusion kernel learning. An ℓ_1 regularization term enforces the sparsity constraint while we recover the sources of diffusion from a single snapshot of the diffusion process. The diffusion kernel is estimated by assuming the process to be as generic as the standard heat diffusion. We show with synthetic data that we can concomitantly learn the diffusion kernel and the sources, given an estimated initialization. We validate our model with cholera mortality and atmospheric tracer diffusion data, showing also that the accuracy of the solution depends on the construction of the graph from the data points.

***Index Terms*—** Source localization, graph, sparsity, optimization

1. INTRODUCTION

Source localization vaguely refers to a wide class of problems in which the spatial origins of some given diffused information are important to identify. Finding the starting point of an epidemic, the source of heat in a sensor network, or the origin of a rumor in a social network are all examples that fit into this category. We aim at introducing an abstract framework for solving this class of problems without knowledge of the number of sources, and with soft assumptions on the information diffusion process. Our framework leverages the structure of the signal to be recovered, namely its sparsity.

An important source of inspiration for our work is the research of Candès and Fernandez-Granda on the super-resolution of point sources [1], [2], [3]. They study the estimation of sparse signals, with support in a subset of $\mathbb{R}$, from low-resolution observations. The measurement, y, is modeled as a convolution of the original sparse signal, x, with a low-pass point-spread function, and the recovery of x from y is cast as a convex optimization problem, with a fidelity term and a sparsity-inducing norm on x. Candès and Fernandez-Granda show that, in the noiseless setting, x can be exactly recovered from y by solving this optimization problem, as long as the spikes in x obey a certain minimum separation constraint [1]. This technique has also been studied in detail by Duval and Peyré in [4].

Our attempt in this work is to cast a similar optimization problem to solve source localization problems on *graphs*. We do this by modeling the source signals as functions whose domain consists on the nodes of the network. In this context, the eigenvectors of the graph Laplacian play a similar role as the Fourier modes on the real line, and the diffusion of the source signals can be modeled as the action of a linear operator which is a function of this graph Laplacian [5]. Unlike Candès and Fernandez-Granda, however, we do not assume in general that the diffusion process is known. Rather, we assume it to be given by a parametrized function of the graph Laplacian and attempt to learn this parameter at the same time as the source locations. We note, however, that this simultaneous learning makes the overall optimization problem nonconvex and initialization-dependent.

Unlike our proposed technique, which relies on a *global* approach of diffusion, most other works in the literature focus on *local* strategies, observing small fractions of nodes, and leveraging information from the detection times at the observed nodes. One such example is the work of Pinto *et al.* [6], who propose a maximum likelihood estimator that is optimal for trees, and otherwise performs best on scale-free networks. Similarly, Feizi *et al.* [7] use maximum likelihood and minimum error estimators to identify the sources, but they improve on the complexity of the algorithm by modeling the diffusion of information among pairs of nodes as depending only on k-shortest-paths between them, which can be too strong an assumption in some cases. More recently, Zhang *et al.* [8] proposed a nonconvex regression learning model for estimating anomalous diffusion sources. They jointly learn the number of sources, and the propagation time and paths by observing the values and detection times on a subset of network nodes. We differ from Zhang *et al.* by casting a different optimization problem, and by observing only a single snapshot of the diffusion process. We should also finally mention NETSLEUTH, by Prakash *et al.* [9], that differs from the aforementioned strategies by employing the Minimum Description Length (MDL) principle to identify the set of source nodes.

Our main contributions can be summarized as follows:

- A generic optimization framework, made possible by using spectral graph theory tools;
- Simultaneous learning of sparse sources and diffusion

978-1-5090-0746-2/16/$31.00

kernel from a single snapshot of the process;
- Specification of an error measure for comparing the recovered sources to the ground truth; and
- A first analysis of how the graph construction can influence the accuracy of the source localization.

2. THEORY

We consider undirected, weighted graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, consisting of a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E}$, and a weighted adjacency matrix W. Each entry W_{ij} of W represents the weight of the edge between nodes $i, j \in \mathcal{V}$, with $W_{ij} = 0$ if vertices i and j are not connected. Because $\mathcal{G}$ is undirected, W is a symmetric matrix. The sparse signal representing the sources of diffusion is a function $x : \mathcal{V} \to \mathbb{R}$ with support in a subset of $\mathcal{V}$.

