Repeated Active Screening of Networks for Diseases

Biswarup Bhattacharya, Han Ching Ou
University of Southern California
Los Angeles, California, USA
bbhattac@usc.edu

Arunesh Sinha
University of Michigan
Ann-Arbor, Michigan, USA

Sze-Chuan Suen, Bistra Dilkina, Milind Tambe
University of Southern California
Los Angeles, California, USA

ABSTRACT

An important means of controlling recurrent infectious diseases is through active screening to detect and treat patients. Disease detection on a large network of individuals is a challenging problem, as the health states of individuals are uncertain and the scale of the problem renders traditional dynamic optimization models impractical. Moreover, efficient use of diagnostic and labor resources is a major concern, especially when the recurrent disease is prevalent in a resource-constrained region. In this paper, we propose a novel active screening model and an algorithm to facilitate active screening for recurrent diseases. Our contributions include: (1) A new approach for modeling SEIS type diseases using a novel belief-state representation, (2) a community and eigenvalue-based algorithm (TRACE) to perform multi-round active screening. We perform extensive experiments on real-world datasets which emulate human contact, and illustrate significant benefits due to TRACE.

CCS CONCEPTS

• **Computing methodologies → Multi-agent planning**; **Partially-observable Markov decision processes**; • **Applied computing → Life and medical sciences**;

KEYWORDS

Public health; SEIS Disease Model; Active Screening; Eigenvalue; Community; Belief states

1 INTRODUCTION

Curable infectious diseases are responsible for millions of deaths every year. Tuberculosis (TB), one such disease, affected over 10 million people worldwide in 2016, and caused over 400,000 deaths in India, the country with the highest TB mortality [28]. While low-cost treatment programs are available, many rely on patients to seek medical care (*passive screening*). However, individuals mistake their symptoms for another condition and not seek care. Public health agencies therefore engage in *active screening*, where individuals in the community are asked to undergo diagnostic tests and are offered treatment if tests return positive results [16].

It is costly to seek out at-risk individuals, and active screening efforts are often limited to high risk groups such as household TB contacts [9]. This method can successfully identify patients [3], and has been extensively evaluated [17]. However, this approach can be challenging to implement widely in resource-constrained regions such as India, as there are large transmission networks of potential patients and the number of health workers is limited. Prior studies show that even when focusing on high-risk TB groups in urban slums in India, the yield can be very small — only 0.8% of screened individuals were diagnosed with TB [9]. With an estimated 1 million undiagnosed TB cases in India, efficient active screening is the need of the hour [9].

Our *first contribution* is a model of the active screening problem which considers the underlying disease dynamics. We focus on recurrent infectious diseases with a latent stage (SEIS model of disease [26]), such as TB. Individuals can be susceptible (S) (currently healthy, but may become exposed), exposed (E), or infected (I). We consider diseases for which there is no means to achieve permanent immunity, either through vaccination or one time infection. As for TB, we assume treatment is effective for both exposed and infected individuals, making the individuals healthy (though again susceptible). Health workers are uncertain about the health state of individuals and have a small budget relative to population size for active screening. To the best of our knowledge, models of multi-round active screening for SEIS diseases are missing in the AI literature.

Our *second contribution* is a novel algorithm—**T**argeted **R**esolution of **A**ctive diseases using **C**ommunities and **E**igenvalues (TRACE)—to guide scalable active screening. In TRACE, we use network community structure to form a community graph, and then we select nodes to screen by maximizing the reduction of the largest eigenvalue of a variant of the community graph. TRACE takes into account the underlying disease dynamics and uncertainty of individuals' health states. TRACE is easily adaptable to most SEIS or SIS type diseases.

We illustrate the benefits of TRACE via extensive testing on real-world human contact networks against various baselines across a wide range of disease parameters (which also demonstrates its applicability to various other diseases).

2 DISEASE MODEL AND BACKGROUND

We first introduce the disease model notations for our problem. An individual can be in one of the following health states: S means that the individual is susceptible to disease (healthy), E means that the individual has been exposed and has latent disease, and I means that the individual is infected. We do not consider an explicit recovered or permanent immunity (R) state in our model, as this have been the focus of many prior studies. In diseases like Hepatitis A and measles, which follow a SEIR or SIR pattern, treated individuals may achieve permanent immunity by entering the recovered state [5, 24]. We focus on recurrent diseases, where permanent immunity is not possible (such as with TB, typhoid, and malaria), represented by SIS [1] or the more general SEIS [26] disease dynamics.

