Source Detection on Graphs

Tobias Weber [a], Volker Kaibel [a], Sebastian Sager [a]

[a] *Department of Mathematics, Otto-von-Guericke Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany*

Abstract

Spreading processes on networks (graphs) have become ubiquitous in modern society with prominent examples such as infections, rumors, excitations, contaminations, or disturbances. Finding the source of such processes based on observations is important and difficult. We abstract the problem mathematically as an optimization problem on graphs. For the deterministic setting we make connections to the metric dimension of a graph and introduce the concept of spread-resolving sets. For the stochastic setting we propose a new algorithm combining parameter estimation and experimental design. We show well-posedness of this algorithm theoretically and via encouraging numerical results on a benchmark library.

Key words: Source inversion; graph theory; linear regression; metric dimension; optimization; optimal experimental design.

1 Introduction

We consider abstract source detection problems for time-dependent data on graphs. Given a graph consisting of nodes and weighted edges, random noise, and a mathematical model for a spreading process on the nodes of the graph we want to determine optimal strategies to query oracles (i.e., observing one or several nodes at a particular time) and to use the available measurement information to infer the most probable source of the spreading process. Depending on the setting, the subproblems may have to be solved while the underlying process is ongoing. The first one can be interpreted as an experimental design problem, focusing on the collection of information. The second one is the *source detection* (or localization, inversion, identification, resolving) problem, exploiting the gained information to identify the source. While both problems are well understood in various circumstances, especially in Euclidean spaces, we are not aware of a rigorous abstract treatment on graphs.

Most source detection problems on graphs have been tackled from a practical viewpoint. In [17] fire localization in buildings and the coast guard LoRaN stations are investigated. Other applications are the classification of chemical compounds [16,17] or the spread of information or diseases [67]. All this is related to the metric dimension of a graph, which is an important concept for source detection on graphs.

Email addresses: tobias.weber@ovgu.de (Tobias Weber), kaibel@ovgu.de (Volker Kaibel), sager@ovgu.de (Sebastian Sager).

Detecting the source of a pollution in water networks is a practically relevant task. Much research in this area is focused on the sensorics, i.e., which chemical or biological markers have to be tested to distinguish different kind of pollution sources [4,64,10,60]. Additionally, location dependent visualization and interpolation schemes are used [22]. Some researchers also try to detect the source in space. In [45] mainly linear simplified models are used in an optimization framework. Here, flow conditions are assumed to be known and then used to calculate time delays of pollution concentration over the network. These delays are used in a quadratic optimization problem to calculate pollution injection profiles over time for all nodes. A similar flow model based approach to the offline problem can be found in [54]. In [25] the online case is considered with the goal to place a minimal number of sensors to identify the source. The spread of computer viruses and fault propagation in information networks have been modeled as spreading phenomenon on a graph as well, [63,19]. In [63] a stochastic model is used to describe the infection between nodes in the network. In [19] ordinary differential equations (ODEs) are used. The spread of epidemics can be modeled as a phenomenon on a graph [59,20,29,5]. The spreading can be described by a stochastic process or by deterministic ODEs, allowing for analysis of, e.g., thresholds that decide if an epidemics dies out or continues spreading, sizes of infected subpopulations, vaccination schemes to suppress outbreaks, or speed of propagation. In our context we are interested in the source detection of epidemics, similar to the approaches using correlation [14], spectrality [27], Bayesian [3] or centrality based estimators [74,50,21].

Similarly, objects that produce sound can be detected with distributed microphones as a sensor network. Applications range from localizing the talker in a room for camera pointing [9,69,13] to surveillance of outside areas (like crossroads, valleys, or industrial facilities) and underwater areas (sonar) [44,18]. Acoustic waves traveling through air usually have a constant velocity. Like in our setting later, the time distance relationship is linear. The least squares approach in [73], which was also applied to sonar, radar, or radio applications, is similar to our approach. The main difference is the use of distances in the Euclidean space and the a priori knowledge of the velocity. Seismic waves were considered in [65,38], focussing on partial differential equations describing the elastodynamics of the ground. The detection of objects in astronomical images is considered in [33]. Examplary medical applications are neural source detection in the brain for epilepsy research [34] and the mapping and prediction of focal cardiac arrhythmias [71].

Methodological research focuses on ill-posed linear problems [52] and parameter estimation for source detection [8]. The problem of choosing graph oracle queries is an experimental design problem. Designing an experiment to optimize an information criteria is called optimal experimental design [26,40]. Different objective functions have been suggested and analyzed. A very early idea was to minimize the variance of the model prediction [66]. This is now referred to as G-optimal and is equivalent to the so called D-optimal criterion [41]. The D-optimal criterion [68] maximizes the determinant of the Fisher information matrix. Another criterion is to minimize the variance of the parameter estimators of the model [24], referred to as A-optimality. Our focus here is different, as we assume an underlying discrete structure, the special metric space of graphs. We introduce basic concepts and survey related work in the following section.

The main contributions of this paper are a unifying problem definition for a wide range of applications, including problems with weighted and/or directed graphs, limited and deterministic/stochastic spreading information, online/offline settings, and linear dynamics. An algorithm is proposed for which convergence is proven and practical performance and robustness over a wide range of test problems with different properties is shown.

The paper is organized as follows. In Section 2 we provide basic definitions and formalize the source detection problem we are interested in. We also provide a simple example graph that will be used throughout the paper for an illustration of concepts. In Section 3 we consider the special deterministic case in which the oracle provides exact measurements. We generalize the metric dimension by introducing the *spread dimension* of a graph as a tool to solve the deterministic case. In Section 4 we discuss the stochastic problem. Based on linear regression and experimental design we present a solution algorithm and discuss its convergence properties in the limit. In Section 5 we present numerical results. In Section 6 we summarize our findings.

2 Source detection problem

We begin with a definition of the considered problem class. First, we define the underlying discrete structure.

Definition 1 (Graph) *We consider a directed weighted graph $G = (V, E)$ with positive edge lengths $\ell(e) > 0$ for all $e \in E$ and shortest-path-distances $d_{i,j}$ with respect to the length function ℓ between nodes i and j for all $i, j \in V$. Let $n := \#V := card(V)$ and $\#E := card(E)$.*

Assumption 2 (Graph) *We assume to have complete knowledge of the graph G and the length function ℓ, and hence also of the distance function d.*

In practical applications, the nodes $i \in V$ correspond to spatial locations where measurements are possible. Examples are communities, airports, cities, or countries for infectious diseases, points on a 3d surface grid of the human heart, or sensors in water distribution networks. The edges correspond to connections between the nodes, along which "something may be passed on". This might, e.g., be a viral load via infections, electrical excitation of cells, or transported and diffused pollutant. In the interest of a simplification, and taking the risk that this term does not intuitively match every application, we will simply use the term *signal* to denote this in a general way in the following.

Definition 3 (Signal Spreading Process) *We consider a dynamic process on a time horizon $\mathcal{T} := [t_s, t_f]$ that originates from an a priori unknown* source *$s \in V$ and spreads the signal via edges to other nodes of the graph. The edge lengths $\ell(e)$ quantify the distances the signal needs to travel to arrive at adjacent nodes.*

Note that the times t_s and t_f are often unknown. The initial time t_s, also called *offset* and indicating when the signal started at source node s, needs to be estimated. The end time t_f is not relevant for the mathematical model. We make some assumptions for the following.

Assumption 4 (Signal Spreading Process) *We assume that*

(1) The source $s \in V$ is unique.
(2) Signal spreading takes place in a diffusive way, i.e., a signal is passed on from a node i to all nodes j that are adjacent to i.
(3) We assume a constant and homogeneous spreading velocity *$1/c > 0$. Hence, for known distances $d_{s,i}$ we have*

$$t_i := t_s + c \cdot d_{s,i}$$

as the arrival time *at node $i \in V$.*

While the first two assumptions are rather technical, the third assumption is an important restriction of the problem class to a linear model. We note that some applications might need less restrictive assumptions. For example, infections or electric conduction on the heart surface do not have a constant velocity in reality. Also, we are not interested here in measuring the strength of the signal, which may be relevant for certain applications. We now look at the available measurement procedure, abstracted as a data oracle.

Definition 5 (Data Oracle) *An oracle allows to query nodes $i \in V$ and obtain measurement data r_i. The r_i indicate times t_i when the signal arrived at node i, but with measurement noise,*

$$r_i = t_i + \epsilon_i$$

Here, $\epsilon_i \in \mathbb{R}$ is a random variable for each $i \in V$. We call the special case of $\epsilon_i = 0 \; \forall \; i \in V$ the deterministic *and the general case the* stochastic version*.*

Assumption 6 (Data Oracle Output) *We assume to know the distributions of the measurement errors ϵ_i for all $i \in V$.*

Assumption 7 (Data Oracle) *We assume that we query the oracle after all relevant times t_i, i.e., data r_i is available at the time of oracle query. In particular, we do not have the possibility to change the process.*

Definition 8 (Source Detection Problem) *We consider a graph, a signal spreading process, and an oracle as specified in Definitions 1, 3, 5 and the above assumptions. We denote the task to minimize the number of oracle queries to determine the source node $s \in V$ (possibly up to a tolerance with respect to graph distance) as the* source detection problem*.*

The queries of the oracle provide (noisy) arrival times r_i, which can be used to infer the unknown offset t_s, the velocity $1/c$, and the source $s \in V$. In this paper, we consider the following general approach to source detection.

