# On the Minimum Differentially Resolving Set Problem for Diffusion Source Inference in Networks

**Chuan Zhou*, Wei-Xue Lu†*, Peng Zhang‡*, Jia Wu‡, Yue Hu*, and Li Guo***
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
‡Centre for Quantum Computation & Intelligent Systems, University of Technology, Sydney, Australia
{zhouchuan, huyue, guoli}@iie.ac.cn, wxlu@amss.ac.cn, {Peng.Zhang, Jia.Wu}@uts.edu.au

**Abstract**

In this paper we theoretically study the minimum Differentially Resolving Set (DRS) problem derived from the classical sensor placement optimization problem in network source locating. A DRS of a graph $G = (V, E)$ is defined as a subset $S \subseteq V$ where any two elements in $V$ can be distinguished by their different *differential characteristic sets* defined on $S$. The minimum DRS problem aims to find a DRS $S$ in the graph $G$ with minimum total weight $\sum_{v \in S} w(v)$. In this paper we establish a group of Integer Linear Programming (ILP) models as the solution. By the weighted set cover theory, we propose an approximation algorithm with the $\Theta(\ln n)$ approximability for the minimum DRS problem on general graphs, where $n$ is the graph size.

## 1 Introduction

In networks, locating diffusion source nodes plays a key role in epidemic control (Shah and Zaman 2011), data provenance estimation (Zhou et al. 2010), virus traceability (Han et al. 2008) and trendsetter identification (Zejnilovic, Gomes, and Sinopoli 2013). Generally, it is often impossible to observe the state of all nodes in a network, and we wish to estimate the diffusion sources from sparsely placed sensors (Pinto, Thiran, and Vetterli 2012). Since it is costly to place sensors in a network, how to select a few sensors for accurate source inference with minimal network search costs is an important question for network management (Zejnilovic, Gomes, and Sinopoli 2013; Seo, Mohapatra, and Abdelzaher 2012).

In this paper we study the problem of selecting a small subset of nodes $S \subseteq V$ in a network $G = (V, E)$ to place sensors for source locating. Under the observation that a sensor placed on a node $v$ can record the 'infected' time when the state of the node $v$ changes by receiving unwanted information (Gomez Rodriguez, Leskovec, and Krause 2010), our goal is to select a subset of nodes $S$ with minimum total cost such that the source can be uniquely located by the differentials of 'infected' times recorded by the sensor set $S$. To this end, we formulate a novel *minimum Differentially Resolving Set (DRS) problem* and connect it to the problem of finding sensors in network source locating.



We focus on the *Computational Complexity* and *Approximability* of the minimal DRS problem, and theoretically propose a group of Integer Linear Programming (ILP) models as the solution and establish a $\Theta(\ln n)$ approximability of this problem on general graphs.

### 1.1 Problem Formulation

We consider a network $G = (V, E)$ as an undirected graph where $|V| = n$ and $|E| = m$. Assume that information is diffused by the shortest path. Let $d(u, v)$ be the shortest distance between nodes $u \in V$ and $v \in V$. If node $u$ is the source and starts a diffusion at an unknown time point $t_0$, then nodes $v$ and $w$ are infected at time points $t_v := t_0 + d(u, v)$ and $t_w := t_0 + d(u, w)$ respectively. The differential time in Eq. (1) cancels time $t_0$ and can be used as a measure to locate diffusion source in the network $G$.

$$\delta(u; v, w) := t_w - t_v = d(u, w) - d(u, v) \tag{1}$$

Let $S \subseteq V$ ($|S| \geq 2$) be the nodes set to place sensors and we set an *anchor node* $v^* \in S$. Assume that node $u$ is the diffusion source, then the *differential characteristic set* for $u$ can be defined as follows,

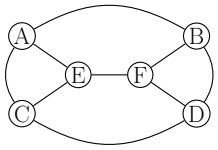
$$\Delta(u; v^*, S) := \Big\{ \big(v, \delta(u; v^*, v)\big) \Big\}_{v \in S \setminus \{v^*\}} \tag{2}$$

Also, we have the following definition,

**Definition 1.** ***(Differentially Resolving Set)*** *A set $S \subseteq V$ with $|S| \geq 2$ is defined as a Differentially Resolving Set (DRS) if there exists an anchor node $v^* \in S$ such that $\Delta(u'; v^*, S) \neq \Delta(u''; v^*, S)$ holds for all different sources $u'$ and $u''$. $v^*$ is called an anchor for $S$, and $V$ is differentially resolved by $\big\{(v^*, v)\big\}_{v \in S \setminus \{v^*\}}$.*

We give a toy example to explain DRS in Fig. 1. Consider the graph $G$ in Fig. 1 (a) and a subset of nodes $S := \{A, B, C\}$ with the anchor node $A$. The differential characteristic sets for nodes $A, B, C, D, E$ and $F$ are $\{(B, 1), (C, 1)\}$, $\{(B, -1), (C, 1)\}$, $\{(B, 1), (C, -1)\}$, $\{(B, -1), (C, -1)\}$, $\{(B, 1), (C, 0)\}$ and $\big\{(B, -1), (C, 0)\big\}$ respectively, which can be reorganized as in Fig. 1 (b). Obviously, any two differential characteristic sets are different. Hence, $S$ is a DRS according to Definition 1.

Intuitively, if a set $S$ is a DRS with respect to an anchor node $v^*$, we have a one-to-one mapping between source




(a) Graph

| Source | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| $t_B - t_A$ | 1 | -1 | 1 | -1 | 1 | -1 |
| $t_C - t_A$ | 1 | 1 | -1 | -1 | 0 | 0 |

(b) Corresponding relation

Figure 1: A toy example of DRS.

