Risk-aware Temporal Cascade Reconstruction to Detect Asymptomatic Cases

Hankyu Jang
Dept. of Computer Science
The University of Iowa
Iowa City, IA, USA
hankyu-jang@uiowa.edu

Shreyas Pai
Dept. of Computer Science
The University of Iowa
Iowa City, IA, USA
shreyas-pai@uiowa.edu

Bijaya Adhikari
Dept. of Computer Science
The University of Iowa
Iowa City, IA, USA
bijaya-adhikari@uiowa.edu

Sriram V. Pemmaraju
Dept. of Computer Science
The University of Iowa
Iowa City, IA, USA
sriram-pemmaraju@uiowa.edu

For the CDC MInD Healthcare Network

***Abstract*—This paper studies the problem of detecting *asymptomatic cases* in a temporal contact network in which multiple outbreaks have occurred. For many infections, asymptomatic cases present a major obstacle to obtaining a precise understanding of infection-spread. We show that the key to detecting asymptomatic cases well, is taking into account both individual risk as well as the likelihood of disease-flow along edges. Most related research has ignored the interplay between these dual aspects influencing disease-spread. We take both aspects into account by formulating the asymptomatic case detection problem as a *Directed Prize-Collecting Steiner Tree* (DIRECTED PCST) problem. We present an approximation-preserving reduction from this problem to the *Directed Steiner Tree* problem and use this reduction to obtain scalable algorithms for the DIRECTED PCST problem. Using these algorithms, we solve instances with more than 1.5M edges obtained from both synthetic and actual fine-grained hospital data. On synthetic data, we demonstrate that our detection methods significantly outperform various baselines (with a gain of $3.6\times$). As an application of our methods, we use a measure of exposure to detected asymptomatic *Clostridioides difficile* (C. diff) *infection* (CDI) cases as an additional feature for the important task of predicting symptomatic CDI cases. In this application, our method outperforms all baselines, including those that don't use asymptomatic CDI cases as a feature and those that use other methods for detecting asymptomatic CDI cases. We also demonstrate that the solutions returned by our approach are clinically meaningful by presenting a case study.**

***Index Terms*—asymptomatic cases, C. diff infections, prize-collecting Steiner tree, temporal contact networks**

I. INTRODUCTION

For many infections, e.g., Zika virus disease, malaria, *Methicillin-resistant Staphylococcus aureus* (MRSA) infection, and *Clostridioides difficile* (C. diff) *infection* (CDI), *asymptomatic* cases present a major obstacle to precisely understand how the infection is spread, and they make implementing effective interventions that much more challenging [12], [19], [26]. Indeed, asymptomatic cases are widely believed to play a substantial role in the spread of COVID-19 [4], and asymptomatic transmission of SARS-CoV-2 has been called the "Achilles' heel" of control strategies for COVID-19.

Ideally, we would like to detect asymptomatic individuals and apply infection-control policies (e.g., quarantine, isolation) to them as well. However, detecting asymptomatic cases is challenging for several reasons. First, since asymptomatic cases do not show symptoms (by definition), only costly, blanket surveillance strategies can detect these cases. Second, asymptomatic cases may not have the same risk factors as symptomatic cases, and therefore risk factors discovered for symptomatic cases may not be a valid proxy for asymptomatic cases. Third, from a data mining point of view, it is hard to learn risk factors for asymptomatic cases because "ground truth" data on asymptomatic cases is essentially non-existent.

The focus of this paper is the detection of asymptomatic cases of *healthcare-associated infections* (HAIs). An HAI is an infection that a patient acquires in a healthcare facility while being treated for another condition. At any given time, 1 in 25 patients in the US has an HAI [18]. CDI and MRSA infection are among the most common HAIs [18]. Some of the experimental results we present are for detecting asymptomatic cases of CDI, but our methods are widely applicable. The main novelty and strength of our approach are that it takes into account both individual risk as well as disease-flow through a contact network. Prior work on detecting "missing infections" (e.g., [23], [27], [28]) has largely ignored individual risk. The main takeaway from our results is that both aspects of disease-spread are critical. When evaluated on large-scale synthetic data and actual hospital data, our approach outperforms methods that ignore either the individual risk or disease-flow.

A. Informal Problem Description

Our input consists of a hospital mobility log that tells us time-stamped locations (e.g., hospital rooms) of patients and *healthcare professionals* (HCPs). We represent this mobility log as a temporal network $\mathcal{G} = (G_1, G_2, \ldots, G_T)$, where $G_i = (V_i, E_i, W_i, \mathbf{F}_i)$ is the static graph that captures interactions at time i. At each time i, the edge set E_i represents the interactions between nodes in V_i and W_i is the associated set of edge weights, representing the "strength" of these interactions. $\mathbf{F}_i[v]$ is the attribute vector for node $v \in V_i$ at time i, representing individual risk factors such as demographics, length of stay, prescriptions, etc. We assume that there is a hidden disease-spread process that starts independently from multiple sources at possibly different times. At each time-stamp i, the set of infected nodes $\mathcal{I}_i \subseteq V_i$ get a single chance to infect their healthy neighbors. A distinguishing feature of our model is

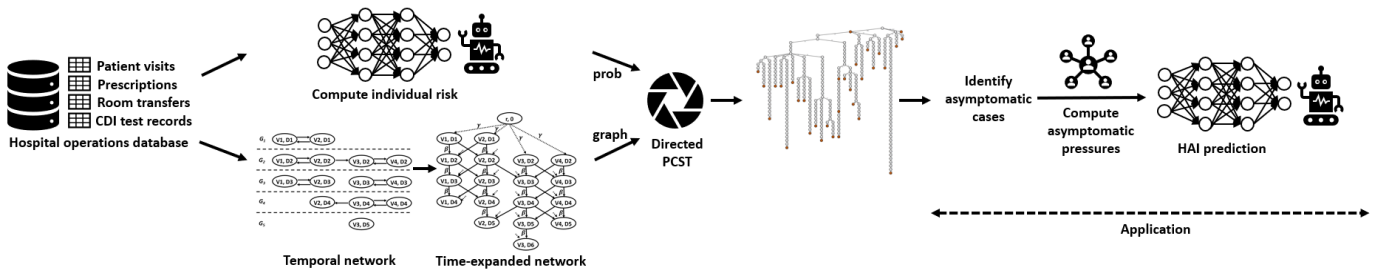

Fig. 1: This schematic shows our overall approach to solving the ASYMPTOMATIC CASE DETECTION problem and applying the solution towards HAI case prediction.

that the attribute vector $\mathbf{F}_i[v]$ influences the likelihood of a node becoming infected. Each infected node also has a single chance to recover. Those nodes which are newly infected and those that fail to recover at time i, are infected at the beginning of time stamp $i+1$. This process continues till time T. Additionally, we are given time-stamped positive test results for an HAI. In other words, for each time i, a subset $S_i \subseteq \mathcal{I}_i$ of the infected nodes are revealed to us and the remaining infected nodes $A_i = \mathcal{I}_i \setminus S_i$ are hidden asymptomatic cases. Our problem can now be stated informally as:

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 $\mathcal{A} = \cup_{i=1}^{T} A_i$.

B. Solution Approach and Contributions

Our overall solution approach to the ASYMPTOMATIC CASE DETECTION problem is shown in Figure 1. We now describe this approach while highlighting our main contributions.

