scientific reports

OPEN

Using spectral characterization to identify healthcare-associated infection (HAI) patients for clinical contact precaution

Jiaming Cui[1,4]✉, Sungjun Cho[1,4], Methun Kamruzzaman[2], Matthew Bielskas[2,3], Anil Vullikanti[2,3] & B. Aditya Prakash[1]✉

Healthcare-associated infections (HAIs) are a major problem in hospital infection control. Although HAIs can be suppressed using contact precautions, such precautions are expensive, and we can only apply them to a small fraction of patients (i.e., a limited budget). In this work, we focus on two clinical problems arising from the limited budget: (a) choosing the best patients to be placed under precaution given a limited budget to minimize the spread (the isolation problem), and (b) choosing the best patients to release when limited budget requires some of the patients to be cleared from precaution (the clearance problem). A critical challenge in addressing them is that HAIs have multiple transmission pathways such that locations can also accumulate 'load' and spread the disease. One of the most common practices when placing patients under contact precautions is the regular clearance of pathogen loads. However, standard propagation models like independent cascade (IC)/susceptible-infectious-susceptible (SIS) cannot capture such mechanisms directly. Hence to account for this challenge, using non-linear system theory, we develop a novel spectral characterization of a recently proposed pathogen load based model, 2-Mode-SIS model, on people/location networks to capture spread dynamics of HAIs. We formulate the two clinical problems using this spectral characterization and develop effective and efficient algorithms for them. Our experiments show that our methods outperform several natural structural and clinical approaches on real-world hospital testbeds and pick meaningful solutions.

Healthcare-associated infections (HAIs) such as Methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridioides difficile* (*C. difficile*) are infections acquired by patients during treatment at healthcare facilities such as hospitals and long-term care homes[1–3]. While under treatment, patients can get exposed to pathogens from other sick patients or contaminated locations like rooms or medical devices. HAIs impose tremendous costs on hospitals, as they tend to extend patients' stays longer than initially planned to reach full recovery and increase the risk of mortality[4–9]. Due to its high health burden, epidemiologists and clinicians have proposed several approaches to control HAI spread in hospitals, such as contact precautions, antimicrobial stewardship, and contact tracing[10–15]. Among these approaches, contact precautions have been shown to be more effective in preventing further nosocomial infections[16, 17], and easier to implement[18]. Therefore, they are most commonly used in most hospitals to control HAI spread[12]. However, real-world contact precautions are expensive[7, 19–23] and therefore they can only be applied to a small fraction of patients (a 'budget'). Further, there are several downsides to contact precautions on patient health outcomes, e.g.,[22, 23]. Therefore, using contact precautions optimally is an important goal. We consider the following two important problems in HAI control with a limited budget, motivated by the above considerations.

Clinical Problem 1 (Isolation Problem, IP) Given a limited budget for precaution, how can we choose the "best" patients to put under precaution?

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA. [2]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904, USA. [3]Department of Computer Science, University of Virginia, Charlottesville, VA 22904, USA. [4]These authors contributed equally: Jiaming Cui and Sungjun Cho. ✉email: jiamingcui1997@gatech.edu; baadityap@cc.gatech.edu

Clinical Problem 2 (Clearance Problem, CP) Assume a hospital initially has a set of patients under precaution, but limited budget requires a subset of these patients to be cleared from precaution. How can we choose the "best" patients to release from precaution?

Intuitively, IP aims to find which patients to put under precautions. In contrast to IP, the setting for CP assumes that a set of patients are already under precautions, and the goal is to find a subset of patients from this set to *clear* from precautions. Although there are some guidelines on removing patients from contact precautions[18, 24] from clinical perspective, this is still a novel problem in the network science research field in terms of problem formulation and optimization: it is the opposite of the well-studied influence maximization problem on social networks[25] since the objective of CP is to minimize the spread while taking off precautions from nodes. This problem is also very different from the influence containment problem[26, 27] since these problems are edge-based, while IP and CP need to remove multiple edges when only isolating/releasing one node. No good formulations exist for the clinical problems above, as we cannot easily assess how each precaution on a specific patient affects the overall epidemic.

It is important to note that healthcare workers (HCWs) and contaminated physical surfaces can also spread the pathogens[28–30] in addition to the direct patient-patient contacts. However, traditional contact networks with only patient-patient contacts cannot capture such pairwise relations among diverse entities. To account for this challenge, heterogeneous contact networks are needed in modeling. Heterogeneous graphs contain multiple types of entities and can represent the diversity among nodes/pairwise relations in real-life networks. They have gathered great interest in recent years with applications in personalized recommendation[31], publication ranking[32], and drug design[33]. Here, we consider heterogeneous contact networks where each node can be one of three types: patients, HCWs, and locations. Besides, the contacts between each node are also changing with time. Hence, such heterogeneous contact networks should also be time-varying to capture the contact changes with time. We note that contact networks are not explicitly known, and need to be inferred. As shown in Fig. 1, we infer a contact network from Electronic Health Records (EHRs) data, which includes a variety of patients' medical records and doctors' notes. Specifically, we infer interactions between HCWs and patients through medicine administration, admission and discharge and flow sheets, and doctor notes from EHRs, and aggregate them into daily heterogeneous contact networks for each day t. See Supplementary Information for more details.

Modeling on heterogeneous contact networks is different from modeling on traditional contact networks. We need to identify the role of HCWs and location nodes in spreading the pathogens and involve this in modeling. We assume that HCWs and locations can also carry some pathogen loads and spread them to patients, but HCWs do not get infected. In this paper, we use a recently proposed model for HAI, the 2-MODE-SIS model proposed by Jang et al.[34], which incorporates pathogens spread through both people and locations as shown in Fig. 2a. Such
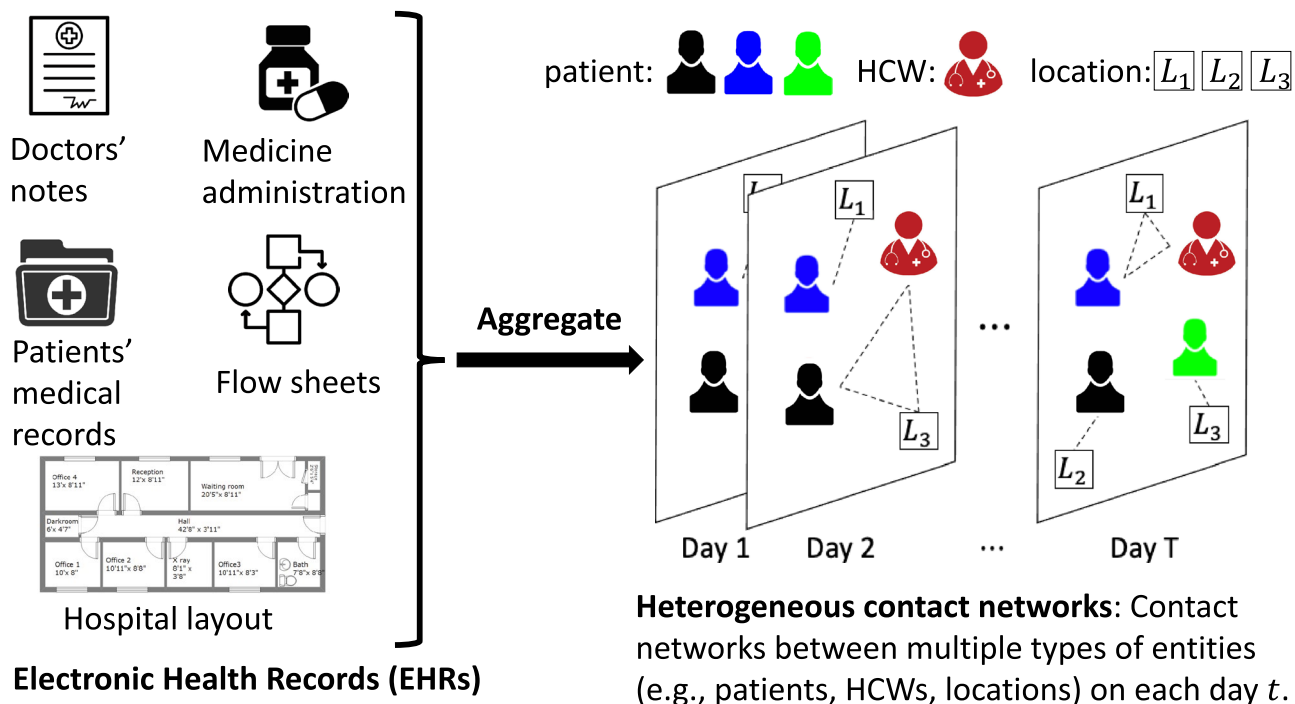