$u$ and the differential characteristic set $\Delta(u; v^*, S)$ as in Eq. (3),

$$u \leftrightarrow \Delta(u; v^*, S). \tag{3}$$

This mapping can be used to detect the source uniquely based on the infected times collected by sensors placed on the set $S \subseteq V$ in three steps:

1. Calculate the differential characteristic set $\Delta(u; v^*, S)$ by $\delta(u; v^*, v) = d(u, v) - d(u, v^*)$ for each potential source node $u$;
2. When a diffusion starts from an unknown node, sensor $v \in S$ can record an infected time $t_v$. By calculating the differential time $\delta(*; v^*, v) = t_v - t_{v^*}$ for each $v \in S\backslash\{v^*\}$, we can establish a differential characteristic set using $\Delta(*; v^*, S) := \big\{\big(v, \delta(*; v^*, v)\big)\big\}_{v \in S\backslash\{v^*\}}$.
3. Compare $\Delta(*; v^*, S)$ with all the characteristic sets established in the $1^{st}$ step and locate the source uniquely.

Recall the example in Fig. 1, we place three sensors at $S = \{A, B, C\}$ to record the 'infected' times. The time records are, say, $t_A = 6$, $t_B = 5$, and $t_C = 5$ respectively, i.e. $t_B - t_A = -1$ and $t_C - t_A = -1$, then we can infer that the diffusion source is node $D$ by the one-to-one mapping as shown in Fig. 1 (b).

The above procedure shows that a DRS can uniquely determine a source node. However, placing sensors often incurs a cost. A natural question is, how to find a DRS with minimal cost? Let the cost (or the weight) of selecting node $v$ be $w(v)$ and denote the family of all DRSs by $\mathfrak{D}$, then the optimization problem can be given as follows,

$$S^* = \arg\min_{S \in \mathfrak{D}} \sum_{v \in S} w(v). \tag{4}$$

The above function aims to select a subset of nodes $S \subseteq V$ with minimum placing or monitoring cost and the source can be uniquely located by the 'infected' times of nodes in $S$. In the sequel, we call Eq. (4) as the **minimum DRS problem**.

## 1.2 Related Work

**Resolvability problem**. The differential resolvability problem extends the well-studied resolvability problem, where a subset of nodes $S \subseteq V$ is a resolving set (RS) of $G$ if there exists a one-to-one mapping between node $u$ and its characteristic set $\big\{\big(v, d(u, v)\big)\big\}_{v \in S}$ for all $u$ in $V$. The minimum cardinality of a RS of $G$ is known as the metric dimension $md(G)$ of $G$, which has been extensively studied due to its theoretical importance and diverse applications (see e.g., (Chartrand and Zhang 2003; Epstein, Levin, and Woeginger 2012) and references therein). The problem of finding minimum RS is NP-hard even for planar graphs, split graphs, bipartite graphs and bounded degree graphs (Díaz et al. 2012; Epstein, Levin, and Woeginger 2012). A lot of research efforts have been devoted to obtaining the exact values or upper bounds of the metric dimensions of special graphic classes (Cáceres et al. 2007).

In the case of diffusion source locating problem, we usually lose the ability to tell how long it took for a diffusion to go from source to sensors, and only know the differences in the arrival times at different nodes in sensor set $S$. To cater for this situation, we introduce DRS, a modified version of RS.

**Source locating problem**. Recently, there have been a surge of researches towards the source locating problem. Shah and Zaman (2011) studied this problem under the SI model and developed a rumor centrality estimator. By employing the rumor centrality estimator, multiple sources detecting was investigated in (Luo, Tay, and Leng 2013), and single source with partial observations was considered in (Karamchandani and Franceschetti 2013). The detection rate of rumor centrality estimator under a priori distribution of the source was evaluated in (Dong, Zhang, and Tan 2013). Besides the SI model, source detection under the SIR model has also been studied (Zhu and Ying 2013; 2014), where a sample-path-based estimator for detecting single source was developed. They later proved that the sample-path-based estimator remains effective with sparse observations (Zhu and Ying 2014). The effectiveness of the sample-path-based estimator under the SIS model with partial observations was investigated in (Luo and Tay 2013). In addition, several other source detecting algorithms have also been proposed recently, including the eigenvalue-based estimator (Prakash, Vreeken, and Faloutsos 2012), the dynamic message-passing algorithm (Lokhov et al. 2014), the fast Monte Carlo algorithm (Agaskar and Lu 2013), and the divide-and-conquer approach (Zang et al. 2015).

Compared with the above source locating methods, we focus on the sensor placement problem, which is abstracted from diffusion source inference, from a combinatorial optimization viewpoint. Our goal is to find a set of sensors with minimum monitoring cost that can guarantee a DRS, i.e., the diffusion source can be uniquely located by the differentials of arrival times recorded by such a set of sensors.

## 1.3 Our Contributions

In this paper, we study the minimum DRS problem in terms of computational complexity and algorithmic approximability for efficient source locating in networks. Our contributions are summarized as follows,

1. We define DRS in Definition 1, formulate the minimum DRS problem in Eq. (4), and connect it to the problem of finding sensors in network source locating.

2. We establish two sufficient and necessary conditions for finding the DRS in a network, i.e., in what condition the partially collected information is sufficient to uniquely identify the diffusion sources (Proposition 2 and Proposition 4).