Definition 9 (Source Detection) *The general* source detection approach *is: Repeat $i = 1 \dots i_{max}$ rounds of*

S1) choosing k_i nodes $S_i = \{i_1, \dots, i_{k_i}\} \subseteq V$,
S2) querying the oracle to obtain $r_{S_i} \in \mathbb{R}^{k_i}$, and
S3) estimating a current best guess for the source $j^* \in V$

If $j^* = s$ holds, then we call the approach successful*. The* source detection problem *is to find a successful approach with a minimal number $N = \sum_{i=1}^{i_{max}} k_i$ of oracle queries.*

The special case of $i_{\max} = 1$ is called the *offline version* of the problem. It corresponds to a situation where it is not possible to do calculations between queries to the oracle. The *online version* for $i_{\max} \geq 2$ is not to be confused with more general concepts in online optimization, such as model predictive or dual control. Assumption 7 relates to the properties of the source detection problem and states that our approach starts after the end of the spreading process at time t_f. Note that some processes such as cardiac excitations have a repetitive nature and a fast timescale, compare [71]. Thus, the results of the considered problem class may find application not only in a posteriori analysis, but also in ongoing processes.

The oracle queries in S2) can be practically difficult and/or expensive, giving rise to our approach to minimize their overall number. Thus, all nodes chosen in S1) have to provide as much information as possible. The problem to identify the corresponding nodes can be seen as an optimal experimental design problem on a graph. The estimation or source inversion problem in S3) can be approached based on regression. Note that the main assumption for this model is a spreading of arrival times from the source s to all other vertices via shortest paths at a constant velocity $1/c > 0$. We also assumed that the answer of the oracle does not depend on the round in which it is queried. According to the classification in [35], the above setting corresponds to *sensor observations* in contrast to the the *snapshot* or *full information* cases.

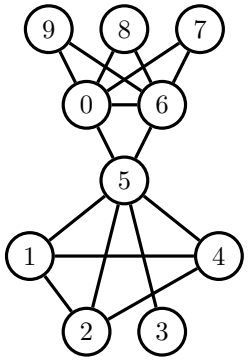
In the interest of simplicity and if not stated otherwise, we will use notation, definitions, and assumptions from this section, without explicit reference. We shall use the following example for illustration throughout this paper.

Example 10 (Graph) *The graph $G = (V, E)$ has nodes*

$$V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

and weights $\ell(e) = 1$ for all undirected edges e in

$$E = \{\{0,5\},\{0,6\},\{0,7\},\{0,8\},\{0,9\},\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{2,5\},\{3,5\},\{4,5\},\{5,6\},\{6,7\},\{6,8\},\{6,9\}\}.$$



i \ j	0	1	2	3	4	5	6	7	8	9
0	0	2	2	2	2	1	1	1	1	1
1	2	0	1	2	1	1	2	3	3	3
2	2	1	0	2	1	1	2	3	3	3
3	2	2	2	0	2	1	2	3	3	3
4	2	1	1	2	0	1	2	3	3	3
5	1	1	1	1	1	0	1	2	2	2
6	1	2	2	2	2	1	0	1	1	1
7	1	3	3	3	3	2	1	0	2	2
8	1	3	3	3	3	2	1	2	0	2
9	1	3	3	3	3	2	1	2	2	0

Fig. 1. Left: visualization of the example graph. Right: symmetric matrix with shortest path distances $d_{i,j}$.

3 Deterministic case

In this section we discuss the *deterministic offline version* of the source detection problem, i.e., $i_{\max} = 1$ and $\epsilon_i = 0 \; \forall \; i \in V$. This problem class deserves special attention, because it is interesting in its own right, it is the idealized limit case of stochastic online versions, and because algorithmic ideas can be iteratively used in more complex settings. Note that it is purely combinatorial, asking for subsets of V for which the oracle answer allows to infer (or resolve) the source.

Of practical relevance is the possibility to verify a source via querying the oracle. Under the assumptions above one possibility is a local enumeration.

Definition 11 (Source Certificate) *A node $s \in V$ is the* source *of the spreading process if and only if t_s is finite and $t_s < t_j$ for all nodes j with $(v_j, v_s) \in E$.*

We start by considering the special case with $t_s = 0$ and $c = 1$, where the oracle returns $r_i = d_{s,i}$ for $i \in V$. To solve the source detection problem we need to choose a minimal cardinality subset of V in S1) for which we question the oracle in S2). The answer shall enable us to calculate the source in S3) no matter which vertex in V actually is the source. This concept is known in graph theory as the metric dimension of a graph [17,16,55,67] and depends on the basis of a graph. Classically, the metric dimension of a graph is defined for unweighted graphs, i.e., $\ell(e) = 1 \; \forall \; e \in E$. We generalize this to weighted graphs (V, E) with weights $\ell(e) > 0$ for $e \in E$.

Definition 12 (B-metric Equivalence) *Given a subset $B \subseteq V$, two nodes $i, j \in V$ are* B-metric equivalent *if $d_{i,k} = d_{j,k} \; \forall \; k \in B$.*

Definition 13 (Metric-Resolving Set) *A set $B \subseteq V$ is* metric resolving*, if $i, j \in V$ are B-metric equivalent if and only if $i = j$.*

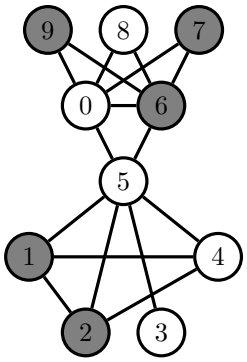
Thus, B is metric-resolving if it uniquely defines all $v \in V$ by their shortest path distances to the elements of B.

Definition 14 (Metric Basis) *A* (metric) basis *B is a metric-resolving set with minimal cardinality.*

Definition 15 (Metric Dimension) *Given a weighted graph $G = (V, E)$,* the metric dimension *is the cardinality of one of its metric bases.*

There are different ways to check if a set B is metric-resolving. Equivalently to Definition 13, one can check if either $\sum_{k \in B} |d_{j,k} - d_{i,k}| = 0$ or (anticipating the stochastic regression case) if $\sum_{k \in B} (d_{j,k} - d_{i,k})^2 = 0$ for all pairs of nodes $i < j \in V$. If the value is strictly positive for (the minimum of) all pairs, then B is metric-resolving.

Example 16 *The graph from Example 10 has metric dimension 5 and one metric basis is $B := \{1, 2, 6, 7, 9\}$. Figure 2 shows that B is a resolving set, as there are no zeros on the off-diagonal. One can show (e.g., by enumeration) that no basis with fewer nodes exists.*



i \ j	0	1	2	3	4	5	6	7	8	9
0	0	2	5	1	4	17	24	16	24	14
1	2	0	5	1	4	17	24	16	24	14
2	5	5	0	2	5	12	13	5	13	9
3	1	1	2	0	3	14	19	11	19	11
4	4	4	5	3	0	5	12	8	12	4
5	17	17	12	14	5	0	5	5	5	1
6	24	24	13	19	12	5	0	4	8	4
7	16	16	5	11	8	5	4	0	4	4
8	24	24	13	19	12	5	8	4	0	4
9	14	14	9	11	4	1	4	4	4	0

Fig. 2. Left: the graph from Example 10 with the basis in gray. Rigth: symmetric matrix with entries $\sum_{k \in B} (d_{j,k} - d_{i,k})^2$ for the metric basis from Example 16.

Deciding whether a graph has metric dimension less than a given value is NP-complete [39]. Hence, determining the metric dimension even of an unweighted graph is difficult [31]. This computational complexity refers to step S1, the experimental design problem. To find a basis one can enumerate all possible vertex sets from small to large cardinality until a basis is found. In [31] an $(1+(1+o(1))\log(n))$-approximation algorithms is given which runs in $\mathcal{O}(n^3)$, with $n = \text{card}(V)$. The metric dimension can not be approximated within $o(\log(n))$ [30]. If a basis has been found in S1) and the oracle queries returned r_k for all $k \in B$ in S2), the source $s \in V$ can be uniquely determined in S3) by calculating $d_{i,k}$ for all $i \in V$ and $k \in B$ and comparing it to $r_k = d_{s,k}$.

Example 17 *Assume that for the basis B from Example 16 the oracle returns $r_{\{1,2,6,7,9\}} = (3, 3, 1, 2, 2)$. Comparison with the full distance table on the right hand side of Figure 1 reveals the source node $s = 8$.*

We are interested in a generalization of this concept to arbitrary and a priori unknown velocity $1/c > 0$ and offset $t_s \in \mathbb{R}$. Again, we want to be able to uniquely determine the source, now for arbitrary $c > 0$, t_s, and $s \in V$. While the concepts of a metric basis and of doubly resolving sets [15,43] can be found in the literature, the spread basis is a novel concept.