Let D be a diagonal matrix with entries $D_{ii} = \sum_j W_{ij}$, and call it the graph's degree matrix. The normalized graph Laplacian [10] is defined then as $\mathcal{L} = I - D^{-1/2} W D^{-1/2}$. By construction, the graph Laplacian is symmetric and positive semidefinite. Therefore, it admits an eigendecomposition, with non-negative eigenvalues, $\mathcal{L} = U \Lambda U^T$, where each column of U is one of $\mathcal{L}$'s eigenvectors, *i.e.*, the graph Fourier modes, and Λ is a diagonal matrix whose entries are the eigenvalues corresponding to each of the eigenvectors in U. We assume, without loss of generality, that the eigenvalues in Λ are ordered, *i.e.*, $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_n = \lambda_{max} \leq 2$, inducing a respective ordering on the columns of U.

We can model the diffusion of the sparse signal x on the graph as the left-multiplication of a function of the Laplacian. The diffusion operator can also be defined on the spectral domain, as a function of the Laplacian eigenvalues. Throughout this work, we will use the heat kernel

$$g_\theta(\lambda) = \exp(-\theta\lambda), \quad \theta > 0. \tag{1}$$

to model information diffusion, but the extension to other parametric diffusion kernels is straightforward. We obtain the corresponding diffusion matrix by simply returning from the spectral domain to the graph domain: $A_\theta = g_\theta(L) = U g_\theta(\Lambda) U^T$.

We can finally state our optimization problem for source localization on graphs as

$$\min_{x,\theta} E(x, \theta) = \min_{x,\theta} \left\{ \gamma \left|\left|x\right|\right|_1 + \frac{\alpha}{2} \left|\left|A_\theta x - b\right|\right|_2^2 \right\}, \tag{2}$$

where b is the observed signal on the graph. We promote sparsity on the sources through the ℓ_1 norm, while the other term accounts for the fidelity with respect to the observations. The parameters γ and α control the trade-off between those two.

In problem (2) we have both x and θ as unknowns, so we take an alternating approach to solving it. At iteration k, we optimize first for x and then for θ:

$$\begin{cases} x_{k+1} = \arg\min_x E(x, \theta_k) \\ \theta_{k+1} = \arg\min_\theta E(x_{k+1}, \theta) \end{cases}, \tag{3}$$

given some initial point (x_0, θ_0). We stop the process when $|E(x_{k+1}, \theta_{k+1}) - E(x_k, \theta_k)| < \epsilon$, for some fixed tolerance $\epsilon > 0$, or if it has attained a given maximum number of iterations.

The first step of (3) is solved by fast iterative shrinkage-thresholding (FISTA) [11], while the second step is solved with a smoothed version of Newton's method (simply adding a proximal term w.r.t. previous estimates θ_k). The algorithms were implemented in MATLAB and are available in the first author's github repository [1]

2.1. Error measure

We specify an error measure based on hop distances similar to the one given in [8]. Let $x : \mathcal{V} \to \mathbb{R}$ be the reference signal, with non-zero values (spikes) on the source nodes, and let $y : \mathcal{V} \to \mathbb{R}$ be the test signal. Let also $\mathcal{A} \subseteq \mathcal{V}$, which we call the active nodes set, contain the nodes with spikes in x. For each $i \in \mathcal{A}$, define a set $\mathcal{N}_i \subseteq \mathcal{V}$ containing the nodes in $\mathcal{V}$ which are closer to i than to any other element of $\mathcal{A}$. Call the set $\mathcal{N}_i$ the influence zone of node i. The distance $h(i, j)$ between two nodes $i, j \in \mathcal{V}$ is measured, in hops, as the shortest path in $\mathcal{G}$ between i and j. The average hop error between those signals can then be written as

$$e(x, y) = \sum_{i \in \mathcal{A}} \frac{\sum_{j \in \mathcal{N}_i} |y(j)|\, h(i, j)}{\sum_{j \in \mathcal{N}_i} |y(j)|}. \tag{4}$$

Each term inside the outermost sum in (4) can be seen as the center of mass of y in the influence zone of an active node i, when the origin of the coordinate system is set to node i. This interpretation makes clear that (4) penalizes both non-sparse test signals, and sparse, but misplaced (with respect to the ground-truth sources) spike signals. As a special case, when $y \equiv 0$ but $x \neq 0$, we set $e(x, y) = \infty$.

