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
Email: {hankyu-jang, bijaya-adhikari, sriram-pemmaraju}@uiowa.edu, {af9pn, hkz8wk, vsakumar}@virginia.edu, {jiamingcui1997, badityap}@gatech.edu

Abstract

Healthcare acquired infections (HAIs) (e.g., Methicillin-resistant *Staphylococcus aureus* infection) have complex transmission pathways, spreading not just via direct person-to-person contacts, but also via contaminated surfaces. Prior work in mathematical epidemiology has led to a class of models – which we call *load sharing* models – that provide a discrete-time, stochastic formalization of HAI-spread on temporal contact networks. The focus of this paper is the *source detection* problem for the load sharing model. The source detection problem has been studied extensively in SEIR type models, but this prior work does not apply to load sharing models.

We show that a natural formulation of the source detection problem for the load sharing model is computationally hard, even to approximate. We then present two alternate formulations that are much more tractable. The tractability of our problems depends crucially on the submodularity of the expected number of infections as a function of the source set. Prior techniques for showing submodularity, such as the "live edge" technique are not applicable for the load sharing model and our key technical contribution is to use a more sophisticated "coupling" technique to show the submodularity result. We propose algorithms for our two problem formulations by extending existing algorithmic results from submodular optimization and combining these with an expectation propagation heuristic for the load sharing model that leads to orders-of-magnitude speedup. We present experimental results on temporal contact networks based on fine-grained EMR data from three different hospitals. Our results on synthetic outbreaks on these networks show that our algorithms outperform baselines by up to 5.97 times. Furthermore, case studies based on hospital outbreaks of *Clostridioides difficile* infection show that our algorithms identify clinically meaningful sources.

Introduction

Healthcare acquired infections (HAIs) such as Methicillin-resistant *Staphylococcus aureus* infection (MRSA) and *Clostridioides difficile* infection (CDI) pose a significant burden on our healthcare infrastructure (Centers for Disease Control and Prevention 2019). Unlike respiratory infections such as SARS-CoV-2 and Influenza, HAIs such as MRSA and CDI have complex transmission pathways (Plipat et al. 2013; Li et al. 2009; Jang et al. 2019) that involve contaminated surfaces in addition to direct person-to-person contacts. As a result, HAIs have been difficult to model, detect, and control. Recent work, e.g., (Plipat et al. 2013; Jang et al. 2019), has shown that standard SEIR type models (Hethcote 2000) are not very suitable for modeling HAI spread; we will describe a *load sharing* model (Jang et al. 2019) later.

Most hospitals have limited testing and when some infections are detected in the hospital, a lot of effort is invested into rapidly identifying the source of infection. This is done so that strategies such as isolation precautions can be imposed in order to limit the spread of the infection. This, of course, corresponds to the classical "source detection" problem, which has been studied extensively in data mining and network science, e.g., (Prakash, Vreeken, and Faloutsos 2014; Shah and Zaman 2010; Lappas et al. 2010). However, all prior work on source detection problems has been restricted to Susceptible-Exposed-Infected-Recovered (SEIR) type compartmental models, and cannot be easily adapted for HAIs. For instance, the Minimum Description Length (MDL) approach of (Prakash, Vreeken, and Faloutsos 2014) is crucially tied to the structure of SEIR type models. Thus the source detection problem remains open for HAIs, and is the focus of our paper.

Motivated by the approach of (Lappas et al. 2010), we use a risk minimization type formulation for the source detection problem on load sharing models for HAIs. We describe this informally below and more formally later. Let $\mathcal{G} = (G_0, G_1, \ldots, G_{T-1})$ be a temporal network, where the network $G_t = (P_t, L_t, E_t)$ represents interactions among a set P_t of people (e.g., patients, nurses, physicians) and between people and a set of locations L_t, in time step t. Let Pos be the set of observed positive HAI cases during the time window $[0, T-1]$, and let $Neg = (\bigcup_t P_t) \setminus Pos$. Let $\mathcal{M}$ be an instance of a load sharing model. For a source set S, let $Inf_{\mathcal{M}}(S) \subseteq \bigcup_t P_t$ denote the (random) subset of people who get infected according to model $\mathcal{M}$ due to disease starting at S. For any $v \in \bigcup_t P_t$, let $\alpha(v, S) :=$ Prob$[v \in Inf_{\mathcal{M}}(S)]$. To measure how good the source set S is, at explaining the observations Pos, we define two quantities: $g(S) := \sum_{v \in Pos} \alpha(v, S)$, the expected number of infections according to $\mathcal{M}$, that are also observed, and $f(S) := \sum_{v \in Neg} \alpha(v, S)$, the expected number of infec-

tions according to $\mathcal{M}$, that are not observed. The overall goal of this paper is to solve the following (informally stated) SOURCEDETECTION problem.

SOURCEDETECTION (SD) (informal)
Given a temporal network $\mathcal{G} = (G_0, G_1, \ldots, G_{T-1})$, a load sharing model $\mathcal{M}$, and a set of observed HAI cases Pos, find a source set S that makes $g(S)$ large while keeping $f(S)$ small.

The main contributions of our paper are as follows:

- **Problem Formulations:** We show that a natural formulation of the SOURCEDETECTION problem for the load sharing model is computationally very hard, even to approximate. We then present two natural, alternate formulations (SD±KNAP and SD±RATIO) that are much more tractable, both in a theoretical and practical sense.
- **Submodularity result**: The tractability of our formulations depends crucially on a submodularity result that we show. We show that for the load sharing model, the expected number of infections that are also observed and the expected number of infections that are not observed, are both submodular functions of the source set. These functions are analogous to the expected number of infections in the Independent Cascade model (Kempe, Kleinberg, and Tardos 2003), which has been shown to be submodular, using the "live edge" technique. However, this technique does not work for our load sharing model, since there is no "live edges" interpretation of infection flow in the load sharing model. Instead, we need to use a more sophisticated *coupling* technique, motivated by (Mossel and Roch 2007). As far as we know, this is the first submodularity result for a disease-spread model involving the transfer and sharing of pathogen loads.
- **Scalable implementations with strong worst-case guarantees**: We use the submodularity of $f(\cdot)$ and $g(\cdot)$ to design multi-criteria approximation algorithms for the SD±KNAP and SD±RATIO problems. The worst case approximation guarantees we obtain depend on a notion of curvature of $f(\cdot)$ and $g(\cdot)$, but for our problem settings the function curvatures are such that we obtain strong guarantees. We significantly improve the running time of our algorithms for the SD±KNAP and SD±RATIO problems using a heuristic that we call *truncated expectation propagation*. This heuristic allows us to shortcut expensive simulations of the load sharing model, and leads to orders-of-magnitude speedup.
- **Experimental Results:** We evaluate our algorithms on real-world contact network datasets from three hospitals. Our experiments on synthetic outbreaks on real hospital contact data show that our approaches significantly outperforms baselines – by up to 6 times. Furthermore, we demonstrate that our approaches identify clinically meaningful sources on an actual in-hospital CDI outbreak.