Figure 1. In this work, we collect the time and location of interactions between patients and HCWs from different tables in the electronic health records (EHRs) dataset, such as medicine administration, admission and discharge and flow sheets, and doctors' notes. We construct daily heterogeneous contact networks from this data (see more details in the Supplementary Information). For example, as shown in the figure on day T, the red healthcare worker and blue patient are in location L_1, the black patient is in location L_2, and the green patient is in location L_3.
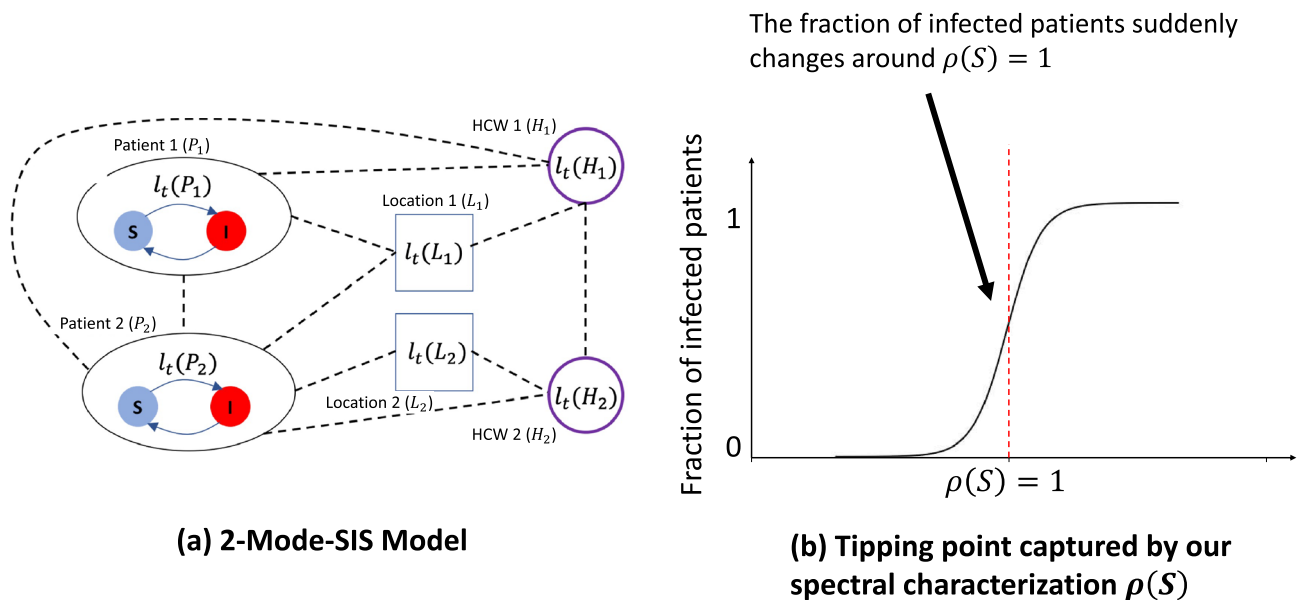
(a) 2-Mode-SIS Model

(b) Tipping point captured by our spectral characterization $\rho(S)$

Figure 2. (**a**) An example of the 2-Mode-SIS model on a heterogeneous contact network: on some day t, there are two patients (P_1, P_2), two HCWs (H_1, H_2), and two locations (L_1, L_2) in the network, and dash lines represent edges in the heterogeneous contact network (e.g., P_1 was in L_1 on day t as well as in contact with P_2 and H_1 on day t). For every node i, their pathogen load is represented as $l_t(i)$ by 2-Mode-SIS model, and these loads can spread via contacts. For patients P_1, P_2, their infection states can transfer between S ($x_t(i) = 0$) and I ($x_t(i) = 1$). See more details in the paper and Supplementary Information. (**b**) Our spectral characterization $\rho(S)$ captures a critical "tipping point": When $\rho(S) > 1$, the fraction of infected patients will be close to 1. When $\rho(S) < 1$, almost none of the patients will be infected (i.e., the epidemic dies out).

pathogen load based models have been used in recent high-profile HAI modeling studies[13, 35–37]. Importantly, they can also model the cleaning practice widely used in contact precautions (e.g., using disinfectants to clean the ward) directly[38–42] compared with traditional independent cascade (IC)[43]/susceptible-infectious-susceptible (SIS)[44] models. However, modeling with 2-Mode-SIS model is non-trivial since there are multiple transmission pathways through multiple kinds of nodes, making its dynamics harder to model.

The main idea of this paper is to use spectral characterization to help formulate the two clinical problems we are interested in as network optimization problems: We give a novel spectral characterization $\rho(S)$ of the 2-Mode-SIS model on underlying time-varying people-location heterogeneous contact networks using stability analysis of discrete-time non-linear dynamical systems[45]. We first approximate the propagation dynamics of the 2-Mode-SIS model on heterogeneous networks, after which we can analyze the stability of its equilibrium point (specifically, *all-zero* point where the disease dies out) and check whether this point is stable or not. As shown in Fig. 2b, we show that this characterization is meaningful in capturing a 'tipping point' for HAI spread dynamics. At this "tipping point", there will be a phase transition for the fraction of infected patients at the end: when $\rho(S) > 1$, the fraction of infected patients will be close to 1. When $\rho(S) < 1$, almost none of the patients will be infected (i.e., the epidemic dies out). This $\rho(S)$ is similar to the basic reproductive number R_0[46–49] for standard SIR-type models, and helps relate the structural and disease parameters to the dynamics. In the following section, we will show how we formulate the clinical problems in detail, and propose our Greedy-Spectral algorithm to address it.

Methods
2-Mode-SIS model
As shown in Fig. 2a, we use the 2-Mode-SIS model[34] to simulate the spread of pathogens (see Procedure 1 for simulation steps). 2-Mode-SIS is pathogen 'load-based' where the patient can be either *susceptible* or *infected*, and the model keeps track of pathogen load on all people and locations using a load vector l_t. Unlike classical SIS models[50], in 2-Mode-SIS model, the probability of a patient i to get infected from susceptible is formulated as a dose-response function proportional to the amount of pathogen on the patient i, or $l_t(i)$, as shown in Procedure 1 step 8 with β as the disease infectivity). Once infected, the patient i sheds α additional pathogen per timestep to his own load, which can later be transferred to neighbors (both people and locations) via edges; this shedding continues until the patient recovers with recovery probability δ. Specifically, in this model, each node carries some amount of pathogens that changes over time, and the exchange of pathogens among the nodes is driven by edges that imply close contact among the nodes. 2-Mode-SIS model uses τ_{ijt} to denote the ratio of pathogen being transferred (or remaining if $i = j$) from node j to node i at time t, and ϕ_i to denote the natural pathogen reduction rate on node i. By constructing a transfer matrix R_t from the transfer ratios τ_{ijt}, reduction rates ϕ_i, and adjacency matrix A_t, the pathogen load updates can be written as a linear operation as in Procedure 1 step 5. Here, x_t is the infection state vector for all patients at time t, where the ith element, $x_t(i)$, corresponds to patient i. α is pathogen shedding rate for infected patients. Note that the column-sums of R_t are restricted to

be less than or equal to 1, which implies that the total amount of pathogen cannot increase after transfer (i.e., $\|R_t l_t\|_1 \leq \|l_t\|_1$). Susceptible patients may still be colonized with pathogen loads and can spread them to others. HCWs and locations are assumed to be non-infectable, but act as pathways of pathogen transfer. More details are provided in Supplementary Information.