- **Directed Prize-Collecting Steiner Tree formulation:** We model the ASYMPTOMATIC CASE DETECTION problem as the *Directed Prize-Collecting Steiner Tree* (DIRECTED PCST) problem. DIRECTED PCST takes two inputs: (i) a *time-expanded network* that models infection flow and observed infections and (ii) individual patients' risks (probabilities) of being colonized. The output to the DIRECTED PCST problem is a tree that uses a combination of edges likely to permit infection-flow and nodes likely to be asymptomatic cases, thus taking into account these dual aspects of disease-spread. We identify nodes in the output tree that are not observed cases as asymptotic cases. Our work seems to be the first to apply the DIRECTED PCST formulation to problems in disease-spread.
- **Scalable algorithms for DIRECTED PCST:** The DIRECTED PCST is computationally very challenging, even to solve approximately [10]. We present a new approximation-preserving reduction from DIRECTED PCST to the *Directed Steiner Tree* (DST) problem. We then leverage this reduction to present three alternative algorithms for DIRECTED PCST: (i) an approximation algorithm via the greedy DST approximation algorithm of Charikar et al. [5], (ii) a flow-based Linear Programming (LP) relaxation, and (iii) a simple and fast heuristic based on *minimum cost arborescence* (MCA). Using these algorithms, we are able to evaluate our approach for detecting asymptomatic cases on a time-expanded network containing more than 1.6 million edges.
- **Learning individual risk:** One of the inputs we provide to the DIRECTED PCST problem is individual patients' risks of being colonized. Learning these risks is a challenging problem due to the absence of "ground truth" data. We present a hypotheses-driven approach to using patients' attributes such as demographics, length of stay, prescriptions, etc., for learning patients' risks of being an asymptomatic CDI case. Our approach can be generalized to other HAIs.
- **Extensive large-scale evaluation:** We present extensive experimental evaluation of our approach on synthetically generated HAI data overlaid on temporal contact networks obtained from fine-grained mobility data from a large public hospital. Our approaches significantly outperform all the baselines, including CuLT [23], a Steiner-tree based approach that ignores individual risk. Our best performing method achieves an F_1-score of 0.281, while our nearest competitor achieves only 0.078.
- **Application to predicting CDI cases:** We present a novel application of our methods to predicting (symptomatic) HAI cases. Using asymptomatic cases identified by our method, we create new features that we call *asymptomatic pressures*, that measure exposure to asymptomatic cases. We then compare models for HAI prediction that include these asymptomatic pressures against (i) models that don't include these pressures and (ii) models that include these pressures, but computed via other methods (e.g., CuLT). We show that using asymptomatic pressures computed by our method as a feature significantly outperforms all other competitors.

II. PROBLEM FORMULATION

In this section, we formalize the ASYMPTOMATIC CASE DETECTION problem. First, we assume that we have learned a function A from the space of feature vectors $\mathbf{F}_i$ to $[0, 1]$, representing probabilities that nodes (which are not positive HAI cases) are asymptomatically infected. Given that no "ground truth" data is available on asymptomatic infections, this by itself is a non-trivial problem. We address this in Section V for CDI, but in principle our methods can be used for any HAI. Second, we transform the temporal network $\mathcal{G} = (G_1, G_2, \ldots, G_T)$ and observed

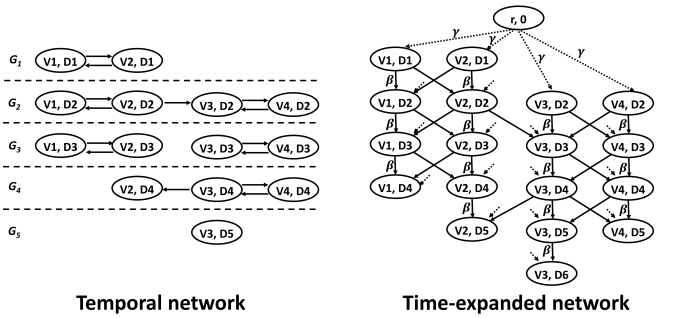

Fig. 2: The temporal graph $\mathcal{G}$ on the left is transformed into the time-expanded network $G_S(V_S, E_S, r, S, W_e, W_v)$ on the right. Even though, in order to avoid clutter, the figure only shows 4 edges leaving node r, there is an edge from r to every node in the graph with weight γ.

case sequence $(S_1, S_2, \ldots, S_T)$ into a time-expanded network $G_S(V_S, E_S, r, S, W_e, W_v)$ with edge weights W_e, node weights W_v, a set $S \subseteq V_S$ of terminals, and a root $r \in V_S$. We describe this transformation below (see Figure 2).

- **Nodes:** Consider V_i, the node set for the time-i contact network G_i. For each node $v \in V_i$, we add two nodes (v, i) and $(v, i+1)$ to V_S. (Note that if $v \in V_i$ and $v \in V_{i+1}$ then $(v, i+1)$ is added only once to V_S.) We use the term *layer* i to denote the subset of all nodes in V_S whose times-stamp label is i.
- **Edges:** For each edge (u, v) in G_i, we create a "cross" edge $((u, i), (v, i+1))$. Additionally, for every $v \in V_i$, we create a "straight" edge $((v, i), (v, i+1))$.
- **Edge weights:** The "cross" edge $((u, i), (v, i+1))$ in E_S inherit its weight from the edge(u, v), i.e., it is assigned weight $W_i(u, v)$. For some parameter, $\beta > 0$, all "straight" edges of the form $((v, i), (v, i+1))$ are assigned weight β. This assignment of edge weights in G_S is denoted by W_e.
- **Node weights:** Each node $(v, i+1)$ in G_s is assigned the probability $A(\mathbf{F}_i[v])$.
- **Terminals:** The set of observed cases $\mathcal{S}$ is designated the set of terminals of the graph G_S.
- **Root:** We add a "dummy" root node r to V_S and connect it to every other node in V_S. For some parameter $\gamma > 0$, we make γ the weight of every edge leaving r. The γ parameter controls the number of connected components in our solution upon removal of r. These connected components are trees and can be interpreted as distinct outbreaks. Larger values of γ will favor few outbreaks in an optimal solution.

An important (and easily verified) observation about G_S is that it is a directed acyclic graph (DAG). This property of G_S will play a crucial role in the efficiency of the algorithms we consider in Section III.

We now formulate a precise version of the ASYMPTOMATIC CASE DETECTION problem as a *Directed Prize-Collecting Steiner Tree* problem (DIRECTED PCST).

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

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

The objective function of the DIRECTED PCST problem aims to balance two weights: one due to *edges included* in the tree and other due to *nodes excluded* from the tree. As a result, an optimal DIRECTED PCST solution T^* uses a combination of low weight edges and high weight nodes. The connection between DIRECTED PCST and the ASYMPTOMATIC CASE DETECTION problem is now natural. Given a tree T^* that is a solution for DIRECTED PCST, we interpret the non-terminal nodes in T^* as likely asymptomatic infections.

The parameter α provides a way of controlling the relative importance of included edge weights versus excluded node weights. A large value of α places more importance on node weights. Setting $\alpha = 0$ yields the DST [5] problem as a special case. While the DIRECTED PCST problem is parameterized by α, the time-expanded network G_S that is input to the problem is parameterized by quantities β and γ. In our experiments, we explore the space of these three parameters.