Due to the space limit, all proofs appear in the Technical Appendix.

Background

Load sharing model

Traditional compartmental models for disease-spread via person-to-person contact (e.g., SI, SIS, SIR, and SEIR) (Hethcote 2000) have a long history, dating back to the early 20th century. However, these have been found to be inadequate for modeling transmission of diseases which need to take the environment into account (Li et al. 2009; Tien and Earn 2010; Plipat et al. 2013; Wang and Ruan 2017; Kraay et al. 2018; Jang et al. 2019), such as HAIs.

We now formally describe a load sharing model that was proposed for MRSA transmission in (Plipat et al. 2013; Jang et al. 2019). For any $y \in P_t \cup L_t$, let $L_y(t)$ denote the HAI pathogen load at node y at time t. Load dynamics can then be described by the following stochastic recurrence:

$$L_y(t+1) = (1-d)L_y(t) - \sum_{x:\{x,y\}\in E_t} \rho_{y,x} \cdot L_y(t) + \sum_{x:\{x,y\}\in E_t} \rho_{x,y} \cdot L_x(t) + I_{inf} \cdot q \quad (1)$$

Here $d \in (0,1)$ is a *die-off* parameter, and the term $(1-d)L_y(t)$ denotes the pathogen load remaining after die-off at time $t+1$. The next two terms represent pathogen transfer. For each time-t edge (interaction) $\{x,y\}$ that y participates in, a fraction $\rho_{y,x} \in (0,1)$ of load $L_y(t)$ is transferred from y to x and a fraction $\rho_{x,y} \in (0,1)$ of load $L_x(t)$ is transferred from x to y. In (Plipat et al. 2013; Jang et al. 2019) the *transfer* parameter $\rho_{x,y}$ depends on the contact area of "touch" interaction, the total area of entity x, and the transfer efficiency between the two touching surfaces. The last term denotes the (stochastic) shedding of pathogen load due to infection. The pathogen load $L_y(t)$ at person node $y \in P_t$ at time t stochastically determines if y becomes infected at time $t+1$. If y is infected at time $t+1$ then, for some *shedding* parameter $q > 0$, y sheds q units of pathogen, i.e., q units are added to y's pathogen load. Whether y becomes infected is determined by a *dose-response function* $p : \mathbb{R}^+ \to [0,1]$. (See (Brouwer et al. 2017) for a systematic study of dose response functions for infectious diseases.) Thus, a person node y becomes infected at time $t+1$ with probability $p(L_y(t))$. The quantity I_{inf} appearing in the last term is the indicator random variable indicating if y is infected in time $t+1$. Thus, Prob$[I_{inf} = 1] = p(L_y(t))$ if $y \in P_t$ and 0 otherwise (i.e., if y is a location).

For ease of exposition we assume the same n nodes are present in all T times stamps[1]. Then, this recurrence can be compactly rewritten as a stochastic matrix recurrence

$$\mathbf{L}(t+1) = (\mathbf{B}(t) + \mathbf{D}(t)) \cdot \mathbf{L}(t) + q \cdot \mathbf{I}(t). \quad (2)$$

Here $\mathbf{L}(t)$ is a length-n vector representing loads at the nodes at time t. $\mathbf{B}(t)$ is an $n \times n$ matrix with entry $[x,y]$ equal to $\rho_{x,y}$ if $\{x,y\}$ is an edge in G_t and 0 otherwise. $\mathbf{D}(t)$ is an $n \times n$ diagonal matrix with non-zero entries $[y,y]$ equal

[1] In reality and in our datasets, patients are admitted and discharged and healthcare professionals may also change. Thus the number of nodes may change from one time step to the next.

to $(1 - d - \sum_x \rho_{y,x})$; in this expression the sum $\sum_x$ is over all neighbors x of y in G_t. We restrict the model parameters so that $0 \leq (d + \sum_x \rho_{y,x}) \leq 1$, thus ensuring that the outgoing load does not cause load at node to become negative. Finally, $\mathbf{I}(t)$ is a length-n random indicator variable vector with entry $[y]$ equal to 1 with probability $p(L_y(t))$ if $y \in P_t$ and 0 otherwise. This models the fact that only people (not locations) can become infected and subsequently shed.

In (Jang et al. 2019), an exponential dose response function ($p(z) = 1 - e^{-\pi z}$) and a truncated linear dose response function ($p(z) = \min(\pi z, 1)$) are evaluated for an *infectivity* parameter $\pi > 0$. See Fig 1 and Table 1 in the Technical Appendix for an illustration of the load sharing model, and symbols and notation we use in this paper, respectively.

Problem Formulations

Intractability of a natural formulation

We start by presenting the hardness of a natural formulation of the source detection problem. Inspired by the k-effectors problem in (Lappas et al. 2010) and the Positive-Negative Partial Set Cover problem (denoted $\pm$PSC) defined in (Miettinen 2008), we define the SD$\pm$PSC problem as follows.

