[ICDM 2021 – IEEE International Conference on Data Mining]

RISK-AWARE TEMPORAL CASCADE RECONSTRUCTION TO DETECT ASYMPTOMATIC CASES

Presenter: Hankyu Jang

Co-authors: Shreyas Pai, Bijaya Adhikari, Sriram V. Pemmaraju

*Funded by CDC MInD Healthcare Network grants and NSF grant

comp|epi
computational epidemiology research

The University of Iowa

Motivation

- For many infections, asymptomatic cases present a major obstacle to precisely understand how the infection is spread
  - COVID-19: a control strategy geared towards asymptomatic infection is regarded as the Achilles' Heel of control strategy [*]
  - C. diff infection (CDI): there is evidence that a substantial fraction (up to 10%) of patients admitted to a healthcare facility are asymptomatic C. diff carriers [-, +]

There is a need for detecting asymptomatic cases!

However, often we don't have ground truth data on the asymptomatic cases, because they don't get tested.

Therefore, we need a method to ***detect asymptomatic cases*** and to ***evaluate them***.

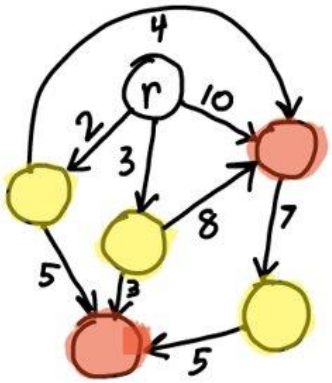
Often, missing infections problem is solved via ***directed Steiner tree*** problem.

[*] M. Gandhi et al, "Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19," N. Engl. J. Med 2020
[-] S. Leekha et al., "Asymptomatic Clostridium difficile colonization in a tertiary care hospital: Admission prevalence and risk factors," AJIC 2013
[+] L. Kyne et al., "Asymptomatic Carriage of Clostridium difficile and Serum Levels of IgG Antibody against Toxin A," N. Engl. J. Med 2000

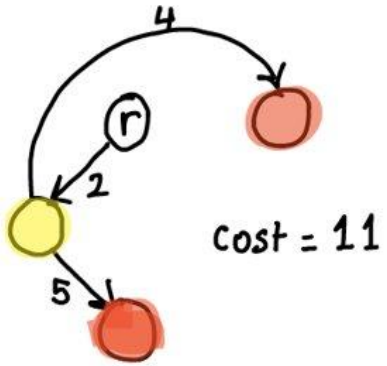
Background – directed Steiner tree (DST)

- The ***directed Steiner tree (DST)*** problem
  - INPUT: A directed graph $G = (V, E)$, an edge weight $w(e)$ for each edge e in E, a special vertex r (root) and a set S of special vertices (terminals)
  - OUTPUT: A directed tree T rooted at r, spanning all terminals S that minimizes

$$\sum_{e \in T} w(e)$$



G = (V, E)



T = (V', E')

DST	Missing infection problem
Input graph	Contact network
Root	Infection source
Terminal	Observed infections
Edge weight w(e)	likelihood of transmission Low w(e) -> high likelihood
Output tree	Infection cascade
Intermediate nodes	Missing infections

Related works – Infer missing infections using DST

- State-of-the-art methods proposed to infer missing infection uses DST [*, -, +]

Limitation: individual susceptibility is ignored

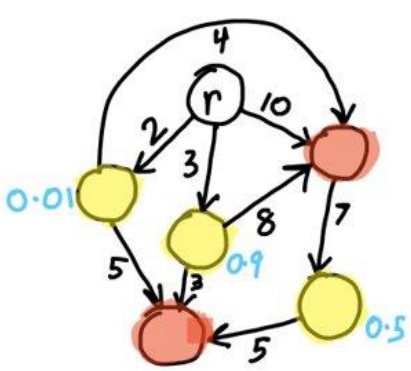
- Can we take into account both the disease-flow and the individual risk?
- What if we had a prior knowledge of ***the likelihood of*** each node being colonized?
- Could DST be optimized to keep highly-likely asymptomatic nodes in the solution?