Spectral characterization for the 2-MODE-SIS model

Approximation via NLDS

We develop a characterization to capture the overall dynamics of the system and derive a spectral characterization for the 2-MODE-SIS model. Our proposed approach is to approximate the propagation dynamics of the 2-MODE-SIS model using the stability analysis of discrete-time non-linear dynamical systems (NLDS) with two coupled states: the pathogen loads l_t and the expectation of infection probabilities $p_t = \mathbb{E}[x_t]$. Let $s_t = \begin{bmatrix} p_t \\ l_t \end{bmatrix} \in \mathbb{R}^{(P+N)}$ be the vector describing the state of the system at timestep t, we can write down the state transition updates at t and use linearity of expectation to get the vector for next timestep as $s_{t+1} = g_t(s_t)$, where g_t is a non-linear mapping.

$$s_{t+1} = g_t(s_t) := \begin{bmatrix} (1-\delta)p_t + \beta l_t(1 - p_t) \\ R_1 l_t + \alpha p_t \end{bmatrix}$$

Stability analysis of the NLDS

With the non-linear equations of our approximated discrete-time NLDS, we can then analyze its asymptotic behavior. The long-term behavior of a dynamical system is dictated by the stability of its equilibrium points, at which the states no longer change over time (i.e., $s^{eq} = g(s^{eq})$). The particular equilibrium point of interest is the *all-zero* point $\mathbf{0}$ where there is no infection as well as no pathogen remaining. Clearly, all g_t map $\mathbf{0}$ onto itself and $\mathbf{0}$ is an equilibrium point. Specifically, we write the Jacobian matrix evaluated at $\mathbf{0}$ as S_t

$$S_t = \nabla_g \big|_{S_t=0} = \frac{\partial s_{t+1}}{\partial s_t} \bigg|_{S_t=0} = \begin{bmatrix} (1-\delta)I_P & \beta I_P & 0 \\ \alpha I_P & R_1 \\ \mathbf{0} \end{bmatrix}$$

Here, I_P represents the identity matrix sized $P \times P$. The following Theorem 1 on asymptotic stability in a discrete-time NLDS shows what condition this equilibrium point is 'stable'[45], i.e. a small perturbation does not cause large deviations (see Supplementary Information for the proof).

Procedure 1 2-MODE-SIS model

1: **Inputs:** $\Theta = \{\alpha, \beta, \delta, \{A_t\}_{t=1}^T, \{\tau_{ijt}\}_{i,j,t}\}$
2: Initialize infection-states x_1 and loads l_1
3: Compute $R_t(i,j) = \begin{cases} \tau_{ijt} A_t(i,j) & \text{if } i \neq j \\ \tau_{ijt} \phi_i & \text{if } i = j \end{cases}$ for all t.
4: **for** $t = 1, \ldots, T$ **do**
5:     Update loads $l_{t+1} = R_t l_t + \alpha x_t$.
6:     **for** each patient i **do**
7:         **if** i is susceptible at time t (i.e., $x_t(i) = 0$) **then**
8:             i gets infected (i.e. $x_{t+1}(i) = 1$) with probability $\beta l_t(i)$.
9:         **else**
10:             i gets susceptible (i.e. $x_{t+1}(i) = 0$) with probability δ.
11:         **end if**
12:     **end for**
13: **end for**

Theorem 1 (Spectral characterization) *If the spectral radius (i.e., the largest eigenvalue) of the system matrix $S = \prod_{t=1}^T S_{T-t+1}$ is less than 1 (i.e. $\rho(S) < 1$), then the all-zero equilibrium of the NLDS is asymptotically stable.*

Theorem 1 demonstrates that $\rho(S) = 1$ is a critical tipping point that there will be a phase transition for the fraction of infected patients at the end: when $\rho(S) > 1$, there is always a fraction of infected patients. When $\rho(S) < 1$, almost no patients will be infected (i.e., the epidemic dies out). The tipping point behavior is similar to that associated with the basic reproductive number R_0. The characterization in Theorem 1 also helps explain the counter-intuitive 'dilution effect' (i.e., the number of infections decreases with increasing link probability between patients) observed in many such load-based models[34, 51, 52]. More explanations and an intuitive example are provided in Supplementary Information.

Applications of $\rho(S)$ for clinical problems

Using $\rho(S)$ to formalize clinical problems
Next, we show how our spectral characterization $\rho(S)$ can help formulate the two clinical problems IP and CP. The intuition is that our spectral characterization well captures the transition dynamics of the HAI breakout, and a lower $\rho(S)$ value should correspond to fewer infections. We first formalize the IP problem as follows:

Formal Isolation Problem (FIP) Let $S(\Theta, P)$ denote the system matrix given input 2-MODE-SIS model parameters Θ (e.g., $\alpha, \beta, \delta \cdots$) and patients in set P under precaution. Given all patients X and a budget constraint k, find the subset of patients P^* for precaution s.t.

$$P^* = \arg\min_{P} \ \rho(S(\Theta, P)) \text{ subject to } P \subset X \text{ and } |P| = k$$

This formulation uses $\rho(S(\Theta, P))$ to capture how isolating patients in P will affect the epidemic. Here, we formally model bringing a patient i under precaution by zeroing out off-diagonal terms in the i-th row and column in each pathogen transfer matrix R_t (i.e., cut off its links with other nodes), and reducing the diagonal terms $R_t(i, i)$ by a multiplicative factor $\phi_{ppe} < 1$ (i.e., reduce its remaining pathogen via cleaning). They are widely used together in hospitals for contact precautions[11, 18, 53]. Note that the current setup can be readily modified to capture other contact precaution practices while maintaining a similar problem formulation. Similarly, we can also formalize the CP problem:

Formal Clearance Problem (FCP) Let $S(\cdot)$ be defined as in FIP. Given an initial set Q of d candidate patients already under precaution ($|Q| = d$) and a budget constraint $k < d$, find the subset of patients P^* for clearance s.t.

$$P^* = \arg\min_{P} \ \rho(S(\Theta, Q \setminus P)) \text{ subject to } P \subset Q \text{ and } |P| = k$$

GREEDY-SPECTRAL algorithm

Numerical issue in computing $\rho(S)$
A natural algorithm for FIP or FCP would try adding (or removing) each candidate patient P under (or from) precaution and then compute by how much $\rho(S)$ changes to incrementally build a feasible solution set while minimizing $\rho(S)$. In practice, however, we find that computing the change in $\rho(S)$ due to a single node removal runs into numerical issues: if there are much more patient-to-non-patient interactions than patient-to-patient interactions, the change in $\rho(S)$ w.r.t. P can be too small to be computed within finite precision. The issue even gets worse when the total number of nodes N and the number of timesteps T increase (see Supplementary Information for a detailed explanation).

Alternative strategy and final algorithm
Hence we propose an alternate strategy to solve FIP and FCP by minimizing an upper bound on $\rho(S)$ that can be computed more precisely. In particular, we define a *system-adjacency* matrix as $A := \prod_{t=1}^{T} A_{T-t+1}$, where A_t is the adjacency matrix for time t, and minimize its the largest eigenvalue $\rho(A)$ as a surrogate to minimize the upper bound of $\rho(S)$. The following theorem shows that $\rho(S)$ is upper bounded by a *monotonic* function of $\rho(A)$, justifying our approach of minimizing $\rho(A)$ instead. We provide a detailed proof for Theorem 2 in Supplementary Information.

Theorem 2 *Let S, A denote the system and system-adjacency matrices of 2-MODE-SIS model with T number of timesteps. Then $\rho(S)$ is upper bounded by a monotone function f of $\rho(A)$.*

Fortunately, $\rho(A)$ does not suffer from the same numerical issue as $\rho(S)$. Therefore, we propose a simple yet effective heuristic algorithm, GREEDY-SPECTRAL, for the two formulated problems: Starting from an empty set, we greedily and iteratively find and add a patient for isolation (or releasing) that leads to the smallest $\rho(A)$ value until the size of the set reaches the given budget.

Ethical statement

This project is approved by the IRB "Controlling healthcare associated infections: an in-silico framework" and meets the criteria of exempt research under 45CFR46.104(d)(4)iii. All methods were performed in accordance with the relevant guidelines and regulations. Ours was a retrospective study and hence informed consent was waived by the approving ethical committee.

