[AAAI 2023]

Detecting Sources of Healthcare Associated Infections

Presenter: Hankyu Jang

Co-authors: Andrew Fu, Jiaming Cui, Methun Kamruzzaman,
B. Aditya Prakash, Anil Vullikanti, Bijaya Adhikari, *Sriram V. Pemmaraju

University of Iowa
comp|epi
computational epidemiology research

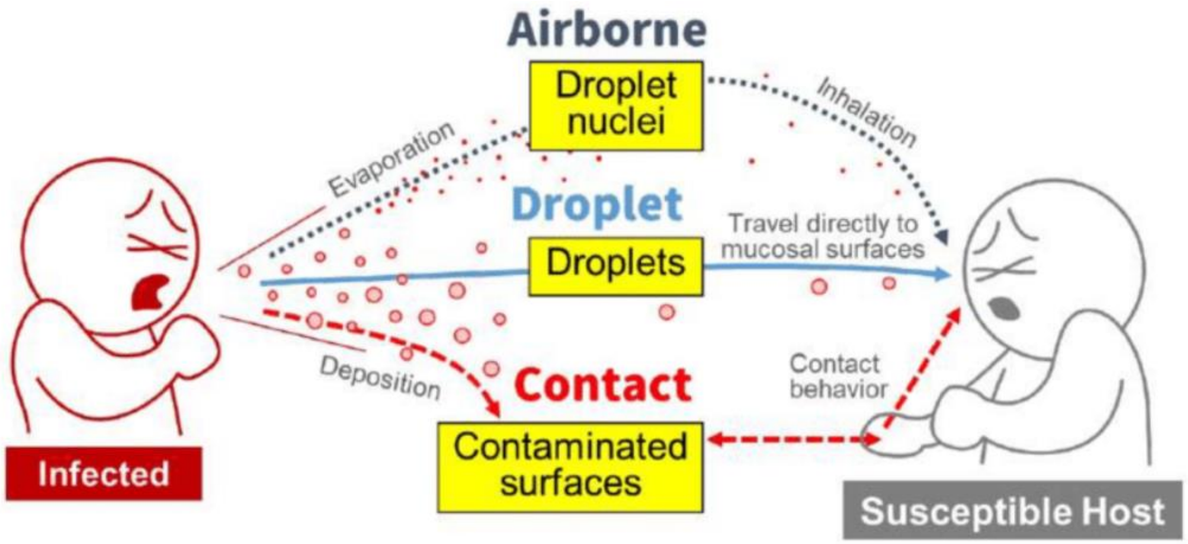
Georgia Tech

University of Virginia

The University of Iowa

Healthcare associated infections

- *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***

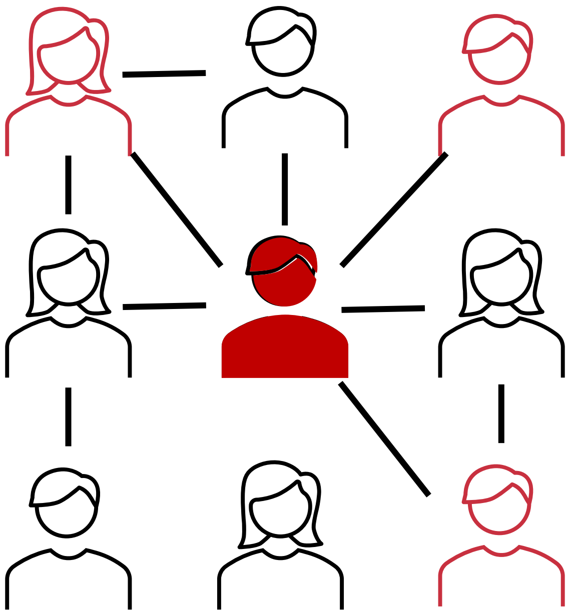


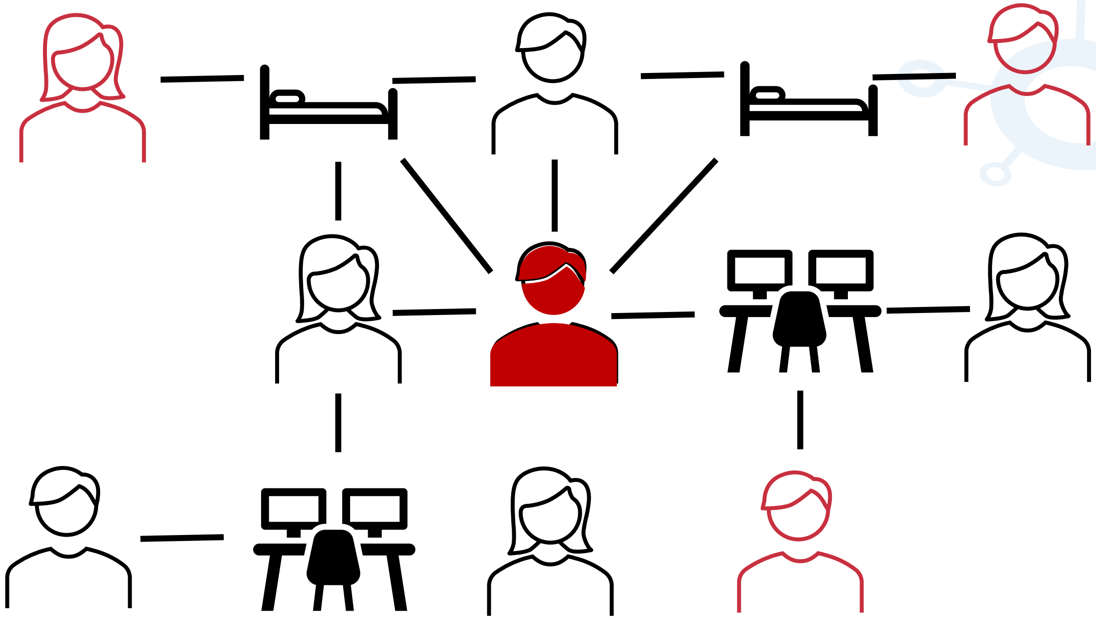
[-]

[*] CDC, "Healthcare-associated infections (hais)," https://www.cdc.gov/winnablebattles/report/HAIs.html.
[-] Gameiro Silva, M. An analysis of the transmission modes of COVID-19 in light of the concepts of Indoor Air Quality. 2020

Motivation

- 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 [+,-,=]





Source detection problem remains open for HAIs, and is the focus of our paper

[-] Shah, D. and Zaman, T. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. SIGMETRICS Perform. Eval. Rev 2010
[=] Lappas, T.; Terzi, E.; Gunopulos, D.; and Mannila, H. Finding Effectors in Social Networks. KDD 2010
[+] ***Prakash BA***, Vreeken J, Faloutsos C. Efficiently spotting the starting points of an epidemic in a large graph. KAIS 2014