SD$\pm$PSC
Given a temporal network $\mathcal{G} = (G_0, G_1, \ldots, G_{T-1})$, a load sharing model $\mathcal{M}$, $0 \leq \tau_1 < \tau_2 < T$, and an observed (positive) set $Pos \subseteq \cup_{t \in [\tau_2, T-1]} P_t$, find a source set $S^* \subseteq \cup_{t \in [0, \tau_1]} P_t$ that minimizes

$$\sum_{v \in Pos} (1 - \alpha(v, S)) + \sum_{v \in Neg} \alpha(v, S). \tag{3}$$

The first term in the objective function is the expected number of observed positive cases *not* infected by an infection starting at source set S and the second term is the expected number of negative cases infected, by an infection starting at source set S. While this objective function is a simple and natural model for the SOURCEDETECTIONproblem, we prove the following hardness of approximation result that shows that no reasonable approximation exists for the problem.

Theorem 1. *For any $\varepsilon > 0$, the SD$\pm$PSC problem does not have an α-approximation for any $\alpha = O(2^{\log^{1-\varepsilon} n^4})$, where n is the number of nodes in G_0, unless NP $\subseteq$ DTIME($n^{polylog(n)}$). This is true even for an instance of the SD$\pm$PSC problem with 3 time stamps.*

In (Lappas et al. 2010) it is claimed that the k-effectors problem does not have a β-approximation for any $\beta > 0$ for the independent cascade (and similar) models. The proof of this claim and the NP-hardness claim it depends on seem incorrect (see the Technical Appendix for details) and so we use a different reduction to prove Theorem 1.

Two tractable problem formulations

In light of the hardness result in the previous section, we present two formulations of the informally stated SOURCEDETECTION problem that can be viewed as computationally tractable surrogates for the problem. The tractability of these formulations relies crucially on the submodularity of the functions f and g, which is shown in Theorem 2.

In the SD$\pm$KNAP problem defined below, instead of including both the positive and negative set of observations in the objective function (as in SD$\pm$PSC), we impose constraints that bound the number of people outside the set of observed cases that are reached by the infection starting at the source set. Let $Pos_t \subseteq P_t$ denote the set of observed cases at time t and let $Neg_t = P_t \setminus Pos_t$. Then let $f_t(S) := \sum_{v \in Neg_t} \alpha(v, S)$ denote the expected number of infections in the negative set at time t.

SD$\pm$KNAP
Given temporal graph $\mathcal{G} = (G_0, G_1, \ldots, G_{T-1})$, where $G_t = (P_t, L_t, E_t)$, integers τ_1, τ_2, $0 \leq \tau_1 < \tau_2 < T$, an observed set $Pos_t \subseteq P_t$ of cases, positive reals k_t, for each $t \in [\tau_2, T-1]$, find $S^* = \arg\max_s g(S)$ such that S satisfies the constraints $f_t(S) \leq k_t$ for all $t \in [\tau_2, T-1]$.

We next define the SD$\pm$RATIO problem. The goal of SD$\pm$RATIO is to find a source set S that maximizes the ratio of $g(S)$ (the expected number of infections in Pos) to $\sum_t \gamma_t \cdot f_t(S)$, a linear combination of the expected number infections in Neg_t for values of t in the observation period $[\tau_2, T-1]$. The coefficients γ_t can be viewed as penalties, which could vary with time – the penalty for getting it wrong in later time steps could be smaller than the penalty of getting it wrong in earlier time steps. This problem is similar in spirit to the problems of maximizing the difference $g(S) - f(S)$. However, as pointed out by Bai et al. (Bai et al. 2016) there is an important difference in the approximability of the ratio problem and the difference problem for submodular functions. While the ratio problem has algorithms with bounded approximation ratio for submodular functions, the difference version does not. This motivates the use of the SD$\pm$RATIO problem for source detection.

SD$\pm$RATIO
Given temporal graph $\mathcal{G} = (G_0, G_1, \ldots, G_{T-1})$, where $G_t = (P_t, L_t, E_t)$, integers τ_1, τ_2, $0 \leq \tau_1 < \tau_2 < T$, an observed set $Pos_t \subseteq P_t$ of cases and positive reals γ_t, for each $t \in [\tau_2, T-1]$, find

$$S^* = \arg\max_s \frac{g(S)}{\sum_{t=\tau_2}^{T-1} \gamma_t \cdot f_t(S)} \tag{4}$$

Methods

Submodularity of expected infections

In this section we show that the function $g(S)$ denoting the expected number of infections in the positive set is a monotone, submodular set function for any load sharing model that uses a concave dose response function. This result holds for the functions $f(S)$ and $f_t(S)$ as well. A concave dose response function models "diminishing response" to marginal increase in pathogen load. It is easy to see that functional

forms such as exponential and linear, mentioned earlier, are concave functions. A key aspect of our proof consists of showing that if loads at nodes are monotone, submodular functions of the source set and the dose response function is concave, then $g(S)$ is submodular. Showing that the loads are submodular needs new ideas because the "live edge" technique used in the classical results of (Kempe, Kleinberg, and Tardos 2003) for the independent cascade model and linear threshold model does not seem to apply here. In particular, there seems to be no apriori setting of the random bits that fixes node-infectivity and shedding, thereby yielding a deterministic version of node loads that can be easily shown to be submodular. To get around this obstacle, we base our proof on the "coupling" technique of (Mossel and Roch 2007). To show the submodularity of loads using the diminishing returns definition of submodularity, we need to consider 4 source sets S, $S+v$, Q, and $Q+v$, where $S \subseteq Q$ and $v \notin Q$. The key idea we use is the coupling of the stochastic decisions made by the 4 disease-spread processes starting from each of these source sets. With this 4-way coupling in place and using induction over time, we are able to show the submodularity of loads. The full proof is in the Technical Appendix.