Table 1: Major variables in the paper

| Variables | Descriptions |
|---|---|
| $G = (V, E)$ | graph $G$ with node set $V$ and edge set $E$ |
| $n$ | number of nodes in the graph $G$ |
| $m$ | number of edges in the graph $G$ |
| $v^*$ | anchor node |
| $DRS$ | differentially resolving set |
| $\mathfrak{D}$ | family of all DRSs |
| $d(u, v)$ | length of the shortest path between $u, v \in V$ |
| $\delta(u; v, w)$ | differential time defined in Eq. (1) |
| $A$ | coefficient square matrix defined in Eq. (5) |
| $IP$ | integer programming model (7) - (9) |
| $\mathbf{y}^*$ | one of the optimum solutions to $IP$ model |
| $ILP_{\neg i}$ | integer linear programming model (10) - (12) |
| $\mathfrak{Y}_{\neg i}$ | family of feasible solutions to $ILP_{\neg i}$ model |
| $\mathbf{y}^*_{\neg i}$ | one of the optimum solutions to $ILP_{\neg i}$ model |
| $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ | set cover problem associated with $ILP_{\neg i}$ |
| $\mathfrak{C}^i$ | family of feasible solutions to $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ |
| $\mathcal{C}^{i*}$ | one of optimum solutions to $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ |
| $w, \widetilde{w}$ | weight defined in Eq. (14) and Eq. (18) |
| $f, g$ | mappings defined in Eq. (16) and Eq. (17) |
| $\mathcal{C}^i_{ga}$ | output of Algorithm 1 |

3. We propose a group of Integer Linear Programming (ILP) models to solve the minimum DRS problem (Theorem 1).

4. We develop a $(1 + o(1)) \ln n$-approximation algorithm to solve the minimum DRS problem in time $O(n^4 \cdot \ln n \cdot OPT)$ (Algorithm 1, Eq. (23) and Theorem 5).

The rest of the paper is organized as follows. Section 2 discusses DRS in detail. Section 3 presents a group of integer linear programming (ILP) models to solve the minimum DRS problem. Section 4 presents the approximation algorithm for general graphs. Section 5 concludes the paper. Table 1 outlines the major variables used.

## 2 The DRS Analysis

Before the discussion of the minimal DRS problem in Eq. (4), in this section we first investigate the properties of DRS, which provide a basic for the latter models.

**Proposition 1.** *If a set $S$ is a DRS with an anchor node $v^* \in S$, then any node $v \in S$ can be the anchor for $S$.*

*Proof.* For $v \in S \backslash \{v^*\}$, we want to prove that $v$ is also a anchor for $S$. Namely, for any $u' \neq u''$, we want to search for a node $\bar{v} \in S$ satisfying $\delta(u'; v, \bar{v}) \neq \delta(u''; v, \bar{v})$. Since $v^*$ is an anchor of $S$, for any $u' \neq u''$, by Definition 1 there exists $v' \in S$ such that $\delta(u'; v^*, v') \neq \delta(u''; v^*, v')$. If $\delta(u'; v^*, v) \neq \delta(u''; v^*, v)$, then let $\bar{v} = v^*$. If $\delta(u'; v^*, v) = \delta(u''; v^*, v)$, combined with $\delta(u'; v^*, v') \neq \delta(u''; v^*, v')$, we have $\delta(u'; v, v') \neq \delta(u''; v, v')$, then let $\bar{v} = v'$. □

Proposition 1 reveals that the differentially resolving property does not depend on the choice of anchor node. Henceforth we will not specify the anchor node when we mention the DRS, except for the application scene of diffusion source inference.

**Proposition 2.** *Let $S \subseteq V$, then $S$ is a DRS if and only if, for any $u' \neq u''$ in $V$, there exists $v' \neq v''$ in $S$ such that $\delta(u'; v', v'') \neq \delta(u''; v', v'')$.*

*Proof.* ($\Rightarrow$) It is just a corollary of Definition 1. ($\Leftarrow$) Fix a node $v^* \in S$. Suppose for any $u' \neq u''$ in $V$, there exists $v' \neq v''$ in $S$ such that $\delta(u'; v', v'') \neq \delta(u''; v', v'')$. We claim that the two equations $\delta(u'; v^*, v'') = \delta(u''; v^*, v'')$ and $\delta(u'; v', v^*) = \delta(u''; v', v^*)$ can not hold simultaneously. Otherwise, we will have $\delta(u'; v', v'') = \delta(u''; v', v'')$, which contradicts the assumption. In other words, either $\delta(u'; v^*, v'') \neq \delta(u''; v^*, v'')$ or $\delta(u'; v', v^*) \neq \delta(u''; v', v^*)$ holds, which means $\Delta(u'; S, v^*) \neq \Delta(u''; S, v^*)$. Hence $S$ is a DRS by Definition 1. □

**Proposition 3.** *Let $S \subseteq S' \subseteq V$. If $S$ is a DRS, then $S'$ is also a DRS.*

*Proof.* It can be easily reached by the necessary and sufficient condition in Proposition 2. □

Proposition 2 establishes a sufficient and necessary condition for finding the DRS in a network. Proposition 3 presents a conduction of differentially resolving property.