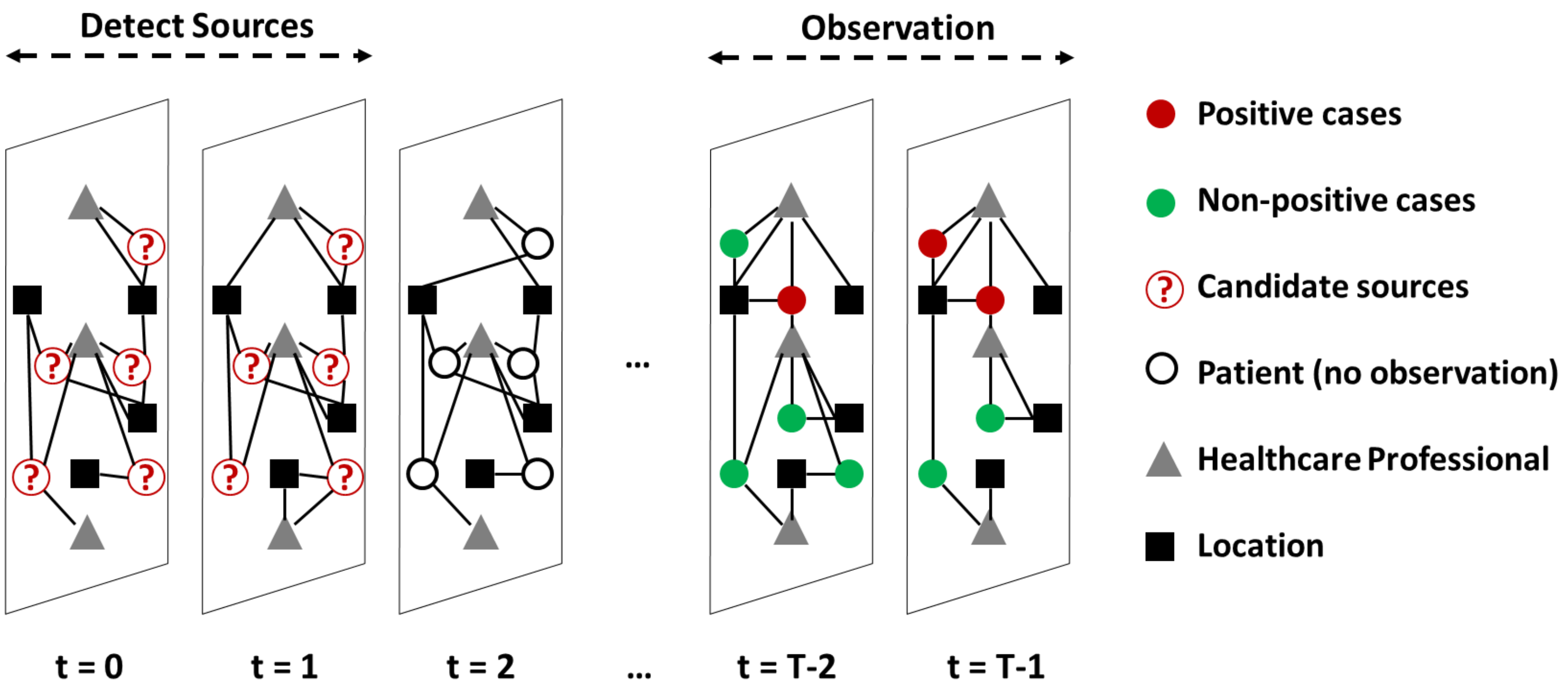
The source detection problem

- Given a temporal network $\mathcal{G} = (G_0, G_1, \dots, G_{T-1})$, 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**



Background: load sharing model

- Traditional compartmental models model disease spread via person-to-person contact
- Recently, disease models that take into account the role of environments were proposed [+, -, =]
- Load sharing model [+]

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

[-] Li S, Eisenberg JN, Spicknall IH, Koopman JS. Dynamics and control of infections transmitted from person to person through the environment. American journal of epidemiology. 2009
[=] Plipat N, Spicknall IH, Koopman JS, Eisenberg JN. The dynamics of methicillin-resistant Staphylococcus aureus exposure in a hospital model and the potential for environmental intervention. BMC infectious diseases. 2013
[+] ***Hankyu Jang***, S. Justice, P. M. Polgreen, A. M. Segre, D. K. Sewell, and ***S. V. Pemmaraju***, "Evaluating Architectural Changes to Alter Pathogen Dynamics in a Dialysis Unit," IEEE/ACM ASONAM 2019

Problem formulation. **SD±PSC**

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 $\underbrace{\sum_{v \epsilon Pos}(1-\alpha(v,\ S))}$ + $\underbrace{\sum_{v \epsilon Neg}\ \alpha(v,\ S)}$

Expected number of ***positive*** cases ***not infected*** by an infection starting at source set S