Theorem 2. *Let τ_1, τ_2 be integers satisfying $0 \leq \tau_1 < \tau_2 < T$. For any concave function $p : \mathbb{R}^+ \to [0, 1]$, and any load sharing model $\mathcal{M}$ using dose response function p, the functions $g(S)$, $f(S)$, and $f_t(S)$ for $t \in [\tau_2, T-1]$ are all monotone and submodular.*

Algorithms for SD±KNAP

Since $g(\cdot)$ and $f_t(\cdot)$ have been shown to be submodular, it follows that SD±KNAP is a problem of maximizing a submodular function subject to multiple submodular knapsack constraints. Even a special case of this problem, with a single cardinality constraint, is well known to be NP-complete (Feige 1998). So we seek an approximation algorithm for SD±KNAP that provides strong worst-case guarantees, but is also practical and scalable. For this purpose we adapt the gradient ascent type framework of (Iyer and Bilmes 2013a,b), which has been proposed for the problem of maximizing a submodular function subject to a *single* submodular knapsack constraint. When used for SD±KNAP, each ascent step of the algorithm requires the solution of a simpler submodular function maximization problem with multiple *linear* constraints. The result of Bilmes and Iyer only works for a single constraint, for which they are able to use a natural greedy algorithm for the ascent step; this does not provide approximation guarantees for the problem with multiple constraints. Multilinear relaxation plus rounding (Kulik, Shachnai, and Tamir 2009) is a well known approach to solving the submodular maximization problem with multiple linear constraints. However, while this approach is theoretically elegant, it does not scale to our input sizes. So we use a simple approach based on the multiplicative update method that has been shown to provide a $O(1)$-factor approximation to the problem in (Azar and Gamzu 2012).

Algorithm 1 shows a high level description of our algorithm MUKnapsackSD for solving SD±KNAP. Given a current solution S, we compute in Line 3, a modular upper bound $\hat{f}_t$ of each submodular function f_t that is guaranteed to be tight at S. In Line 4, we use the multiplicative update algorithm to solve the problem of maximizing the submodular function $g(X)$ subject to linear constraints $\hat{f}_t(X) \leq k_t$. Below we show that MUKnapsackSD yields the following multicriteria approximation guarantee, where $K_{f_t} := max\{|X| : f_t(X) \leq k_t\}$ and κ_{f_t} is the *curvature* of the set function $f_t : 2^V \to \mathbb{R}$ defined as $\kappa_{f_t} := 1 - \min_{v \in V} \frac{f_t(v|V \setminus v)}{f_t(v)}$.

Algorithm 1: MUKnapsackSD

Input: $\mathcal{G}$, τ_1, τ_2, Pos, and k_t for each $t \in [\tau_2, T-1]$
Output: A seed set S

1: $S \leftarrow \emptyset$
2: **while** S has not converged **do**
3: Compute a linear function $\hat{f}_t$ defined as

$$\hat{f}_t(Y) := f_t(S) - \sum_{j \in S \setminus Y} f_t(j|S \setminus j) + \sum_{j \in Y \setminus S} f_t(j|\emptyset)$$

4: $S \leftarrow$ MultiplicativeUpdate$(g, \hat{f}_t, \forall t \in [\tau_2, T-1])$
5: **end while**
6: **return** S

Theorem 3. *Algorithm MUKnapsackSD returns a solution S_A such that $g(S_A) \geq \frac{1}{2(e(T-\tau_2)+1)} \cdot g(\hat{S})$. Here, $\hat{S}$ is an optimal solution to the problem of maximizing $g(S)$ subject to constraints*

$$f_t(S) \leq n_t \frac{1 + (K_{f_t} - 1)(1 - \kappa_{f_t})}{K_{f_t}}$$

for all $t \in [\tau_2, T-1]$.

Our observations of positive cases typically come from a small time window $[\tau_2, T-1]$. For example, in our experiments this time window has size 2. Thus, for all practical purposes, the fraction $1/2(e(T-\tau_2)+1)$ in the approximation ratio is a constant. In our setting, k_t tends to be fairly small because we don't want to allow the infection of too many negative cases. As a result, K_{f_t} will also tend to be quite small. Since the fraction $\frac{1+(K_{f_t}-1)(1-\kappa_{f_t})}{K_{f_t}}$ is bounded below by $1/K_{f_t}$, it is reasonable to expect the constraint violation in the multicriteria approximation is only a constant factor.

Additionally, we also implemented an algorithm called GreedyKnapsackSD, where we use a greedy heuristic instead of multiplicative update, for the ascent step. See the Technical Appendix for further details.

Algorithms for SD±RATIO

We use the simple, greedy algorithm from (Bai et al. 2016) as solution to SD±RATIO. For the problem of maximizing the ratio $g(S)/f(S)$, it is shown in (Bai et al. 2016) that their algorithm yields an approximation ratio of $1 -$

$1/e^{1-\kappa_f}$. For the SD±RATIO, this immediately translates to an approximation ratio of $1 - 1/e^{1-\kappa_{f'}}$, where $f'(S) = \sum_{t=\tau_2}^{T-1} \gamma_t f_t(S)$. While this approximation factor may be tight in the worst case, in the following we show that it can be restated in terms of the curvature of f' at the *optimal solution*. This quantity can be in general much smaller than $\kappa_{f'}$ and can therefore lead to a much better approximation ratio. For a set function $f : 2^V \to \mathbb{R}$ and $X \subseteq V$ define the *curvature of f at X*, denoted $\kappa_f(X)$, as

$$\kappa_f(X) := 1 - \min_{v \in X} \frac{f(v \mid X \setminus v)}{f(v)}.$$

In our setting, it is reasonable to expect $f(v \mid V \setminus v)$ to be near-0 for some $v \in V$. This is because $f(v \mid V \setminus v)$ represents the marginal increase in the expected number of infected when the set of sources contains all but one nodes and we add that last node as a source. On the other hand, for most HAI applications we are interested in, S^* will be small and $f(v \mid S^* \setminus v)$ may be relatively large for all $v \in S^*$. As a result, we expect that in practice $\kappa_{f'}(S^*) \ll \kappa_{f'}$. We use the name GreedyRatioSD to denote the greedy algorithm that we implemented for SD±RATIO problem.

