Provable Sensor Sets for Epidemic Detection over Networks with Minimum Delay

Jack Heavey[1], **Jiaming Cui**[2], **Chen Chen**[1], **B. Aditya Prakash**[2], **Anil Vullikanti**[1]

[1] University of Virginia
[2] Georgia Institute of Technology
{jch7jm, zrh6du, vsakumar}@virginia.edu, {jiamingcui1997, badityap}@gatech.edu

Abstract

The efficient detection of outbreaks and other cascading phenomena is a fundamental problem in a number of domains, including disease spread, social networks, and infrastructure networks. In such settings, monitoring and testing a small group of pre-selected susceptible population (i.e., sensor set) is often the preferred testing regime—we refer to this as the MinDelSS problem. Prior methods for minimizing the detection time rely on greedy algorithms using submodularity. We show that this approach can lead to sometimes lead to a worse approximation for minimizing the detection time than desired. We also show that MinDelSS is hard to approximate within an $O(n^{1-1/\gamma})$-factor for any constant $\gamma \geq 2$ (n is the number of nodes in the graph), which motivates our bicriteria approximations. We present the algorithm ROUNDSENSOR, which gives a rigorous worst case $O(\log n)$-factor for the detection time, while violating the budget by a factor of $O(\log^2 n)$. Our algorithm is based on the sample average approximation technique from stochastic optimization, combined with linear programming and rounding. We evaluate our algorithm on several networks, including hospital contact networks, which validates its effectiveness in real settings.

1 Introduction

Recurring disease outbreaks, including the COVID-19 pandemic and the emerging variants, illustrate the importance of good surveillance in order to detect the onset of a disease outbreak in the population. Diverse mechanisms exist for testing, ranging from syndromic surveillance (CDC 2021) based on symptoms to more accurate PCR based tests, which can detect fragments of the pathogen. However, such testing is expensive, both in terms of personnel and equipment needed. This is also true in a hospital setting, where hospital acquired infections, such as MRSA, pose significant health burden (Stone 2009; Leclère et al. 2017). Problems of detecting a spreading process also arise more generally in a number of other applications, such as water networks and blog networks (Leskovec et al. 2007).

Such surveillance problems involve finding a "sensor set" S, such that monitoring nodes in S is sufficient to detect the disease in the network. As in prior work (Leskovec et al. 2007; Adhikari et al. 2019), we assume that either a set of cascades (or possible scenarios) of disease transmission in a graph are given, or the cascades are sampled from an SIR process on the network. Standard metrics for effective surveillance are probability of detection and the expected delay in detection. Submodularity has been the primary technique in finding good algorithms in prior work (Leskovec et al. 2007; Adhikari et al. 2019). While the detection probability is submodular, the detection time is not (also, it is a minimization problem). (Leskovec et al. 2007) show that a slightly different "penalty reduction" variation of the detection time objective is submodular. As a result, a greedy algorithm gives a $(1 - 1/e)$-factor approximation.

However, as we show, a solution computed by the greedy algorithm for maximizing such a penalty reduction objective can be highly suboptimal with respect to the detection time—this is especially important in regimes when the detection time is low, which is the more realistic setting in disease surveillance. No prior results are known for directly minimizing the detection time. Further, in an SIR model, minimizing the expected detection time is a stochastic optimization problem, which hasn't been considered before.

Our contributions.

- We study the problem MinDelSS of finding a sensor set with a given budget, that directly minimizes the detection time; the cascades can be either specified as input, or sampled, in the case of the SIR model for epidemic spread. We show that the greedy approach of (Leskovec et al. 2007) can have an approximation factor of $\Theta(n)$, in general, where n denotes the number of nodes in the graph. We prove that MinDelSS is hard to approximate within an $O(n^{1-1/\gamma})$-factor, for any constant $\gamma \geq 2$, which contrasts with the complexity of maximizing the detection probability or the penalty reduction approach for detection time (Leskovec et al. 2007); this hardness holds even when the cascades are sampled from the SIR process.
- We design a bicriteria approximation algorithm, ROUNDSENSOR, which gives a rigorous worst case $O(\log n)$-factor for the average delay for a set of cascades, while violating the budget by a factor of $O(\log^2 n)$. Note that in light of the above approximation hardness, we believe such a bicriteria approximation is the best option for finding effective algorithms. We

combine this with the sample average approximation technique from stochastic optimization and derive similar bounds on the expected detection time in the SIR model.

- We evaluate our algorithms on multiple real world networks, including four contact networks between patients and health care workers in a hospital. Two of these are quite novel, and constructed using patient electronic medical record (EMR) data; additionally, we have networks for a pre-COVID and COVID period. *Our results show that the empirical performance is significantly better than all our theoretical worst case bounds*, including the approximation factor for the average delay, the violation in budget, and the number of sampled cascades needed in the case of the SIR model. We also find interesting structural differences in the solutions for the pre-COVID and COVID networks.

Due to the space constraint, we omit some of the technical and experimental details; these are presented in the Supplementary material.

2 Related Work

As mentioned earlier, there has been a lot of work on surveillance and finding sensor sets in different applications, including water networks, air pollution, blogs networks and epidemic surveillance, e.g., (Leskovec et al. 2007; Shao et al. 2018; Christakis and Fowler 2010; Adhikari et al. 2019; Leclère et al. 2017; Hsieh, Lin, and Zheng 2015). (Leclère et al. 2017) describe diverse approaches for outbreak detection for hospital acquired infections–these include statistical process control, scan statistics, traditional statistical models, and data mining methods. (Christakis and Fowler 2010) introduce the problem of finding sensor sets that give good lead time for the peak time of an epidemic—they show that a sensor set chosen based on popular friends of a random set provides good lead time; this was improved by (Shao et al. 2018) using an approach based on dominator sets. Sensors have also been deployed on social networks (Lerman, Yan, and Wu 2016) like Twitter to detect major events. For example, (Kryvasheyeu et al. 2015) studies the detectable patterns of user Twitter messages during Hurricane Sandy to track the disaster. (Sakaki, Okazaki, and Matsuo 2010) treats every user in Twitter as a sensor to construct an earthquake reporting system while (Zhang et al. 2017; Paul, Peng, and Li 2019) sense the geo-tagged tweets to detect events.