[*] P. Rozenshtein et al., "Reconstructing an Epidemic Over Time," KDD 2016
[-] H. Xiao et al., "Reconstructing a cascade from temporal observations," SDM 2018
[+] H. Xiao et al., "Robust Cascade Reconstruction by Steiner Tree Sampling," ICDM 2018

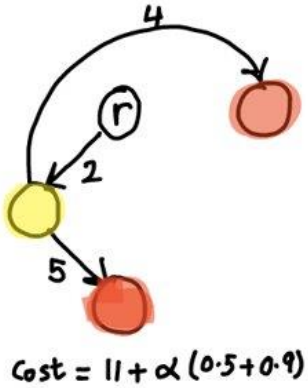
Solution approach and contributions

- We formulate asymptomatic case detection problem as a ***Directed Prize-Collecting Steiner Tree*** problem (Directed PCST)
  - **INPUT**: A directed graph $G = (V, E)$, an edge weight $w(e)$ for each edge e in E, a special vertex r (root) and a set S of special vertices (terminals) and a ***node weight*** $p(v)$ ***for*** $v \in V$
  - **OUTPUT**: A directed tree T rooted at r, spanning all terminals that minimizes

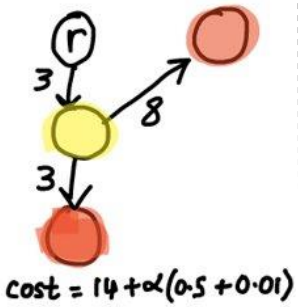
$$\sum_{e \in T} w(e) + \alpha \sum_{v \notin T} p(v)$$



G = (V, E)



T = (V', E')



T = (V', E')

Directed PCST	Missing infection problem
Input graph	Contact network
Root	Infection source
Terminal	Observed infections
Edge weight w(e)	Likelihood of transmission Low w(e) -> high likelihood
Output tree	Infection cascade
Intermediate nodes	Missing infections
Node weight p(v)	***The likelihood of*** node being an asymptomatic

Problem formulation

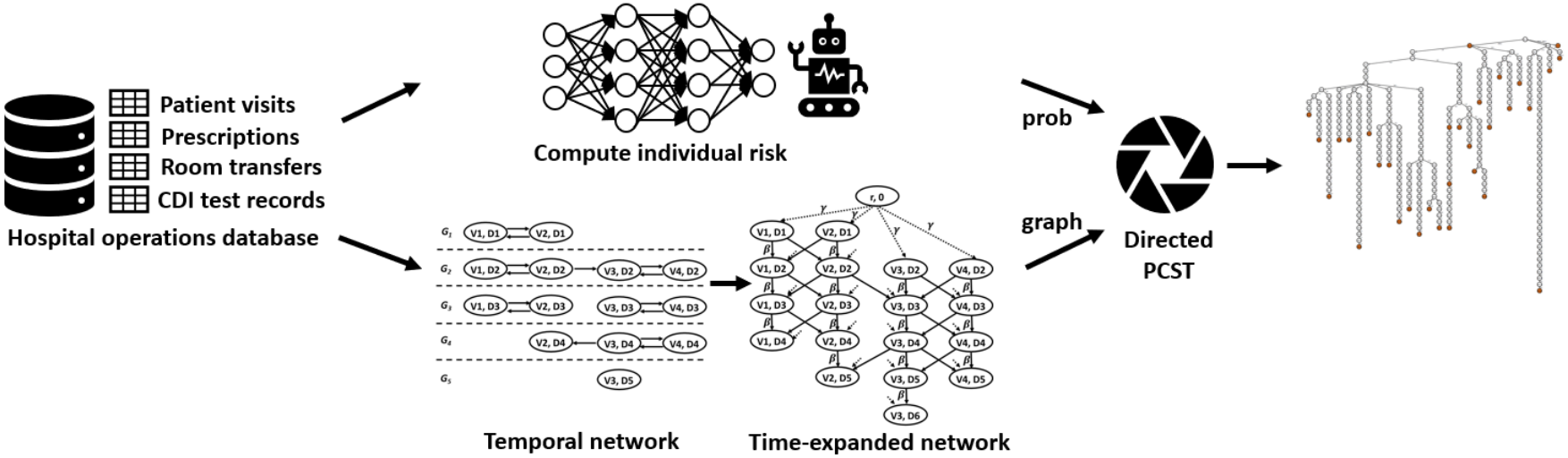
ASYMPTOMATIC CASE DETECTION
Given a temporal network $\mathcal{G} = (G_1, G_2, \ldots, G_T)$ and a sequence $(S_1, S_2, \ldots, S_T)$ of observed cases, find the asymptomatic cases $\mathcal{A} = \cup_{i=1}^{T} A_i$.

