Detecting Sources of Healthcare Associated Infections

Hankyu Jang[1], Andrew Fu[2], Jiaming Cui[3], Methun Kamruzzaman[4],
B. Aditya Prakash[3], Anil Vullikanti[2,4], Bijaya Adhikari[1], *Sriram V. Pemmaraju[1]
[1] Department of Computer Science, University of Iowa [2] Department of Computer Science, University of Virginia
[3] College of Computing, Georgia Institute of Technology [4] Biocomplexity Institute, University of Virginia
for the CDC MInD-healthcare group

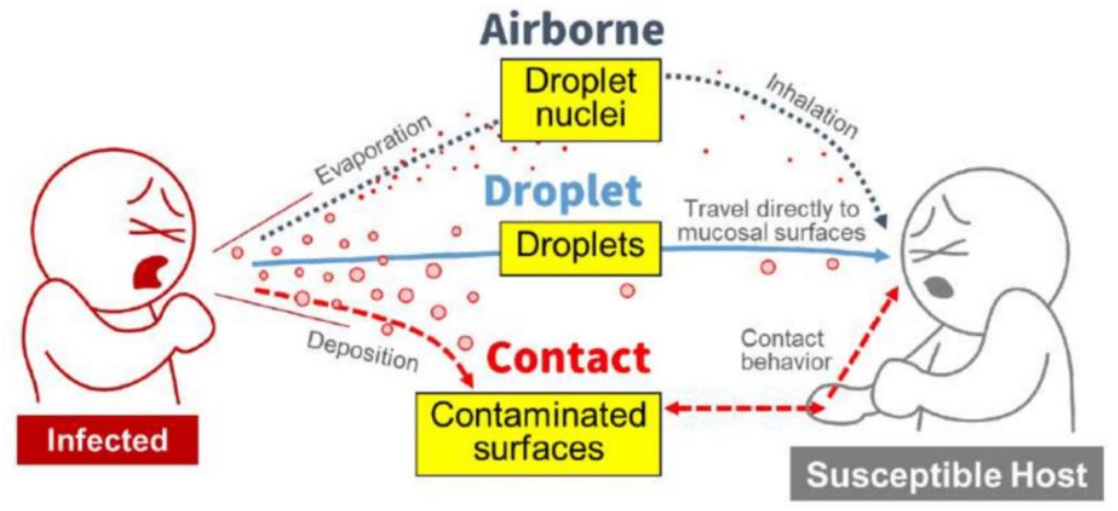
University of Iowa
comp|epi
computational epidemiology research

Introduction

Healthcare-associated infections (HAIs): infections that spread in healthcare settings

- Each year, roughly 4% of patients in the US are diagnosed with HAI
- Immunocompromised patients are at risk of HAI, and infections can lead to severe outcomes

Common HAIs, such as Methicillin-resistant Staphylococcus aureus (MRSA) infection or Clostridioides difficile infection (CDI) spread via ***contact***



When some HAIs are detected, a lot of effort is invested into rapidly identifying the source of infection. This corresponds to the classical ***source detection*** problem. Previous works in source detection are restricted to disease models that spread via person-to-person contacts. However, these methods are not suitable for infections where environment plays an important role.

Hence, source detection problem remains open for HAIs, and is the focus of our paper.

Data

- Daily interactions between healthcare personnel (HCP), patients, and locations
- 31 daily snapshots each of the datasets

Hospital	Number of nodes	Number of edges (/ day)	Interactions captured in
UIHC[1] whole graph	10.4 K	13.8 K	UIHC, the whole hospital
UIHC unit	0.8 K	0.5 K	A unit in UIHC with the most number of CDI cases
UVA[2] pre COVID	2.4 K	0.4 K	Cardiology department, 2011
UVA post COVID	0.9 K	0.4 K	Cardiology department, 2020
Carilion	2.3 K	29.6 K	Carilion Hospital in VA. Public dataset

1 UIHC: University of Iowa Hospitals and Clinics
2 UVA: University of Virginia Hospital

Sponsors

NSF
NIH National Institutes of Health
CDC Centers for Disease Control and Prevention