Much of the work on surveillance in the AI and ML literature that provide rigorous performance guarantees is based on using submodular optimization. (Leskovec et al. 2007) introduce different metrics for detection by sensor sets with a fixed budget, and reduce them to submodular maximization, including for detection time, as mentioned earlier; this allows a simple greedy algorithm gives a $(1 - 1/e)$-approximation (Nemhauser, Wolsey, and Fisher 1978). (Adhikari et al. 2019) extend this problem to surveillance schedules, instead of a fixed sensor sets, i.e., nodes could be tested at different times, and develop approximation algorithms using submodular functions on a lattice. More complex cost constraints for the sensor set than cardinality have also been considered, e.g., (Krause et al. 2008).

3 Preliminaries

Cascades and testing. A cascade $H = (V, E')$ is a subgraph of an undirected graph $G = (V, E)$ which is initiated at a node $s(H) \in V$. A node v that is reachable from $s(H)$ and is at distance $t(v, H) - 1$ in H is said to be infected at time $t(v, H)$ (the source $s(H)$ is infected at time 1). If v is not reachable from $s(H)$ in H, we define $t(v, H) = n+1$ where n is the number of nodes within the graph. We consider the following setting for testing— if a node v gets infected at time t' and is tested at any time $t \geq t'$, its prior infection status gets detected. In the context of a disease spread, this could be viewed as a model for antibody tests. A *sensor set* S is a subset of nodes which get tested every day; the detection time $T(S, H)$ for sensor set S in cascade H, denoted by $T(S, H) = \min_{v \in S} t(v, H)$ is the minimum time at which any node in S gets infected in H (recall that the time is $n+1$ in case no node in S gets infected in H). We assume a set of N cascades $H_1, \ldots, H_N$ are given as input. The objective of interest is the *average detection time with respect to* S, denoted by $T_{avg}(S) = \frac{1}{N} \sum_H T(S, H)$.

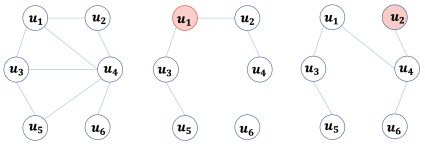

Figure 1: (a) Contact graph G with $V = \{u_1, \ldots, u_6\}$, and edges shown by solid lines; (b) Cascade H_1 with node u_1 as the source in this outcome, and the solid edges correspond to those on which the disease spread. Node u_6 does not get infected in this outcome; (c) Cascade H_2 with node u_2 as a source.

Example. In the example in Figure 1(b) with budget $k = 1$, node u_4 gets infected at time $t = 3$, i.e., $t(u_4, H_1) = 3$, and so if it is tested any time $t' \geq 3$, the infection would be detected. For the sensor set $S_1 = \{u_4\}$, we have $T(S_1, H_1) = 3$ and $T(S_1, H_2) = 2$. For $S_2 = \{u_6\}$, we have $T(S_2, H_1) = 7$ and $T(S_2, H_2) = 3$. For an input with cascades H_1, H_2, we have $T_{avg}(S_1) = 2.5$ and $T_{avg}(S_2) = 5$. For this instance, $S^* = \{u_2\}$ is the optimal solution with $T_{avg}(S^*) = 1.5$.

Min Delay Sensor Set (MinDelSS) problem. Given a graph $G = (V, E)$, a set of cascades $H_1, \ldots, H_N$, and a budget parameter k, find a sensor set S of size at most k such that $T_{avg}(S)$ is minimized.

Approximation algorithms. We say that S is an (α, β)-bicriteria approximation for MinDelSS if $T_{avg}(S) \leq \alpha T_{avg}(S^*)$ and $|S| \leq \beta k$, where S^* is an optimal solution to the instance of MinDelSS.

Detection in an SIR model. A specific instance of MinDelSS is in the context of the spread of a disease on a network. We consider an SIR model on a graph $G = (V, E)$, in which the disease spreads from an infectious node u to a

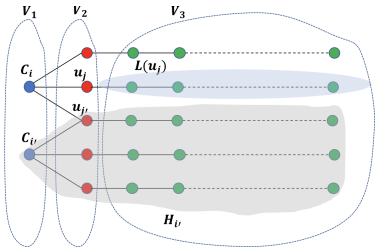

Figure 2: Reduction from the Hitting Set problem to MinDelSS. The cascade $H_{i'}$ consists of the node corresponding to $C_{i'}$, the set of nodes in $C_{i'}$, and $\cup_{u_{j''} \in C_{i'}} L(u_{j''})$

susceptible neighbor $v \in N(u)$ with probability p_{uv}, independently. We assume p_u^0 is the probability that the disease starts at node u; $\sum_u p_u^0 = 1$. In this setting, a cascade H is a random subgraph on which the disease spreads. In this setting, our objective is to minimize $\mathbb{E}[T(S)]$, where the expectation is over the stochastic disease outcomes H.

Notation	Definition
N	Number of cascades
$t(v, H)$	Time at which node v is infected in cascade H
$T(S, H)$	Detection time in cascade H for sensor set S
$T_{avg}(S)$	Average detection time for S
p_{uv}	Probability that node u infects susceptible neighbor v in the SIR model
p_u^0	Probability that the disease starts at node u
$\mathbb{E}[T(S)]$	Expected detection time when cascades are sampled using the SIR model
k	Budget for sensor set

Table 1: Summary of notation used in the paper.

4 Inefficiency of greedy and computational hardness of MinDelSS

4.1 Greedy can be very inefficient

(Leskovec et al. 2007) propose a greedy algorithm which attempts to find a set S that maximizes $\pi(S) = \frac{1}{N}\sum_i (n + 1 - T(S, H_i))$ for the set of cascades, which is equal to $n+1-T_{avg}(S)$. They show that $\pi(S)$ is submodular, which implies that a greedy algorithm gives a $(1 - 1/e)$ approximation to the optimal solution. However, maximizing $\pi(\cdot)$ doesn't imply minimization of the delay, as shown below; the proof is presented in the Supplementary material.

Lemma 1. *There exist instances where the solution S_g computed by the greedy algorithm of (Leskovec et al. 2007) has $T_{avg}(S_g) = \Omega(nT_{avg}(S^*))$, where S^* denotes an optimal solution.*

4.2 Computational hardness of MinDelSS

In contrast to the $(1 - 1/e)$-factor approximation for maximizing detection probability, or the penalty reduction approach for detection time (Leskovec et al. 2007), MinDelSS is much harder, both when the cascades are specified, and in the SIR model.