## 3 Programming Models

We consider a network $G = (V, E)$ as an undirected graph where $|V| = n$ and $|E| = m$. For the sake of simplicity, we define $V := \{1, 2, ..., n\}$. Let $d(u, v)$ be the length of the shortest $u - v$ path for all $u, v \in V$. The coefficient square matrix $A$ of $n(n-1)/2$-order is defined as follows,

$$A_{(u,v)(i,j)} = \begin{cases} 1, & \text{if } \delta(u; i, j) \neq \delta(v; i, j) \\ 0, & \text{if } \delta(u; i, j) = \delta(v; i, j) \end{cases} \tag{5}$$

where $1 \leq u < v \leq n$, $1 \leq i < j \leq n$. Let $y_i$ in Eq. (6) denote whether a node $i$ belongs to a DRS $S$ for $1 \leq i \leq n$,

$$y_i = \begin{cases} 1, & \text{if } i \in S \\ 0, & \text{if } i \notin S \end{cases} \tag{6}$$

i.e., $y_i = 1_S(i)$. *From now on the set $S$ is also viewed as a $\{0, 1\}$-valued vector $\mathbf{y} = (y_i)_{i=1}^n$ of $n$-dimension.* The **integer programming ($IP$) model** for the minimum DRS problem in Eq. (4) can be formulated as follows,

$$(IP): \quad \min \sum_{k=1}^{n} w(k) \cdot y_k \tag{7}$$

subject to:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} A_{(u,v)(i,j)} \cdot y_i \cdot y_j \geq 1, \ 1 \leq u < v \leq n \tag{8}$$

$$y_k \in \{0, 1\}, \ 1 \leq k \leq n \tag{9}$$

The following proposition shows that a feasible solution of (8) and (9) defines a DRS $S$ of $G$, and vice versa.

**Proposition 4.** *A set $S$ is a DRS of $G$ if and only if constraints (8) and (9) are satisfied. In other words, the IP model is equivalent to the minimal DRS problem.*

*Proof.* ($\Rightarrow$) Suppose that $S$ is a DRS, then for any $u \neq v$ there exist $i \neq j$ in $S$ such that $\delta(u;i,j) \neq \delta(v;i,j)$ by Proposition 2. Without loss of generality we may assume that $u < v$ and $i < j$. It follows that $A_{(u,v),(i,j)} = 1$ and consequently constraints (8) and (9) are satisfied.

($\Leftarrow$) According to (6) set $S = \{i \in V | y_i = 1\}$. If constraints (8) are satisfied, then for each $1 \leq u < v \leq n$, there exist $i < j$ in $V$, such that $A_{(u,v),(i,j)} \cdot y_i \cdot y_j = 1$, which implies $i, j \in S$ and $A_{(u,v),(i,j)} = 1$ (i.e. $\delta(u;i,j) \neq \delta(v;i,j)$). By Proposition 2, it follows that the set $S$ is a DRS of $G$. □

Note that IP model $(7) - (9)$ has $n$ variables and $n^2/2 - n/2$ non-linear constraints. Since non-linear optimization problem is very hard to handle, the IP model $(7) - (9)$ has serious limitations on solving the minimal DRS problem. To address the issue, we break up the IP model into a series of **integer linear programming ($ILP_{\neg *}$) models**. The core idea is to make some $y_i$ fixed (equal to 1) and optimize other $\{y_k\}_{k \neq i}$. Specifically, for each $i \in V$, we consider the following formulation.

$$(ILP_{\neg i}): \quad \min \sum_{k=1}^{n} w(k) \cdot y_k \tag{10}$$

subject to:

$$\sum_{k=1}^{n} A_{(u,v)(i,k)} \cdot y_k \geq 1,\ 1 \leq u < v \leq n \tag{11}$$

$$y_i = 1,\ \ y_k \in \{0,1\},\ k \neq i \tag{12}$$

where $A_{(u,v)(i,i)} := 0$ and $A_{(u,v)(i,k)} := A_{(u,v)(k,i)}$ in the case of $i > k$.

**Proposition 5.** *If constraints* (11) *and* (12) *are satisfied, then the set $S = \{k \in V | y_k = 1\}$ is a DRS of $G$.*

*Proof.* If constraints (11) and (12) are satisfied, then for each $1 \leq u < v \leq n$, there exist $k \neq i$ in $V$, such that $A_{(u,v),(i,k)} \cdot y_k = 1$, which implies $i, k \in S$ and $A_{(u,v),(i,k)} = 1$ (i.e. $\delta(u;i,k) \neq \delta(v;i,k)$). By Proposition 2, it follows that the set $S$ is a DRS of $G$. □

Now we have established a series of linear models $\{ILP_{\neg i}\}_{i=1}^{n}$ drawn from non-linear model $IP$. Such a transformation leads to a fundamental question that, what is the relationship between them?

**Theorem 1.** *Let $\mathbf{y}^*_{\neg i}$ denote one of the optimum solutions to $ILP_{\neg i}$ model* (10) $-$ (12)*, then we have that*

$$\boldsymbol{y}^* := \arg \min_{\boldsymbol{y} \in \{\mathbf{y}^*_{\neg i}\}_{i=1}^{n}} \sum_{k=1}^{n} w(k) \cdot y_k \tag{13}$$

*is one of the optimum solutions to $IP$ model* $(7) - (9)$*. In Eq. (13) the bold letter $\mathbf{y} = (y_k)_{k=1}^{n}$ stands for a $\{0,1\}$-valued vector of $n$-dimension.*

*Proof.* For the sake of simplicity, we define

$$w(\mathbf{y}) := \sum_{k=1}^{n} w(k) \cdot y_k. \tag{14}$$

From Proposition 5, each $\mathbf{y}^*_{\neg i}$ corresponds to a DRS, and so do $\mathbf{y}^*$. Hence it suffices to prove that

$$w(\mathbf{y}^*) \leq w(\mathbf{y})$$

holds for any $\mathbf{y}$ satisfying constraints (8) and (9). Assume $y_{i_0} = 1$, then $\mathbf{y}$ satisfies the constraints (11) and (12) with $i$ replaced by $i_0$, i.e., $\mathbf{y}$ is a feasible solution of $(ILP_{\neg i_0})$. Therefore, we have

$$w(\mathbf{y}^*) \leq w(\mathbf{y}^*_{\neg i_0}) \leq w(\mathbf{y})$$

where the first '$\leq$' is from Eq. (13) and the second '$\leq$' stems from that $\mathbf{y}$ is a feasible solution of $(ILP_{\neg i_0})$. □

So far, we have proposed a series of linear models $\{ILP_{\neg i}\}_{i=1}^{n}$ to solve the minimal DRS problem. Specifically, we solve linear models $\{ILP_{\neg i}\}_{i=1}^{n}$ separately, get the optimum solution $\mathbf{y}^*_{\neg i}$ to each model, and select the one with minimal weight, i.e., $\mathbf{y}^* := \arg\min_{\mathbf{y} \in \{\mathbf{y}^*_{\neg i}\}_{i=1}^{n}} w(\mathbf{y})$, as the solution to minimal DRS problem. Theorem 1 explains that $\mathbf{y}^*$ is optimum.

