Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in DISCRETE MATHEMATICS

Electronic Notes in Discrete Mathematics 50 (2015) 65–70

www.elsevier.com/locate/endm

Budgeted sensor placement for source localization on trees

L. E. Celis [a,1] F. Pavetić [b,2] B. Spinelli [a,3] P. Thiran [a,4]

[a] *School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland*

[b] *Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*

Abstract

We address the problem of choosing a fixed number of sensor vertices in a graph in order to detect the source of a partially-observed diffusion process on the graph itself. Building on the definition of *double resolvability* we introduce a notion of *vertex resolvability*. For the case of tree graphs we give polynomial time algorithms for both finding the sensors that maximize the probability of correct detection of the source and for identifying the sensor set that minimizes the expected distance between the real source and the estimated one.

Keywords: Double Resolvability, Sensor Placement, Source Localization.

[1] Email: elisa.celis@epfl.ch.

[2] Email: fpavetic@google.com. The work has been done during the author's enrollment in the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia and prior to the current employment by Google Switzerland GmbH.

[3] Email: brunella.spinelli@epfl.ch. The author was supported in part by the Bill & Melinda Gates Foundation, under Grant No. OPP1070273.

[4] Email: patrick.thiran@epfl.ch.

http://dx.doi.org/10.1016/j.endm.2015.07.012

1 Introduction

The problem of source localization has received considerable attention in the last years. Most approaches, starting with the seminal work by Shah and Zaman [1], rely on knowing the state of the entire network at a given instant in time. Pinto [2] introduced a model that instead, estimates the source based on a sparse set of sensor vertices in the graph. However, the set of sensors is often assumed to be given. In this work we consider the problem of selecting *which* vertices to observe given a *budget* on the number (or cost) of allowed sensors in order to optimize source detection.

Consider a graph $\mathcal{G}(V, E)$ with weighted edges. We say that a vertex gets *infected* when it is reached by a diffusion process on the graph; the moment at which this happens is called the *infection time*. Given a budget $k \in \mathbb{N}$, we are interested in finding $S \subseteq V$ of size k such that the infection times of the vertices of S maximize the accuracy in source identification. Specifically, our algorithm allows us to identify a set of *candidate source vertices*.

Depending on the context in which we want to localize the source of a diffusion, we could be more interested in maximizing the chances of an exact identification of the source or in minimizing, in average, the distance between the real source s^* and the estimated source $\hat{s}$. Hence, assuming that s^* can appear uniformly at random in V, we consider two metrics: (i) the *error probability*, i.e., $\mathcal{P}_e = \mathcal{P}(\hat{s} \neq s^*)$; (ii) the *expected distance between the real source* s^* *and the estimated source* $\hat{s}$, i.e., $\mathbb{E}[d(s^*, \hat{s})]$, where d is the weighted distance between two vertices in the graph. It is easy to see that the two metrics may require different sets of sensors.

An approximation algorithm for minimising the cardinality of the sensor set that *perfectly* detects the source was given [3] using the connection to the Doubly Resolving Set (DRS) of a graph [5]. A similar question was first considered by [4]when *starting time* of the diffusion is known. In both cases, budget constraints are not considered.

Based on the definition of DRS, we introduce a concept of *vertex resolvability* and show that the performance of a set of sensor vertices with respect to the error probability is directly linked to the number of *unresolved* vertices. For the case of a tree graph $\mathcal{T}$ of size n, we design an $O(nk^2)$ dynamic-programming algorithm to find k sensor vertices that minimize the number of unresolvable vertices and hence the error probability. Minimizing the expected error distance, was, to the best of our knowledge, never considered before. Also for this metric, we show that if $\mathcal{G}$ is a tree, an optimal set S can be found with a polynomial-time algorithm.

2 Preliminaries

Consider a weighted graph $\mathcal{G}(V, E)$ that models a contact network. A diffusion process on $\mathcal{G}$ is started by a single unknown vertex s^*, the *source*, at an unknown time t^*. If a vertex v becomes infected at time t_v, each non-infected neighbor u of v gets infected at time $t_u = t_v + w_{u,v}$ where $w_{u,v} \in \mathcal{R}_+$ is the weight of edge (u, v). For every vertex s in the *sensor set* S the infection time t_s is known. We estimate the position of the source based only on the infection times $\{t_s, s \in S\}$. The time at which the diffusion starts is unknown, hence a single observed infection time does not give any information about the position of the source. We choose one sensor, say $s_1 \in S$ as *reference point* and define a vector of relative observed times as follows.