Theorem 1. *No polynomial time $O(n^{1-1/\gamma})$-approximation to MinDelSS is possible, unless P=NP, for a constant $\gamma \geq 2$.*

Our proof is by a reduction from the Hitting Set problem. An instance $\mathcal{H}$ of Hitting Set is a tuple $\mathcal{H} = (U = \{u_1, \ldots, u_r\}, \mathcal{C} = \{C_1, \ldots, C_m\})$, where each $C_i \subset U$; without loss of generality, we can assume $m = \Theta(r)$. A subset $S \subset U$ is a hitting set for $\mathcal{H}$ if for all i, we have $S \cap C_i \neq \emptyset$. The decision question is, given a bound k, is there a hitting set S for $\mathcal{H}$ with $|S| \leq k$? We construct an undirected graph $G = (V, E)$ from $\mathcal{H}$ in the following manner (see Figure 2 for an illustration). We assume γ is an integer; else we consider $\lceil \gamma \rceil$.

- We have $V = V_1 \cup V_2 \cup V_3$, where $V_1 = \mathcal{C}$, and $V_2 = U$. For each $u_j \in V_2$, let $L(u_j)$ denote a set of r^γ nodes. The set $V_3 = \cup_j L(u_j)$. Then $n = |V| = m + r + r^{\gamma+1}$.
- The set E of edges is constructed in the following manner: if $u_j \in C_i$, we have edge (u_j, C_i). For each u_j, the set $L(u_j)$ forms a path, which is connected to u_j.
- We construct m cascades $H_1, \ldots, H_m$, with H_i consisting of the subgraph induced by the set of nodes $V_i = \{C_i\} \cup C_i \cup_{u_j \in C_i} L(u_j)$.

Lemma 2. *If there exists a hitting set for $\mathcal{H}$ of size k, then there exists an optimal sensor set S^* for G with $|S^*| \leq k$ such that $T_{avg}(S^*) \leq 2$.*

Proof. Let $S \subset U$ be a hitting set for $\mathcal{H}$ with $|S| = k$. Then, for each C_i, we have $C_i \cap S \neq \emptyset$. Due to the structure of H_i, it follows that $T(S, H_i) = 2$, which implies $T_{avg}(S) = 2$. □

Lemma 3. *There exists a constant c such that if there exists a sensor set S for G with $T_{avg}(S) < cn^{1-1/\gamma}$, then there exists (1) a hitting set for $\mathcal{H}$ of size at most $|S|$, and (2) a sensor set S^* for G with $T_{avg}(S^*) \leq 2$.*

Proof. First, observe that without loss of generality, we can assume that $S \cap L(u_j) = \emptyset$ for all u_j; if not, and there exists a node $v \in S \cap L(u_j)$, the set $S \cup \{u_j\} - \{v\}$ has expected delay no larger than $T_{avg}(S)$.

Next, suppose there exists C_i such that $(\{C_i\} \cup C_i) \cap S = \emptyset$. Then, for the sample H_i, we have $T(S, H_i) \geq r$, which implies $T_{avg}(S) \geq n/m \geq r^\gamma \geq cn^{1-1/\gamma}$ for a constant c,. Thus, if $T_{avg}(S) < cn^{1-1/\gamma}$, it must follow that for each C_i, we have $(\{C_i\} \cup C_i) \cap S \neq \emptyset$. Construct a set $S' \subset U$ in the following manner: if $u_j \in S$, add u_j to S', and if $C_i \in S$, add any $u_j \in C_i$ to S'. It follows that for all C_i, $C_i \cap S' \neq \emptyset$, and so S' is a hitting set for $\mathcal{H}$. Further, $|S'| \leq |S|$, and since S' is a hitting set, $T_{avg}(S') \leq 2$. Thus, the lemma follows. □

Lemma 3 implies that if S^* is an optimal solution for graph G, then either $T_{avg}(S^*) \leq 2$ or $T_{avg}(S^*) \geq cn^{1-1/\gamma}$.

Proof. (of Theorem 1) Suppose we have a polynomial time algorithm with approximation factor less than $\frac{cn^{1-1/\gamma}}{2}$, where c is the constant in Lemma 3. Let S be the sensor set computed by such an algorithm for the instance G. Let

S^* be an optimal solution for G. Let S_h^* be an optimal hitting set for $\mathcal{H}$. Then, $T_{avg}(S) < \frac{cn^{1-1/\gamma}}{2} T_{avg}(S^*)$. Recall that the decision question for the hitting set problem is whether $|S_h^*| \leq k$. From Lemmas 2 and 3, it follows that either $T_{avg}(S^*) \leq 2$ or $T_{avg}(S^*) \geq cn^{1-1/\gamma}$. In the former case, we would have $T_{avg}(S) < cn^{1-1/\gamma}$, and in the latter case we would have $T_{avg}(S) \geq cn^{1-1/\gamma}$. Thus, if $T_{avg}(S) < cn^{1-1/\gamma}$, it follows that $T_{avg}(S^*) \leq 2$, and $|S_h^*| \leq |S^*| \leq k$. On the other hand, if $T_{avg}(S) \geq cn^{1-1/\gamma}$, it follows that $T_{avg}(S^*) > 2$, and by Lemma 2, $|S_h^*| > k$. Therefore, a polynomial time algorithm with approximation factor less than $cn^{1-1/\gamma}/2$ allows us to solve the hitting set problem, and the theorem follows. □

Hardness for MinDelSS in the SIR model. Our reduction can be adapted to show that the hardness holds even in the SIR model. The details are presented in the Supplementary Information.

Theorem 2. *No polynomial time $O(n^{1-1/\gamma})$-approximation to MinDelSS is possible for any constant $\gamma \geq 2$, unless P=NP, even when the cascades are sampled from the SIR model.*

5 Our approach

We first present ROUNDSENSOR for the setting in which a set of cascades $H_1, \ldots, H_N$ is given as input, and the goal is to find a sensor set S to minimize $T_{avg}(S)$. Later, we show how this can be extended to an SIR model with the goal of minimizing $T_{avg}(S)$.

5.1 ROUNDSENSOR under given cascades

Our algorithm, ROUNDSENSOR, is based on linear programming relaxation and randomized rounding. We first start with the following integer program (IP) with two kinds of variables: the variables x_u indicate that node u is picked in the sensor set, and y_{id} indicates that in the sample H_i, a node in the set V_{id} is in the sensor set, where V_{id} is the set of nodes at distance $d-1$ from the source $s(H_i)$ (i.e., the detection time is d).

$$\min \sum_{d=0}^{n+1} \frac{1}{N} \sum_{i=1}^{N} y_{id} \cdot d \tag{1}$$

$$\text{for all } i, d: \sum_{u \in V_{id}} x_u \geq y_{id} \tag{2}$$

$$\sum_{u} x_u \leq k \tag{3}$$

$$\text{for all } i: \sum_{d} y_{id} = 1 \tag{4}$$

$$x_u, y_{id} \in \{0, 1\} \tag{5}$$

The constraint (2) indicates that if $y_{id} = 1$ (i.e., the detection time is d), then some node in V_{id} is picked. The constraint (4) ensures that $y_{id} = 1$ for exactly one value of d (which will be the minimum, because of the objective).