3. EXPERIMENTS

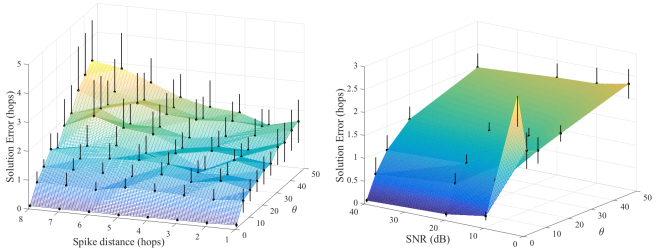
3.1. Sensor Graph

Sensor graphs are constructed by first picking random points on the plane and then connecting each one to its k-nearest neighbors (k-NN). The weight of the edge between two nodes, i and j, is given by $exp(\frac{-d(i,j)^2}{\sigma^2})$, where $d(i, j)$ is the Euclidean distance between their respective coordinates, and σ^2 is a scaling factor.

We first analyze the accuracy of the solution to (2) with respect to both spike distances, h, and diffusion times, θ. For each pair (h, θ), and for each trial, we first pick at random two spikes, h hops away from one another, on a 250-node sensor graph. We then diffuse this spike signal using (1) with parameter θ to obtain observation b. We recover a sparse signal from this observation by solving (2), and measure the error of this solution with respect to the originally drawn spikes. Figure 1a shows the average and standard deviations, over 32 trials, of

[1] https://github.com/rodrigo-pena/src-localization-graphs

the hop error (4) of the recovered solution for different pairs (h, θ). We see that the hop distance between the sources of diffusion does not seem to affect the accuracy of the solution, while the diffusion time θ has a lot of influence on it.



(a) Average hop error as a function of source distance and diffusion time θ.

(b) Average hop error as a function of the SNR and the diffusion time θ.

Fig. 1: Experiments on the sensor graph.

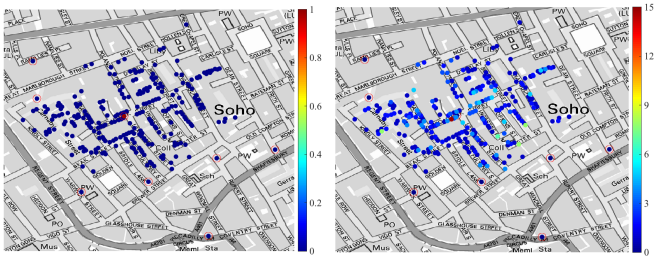
In our next experiment, we analyze the influence of noise. At each trial, we randomly draw two spikes 6 hops away from one another and diffuse them as before. We then add normally distributed noise to b, varying the levels of Signal-to-Noise-Ratio (SNR). Figure 1b shows the average and standard deviations, over 32 trials, of the hop error of the solution for different pairs (SNR, θ). The solver is robust to noise, given that the error grows smoothly with the noise level. We also confirm that a low diffusion time θ is important for accurately localizing the sources of the diffusion. This intuitively makes sense, as one can imagine it is hard to infer the past from observations in the distant future.

3.2. John Snow's Cholera Data

In 1854, a major outbreak of cholera in the district of Soho, in London, inspired a ground-breaking study by the father of modern epidemiology, physician John Snow [12]. Being skeptical of the then dominant miasma theory of diseases, he hypothesized that cholera was transmitted by water, which was reinforced by observing that the infected people's residences seemed to cluster around a water pump on Broad Street.

Snow used a map of the region to illustrate how the cases of cholera were distributed around the infected pump, data which was recently converted into modern Geographic Information System (GIS) format and made available in a blog post by Robin Wilson [13]. In this map, there are 250 points marked with deaths by cholera. Each of these points has an associated death count, ranging from 1 to 15, resulting in a total of 489 deaths. There are also 8 special points in the map, corresponding to the closest water pumps in the area during that time. In Figure 2a, we can see where the infected pump is located, and in Figure 2b, we plot the death counts as a heat map. If we make no distinction between water pump nodes and death nodes, can we automatically recover the position of the infected pump?

We first construct a k-NN graph from the data points. The distance between two points is computed as the length of the shortest path between them on the roads of the map. This seems more in tune with the context of the problem than simply computing the Euclidean length of the line segment between the points. The edge weights are given by an exponential kernel similar to the one used for the sensor graph. We assume the number of deaths at each node is a signal that diffuses on the graph according to a linear model. Our goal then is to recover the sparse signal x (source of infection) that generated the observation b (death counts) after being diffused by some A_θ computed from (1).



(a) Non-zero signal identifying the infected pump (ground-truth).

(b) Observed death count on each node.

Fig. 2: John Snow's GIS cholera death data. The red circles indicate the water pumps.

Modeling cholera transmission merely as a heat diffusion is a heavy simplification, which makes it hard to obtain a satisfying source localization when solving the full non-convex problem (2). We have discovered, however, that if we choose a good value for θ and fix it while solving (3), we manage to reliably recover the location of the infected pump. Another aspect that seemed to noticeably influence the accuracy of the solution was the choice of k when building the k-NN graph. In order to investigate that, we ran our solver for varying values of k and observed what was the average hop error of the converging solution in each case. These results can be seen on the blue curve on Figure 3.