Expected number of ***negative*** cases ***infected*** by an infection starting at source set S

The objective function is a simple and natural model for the Source Detection problem

However, ***no reasonable approximation exists*** for the problem

Proof of hardness of approximation is in the paper

Tractable problem formulation. **SD±KNAP**

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 $f_{T-2}(S) \leq k_{T-2}$ and $f_{T-1}(S) \leq k_{T-1}$

Tractable problem formulation. **SD±RATIO**

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

These tractable problem formulations depend on the ***submodularity*** of the functions f and g

Submodularity

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

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

- 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 (i) *loads at nodes are monotone, submodular functions of the source set* and *(ii) 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 [-] Azar-Gamzu, multiplicative update [=]
  - SD±RATIO : Adapted from Bai et. al, greedy for maximizing ratio of submodular functions [*]

[+] Mossel E, Roch S. On the submodularity of influence in social networks. ACM STOC 2007
[=] Azar Y, Gamzu I. Efficient submodular function maximization under linear packing constraints. ICALP 2012
[-] Iyer RK, Bilmes JA. Submodular optimization with submodular cover and submodular knapsack constraints. NIPS 2013
[*] Bai W, Iyer R, Wei K, Bilmes J. Algorithms for optimizing the ratio of submodular functions. ICML 2016

Data

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

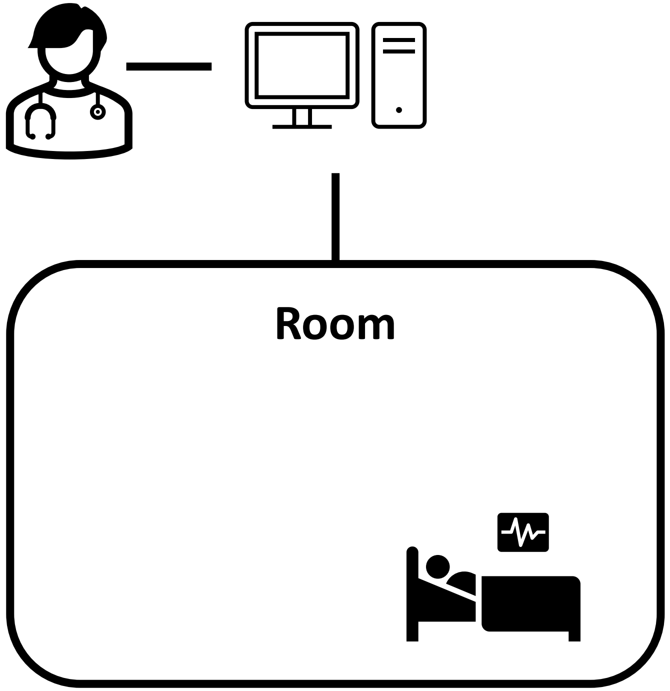
Hospital	Number of nodes	Number of edges (/ day)	Description
UIHC[1] whole graph	10.4 K	13.8 K	Interactions captured in UIHC, the whole hospital
UIHC unit	0.8 K	0.5 K	A unit in UIHC with the most number of CDI cases
UVA[2] pre COVID	2.4 K	0.4 K	Interactions recorded in Cardiology department, 2011
UVA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