Lemma 4. *The above integer program (IP) is valid, i.e., if x, y is an optimal solution to the IP, then $T_{avg}(S^*) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{n+1} y_{id} \cdot d$, where S^* is an optimal solution.*

Proof. Consider a solution x^*, y^* corresponding to S^* defined in the following manner: if $u \in S^*$, we have $x_u^* = 1$, else $x_u^* = 0$. If $T(S^*, H_i) = d$, we set $y_{id}^* = 1$, else $y_{id}^* = 0$. It is easy to verify that x^*, y^* is a feasible solution to (IP). Similarly, any feasible solution x, y corresponds to a sensor set S. Therefore, the lemma follows. □

Relaxation and rounding. Solving (IP) is infeasible for large instances. We relax the difficult integrality constraint (5) to get a linear program, and use the technique of randomized rounding to get an approximate solution, as summarized in Algorithm 1.

Algorithm 1: ROUNDSENSOR
Input: $G = (V, E), k, H_1, \ldots, H_N$
Output: Sensor set S

1: Solve the LP obtained by relaxing the constraints (5) of IP to $x_u, y_{id} \in [0, 1]$ for all u, i, d.
2: Let x, y be the optimum fractional solution to the above LP. For each node u, add u to S_r with probability $x'_u = \min\{1, x_u \log(n+1) \log(Nn)\}$
3: Return S_r

***Main ideas behind the analysis of* ROUNDSENSOR**

- Let $U_j = \{2^j, \ldots, \min\{n+1, 2^{j+1}-1\}\}$, for $0 \leq j \leq \log(n+1)$. While the variables in the LP can be fractional, and y_{id} need not be 1, we can show that there is a set $U_{d(i)}$ for each sample H_i, such that the sum of y_{id} for $d \in U_{d(i)}$ is not too small (Lemma 5).
- For each $d(i)$, it follows that $\sum_{d \in U_{d(i)}} \sum_{u \in V_{id}} x'_u \geq \log(Nn)$. As a result, the randomized rounding in Step 3 of ROUNDSENSOR ensures that at least one node $u \in \cup_{d \in U_{d(i)}} V_{id}$ is picked in the solution (Lemma 6).
- Finally, the scaling used to construct the variables x'_u and the definition of the sets U_j give bicriteria bounds on $|S_r|$ and $T_{avg}(S_r)$ (Theorem 3).

Lemma 5. *Let x, y denote the solution to the LP. Then, for all $i = 1, \ldots, N$, there exists $d(i) \leq \log(n+1)$ such that $\sum_{d \in U_{d(i)}} y_{id} \geq \frac{1}{\log(n+1)}$.*

Proof. The sets U_j are disjoint, and $\sum_{d=1}^{n+1} y_{id} = \sum_{j \geq 0} \sum_{d \in U_j} y_{id} = 1$, which implies there exists $d(i)$ such that $\sum_{d \in U_{d(i)}} y_{id} \geq \frac{1}{\log(n+1)}$. □

Lemma 6. *Let S_r denote the solution picked by our algorithm. For each cascade H_i, let $d(i)$ be the index, as in Lemma 5. Then, with probability at least $1 - \frac{1}{n}$, for all $i = 1, \ldots, N$, we have $T(S_r, H_i) \leq 2^{d(i)+1}$.*

Proof. Let $B_{id(i)} = \cup_{d \in U_{d(i)}} V_{id}$. We will show below that for all i, we have $S_r \cap B_{id(i)} \neq \emptyset$ with probability at least $1 - \frac{1}{n^2}$. Observe that if $S_r \cap B_{id(i)} \neq \emptyset$, we have $T(S, H_i) \leq 2^{d(i)+1}$, and the lemma follows.

Consider any fixed i. Observe that $\Pr[S_r \cap B_{id(i)} = \emptyset] = \prod_{u \in B_{id(i)}} (1 - x'_u)$. We have two cases. First, suppose there is a node $u \in B_{id(i)}$ such that $x'_u = 1$, then $\Pr[S_r \cap B_{id(i)} = \emptyset] = 0$, which implies $S_r \cap B_{id(i)} \neq \emptyset$.

Second, if $x'_u < 1$ for all $u \in B_{id(i)}$, we have $x'_u = \log(n+1)\log(Nn) \cdot x_u$. In this case, $\Pr[S_r \cap B_{id(i)} = \emptyset] = \prod_{u \in B_{id(i)}} (1 - x'_u) \leq \prod_{u \in B_{id(i)}} exp(-x'_u) = exp(-\sum_{u \in B_{id(i)}} x'_u)$. We have

$$\begin{aligned}\sum_{u \in B_{id(i)}} x'_u &= \log(n+1)\log(Nn) \sum_{d \in U_{d(i)}} \sum_{u \in V_{id}} x_u \\ &\geq \log(n+1)\log(Nn) \sum_{d \in U_{d(i)}} y_{id} \\ &\geq \log(Nn),\end{aligned}$$

where the first inequality follows because of the constraint $\sum_{u \in V_{id}} x_u \geq y_{id}$ in the LP, and the second inequality follows from Lemma 5.

Therefore, $\Pr[S_r \cap B_{id(i)} = \emptyset] \leq exp(-\sum_{u \in B_{id(i)}} x'_u) \leq \frac{1}{Nn}$, using the above bound on $\sum_{u \in B_{id(i)}} x'_u$. This implies that in both the cases, we have $\Pr[S_r \cap B_{id(i)} = \emptyset] \leq \frac{1}{Nn}$. By a union bound over all the cascades i, we have $\Pr[\text{There exists } i \text{ such that } S_r \cap B_{id(i)} = \emptyset] \leq \frac{N}{Nn} = \frac{1}{n}$, and the Lemma follows. □

Theorem 3. *Let S_r be the set of nodes selected by the above algorithm. With probability at least $1 - \frac{2}{n}$, we have: (1) $|S_r| \leq k \cdot 2\log(Nn)\log(n+1)$, and (2) $T_{avg}(S_r) \leq 2\log(n+1) T_{avg}(S^*)$, where S^* is an optimal solution.*

Proof. Let $X_u = 1$ with probability x'_u. Then, $|S_r| = \sum_u X_u$, and $\mathbb{E}[|S_r|] = \sum_u x'_u \geq \log(Nn)\log(n+1) \cdot k$. Applying the Chernoff bound (Theorem 5 in the Appendix), we have $\Pr[|S_r| > 6k\log(Nn)\log(n+1)] \leq 2^{-6k\log(Nn)\log(n+1)} \leq 2^{-6\log(n+1)} \leq 1/n$.