Since each $ILP_{\neg *}$ model has $n - 1$ variables and $n^2/2 - n/2$ linear constraints, the mathematical programming model can be solved to optimality only for low dimension problems.

# 4 An Approximation Algorithm

In this section, we propose an algorithm for the minimal DRS problem in general graphs that can achieve an approximation ratio of $(1 + o(1)) \ln n$. We first transform each $ILP_{\neg *}$ model to a weighted set cover problem and then use the greedy algorithm to solve the weighted set cover problem within logarithmic approximation.

## 4.1 Transformation to weighted set cover problems

Generally speaking, a weighted set cover problem consists of a universe set $U$, a family of sets $\mathcal{S} = \{S_1, \ldots, S_m\}$ satisfying $U \subseteq \cup\mathcal{S} := \cup_{k=1}^{m} S_k$, and a weight $w(S_i)$ assigned to each set $S_i$. The goal is to find a subfamily of sets $\mathcal{C} \subseteq \mathcal{S}$ such that $U \subseteq \cup\mathcal{C}$ so as to minimize the total weight of the sets in $\mathcal{C}$ (Cygan, Kowalik, and Wykurz 2009). We denote a weighted set cover problem by an ordered triple $(U, \mathcal{S}, w)$ hereafter.

Back to the $ILP_{\neg i}$ model, we establish a new weighted set cover problem equivalent to it. Specifically we construct the universe set $\widetilde{U}$ as follows

$$\widetilde{U} := \big\{(u,v) \big| 1 \leq u < v \leq n\big\}$$

Define a family of sets $\mathcal{S}^i = \{S_1^i, \ldots, S_n^i\}$ like this

$$S_k^i := \big\{(u,v) \big| A_{(u,v)(i,k)} = 1\big\}$$

for each $k \in \{1, \ldots, n\}$. Obviously, we have $S_i^i = \emptyset$ here. We assign each set $S_k^i$ a weight $w(k)$, i.e.,

$$\widetilde{w}(S_k^i) := w(k).$$

**Proposition 6.** *For the constructed universe set $\widetilde{U}$ and family of sets $\mathcal{S}^i = \{S_1^i, \ldots, S_n^i\}$, we have $\widetilde{U} = \cup_{k=1}^{n} S_k^i$ holds, i.e., the ordered triple $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ defines a weighted set cover problem.*

*Proof.* It suffices to prove that, for any $(u,v) \in \widetilde{U}$, there exists some $k \in \{1,\ldots,n\}$ such that $(u,v) \in S^i_k$, i.e. $A_{(u,v)(i,k)} = 1$. First, we can easily verify that

$$\delta(u;u,v) \neq \delta(v;u,v) \tag{15}$$

Second, we claim that the two equations $\delta(u;i,u) = \delta(v;i,u)$ and $\delta(u;i,v) = \delta(v;i,v)$ can not hold simultaneously. Otherwise, we will have $\delta(u;u,v) = \delta(v;u,v)$, which contradicts Eq. (15). In other words, either $\delta(u;i,u) \neq \delta(v;i,u)$ or $\delta(u;i,v) \neq \delta(v;i,v)$ holds, which means that there exists such a $k$ satisfying $A_{(u,v)(i,k)} = 1$. □

**Proposition 7.** *There exists a one-to-one correspondence between the family $\mathfrak{Y}_{\neg i}$ of all feasible solutions to the integer linear programming $(ILP_{\neg i})$ model and the family $\mathfrak{C}^i$ of all feasible solutions to the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem. Specifically there exist two mappings $f : \mathfrak{Y}_{\neg i} \to \mathfrak{C}^i$ and $g : \mathfrak{C}^i \to \mathfrak{Y}_{\neg i}$ such that $f \circ g$ and $g \circ f$ are both identity mappings.*

*Proof.* We can define the mappings $f$ and $g$ as follow

$$f(\mathbf{y}) := \{S^i_k\}_{k\in\{j|y_j=1\}\setminus\{i\}} \tag{16}$$

and

$$g(\mathcal{C}^i) := \big(1_{\mathcal{C}^i\cup\{S^i_i\}}(S^i_k)\big)^n_{k=1} \tag{17}$$

for each $\mathbf{y} \in \mathfrak{Y}_{\neg i}$ and $\mathcal{C}^i \in \mathfrak{C}^i$. It is easy to verify that both $f$ and $g$ are well defined, i.e., $f(\mathbf{y}) \in \mathfrak{C}^i$ for each $\mathbf{y} \in \mathfrak{Y}_{\neg i}$ and $g(\mathcal{C}^i) \in \mathfrak{Y}_{\neg i}$ for each $\mathcal{C}^i \in \mathfrak{C}^i$. From the definitions, it is trivial that $f \circ g$ and $g \circ f$ are both identity mappings. □