Disease Model: We adopt a SEIS model [26] for modeling the disease dynamics. TB and many other diseases follow a SEIS pattern, where treated individuals can relapse or become reinfected. The disease dynamics are therefore given by:

$$\text{Susceptible } (S) \xrightarrow{\alpha} \text{Exposed } (E)$$
$$\text{Exposed } (E) \xrightarrow{\beta} \text{Infected } (I)$$
$$\text{Infected } (I) \xrightarrow{c} \text{Susceptible } (S)$$

In the context of a graph of individuals, α is the edge-wise fixed probability of a susceptible (S) individual (node) being exposed (E) to the disease from an infected (I) neighbor, β is the fixed probability of an exposed (E) individual (node) becoming infected (I), and c is the probability of an infected (I) individual (node) voluntarily seeking and successfully completing treatment and returning to the susceptible S stage. We assume that the treatment takes place in one time period, where a period represents the duration needed for a complete treatment regimen (~half a year for TB).

Prior Approaches for Active Screening: Most previous work on active screening deals primarily with SIR or SEIR type diseases, often referred to as the *Vaccination Problem* [5, 24, 27, 30**?**], where permanent immunization (entry into R state) can be viewed as removing nodes from the graph [2, 20, 25]. Exploiting this idea, [20, 25] focus on immunization ahead of an epidemic and suggest a heuristic method of removing a set of k nodes based on the eigenvalues of the adjacency matrix. [30] considers the problem of selecting the best k nodes to immunize in a network after the disease has started to spread. These methods assume that the exact status of each node is known and deal with a single round of vaccination or screening. However, our paper focuses on multi-round active screening of SEIS diseases, where the complexity increases substantially due to lack of permanent immunity, existence of a latent stage, and uncertainty about the health states of all individuals. To the best of our knowledge, this complex setting has not been attempted previously in the AI literature. Generally, the problem of minimizing disease spread is different from the well-studied problem of influence maximization [**? ?**] as well, where one optimizes the selection of seeds or starting nodes for maximizing spread, as opposed to optimizing the selection of nodes on which to intervene in order to minimize spread.

3 ACTIVE SCREENING MODEL FORMULATION

Setup. We define k *active screening agents* that are to be deployed at every timestep t to diagnose and treat I and E individuals. Individuals are part of a contact network $G(V, E)$, and infection spreads via the edges in the network. There are $|V|$ individuals, and $N(i)$ denotes neighbors of individual i in the network. The network structure (graph) is known from the beginning ($t = 0$). Each individual (node) in the network is in one of the health states $\{S, E, I\}$. Let s_i^t denote the state of individual i at time t. In every round, the agents can either choose to screen a node i (action $a_i = 1$) or not ($a_i = 0$). Only k nodes can be screened in one round. A screened node is observed to be in state S, E, or I, and an unscreened node generates no observation. The agents maintain a belief about the state of every individual, starting with no information at $t = 0$. The beliefs about the health states evolve over time as the agents gain information about individuals (detailed later in this section).

Transition Dynamics. The probability of an individual undergoing a change in health state is given by:

$$T^0 = \begin{array}{c|ccc} & S & E & I \\ \hline S & q_j & 1-q_j & 0 \\ E & 0 & 1-\beta & \beta \\ I & c & 0 & 1-c \end{array},$$

$$T^1 = \begin{array}{c|ccc} & S & E & I \\ \hline S & q_j & 1-q_j & 0 \\ E & 1 & 0 & 0 \\ I & 1 & 0 & 0 \end{array},$$

$$\text{and } q_j = (1-\alpha)^{|\{k \in N(j) \mid s_k^t = I\}|}$$

where, T^0 is the probability matrix for non-screened individuals and T^1 is the probability matrix for screened individuals. The rows denote the state at time t and the columns denote the state at $t + 1$. The transition probabilities follow the disease dynamics described earlier. In particular, q_j captures the probability that node j does not become exposed from his infected neighbors $\{k \in N(j) \mid s_k^t = I\}$. Both I and E individuals who are screened can be treated, but we assume E individuals do not seek treatment voluntarily since their disease is latent unlike I individuals who seek treatment voluntarily with the probability c. For model simplicity, we assume S individuals cannot directly transition directly to I state. This is not an extreme assumption for TB, where the overall duration with latent TB can be much longer than the round length (6 months).

Objective. Finally, we define a *reward* function $R(s^t) = \sum_j R(s_j^t)$, where $R(s_j^t)$ is defined as follows.

$$R(s_j^t) = \begin{cases} +1, & s_j^t = S \\ 0, & \text{otherwise} \end{cases}$$

The objective of the model is to choose the budget-limited actions at each time step in order to maximize the number of susceptible individuals over T time-steps: $\max \sum_{t=0}^{T} R(s^t)$, interpreted as the total number of disease-free half years [12]. This is closely related to another well known public health metric — QALY [**? ?**], where additionally a +1 reward is given to E individuals and a +0.66 reward to I individuals. We focus on maximizing health outcomes in this study and leave cost considerations to future work. An important part of the model is our belief updating approach, which is described next.