Definition 2.1 [observation vector] Let $\mathcal{G}$ be a graph, $S \subseteq V$, $|S| = k$ a set of sensors and $\{t_s, s \in S\}$ the infection times observed during a diffusion process. Then $\boldsymbol{\tau} \in \mathbb{R}^{k-1}$ where $\tau_i = t_{s_{i+1}} - t_{s_1}$, $i \in [k-1]$, is the observation vector associated to the diffusion process.

Definition 2.2 [distance vector] Let $\mathcal{G}$ be a graph, $S \subseteq V$, $|S| = k$ a set of sensors, for each candidate source s we define its distance vector as $\mathbf{d}_s$ where $\mathbf{d}_{s,i} = d(s_{i+1}, s) - d(s_1, s)$, $i \in [k-1]$, and d is the weighted graph distance.

Definition 2.3 [resolved / unresolved vertex] A vertex u is resolved by a set S if $\mathbf{d}_u \neq \mathbf{d}_v$ for all $v \in V$, $v \neq u$, and unresolved otherwise.

Note that $u \sim v$ iff $\mathbf{d}_u = \mathbf{d}_v$ is an equivalence relation and we call $[u]_S$ the class of vertices equivalent to u. A set S such that $[u]_S = \{u\}$ for each $u \in V$ is a Double Resolving Set, as clarified by the definition and lemma below.

Definition 2.4 [Double Resolvability] Given a graph $\mathcal{G}$, $S \subseteq V$ is said to doubly resolve $\mathcal{G}$ if for any $x, y \in V$ there exist $u, v \in S$ s.t. $d(x, u) - d(x, v) \neq d(y, u) - d(y, v)$. Such a subset S is a Double Resolving Set for $\mathcal{G}$ (DRS).

Lemma 2.5 (Lemma 3.1 in [3]) *Let $S \subseteq V$ and fix $s \in S$. Then every vertex $u \in V$, $u \neq s$ resolved by S is resolved by a pair in $\{(s, v) : v \in S \backslash \{s\}\}$.*

As a consequence of Lemma 2.5, the definition of resolved and unresolved vertices above does not depend on the choice of *reference point* $s_1 \in S$.

Lemma 2.6 *Let $\mathcal{T}$ be a tree with n vertices, $S \subseteq V$.*

(i) *Let $u \in S$ a non-leaf vertex or $u \in V \backslash S$. If there do not exist $s_1, s_2 \in S$,*

$s_1, s_2 \neq u$, *s.t.* $u \in \mathcal{P}(s_1, s_2)$ [5] *, then* u *is not resolved by* S*;*

(ii) *let* $u \in V$ *and consider* $\mathcal{T}$ *as rooted at* u*. If every subtree* $\mathcal{T}_c$ *rooted at a child* c *of* u *contains at least one vertex of* S*, then* u *is resolved;*

(iii) *let* $|S| > 1$*: a leaf-vertex* ℓ *is resolved if and only if* $\ell \in S$*.*

If the source of the diffusion is s^* and the observation vector is $\boldsymbol{\tau}$, then all vertices in $[s^*]$ are *candidate source vertices* because their distance vectors are equal to $\boldsymbol{\tau}$. Given a prior distribution π on V for the position of s^*, we select an approximated source $\hat{s}$ by sampling the conditional distribution $\pi|_{[s^*]}$.

Remark 2.7 On trees, this model and estimator tolerate a uniformly bounded amount of noise in the transmission delays: in fact, the estimation of the source would have the same accuracy if for a vertex u infected by its neighbor v, $t_u = t_v + w_{u,v} + X_{u,v}$ where $X_{u,v} \in [-\varepsilon, \varepsilon]$ is a random variable and $\varepsilon < \min_{(u,v)\in E}[w_{u,v}/\text{diameter}(\mathcal{G})]$.