Definition 18 (B-spread Equivalence) *For $B \subseteq V$, two nodes $i, j \in V$ are* B-spread equivalent *if*

$$\exists t_i, t_j \in \mathbb{R}, c_i, c_j > 0 : \quad t_i + c_i d_{i,k} = t_j + c_j \; d_{j,k} \qquad k \in B.$$

Note that with the choice of $t = (t_j - t_i)/c_i$ and $c = c_j/c_i$ this is equivalent to

$$\exists t \in \mathbb{R}, c > 0 : \quad d_{i,k} = t + c \; d_{j,k} \quad \forall \; k \in B.$$

Definition 19 (Spread-Resolving Set) *A set $B \subseteq V$ is* spread-resolving, *if $i, j \in V$ are B-spread equivalent if and only if $i = j$.*

Definition 20 (Spread Basis) *A* spread basis *B is a spread-resolving set with minimal cardinality.*

Definition 21 (Spread Dimension) *Given a weighted graph $G = (V, E)$,* the spread dimension *is the cardinality of one of its spread bases.*

Although the interpretation of a velocity $1/c$ is not well posed for $c = 0$, we will require $c \geq 0$ instead of $c > 0$ in the following minimization problems to avoid open sets. To find a spread basis we consider the objective function

$$J_j(t, c, r_S) = \sum_{k \in S} (t + c\, d_{j,k} - r_k)^2. \tag{1}$$

Minimizing this objective with $r_k = d_{i,k}$ and constraint $c \geq 0$ results in an optimal objective value $\phi_{i,j}(S)$ depending on i, j, and S. As above, an equivalent criterion to check if a set $S \subset V$ is spread-resolving is to check if

$$\phi^*(S) = \min_{i,j \neq i \in V} \phi_{i,j}(S) = \min_{i,j \neq i \in V} \min_{t, c \geq 0} J_j(t, c, d_{i,S}) \tag{2}$$

is strictly positive, compare Example 23.

Proposition 22 (Sign symmetric objective) *For any subset S of V the objective values $\phi_{i,j}(S)$ are sign symmetric, i.e., for $i, j \in V$ we have*

$$\begin{array}{rcl} \phi_{i,j}(S) = 0 & \iff & \phi_{j,i}(S) = 0 \\ \phi_{i,j}(S) > 0 & \iff & \phi_{j,i}(S) > 0. \end{array} \tag{3}$$

PROOF. For $\phi_{i,j}(S) = 0$ with $c > 0$ and t_s we have by equations (1) and (2):

$$c\, d_{j,k} + t_s = d_{i,k} \,\forall\, k \in S.$$

Reformulating this results in

$$1/c\, d_{i,k} - t_s/c = d_{j,k} \,\forall\, k \in S$$

with new slope $1/c > 0$ and offset $-t_s/c$. Then by equations (1) and (2) again we have $\phi_{j,i}(S) = 0$. The second part follows from the first due to $\phi_{j,i}(S) \geq 0$.

Example 23 *For the graph from Example 10 the metric basis $B = \{1, 2, 6, 7, 9\}$ is not spread-resolving. E.g., for $t_s = c = 1$ we have $d_{8,k} = t_s + d_{6_k}$. The graph has spread dimension 7 and $B^{sp} := \{1, 2, 3, 6, 7, 8, 9\}$ ia a spread basis. Figure 3 shows that B^{sp} is spread-resolving, as only diagonal values are zero, and that B is not spread-resolving, as $\phi_{6,8}(B) = \phi_{8,6}(B) = 0$.*

i \ j	0	1	2	3	4	5	6	7	8	9
0	0.00	8.00	8.00	6.86	4.86	1.71	0.75	2.75	2.75	2.75
1	1.71	0.00	1.88	5.73	0.36	0.59	3.43	8.00	8.00	8.00
2	1.71	1.88	0.00	5.73	0.36	0.59	3.43	8.00	8.00	8.00
3	1.71	6.69	6.69	0.00	3.67	0.75	3.43	8.00	8.00	8.00
4	1.71	0.59	0.59	5.18	0.00	0.35	3.43	8.00	8.00	8.00
5	1.71	2.75	2.75	3.00	1.00	0.00	3.43	8.00	8.00	8.00
6	0.37	8.00	8.00	6.86	4.86	1.71	0.00	3.33	3.33	3.33
7	0.59	8.00	8.00	6.86	4.86	1.71	1.43	0.00	6.00	6.00
8	0.59	8.00	8.00	6.86	4.86	1.71	1.43	6.00	0.00	6.00
9	0.59	8.00	8.00	6.86	4.86	1.71	1.43	6.00	6.00	0.00

i \ j	0	1	2	3	4	5	6	7	8	9
0	0.00	6.80	6.80	1.20	4.00	1.20	0.67	2.00	0.67	2.00
1	1.20	0.00	1.85	0.35	0.32	0.35	2.80	6.80	2.80	6.80
2	1.20	1.85	0.00	0.35	0.32	0.35	2.80	6.80	2.80	6.80
3	1.20	2.00	2.00	0.00	0.67	0.00	2.80	6.80	2.80	6.80
4	1.20	0.55	0.55	0.20	0.00	0.20	2.80	6.80	2.80	6.80
5	1.20	2.00	2.00	0.00	0.67	0.00	2.80	6.80	2.80	6.80
6	0.29	6.80	6.80	1.20	4.00	1.20	0.00	3.14	0.00	3.14
7	0.35	6.80	6.80	1.20	4.00	1.20	1.29	0.00	1.29	5.65
8	0.29	6.80	6.80	1.20	4.00	1.20	0.00	3.14	0.00	3.14
9	0.35	6.80	6.80	1.20	4.00	1.20	1.29	5.65	1.29	0.00

Fig. 3. Top: matrix with objective values $\phi_{i,j}(B^{\text{sp}})$ for the spread-basis from Example 23. Bottom: matrix with objective values $\phi_{i,j}(B)$ for the metric basis from Example 16. Both are not symmetric, as $\phi_{i,j}$ can differ from $\phi_{j,i}$.

Our considerations suggest a (not necessarily efficient) approach to find a spread basis. In (2), one can detect infeasibility or solve the inner minimization problem analytically and enumerate the outer minimization problem over all (modulo symmetry because of Proposition 22) pairs of nodes and all subsets S of V. Checking if $\phi^*(S) > 0$ allows to find a spread-resolving set S of minimal cardinality, similar to the metric case.

To close the section, we collect some results on bounds for the spread dimension. We are interested in behavior for large $n = \#V$, hence we assume $n \geq 4$ to avoid the discussion of special cases for the following results. If all edges have equal length, a trivial upper bound on the spread dimension is $n - 1$. This bound is active for the special case of complete graphs.

Proposition 24 (Dimension of Complete Graphs) *Let G be a complete graph with equal weight $\ell > 0$ on all edges. Then the spread dimension of G is $n - 1$.*

PROOF. We have $r_s = t_s$ and the same oracle answer $r_i = t_s + c\ell > t_s$ for all $i \in V \backslash \{s\}$ and for all choices t_s and $c > 0$. Hence, the source s can only be identified if either $s \in B$, or if s is the only node in $V \backslash B$. As the spread basis needs to identify all possible s, we have necessarily $\text{card}(B) = n - 1$.

If the edge weights are not identical, we may even need all n nodes in the spread basis. Thus, complete graphs are the worst case in terms of an upper bound for the spread dimension. However, also other topologies, such as star graphs, may have large spread dimensions.

By definition, the metric dimension is a lower bound for the spread dimension. Furthermore, we have the following lower bound for all graphs.

Proposition 25 (Lower Bound for Dimension) *Let G be a graph as in Definition 1 with $n \geq 4$. Then the spread dimension of G is at least 3.*

PROOF. Assume a spread basis $B = \{i, j\}$ of cardinality two. Choose $v, w \in V \backslash B$ with $v \neq w$.

If $d_{v,i} = d_{v,j}$ and $d_{w,i} = d_{w,j}$ hold then with

$$c := \frac{d_{v,i}}{d_{w,i}} = \frac{d_{v,j}}{d_{w,j}}$$

we obtain $d_{v,i} = cd_{w,i}$ and $d_{v,j} = cd_{w,j}$, contradicting B being spread-resolving.

Let hence w.l.o.g. $v \in V \backslash B$ be such that $d_{v,i} > d_{v,j}$. Now we can choose $t_s = -d_{v,j}c$ and $c = \frac{d_{j,i}}{d_{v,i} - d_{v,j}} > 0$ and obtain for $v, j \in V$

$$d_{j,k} = d_{v,k}c + t_s = (d_{v,k} - d_{v,j})\frac{d_{j,i}}{d_{v,i} - d_{v,j}} \quad \forall\, k \in B,$$

contradicting Definition 19 of a spread-resolving set.

Again, there are graphs for which this bound is sharp, independent of n.

Proposition 26 (Lower Bound 3 is Active) *Let $V = \{1, \ldots, n\}$ and $E = \{\{1, 2\}, \{2, 3\}, \ldots, \{n-1, n\}\}$ for $n \geq 4$. The chain graph has spread dimension 3 and $B = \{1, 2, n\}$ is a basis.*

PROOF. Let $B = \{1, 2, n\}$. First we note that the distance between two nodes $i, j \in V$ is $d_{i,j} = |i - j|$.

Let w.l.o.g. $a < b \in V$ and $\Delta = b - a > 0$. We consider the three equations

$$d_{a,k} = t_s + c\, d_{b,k} \quad \forall\, k \in B$$

from Definition 18 and show that no $t_s, c > 0$ exist which satisfy all of them.