Results

In this section, we demonstrate the performance of our GREEDY-SPECTRAL algorithm for clinical problems using MRSA (a kind of HAI) as an example. Here, we use two time-varying heterogeneous contact networks constructed from clinical metadata (e.g., inpatient Data, doctor's Notes, and MedAdmin Data) from the Epic-based SQL database at the University of Virginia (UVA) hospital to showcase our GREEDY-SPECTRAL algorithm could achieve robust good performance. One represents 294 days before COVID-19 (May 2019–Feb 2020), referred as the UVA-PRECOVID network, and another represents 294 days since the pandemic (May 2020–Feb 2021), referred as the UVA-COVID network. In these two networks, MRSA rates significantly declined in the UVA-COVID period, while networks became slightly denser, though many local properties (e.g., degree distribution, local clustering)
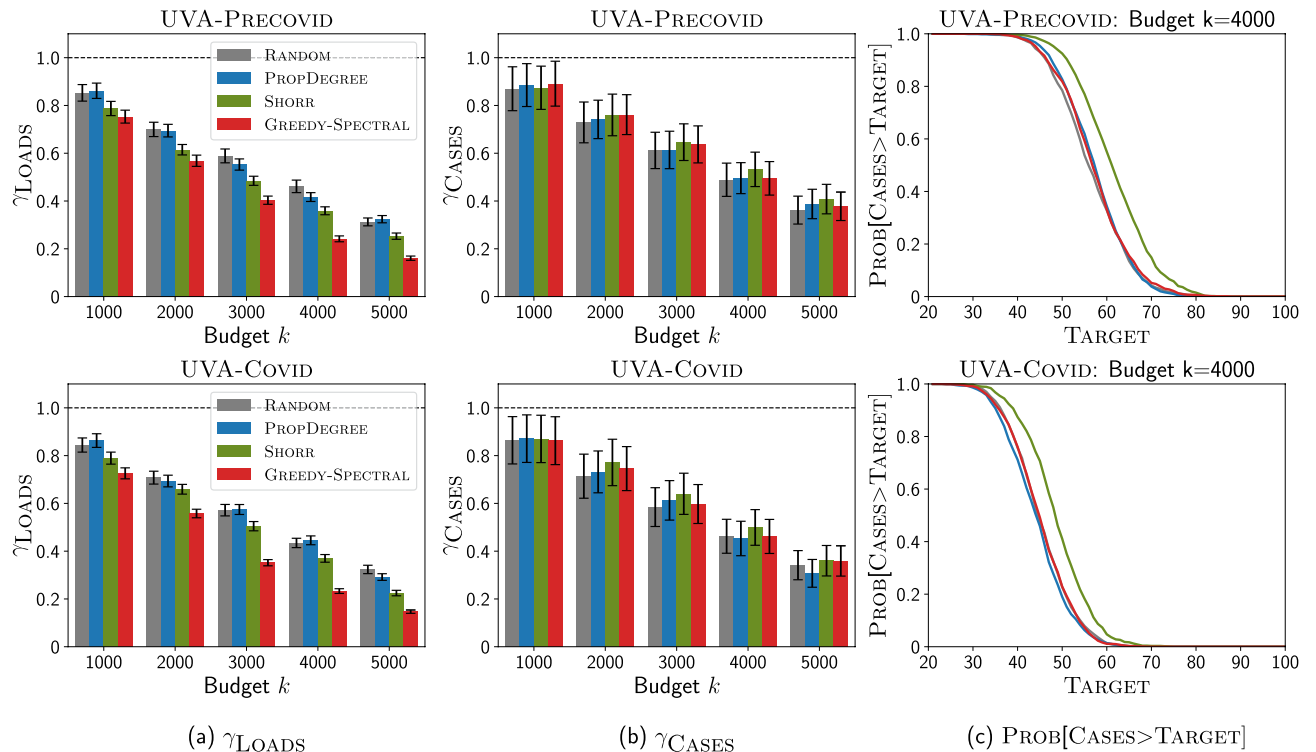
(a) γ_{LOADS}

(b) γ_{CASES}

(c) PROB[CASES>TARGET]

Figure 3. GREEDY-SPECTRAL achieves lower non-patient pathogen loads but similar number of new MRSA cases than other baselines under high precaution and cleaning scenario for isolation problem (IP). Top is for UVA-PRECOVID dataset, down is for UVA-COVID dataset. (**a**) Ratio of non-patient pathogen loads γ_{Loads} (y-axis) with varying budget k (x-axis). Lower is better. The black bars show the standard error. (**b**) Ratio of new MRSA cases γ_{Cases} (y-axis) with varying budget k (x-axis). Lower is better. (**c**) Prob[Cases > Target] (y-axis) with varying TARGET (x-axis) for budget $k = 4000$. Lower is better. See Supplementary Information for more experiment results.

were the same[54]. We first demonstrate the performance of GREEDY-SPECTRAL algorithm and several competing baselines in suppressing the MRSA outbreak for both high and relaxed precaution and cleaning scenarios (we will explain the two scenarios in later sections). We then investigate the patients picked by GREEDY-SPECTRAL using their electronic health records (EHRs) to see if they are meaningful from a clinical perspective.

Performance for clinical problems
High precaution and cleaning scenario
We first use the parameterization by calibrating 2-MODE-SIS model to the weekly number of new tested positive MRSA cases collected from EHRs data for experiments (see Supplementary Information fore more details). This scenario corresponds to the normal level of precaution and cleaning practices in healthcare settings[55], as in UVA, to avoid the outbreak of HAI including MRSA. We discuss the performance of GREEDY-SPECTRAL and other baselines under this scenario for the isolation problem (we show the results for clearance problem in Supplementary Information.)