III. SCALABLE ALGORITHMS FOR DIRECTED PCST

The DIRECTED PCST problem is computationally very challenging. In fact, its special case, the DST problem is also very challenging. Not only is DST NP-complete, it is also difficult to solve approximately (see Halperin and Krauthgamer [10]). While there are constant-factor approximation algorithms for the undirected version of PCST [2], except for the message-passing heuristic [25] (which provides no approximation guarantee), nothing seems to be known for DIRECTED PCST. In fact, this is the situation not just for arbitrary directed graphs, but also for DAGs [30] (see also Theorem 1 in [22]).

We present the following three approaches to solving the DIRECTED PCST. All three approaches depend on an approximation-preserving reduction from DIRECTED PCST to DST, that we provide in Section III-A. (i) We use the greedy DST approximation algorithm of Charikar et al. [5] to approximately solve DIRECTED PCST (Section III-B). (ii) We solve a flow-based LP relaxation of DST [22] (Section III-C). Even though solution returned by the LP is fractional, as we show below, it can still be meaningfully interpreted in the context of the ASYMPTOMATIC CASE DETECTION problem. (iii) We solve the MCA problem on the metric graph induced by the terminal set S and the root r (Section III-D). Even though this approach does not come with a provable approximation guarantee, our experimental results indicate that this is a fast algorithm that outputs near-optimal solution.

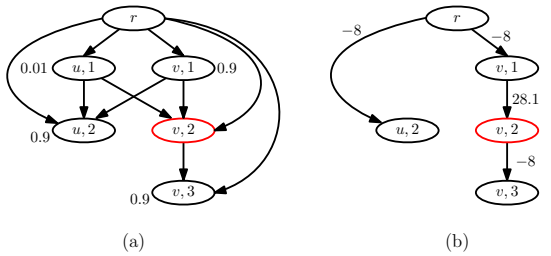

Fig. 3: (a) A time-expanded network G_S with terminal set $S = \{(v, 2)\}$ (shown in red) and node weights is shown. To obtain edge weights assume that $\beta = \gamma = 1$ and $W_e((u, 1), (v, 2)) = W_e((v, 1), (u, 2)) = 1$. Suppose we want to solve DIRECTED PCST with $\alpha = 10$. (b) After reducing DIRECTED PCST to DST, we get a graph G' with modified edge weights and no node weights. An optimal directed Steiner tree T on graph G' is shown on the right; $W_{DST}(T, G') = 3 \times (-8) + 28.1 = 4.1$. Also note that $W_{PCST}(T, G_S) = 4 \times 1 + 10 \times 0.01 = 4.1$, showing that $W_{DST}(T, G') = W_{PCST}(T, G_S)$. Note that since edge weights can be negative, leaves of an optimal directed Steiner tree need not be terminals.

A. *Reducing* DIRECTED PCST *to Directed Steiner Tree*

We reduce DIRECTED PCST to DST as follows. Let $E_S \subseteq E$ denote the edge set $\{(a, b) \in E \mid b \in S\}$. Let $T := \sum_{a \in V} W_v(a)$. From G_S we create a new graph $G'(V, E, r, S, W'_e)$ with *only* edge weights, given by the function $W'_e : E \to \mathbb{R}$, such that for all $(a, b) \in E$

$$W'_e(a, b) = \begin{cases} W_e(a, b) - \alpha \cdot W_v(b), & \text{for } (a, b) \in E \setminus E_S \\ W_e(a, b) + \alpha \cdot \frac{T}{|S|}, & \text{for } (a, b) \in E_S. \end{cases}$$

Note that the new edge weights $W'_e(a, b)$ can be negative, especially for large α. For any directed tree T in G that is rooted at r and spans S, let $W_{PCST}(T, G)$ denote the objective function value (i.e., the expression in (1)) of tree T for the DIRECTED PCST problem on graph G_S. For any directed tree T in G' that is rooted at r and spans S, let $W_{DST}(T, G')$ denote the objective function value of tree T for the DST problem on graph G'. The reduction is illustrated in Figure 3. We have the following lemma.

Lemma 1: For any directed tree $T(V_T, E_T)$, $V_T \subseteq V$, $E_T \subseteq E$, rooted at r and spanning S, $W_{DST}(T, G') = W_{PCST}(T, G)$. Furthermore, if T is an optimal directed Steiner tree for G', then T is also an optimal prize-collecting Steiner tree for G.

In fact, we prove a stronger, approximation-preserving relation between the two problems as shown by the following lemma.

Lemma 2: For any $\rho \geq 1$, if a tree T is a ρ-approximate directed Steiner tree for G', then T is a ρ-approximate directed Prize-Collecting Steiner tree for G.

Proofs for Lemma 1 and Lemma 2 are omitted due to space constraints.

B. *Greedily Solving DST Approximately*

Having reduced DIRECTED PCST to DST, we use the clever, greedy algorithm of Charikar et al. [5] to obtain an approximation algorithm for DIRECTED PCST. The Charikar et al. algorithm achieves an $O(i^2 k^{1/i})$ approximation ratio in time $O(n^i k^{2i})$ for any fixed $i \geq 1$ where k is the number of terminals. Setting $i = 1$ gives an $O(k)$ approximation algorithm in $O(nk^2)$ time and setting $i = 2$ gives an $O(\sqrt{k})$ approximation algorithm in $O(n^2 k^4)$ time. We use GREEDY$_i$ to denote the Charikar et al. algorithm with parameter i.

C. *Using an LP Relaxation of DST*

Given a directed graph $G'(V, E, r, S, W'_e)$, with a root node $r \in V$, terminal set $S \subseteq V$, and edge-weight function $W'_e : E \to \mathbb{R}$, the DST problem can be modeled by the following flow-based integer linear program (ILP) [22]. The variables $f_{s,e} \in \{0, 1\}$ for each $s \in S$ and $e \in E$ represent the presence of 1 unit of flow from the root r to terminal s via edge e. The variable $y_e \in \{0, 1\}$ indicates the use of edge e for some flow; this meaning is enforced by the fourth constraint set, $f_{s,e} \leq y_e$. For any node $v \in V$, $\delta^+(v)$ (respectively, $\delta^-(v)$) is the set of edges leaving (respectively, entering) v. The first constraint ensures that there is 1 unit of flow leaving the root, for each terminal s. The second constraint ensures that there is 1 unit of flow intended for s entering each terminal s. The third constraint ensures flow conservation, of flows intended for all terminals, at all nodes. The fifth constraint set ($\sum_{e \in \delta^-(v)} y_e \leq 1$) ensures that the paths induced by the flows form a tree.

$$\min \sum_{e \in E} W'_e(e) \cdot y_e$$

$$\text{s.t.} \quad \sum_{e \in \delta^+(r)} f_{s,e} - \sum_{e \in \delta^-(r)} f_{s,e} = 1 \ \forall s \in S$$

$$\sum_{e \in \delta^+(s)} f_{s,e} - \sum_{e \in \delta^-(s)} f_{s,e} = -1 \ \forall s \in S$$

$$\sum_{e \in \delta^+(v)} f_{s,e} - \sum_{e \in \delta^-(v)} f_{s,e} = 0 \ \forall v \in V \setminus \{r\}, s \in S \setminus \{v\}$$

$$f_{s,e} \leq y_e \ \forall s \in S, e \in E$$

$$\sum_{e \in \delta^-(v)} y_e \leq 1 \ \forall v \in V$$

$$f_{s,e} \in \{0, 1\} \ \forall s \in S, e \in E$$

$$y_e \in \{0, 1\} \ \forall e \in E$$

It is easy to verify that this ILP models DST. An LP relaxation of this ILP is obtained by replacing the two sets of integrality constraints at the end of the program by $0 \leq f_{s,e} \leq 1$ and $0 \leq y_e \leq 1$ for all $s \in S$, $e \in E$. While solving the ILP optimally is not computationally feasible, solving this LP relaxation is. The following theorem formalizes the connection between DIRECTED PCST and this LP relaxation and indicates how we use the LP relaxation. The proof is omitted due to space constraints.