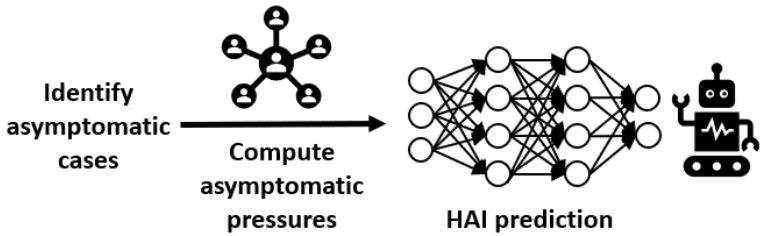
DIRECTED PRIZE-COLLECTING STEINER TREE (DIRECTED PCST)
Given $G_S(V, E, r, S, W_e, W_v)$ and a parameter $\alpha > 0$, find a tree $T^*(V^*, E^*)$ rooted at r and spanning terminal set S, such that

$$T^* = \arg\min_{T} \sum_{(a,b)\in E(T)} W_e(a,b) + \alpha \cdot \sum_{a \in V \setminus V(T)} W_v(a) \quad (1)$$

Solution approach and contributions

- We formulate asymptomatic case detection problem as a ***Directed Prize-Collecting Steiner Tree*** problem (Directed PCST)