Figure 3 shows how GREEDY-SPECTRAL performs in minimizing non-patient pathogen loads and the number of new MRSA cases compared with other baselines: (1) RANDOM will randomly pick k patients, (2) PROPDEGREE will pick k patients with the probability of picking patients i proportional to its degree, (3) SHORR will pick k patients with the highest so-called Shorr scores[56] (see Supplementary Information for more details). Note that SHORR was developed to assess the clinical risk of an admitted patient acquiring MRSA, better performance compared with SHORR means GREEDY-SPECTRAL is a more effective contact precaution strategy. As shown in Fig. 3a, GREEDY-SPECTRAL is more effective in reducing the amount of non-patient pathogen loads than other baselines. Here, γ_{Loads} is defined as the ratio of non-patient (HCWs and location) pathogen loads against $k = 0$ situation. For example, when $k = 5000$, γ_{Loads} for GREEDY-SPECTRAL is $\frac{\text{Loads}_{k=5000,\text{Greedy-Spectral}}}{\text{Loads}_{k=0}}$, where $\text{Loads}_{k=5000,\text{Greedy-Spectral}}$ and $\text{Loads}_{k=0}$ are the sum of HCWs pathogen loads and location pathogen loads when selecting 5000 patients for contact precaution using GREEDY-SPECTRAL, and when no patients are selected ($k = 0$) respectively. In Fig. 3a, when $k = 5000$, γ_{Loads} for GREEDY-SPECTRAL is 9.2% (7.8%) lower than for SHORR on UVA-PRECOVID (UVA-COVID), this indicates that GREEDY-SPECTRAL is more effective in reducing the non-patient pathogen loads than other baselines (similar conclusion can also be drawn for other budget k). However, in Fig. 3b, we observe that our algorithm still gives a similar ratio of new MRSA cases γ_{Cases}, which is defined similarly as the ratio of the number of new MRSA cases against $k = 0$ situation. In Fig. 3c, it also gives similar probability that the number of new MRSA cases is larger than TARGET, Prob[Cases > Target] for varying Target (x-axis) as other baselines.
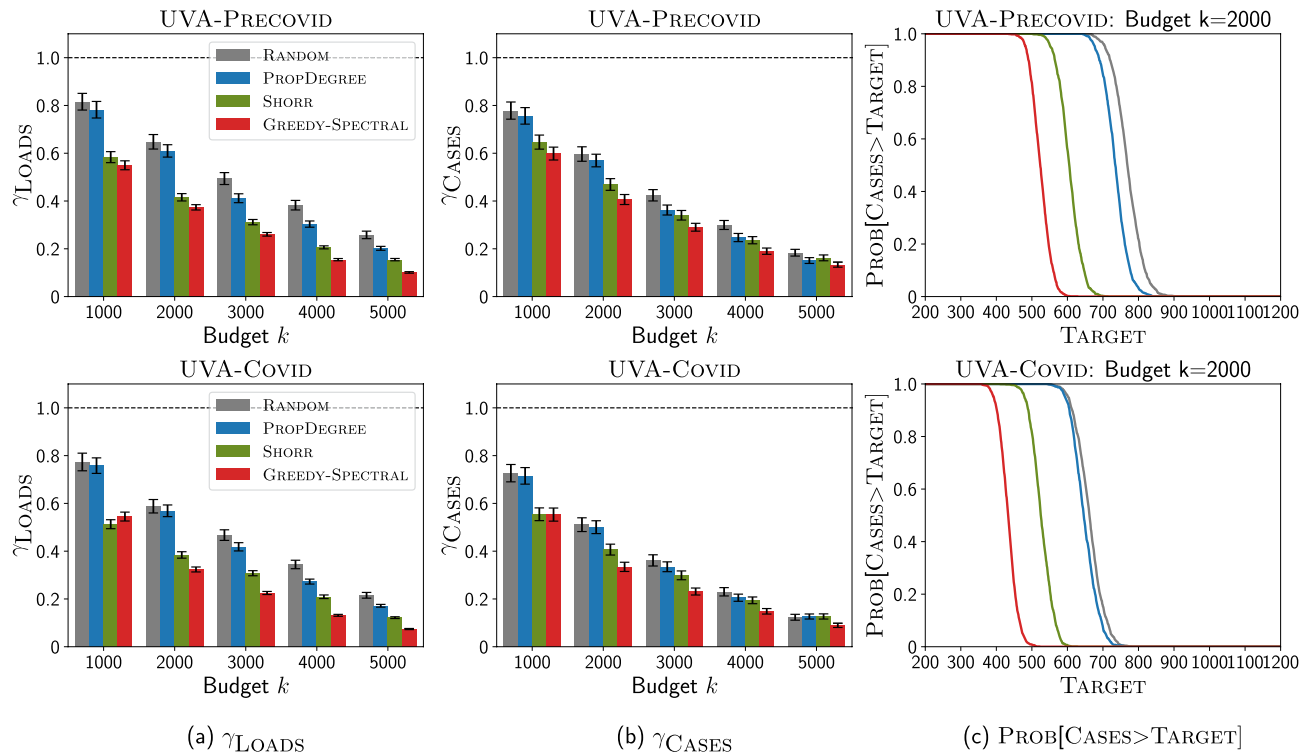
Figure 4. GREEDY-SPECTRAL achieves lower non-patient pathogen loads and number of new MRSA cases than other baselines under relaxed precaution and cleaning scenario for isolation problem (IP). Top is for UVA-PRECOVID dataset, down is for UVA-COVID dataset. (**a**) Ratio of non-patient pathogen loads γ_{Loads} (y-axis) with varying budget k (x-axis). (**b**) Ratio of new MRSA cases γ_{Cases} (y-axis) with varying budget k (x-axis). (**c**) Prob[Cases > Target] (y-axis) with varying TARGET (x-axis) for budget $k = 2000$. See Supplementary Information for more experiment results.

Here, lower loads achieved by GREEDY-SPECTRAL are not leading to fewer cases, which is likely because of the isolation practices currently in use.

Relaxed precaution and cleaning scenario
Next, we investigate how Greedy-Spectral and other baselines perform in settings where precautions and cleaning practices are relaxed. Hospitals all over the world have already faced shortages induced by the COVID-19 pandemic, including personal protective equipment, nasal swabs for testing, laboratory diagnostic capacity, and medical personnel, e.g.,[57, 58]. Further, shortages in nursing staff have predated the COVID-19 pandemic, and is known to be associated with increased risk of HAI[59]. In the event of such shortages, it is hard to maintain the same level of precautions and cleaning policies, and the number of cases is likely to increase. Optimizing interventions becomes even more important in this setting. In this paper, we use the pathogen shedding rate α in the 2-MODE-SIS model to capture the influence of relaxed cleaning practices, and we use α leading to different numbers of cases when $k = 0$ to capture different levels of relaxation of cleaning practices.

Isolation problem results We first focus on isolation problem (IP) under relaxed precaution and cleaning scenario. Here, we demonstrate our GREEDY-SPECTRAL algorithm and other baselines by varying k but fixing α such that the number of cases is 1000 when $k = 0$. We provide more results in Supplementary Information.

Fig. 4 shows the results with varying k. As shown in Fig. 4, GREEDY-SPECTRAL achieves lower non-patient pathogen loads and number of new MRSA cases than other baselines. In Fig. 4a, it achieves lower γ_{Loads}. For example, when $k = 2000$, it leads to 4.2% (6.1%) lower loads than SHORR, and 23.6% (24.5%) lower loads than PROPDEGREE on UVA-PRECOVID (UVA-COVID). Furthermore, our algorithm also leads to lower γ_{Cases} as shown in Fig. 4b. When $k = 2000$, it achieves 6.3% (7.2%) lower cases than SHORR, and 16.3% (16.6%) lower cases than PROPDEGREE on UVA-PRECOVID (UVA-COVID). This indicates that our algorithm leads to around 60–70 fewer MRSA cases than SHORR and 160–170 fewer cases than PROPDEGREE (since the number of cases is 1000 when $k = 0$). In Fig. 4c, we show how Prob[Cases > Target] changes with varying Target. Here, GREEDY-SPECTRAL always has lowest Prob[Cases > Target]. For example, for UVA-PRECOVID (top row) and Target = 600, Prob[Cases > Target] for our algorithm is 0, indicating that it can always lead to less than 600 cases. However, SHORR has 56.1% probability of having more than 600 cases, and PROPDEGREE and RANDOM have 100% probability of having more than 600 cases. This demonstrates the effectiveness of GREEDY-SPECTRAL algorithm in leading to fewer cases.

Clearance problem results We repeat the experiments shown for isolation problem under relaxed precaution and cleaning scenario above to clearance problem. Similarly, we demonstrate our GREEDY-SPECTRAL algorithm
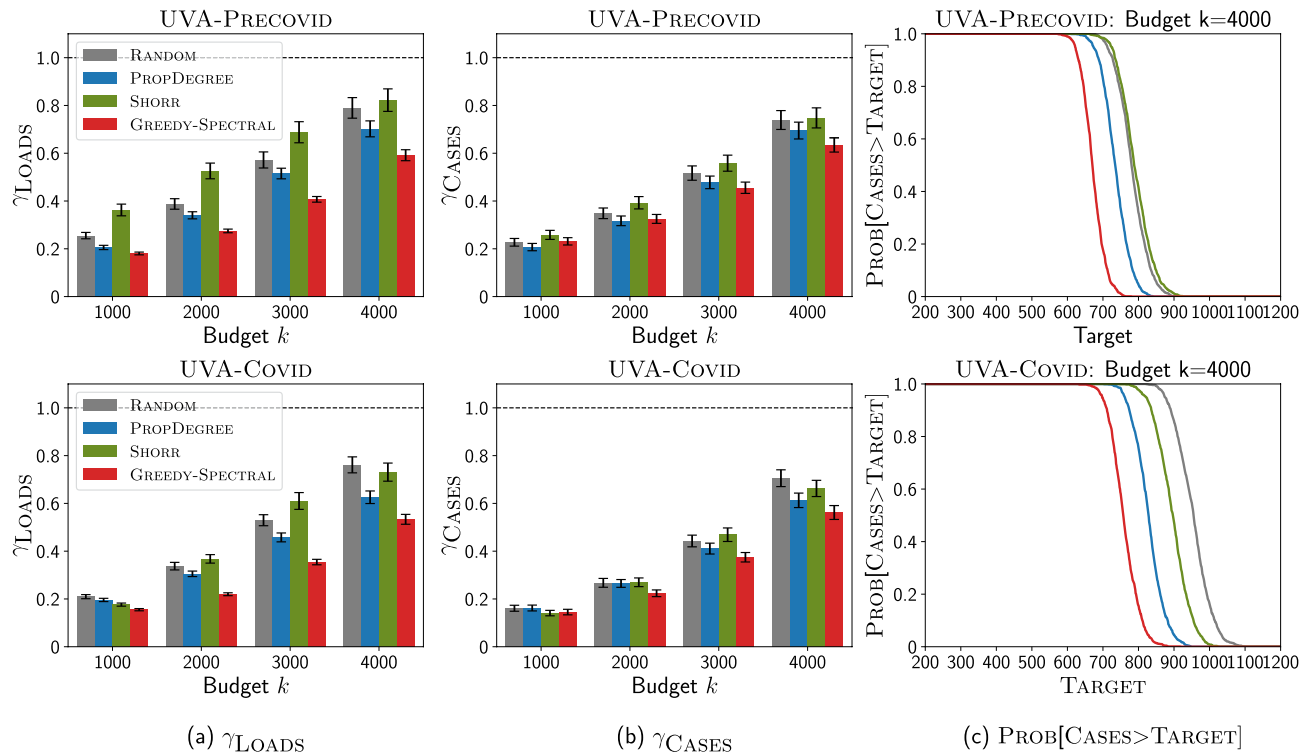
7