Theorem 1: Let T^* be an optimal directed Prize-Collecting Steiner tree for input $G_S(V, E, r, S, W_e, W_v)$ and parameter $\alpha > 0$. Let the graph $G'(V, E, r, S, W'_e)$ be obtained from

G_S via the reduction in Section III-A. Let W^* be the cost of the solution returned by above LP relaxation on G'. Then $W^* \leq W_{PCST}(T^*, G)$.

At first glance it may be unclear if a fractional solution to the LP relaxation has a useful interpretation in the context of identifying asymptomatic cases. We propose the following interpretation. For an integral solution to the LP, for each non-terminal node v, either $\sum_{e \in \delta^-(v)} y_e = 1$ or $\sum_{e \in \delta^-(v)} y_e = 0$. If the former is true, then v is in the tree and consider an asymptomatic case. For a fractional solution to the LP, for each non-terminal node v, $0 \leq \sum_{e \in \delta^-(v)} y_e \leq 1$, and we interpret $\sum_{e \in \delta^-(v)} y_e$ as the probability that v is an asymptomatic case. This idea is inspired by the technique of randomized rounding [20] for obtaining good integral solutions from optimal fractional solutions.

D. Minimum Cost Arborescence Heuristic

Given an edge-weighted directed graph $G(V, E, W_e)$ and a vertex $r \in V$, an *arborescence* (rooted at r) is a tree T such that (1) T is a spanning tree of G if we ignore the direction of edges and (2) there is a directed unique path in T from r to each other node $v \in V$. An MCA is an arborescence of smallest total weight.

For general directed graphs we can compute an MCA in $O(m + n \log n)$ time due to Gabow et al. [9]. This improves the naive implementation that runs in $O(nm)$ time. For DAGs, this algorithm can be simplified to run in $O(m+n)$ time. The algorithm is simply this: for each $v \neq r$ add to the solution the edge incoming into v with minimum edge weight (breaking ties arbitrarily).

We use an MCA algorithm to produce a directed Prize-Collecting Steiner tree on $G_S(V, E, r, S, W_e, W_v)$ as follows. We first transform G_S into $G'(V, E, r, S, W'_e)$ as per the reduction from DIRECTED PCST to DST from Section III-A. From G' we construct a new directed graph H whose vertex set is $S \cup \{r\}$. We add an edge (u, v) to H iff there is directed path in G' from u to v. The weight assigned to (u, v) in H is the shortest path distance from u to v in G'. It is easy to verify that since G' is a DAG, H is also a DAG. To obtain a directed Steiner tree on G', we compute an MCA on H, replace each edge (u, v) in the MCA by a shortest path in G' from u to v, and finally return tree obtained by taking the union of these shortest paths.

IV. DATA

Our experimental results use an extensive, fine-grained hospital operations data set collected from a large, public, tertiary-care teaching hospital. The subset of these data used in this paper consist of architecture data (complete set of CAD files for a 3.2M square feet facility), admission-discharge-transfer data (273K inpatient hospitalizations between 2003 and 2013), prescription data (7.8M prescriptions), and surveillance data (2K positive CDI lab tests between 2005 and 2011). Using these data and given a size-T time window, we construct a temporal network $\mathcal{G} = (G_1, G_2, \ldots, G_T)$ and a sequence of observed cases $(S_1, S_2, \ldots, S_T)$, as described next.

Note: All individuals present in our data (patients and HCPs) are completely anonymous. For this reason, this project was human subjects research exempt.

A. Constructing Temporal Graph from Raw Data

The subset of data relevant to our experiments consists of the following elements:

1) A collection X of patient visits. Each visit $x \in X$ spans a sequence of consecutive days denoted by the range $[s(x), e(x)]$ and for each day $d \in [s(x), e(x)]$ of a visit x there is an associated location (patient room) denoted $\ell(x, d)$.
2) The set of locations is denoted by L and there is a distance metric $D : L \times L \to \mathbb{R}^+$ defined on this set. We use discretized CAD drawings of the facility to obtain a "walking distance" metric between all pairs of rooms in the hospital. This is represented by D.
3) A partition of $X = C \cup N$, into CDI visits and non-CDI visits. For each CDI visit $x \in C$, there is a day $d \in [s(x), e(x)]$ that corresponds to a positive CDI test; we denote this day of positive test by $d^+(x)$.
4) For each visit $x \in X$, we have associated demographic features and whether there was a previous visit to the hospital within 60 days. In addition, we have features that change over time. Specifically, for each day $i \in [s(x), e(x)]$, we have the length of stay (from admission time to day i) and a list of high-risk antibiotics and gastric acid suppressors prescribed to the patient for day i of the visit. Finally, we also have "exposure" features, i.e., counts of the number of other CDI patients in the same room or unit.

We now fix a time window size T (in days) and without loss of generality assume that the days in this time window are labeled $1, 2, \ldots, T$. For each $i = 1, 2, \ldots, T$, we construct the directed network $G_i = (V_i, E_i, W_i, \mathbf{F}_i)$ and observed case sequence $(S_1, S_2, \ldots, S_T)$ as follows.

- **Node set** V_i**:** If $i \in [s(x), e(x)]$ for a patient visit $x \in X$, we add x to V_i. In other words, V_i is the set of all patient visits that are taking place on day i.
- **Edge set** E_i**:** For every $x, y \in V_i$, if locations $\ell(x, i)$ and $\ell(y, i)$ *belong to the same hospital unit*, then we add two directed edges (x, y) and (y, x) to E_i. In other words, all ordered pairs of nodes in V_i that are located (on day i) in the same unit are connected by edges. If locations $\ell(x, i)$ and $\ell(y, i)$ do not belong to the same unit, then for small probability $p \in [0, 1]$ (e.g., $p = 0.01$), we randomly (and independently) add edge (x, y) to E_i. These "long-distance" edges model "weak ties" induced by HCPs (especially physicians) who travel between units. Note that the preponderance of HCP mobility is within units.
- **Weights** W_i**:** For every edge $(x, y) \in E_i$, we set $W_e(x, y) = D(\ell(x, i), \ell(y, i))$. In other words, edge weights simply represent physical distance between pairs of hospital rooms.
- **Feature vector** F_i**:** For every node $x \in V_i$, we set $F_i[x]$ using the features described earlier in item (4).
- **Observed cases** S_i**:** For any node $x \in V_i$, if $x \in C$, i.e., x is a CDI visit, and $d^+(x) = i$, then x is added to S_i.