Theorem 4. *The algorithm GreedyRatioSD yields a $(1 - 1/e^{1-\kappa_{f'}(S^*)})$-approximation for solving SD±RATIO, where S^* is an optimal solution.*

It is possible for the $g(S)/f(S)$ ratio to be very high, even when $g(S)$ is very small. Thus, GreedyRatioSD may return a solution S with very small value of $g(S)$, which may be considered unsatisfactory for the SD±RATIO problem. So we have also implemented a variant of GreedyRatioSD, called CoverRatioSD, that is forced to return a solution S for which $g(S)$ is at least a prescribed fraction of the number of observed cases. See the Technical Appendix for more details.

Truncated expected load propagation

A significant bottleneck in the running time of all our algorithms is the fact that the functions f and g are stochastic and obtaining good estimates requires a substantial number of simulations. We use a simple *expected load propogation* heuristic, described below, that allows us to shortcut costly simulations, while preserving solution quality.

Given loads at all nodes at time t, the expected load $\mathbb{E}[L_y(t+1)]$ at node y at time $t+1$ can be written as

$$(1-d)L_y(t) + \sum_x \left(\rho_{y,x} L_x(t) - \rho_{x,y} L_y(t)\right) + q \cdot p(L_y(t)).$$

Pushing the expectations through the right hand side and *approximating* the value of the dose response function $p(L_y(t))$ by $p(\mathbb{E}[L_y(t)])$, we get a deterministic recurrence

$$\begin{aligned}\mathbb{E}[L_y(t+1)] = \quad & (1-d)\mathbb{E}[L_y(t)] + \sum_x (\rho_{y,x}\mathbb{E}[L_x(t)] \\ & - \sum_x \rho_{x,y}\mathbb{E}[L_y(t)]) + q \cdot p(\mathbb{E}[L_y(t)])\end{aligned}$$

We use this recurrence to propagate expected loads from time stamp 0 to time stamp $\tau_2 - 1$ and then run simulations only for time stamps in $[\tau_2, T]$ in order to find the set $Inf([\tau_2, T]|S, \mathcal{M})$. Due to a somewhat subtle issue with this heuristic, we implement a "truncated" version of it, described in the Technical Appendix. As shown in Figure 3, this enables 17× speedup of our algorithms, without a noticeable degradation in quality of solution.

Experiments

We design extensive experiments to compare and contrast the performance of our algorithms against baselines in a variety of outbreak scenarios. Our code and public data is available for academic purposes [2].

Baselines: Despite there not being a direct competitor, we compare performance of our algorithms against a wide range of methods. CuLT (Rozenshtein et al. 2016) is the state-of-the-art cascade reconstruction approach that uses a directed Steiner-tree algorithm to span the observed infection set Pos. Similarly, NetSleuth (Prakash, Vreeken, and Faloutsos 2012) is an MDL-based approach which returns a set of source nodes that "best" describe the observations. PathFinder simply selects top-k candidate source nodes with the highest number of reachable nodes in Pos. Finally, LOS (Length of Stay) selects patients from the first few time-stamps as sources, who remain longest in the hospital.

Data: We run our methods and the baselines in real and simulated HAI outbreaks on a number of datasets. We used a total of 31 daily snapshots in each of our datasets. UIHC (10.4K nodes, 13.8K edges per day) consists of daily interactions between healthcare workers (HCWs), patients, and locations within the University of Iowa Hospitals and Clinics. The interactions were reconstructed from HCW logins and patient admission-discharge-transfer (ADT) records. UIHCUNIT (789 nodes, 526 edges per day) is a subset of UIHC. It corresponds to the unit with the history of highest number of CDI cases. UVAPRECOVID (2.4K nodes, 430 edges per day) consists of interactions between patients, HCWs and locations in the Cardiology department of the University of Virginia Hospital. These interactions were recorded in March 2011. UVAPOSTCOVID (0.9K nodes, 396 edges per day) was collected in the same unit as UVAPRECOVID, but in January 2020. Note that COVID-19 pandemic had already started when the interactions in this dataset occurred. Finally, CARILION (Jiménez, Lewis, and Eubank 2012) (2.3K nodes, 30K edges per day) consists of daily snapshots of interactions generated from mobility log obtained from Carilion Hospital in Roanoke, VA. There are a total of 72K unique locations and 89K unique individuals in this dataset. We extract a densely connected subgraph from the largest connected component of the dynamic interaction network.

Quality of the detected sources

Our goal here is to quantify the goodness of the sources inferred by our approaches and the baselines. Although we have real HAI outbreaks in some of our datasets, true "ground truth" sources are unavailable. Hence, in these sets of experiments we rely on simulated outbreaks. We run a

[2] https://github.com/HankyuJang/Detecting-Sources-of-Healthcare-Associated-Infections

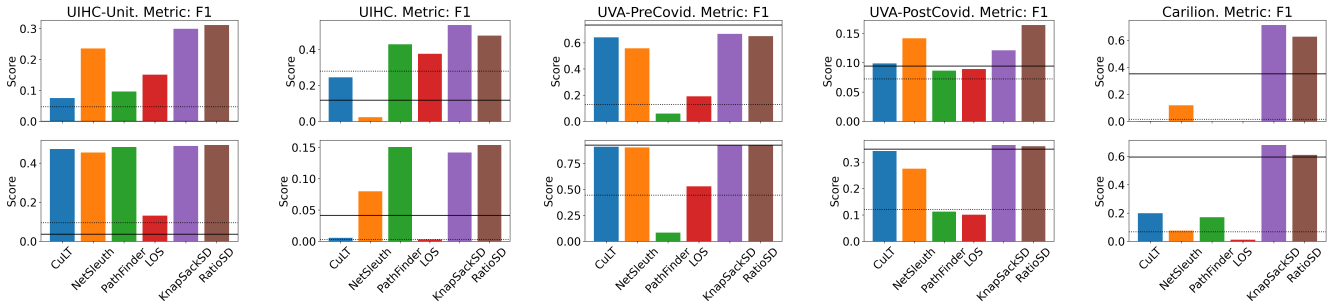

Figure 1: The performance of our approaches KnapsackSD and RatioSD and the baselines on various datasets in terms of F1-score (see the Technical Appendix for MCC results). The size of the ground-truth source set in the top row is 2 and that of the bottom row is 6. Only our approaches perform consistently well across different settings.

well-calibrated version of the load sharing model (see Section "Background") with an arbitrary set S^+ of sources selected from among nodes appearing in time window $[0, \tau_1]$. From this run we obtain the sets of infections Pos and non-infections Neg for the time window $[\tau_2, T-1]$. Each method m is given sets Pos and Neg and returns a source set S^m.