If $a > 1$, we have the distances of a to the basis as $d_{a,1} = a - 1$, $d_{a,2} = a - 2$, $d_{a,n} = n - a$ and distances of b accordingly $d_{b,1} = a - 1 + \Delta$, $d_{b,2} = a - 2 + \Delta$, $d_{b,n} = n - a - \Delta$. While the first two equations result in $t_s = -\Delta$ and $c = 1$, the equation for $k = n$ is incorrect with these values.

If $a = 1$, we have distances $0, 1, n-1$ and $b-1, b-2, n-b$, respectively. Here the first two equations result in $t_s = b-1$ and $c = -1$, but negative c values are not permitted.

Thus, B is a spread-resolving set. With Proposition 25, it is also an spread basis.

Summarizing, the spread dimension can be anything between 3 and n for graphs with n nodes. The examples of chain and star graphs show that it is not the absolute number of edges, but rather the graph topology that impacts the spread dimension. Tailored results for specific graph topologies are interesting, but beyond the scope of this paper.

4 Stochastic case

In this section we consider the more general case with normally distributed random measurement errors ϵ_i. The measurements r_k become random variables, and thus also the estimated parameters t_s, c, and $s \in V$. It may make sense to measure multiple times at a particular node. We start by updating some definitions.

Definition 27 (Stochastic Source Certificate) *For a given $\alpha \in (0, 1)$ we call a node s the* probable source *of the spreading process, if t_s is finite and if a given statistical test passes with an error probability $1 - (1 - \alpha)^{(1/N)}$ for the hypothesis $t_s < t_j$ for all nodes j for which an edge $(v_j, s) \in E$. Here N is the number of edges $(v_j, s) \in E$.*

We start by looking at the source inversion (resolving) problem S3) from Definition 9.

Definition 28 (Source Estimator) *Given a multiset (nodes can be queried multiple times) of nodes $\widehat{S}$ and the corresponding oracle answers $r_{\widehat{S}}$, we define the source estimator similar to* (1) *as*

$$j^* := \arg\min_{j \in V} J^*_{j,\widehat{S}} := \arg\min_{j \in V} \min_{t_s, c \geq 0} J_{j,\widehat{S}}(t_s, c) \tag{4}$$

$$:= \arg\min_{j \in V} \min_{t_s, c \geq 0} \sum_{k \in \widehat{S}} (c\, d_{j,k} + t_s - r_k)^2 \tag{5}$$

as the most likely source for $\widehat{S}$ in a least squares sense.

For fixed source estimate $j \in V$, the solution of the linear regression problem can be derived analytically [11, pages

4–5 for the unconstrained solution] as

$$c(j,\widehat{S}) = \max\left(0, \frac{\sum_{i\in\widehat{S}}(r_i-\bar{r})(d_{j,i}-\bar{d}(j))}{\sum_{i\in\widehat{S}}(d_{j,i}-\bar{d}(j))^2}\right), \quad (6)$$

$$t_s(j,\widehat{S}) = \bar{r} - c(j,\widehat{S})\,\bar{d}(j) \quad (7)$$

with $\bar{r} = \text{mean}(r_{\widehat{S}})$ and $\bar{d}(j) = \text{mean}(d_{j,\widehat{S}})$. This allows to evalute $J^*_{j,\widehat{S}}$ for all $j \in V$ and derive an estimate j^* via enumeration, similar to the deterministic case. If the source can be resolved with a certain probability depends obviously on the choice of the multiset $\widehat{S}$.

Definition 29 (Stochastic Spread-Resolving Set) *Given $a, b \in \mathbb{R}_+$, a source estimate $j^* \in V$, a resolution radius $\gamma > 0$, and values $J^*_{j,\widehat{S}}\ \forall\, j \in V$, we define*

$$B^{j^*}_\gamma = \{ i \in V :\ d_{j*,i} \le \gamma \} \quad (8)$$

and call the multiset $\widehat{S}$ stochastically spread-resolving *(SSR), if an F-Test is successful for a confidence α with*

$$\frac{\left(\min_{j\in V\setminus B^{j^*}_\gamma} J^*_{j,\widehat{S}}\right) - J^*_{j^*,\widehat{S}}}{J^*_{j^*,\widehat{S}}}\ \frac{b}{a} \ge F^{-1}_{a,b}(\alpha) \quad (9)$$

Note that this approach is heuristic, because the statistic is not F-distributed.

Example 30 (Stochastic Source Inversion) *We consider our example graph with oracle queries at $\widehat{S} = \{1,4,6,7,9\}$ resulting in*

$$r_{\widehat{S}} = (1.44327, 0.31493, 3.43784, 5.48041, 4.77700).$$

We can calculate best fit regression lines for all ten nodes:

node	c	t_s	*objective value*
$0,6,7,8,9$	$1e-10$	3.09069	19.09375
1	1.49215	0.40482	3.95344
2	2.12480	−1.15892	1.03461
3	3.39669	−5.06137	5.24874
4	1.65808	0.10614	0.39891
5	3.39669	−1.66468	5.24874

The smallest objective value is obtained for node 4. However, $\widehat{S}$ does not spread-resolve nodes 3 and 5 (e.g., for $t_s = c = 1$ we have $d_{3,k} = t_s + c d_{5,k}$), resulting in not distinguishable optimal solutions (objective, c) with different t_s. For $a = b = 1$, $\alpha = 0.05$ and the ball $B^2_{1.5} = \{1,2,4\}$ the F-test fails with 12.158 and a cutoff value of 161.45. Thus $\widehat{S}$ is not SSR, and 4 is not a probable source, which is accurate as r_B was simulated for $s = 2, c = 2, t_s = -1$, and a standard deviation of 1.

For the experimental design problem S1) in Definition 9 we use A-optimality, i.e., we choose oracle queries that minimize the following function.

Definition 31 (Set Variance) *For a given multiset $\widehat{S}$, variances σ_j, and $\lambda \in [0,1]$ we define the* set variance

$$\Phi(\widehat{S}) := \sum_{j\in V} \lambda \frac{Var[c(j,\widehat{S})]}{\sigma_j^2} + (1-\lambda)\frac{Var[t_s(j,\widehat{S})]}{\sigma_j^2}, \quad (10)$$

calculated using the variances of the parameter estimates (6-7) according to [72, Section 2.4],

$$Var[c(j,\widehat{S})] = \frac{\sigma^2}{\sum_{i\in\widehat{S}}(d_{j,i}-\bar{d}_j)^2}$$

$$Var[t_s(j,\widehat{S})] = \frac{\sigma^2 \sum_{i\in\widehat{S}} d^2_{j,i}}{|\widehat{S}| \sum_{i\in\widehat{V}}(d_{j,i}-\bar{d}_j)^2}$$

with an unknown, but fixed σ^2.

To prove convergence, and also to avoid observed unwanted numerical behavior, we restrict the multiplicities of the multiset $\widehat{S}$. The number of queries per node must not differ by more than 1. This avoids that specific nodes are queried significantly more often than others.

Definition 32 (Feasible Oracle Queries) *Let V be given. A multiset $\widehat{S}$ of V is called* feasible*, if the multiplicities of all $i \in V$ within $\widehat{S}$ do not differ by more than 1. We denote by $V^{\widehat{S}}$ the subset of V containing all nodes that can be added to a feasible $\widehat{S}$ and maintain feasibility.*

With this setup, we can now formulate a source detection algorithm realizing Definition 9.

The goal of Algorithm 1 is to find a probable source j^* with a small number of oracle queries, assuming considerable practical costs (e.g., increased risk of side effects for intracardiac measurements). Concerning the computational complexity per iteration of Algorithm 1, the main calculations happen in Lines 5, 6, and 10. The inner optimization problems can be solved analytically, compare (6-7), with an effort proportional to $|V|$. This is similar to calculating the set variance in (10). The overall effort to evaluate all objective functions $J^*_{j,\widehat{S}}$ and minimizing over $V \setminus B^{j^*}_\gamma$ in Line 8 and over $V^{\widehat{S}}$ in Line 10 is then propotional to $|V|^2$, where clever look-up tables can be applied to increase performance. Note that the distance resolution in Line 7 is calculated by dividing the estimated standard deviation by the estimated slope $c(j^*,\widehat{S})$.