From this temporal network $\mathcal{G}$, we obtain a time-expanded network $G_S(V_S, E_S, r, S, W_e, W_v)$ as described in Section II. Note that learning the node weights W_v, which represent the probability of a patient being an asymptomatic, is described in Section V.

B. Generating synthetic "ground truth" data

Since we do not have "ground truth" data on asymptomatic infections, we also use the data described in the previous section to generate synthetic data via a partially hidden, "biased" *susceptible-infectious-susceptible* (SIS) process. The SIS model is designed for infections with no long-lasting immunity; any susceptible agent can get infected with a probability upon contact with an infectious agent, and an infected agent returns to a susceptible state with some probability. Here we implement a *biased* version of the SIS process. Every node v has an assigned probability (i.e., a bias) that represents the individual node v's risk of being an asymptomatic case; node v participates in the process with this bias. We now describe this process in more detail, carefully differentiating between aspects that are hidden and aspects that are revealed.

Our implementation of the biased SIS process has three main steps.

1) **Generating biases and susceptible nodes.** Recall that in G_S, each node v is assigned a probability $W_v(v)$ that represents the individual node v's risk of being an asymptomatic case. From the distribution of the W_v values, we first learn a probability density function using *kernel density estimation* (KDE). Next, for every visit $x \in X$, we sample a bias $W_x \in [0, 1]$ from the estimated probability density function and then set a bit s_x to 1, independently, with probability proportional to W_x. We then project these quantities, associated with visits, onto individual nodes in G_s. Specifically, every node (x, i) in G_S that corresponds to visit x is assigned the probability W_x; i.e., $W_v(x, i) = W_x$, and we set the state of (x, i) to "susceptible" if $s_x = 1$. For every node (x, i) in G_S, the probability $W_v(x, i)$ is revealed to us, but the state of the node remains hidden.
2) **Running the biased SIS process.** Now, for some positive integer parameter k, we pick k infection sources at random from the set of susceptible nodes. Then we run an SIS process starting at each of these k sources. Two aspects of the SIS process are worth noting: (i) only the susceptible nodes participate in the SIS process and (ii) the probability of infection flowing along an edge $((x, t), (y, t + 1))$ in G_S is inversely proportional to the edge weight $W_e((x, t), (y, t + 1))$. Note that the k sources and the SIS process is entirely hidden from us.
3) **Revealing observed infections.** After running the SIS process, we visit every infected node (x, i) in each of the k infection trees and with a fixed probability q, we independently *reveal* (x, i) to be infected. These revealed infected nodes form the observed set of infections and we use these as the set S of terminals. We then prune each infection tree so that all leaves are revealed infected nodes. The nodes in each pruned tree that are not revealed to be infected are considered asymptomatically infected.

In summary, the biases and observed infections are revealed to our algorithms, but everything else is hidden. However, we are able to evaluate the performance of our algorithms because the above-described process also provides us with "ground truth" asymptomatic cases. We emphasize that a critical aspect of our setup is the fact that the SIS process is influenced by the revealed biases. We also note that our experimental setup is quite flexible. For example, the simple functions that govern the relationship between node biases and node susceptibility or edge weights and the likelihood of infection flow along edges can be easily replaced by other, more complicated functions.

For the time-expanded network G_S obtained from 1 month of data (20.9K nodes and 0.5M edges), for values of $\beta = 1, 2, 4$, we generate 18, 20, and 17 terminal nodes respectively, and 40, 49, and 47 asymptomatic cases, respectively.

V. Learning Individual Risks

In this section, we describe the training of a model that takes as input the feature vector $\mathbf{F}_i[x]$ of each node $x \notin S_i$ in graph G_i (i.e., nodes not observed to be infected) and estimates the likelihood of x being an asymptomatic CDI case. As mentioned earlier, the fundamental obstacle to training this model is the fact that our data lacks "ground truth" labels. So we use two simple and well-motivated hypotheses about how asymptomatic CDI cases may relate to observed CDI cases in order to train our model.

- **Hypothesis 1**: Asymptomatic CDI cases and observed CDI cases have similar risk profiles.
- **Hypothesis 2**: The mechanism for acquiring (symptomatic) CDI consists of the patient first being an asymptomatic CDI cases and then being prescribed high-risk antibiotics.

Hypothesis 1 is justified by studies (e.g., [17]) that show asymptomatic colonization has risk factors such as previous CDI, antibiotic exposure, and hospital stay and these are also risk factors associated with symptomatic CDI [6], [7]. Hypothesis 2 is justified by mechanistic models for CDI (e.g., [29]) that often attribute the transition from asymptomatic CDI to symptomatic CDI to the use of additional high-risk antibiotics. The two simple hypotheses are quite powerful in that they allow us to train different asympomatic CDI case prediction models that we can then evaluate. Using the trained models we can obtain, for each non-terminal node (x, i) in the time-expanded network G_S, a probability $W_v(x, i)$ of node (x, i) being an asymptomatic CDI case. These probabilities serve as node weights of the time-expanded network G_S provided as input to DIRECTED PCST.

Hypothesis 1 implies that we can train our model using observed CDI cases as instance labels. Then, patients who are assigned a high probability by a model trained in this manner, but are not CDI cases, are inferred to be asymptomatic CDI cases. Variants of this model can be obtained by using different subsets of features. More specifically, we partition the set of features into three groups: (i) *baseline* feature set B, consisting of length of stay, age, gender, prior UIHC visit

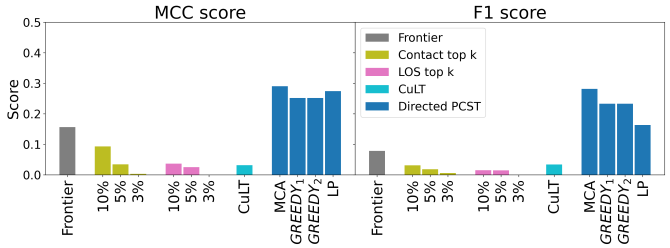

Fig. 4: Performance of all the methods in the synthetic data as measured by MCC (left) and F_1-score (right). All of our proposed methods MCA, GREEDY$_1$, GREEDY$_2$, and LP (in blue) comprehensively outperform the baselines.

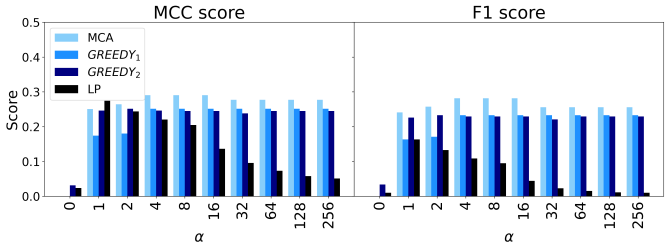

Fig. 5: The effect of varying α on the performance measured by MCC (left) and F_1-score (right). We see a sharp increase of the performance from $\alpha = 0$ to $\alpha = 1$, followed by a minor increase, then a gradual decrease as α increases.

within 60 days, and the use of gastric acid suppressors, (ii) *colonization pressure* feature set CP, consisting of different measures of exposure to other observed CDI cases and (iii) *antibiotics* (*ABXs*) feature set ABX, consisting of the use of high-risk ABXs. For each of the 4 subsets $S \subseteq \{CP, ABX\}$, we train a model on the feature set $\{B\} \cup S$.