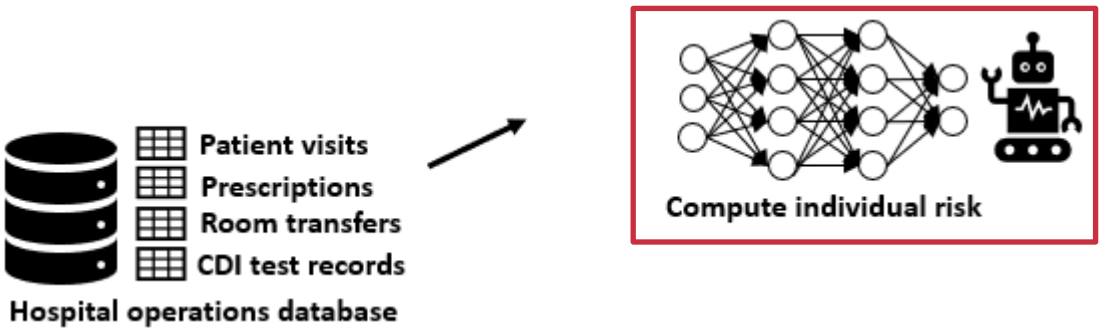


Application

compepi
computational epidemiology research

The University of Iowa

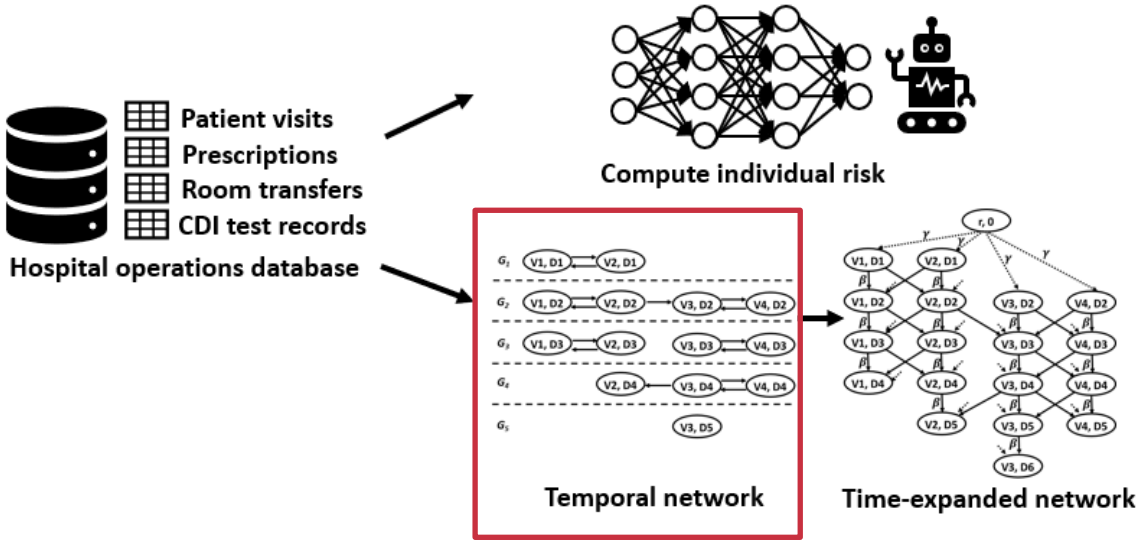
Compute individual risk



Challenge: the data lacks "***ground truth***" labels for asymptomatic carriage

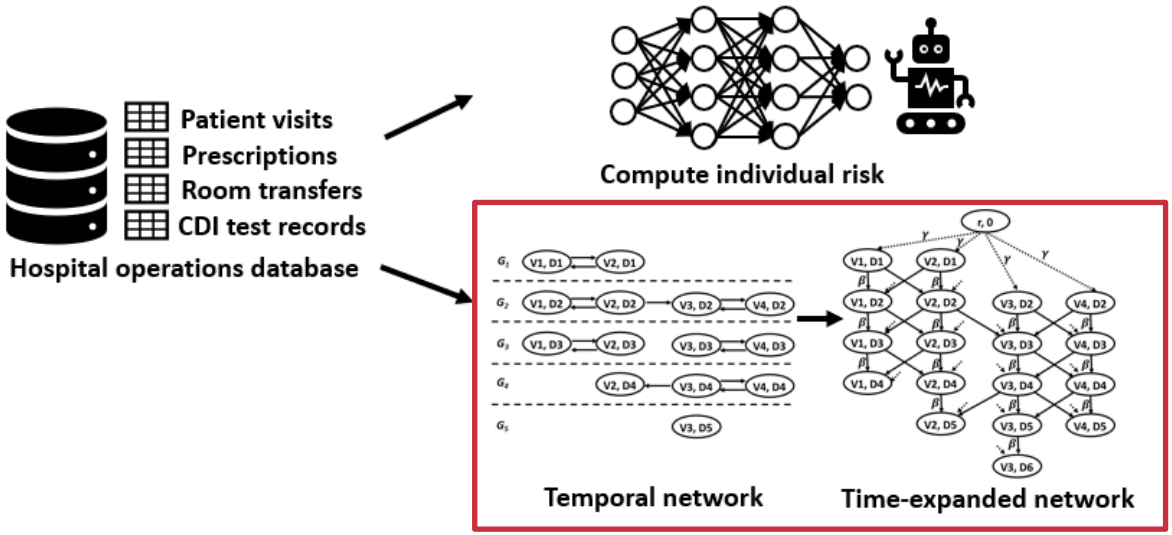
- We make two hypotheses to enable model training
  - Hypothesis 1: asymptomatic cases and CDI cases have similar risk profiles
  - Hypothesis 2: the patient must first be an asymptomatic, then prescribed to high-risk antibiotic to acquire CDI
- After training, we use the output probability as ***the asymptomatic likelihood***