For all i, we have $\sum_{d \in U_{d(i)}} dy_{id} \geq 2^{d(i)} \sum_{d \in U_{d(i)}} y_{id} \geq \frac{2^{d(i)}}{\log(n+1)} \geq \frac{T(S_r, H_i)}{2\log(n+1)}$, with probability at least $1 - 1/n$, where the first inequality follows from the definition of the set $U_{d(i)}$, the second inequality follows from Lemma 5, and the third inequality follows from Lemma 6. This implies that with probability at least $1 - 1/n$, for all i, $T(S_r, H_i) \leq 2\log(n+1) \cdot \sum_{d \in U_{d(i)}} dy_{id} \leq 2\log(n+1) \cdot \sum_d dy_{id}$. Therefore, with probability at least $1 - 1/n^2$, $T_{avg}(S_r) = \frac{1}{N}\sum_i T(S_r, H_i) \leq \frac{1}{N}\sum_i 2\log(n+1) \cdot \sum_d dy_{id} \leq 2\log(n+1) T_{avg}(S^*)$, since the LP value is a lower bound for $T_{avg}(S^*)$. The probability that either the bound for $|S_r|$ or $T_{avg}(S_r)$ is not satisfied is at most $\frac{2}{n}$, and therefore, the theorem follows. □

5.2 Minimizing $\mathbb{E}[T(S)]$ in the SIR model

In the context of disease surveillance, the disease spread is modeled by an SIR process on the network. The cascades are not specified ahead of time, but are sampled according to the SIR process, as defined in Section 3. We use the sample average technique from stochastic optimization and show that if we run ROUNDSENSOR on a polynomial set of cascades, we also get a good approximation for the expected detection time; we refer to this as Algorithm ROUNDSENSORSIR.

Algorithm 2: ROUNDSENSORSIR
Input: $G = (V, E), k$
Output: Sensor set S_r

1: Sample $N = \Omega(\frac{3}{\epsilon^2} n(n+1)\log n)$ cascades $H_1, \ldots, H_N$ using the SIR process
2: Return $S_r = \text{ROUNDSENSOR}(G, k, H_1, \ldots, H_N)$

Recall that $T(S, H)$ is the delay associated with set S for the disease outcome H. Let $\hat{S} = \text{argmin}_S T_{avg}(S)$ denote an optimal solution for the cascades $H_1, \ldots, H_N$. Let $S^* = \text{argmin}_S \mathbb{E}[T(S)]$ denote an optimal solution (which minimizes the expected detection time).

Lemma 7. *Let $N \geq \frac{3}{\epsilon^2} n(n+1)\log n$ for $\epsilon \in (0,1)$. For any $\epsilon \in (0,1)$, $\Pr[$there exists S such that $T_{avg}(S) \notin [(1-\epsilon)\mathbb{E}[T(S)], (1+\epsilon)\mathbb{E}[T(S)]] \leq 1/n^2$*

Proof. For any fixed S, we have $N\frac{T_{avg}(S)}{n+1} = \sum_{i=1}^{N} \frac{T(S,H_i)}{n+1}$. By the definition of $T(S, H_i)$, we have $Z_i = \frac{T(S,H_i)}{n+1} \in [0,1]$. Also, $\mathbb{E}[T(S, H_i)] = \mathbb{E}[T(S)]$, which implies $\mathbb{E}[NT_{avg}(S)] = N\mathbb{E}[T(S)]$. Applying Theorem 5 to $Z = N\frac{T_{avg}(S)}{n+1}$, we have

$$\begin{aligned}&\Pr\left[T_{avg}(S) \notin [(1-\epsilon)\mathbb{E}[T(S)], (1+\epsilon)\mathbb{E}[T(S)]\right] \qquad (6)\\ =&\Pr\left[N\frac{T_{avg}(S)}{n+1} \notin [(1-\epsilon)N\frac{\mathbb{E}[T(S)]}{n+1}, (1+\epsilon)N\frac{\mathbb{E}[T(S)]}{n+1}\right]\\ \leq& 2exp(-\epsilon^2 N\mathbb{E}[T(S)]/(3(n+1)))\\ \leq& exp(-n\log n)\end{aligned}$$

if $N \geq \frac{3}{\epsilon^2} n(n+1)\log n$, since $\mathbb{E}[T(S)] \geq 1$. By a union bound over all $S \subset V$, it follows that the probability that there exists S with $T_{avg}(S) \notin [(1-\epsilon)\mathbb{E}[T(S)], (1+\epsilon)\mathbb{E}[T(S)]$ is at most $2^n exp(-n\log n) \leq 1/n^2$. □

Theorem 4. *Let $N \geq \frac{3}{\epsilon^2} n(n+1)\log n$ for $\epsilon \in (0,1)$. With probability at least $1 - \frac{3}{n}$, (1) $|S_r| \leq k \cdot 7\log(\frac{1}{\epsilon})\log^2(n$ and (2) $\mathbb{E}[T(S_r)] \leq 2(1+\epsilon)\log(n+1)\mathbb{E}[T(S^*)]$, where S^* is an optimal solution.*

Proof. Plugging in the bound for $N = \frac{3}{\epsilon^2} n(n+1)$ in Theorem 3, we have $|S_r| \leq k \cdot 2\log(\frac{3}{\epsilon^2} n^2(n+1)\log n)\log(n+1) \leq (2 + c_1)\log(\frac{3+c_1}{\epsilon^2} n^3)\log(n) \leq (6 + c_2)\log^2(n)\log(\frac{1}{\epsilon})$ for small constants c_1, c_2, as long as $n \geq n_0$. This can, in turn, implies $|S_r| \leq k \cdot 7\log^2(n)\log(\frac{1}{\epsilon})$, for n larger than a constant.