**Theorem 2.** *Assume a subfamily $\mathcal{C}^{i*} \subseteq \mathcal{S}^i$ is one of optimum solutions to the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem, then $g(\mathcal{C}^{i*})$ is one of optimum solutions to $ILP_{\neg i}$ model.*

*Conversely, assume $\mathbf{y}^*_{\neg i}$ is one of optimum solutions to $ILP_{\neg i}$ model, then $f(\mathbf{y}^*_{\neg i})$ is one of optimum solutions to the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem.*

*Proof.* For the sake of simplicity, we denote

$$\widetilde{w}(\mathcal{C}^i) := \sum_{S^i_k\in\mathcal{C}^i} \widetilde{w}(S^i_k) \tag{18}$$

which satisfies that

$$\widetilde{w}(\mathcal{C}^i) = w\big(g(\mathcal{C}^i)\big) - w(i) \tag{19}$$

for each $\mathcal{C}^i \in \mathfrak{C}^i$ and

$$\widetilde{w}\big(f(\mathbf{y})\big) = w(\mathbf{y}) - w(i) \tag{20}$$

for each $\mathbf{y} \in \mathfrak{Y}_{\neg i}$. If $\mathcal{C}^{i*}$ is one of optimum solutions to the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem, for any $\mathbf{y} \in \mathfrak{Y}_{\neg i}$, we have

$$\begin{aligned} w(g(\mathcal{C}^{i*})) &= \widetilde{w}(\mathcal{C}^{i*}) + w(i) \\ &\leq \widetilde{w}(f(\mathbf{y})) + w(i) \\ &= w(\mathbf{y}). \end{aligned}$$

Hence $g(\mathcal{C}^{i*})$ is one of optimum solutions to $ILP_{\neg i}$ model.

Conversely, if $\mathbf{y}^*_{\neg i}$ is one of optimum solutions to $ILP_{\neg i}$ model, for any $\mathcal{C}^i \in \mathfrak{C}^i$, we have

$$\begin{aligned} \widetilde{w}(f(\mathbf{y}^*_{\neg i})) &= w(\mathbf{y}^*_{\neg i}) - w(i) \\ &\leq w(g(\mathcal{C}^i)) - w(i) \\ &= \widetilde{w}(\mathcal{C}^i). \end{aligned}$$

Hence $f(\mathbf{y}^*_{\neg i})$ is one of optimum solutions to the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem. □

Proposition 7 and Theorem 2 tell us that we can approximate the $ILP_{\neg i}$ model from the view of the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem.

## 4.2 Approximation of the $ILP_{\neg i}$ model

For a large network, any straightforward method for exactly solving the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem suffers from combinatorial explosion. Therefore, we consider approximately solving the problem with greedy algorithm, which is an efficient method to find a good approximate solution.

We first define the cover range function $\mathbb{C}$ as follows

$$\mathbb{C}(\mathcal{C}^i) := \bigcup_{S^i_\cdot\in\mathcal{C}^i} S^i_\cdot$$

for each subfamily $\mathcal{C}^i \subseteq \mathcal{S}^i$. Let $|\mathbb{C}(\mathcal{C}^i)|$ denote the number of elements in $\mathbb{C}(\mathcal{C}^i)$. Define the marginal increment $\Delta$ of the cover range function $\mathbb{C}$ as follows

$$\Delta(\mathcal{C}^i, S^i_\cdot) := \Big|\mathbb{C}\big(\mathcal{C}^i \cup \{S^i_\cdot\}\big)\Big| - \Big|\mathbb{C}\big(\mathcal{C}^i\big)\Big| \tag{21}$$

Now we present the greedy algorithm (Algorithm 1) for the weighted set cover $(\widetilde{U},\mathcal{S}^i,\widetilde{w})$ problem. This algorithm iteratively selects a new set $S^i_*$ that maximizes the weighted incremental change of $|\mathbb{C}(\cdot)|$, to be included into the cover $\mathcal{C}^i$, until the universe set $\widetilde{U}$ is covered.

**Algorithm 1:** Finding a minimum weighted cover $\mathcal{C}^i_{ga}$

1: initial $\mathcal{C}^i_{ga} \leftarrow \varnothing$
2: **while** $\mathbb{C}(\mathcal{C}^i_{ga}) \neq \widetilde{U}$ **do**
3: $S^i_* \leftarrow \arg\max_{S^i_\cdot\in\mathcal{S}^i\setminus\mathcal{C}^i_{ga}} \frac{\Delta(\mathcal{C}^i_{ga},S^i_\cdot)}{\widetilde{w}(S^i_\cdot)}$
4: $\mathcal{C}^i_{ga} \leftarrow \mathcal{C}^i_{ga} \cup \{S^i_*\}$
5: **end while**
6: output $\mathcal{C}^i_{ga}$

**Proposition 8.** *The marginal increment $\Delta$ is submodular in the sense that*

$$\Delta(\mathcal{C}^i_1, S^i_\cdot) \geq \Delta(\mathcal{C}^i_2, S^i_\cdot) \tag{22}$$

*for any $\mathcal{C}^i_1 \subseteq \mathcal{C}^i_2$ and $S^i_\cdot \in \mathcal{S}^i\setminus\mathcal{C}^i_2$.*

*Proof.* By the definition in Eq. (21), it is trivial to verify that

$$\Delta(\mathcal{C}^i, S^i_\cdot) = \Big|S^i_\cdot \Big\backslash \mathbb{C}(\mathcal{C}^i)\Big|$$