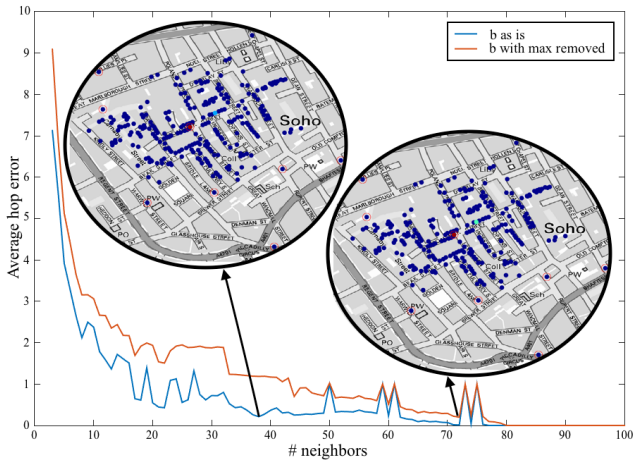


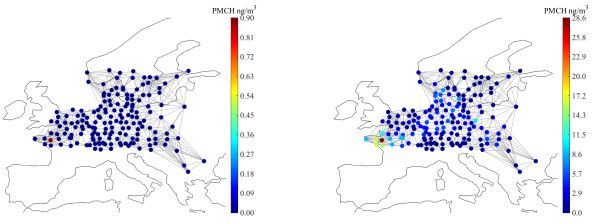
Fig. 3: Snow GIS Graph. Average hop error as a function of the number of neighbors of each node.

Even though we manage to satisfyingly identify the in-

fected pump in the above tests (assuming a proper k was chosen), one could argue that simply choosing the node with the maximum death count in Figure 2b as the source of infection would guarantee a solution with a small hop error. We can account for this particularity of the data by simply removing this "outlier" from our observations b. We do this by either multiplying the term inside the ℓ_2 norm in (2) by an observation mask which is zero at the index of the node signal to be removed, or by substituting this outlier by the interpolation of the signal values on the neighboring nodes. Both those strategies have similar results, so we will not bother comparing them here. The orange curve on Figure 3 shows the results of the same experiment relative to the blue curve, but this time using the outlier-removed observation. As we can see, the solution is robust to this removal, although one might notice that we need a denser graph to attain similar accuracy levels as before.

3.3. European Tracer Experiment (ETEX) Data

In the years that followed the Chernobyl accident, the European Community became very interested in monitoring and modeling the atmospheric transport of chemicals, in particular of radionuclides. One of the studies that were devised in this context was the European Tracer Experiment (ETEX) [14], which took place in 1994. It consisted of two different runs of the same protocol: release, from near Rennes, FR, easily identifiable tracers (perfluorocarbons) on the atmosphere, and sample their concentration, over a period of 72 consecutive hours, at 168 ground-level stations in Western and Eastern Europe. Figure 4 shows a network of these ground stations, overlaid on a map of Europe. The network is assembled as a k-NN graph, but this time the distance between points is simply computed as the Euclidean distance between their coordinates on the globe.



(a) Non-zero signal identifying the tracer release site (ground-truth).

(b) Observed cumulative tracer concentration on each node.

Fig. 4: ETEX tracer concentration data.

We model the observations b as the cumulative tracer concentration on each of the nodes. Some stations had invalid samples, or samples that couldn't be quantified. To account for that, we set the concentration values on these nodes as an interpolation of the values on their neighbors. As before, we assume the diffusion is performed by a heat kernel, and fix a given θ while solving (2) only for x. The goal is to recover a sparse signal x which is non-zero only at the Rennes station.

We were also interested here in observing how the graph construction affects the accuracy of the solution. The blue curve on Figure 5 shows the average hop error of the output of our solver for varying graph densities.

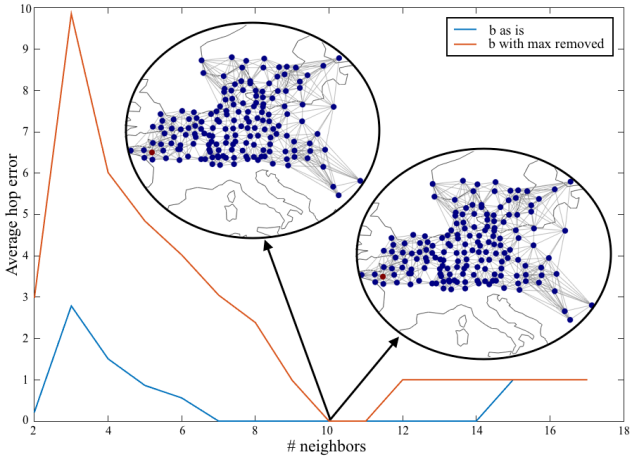


Fig. 5: ETEX Graph. Average hop error as a function of the number of neighbors of each node.