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

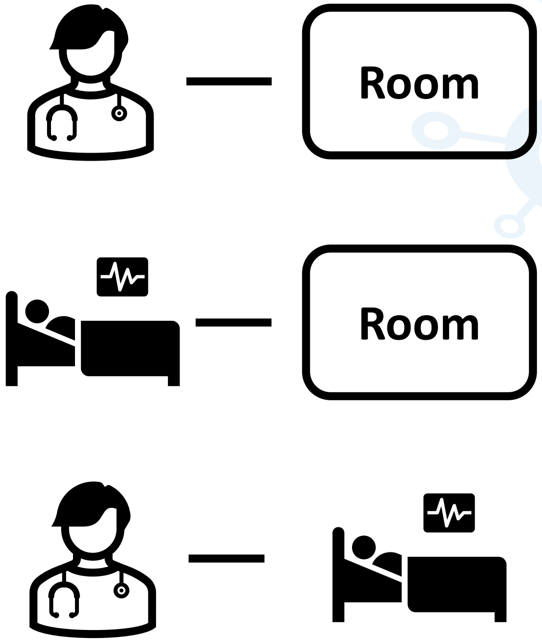
UIHC contact network

- HCP mobility: HCP terminal logins data
- Patient mobility: Admission-discharge-transfer (ADT) data

Nodes



Edges



Experiments

Experiment set up

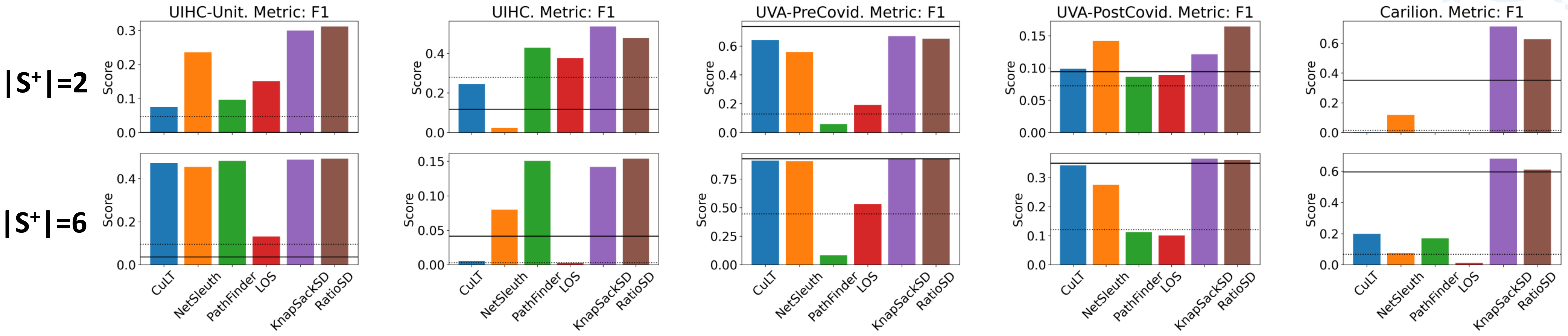
- Simulate outbreaks from randomly selected sources S_{GT}, and record observations Pos and Neg
- Then detect sources S_M using our algorithms and baselines

Evaluation

- A straightforward metric is to measure intersection of S_{GT} and S_M
  - In general, it is impossible for any algorithm M to do well w.r.t. this metric
  - S_{GT} may do poor in explaining Pos and Neg
- 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. Metric: Pos, Neg ***Overlap***

- KnapsackSD and RatioSD consistently outperform all baselines
- Other baselines, e.g., CuLT [+] and NetSleuth [-] are inconsistent
  - Baselines do not capture instances where multiple pathways plays role in infections
  - In such settings, their performance may drop