Hypothesis 2 has the following useful implication. Suppose A is the subset of patients who were prescribed high-risk antibiotics during their visit. Then, the subset $A_{CDI} \subseteq A$, consisting of patients who tested positive for CDI is exactly identical to the subset of A of patients who were asymptomatic CDI cases (prior to receiving antibiotics) and $A \setminus A_{CDI}$ is exactly the subset of A of patients who are not asymptomatic CDI cases. Thus for the patients who were prescribed high-risk antibiotics during their visit, the "observed case" label corresponds exactly to the "asymptomatic case" label and we can train our models on this subset of the data. Just like for Hypothesis 1, we train 4 models by considering different subsets of features in addition to the baseline set of features.

VI. Experiments

We now present an extensive evaluation of the accuracy and efficiency of our proposed methods on a large-scale synthetic data. We also leverage our approach for an important application, (symptomatic) CDI case prediction, on a real hospital operations data described in Section IV. Our code and synthetic data are available for academic purposes [1]. All of our experiments were conducted on a Intel(R) Xeon(R) machine with 528GB memory.

Baselines: Since this is the first work on detecting asymptomatic cases in a temporal network that takes individual risks into account, there are no directly comparable methods. However, we compare the performance of our approach against the following natural baselines and state-of-the-art approach for a closely related task.

- **Frontier:** Nodes neighboring the known symptomatic cases could potentially be carriers. This method selects the neighbors of the terminal nodes as asymptomatic cases. Precisely, we mark a node as asymptomatic in the time-expanded network G_S if it has a directed edge to a terminal node.
- **Contact top k:** People with frequent contacts with others are likely to be exposed to infectious pathogens. This method selects top $k\%$, for $k \in \{3, 5, 10\}$, high-contact nodes based on the out-degree in G_S. We explore k up to 10, based on a study that found up to 10% of admitted patients fwere asymptomatic C. diff carriers [13].
- **LOS top k:** As length of stay (LOS) of patients in the hospital increases, there is a higher chance for the patient to contract infectious agents. For example, LOS is known to be a risk factor for HAIs [6], [7]. Here, we select top $k\%$, for $k \in \{3, 5, 10\}$, nodes based on the LOS.
- **CuLT:** This is the state-of-the-art Steiner-tree-based missing infection detection approach [23]. Note that algorithms that *CuLT* uses are just a special case of our *Greedy* approaches, where there are no node weights.

A. Performance on the Synthetic Data

We perform a series of experiments on the synthetic data described in Section IV-B to evaluate the performance of our algorithms. We start with a subset of the hospital data restricted to 1 month (Jan 2010) and first derive a temporal network (as described in Section IV-A) from this data and then a time-expanded network from this temporal network (as described in Section II). This yields a time-expanded network with more than 20.9K nodes and 0.5M edges We then run the biased SIS simulation described in Section IV-B on this time-expanded network, starting from multiple sources, to obtain a set of observed symptomatic cases $\mathcal{S}_{GT}$ and a set of asymptomatic cases $\mathcal{A}_{GT}$ as described in Section IV-B. Note that the point of using synthetic data is that it provides us with "ground truth" on asymptomatic cases.

1) Comparison with Baselines: The first experiment we conduct is designed to measure effectiveness of our approaches as compared to the baselines. We measure success for method m based on the overlap of the asymptomatic cases $\mathcal{A}_m$ it infers and the ground truth $\mathcal{A}_{GT}$. We use *Matthews correlation coefficient* (MCC) [16] and F_1-score as metrics to quantify success of the methods we evalute. Note that we tune hyper-parameters including α, β, and γ for each method and report the best performance. The final result is presented in Figure 4. As shown in the figure, our proposed approaches significantly outperform all the baselines in terms of both MCC and F_1-score. The result implies that our approaches recover as

[1] https://github.com/HankyuJang/directed-PCST-asymptomatic-detection

many ground truth asymptomatic cases as any method, while maintaining a very high accuracy. The high margin of the discrepancy between the performance of our approaches and the baselines could be attributed to the fact that our approach finds the right balance between the likelihood of node being exposed to the disease (via edge weights) and the likelihood node developing symptoms (via node weights). The superiority of our approach over *CuLT* highlights the importance of taking individual risks into account in detecting asymptomatic cases. This is a key takeaway from our results.

2) The Effect of α on the Performance: Next we study the effect of the parameter α on the performance. Note that the smaller values of α give higher weight to the edge costs while the larger values give higher importance to the node weights, forcing our algorithms to pick nodes with higher probabilities. In this experiment, we quantify the effect of varying α on the performance. To this end, we run our approaches on the synthetic graph and obtain a set of recovered asymptomatic cases for $\alpha = 0$. We then repeat this process for different values of α. We compute MCC and the F_1-score for the inferred asymptomatic cases for each value of α. The result is presented in Figure 5. We observe that for $\alpha = 0$, when the node probabilities have no effect on the solution, the performance is poor (as expected). On the other hand, for positive values of α, the performance is much better for all our methods. An immediate takeaway from this result is that individual risks are an important aspect of disease-spread. As we further increase α to values greater than 1, for the three methods that return an integral solution, there is a slight but gradual degradation in performance as more and more emphasis is placed on the node weights. The LP-based solution is much more sensitive to α and degrades relatively quickly as α increases. We suspect that the flexibility of being able to return a fractional solution allows the LP to disregard the constraint (solution to be a tree) and this degrades performance as α increases and there are more negative weight edges in the underlying graph. This experiment demonstrates that for small positive values of α where the importance of node-weights and edge-weights are balanced, we achieve the best performance, highlighting the importance of incorporating individual risks in detecting asymptomatic cases.

B. Scalability and Accuracy Trade-off

As a step towards running our experiments on a larger time-expanded network, we next study the trade-off between scalability and the solution cost (summation of edge weights) of our proposed approaches on the synthetic data. In Table I, we summarize the performance, in terms of the objective function cost our algorithms are minimizing and the time (in seconds) for each of our four methods. As expected, the optimal LP solution achieves the lowest cost, since its solution cost is guaranteed to be a lower bound on the cost of any integral directed Steiner tree (see Theorem 1). A pleasant surprise is that MCA returns a solution that is just a little bit larger than the LP solution, implying that the MCA returns a near-optimal solution. The cost of the tree

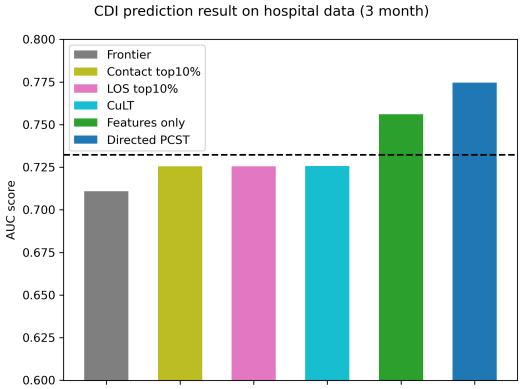

Fig. 6: Performance of the CDI prediction task measured by AUC. Our proposed approach DIRECTED PCST (in blue) outperforms the other methods as well as not using asymptomatic pressure features (dashed line).

returned by GREEDY$_2$ is reasonable (within 2 times OPT), while GREEDY$_1$ returns a tree with relatively larger cost. With regards to running time, the LP-based solution has a large running time making it unscalable to large graphs. This is because even though an LP can be solved in polynomial time, the size of the flow-based LP (see Section III-C) is much larger than the size of the underlying graph because there is a flow variable $f_{s,e}$ for every edge e and every terminal s. GREEDY$_2$ achieves reasonable running time while maintaining performance. It too, however, is too expensive for large graphs. On the other hand, our other two heuristics GREEDY$_1$ and MCA take a fraction of a second to compute the solution. Note that the MCA algorithm was made possible because we ensured that the time-expanded graph is a DAG.