- UIowa startup
- Georgia Tech
- Facebook faculty research award

Source Detection Problem

Given a temporal network $\mathcal{G}$=(G_0,G_1,…,G_(T−1)),
a load sharing model M, and a set of observed cases

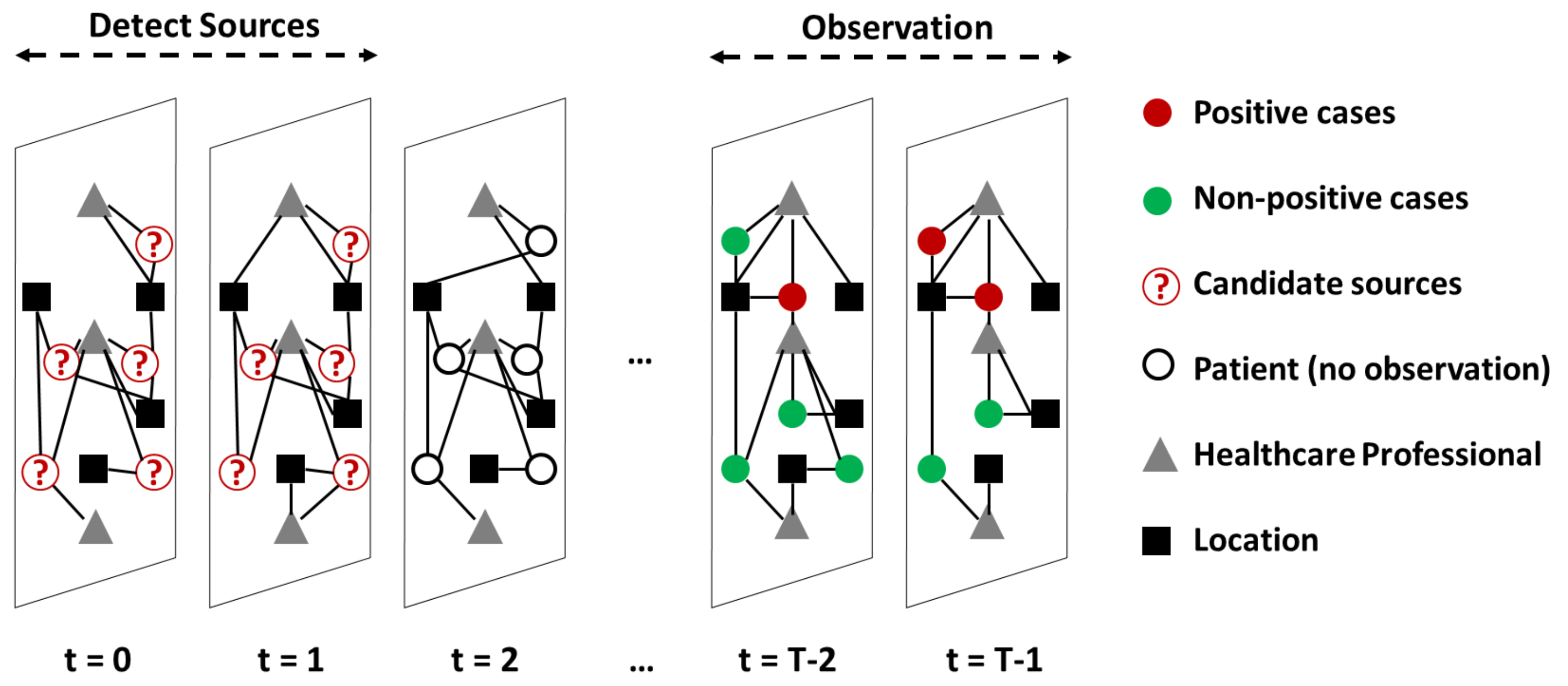
Find a source set S

- that makes $g(S)$ large
- while keeping $f(S)$ small

$\alpha(v, S)$: Probability of v that get infected according to M due to disease starting at S