Figure 5. GREEDY-SPECTRAL achieves lower non-patient pathogen loads and number of new MRSA cases than other baselines under relaxed precaution and cleaning scenario for clearance problem (CP). Top is for UVA-PRECOVID dataset, down is for UVA-COVID dataset. (**a**) Ratio of non-patient pathogen loads γ_{Loads} (y-axis) with varying releasing budget k (x-axis). (**b**) Ratio of new MRSA cases γ_{Cases} (y-axis) with varying releasing budget k (x-axis). (**c**) Prob[Cases > Target] (y-axis) with varying TARGET (x-axis) for releasing budget $k = 4000$. See Supplementary Information for more experiment results.

and other baselines by varying k but fixing α such that the number of cases is 1000 when $k = 0$. We also provide more results in Supplementary Information.

Figure 5 shows the results with varying k. As shown in Fig. 5, GREEDY-SPECTRAL achieves lower non-patient pathogen loads and number of new MRSA cases than other baselines for the clearance problem. In Fig. 5a, it achieves lower non-patient pathogen loads. Furthermore, it also leads to a lower number of new MRSA cases as shown in Fig. 5b. For example, when $k = 4000$ (i.e. releasing 4000 patients) for UVA-COVID (bottom row), our algorithm leads to 8.5% lower loads and 6.9% lower γ_{Cases} than other baselines. Besides, it also has lowest Prob[Cases > Target] in Fig. 5c. For example for UVA-PRECOVID (top row), when releasing $k = 4000$ patients and Target = 700, GREEDY-SPECTRAL has only 12.3% probability of having more than 700 cases. Instead, other baselines have 100% probability of having more than 700 cases.

Comparison with clinical heuristic precaution policies
Many clinical heuristic precaution policies are also used in practice[60, 61]. Usually, these policies tend to pick patients with high MRSA risks. For example, patients with surgery history or invasive device use may be considered as having high risk and may be picked for precaution. We next leverage EHRs to implement these clinical heuristic policies and compare them with our GREEDY-SPECTRAL algorithm. The EHRs of patients in the UVA dataset include information like surgery history or underlying diseases (for brevity, we call them 'features') for each patient. Specifically, we compare with the following four heuristic policies: isolate all emergency patients, isolate all patients with surgery history within 90 days, isolate all patients with invasive device use history within 90 days, and isolate all patients with dialysis history within 90 days. The four features used in these policies are all considered to be highly risky[56, 60, 61]. As shown in Fig. 6 for UVA-PRECOVID, GREEDY-SPECTRAL can also suppress MRSA outbreak better than these clinical heuristic precaution policies. Here, horizontal dashed lines represent the simulated number of new cases for these four policies. Scattered dots correspond to the number of patients isolated by each policy. We can see the line for our algorithm is below all scattered dots. This indicates that when isolating the same number of patients, GREEDY-SPECTRAL leads to fewer MRSA cases than these policies.

Case study: analysis on isolated patients

We also investigate the patients picked by our GREEDY-SPECTRAL algorithm from a clinical perspective. In Table 1, we use PROPDEGREE, SHORR, and GREEDY-SPECTRAL to pick 500 patients for precaution for UVA-PRECOVID, and then show the fraction of patients with highly risky features[60, 61] (note that one patient may have multiple features and can be counted multiple times). We also list the fraction of patients with these features in
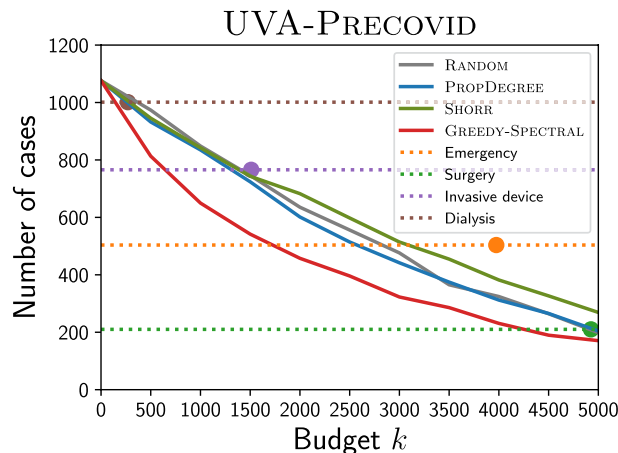
Figure 6. Comparison against clinical heuristic precaution policies. The x-axis is budget *k*, the y-axis is the number of cases. The horizontal dashed lines represent the simulated number of MRSA cases with the these policies, and scattered dots correspond to the number of patients isolated by these policies.

Features	Population	PropDegree	Shorr	Greedy-Spectral
Surgery (within 90 days)	0.601	0.614	0.696	0.750
Surgery (90+ days)	0.377	0.434	0.572	0.572
Invasive Device (within 90 days)	0.233	0.264	0.472	0.470
Invasive Device (90+ days)	0.171	0.180	0.356	0.382
Dialysis (within 90 days)	0.041	0.050	0.184	0.136
Dialysis (90+ days)	0.031	0.042	0.160	0.096
ICU	0.281	0.326	0.508	0.402

Table 1. Fraction of patients with different features.

Notations	UVA-Precovid	UVA-Covid
Surgery (within 90 days)	0.750	0.752
Surgery (90+ days)	0.572	0.6
Invasive device (within 90 days)	0.470	0.532
Invasive device (90+ days)	0.382	0.342
Dialysis (within 90 days)	0.136	0.164
Dialysis (90+ days)	0.096	0.086
ICU	0.402	0.48

Table 2. Fraction of patients with different features for Greedy-Spectral on UVA-Precovid and UVA-Covid dataset.

population for comparison. As shown in Table 1, our algorithm GREEDY-SPECTRAL picks meaningful patients with these features for precaution. For example, compared with PROPDEGREE and fraction in population, it picks many more patients with invasive device use history and dialysis history. Compared with SHORR, GREEDY-SPECTRAL highlights patients having surgery within 90 days more while emphasizing dialysis patients less. Additionally, GREEDY-SPECTRAL also captures the structural information of the heterogeneous contact networks well since the average degree of the patients picked by it is larger than other baselines. This explains why it performs better than SHORR in suppressing MRSA outbreaks. *Note that our algorithm uses only network information, and not patient level feature information, but it still identifies the correct patients at high risk and performs better in minimizing the MRSA spread as shown above.*

In Table 2, we investigate the differences in the type of patients picked in the UVA-PRECOVID and UVA-COVID period. Here, we list the fraction of patients with highly risky features picked by GREEDY-SPECTRAL in both periods. As shown in Table 2, our algorithm is picking 13.2% more patients with invasive device use history within 90 days, 20.6% more patients with dialysis history within 90 days and 19.4% more patients in

ICU in UVA-Covid than UVA-Precovid. For example, the increasing number of patients picked in ICU by our algorithm can be explained by the fact that ICU is getting more crowded. According to the EHRs data, there are more contacts in the UVA-Covid period than in the UVA-Precovid period (average degree of patients in ICU: 115.33 (UVA-Precovid) va. 125.24 (UVA-Covid)). Additionally, the average degree of the patients picked by our algorithm for UVA-Covid period is also larger than UVA-Precovid.