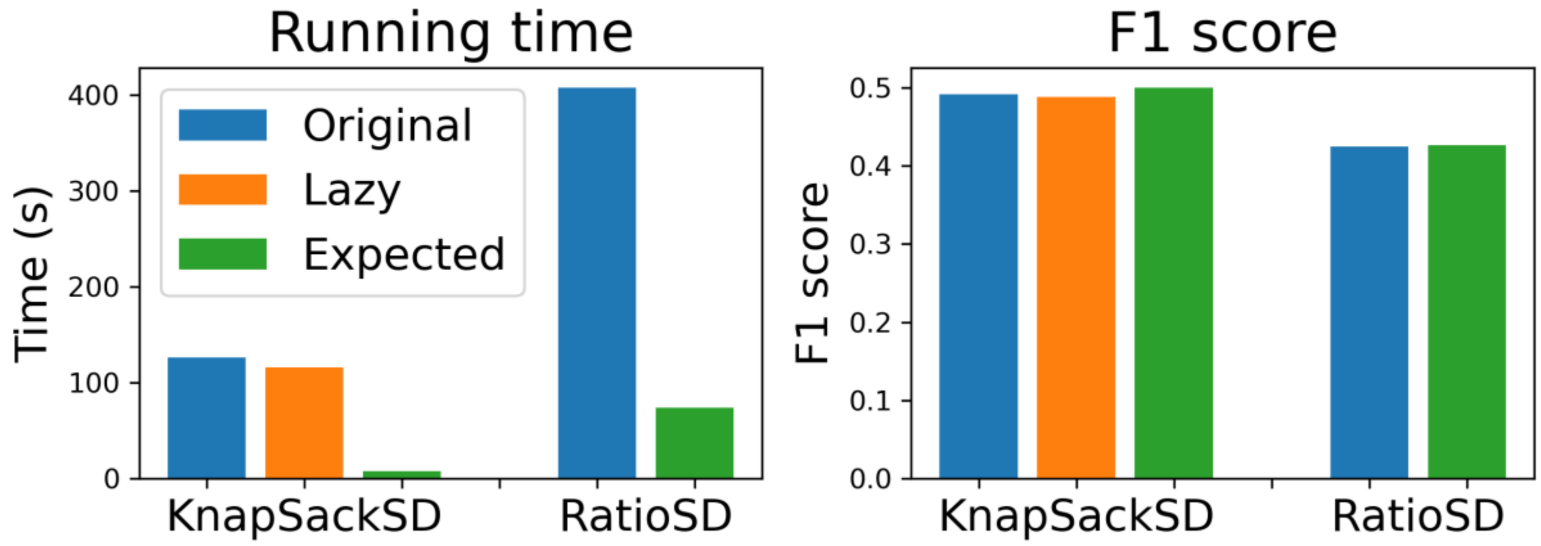
[-] ***Prakash BA***, Vreeken J, Faloutsos C. Spotting culprits in epidemics: How many and which ones?. ICDM 2012
[+] Rozenshtein P, Gionis A, ***Prakash BA***, Vreeken J. Reconstructing an epidemic over time. KDD 2016

Speed up via expected load propagation

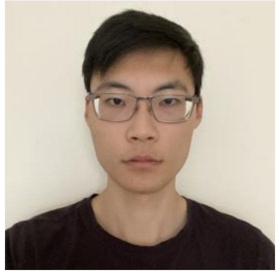
- Since f and g are ***stochastic***, which requires a substantial number of simulations to get good estimates
- We propose ***expected load propagation*** heuristic

$$\mathbb{E}[L_y(t+1)] = \underbrace{(1-d)\mathbb{E}[L_y(t)]}_{\text{Load remaining after natural decay}} + \underbrace{\sum_x (\rho_{y,x}\mathbb{E}[L_x(t)]}_{\text{Incoming load}} \underbrace{- \sum_x \rho_{x,y}\mathbb{E}[L_y(t)])}_{\text{Outgoing load}} + \underbrace{q \cdot p(\mathbb{E}[L_y(t)])}_{\text{Probabilistic shedding}}$$

Pos, Neg **Overlap**



Thank you!

Hankyu Jang	Andrew Fu	Jiaming Cui	Methun Kamruzzaman	B.Aditya Prakash	Anil Vullikanti	Bijaya Adhikari	* Sriram V. Pemmaraju
U Iowa	U Virginia	Georgia Tech	U Virginia	Georgia Tech	U Virginia	U Iowa	U Iowa
							

Sponsors

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

UIowa startup, Georgia Tech,
Facebook faculty research award

Collaborative work

Georgia Tech
University of Virginia
The University of Iowa

* **Corresponding author:** Sriram V. Pemmaraju. Department of Computer Science, University of Iowa. sriram-pemmaraju@uiowa.edu