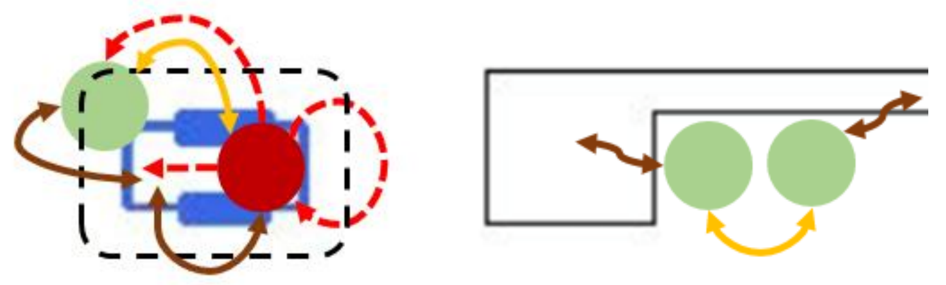
$g(S) = \sum_{v \in Pos} \alpha(v,S)$: Expected number of infections among ● **Positive cases**

$f(S) = \sum_{v \in Neg} \alpha(v,S)$: Expected number of infections among ● **Non-positive cases**



Load Sharing Model

$$L_y(t+1) = \underbrace{(1-d)L_y(t)}_{\text{Load remaining after natural decay}} - \underbrace{\sum_{x:\{x,y\} \in E_t} \rho_{y,x} \cdot L_y(t)}_{\text{Outgoing load}} + \underbrace{\sum_{x:\{x,y\} \in E_t} \rho_{x,y} \cdot L_x(t)}_{\text{Incoming load}} + \underbrace{I_{inf} \cdot q}_{\text{Shedding}}$$



Problem Formulation

Source Detection Positive-Negative Partial Set Cover (SD±PSC)

Given

- a temporal network $G = (G_0, G_1, \dots, G_{T-1})$
- a load sharing model M
- an observed positive set Pos in time $T-2$ and $T-1$

Find a source set S^* in time 0 and 1

That minimizes $\underbrace{\sum_{v \epsilon Pos}(1-\alpha(v,\ S))}_{\text{Expected number of positive cases not infected by an infection starting at source set S}} + \underbrace{\sum_{v \epsilon Neg}\ \alpha(v,\ S)}_{\text{Expected number of negative cases infected by an infection starting at source set S}}$

Source Detection Positive-Negative Knapsack (SD±KNAP)

Given

- a temporal network $G = (G_0, G_1, \dots, G_{T-1})$
- a load sharing model M
- an observed positive set Pos in time $T-2$ and $T-1$
- parameters $k_{T-2}, k_{T-1} \in \mathbb{R}^+$

Find a source set S^* in time 0 and 1

That maximizes $g(S)$

- such that S satisfies constraints $f_{T-2}(S) \le k_{T-2}$ and $f_{T-1}(S) \le k_{T-1}$

Source Detection Positive-Negative Ratio (SD±RATIO)

Given

- a temporal network $G = (G_0, G_1, \dots, G_{T-1})$
- a load sharing model M
- an observed positive set Pos in time $T-2$ and $T-1$
- parameters $\gamma_{T-2}, \gamma_{T-1} \in \mathbb{R}^+$

Find a source set S^* in time 0 and 1

That maximizes $\dfrac{g(S)}{\gamma_{T-2} \cdot f_{T-2}(S) + \gamma_{T-1} \cdot f_{T-1}(S)}$

Note: the source detection period and the observation period is 2 timesteps, but the paper considers the general problem.
The objective function is a simple and natural model for the Source Detection problem. However, ***no reasonable approximation exists*** for the problem. Due to the hardness of the problem, we present two computationally tractable surrogates for the problem.

Submodularity

Set function $f: 2^V \to \mathbb{R}$ is submodular if it satisfies

$$f(S \cup \{e\}) - f(S) \ge f(T \cup \{e\}) - f(T)\ ,\ S \subseteq T \subseteq V,\ e \in V \backslash \mathrm{T}$$

The core of our contribution is showing $g(S)$, $f(S)$ and $f_t(S)$ are monotone and submodular set functions

- The key aspect is showing that if
  - *loads at nodes are monotone, submodular functions of the source set*
  - *the dose response function is concave*
  - then $g(S)$ is submodular
- Proof uses 'coupling' technique
  - We couple the stochastic decisions made from 4 source sets
  - $S, S+\{v\}, Q, Q+\{v\}$, where $S \subseteq Q$ and $v \notin Q$
- The submodularity in the objective functions allows access to various algorithmic approaches
  - **SD±KNAP** : Adapted from Bilmes-Iyer gradient ascent framework and Azar-Gamzu multiplicative update
  - **SD±RATIO** : Adapted from Bai et. al, greedy for maximizing ratio of submodular functions