3 Error Probability Minimization

Proposition 3.1 *Let* $\mathcal{G}$ *be a graph of size* n*,* $S \subseteq V$ *and uniform prior* π*. The probability of error* $\mathcal{P}_e(S)$ *is given by* $\mathcal{P}_e(S) = \frac{1}{n}\sum_{[u]_S \subseteq V}(|[u]_S| - 1)$.

If q is the number of equivalence classes we have $\mathcal{P}_e = 1 - q/n$ and it is clear that the error probability is minimized if the number of equivalence classes is maximized. Looking back to Lemma 2.6, if the graph is a tree $\mathcal{T}$, $\mathcal{P}_e$ is 0 if a sensor is placed on each leaf. [6] In fact, the minimum k required for $\mathcal{P}_e = 0$ is the number of leaves ℓ. Moreover, if $k < \ell$, the vertices that minimize $\mathcal{P}_e$ are a subset of the leaves of $\mathcal{T}$. This suggests that, given a tree and a sensor set, if we root the tree at an arbitrary vertex it is possible to compute $\mathcal{P}_e$ as the sum of the probabilities of error of the different subtrees. Building on this observation we prove that, for any n-vertex tree $\mathcal{T}$ and budget $k \in \mathbb{N}$, a set S^k_{opt} that minimizes $\mathcal{P}_e$ can be found with a recursive algorithm of total complexity $O(k^2 n)$.

Theorem 3.2 *Let* $\mathcal{T}$ *be a tree with* n *vertices and* ℓ *leaves and let the prior* π *be uniform. If* $k \geq \ell$*, the leaf set is an optimal sensor set. If* $k \in [\ell - 1]$*, there exists an algorithm that finds* $S^k_{opt} \in \text{argmin}_{|S|=k}\mathcal{P}_e(S)$ *in time* $O(nk^2)$.

Proof. [Correctness] The statement is trivial for $k \geq \ell$ as the set of leaves resolves all the vertices. If $k < \ell$, call $\mathcal{T}_r$ the tree obtained rooting $\mathcal{T}$ at an

[5] $\mathcal{P}(s_1, s_2)$ denotes the unique shortest path between s_1 and s_2 on a tree $\mathcal{T}$.

[6] See also [3] for a different proof.

arbitrary non-leaf vertex r. We claim that S^k_{opt} is obtained through the main function of Algorithm 1, i.e., by computing OptErr$(\mathcal{T}_r, k)$. We prove the statement by strong induction on the height of the tree.

Fix a budget k' and let $p(\mathcal{T}_x, k')$ be the contribution to the error probability from $\mathcal{T}_x$ assuming k' sensors are placed optimally in $\mathcal{T}_x$. The base case is a subtree $\mathcal{T}_x$ of height 0, i.e., a leaf: if $k' \geq 1$ then we can place a sensor directly on the leaf. If there is at least one other sensor in $\mathcal{T}_r$ (if $k' < k$), we can resolve it due to Lemma 2.6(iii). If $k' \in \{0, k\}$ then we cannot resolve it and $p(\mathcal{T}_x, 0) = 1/n$. Now consider the general case of a rooted tree $\mathcal{T}_x$ of height $h > 0$, and assume we can find $p(\mathcal{T}_i, k'_i)$ for all trees $\mathcal{T}_i$ of height less than h. If $k' = 0$, then $p(\mathcal{T}_x, 0) = |T_x|/n$ since we have no way to distinguish between any vertices in $\mathcal{T}_x$. Otherwise, we recurse over all possible partitions k' between the subtrees rooted at the children of x.[7] In particular, if g is the number of children of x and $\mathcal{T}_{x,i}$, for $i \in [g]$, denotes the subtree rooted at the ith child of x, any configuration of k' sensors in $\mathcal{T}_x$ has $0 \leq k'_i \leq k'$ sensors in subtree $\mathcal{T}_{x,i}$ with $\sum_{i=1}^{g} k'_i = k'$. Hence, $p(\mathcal{T}_x, k') = \sum_{k'_i=0}(|\mathcal{T}_{x,i}|/n) + \sum_{k'_i \neq 0} p(\mathcal{T}_{x,i}, k'_i)$. In fact, x is equivalent to all vertices in the subtrees $\mathcal{T}_{x,i}$ (if any) for which $k'_i = 0$ and $|[x]| - 1 = \sum_{k'_i=0} |\mathcal{T}_{x,i}|$. Since the height of each $\mathcal{T}_{x,i}$ is less than h, by the induction hypothesis we can compute the optimal $p(\mathcal{T}_{x,i}, k'_i)$, and hence $p(\mathcal{T}_x, k')$. By induction, this concludes the proof. □