Construct temporal network



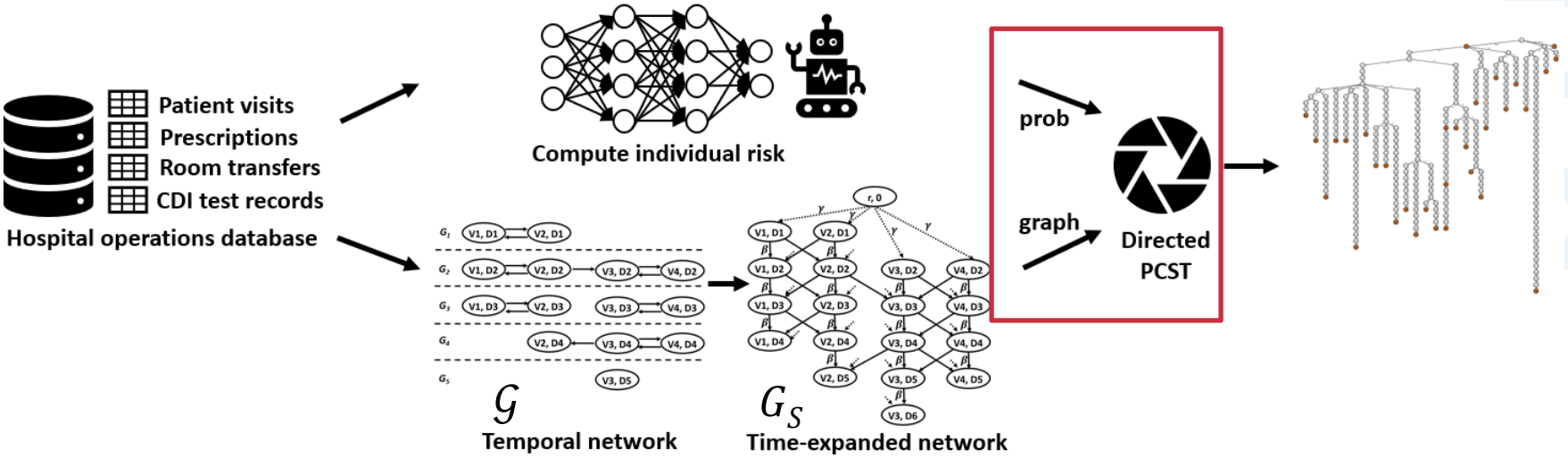
- **Nodes**: a day of a patient visit
- **Node weight**: *the asymptomatic likelihood*
- **Edges**: edge if two patients visited the same unit
- **Edge weight**: physical distance between the two patients in terms of room
- **Observed cases**: CDI positive case