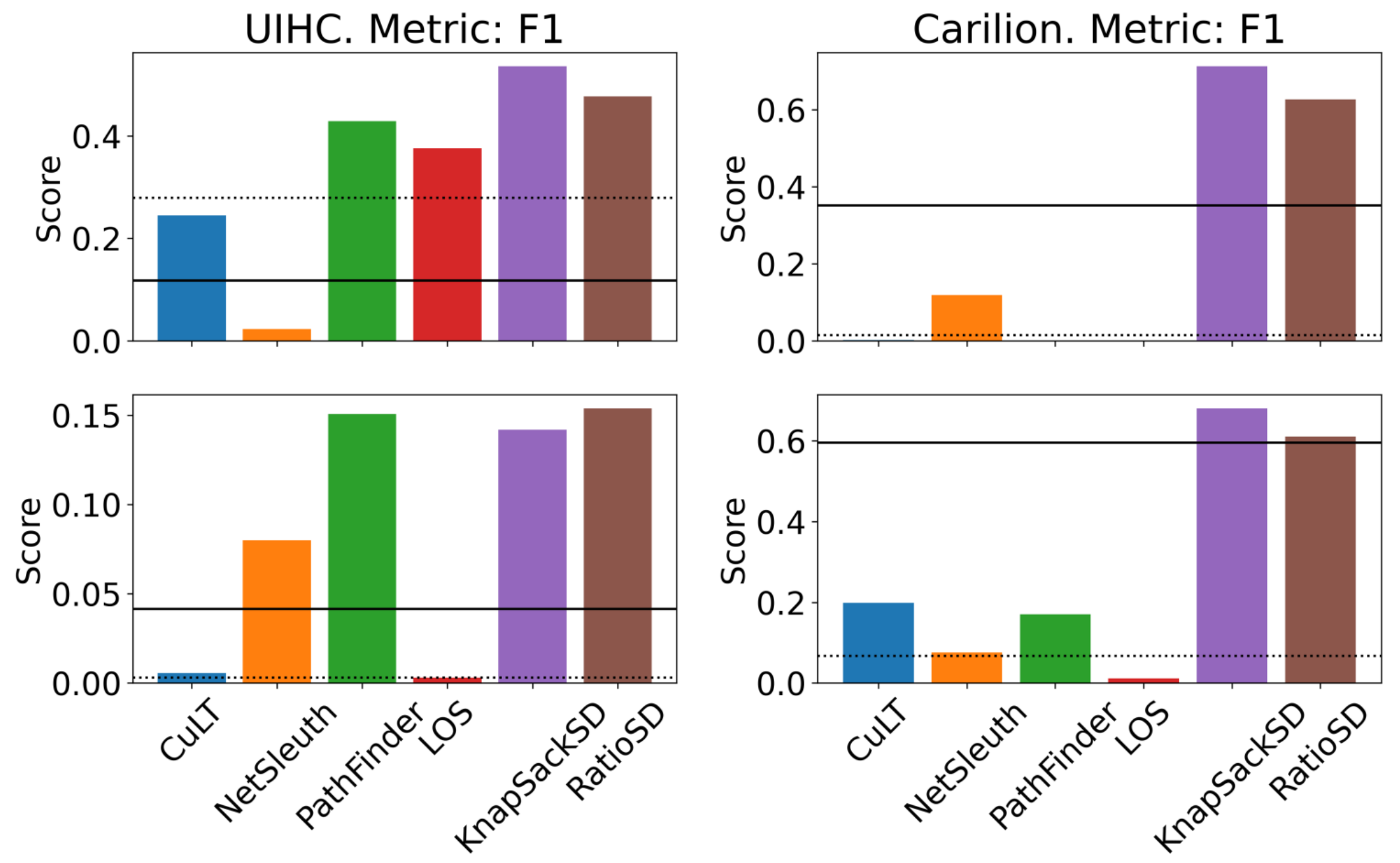
Experiments

Set up: Simulate outbreaks from randomly selected sources S_{GT}, and record observations Pos and Neg
Then detect sources S_M using our algorithms and baselines
Evaluation: A straightforward metric is to measure intersection of S_{GT} and S_M. However, it is impossible for any algorithm to do well w.r.t. this metric due to the stochasticity of the load sharing model. Hence, we use two other natural metrics