Next, from Lemma 7, we have: (1) $T_{avg}(S_r) \geq (1-\epsilon)\mathbb{E}[T(S_r)$ with probability at least $1 - \frac{1}{n^2}$, and (2) $T_{avg}(S^*) \leq (1+\epsilon)\mathbb{E}[T(S^*)]$ with probability at least $1 - \frac{1}{n}$. From Theorem 3, we also have $T_{avg}(S_r) \leq 2\log(n +$

$1)T_{avg}(\hat{S}) \leq 2\log(n+1)T_{avg}(S^*)$, where $\hat{S}$ is an optimal solution for the N samples, and S^* is an optimal solution for the expected detection time. Putting all of them together, with probability at least $1-\frac{2}{n}$, we have $\mathbb{E}[T(S_r)] \leq \frac{1}{1-\epsilon}T_{avg}(S_r) \leq \frac{1}{1-\epsilon}2\log(n+1)T_{avg}(S^*) \leq 2\log(n+1)\frac{1+\epsilon}{1-\epsilon}\mathbb{E}[T(S^*)]$.

Combining both these parts, the bounds on $|S_r|$ and $\mathbb{E}[T(S_r)]$ both hold with probability at leasst $1-\frac{3}{n}$. □

This algorithm's runtime performance is dominated by the amount of time it takes to solve the relaxed linear program, which has $O(Nn)$ constraints and $O(Nn)$ variables. This gives us a worst case running time of $O((Nn)^{2.5})$, although many modern Linear Program solvers can solve these more efficiently.

6 Experiments

We study the following questions

- ***Effectiveness***: how does the approximation factor of RoundSensor in practice compare with the theoretical worst case bounds in Theorems 3? How do these compare with other baselines?
- ***Efficiency in the SIR model***: Theorem 4 requires $N = \Omega(n^2 \log n)$ sampled cascades. How many are sufficient in practice?
- ***Impact of transmission probability in the SIR models***: as the transmission probability increases, the problem becomes easier. What is the impact on the performance of RoundSensor?
- ***Case study***: what are structural properties of the solutions, e.g., what type of nodes are picked?

Due to the space constraint, we present results for a subset of the datasets here; additional analysis is described in the Supplementary material.

6.1 Datasets and baselines

We evaluate our algorithms on diverse kinds of real world data sets, as summarized in Table 2. These include four hospital contact networks.

1. arXiv High Energy Physics-Theory (HEP-TH) collaboration network from January 1993 to April 2004 (Leskovec and Krevl 2014). We only use the largest connected component in this network.
2. Battle of the Water Sensor Networks (BWSN) (Ostfeld et al. 2008).
3. Hospital network from Lyon: this is a temporal contact network of a hospital based in Lyon, France (Vanhems et al. 2013). The data set tracks 46 health care workers and 29 patients from December 6, 2010 at 1pm to December 10, 2010 at 2pm.
4. Carilion network: constructed from patient and provider contacts in the Carilion hospital in Roanoke, Virginia (Adhikari et al. 2019). We only use the network data for a 2 month period.
5. UVA pre-COVID and COVID networks: these are anonymized contact networks from electronic medical record (EMR) data of hospitalized patients at the University of Virginia (UVA) hospital. Nodes are patients and health care providers, while edges represent co-location based contacts. The Pre-COVID network spans the period 4/1/2018—10/28/2018, and the COVID network spans the period 10/14/2020—5/21/2021.

Graph Name	Number of nodes	Number of edges
Arxiv HEP-TH	8638	24827
Water network	12523	14822
Lyon Hospital Ward	75	1138
Carilion hospital network	11413	25663
UVA Pre-COVID	10789	291881
UVA COVID	9949	399495

Table 2: Data sets used for experiments

Baselines. We compare with the following baselines

- Random: select k nodes randomly with uniform probability distribution
- Degree: select top k nodes with highest degrees
- GREEDY: select k nodes with greedy scheme by (Leskovec et al. 2007).

6.2 Effectiveness

Figure 3 shows that RoundSensor performs better than the three baselines that we have selected. While this greedy approach does offer certainty in the budget allowance and a deterministic outcome, we see that RoundSensor offers increasing benefits as our budget increases, up to a 9% lower mean detection time after rounding. Note that these baselines have budget $k' = |S_r|$ after rounding to ensure a fair comparison.

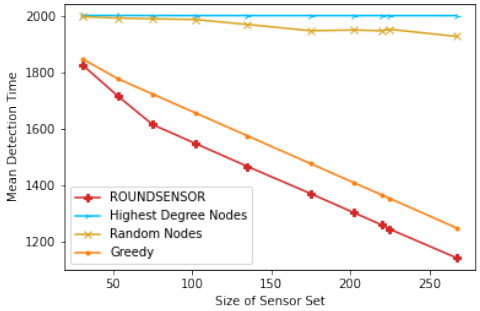

Figure 3: Mean detection time vs the sensor set size for the UVA COVID network, comparing RoundSensor and other baselines.

Next, we consider the approximation ratios. Recall that RoundSensor gives a bicriteria approximation (Theorem 3), and we evaluate both $T_{avg}(S_r)/T_{avg}(S^*)$ (the approximation ratio with respect to the objective), where S^* is an optimal solution, and $\frac{|S_r|}{k}$ (the violation in budget). We

cannot calculate the exact approximation ratio, since S^* is unknown, but the ratio between $T_{avg}(S_r)$ and the LP objective gives an upper bound on the approximation ratio. Figure 4 (left) shows that the approximation ratio achieved by ROUNDSENSOR, with respect to the objective value is less than 1.5—this is a significant contrast with the worst case $O(\log n)$ approximation factor we prove in Theorem 3.

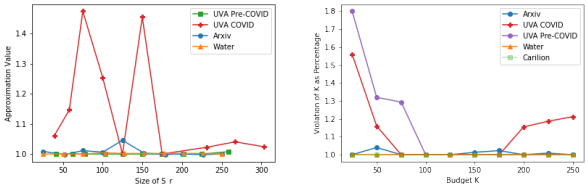

Figure 4: (Left)Upper bound on the approximation ratio achieved by ROUNDSENSOR vs $|S_r|$. Note that the maximum value on the y-axis is 1.5. (Right) Violation in budget for ROUNDSENSOR vs k, for a transmission probability of 0.15

Finally, Figure 4 (right) shows the violation in budget for ROUNDSENSOR. Note that for all values of k, the violation is at most 1.35, which is a significant contrast with the $O(\log^2 n)$ bound in Theorem 3.

6.3 Number of sampled cascades needed in the SIR model

Recall that Theorem 4 gives a bound of $N \geq \frac{3}{\epsilon^2} n(n+1)\log n$ for the analysis of the SAA technique. Figure 5 shows the impact of varying N. For low values of N, the objective value is low, since choosing nodes close to the sampled sources is a good strategy. We find that the plots plateau off well before $N = n$, which suggests that linear number of samples are adequate in practice.