Temporal network -> Time-expanded network



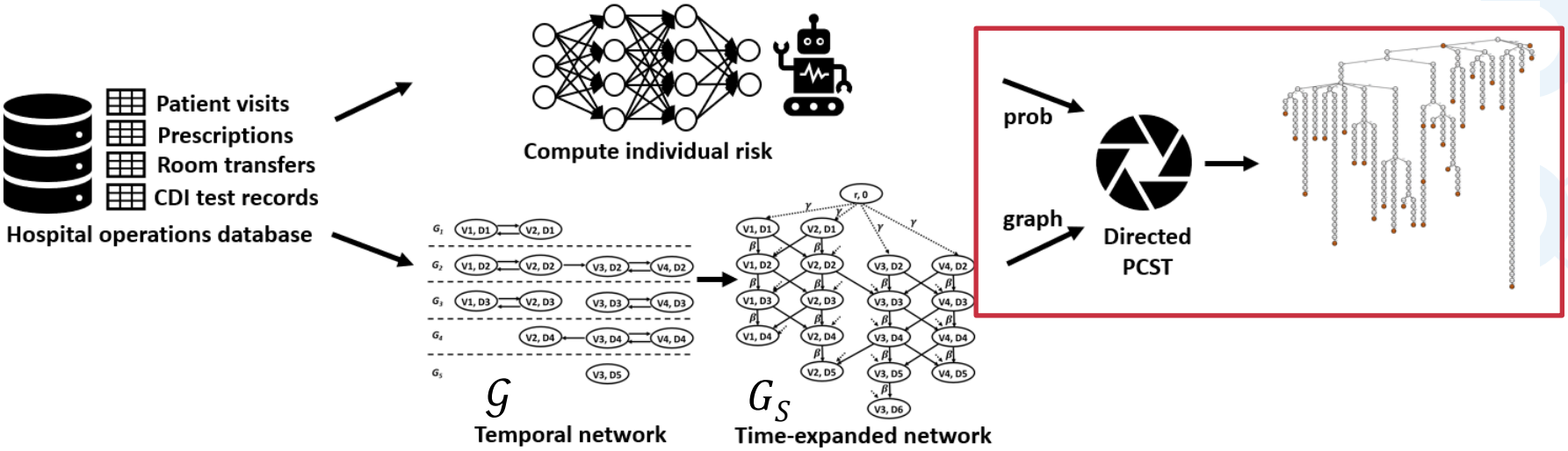
- **Root**: a ***dummy node***. Connect to all the nodes with weight γ.
- **Edges**: add ***cross edges*** for contacts between patients (weight is inherited). Connect nodes from same visit over time with weight β
- **Terminals**: observed cases
- **Node weight**: *the asymptomatic likelihood*

Reduce Directed PCST to DST



- ***Directed PCST*** is computationally very challenging, even to approximate
- Therefore, we reduce ***Directed PCST*** to ***DST*** to obtain scalable algorithms:
  - from G_S we create a new graph G' with *only* edge weights by $W_e(a,b) - \alpha \cdot W_v(b)$
- We show the following:
  - the optimal DST in G' is the optimal directed PCST in G_S
  - the approximation factor is preserved in the reduction

Scalable algorithms for Directed PCST



- We propose three approximation algorithms to solve DST on G'
  - Greedy algorithm [+], linear programming (LP), and minimum cost arborescence (MCA)
- In the solution tree, we interpret the ***nodes in the path*** from the **root** (dummy node) to the **terminals** (CDI cases) as ***asymptomatic cases***

[+] M. Charikar et al., Approximation algorithms for directed steiner problems. J of Algorithms 1999