- Pos, Neg ***Overlap***: Measure overlap between "Pos and Neg" and "the positive set and negative set caused by outbreaks starting from S_M"
- S_{GT} ***Distance***: Compute distance between S_{GT} and S_M

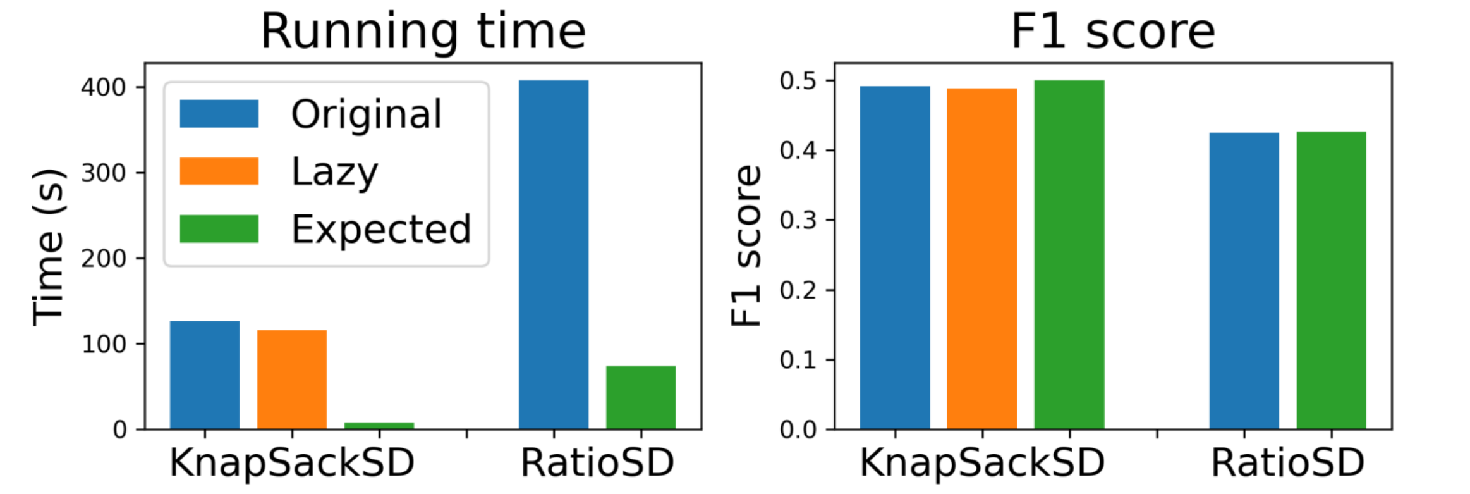
Results

Pos, Neg ***Overlap***



Speed up: f and g are ***stochastic***, which requires a substantial number of simulations to get good estimates. Hence, we propose ***expected load propagation*** heuristic

$$\mathbb{E}[L_y(t+1)] = (1-d)\mathbb{E}[L_y(t)] + \sum_x (\rho_{y,x}\mathbb{E}[L_x(t)] - \sum_x \rho_{x,y}\mathbb{E}[L_y(t)]) + \underbrace{q \cdot p(\mathbb{E}[L_y(t)])}_{\text{Probabilistic shedding}}$$



Conclusion

We consider the well-known source detection problem, but for a new and fundamentally different disease-spread model called the load sharing model. We show that a natural formulation of the problem is intractable, but present two tractable formulations. The tractability of these formulations critically depends on the submodularity of the expected number of infections as a function of the source set. We show submodularity despite not being able to use standard techniques such as the "live edge" technique. We design scalable algorithms that leverage submodularity and speed these up significantly by using a novel heuristic. Experiments on real and simulated outbreaks on three different hospital data show significant advantages of our approach over the baselines.