Discussion

We develop a spectral characterization for 2-Mode-SIS[34] for modeling the spread of HAIs, such as MRSA pathogens, which captures the long-term dynamics of pathogen spread by taking into account both the contact network structure and disease parameters. This model has been proposed as an alternative to the standard SIS models, but has not been studied rigorously before. We also study two important clinical problems arising from control of HAIs, and show how these can be formalized using our spectral characterization. We design Greedy-Spectral, an algorithm for solving these clinical problems, which greedily and iteratively finds and adds a patient to isolate (or relese) until reaching the budget.

Overall, our results demonstrate that our spectral characterization and Greedy-Spectral algorithm hold sufficient potential for minimizing the spread of MRSA. In experiments, we compare Greedy-Spectral with several baselines including Shorr, which picks patients for contact precautions using their features collected from Electronic Health Records. For the high precaution and cleaning scenario, which corresponds to the normal level of pathogens and cleaning practices in healthcare settings, our simulations suggest that the Greedy-Spectral algorithm outperforms the Shorr by achieving lower non-patient pathogen loads. For the relaxed precaution and cleaning scenario, which captures the shortage of resources (e.g., personal protective equipment, laboratory diagnostic capacity, and medical personnel in the COVID-19 pandemic[57, 58]), our simulations indicate that Greedy-Spectral algorithm not only reduces pathogen loads but also leads to fewer MRSA cases compared with Shorr. Additionally, our algorithm uses only the network information and not the patient-level feature information. However, it still identifies the correct patients at high risk and performs better in minimizing the MRSA spread as shown above.

Our Greedy-Spectral algorithm is likely to be helpful in controlling MRSA and other healthcare-associated infections. As demonstrated in our simulations, our algorithm uses only the network information rather than patient-level features, yet it successfully identifies high-risk patients and outperforms Shorr and other baselines in minimizing MRSA spread. In practice, the Greedy-Spectral algorithm can be implemented by constructing contact networks using EHR data such as inpatient data, doctors' notes, and MedAdmin data, which are as similarly available as patient-level features. Some research also suggests that wearable RFID sensors can be used to construct networks[62, 63]. With the contact networks, people can then run our Greedy-Spectral algorithm to decide the most suitable patients to put under or released from precaution (corresponding to the isolation problem and clearance probelm respectively). We expect our approach will assist epidemiologists and clinicians in identifying appropriate patients for contact precautions, ultimately aiding in the control of MRSA transmission. Our spectral characterization and Greedy-Spectral algorithm can also be extended to other diseases and epidemiological models.

One of the limitations of this work is that the 2-Mode-SIS model we use involves only two states: susceptible and infected. Real-world MRSA infections, however, manifest in various forms (e.g., bloodstream infections and pneumonia)[64], while the 2-Mode-SIS model simplifies them into a single infected state. In addition, the 2-Mode-SIS model assumes that any susceptible patient can be colonized instead of explicitly incorporating a colonization state in it itself. One potential extension of this work could extend our spectral characterization analysis to more complex models that could distinguish different kinds of MRSA infections and colonization. Our approach can also be adapted for other healthcare-associated infections, such as *C. difficile*, by substituting the 2-Mode-SIS model with other models. Another limitation is that we assume patients selected for contact precautions remain isolated throughout the entire period. However, in actual clinical practice, patients may be released from contact precautions when they recover or are discharged from the hospital. Future work could consider the possibility of isolating or releasing patients at any time step to gain more flexibility. Although contact networks and other parts of EHR data are very sensitive datasets, the privacy concerns can be mitigated by running our Greedy-Spectral algorithm on anonymous contact networks to estimate the impact of the optimal solution, and use characteristics of the solution in defining actual strategies.

Data availability

The simulation outputs of our model are available in https://doi.org/10.5281/zenodo.7811397. The electronic health record (EHR) data used in developing the models is not available since it is highly sensitive, and we do not have permission to release it.

Code availability

Code have been deposited in https://doi.org/10.5281/zenodo.7811397.

References

1. Kallen, A. J. *et al.* Health care-associated invasive mrsa infections, 2005–2008. *JAMA* **304**, 641–647 (2010).
2. Fernández-Gracia, J., Onnela, J.-P., Barnett, M. L., Eguíluz, V. M. & Christakis, N. A. Influence of a patient transfer network of us inpatient facilities on the incidence of nosocomial infections. *Sci. Rep.* **7**, 2930 (2017).
3. Rocha, L. E. *et al.* Dynamic contact networks of patients and mrsa spread in hospitals. *Sci. Rep.* **10**, 1–10 (2020).