Given the general applicability of Algorithm 1 and the stochasticity of the task, we can not expect that the algo-

Algorithm 1 Stochastic Source Detection

Input: Graph (V, E) with shortest distances d, access to oracle $\mathcal{V}$, parameters a, b, α, λ, variance weights σ_j

Output: Probable source j^*, SOVR set $\widehat{S}$

1: $i_1, i_2 \leftarrow \arg\min_{i_1 \neq i_2 \in V} \Phi(\{i_1, i_2\})$ ▷ See Def. 31
2: $\widehat{S} \leftarrow \{i_1, i_2\}$ ▷ Initialize set $\widehat{S}$
3: **for** i in $3 \dots i_{\max}$ **do**
4: $r_{\widehat{S}} \leftarrow \mathcal{V}(\widehat{S})$ ▷ Update oracle $\mathcal{V}$ query
5: Calculate $c(j, \widehat{S}), t_s(j, \widehat{S}) \; \forall \; j \in V$ ▷ See (6-7)
6: Calculate objectives $J^*_{j,\widehat{S}}$ and j^* ▷ See (4-5)
7: Calculate $\gamma = \frac{J^*_{j^*,\widehat{S}}}{(|\widehat{S}|-2)c(j^*,\widehat{S})}$ ▷ For SOVR test
8: **if** $\widehat{S}$ is SOVR **then** ▷ See (8-9)
9: **break**
10: $i^+ \leftarrow \arg\min_{j \in V^{\widehat{S}}} \Phi(\widehat{S} \cup \{j\})$ ▷ See Defs. (31-32)
11: $\widehat{S} \leftarrow \widehat{S} \cup \{i^+\}$ ▷ Add node to $\widehat{S}$

rithm has a deterministic bound on the number of necessary iterations. However, the well-posedness follows from the following result.

Corollary 33 (Convergence in the limit) *Assume we remove Lines 8–9 in Algorithm 1. Then there is an i_{max} such that the output of Algorithm 1 is $j^* = s$.*

PROOF. In Line 6 of Algorithm 1 we calculate (4-5)

$$j^* = \arg\min_{j \in V} \; J^*_{j,\widehat{S}}.$$

We want to show that $j^* = s$, i.e., that

$$J^*_{j,\widehat{S}} = \min_{t_s, c \geq 0} \sum_{k \in \widehat{S}} (c\, d_{j,k} + t_s - r_k)^2$$

is smallest for $j = s$, if $\widehat{S}$ is large enough. As $\widehat{S}$ is augmented by one node in every iteration in Line 11, this correlates to a longer runtime and a larger $i_{\max}$.

We use Definition 5 and the true model for s for

$$r_k = d_{s,k} c + t_s + \epsilon_k$$

and the analytical solutions (6-7) to obtain

$$\begin{aligned} J^*_{j,\widehat{S}} &= \sum_{k \in \widehat{S}} (c(j, \widehat{S})\, d_{j,k} + t_s(j, \widehat{S}) - r_k)^2 \\ &= \sum_{k \in \widehat{S}} (c(j, \widehat{S})\, d_{j,k} + \bar{r} - c(j, \widehat{S})\; \bar{d}(j) - r_k)^2 \\ &= \sum_{k \in \widehat{S}} (c(j, \widehat{S})\, (d_{j,k} - \bar{d}(j)) + (\bar{r} - r_k))^2 \\ &= \sum_{k \in \widehat{S}} \Big(c(j, \widehat{S})\, (d_{j,k} - \bar{d}(j)) \\ &\qquad - c(d_{s,k} - \bar{d}(s)) - (\epsilon_k - \bar{\epsilon}) \Big)^2 . \end{aligned}$$

We look at $c(j, \widehat{S})$ seperately and use

$$f(x_{\widehat{S}}, j) = \frac{\sum_{i \in \widehat{S}} (x_i - \bar{x})(d_{j,i} - \bar{d}(j))}{\sum_{i \in \widehat{S}} (d_{j,i} - \bar{d}(j))^2}$$

as abbrevation:

$$\begin{aligned} c(j, \widehat{S}) &= \max\Big(0, f(r_{\widehat{S}}, j)\Big) \\ &= \max\Big(0, c\; f(d_{s,\widehat{S}}, j) + f(\epsilon_{\widehat{S}}, j)\Big) \end{aligned}$$

The term $\bar{\epsilon}$ in $f(\epsilon_{\widehat{S}}, j)$ is a Gaussian distribution $\mathcal{N}(0, \sigma^2/|\widehat{S}|)$. As the probability $P(|\bar{\epsilon}| < \gamma), \gamma > 0$ tends towards one there is no influence of this term in the limit. Then $f(\epsilon_{\widehat{S}}, j)$ can be rewritten with notation $g_i = (d_{j,i} - \bar{d}(j)) / \sum_{i \in \widehat{S}} (d_{j,i} - \bar{d}(j))^2$ as $\sum_{i \in \widehat{S}} \epsilon_i g_i \sim \mathcal{N}(0, \hat{\sigma}^2)$ with variance

$$\hat{\sigma}^2 = \sigma^2 \sum_{i \in \widehat{S}} g_i^2 = \sigma^2 / \sum_{i \in \widehat{S}} (d_{j,i} - \bar{d}(j))^2.$$

As also this Gaussian stochastically converges towards 0 (as $\bar{\epsilon}$ above), it has no influence. Inserting the remaining parts of $c(j, \widehat{S})$ back into the objective yields

$$J^*_{j,\widehat{S}} = \sum_{k \in \widehat{S}} (h_{j,k} - \epsilon_k)^2 = \sum_{k \in \widehat{S}} h_{j,k}^2 - 2 h_{j,k} \epsilon_k + \epsilon_k^2.$$

As above, the term $\bar{\epsilon}$ is neglected because of its stochastic convergence to zero and we used

$$h_{j,k} = \max\Big(0, c\; f(d_{s,\widehat{S}}, j)\Big)\; \big(d_{j,k} - \bar{d}(j)\big) - c\,\big(d_{s,k} - \bar{d}(s)\big)$$

The difference between the true source objective and any other objective is in this term. Because $f(d_{s,\widehat{S}}, s) = 1$ we have $h_{s,k} = 0$.

For all other objectives the term $\sum_{k\in\widehat{S}} h_{j,k}^2$ grows at least linear in the size of $\widehat{S}$ because with V as spread-resolving set $\sum_{k\in V} h_{j,k}^2$ is bounded from below by a positive value and we add elements to $\widehat{S}$ in chunks of V.

The term $-2\sum_{k\in\widehat{S}} h_{j,k}\epsilon_k$ is Gaussian $\mathcal{N}(0, \tilde{\sigma}^2)$ with variance $\tilde{\sigma}^2 = 4\sigma^2\sum_{k\in\widehat{S}} h_{j,k}^2$ which grows at most linearly in the size of $\widehat{S}$ because $\sum_{k\in V} h_{j,k}^2$ is bounded from above.

The last term is $\widetilde{\chi}^2$ distributed with $|\widehat{S}|$ degrees of freedom. In the limit this tends to a Gaussian distribution with mean $|\widehat{S}|$ and variance $2|\widehat{S}|$. For the true source objective this is the only existing term.

Substracting the true source objective from any other objective the result is Gaussian $\mathcal{N}(\mu, \hat{\sigma}^2)$ with $\mu = \sum_{k\in\widehat{S}} h_{j,k}^2$ and $\hat{\sigma}^2 = \tilde{\sigma}^2 + 2|\widehat{S}|$. The probability that it is greater than zero tends to one because the mean grows at least linearly and the variance grows at most linearly.

5 Numerical Results

5.1 Implementation

We have implemented Algorithm 1 in `octave 5.2.0` [23]. The code is available with a permissive license on the website `https:\\github.com\TobiasWeber\IMLR`.

The implementation is a set of `octave` functions and scripts that should work in any `octave` installation with the `statistics` package. For Algorithm 1 only core `octave` was used. Data structures for graph representation were taken from the `octave` network toolbox [12], where also simple graph information and manipulation algorithms can be found. However, the algorihm works standalone as we implemented a different shortest path algorithm for efficiency reasons. It is closely related to the fast matrix multiplication shortest path algorithms and more suited to the `octave` programming language than the Dijkstra algorithm in the toolbox. For the numerical random scenarios and errors we use the statistics package of `octave`.

Remark 34 (Treating infinities) *There are two different sources of infinite values.*

For directed graphs that are not strongly connected or graphs that are not connected, some pairs of nodes might have no shortest path between them, or just in one direction. The infinity pattern can be exploited to find the source by a clustering into connected subgraphs. These subgraphs can be used in step S3), and determined in an extra run of S1) by replacing infinity by 1 *and finite values by* 0*.*

Also the variances Var$[c(j, \widehat{S})]$ and Var$[t_s(j, \widehat{S})]$ in Def. 31 may be infinite. As we are minimizing, this is not a problem though, if implemented carefully. If all variances in $V^{\widehat{S}}$ are infinite, we "minimize" by counting the non finite values in the sum over the variances and choose the "solution" with the least infinities (or NaNs).

In the following and if not stated otherwise, we use hyperparameters $\alpha = 0.05$, $a = 1$, and $b = |\widehat{S}| - 4$ (see Def. 29) and $\lambda = 0$ and $\sigma_j = 1 + J^*_{j,\widehat{S}}$ (see Def. 31). Here, the σ_j were chosen to have larger weight on nodes with a smaller objective function value.