Experiments

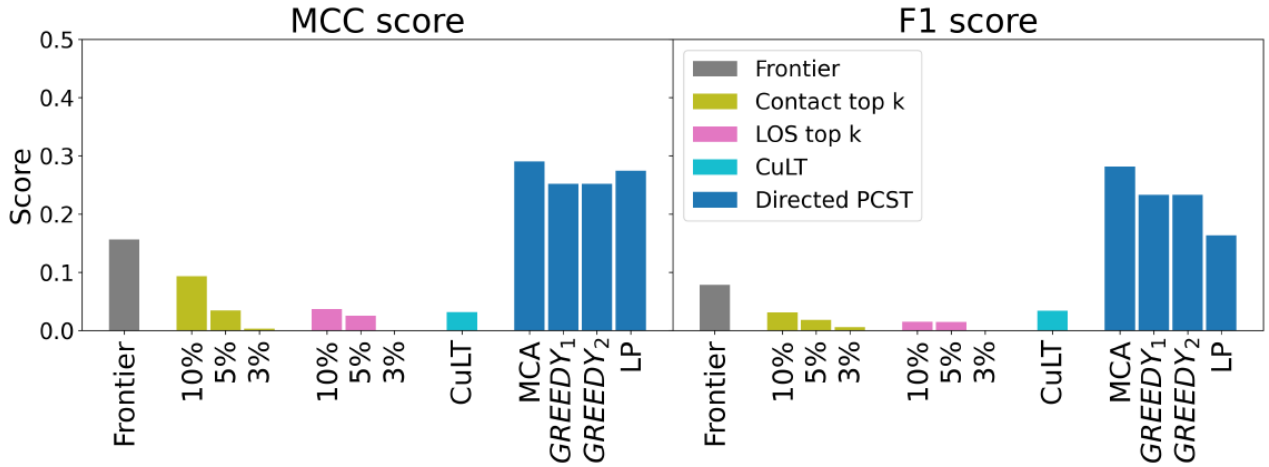
Baseline methods

- **Frontier**: Select the neighbors of the terminal nodes as symptomatic cases
- **Contact top k**: Select top k% high-contact nodes based on the out-degree in the time expanded network
- **Length of stay (LOS) top k**: Select top k nodes based on the LOS
- **CuLT** [*]: state-of-the-art Steiner-tree-based missing infection detection approach. Note that algorithms that CuLT uses are just a special case of our Greedy approaches, where there are no node weights

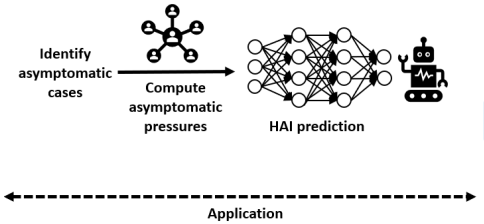
[*] P. Rozenshtein et al., "Reconstructing an Epidemic Over Time," KDD 2016

Performance on the synthetic data

- We use one month of patient data to generate a time-expanded network
  - 20.9 K nodes, 0.5 M edges
- We run biased SIS simulation from multiple sources to obtain a set of observed *symptomatic cases* and a set of *asymptomatic cases*
- We measure success based on overlap of inferred asymptomatic cases and the ground truth asymptomatic cases



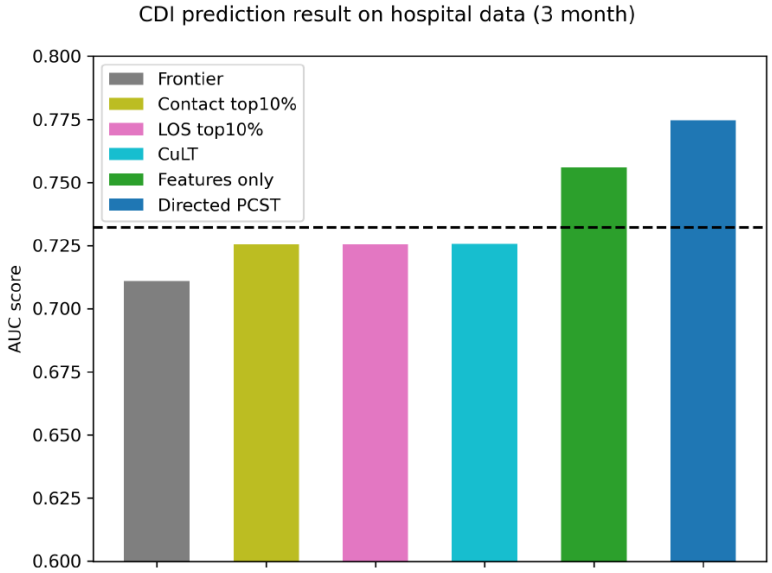
Application: CDI case prediction



- We use the inferred asymptomatic cases to predict the symptomatic CDI cases
- We train a neural network with two types of features
  - Standard risk factors of CDI
  - *asymptomatic pressures*: measures the exposure to the newly identified asymptomatic CDI cases
- We use three month of patient data: 60.9 K nodes, 1.6M edges

  and investigate if adding asymptomatic pressures improves the performance

- Our proposed approach via the Directed PCST problem (in blue) outperform all the baselines
- Adding ***asymptomatic pressures*** from our method improves the symptomatic cases prediction task



comp|epi
computational epidemiology research