The ETEX data also suffers from a similar bias as Snow's data: simply picking the node with the maximum tracer concentration gives us the true source. Once again, we masked or interpolated this value, and ran our solver for different values of k in the graph construction. This is illustrated by the orange curve on Figure 5, and we see that the results are robust to this information removal.

4. CONCLUSION

We have introduced a framework for solving source localization problems on graphs that is robust to noise and to source distance, but can be very sensitive to the diffusion time of the heat kernel. This sensitivity can perhaps be attenuated if we allow observations at different time steps throughout the diffusion process. In a future work, perhaps more complex diffusion models could be tested when dealing with real data, in an attempt to get a better behavior from the non-convex problem. For instance, Bertuzzo *et al.* [15] develop a fairly detailed dynamic model of cholera epidemics on networks. As a final note, we have also seen that the results depend on the construction of the graph from the given data. Graph construction from data points is still an open problem, but there are some interesting works on the area, e.g., [16], [17], [18].

5. ACKNOWLEDGMENTS

The first author would like to thank Vassilis Kalofolias for the discussions about the problem and its implementation. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement n° 607290 SpaRTaN.

6. REFERENCES

[1] C. Fernandez-Granda, "Super-resolution of point sources via convex programming," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, Dec 2015, pp. 41–44.

[2] Emmanuel J Candès and Carlos Fernandez-Granda, "Towards a Mathematical Theory of Super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, June 2014.

[3] Emmanuel J Candès and Carlos Fernandez-Granda, "Super-Resolution from Noisy Data," *Journal of Fourier Analysis and Applications*, vol. 19, no. 6, pp. 1229–1254, 2013.

[4] Vincent Duval and Gabriel Peyré, "Exact support recovery for sparse spikes deconvolution," *Found. Comput. Math.*, vol. 15, no. 5, pp. 1315–1355, Oct. 2015.

[5] David I Shuman, S K Narang, P Frossard, A Ortega, and P Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, Apr. 2013.

[6] Pedro C Pinto, Patrick Thiran, and Martin Vetterli, "Locating the Source of Diffusion in Large-Scale Networks," *Physical Review Letters*, vol. 109, no. 6, 2012.

[7] Soheil Feizi, Ken Duffy, Manolis Kellis, and Muriel Medard, "Network infusion to infer information sources in networks," Tech. Rep. MIT-CSAIL-TR-2014-028, MIT, Dec. 2014.

[8] Peng Zhang, Jing He, Guodong Long, Guangyan Huang, and Chengqi Zhang, "Towards Anomalous Diffusion Sources Detection in a Large Network," *ACM Transactions on Internet Technology*, vol. 16, Jan. 2016.

[9] B Aditya Prakash, Jilles Vreeken, and Christos Faloutsos, "Spotting Culprits in Epidemics: How Many and Which Ones?," *ICDM*, pp. 11–20, 2012.

[10] Fan Chung, "Spectral Graph Theory," vol. 92, Dec. 1996.

[11] Amir Beck and Marc Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[12] John Snow, "Dr. Snow Report," *Report on the Cholera Outbreak in the Parish of St. James, Westminster, during the Autumn of 1854*, pp. 97–120, July 1855.

[13] Robin Wilson, "John Snow's famous cholera analysis data in modern GIS formats," 2012, [Online]. Available: http://blog.rtwilson.com/john-snows-famous-cholera-analysis-data-in-modern-gis-formats/.

[14] "European Tracer Experiment (ETEX)," 1995, [Online]. Available: https://rem.jrc.ec.europa.eu/RemWeb/etex/.

[15] E Bertuzzo, R Casagrandi, M Gatto, I Rodriguez-Iturbe, and A Rinaldo, "On spatially explicit models of cholera epidemics," *Journal of The Royal Society Interface*, vol. 7, no. 43, pp. 321–333, Feb. 2010.

[16] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, July 2008.

[17] Xiaowen Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Laplacian matrix learning for smooth graph signal representation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 3736–3740.

[18] Vassilis Kalofolias, "How to learn a graph from smooth signals," in *th International Conference on Artificial Intelligence and Statistics AISTATS*, Cadiz, Spain, 2016.