Proof. [Complexity] A call to OptErrChildren is determined by the root x of the subtree, the subset c of its children considered and the budget $k' \leq k$. The possible values for the pair x, c is the number of edges $n-1$. In fact, we can assume that the children are ordered and the possible partitions are of the form (c, children at the right of c) so the number of pairs (x, c) is bounded by $n-1$. Hence, there are $O(nk)$ possible calls of OptErrChildren. Combining this with the minimization on $m \leq k$ sensors sent to the leftmost sub-tree, the complexity is $O(nk^2)$. □

Algorithm 1 *Minimizes $\mathcal{P}_e$ for initial budget k on a tree of size n*

OptErr($\mathcal{T}_x$, k')
if $k' = 0$ **return** $|\mathcal{T}_x|/n$
else if $|\mathcal{T}_x| = 1$
  if $1 \leq k' < k$ **return** 0, **else return** $1/n$
return OptErrChildren($\mathcal{T}_x$, k', *children*(x))
OptErrChildren($\mathcal{T}_x$, k', C)
if $|C| = 0$ **return** 0, **else if** $k' = 0$ **return** $\sum_{c \in C} |subtree(c)|/n$

[7] See the function OptErrChildren in Algorithm 1.

f ← *first child, r* ← *other children, results* ← {}
for *m from* 0 *to k′*
  results ← *results* ∪ {OptErr($\mathcal{T}_f, m$)+ OptErrChildren($\mathcal{T}_x, k' - m, r$)}
return min{*results*}

4 Expected Distance Minimization

If $\mathcal{G}$ is a graph of size n with weighted distance d, S the set of sensors, $|S| = k$, and the prior π uniform, the expected distance between the real source s^* and the estimated source $\hat{s}$ is $\mathbb{E}[d(s^*, \hat{s})] = \frac{1}{n} \sum_{[u]_S} \left(\sum_{s,t \in [u]_S} \frac{d(s,t)}{|[u]|_S} \right)$.
In this case, the contribution of each unresolved vertex depends on the sum of distances between the vertices in an equivalence class *in addition to* the size of the class; this makes the problem more challenging. It can be proven that if $\mathcal{T}$ is a tree of size n with maximum degree D, ℓ leaves and uniform prior π, the leaf set minimizes $\mathbb{E}_S[d(s^*, \hat{s})]$ when $k \geq \ell$; if $k \in [\ell - 1]$, there exists an algorithm that finds $S^k_{opt} \in \text{argmin}_{|S|=k} \mathbb{E}_S[d(s^*, \hat{s})]$ in time $O(2^D n k^2)$.

5 Future Work

An important open problem is extending our results to general graphs. Other interesting directions include optimizing *worst case* metrics rather than *average case* metrics, accounting for noisy infection delays and transmission failures, and non-uniform prior distributions on the position of s^*.

References

[1] Shah, D., and T. Zaman, *Rumors in a network: who's the culprit?*, IEEE Transactions on information theory **57**(8) (2011), 5163-5181.

[2] Pinto, P., P. Thiran, and M. Vetterli, *Locating the Source of Diffusion in Large-Scale Networks*, Physical Review Letters, **109** (2012), 068702.

[3] Chen, X., X. Hu, and C. Wang, *Approximability of the Minimum Weighted Doubly Resolving Set Problem*, Proc. 20^{th} Int. Computing & Combinatorics Conf. (2014), 357-368.

[4] Zejnilovic, S., J.P. Gomes, and B. Sinopoli, *Network observability and localization of the source of diffusion based on a subset of vertices*, Proc. of the 51^{st} Allerton Conf. on Communication, Control & Computing (2013), 847-852.

[5] Cáceres, J., M.C. Hernando, M. Mora, I.M. Pelayo, M.L. Puertas, C. Seara, and D.R. Wood, *On the metric dimension of cartesian products of graphs*, SIAM J. Discrete Math. **21**(2) (2007), 423-441.