5.2 Illustration on Example Graph

We use our example graph with the same spreading process as in Example 30 ($s = 2, c = 2, t_s = -1$) to illustrate the behavior of Algorithm 1. The output for an instance with "average" behavior in terms of iteration count is as follows.

iter	i^+	j^*	$c(j^*, \widehat{S})$	$t_s(j^*, \widehat{S})$	$\frac{J^*_{j^*,\widehat{S}}}{\lvert\widehat{S}\rvert}$	$\min_{j\in V\setminus B^{j^*}_\gamma} \frac{J^*_{j,\widehat{S}}}{\lvert\widehat{S}\rvert}$	α^*
1, 2	0, 5	5	3.34	−0.38	0.00	0.00	1.0000
3	1	2	3.79	−4.62	0.14	0.14	1.0000
4	7	1	2.09	−1.67	0.25	0.32	1.0000
5	2	4	2.97	−3.89	0.37	0.55	0.6137
6	4	2	2.25	−2.19	0.52	0.72	0.4716
7	8	2	2.44	−2.34	0.53	0.72	0.3800
8	3	2	2.48	−2.30	0.53	0.70	0.3255
9	9	2	2.71	−2.52	0.74	0.90	0.3434
10	6	2	2.72	−2.48	0.70	0.85	0.3054
11	2	2	2.56	−2.10	0.73	0.95	0.1905
12	1	2	2.60	−2.24	0.72	1.21	0.0494

The initialization in Line 1 results in $\{i_1, i_2\} = \{0, 5\}$. Until iteration 10 all nodes are selected once. In iterations 11 and 12 a second query at nodes 2 and 1 results in objective function values that are far enough apart such that the heuristic termination criterion (9) is fulfilled. In this instance the feasibility requirement in Def. 32 leads to oracle queries that might not be necessary. The last column depicts the converging α^* value, obtained by evaluating (9) (stopping criterion is that α^* is below 0.05). The parameters $c(j^*, \widehat{S})$ and $t_s(j^*, \widehat{S})$ converge slowly towards the real values and are still inaccurate at termination. The resulting regression line is a good fit for the measurements, though, compare Figure 4.

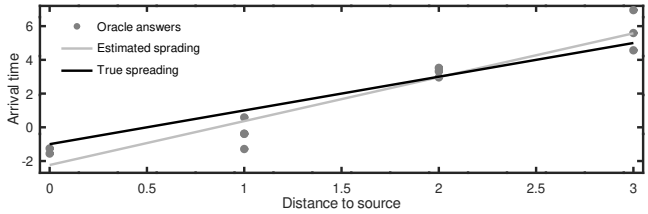


Fig. 4. Regression after iteration 12 at the true (and estimated) source node s. Despite significant outliers, the estimated spreading approximates the true spreading quite well.

In summary, the strict termination criterion and feasibility requirement for oracle queries seem to be robust and avoid early termination, even when by chance a good fit is achieved as in the iterations 2 and 3.

5.3 Problem Instances

We used graph instances that we collected in three sets.

Col. The first set is an operations research library [7] with 30 instances from [28] and 79 DIMACS graph coloring instances [37] from different sources, e.g., [36,46].

Misc. The second set comprises miscellaneous instances. It contains three simple water network instances from the `epanet` software for modeling of water networks [61], eight train networks from the `lintim` software [1], five instances from Mark Newman's webpage [58] (original sources [70,51,75,42]), seven instances from the Weizmann laboratory collection of complex networks [2,53,57,56], and 41 instances from the Pajek dataset [6].

Snap. The last set of instances is a subset of the Stanford large network dataset collection (Snap). It contains 16 graphs derived from an internet topology [47], 9.629 graphs describing user interaction on the music streaming service `deezer` [62], five graphs describing email interactions of members in a large European research institution [48], and ten graphs as friend networks of `Facebook` users [49].

All instances were chosen to have up to 1000 nodes and a possible spreading process application. Some of the instances are directed graphs, some are weighted, others are not connected. If not provided, all edge weights were set to one. For all of the different sources and their different graph formats parsers in `octave` are available. We conducted 100 randomized test runs on each instance (only the 9629 `deezer` graphs were only run once). For each run s was chosen randomly, just as c (exponential distribution with mean 1) and t_s (Gaussian with mean 0 and standard deviation 10). We used $\sigma = \frac{1}{5}c$ as standard deviation for the random error of oracle queries.

5.4 Benchmark Library: Convergence

To evaluate the convergence behavior of Algorithm 1, we assess the quality of the detected source j^* in comparison to the true source s_k of instance k. We use the following normalization and evaluation measure.

Definition 35 (Normalization) *Let Q be the set of all instances (test runs) and $k \in Q$ a specific one.*

Let i_k be the number of iterations until Algorithm 1 terminated for instance k. We then transform all iterations $i \in \{i_{min} = 3, \dots, i_k\}$ to normalized iterations *i_n via $i_n = (i - i_{min})/(i_k - i_{min}) \in [0,1]$, omitting the dependence on k for notational simplicity. Then we define*

$$q(k, i_n) := \frac{\left|\left\{ j \in V : \; J^*_{j,\widehat{S}} \leq J^*_{s,\widehat{S}} \right\}\right|}{|\,V\,|} \qquad (11)$$

as the uniqueness level *of a given true source s for an oracle query set $\widehat{S}$. It depends on the instance $k \in Q$ and on the normalized iteration counter i_n of Algorithm 1. The level $q(k, i_n) \in \left[\frac{1}{|V|}, 1\right]$ is evaluated for the least squares function $J^*_{j,\widehat{S}_{k,i_n}}$. If $q(k, i_n) = \frac{1}{|V|}$ then $j^* = s$.*

First, Table 1 shows the median and mean distances between j^* and s after termination (i.e., $i_n = 1$) of Algorithm 1, indicating its accuracy. There are no significant differences between the test sets, indicating the general applicability of Algorithm 1. As the following results are very similar for all test sets, we present them from now on for Q as the union of the test sets Col, Misc, and Snap.

Table 1
Distance between j^* at $i_n = 1$ and s for different test sets Q. Note that for test set Misc with weighted graphs the distances of j^* to s were divided by the maximum non-infinite shortest path lengths, and the special infinity treatment was applied, see Remark 34. Mostly, Algorithm 1 returned $j^* \approx s$.

Set Q	Median	Mean	Max	# Inf
Col	0.000	0.0183	2.000	0
Misc	0.000	0.0036	0.137	1
Snap	0.000	0.0078	3.000	0

Second, to investigate the efficiency of Algorithm 1, we illustrate in Figure 5 the uniqueness level q as a function of normalized iterations and instances $k \in Q$. Both plots indicate that for the chosen sets $\widehat{S}_{k,i_n}$ the termination criterion is a good choice and that Algorithm 1 is well-posed in the sense that at termination, the true source s_k is detected with high probability (as $q(k, i_n = 1) \approx 0$ for almost all $k \in Q$). From Figure 5 we deduce on the one hand that an earlier termination of Algorithm 1, as seemed plausible from the example in Section 5.2, would often result in j^* that are not minimal with respect to

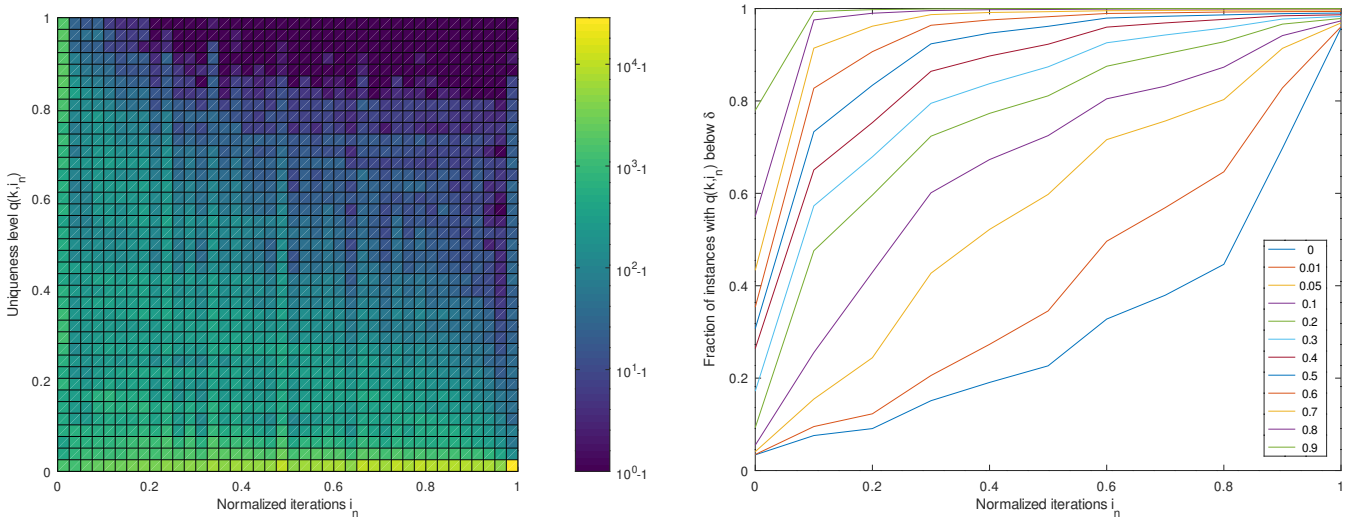


Fig. 5. Using Definition 35, $q(k, i_n)$ is visualized for all instances $k \in Q$. Left: a color gradient indicates for how many instances $k \in Q$ the value $q(k, i_n)$ is in a given box. While for small iteration numbers i_n the values $J^*_{s_k,\widehat{S}_{k,i_n}}$ are almost randomly distributed among all $j \in V$, for large iterations we have $J^*_{s_k,\widehat{S}_{k,i_n}} \leq J^*_{j,\widehat{S}_{k,i_n}}$ for almost all instances k and all $j \in V$, indicating the probable proximity of j^* to the true source s_k of instance k. Right: for different values of δ the lines plot the fraction $|Q_1(i_n)|/|Q|$ with $Q_1(i_n) := \{k \in Q : \ q(k, i_n) \leq \delta + \frac{1}{|V|}\}$. E.g., for $\delta = 0$ the lowest blue line depicts the fraction of instances for which at iteration i_n the source s_k was the unique minimizer of J^*, increasing from approximately 5% to 95%.

the least squares regression. On the other hand, more iterations are not necessary.