Based on this observation, it follows that

$$\left|S^i_{\cdot}\Big\backslash \mathbb{C}(\mathcal{C}^i_1)\right| \geq \left|S^i_{\cdot}\Big\backslash \mathbb{C}(\mathcal{C}^i_2)\right|$$

for any $\mathcal{C}^i_1 \subseteq \mathcal{C}^i_2$ and $S^i_{\cdot} \in \mathcal{S}^i \backslash \mathcal{C}^i_2$, since we have $\mathbb{C}(\mathcal{C}^i_1) \subseteq \mathbb{C}(\mathcal{C}^i_2)$. Hence Eq. (22) is reached and $\Delta$ is submodular. □

**Theorem 3.** *The time complexity of Algorithm 1 is $O(n^3 \cdot \ln n \cdot OPT)$ for the weighted set cover $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ problem with an approximation ratio of $(1 + o(1)) \ln n$, where $OPT$ is the size of the optimum solution to $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$.*

*Proof.* By the submodularity of $\Delta$, the classical weighted set cover theory (Cygan, Kowalik, and Wykurz 2009) ensures the result on the approximation ratio of $(1+o(1)) \ln n$. To implement Algorithm 1, we first design a table (denoted as TABLE I) to store the distance values $\{d(u,v)\}_{u,v \in V}$ between each pair of nodes in $G$, which takes time of $O(n^2)$. Based on TABLE I, we design another table $\{A_{(u,v)(i,k)}\}$ with $i$ fixed (denoted as TABLE II), which also takes time of $O(n^3)$. By using TABLE II, it needs time $O(n^3)$ to select each $S^i_*$ in the third row of Algorithm 1. Note that the whole loop takes at most $O(\ln n \cdot OPT)$ times in the worst case. Overall, the algorithm runs in $O(n^3 \cdot \ln n)$ time. □

**Theorem 4.** ***(Approximation of the $ILP_{\neg i}$ model)*** *Denote the output of the greedy algorithm (Algorithm 1) as $\mathcal{C}^i_{ga}$, then $g(\mathcal{C}^i_{ga})$ can approximate the $ILP_{\neg i}$ model with a ratio of $(1 + o(1)) \ln n$.*

*Proof.* From Proposition 7, we know that $g(\mathcal{C}^i_{ga})$ is a feasible solution to the $(ILP_{\neg i})$ model. Let $\mathbf{y}^*$ be any optimum solution to the $(ILP_{\neg i})$ model. By Theorem 2, $f(\mathbf{y}^*)$ is one of optimum solutions to the weighted set cover $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ problem. Hence, we have

$$\begin{aligned} w\big(g(\mathcal{C}^i_{ga})\big) &= \widetilde{w}(\mathcal{C}^i_{ga}) + w(i) \\ &\leq (1 + o(1)) \ln n \cdot \widetilde{w}(f(\mathbf{y}^*)) + w(i) \\ &\leq (1 + o(1)) \ln n \cdot \big(\widetilde{w}(f(\mathbf{y}^*)) + w(i)\big) \\ &= (1 + o(1)) \ln n \cdot w\big(g \circ f(\mathbf{y}^*)\big) \\ &= (1 + o(1)) \ln n \cdot w\big(\mathbf{y}^*\big) \end{aligned}$$

where the first two '='s are from Eq. (19), the third '=' comes from Proposition 7 that $g \circ f$ is an identity mapping, the 1st '$\leq$' stems from Theorem 3, and the second '$\leq$' is from simple scale method. □

### 4.3 Approximation of the minimal DRS problem

Given the integer linear programming $(ILP_{\neg i})$ model, we can first employ Algorithm 1 to output an approximate solution $\mathcal{C}^i_{ga}$ to the weighted set cover $(\widetilde{U}, \mathcal{S}^i, \widetilde{w})$ problem. By Theorem 4, we know that $g(\mathcal{C}^i_{ga})$ can approximate the $ILP_{\neg i}$ model with a ratio $(1 + o(1)) \ln n$.

Similarly, we can run Algorithm 1 for each $i \in \{1, \ldots, n\}$ to obtain $n$ approximate solution $\{\mathcal{C}^i_{ga}\}^n_{i=1}$ to different weighted set cover problems. After the $g$-transformation in Eq. (17), we can obtain $n$ approximate solutions $\{g(\mathcal{C}^i_{ga})\}^n_{i=1}$ to the $n$ integer linear programming $\{ILP_{\neg i}\}^n_{i=1}$ models respectively.

Inspired by Theorem 1, we select the one, say $g(\mathcal{C}^{i_0}_{ga})$, which has the minimum weight, i.e.,

$$w(g(\mathcal{C}^{i_0}_{ga})) = \min\Big\{w\big(g(\mathcal{C}^i_{ga})\big) : i = 1, \ldots, n\Big\} \qquad (23)$$

as the approximate solution to the IP model $(7) - (9)$. In other words, we may select $g(\mathcal{C}^{i_0}_{ga}) \in \{0,1\}^n$ as the approximate solution to the minimal DRS problem by Proposition 4. According to the definition of $g(\mathcal{C}^{i_0}_{ga})$, we obtain Theorem 5 as follows.