TABLE I: The mean cost of the solution and the mean elapsed time in seconds for each method. s.t.dev in the parenthesis. The values are obtained from experiments by keeping $\alpha = 0$, but varying other parameters β and γ.

	Cost	Time
MCA	66.422 (65.56)	0.006 (0.0)
GREEDY$_1$	880.0 (1048.11)	0.001 (0.0)
GREEDY$_2$	100.678 (96.68)	1830.093 (2385.95)
LP	61.111 (62.94)	3807.492 (2302.26)

VII. APPLICATION: CDI CASE PREDICTION

Next we apply our method for asymptomatic case detection to the important task of predicting symptomatic CDI cases on actual hospital data. Specifically, we use a measure of exposure to detected asymptomatic CDI cases as an additional feature in a CDI prediction model. Since we do have "ground truth" CDI cases in our hospital data, we are able to compare our approach to other proposed methods for CDI prediction. Predicting CDI cases early is important task for many clinical reasons. For example, it can be used for doing early and more targeted testing of patients and initiating additional cleaning procedures at targeted locations so as to reduce CDI spread.

We expand the 1-month time window to 3-months and obtain a time-expanded network with more than 60.9K nodes

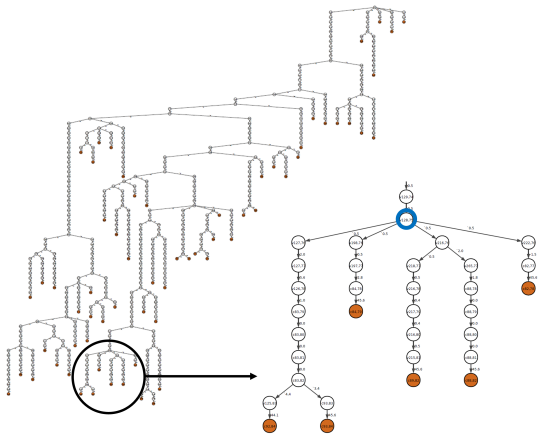

Fig. 7: A solution tree of the DIRECTED PCST of the experiment on the hospital data. Nodes in orange denote terminals.

and 1.6M edges Given the size of the this graph, and given our results on the cost-time tradeoff, we only use GREEDY$_1$ and MCA as algorithms for DIRECTED PCST.

Here, we first run all the methods to infer asymptomatic cases in the 3-month time-expanded network. Then we leverage the inferred asymptomatic cases to predict the symptomatic CDI cases. To do so, we train a neural network with two types of data: (i) standard risk factors of CDI (e.g., [7]) and (ii) additional features, that we call *asymptomatic pressures*, that measure the exposure to the newly identified asymptomatic CDI cases. Note that the additional features in (ii) are generated from the solutions of our approaches and the baselines. For each method, we investigate if adding exposure features to asymptomatic CDI cases improves performance of the neural model in predicting symptomatic cases. Since we have the ground-truth symptomatic cases, we use the area under ROC curve (AUC) to quantify the effectiveness of each method of predicting symptomatic cases. We split the data temporally into equally sized training and test sets. We further split training set into training (80%) and validation (20%) sets. Since the methods differ from each other only in asymptomatic pressures, their performance on symptomatic case prediction can be viewed as a proxy measure of how accurate the choice of asymptomatic cases was.

Figure 6 shows the CDI case detection results. The horizontal dashed line is the baseline performance that do not use asymptomatic pressures as features. One would expect adding extra information regarding exposure to the asymptomatic cases would only improve the performance. Hence, we interpret a method to have detected potential asymptomatic carriers correctly, if the performance of the symptomatic cases classification surpass the dashed line after adding additional features measuring the exposure to inferred asymptomatic cases. We note that all the natural baselines described earlier actually deteriorate the performance, as these baselines are not able to identify the asymptomatic carriers correctly. Furthermore, we compare our method against two alternative "extremes" for identifying asymptomatic cases: (i) *CuLT*: which uses a low-cost directed Steiner tree, while ignoring node weights completely and (ii) *Features only*: which uses node weights representing individual risks, while ignoring edges completely. For the *CuLT* method, adding asymptomatic pressures degrades performance relative to the baseline. The *Features only* method is helped by the use of asymptomatic pressures, but not as much as our method. The results in Fig 6 show that our proposed approach via the DIRECTED PCST problem (in blue) performs better than all the baselines and improves over the horizontal dashed line in terms of the AUC on the symptomatic cases prediction task. These results indicate that our approach is indeed able to infer likely asymptomatic cases in real outbreaks even when the "ground truth" data on asymptomatic cases is not available. The success in this task serves as a further, though indirect, evidence of the fact that our approach detects asymptomatic cases accurately.

VIII. CASE STUDY

We perform a case study on the real hospital data to demonstrate that the asymptomatic cases inferred by our approachs are meaningful. Here we chose our MCA algorithm with parameters $\alpha = 2, \beta = 2, \gamma = 0$. Fig 7 shows the solution tree returned by MCA for our directed PCST problem. There were 97 CDI cases in the period of 3 months and our solution partitioned these into 38 outbreaks. One of these outbreaks is the "giant" outbreak, shown in Fig 7 as emanating from the leftmost child of the "dummy" root. There are 4 minor outbreaks, also shown in Fig 7. The remaining 33 outbreaks are just isolated cases and not depicted in the figure. In the figure, the intermediate nodes connecting the terminals to the root are inferred to be the asymptomatic cases.

Upon exploring the data in further detail, we discovered an inferred asymptomatic case (node in blue in the highlighted sub-tree) had visited the hospital for a major surgery for a disease unrelated to CDI. The asymptomatic patient was transferred into the hospital from an acute-care hospital which provides inpatient medical care. Since exposure to healthcare settings is an important risk factor for CDI, it is likely that this asymptomatic patient was exposed to C. diff there. It turns out that this asymptomatic case and the four children of this node had visited the same location; the children nodes may have contracted C. diff at the location.

IX. RELATED WORK

Several approaches have been proposed for outbreak detection [1], [21], infection prediction [15], disease modeling control [11], and epidemic surveillance [3] in temporal networks. See [15] for a survey of existing works in this space.