5.5 Benchmark Library: Number of Iterations

In this section we have a closer look at how the iteration numbers i_k until Algorithm 1 terminated relate to properties of the graphs. Figure 6 (left) shows them for different graph sizes $n = |V|$. For most $k \in Q$ we have $i_k \leq \frac{1}{2}n$. As the spread dimension is the number of necessary oracle queries (iterations in the online case), this is plausible when looking at the upper bound from Proposition 24 for complete graphs. Note that the spread dimension is not a strict lower bound on i_k due to the advantage that in the online setting we can place oracle queries with knowledge gained in previous iterations.

For small graphs ($n \leq 60$), we could determine the spread dimension by brute force enumeration. The result in Figure 6 (right) confirms the impression that the number of iterations of Algorithm 1 is in many cases below the spread dimension, and only in few cases above. Thus it seems valid to see i_k, at least for the chosen variance of measurement errors, as an approximation of the spread dimension.

This result does not consider other graph properties. An investigation of the topological diameter and of the connectivity (number of edges divided by n) of the graph did not reveal obvious correlations (negative results are not shown here, the color gradients were rather erratic). Known results for the metric dimension β, which is a lower bound for the spread dimension as discussed in Section 4, indicate that the graph topology could have a strong impact (on the lower and not necessarily active bound). E.g., for the diameter d it was shown that

$$n \leq \left(\left\lfloor \frac{2d}{3} \right\rfloor + 1 \right)^{\beta} + \beta \sum_{i=1}^{\lceil d/3 \rceil} (2i-1)^{\beta-1}$$

by [32, Theorem 3.1]. Also the simpler, but less strict inequality

$$n \leq d^{\beta} + \beta$$

from [39] emphasizes the role of the diameter. Not finding a correlation between the diameter d and i_k might indicate that the diameter is not as relevant for the spread dimension as it is for the metric dimension or that i_k differs from the spread dimension for specific graphs. Note also that the spread dimension depends on the edge weights of the graph. Two graphs with the same edge sets can have different spread dimensions, if the edge weights are different. The connection between graph properties on the one hand and spread dimension and iteration numbers on the other hand should be investigated in future research.

6 Conclusion

We formalized the source detection problem in graphs and discussed some of its theoretical properties. We showed that the well-known concept of the metric basis and metric dimension of a graph is related to the specific case of deterministic source detection. We generalized

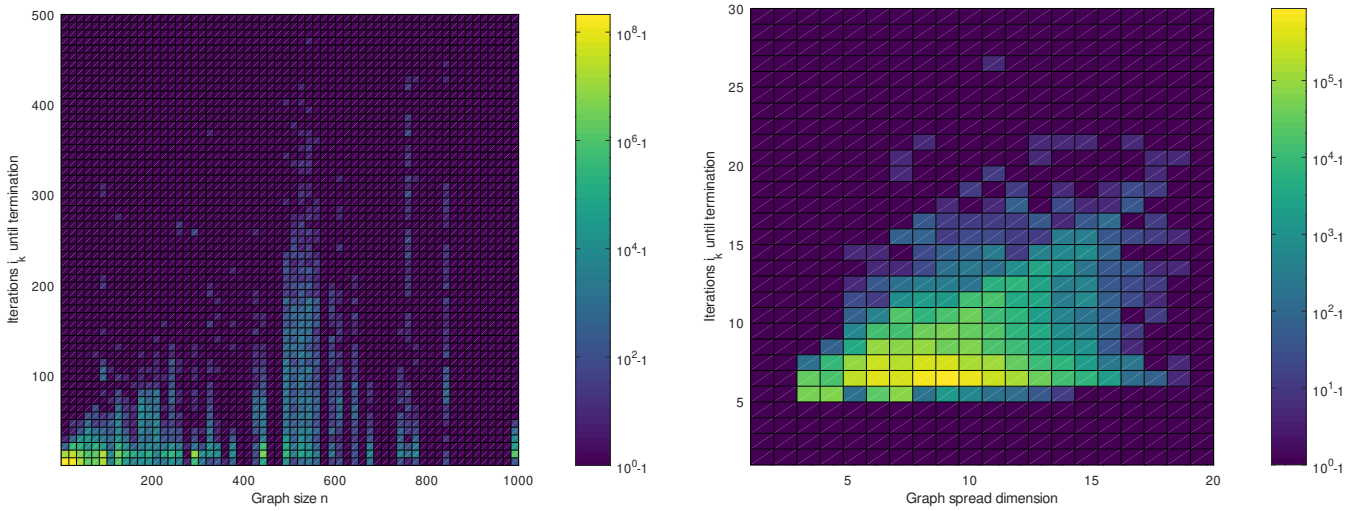


Fig. 6. Color gradients showing how many instances $k \in Q$ needed (on the y-axis) how many iterations until termination of Algorithm 1. Left: Plotted over the graph size $n = |V|$, suggesting a linear relation $i_k \leq c_1 n$ for a constant c_1 and most $k \in Q$. Right: Plotted over the spread dimension β, suggesting a linear relation $i_k \leq c_2 \beta$ for a constant c_2 and most $k \in Q_{\text{small}}$ where $Q_{\text{small}} \subseteq Q$ contains all graphs in Q with $n \leq 60$. Up to this size a brute force enumeration of the spread dimension was computationally feasible.

this concept towards spreading processes by the novel concept of the spread dimension, and towards stochastic measurement errors via stochastic spread-resolving sets. These concepts can and should be further investigated from a graph theoretic point of view.

As a second main contribution, we developed a new solution algorithm to the problem. Algorithm 1 can solve problems belonging to a very general problem class. It consists of a source estimation and of an experimental design part. For the source estimation we proved convergence to the true source under a constraint on the query set $\widehat{S}$. To choose oracle queries for this set we proposed a heuristic solution based on the concept of A-optimality in experimental design and a constraint on multiple measurements. It is unclear how this heuristic can be further improved. A detailed study on relaxing Definition 32, e.g., by allowing the numbers of queries to differ by $k > 1$, seems promising.

A third contribution is a numerical study based on a novel benchmark library for source detection problems. The results indicate well-posedness of Algorithm 1 and a relation between the graph size n and the spread dimension on the one hand, and the number of iterations until termination on the other.

Summarizing, Algorithm 1 proved to be robust over a wide variety of different graphs (weighted/unweighted, directed/undirected, connected/unconnected) and for stochastic disturbances. So far, two of the three main parts (termination and choice of $\widehat{S}$) are only heuristic, whereas the link to quadratic regression is well-posed in the limit. Another future direction of research is hence the application to practical source detection problems with real world data. Also larger problem instances with $n > 1000$ and special cases should be further investigated. For those it will be worthwhile to investigate graph decomposition approaches. Promising concepts from graph theory are the modular decomposition and to a lesser degree the split decomposition of a graph. A very simple and efficient meta algorithm for a graph with more than one (strongly) connected component is to query a node in each connected component once, until the oracle answer is finite. Then the source in this connected component and the rest of the search can continue there, e.g., with Algorithm 1.

In general, the class of source detection on graphs is important from a practical point of view, especially given the omnipresence of networks in modern life. At the same time, this new problem class may stimulate theoretical and algorithmic research.

Acknowledgements

The authors have received funding from the German Research Foundation under GRK 2297 MathCoRe (project No. 314838170), which is gratefully acknowledged.

References

[1] Sebastian Albert, Julius Pätzold, Alexander Schiewe, Philine Schiewe, and Anita Schöbel. Documentation for lintim 2020.02, 2020.

[2] Uri Alon. Collection of complex networks.

[3] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11):118701, 2014.

[4] Andy Baker, Roger Inverarity, Martin Charlton, and Susie Richmond. Detecting river pollution using fluorescence spectrophotometry: case studies from the ouseburn, ne england. *Environmental Pollution*, 124(1):57–70, 2003.

[5] Frank G Ball and Owen D Lyne. Optimal vaccination policies for stochastic epidemics among a population of households. *Mathematical biosciences*, 177:333–354, 2002.

[6] Vladimir Batagelj and Andrej Mrvar. Pajek datasets. `http://vlado.fmf.uni-lj.si/pub/networks/data/`, 2006.

[7] J. E. Beasley. Or-library: Distributing test problems by electronic mail. *The Journal of the Operational Research Society*, 41(11):1069–1072, 1990.

[8] Amir Beck, Petre Stoica, and Jian Li. Exact and approximate solutions of source localization problems. *IEEE Transactions on signal processing*, 56(5):1770–1778, 2008.

[9] Jacob Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *The Journal of the Acoustical Society of America*, 107(1):384–391, 2000.