4. Stone, P. W. Economic burden of healthcare-associated infections: An American perspective. *Expert Rev Pharmacoecon. Outcomes Res.* **9**, 417–422 (2009).
5. Sharma, A. *et al.* The cost of contact precautions: A systematic analysis. *Can. J. Infect. Control* **35**, 163–169 (2020).
6. Roth, J. A. *et al.* Direct costs of a contact isolation day: .A prospective cost analysis at a swiss university hospital. *Infect. Control Hosp. Epidemiol* **39**, 101–103 (2018).
7. Valiquette, L., Chakra, C. N. A. & Laupland, K. B. Financial impact of health care-associated infections: When money talks. *Can. J. Infect. Dis. Med. Microbiol.* **25**, 71–74 (2014).
8. Howden, B. P. *et al.* Staphylococcus aureus host interactions and adaptation. *Nat. Rev. Microbiol.* **21**, 380–395 (2023).
9. Lowy, F. D. Staphylococcus aureus infections. *N. Engl. J. Med.* **339**, 520–532 (1998).
10. Sjoberg, A. M. *et al.* One health interprofessional stewardship to combat antimicrobial resistance. *Nat. Med.* **29**, 512–513 (2023).
11. Marshall, C., Richards, M. & McBryde, E. Do active surveillance and contact precautions reduce mrsa acquisition? A prospective interrupted time series. *PLoS ONE* **8**, e58112 (2013).
12. Haque, M. *et al.* Strategies to prevent healthcare-associated infections: A narrative overview. *Risk Manag. Healthc. Policy* **13**, 1765–1780 (2020).
13. Pei, S., Liljeros, F. & Shaman, J. Identifying asymptomatic spreaders of antimicrobial-resistant pathogens in hospital settings. *Proc. Natl. Acad. Sci.* **118**, e2111190118 (2021).
14. Karanika, S., Paudel, S., Grigoras, C., Kalbasi, A. & Mylonakis, E. Systematic review and meta-analysis of clinical and economic outcomes from the implementation of hospital-based antimicrobial stewardship programs. *Antimicrob. Agents Chemother.* **60**, 4840–4852 (2016).
15. Mouajou, V., Adams, K., DeLisle, G. & Quach, C. Hand hygiene compliance in the prevention of hospital-acquired infections: A systematic review. *J. Hosp. Infect.* **119**, 33–48 (2022).
16. Pittet, D. *et al.* Evidence-based model for hand transmission during patient care and the role of improved practices. *Lancet. Infect. Dis* **6**, 641–652 (2006).
17. Gawande, A. On washing hands. *N. Engl. J. Med.* **350**, 1283–1286 (2004).
18. Siegel, J. D., Rhinehart, E., Jackson, M. & Chiarello, L. 2007 guideline for isolation precautions: Preventing transmission of infectious agents in health care settings. *Am. J. Infect. Control* **35**, S65–S164 (2007).
19. Doebbeling, B. N. & Wenzel, R. P. The direct costs of universal precautions in a teaching hospital. *JAMA* **264**, 2083–2087 (1990).
20. Martin, E. M. *et al.* Noninfectious hospital adverse events decline after elimination of contact precautions for mrsa and vre. *infection control & hospital epidemiology* **39**, 788–796 (2018).
21. Evans, H. L. *et al.* Contact isolation in surgical patients: A barrier to care?. *Surgery* **134**, 180–188 (2003).
22. Day, H. *et al.* Do contact precautions cause depression? A two-year study at a tertiary care medical centre. *J. Hosp. Infect.* **79**, 103–107 (2011).
23. Mehrotra, P. *et al.* Effects of contact precautions on patient perception of care and satisfaction: A prospective cohort study. *Infect. Control Hosp. Epidemiol.* **34**, 1087–1093 (2013).
24. England, P. H. Guidance for stepdown of infection control precautions and discharging covid-190 patients (2020).
25. Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146 (2003).
26. Tong, H., Prakash, B. A., Eliassi-Rad, T., Faloutsos, M. & Faloutsos, C. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 245–254 (2012).
27. Zhang, Y. & Prakash, B. A. Dava: Distributing vaccines over networks under prior information. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, 46–54 (SIAM, 2014).
28. Bhalla, A. *et al.* Acquisition of nosocomial pathogens on hands after contact with environmental surfaces near hospitalized patients. *Infect. Control Hosp. Epidemiol.* **25**, 164–167 (2004).
29. Albrich, W. C. & Harbarth, S. Health-care workers: Source, vector, or victim of MRSA?. *Lancet. Infect. Dis* **8**, 289–301 (2008).
30. Snyder, G. M. *et al.* Detection of methicillin-resistant staphylococcus aureus and vancomycin-resistant enterococci on the gowns and gloves of healthcare workers. *Infect. Control Hosp. Epidemiol.* **29**, 583–589 (2008).
31. Hu, B., Shi, C., Zhao, W. X. & Yu, P. S. Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1531–1540 (2018).
32. Sun, Y., Han, J., Yan, X., Yu, P. S. & Wu, T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proc. VLDB Endow.* **4**, 992–1003 (2011).
33. Zhang, F., Wang, M., Xi, J., Yang, J. & Li, A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* **8**, 1–9 (2018).
34. Jang, H. *et al.* Evaluating architectural changes to alter pathogen dynamics in a dialysis unit. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 961–968 (IEEE, 2019).
35. Chandrasekaran, S. & Jiang, S. C. A dose response model for staphylococcus aureus. *Sci. Rep.* **11**, 1–10 (2021).
36. Adhikari, B., Lewis, B., Vullikanti, A., Jiménez, J. M. & Prakash, B. A. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. *PLoS Comput. Biol.* **15**, e1007284 (2019).
37. Montella, E. *et al.* Healthcare associated infections in the neonatal intensive care unit of the "federico ii" university hospital: Statistical analysis and study of risk factors. In *2021 International Symposium on Biomedical Engineering and Computational Biology*, BECB 2021, https://doi.org/10.1145/3502060.3503650 (Association for Computing Machinery, 2022).
38. Plipat, N., Spicknall, I. H., Koopman, J. S. & Eisenberg, J. N. The dynamics of methicillin-resistant staphylococcus aureus exposure in a hospital model and the potential for environmental intervention. *BMC Infect. Dis.* **13**, 1–11 (2013).
39. Rubin, M. A. *et al.* A simulation-based assessment of strategies to control clostridium difficile transmission and infection. *PLoS ONE* **8**, e80671 (2013).
40. Nelson, R. E. *et al.* An economic analysis of strategies to control clostridium difficile transmission and infection using an agent-based simulation model. *PLoS ONE* **11**, e0152248 (2016).
41. Coia, J. *et al.* Guidelines for the control and prevention of meticillin-resistant staphylococcus aureus (MRSA) in healthcare facilities. *J. Hosp. Infect.* **63**, S1–S44 (2006).
42. of Disease Prevention, O. & Promotion, H. Health Care-associated Infections. https://www.cdc.gov/mrsa/community/environment/index.htm.
43. Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E. & Guo, R. The independent cascade and linear threshold models. In *Diffusion in Social Networks* (eds Shakarian, P. *et al.*) 35–48 (Springer, 2015).
44. Cabrera, M., Córdova-Lepe, F., Gutiérrez-Jara, J. P. & Vogt-Geisse, K. An sir-type epidemiological model that integrates social distancing as a dynamic law based on point prevalence and socio-behavioral factors. *Sci. Rep.* **11**, 10170 (2021).
45. Prakash, B. A., Chakrabarti, D., Faloutsos, M., Valler, N. & Faloutsos, C. Threshold conditions for arbitrary cascade models on arbitrary networks. In *2011 IEEE 11th International Conference on Data Mining*, 537–546 (IEEE Computer Society, 2011).
46. Hasan, A. *et al.* A new estimation method for covid-19 time-varying reproduction number using active cases. *Sci. Rep.* **12**, 6675 (2022).
47. Liu, Q.-H. *et al.* Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl. Acad. Sci.* **115**, 12680–12685 (2018).

48. Alvarez, L., Colom, M., Morel, J.-D. & Morel, J.-M. Computing the daily reproduction number of covid-19 by inverting the renewal equation using a variational technique. *Proc. Natl. Acad. Sci.* **118**, e2105112118 (2021).
49. Inglesby, T. V. Public health measures and the reproduction number of sars-cov-2. *JAMA* **323**, 2186–2187 (2020).
50. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
51. Pujol, J. M., Eisenberg, J. E., Haas, C. N. & Koopman, J. S. The effect of ongoing exposure dynamics in dose response relationships. *PLoS Comput. Biol.* **5**, e1000399 (2009).
52. Schmidt, K. A. & Ostfeld, R. S. Biodiversity and the dilution effect in disease ecology. *Ecology* **82**, 609–619 (2001).
53. Bartsch, S. M. *et al.* Modeling interventions to reduce the spread of multidrug-resistant organisms between health care facilities in a region. *JAMA Netw. Open* **4**, e2119212–e2119212 (2021).
54. Madden, G. *et al.* Deciphering COVID-19-associated effects on hospital MRSA transmission and social networks. *Open Forum Infect. Dis.* **8**, S104–S106. https://doi.org/10.1093/ofid/ofab466.173 (2021).
55. for Disease Control, C. & Prevention. Multidrug-Resistant Organisms (mdro) Management. https://www.cdc.gov/infectioncontrol/guidelines/mdro/index.html.
56. Shorr, A. F. *et al.* A risk score for identifying methicillin-resistant staphylococcus aureus in patients presenting to the hospital with pneumonia. *BMC Infect. Dis.* **13**, 1–7 (2013).
57. West, N. E., Cheong, W.-F., Boone, E. & Moat, N. E. Impact of the covid-19 pandemic: A perspective from industry. *Euro. Heart J. Suppl.* **22**, P56–P59 (2020).
58. Knezevic, C. E. *et al.* Rising to the challenge: Shortages in laboratory medicine. *Clin Chem.* **8**(12), 1486–1492 (2022).
59. Nurse staffing and health care-associated infections. Proceedings from a working group meeting. *Am. J. Infect. Control* **30**, 199–206. https://doi.org/10.1067/mic.2002.123416 (2002).
60. for Disease Control, C. & Prevention. Health care-associated infections. https://www.cdc.gov/mrsa/community/patients.html.
61. Naimi, T. S. *et al.* Comparison of community-and health care-associated methicillin-resistant staphylococcus aureus infection. *JAMA* **290**, 2976–2984 (2003).
62. Vanhems, P. *et al.* Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE* **8**, e73970 (2013).
63. Valdano, E., Poletto, C., Boëlle, P.-Y. & Colizza, V. Reorganization of nurse scheduling reduces the risk of healthcare associated infections. *Sci. Rep.* **11**, 1–12 (2021).
64. for Disease Control, C. & Prevention. Health Care-Associated Infections. https://www.cdc.gov/mrsa/community/index.html.

Author contributions

J.C., S.C., A.V., B.A.P. conceived the study; J.C., S.C., M.K., M.B. conducted the experiments; J.C., S.C., A.V., B.A.P. analysed the results; J.C., S.C., A.V., B.A.P. reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-41852-5.

Correspondence and requests for materials should be addressed to J.C. or B.A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.