A straightforward metric to measure success is the intersection between S^m and S^+. However, it is in general impossible for any algorithm to do well with respect to this metric because the "ground truth" source set S^+ may be quite poor in explaining the observed cases and non-observed cases relative to other source sets. In fact, we see this in our experiments, where we consistently discover source sets that have much higher probability of leading to observed cases and avoiding non-observed cases than the "ground truth" source set. So we use two other natural metrics to measure performance. First, we propose to measure the overlap between the ground truth Pos and Neg sets and the sets of infections Pos_m and non-infections Neg_m caused by an outbreak starting from the source set S^m selected by method m. Second, we use the "distance" between S^+ and S^m as a metric of success (see the Technical Appendix for results from this second metric).

We run 100 simulations starting from S^m for each method m to obtain Pos_m^i and Neg_m^i for $1 \leq i \leq 100$. From these, we compute the averaged true positive ATP_m, true negative ATN_m, false positive AFP_m, and false negative AFN_m for each method m. Finally, we compute average F1-Score and average MCC score (see the Technical Appendix for precise definitions of these scores). Our experiments on all the datasets set $|S^+|$ to 2 and 6. We set $\tau_1 = 1$ (source set comes from the first two time steps) and $\tau_2 = 29$ (observations come from last two time steps). In all settings, the calibrated simulations yield about 5-10% infection in the last snapshot, as is the case for a typical HAI outbreak (Jarvis et al. 2007; Clabots et al. 1992). Our results are summarized in Figure 1.

In the rest of the paper and in the figures, we use KnapsackSD to denote our better-performing algorithm for SD±KNAP (we implemented two) and RatioSD to denote our better-performing algorithm for SD±RATIO (we implemented two). The solid horizontal line represents performance of the ground truth seed set S^+ and the dotted horizontal line represents performance of a random seed set averaged over 30 times. These two lines represent the empirical "hardness" of the problem instance; a poor performance of ground truth indicates that the problem instance is hard and conversely a good performance of random source sets represent easier problem instance. The primary take-away from the results is that both of our approaches KnapsackSD and RatioSD consistently outperform all baselines regardless of instance hardness and performance metric. One can observe that CuLT and NetSleuth are inconsistent. We hypothesize that in settings where load transfer along multiple pathways plays role in infections, their performance drop as they are unable to model this phenomenon. Performance of other two baselines LOS and PathFinder are also similar. See the Technical Appendix for additional results.

Case study on a CDI outbreak

We perform a case study in UIHCUNIT to qualitatively explore the sources detected by our approaches in an actual CDI outbreak. In our UIHCUNIT dataset, there were a total of five CDI positive tests. One positive test (with anonymized ID 214) was recorded on day 1, whereas the remaining 4 were on day 25. We make the 4 observed positive cases on day 25 available to both of our methods and ask them to infer sources in days 0 and 1. The sources returned by our approaches did not include the infection observed in time 1. However, they included patients with (anonymized) IDs 753 and 409 (see Fig 2). Digging deeper into these patients medical history, we find that both patients had high co-morbidity scores at admission time. Patient 753's visit was later marked as a high severity admission by the Agency of Health Research and Quality (AHRQ). We also find that patient 409 was transferred to UIHC from an external acute-care hospital with inpatient facilities. In other words, both patients identified as sources had several of the known risk factors for CDI.

To check the quality of patients 409 and 753 as potential sources, we simulated two sets of 1000 outbreaks each with source set $\{409, 753\}$ and source set $\{214\}$. We observe that simulations starting from seeds detected from our

methods (409 and 753) are much better at "hitting" observed infections on day 25: 807 times for $\{409, 753\}$ versus 373 times for $\{214\}$. Similarly, simulations starting from with source set $\{409, 753\}$ "hit" unobserved infections a total of 2228 times versus 2765 times by simulations with source set $\{214\}$. Figure 2 illustrates one simulation with source set $\{409, 753\}$. Our conclusion is that patients 409 and 753 are clinically meaningful potential sources.

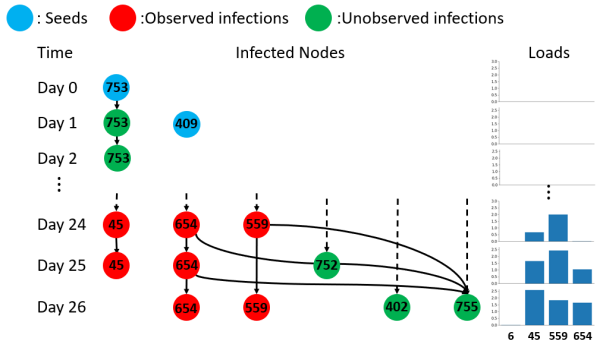

Figure 2: An illustration of one simulation for the CDI outbreak in UIHCUNIT. The sources (blue nodes) $\{409, 753\}$ identified by our algorithm were able to infect 3 out of 4 observed infections (red nodes) on day 25. Additionally, in this simulation there were only 3 spurious infections (green nodes) out of a total 43 patients not observed to be cases.

Speed-up due to expected load propagation

A big computational bottleneck for our algorithms is the number of simulations required to get good estimates of the stochastic functions $g(S)$ and $f(S)$. The truncated expectation propagation heuristic described earlier allowed us to shortcut costly simulations and was a key factor in our being able run all our experiments. Next we compute the speed-ups due to the lazy evaluation (Minoux 1978) and expected load propagation on UIHCUNIT. To show the kind of speed-ups this heuristic leads to, we demonstrate some experiments on UIHCUNIT. We start by running the vanilla version of KnapsackSD, its "lazy greedy" version, and a version augmented with both "lazy greedy" and truncated expected load propagation. The results are presented in Figure 3 As we can see lazy evaluation leads to the 1.09× speedup while lazy + expected load prorogation leads to 17.2× speedup. For RatioSD, we compare the performance of the vanilla version and the one augmented with truncated expected load propagation. Note that RatioSD does not have a lazy version. Here the expected load propagation version results in 5.6× speed up. These results highlight the importance of expected load propagation on speeding up our algorithms. While "lazy greedy" provides non-trivial speed-up, it is the truncated expected load propagation heuristic in addition to the "lazy greedy" that leads to an enormous speed-up.