6.4 Case study

We analyze the mean detection time in the UVA contact networks in Figure 8. We observe that the detection time reduces in the COVID period, for the same transmission probability. While there are likely many factors at play, we note that the COVID network is much denser (average density of 40.15) than the pre-COVID network (average density

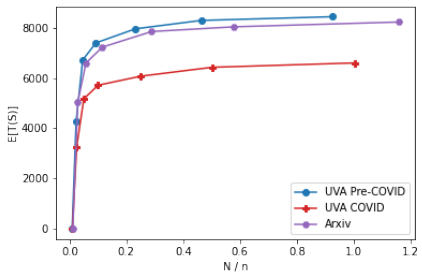

Figure 5: Mean detection time estimated from N samples vs N/n for different networks. Linear number of samples suffice for getting a good estimate.

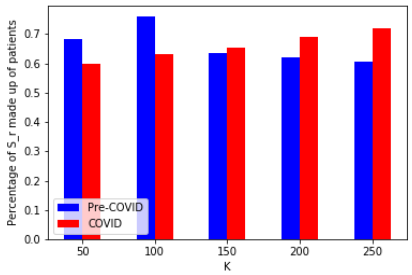

Figure 6: The proportion of S_r that consists of patients in the hospital network for pre-COVID and COVID data sets for different initial budgets k), for transmission probability $p = 0.15$.

27.05). Higher density generally increases the ease of disease transmission, e.g.,(Newman 2003). A deeper analysis is needed to understand this better.

Figure 6 shows how the makeup of S_r (namely, the fraction consisting of patients) changed from the pre-COVID to the COVID period. There are noticeable differences in the composition of the sensor sets, which changes with k. In particular, during the COVID period, the fraction of patients increases with k, becoming a more disproportionate size of the set, while in our pre-COVID data set, we see that decrease as it gets larger.

7 Conclusions

We present the first approximation algorithms for the MinDelSS, both when the cascades are specified, or sampled from an SIR process. Our algorithms give bicriteria approximation guarantees, which is inevitable in light of the computational hardness we prove for MinDelSS. Our experiments on diverse networks, including four hospital networks show that our method is quite effective. Identifying surrogates of the nodes picked in our near-optimal solutions can be useful in designing more implementable solutions. In practice, more general surveillance strategies need to be considered, in which a node is not tested daily, but with some rate. Our approach can be extended to settings where schedules are periodic.

Acknowledgments

We thank the anonymous reviewers for their detailed feedback, which helped strengthen the results in Section 4.2. This paper is based on work partially supported by NSF (Expeditions CCF-1918770 and CCF-1918656, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, Medium IIS-2106961, CCF-2115126, IIS-1931628, IIS-1955797), CDC MInD U01CK000589, NIH R01GM109718, DTRA subcontract/ARA S-D00189-15-TO-01-UVA, ORNL, faculty award from Facebook, and funds/computing resources from Georgia Tech.

References

Adhikari, B.; Lewis, B.; Vullikanti, A.; Jimenez, J. M.; and Prakash, B. A. 2019. Fast and Near-Optimal Monitoring for Healthcare Acquired Infection Outbreaks. *PLoS Computational Biology*.

CDC. 2021. National Syndromic Surveillance Program (NSSP). https://www.cdc.gov/nssp/overview.html.

Christakis, N. A.; and Fowler, J. H. 2010. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9): e12948.

Dubhashi, D. P.; and Panconesi, A. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press. ISBN 978-0-521-88427-3.

Hsieh, H.-P.; Lin, S.-D.; and Zheng, Y. 2015. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 437–446. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Krause, A.; McMahan, H.; Guestrin, C.; and Gupta, A. 2008. Robust Submodular Observation Selection. *Journal of Machine Learning Research*, 9: 2761–2801.

Kryvasheyeu, Y.; Chen, H.; Moro, E.; Van Hentenryck, P.; and Cebrian, M. 2015. Performance of social network sensors during Hurricane Sandy. *PLoS one*, 10(2): e0117288.

Leclère, B.; Buckeridge, D. L.; Boëlle, P.-Y.; Astagneau, P.; and Lepelletier, D. 2017. Automated detection of hospital outbreaks: A systematic review of methods. *PloS one*, 12(4): e0176438.

Lerman, K.; Yan, X.; and Wu, X.-Z. 2016. The" majority illusion" in social networks. *PloS one*, 11(2): e0147617.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429. ACM.

Leskovec, J.; and Krevl, A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An Analysis of Approximations for Maximizing Submodular Set Functions–I. *Math. Program.*, 14(1): 265–294.

Newman, M. 2003. The Structure and Function of Complex Networks. *SIAM Review*, 45: 167–256.

Ostfeld, A.; et al. 2008. The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *Journal of Water Resources Planning and Management*.

Paul, D.; Peng, Y.; and Li, F. 2019. Bursty event detection throughout histories. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 1370–1381. IEEE.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860.

Shao, H.; Hossain, K.; Wu, H.; Khan, M.; Vullikanti, A.; Prakash, B. A.; Marathe, M.; and Ramakrishnan, N. 2018. Forecasting the Flu: designing social network sensors for epidemics. *SIGKDD epiDAMIK Workshop*.

Stone, P. W. 2009. Economic burden of healthcare-associated infections: an American perspective. *Expert review of pharmacoeconomics & outcomes research*, 9(5): 417–422.

Vanhems, P.; Barrat, A.; Cattuto, C.; Pinton, J.; Khanafer, N.; Regis, C.; Kim, B.; Comte, B.; and Voirin, N. 2013. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE*, 8(9): e73970.

Zhang, C.; Liu, L.; Lei, D.; Yuan, Q.; Zhuang, H.; Hanratty, T.; and Han, J. 2017. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 595–604.

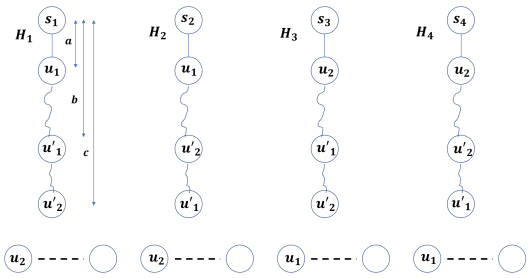

Figure 7: The greedy algorithm can have an approximation factor $\Omega(n)$.