**Theorem 5.** *The minimal DRS problem can be approximated in $O(n^4 \cdot \ln n \cdot OPT)$ time within an accuracy ratio of $(1 + o(1)) \ln n$.*

*Proof.* **Proof.** According to Theorem 3 and Eq. (23), the time complexity for getting $g(\mathcal{C}^{i_0}_{ga})$ is $O(n^4 \cdot \ln n \cdot OPT)$. Let $\mathbf{y}^*$ be an optimal solution to the minimal DRS problem in Eq. (4). Based on Theorem 1, there exists some $j_0 \in \{1, \ldots, n\}$ such that $w(\mathbf{y}^*_{\neg i_0}) = w(\mathbf{y}^*)$, where $\mathbf{y}^*_{\neg j_0}$ is one of optimum solutions to the $(ILP_{\neg j_0})$ model. Based on Theorem 4, we have

$$\begin{aligned} w(g(\mathcal{C}^{i_0}_{ga})) &\leq w(g(\mathcal{C}^{j_0}_{ga})) \\ &\leq (1 + o(1)) \ln n \cdot w(\mathbf{y}^*_{\neg j_0}) \\ &= (1 + o(1)) \ln n \cdot w(\mathbf{y}^*) \end{aligned}$$

where the first '$\leq$' is from Eq. (23), the second '$\leq$' comes from Theorem 4, and the last '=' stems from the choice of $\mathbf{y}^*_{\neg i_0}$ above. Hence Theorem 5 holds. □

## 5 Conclusions

In this paper we theoretically studied the minimum differentially resolving set (DRS) problem derived from the sensor placement optimization for diffusion source inference in networks. We presented a group of integer linear programming (ILP) models as the solution. We presented an algorithm with time complexity of $O(n^4 \cdot \ln n \cdot OPT)$ for general minimal DRS problem within an accuracy ratio of $(1+o(1)) \ln n$.

In the future, we will study the minimal DRS problem under stochastic diffusion models (Bailey and others 1975), develop more efficient and effective heuristics, and conduct experiments on real-world social network data to evaluate the performance of the proposed methods.

## Acknowledgments

We are highly grateful to the anonymous reviewers for their valuable suggestions and comments. This work was supported by the NSFC (No. 61502479, 61370025, and 61272427), 973 project (No. 2013CB329605), Xinjiang Uygur Autonomous Region Science and Technology Project (No. 201230123), Strategic Leading Science and Technology Projects of CAS (No.XDA06030200), and Australia ARC Discovery Project (DP140102206).

## References


Agaskar, A., and Lu, Y. M. 2013. A fast monte carlo algorithm for source localization on graphs. In *SPIE Optical Engineering+ Applications*, 88581N–88581N. International Society for Optics and Photonics.

Bailey, N. T., et al. 1975. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Cáceres, J.; Hernando, C.; Mora, M.; Pelayo, I. M.; Puertas, M. L.; Seara, C.; and Wood, D. R. 2007. On the metric dimension of cartesian products of graphs. *SIAM Journal on Discrete Mathematics* 21(2):423–441.

Chartrand, G., and Zhang, P. 2003. The theory and applications of resolvability in graphs. *Congressus Numerantium* 47–68.

Cygan, M.; Kowalik, Ł.; and Wykurz, M. 2009. Exponential-time approximation of weighted set cover. *Information Processing Letters* 109(16):957–961.

Díaz, J.; Pottonen, O.; Serna, M.; and Van Leeuwen, E. J. 2012. On the complexity of metric dimension. In *Algorithms–ESA 2012*. Springer. 419–430.

Dong, W.; Zhang, W.; and Tan, C. W. 2013. Rooting out the rumor culprit from suspects. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2671–2675. IEEE.

Epstein, L.; Levin, A.; and Woeginger, G. J. 2012. The (weighted) metric dimension of graphs: hard and easy cases. In *Graph-Theoretic Concepts in Computer Science*, 114–125. Springer.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1019–1028. ACM.

Han, L.; Han, S.; Deng, Q.; Yu, J.; and He, Y. 2008. Source tracing and pursuing of network virus. In *Computer and Information Technology, IEEE 8th International Conference on*, 230–235. IEEE.

Karamchandani, N., and Franceschetti, M. 2013. Rumor source detection under probabilistic sampling. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2184–2188. IEEE.

Lokhov, A. Y.; Mézard, M.; Ohta, H.; and Zdeborová, L. 2014. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E* 90(1):012801.

Luo, W., and Tay, W. 2013. Estimating infection sources in a network with incomplete observations,. In *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP), Austin, TX*, 301–304.

Luo, W.; Tay, W. P.; and Leng, M. 2013. Identifying infection sources and regions in large networks. *Signal Processing, IEEE Transactions on* 61(11):2850–2865.

Pinto, P. C.; Thiran, P.; and Vetterli, M. 2012. Locating the source of diffusion in large-scale networks. *Physical review letters* 109(6):068702.

Prakash, B. A.; Vreeken, J.; and Faloutsos, C. 2012. Spotting culprits in epidemics: How many and which ones? In *ICDM*, volume 12, 11–20.

Seo, E.; Mohapatra, P.; and Abdelzaher, T. 2012. Identifying rumors and their sources in social networks. In *SPIE Defense, Security, and Sensing*, 83891I–83891I. International Society for Optics and Photonics.

Shah, D., and Zaman, T. 2011. Rumors in a network: Who's the culprit? *Information Theory, IEEE Transactions on* 57(8):5163–5181.

Zang, W.; Zhang, P.; Zhou, C.; and Guo, L. 2015. Locating multiple sources in social networks under the sir model: A divide-and-conquer approach. *Journal of Computational Science* 10:278–287.

Zejnilovic, S.; Gomes, J. P.; and Sinopoli, B. 2013. Network observability and localization of the source of diffusion based on a subset of nodes. In *Allerton*, 847–852.

Zhou, W.; Sherr, M.; Tao, T.; Li, X.; Loo, B. T.; and Mao, Y. 2010. Efficient querying and maintenance of network provenance at internet-scale. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 615–626. ACM.

Zhu, K., and Ying, L. 2013. Information source detection in the sir model: A sample path based approach. In *Information Theory and Applications Workshop (ITA), 2013*, 1–9. IEEE.

Zhu, K., and Ying, L. 2014. A robust information source estimator with sparse observations. *Computational Social Networks* 1(1):1–21.