Related Works

Data Mining for Hospital Acquired Infections: Several recent works have leveraged data mining and machine learning techniques for HAI related problems. These include outbreak detection (Adhikari et al. 2019), case predictions (Li et al. 2019; Brodzicki et al. 2020), and inferring latent cases (Makar, Guttag, and Wiens 2018; Jang et al. 2021). Jang et al. used agent based simulation to determine the effect of architectural changes on methicin-resistant Staphyloccus aureus (MRSA) outbreaks (Jang et al. 2019).

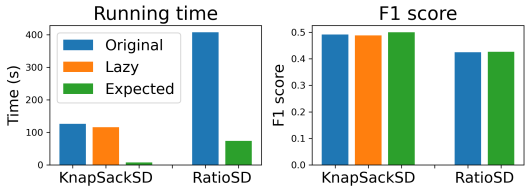

Figure 3: Running time and F1 score for various versions of KnapsackSD and RatioSD on UIHCUNIT. We achieved dramatic speedups by employing expected load propagation and lazy evaluations, with no sacrifice on the performance.

Source Detection: Detecting patient-zero in epidemics (Lappas et al. 2010; Shah and Zaman 2011; Jiang et al. 2016) has been well studied for the Independent Cascade (IC) and the Susceptible-Infected (SI) models. Several past works employ maximum likelihood estimation (Shah and Zaman 2011, 2012),Minimum Description Length (Prakash, Vreeken, and Faloutsos 2012), entropy (Zhang et al. 2021). Other commonly used approaches include label propagation (Wang et al. 2017) and monte carlo algorithms (Agaskar and Lu 2013) and divide and conquer using a reverse model (Zang et al. 2015).

Dynamics over Temporal Networks: Dynamical Processes over temporal networks have been well studied including in information diffusion (Tong et al. 2016), epidemiology (Volz and Meyers 2007), and malwares (Wen et al. 2012). Epidemic threshold on temporal networks are also well studied (Prakash et al. 2010; Leitch, Alexander, and Sengupta 2019; Valdano et al. 2015). Subsequently, dynamics over temporal networks have been used for many applications including viral marketing (Zhuang et al. 2013), community detection (Sattari and Zamanifar 2018), network compression (Adhikari et al. 2017), and so on.

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 showed 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 were able to 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. Extensive experiments on real and simulated outbreaks on three different hospital contact networks demonstrate significant advantages of our approach over the baselines.

Acknowledgements

This work was partially supported by the NSF (Expeditions CCF-1918770 and CCF-1918656, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, Medium IIS-1955939, Medium IIS-2106961, PIPP CCF-2200269, IIS-1955797), NIH 2R01GM109718-07, CDC MInD Healthcare network cooperative agreements U01-CK000594 and U01CK000589, and the associated COVID-19 supplemental funding, startup from the University of Iowa, faculty research award from Facebook and funds/computing resources from Georgia Tech. The University of Iowa authors acknowledge feedback received from members of the Computational Epidemiology research group at the University of Iowa. The authors also thank anonymous reviewers whose feedback improved the quality of the paper.

References

Adhikari, B.; Lewis, B.; Vullikanti, A.; Jiménez, J. M.; and Prakash, B. A. 2019. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. *PLoS computational biology*, 15(9): e1007284.

Adhikari, B.; Zhang, Y.; Bharadwaj, A.; and Prakash, B. A. 2017. Condensing temporal networks using propagation. In *SDM 2017*, 417–425. SIAM.

Agaskar, A.; and Lu, Y. M. 2013. A fast Monte Carlo algorithm for source localization on graphs. In *Wavelets and Sparsity XV*, volume 8858, 88581N. SPIE.

Azar, Y.; and Gamzu, I. 2012. Efficient Submodular Function Maximization under Linear Packing Constraints. In *ICALP*.

Bai, W.; Iyer, R.; Wei, K.; and Bilmes, J. 2016. Algorithms for optimizing the ratio of submodular functions. In *ICML*, 2751–2759. PMLR.

Brodzicki, A.; Jaworek-Korjakowska, J.; Kleczek, P.; Garland, M.; and Bogyo, M. 2020. Pre-trained deep convolutional neural network for clostridioides difficile bacteria cytotoxicity classification based on fluorescence images. *Sensors*, 20(23): 6713.

Brouwer, A. F.; Weir, M. H.; Eisenberg, M. C.; Meza, R.; and Eisenberg, J. N. S. 2017. Dose-response relationships for environmentally mediated infectious disease transmission models. *PLOS CompBio*, 13(4): 1–28.

Centers for Disease Control and Prevention. 2019. Antibiotic Resistance Threats in the United States.

Clabots, C. R.; Johnson, S.; Olson, M. M.; Peterson, L. R.; and Gerding, D. N. 1992. Acquisition of Clostridium difficile by hospitalized patients: evidence for colonized new admissions as a source of infection. *Journal of infectious diseases*, 166(3): 561–567.

Feige, U. 1998. A Threshold of Ln n for Approximating Set Cover. *J. ACM*, 45(4): 634–652.

Hethcote, H. W. 2000. The Mathematics of Infectious Diseases. *SIAM Rev.*

Iyer, R.; and Bilmes, J. 2013a. Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints. In *NeurIPS*.

Iyer, R.; and Bilmes, J. 2013b. Submodular optimization with submodular cover and submodular knapsack constraints. *arXiv*.

Jang, H.; Justice, S.; Polgreen, P. M.; Segre, A. M.; Sewell, D. K.; and Pemmaraju, S. V. 2019. Evaluating architectural changes to alter pathogen dynamics in a dialysis unit. In *ASONAM*, 961–968. IEEE.