[10] Anne E Bernhard and Katharine G Field. Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16s ribosomal dna genetic markers from fecal anaerobes. *Appl. Environ. Microbiol.*, 66(4):1587–1594, 2000.

[11] Nicholas H Bingham and John M Fry. *Regression: Linear models in statistics*. Springer Science & Business Media, 2010.

[12] Gergana Bounova. Octave network toolbox, September 2016.

[13] Michael S Brandstein. A pitch-based approach to time-delay estimation of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4–pp. IEEE, 1997.

[14] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *science*, 342(6164):1337–1342, 2013.

[15] José Cáceres, Carmen Hernando, Merce Mora, Ignacio M Pelayo, María L Puertas, Carlos Seara, and David R Wood. On the metric dimension of cartesian products of graphs. *SIAM journal on discrete mathematics*, 21(2):423–441, 2007.

[16] Gary Chartrand, Linda Eroh, Mark A Johnson, and Ortrud R Oellermann. Resolvability in graphs and the metric dimension of a graph. *Discrete Applied Mathematics*, 105(1-3):99–113, 2000.

[17] Gary Chartrand and Ping Zhang. The theory and applications of resolvability in graphs: A survey. 160, 01 2003.

[18] Joe C Chen, Kung Yao, and Ralph E Hudson. Source localization and beamforming. *IEEE Signal Processing Magazine*, 19(2):30–39, 2002.

[19] Edward G Coffman Jr, Zihui Ge, Vishal Misra, and Don Towsley. Network resilience: exploring cascading failures within bgp. In *Proc. 40th Annual Allerton Conference on Communications, Computing and Control*, 2002.

[20] Vittoria Colizza and Alessandro Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Physical review letters*, 99(14):148701, 2007.

[21] Cesar Henrique Comin and Luciano da Fontoura Costa. Identifying the starting point of a spreading process in complex networks. *Physical Review E*, 84(5):056105, 2011.

[22] SD Costanzo, MJ O'donohue, WC Dennison, NR Loneragan, and M Thomas. A new approach for detecting and mapping sewage impacts. *Marine Pollution Bulletin*, 42(2):149–156, 2001.

[23] John W. Eaton, David Bateman, Søren Hauberg, and Rik Wehbring. *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, 2020.

[24] Gustav Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262, 1952.

[25] Demetrios G Eliades and Marios M Polycarpou. Fault isolation and impact evaluation of water distribution network contamination. *IFAC Proceedings Volumes*, 44(1):4827–4832, 2011.

[26] Valerii Fedorov. Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589, 2010.

[27] Vincenzo Fioriti and Marta Chinnici. Predicting the sources of an outbreak with a spectral technique. *arXiv preprint arXiv:1211.2333*, 2012.

[28] Charles Fleurent and Jacques A Ferland. Genetic and hybrid algorithms for graph coloring. *Annals of Operations Research*, 63(3):437–461, 1996.

[29] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 2, pages 1455–1466. IEEE, 2005.

[30] Sepp Hartung and André Nichterlein. On the parameterized and approximation hardness of metric dimension. In *2013 IEEE Conference on Computational Complexity*, pages 266–276. IEEE, 2013.

[31] Mathias Hauptmann, Richard Schmied, and Claus Viehmann. Approximation complexity of metric dimension problem. *Journal of Discrete Algorithms*, 14:214–222, 2012.

[32] Carmen Hernando, Merce Mora, Ignacio M Pelayo, Carlos Seara, and David R Wood. Extremal graph theory for metric dimension and diameter. *Electronic Notes in Discrete Mathematics*, 29:339–343, 2007.

[33] Andrew M Hopkins, CJ Miller, AJ Connolly, Christopher Genovese, Robert C Nichol, and Larry Wasserman. A new source detection algorithm using the false-discovery rate. *The Astronomical Journal*, 123(2):1086, 2002.

[34] Munsif Ali Jatoi, Nidal Kamel, Aamir Saeed Malik, Ibrahima Faye, and Tahamina Begum. A survey of methods used for source localization using eeg signals. *Biomedical Signal Processing and Control*, 11:42–52, 2014.

[35] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, Wanlei Zhou, and Ekram Hossain. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys and Tutorials*, 17(9), 2014.

[36] David S Johnson, Cecilia R Aragon, Lyle A McGeoch, and Catherine Schevon. Optimization by simulated annealing: an experimental evaluation; part ii, graph coloring and number partitioning. *Operations research*, 39(3):378–406, 1991.

[37] David S Johnson and Michael A Trick. *Cliques, coloring, and satisfiability: second DIMACS implementation challenge, October 11-13, 1993*, volume 26. American Mathematical Soc., 1996.

[38] Hiroo Kanamori and Luis Rivera. Source inversion of w phase: speeding up seismic tsunami warning. *Geophysical Journal International*, 175(1):222–238, 2008.

[39] Samir Khuller, Balaji Raghavachari, and Azriel Rosenfeld. Landmarks in graphs. *Discrete Applied Mathematics*, 70(3):217–229, 1996.

[40] Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 272–319, 1959.

[41] Jack Kiefer and Jacob Wolfowitz. Optimum designs in regression problems. *The Annals of Mathematical Statistics*, pages 271–294, 1959.

[42] Donald Ervin Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. AcM Press New York, 1993.

[43] Jozef Kratica, Mirjana Čangalović, and Vera Kovačević-Vujčić. Computing minimal doubly resolving sets of graphs. *Computers & Operations Research*, 36(7):2149–2159, 2009.

[44] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE signal processing magazine*, 13(4):67–94, 1996.

[45] CD Laird, LT Biegler, BG van Bloemen Waanders, and RA Bartlett. Time dependent contamination source determination for municipal water networks using large scale optimization. *Journal of Water Resources Planning and Management*, 2003.

[46] Frank Thomson Leighton. A graph coloring algorithm for large scheduling problems. *Journal of research of the national bureau of standards*, 84(6):489–506, 1979.

[47] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, 2005.

[48] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[49] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[50] Wuqiong Luo, Wee Peng Tay, and Mei Leng. How to identify an infection source with limited observations. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):586–597, 2014.

[51] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[52] Dmitry Malioutov, Müjdat Cetin, and Alan S Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE transactions on signal processing*, 53(8):3010–3022, 2005.

[53] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[54] Malvin S Marlim and Doosun Kang. Identifying contaminant intrusion in water distribution networks under water flow and sensor report time uncertainties. *Water*, 12(11):3179, 2020.

[55] Robert A Melter and Ioan Tomescu. Metric bases in digital geometry. *Computer Vision, Graphics, and Image Processing*, 25(1):113–121, 1984.

[56] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[57] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[58] Mark E.J. Newman. Network data. `http://www-personal.umich.edu/~mejn/netdata/`.

[59] Mark E.J. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):116–128, 2002.

[60] James A Noblet, Diana L Young, Eddy Y Zeng, and Semsi Ensari. Use of fecal steroids to infer the sources of fecal indicator bacteria in the lower santa ana river watershed, california: sewage is unlikely a significant source. *Environmental science & technology*, 38(22):6002–6008, 2004.

[61] Lewis A Rossman et al. Epanet 2: users manual. 2000.

[62] Benedek Rozemberczki, Oliver Kiss, and Rik Sarkar. An api oriented open-source python framework for unsupervised learning on graphs, 2020.

[63] Devavrat Shah and Tauhid Zaman. Detecting sources of computer viruses in networks: Theory and experiment. *SIGMETRICS Perform. Eval. Rev.*, 38(1):203–214, June 2010.

[64] JPS Sidhu, W Ahmed, W Gernjak, R Aryal, D McCarthy, A Palmer, P Kolotelo, and S Toze. Sewage pollution in urban stormwater runoff as evident from the widespread presence of multiple microbial and chemical source tracking markers. *Science of the Total Environment*, 463:488–496, 2013.

[65] SK Singh, M Ordaz, JF Pacheco, and F Courboulex. A simple source inversion scheme for displacement seismograms recorded at short distances. *Journal of seismology*, 4(3):267–284, 2000.

[66] Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

[67] Richard C Tillquist, Rafael M Frongillo, and Manuel E Lladser. Getting the lay of the land in discrete space: A survey of metric dimension and its applications. *arXiv preprint arXiv:2104.07201*, 2021.

[68] Abraham Wald. On the efficient design of statistical investigations. *The Annals of Mathematical Statistics*, 14(2):134–140, 1943.

[69] Hong Wang and Peter Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, pages 187–190. IEEE, 1997.

[70] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[71] Tobias Weber, Hugo A Katus, Sebastian Sager, and Eberhard P Scholz. Novel algorithm for accelerated electroanatomic mapping and prediction of earliest activation of focal cardiac arrhythmias using mathematical optimization. *Heart rhythm*, 14(6):875–882, 2017.

[72] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[73] Kung Yao, Ralph E Hudson, Chris W Reed, Daching Chen, and Flavio Lorenzelli. Blind beamforming on a randomly distributed sensor array system. *IEEE Journal on Selected Areas in Communications*, 16(8):1555–1567, 1998.

[74] Pei-Duo Yu, Chee Wei Tan, and Hung-Lin Fu. Epidemic source detection in contact tracing networks: Epidemic centrality in graphs and message-passing algorithms. *arXiv preprint arXiv:2201.06751*, 2022.

[75] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.