There has been much interest in reconstructing epidemic outbreaks over time. Farajtabar et al. [8] use two-stage framework that learns the diffusion model and identifies the source that maximizes the likelihood of observing the cascade as per the learned diffusion model, as a solution to the source identification which is a special case of missing infection problem. Sundareisan et al. [24] propose Netfill that recovers missing infections under SI model given snapshots of infections over

time, using minimum description length principle [24]. There is also some work that leverage Steiner trees to infer missing infections. Rozenshtein et al. [23] use a Steiner tree based approach to reconstruct epidemic cascades on a streaming contact network for SI-like model. Xiao et al. [28] solve a related problem of inferring missing infections in static networks and in a different paper Xiao et al. [27] propose a sampling based approach for a robust cascade reconstruction. None of these approaches incorporate individual risks. Note that our problem of detecting asymptomatic cases is inherently different from the problem of inferring missing infections because individual attributes play a significant role in determining whether or not an individual is asymptomatically colonized. Makar et al. [14] present a latent state modelling approach to detect asymptomatic carriers. However, they assume that the underlying network is static and the disease does not spread through a chain of infections. In this paper, we solve the problem in a more general setting, where the underlying network is dynamic and disease spreads through a chain of asymptomatic cases.

X. Conclusion

This paper studies the problem of detecting asymptomatic cases in a temporal network in which outbreaks have occurred. We show that taking into account both individual risk and the likelihood of disease-flow along edges, leads to improved detection. We formulate the asymptomatic case detection problem as a Directed Prize-Collecting Steiner Tree problem. and show an approximation-preserving reduction from this problem to the Directed Steiner Tree problem and then use this reduction to obtain scalable algorithms for the Directed Prize-Collecting Steiner Tree problem. We then solve large instances of this problem on both synthetic data and actual hospital data and demonstrate that our detection methods outperform various baselines, including baselines that ignore either the individual risk or edge characteristics. We also demonstrate that the solutions returned by our approach are clinically meaningful by conducting a case study.

XI. Acknowledgement

This project was funded by CDC MInD Healthcare Network grants U01CK000531 and U01CK000594 and NSF grant 1955939. The authors acknowledge feedback from other University of Iowa CompEpi group members.

References

[1] B. Adhikari, B. Lewis, A. Vullikanti, J. M. Jiménez, and B. A. Prakash. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. *PLoS Comp Bio*, 15(9):e1007284, 2019.

[2] A. Archer, M. Bateni, M. Hajiaghayi, and H. Karloff. Improved approximation algorithms for prize-collecting Steiner tree and TSP. *SICOMP*, 40(2):309–332, 2011.

[3] Y. Bai, B. Yang, L. Lin, J. L. Herrera, Z. Du, and P. Holme. Optimizing sentinel surveillance in temporal network epidemiology. *Scientific reports*, 7(1):1–10, 2017.

[4] D. C. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri, A. M. Ipekci, G. Salanti, and N. Low. The role of asymptomatic sars-cov-2 infections: rapid living systematic review and meta-analysis. *medRxiv*, 2020.

[5] M. Charikar, C. Chekuri, T. yat Cheung, Z. Dai, A. Goel, S. Guha, and M. Li. Approximation algorithms for directed steiner problems. *J of Algorithms*, 33(1):73 – 91, 1999.

[6] E. R. Dubberke, K. A. Reske, S. Seiler, T. Hink, J. H. Kwon, and C.-A. D. Burnham. Risk factors for acquisition and loss of clostridium difficile colonization in hospitalized patients. *Antimicrobial Agents and Chemotherapy*, 2015.

[7] E. R. Dubberke, Y. Yan, K. A. Reske, A. M. Butler, J. Doherty, V. Pham, and V. J. Fraser. Development and validation of a clostridium difficile infection risk prediction model. *ICHE*, 32(4):360–366, 2011.

[8] M. Farajtabar, M. G. Rodriguez, M. Zamani, N. Du, H. Zha, and L. Song. Back to the past: Source identification in diffusion networks from partially observed cascades. In *AISTATS*, 2015.

[9] H. N. Gabow, Z. Galil, T. Spencer, and R. E. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6(2):109–122, 1986.

[10] E. Halperin and R. Krauthgamer. Polylogarithmic inapproximability. In *STOC*, page 585–594, 2003.

[11] H. 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. In *ASONAM*, 2019.

[12] L. Kyne, M. Warny, A. Qamar, and C. P. Kelly. Asymptomatic carriage of clostridium difficile and serum levels of igg antibody against toxin a. *NEJM*, 342(6):390–397, 2000.

[13] S. Leekha, K. C. Aronhalt, L. M. Sloan, R. Patel, and R. Orenstein. Asymptomatic clostridium difficile colonization in a tertiary care hospital: admission prevalence and risk factors. *American journal of infection control*, 41(5):390–393, May 2013.

[14] M. Makar, J. Guttag, and J. Wiens. Learning the probability of activation in the presence of latent spreaders. In *AAAI*, volume 32, 2018.

[15] N. Masuda and P. Holme. Predicting and controlling infectious disease epidemics using temporal networks. *F1000prime reports*, 5, 2013.

[16] B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975.

[17] K. Nissle, D. Kopf, and A. Rösler. Asymptomatic and yet c. difficile-toxin positive? prevalence and risk factors of carriers of toxigenic clostridium difficile among geriatric in-patients. *BMC Geriatrics*, 2016.

[18] U. D. of Health and H. Services. *Health Care-Associated Infections*, Jan 15, 2020 (accessed June 10, 2020).

[19] I. Potasman. Asymptomatic infections: The hidden epidemic. *Int J Clin Res Trials*, 2, 2017.

[20] P. Raghavan and C. D. Tompson. Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 1987.

[21] B. Y. Reis, I. S. Kohane, and K. D. Mandl. An epidemiological network model for disease outbreak detection. *PLoS Med*, 4(6):e210, 2007.

[22] T. Rothvoß. Directed steiner tree and the lasserre hierarchy. *CoRR*, 2011.

[23] P. Rozenshtein, A. Gionis, B. A. Prakash, and J. Vreeken. Reconstructing an epidemic over time. In *ACM SIGKDD*, pages 1835–1844, 2016.

[24] S. Sundareisan, J. Vreeken, and B. A. Prakash. Hidden hazards: Finding missing nodes in large graph epidemics. In *SDM*, pages 415–423, 2015.

[25] N. Tuncbag, A. Braunstein, A. Pagnani, S. S. Huang, J. Chayes, C. Borgs, R. Zecchina, and E. Fraenkel. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J Comput Biol*, 20(2):124–36, Feb 2013.

[26] C. J. Worby, D. Jeyaratnam, J. V. Robotham, T. Kypraios, P. D. O'neill, D. De Angelis, G. French, and B. S. Cooper. Estimating the effectiveness of isolation and decolonization measures in reducing transmission of methicillin-resistant staphylococcus aureus in hospital general wards. *AJE*, 177(11):1306–1313, 2013.

[27] H. Xiao, C. Aslay, and A. Gionis. Robust cascade reconstruction by steiner tree sampling. In *ICDM*, pages 637–646, 2018.

[28] H. Xiao, P. Rozenshtein, N. Tatti, and A. Gionis. Reconstructing a cascade from temporal observations. In *SDM*, pages 666–674, 2018.

[29] L. Yakob, T. V. Riley, D. L. Paterson, and A. C. Clements. Clostridium difficile exposure as an insidious source of infection in healthcare settings: an epidemiological model. *BMC Infectious Diseases*, 13(376), 2013.

[30] A. Zelikovsky. A series of approximation algorithms for the acyclic directed steiner tree problem. *Algorithmica*, 18(1):99–110, May 1997.