Jang, H.; Pai, S.; Adhikari, B.; and Pemmaraju, S. V. 2021. Risk-aware Temporal Cascade Reconstruction to Detect Asymptomatic Cases. In *ICDM*.

Jarvis, W. R.; Schlosser, J.; Chinn, R. Y.; Tweeten, S.; and Jackson, M. 2007. National prevalence of methicillin-resistant Staphylococcus aureus in inpatients at US health care facilities, 2006. *American journal of infection control*, 35(10): 631–637.

Jiang, J.; Wen, S.; Yu, S.; Xiang, Y.; and Zhou, W. 2016. Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, 19(1): 465–481.

Jiménez, J. M.; Lewis, B.; and Eubank, S. 2012. Hospitals as complex social systems: Agent-based simulations of hospital-acquired infections. In *International Conference on Complex Sciences*, 165–178. Springer.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the Spread of Influence through a Social Network. In *KDD*, KDD '03, 137–146. ACM. ISBN 1581137370.

Kraay, A. N.; Hayashi, M. A.; Hernandez-Ceron, N.; Spicknall, I. H.; Eisenberg, M. C.; Meza, R.; and Eisenberg, J. N. 2018. Fomite-mediated transmission as a sufficient pathway: a comparative analysis across three viral pathogens. *BMC infectious diseases*, 18(1): 1–13.

Kulik, A.; Shachnai, H.; and Tamir, T. 2009. Maximizing Submodular Set Functions Subject to Multiple Linear Constraints. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, 545–554. USA: Society for Industrial and Applied Mathematics.

Lappas, T.; Terzi, E.; Gunopulos, D.; and Mannila, H. 2010. Finding Effectors in Social Networks. In *KDD*, KDD '10. ACM. ISBN 9781450300551.

Leitch, J.; Alexander, K. A.; and Sengupta, S. 2019. Toward epidemic thresholds on temporal networks: a review and open questions. *Appl. Netw. Sci.*

Li, B. Y.; Oh, J.; Young, V. B.; Rao, K.; and Wiens, J. 2019. Using machine learning and the electronic health record to predict complicated Clostridium difficile infection. In *OFID*, volume 6, ofz186. Oxford University Press US.

Li, S.; Eisenberg, J. N.; Spicknall, I. H.; and Koopman, J. S. 2009. Dynamics and control of infections transmitted from person to person through the environment. *American journal of epidemiology*, 170(2): 257–265.

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

Miettinen, P. 2008. On the Positive–Negative Partial Set Cover Problem. *Inf. Process. Lett.*, 108(4): 219–221.

Minoux, M. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*, 234–243. Springer.

Mossel, E.; and Roch, S. 2007. On the Submodularity of Influence in Social Networks. In *STOC*. ACM. ISBN 9781595936318.

Plipat, N.; Spicknall, I. H.; Koopman, J. S.; and Eisenberg, J. N. 2013. The dynamics of methicillin-resistant Staphylococcus aureus exposure in a hospital model and the potential for environmental intervention. *BMC infectious diseases*, 13(1): 595.

Prakash, B. A.; Tong, H.; Valler, N.; Faloutsos, M.; and Faloutsos, C. 2010. Virus propagation on time-varying networks: Theory and immunization algorithms. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 99–114. Springer.

Prakash, B. A.; Vreeken, J.; and Faloutsos, C. 2012. Spotting culprits in epidemics: How many and which ones? In *ICDM*. IEEE.

Prakash, B. A.; Vreeken, J.; and Faloutsos, C. 2014. Efficiently spotting the starting points of an epidemic in a large graph. *Knowl. Inf. Syst.*

Rozenshtein, P.; Gionis, A.; Prakash, B. A.; and Vreeken, J. 2016. Reconstructing an epidemic over time. In *KDD*.

Sattari, M.; and Zamanifar, K. 2018. A spreading activation-based label propagation algorithm for overlapping community detection in dynamic social networks. *Data & Knowledge Engineering*, 113: 155–170.

Shah, D.; and Zaman, T. 2010. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. *SIGMETRICS Perform. Eval. Rev.*

Shah, D.; and Zaman, T. 2011. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8): 5163–5181.

Shah, D.; and Zaman, T. 2012. Rumor centrality: a universal source detector. In *SIGMETRICS/Performance*.

Tien, J. H.; and Earn, D. J. 2010. Multiple transmission pathways and disease dynamics in a waterborne pathogen model. *Bulletin of mathematical biology*, 72(6): 1506–1533.

Tong, G.; Wu, W.; Tang, S.; and Du, D.-Z. 2016. Adaptive influence maximization in dynamic social networks. *Trans Netw*.

Valdano, E.; Ferreri, L.; Poletto, C.; and Colizza, V. 2015. Analytical computation of the epidemic threshold on temporal networks. *Physical Review X*, 5(2): 021005.

Volz, E.; and Meyers, L. A. 2007. Susceptible–infected–recovered epidemics in dynamic contact networks. *Proc. Royal Soc. B*.

Wang, L.; and Ruan, S. 2017. Modeling nosocomial infections of methicillin-resistant Staphylococcus aureus with environment contamination. *Scientific reports*, 7(1): 1–12.

Wang, Z.; Wang, C.; Pei, J.; and Ye, X. 2017. Multiple source detection without knowing the underlying propagation model. In *AAAI*.

Wen, S.; Zhou, W.; Zhang, J.; Xiang, Y.; Zhou, W.; and Jia, W. 2012. Modeling propagation dynamics of social network worms. *TPDS*.

Zang, W.; Zhang, P.; Zhou, C.; and Guo, L. 2015. Locating multiple sources in social networks under the SIR model: A divide-and-conquer approach. *Journal of Computational Science*, 10: 278–287.

Zhang, C.; Guo, Q.; Fu, L.; Gan, X.; and Wang, X. 2021. ITE: A Structural Entropy Based Approach for Source Detection. In *IEEE INFOCOM*.

Zhuang, H.; Sun, Y.; Tang, J.; Zhang, J.; and Sun, X. 2013. Influence maximization in dynamic social networks. In *ICDM*. IEEE.