8 Supplementary material

We use the following version of the Chernoff bound.

Theorem 5. *(Theorem 1.1 of (Dubhashi and Panconesi 2009)) Let $Z = \sum_{i=1}^{n} Z_i$, where Z_i are independently distributed random variables in $[0,1]$. Then, for any $\epsilon \in (0,1)$, we have $\Pr[Z \notin [(1-\epsilon)E[Z], (1+\epsilon)E[Z]]] \leq 2exp(-\epsilon^2 E[Z]/3)$. Also, for any $t > 2eE[Z]$, $\Pr[Z > t] \leq 2^{-t}$.*

8.1 Proof of Lemma 1

Proof. Consider the instance shown in Figure 7, which consists of four cascades $H_1, \ldots, H_4$ on a graph $G = (V, E)$ with n nodes. Let $A = \{u_1, u_2, u'_1, u'_2\}$ denote four special nodes. Let $V_1 \cup V_2 \cup V_3 \cup V_4$ be a partition of $V - A$ of size $(n/4 - 1)$ each. For each $i = 1, \ldots, 4$, the cascade H_i has the following structure: (1) it consists of a tree on the set V_i with a source s_i, (2) the $3(n/4-1)$ nodes in $\cup_{j \neq i} V_j$ are all isolated nodes, and (3) the nodes u_1, u_2, u'_1, u'_2 have distances from s_i as shown in Figure 7. For instance, in H_1, u_1, u'_1 and u'_2 are at distances a, b and c, respectively, from s_i, whereas u_2 is an isolated node.

We first argue that when the budget $k = 2$, the greedy algorithm picks the set $\{u'_1, u'_2\}$. Observe that $\pi(\{u_1\}) = \pi(\{u_2\}) = \frac{1}{4}(n - a + n - a + 0 + 0) = \frac{1}{2}(n - a)$, whereas $\pi(\{u'_1\}) = \pi(\{u'_2\}) = \frac{1}{4}(n - b + n - c + n - b + n - c) = \frac{1}{2}(n - b + n - c)$. By ensuring that $n - b + n - c \geq n - a$, i.e., $b + c \leq n + a$, we have $\pi(\{u'_1\} \geq \pi(\{u_1\}) = \pi(\{u_2\})$. Further for any node $u \notin A$, we have $\pi(\{u\} \leq n/4$, since u is an isolated node in three of the cascades. This implies $\pi(\{u'_1\}) \geq \pi(\{u\})$ for any node u, and so the greedy algorithm picks node u in the first iteration. Next, $\pi(\{u'_1, u_1\}) = \pi(\{u'_1, u_2\}) = \frac{1}{4}(n-a+n-a+n-b+n-c)$, whereas $\pi(\{u'_1, u'_2\}) = \frac{1}{4}(n - b + n - b + n - b + n - b) = (n - b)$. Choosing a, b, c such that $3(n - b) \geq 2(n - a) + (n - c)$, i.e., $b \leq \frac{2a}{3} + \frac{c}{3}$, ensures that $\pi(\{u'_1, u'_2\}) \geq \pi(\{u'_1, u_1\}) = \pi(\{u'_1, u_2\})$. Further, for any other node $u \notin A$, $\pi(\{u'_1, u\}) = \frac{1}{4}(n - b + n - c + n - b) + \frac{n}{4}$, since u is an isolated node in three cascades. By choosing b, c such that $2b \leq \frac{3n}{4} + c$, it follows that $\pi(\{u'_1, u'_2\}) \geq \pi(\{u'_1, u\})$ for any such node $u \notin A$.

We choose $a = 2, b = \frac{n}{32}, c = \frac{n}{8}$, which satisfies all the conditions above. Therefore, for the above conditions on a, b, c, the greedy algorithm picks the set $S_g = \{u'_1, u'_2\}$. The optimal solution is $\{u_1, u_2\}$. This implies $T_{avg}(S_g) = b + 1 = \Theta(n) T_{avg}(S^*)$. □

8.2 Hardness for MinDelSS in the SIR model

Our reduction can be adapted to show that the hardness holds even in the SIR model. We construct the same instance G as before. For all C_i, u_j such that $u_j \in C_i$, we have $p_{C_i,u_j} = 1$ and $p_{u_j,C_i} = 0$. For all the remaining edges $(u, v) \in E$ (i.e., (u, v) not of the form (C_i, u_j)), we have $p_{uv} = 1$ We have $p^0_{C_i} = 1/m$ for all C_i, and $p^0_v = 0$ for all $v \notin V_1$. Note that when we run the SIR process on G, there are exactly m possible outcomes $H_1, \ldots, H_m$, where H_i is the outcome if node C_i is the source. Due to the way the edge probabilities are defined, H_i consists of the edges (C_i, u_j) and the paths $L(u_j)$ for each $u_j \in C_i$. For a sensor set S, we have $\mathbb{E}[T(S)] = \frac{1}{m} \sum_{i=1}^{m} T(S, H_i)$.

8.3 Impact of transmission probability in the SIR model

The MinDelSS problem for the SIR model becomes easier as the transmission probability increases, since the outbreak size increases. Figure 8 shows that the mean detection time decreases with the budget for different values of transmission probability p. We observe that the value of p has a very significant impact on this variation in the UVA pre-COVID and COVID networks. This is more clear in Figure 9, which shows the mean detection time vs p, for different values of k. We observe steep decline in the objective, as p increases. We also see in figure 9 that our budget k is much less significant than our probability of transmission p for determining the mean detection time in outbreaks.

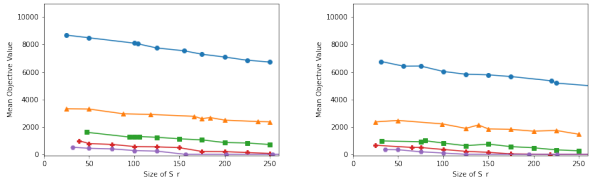

Figure 8: Mean detection time vs $|S_r|$ in the UVA pre-COVID (left) and COVID (right) networks, for different transmission probability values.

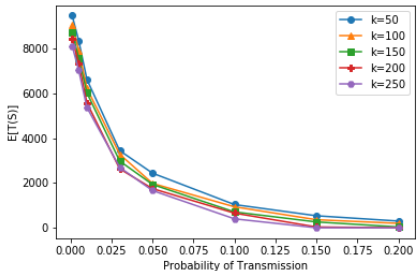

Figure 9: Mean detection time vs the transmission probability for different values of k for the UVA pre-COVID network.