Belief States. We do not know the true health states of every individual at all times perfectly. We therefore model our belief of node i's health state as $\mathbf{b}_i^t = [b_{i,S}^t, b_{i,E}^t, b_{i,I}^t]$, where $b_{i,j}^t$ is the probability node i is in state j. This marginal representation of health state belief for each node i addresses scalability issues, as representations of the joint distribution of health state beliefs over all nodes can be prohibitively large. We assume marginal beliefs $\mathbf{b}_i^t$'s can be updated independently at each node. Such independence assumptions have been made in prior literature on the spread of contagion [8, 18] and experimentally found to have a minimal effect on outcomes.

Belief Update. We assume perfect observability of the health state s_i^t of any node when it is screened. We cannot observe the health state of a node at time t if we do not screen it at time t.

We update the belief for each individual (node) i who voluntarily come to the clinic to an intermediate belief state $\bar{\mathbf{b}}_i^t = [0, 0, 1]$. We also update the beliefs of actively screened individuals to an intermediate belief state $\bar{\mathbf{b}}_i^t \sim s_i^t$. We update the intermediate beliefs of the remaining individuals as:

$$\bar{\mathbf{b}}_i^t = \frac{[b_{i,S}^t, b_{i,E}^t, (1-c)b_{i,I}^t]}{b_{i,S}^t + b_{i,E}^t + (1-c)b_{i,I}^t}$$

For each node i that voluntarily came to a clinic or was actively screened, the final belief update is: $\mathbf{b}_i^{t+1} = [1, 0, 0]$ because the node will be successfully treated and returned to the susceptible state if it was in E or I state. For the remaining nodes, we update to $\mathbf{b}_i^{t+1}$ as follows:

$$\mathbf{b}_i^{t+1} = \bar{\mathbf{b}}_i^t \, \mathbf{\Gamma^t}, \text{ where}$$

$$\mathbf{\Gamma^t} = \begin{bmatrix} w_i^t & 1 - w_i^t & 0 \\ 0 & 1 - \beta & \beta \\ c & 0 & 1 - c \end{bmatrix}, \; w_i^t = \prod_{j \in N(i)} (1 - \alpha \bar{b}_{j,I}^t).$$

This belief update procedure is an important and novel aspect of our proposed active screening model.

While the our model can be interpreted as a POMDP, it is slightly different from standard POMDP models, since in the active screening setting a screening action results in observing the current health states of the individual and not the individual's transitioned state. This difference can be handled straightforwardly, as in [4, 19] using a modified value iteration technique. However, we show in Section 6 that known POMDP approaches are not scalable for our problem.

4 MOTIVATION FOR TRACE

Given the problem setup, we motivate the need for the TRACE algorithm by showing that many prior approaches or simple extensions do not achieve the desired goal.

4.1 Eigenvalue Based Prior Approach

We first consider the circumstances under which diseases or epidemics die out on their own. In the absence of any intervention (action), the system is a discrete non-linear dynamical system. Such systems have been studied in prior work, and the following has been shown:

Proposition 1. *[18] Let λ_A^* denote the largest eigenvalue of the adjacency matrix A of the underlying graph, otherwise known as the spectral radius. Then, the epidemic dies out if and only if*

$$\frac{\alpha}{c} < \frac{1}{\lambda_A^*} \; \textit{and} \; \beta \neq 0 \; .$$

Remark: An observation is that the bound on λ_A^* above is same as derived for SIS model (without exposed E state) in earlier work [8]. This is because in the SEIS model, the E state must eventually become I if $\beta \neq 0$; thus, in the long run, E behaves similarly to I when $\beta \neq 0$ and there is no active intervention.

Permanent immunization can be viewed as removing nodes. Given the result above, one would wish to select the set of k nodes that reduces the largest eigenvalue the most. This is a NP-complete problem. [20, 25] suggest a heuristic that greedily removes k nodes one at a time, each time selecting the node that maximizes the reduction in the largest eigenvalue.

We also observe that the underlying problem is extremely hard to solve. In SIS networks, computing an individual's probability of infection and computing the expected number of infections are NP-hard [13, 21]. SIS is the relaxed version of the SEIS model, where $\beta = 1$. It is also known from [27] that given a network and limited resources, finding the optimal strategy for vaccinating a limited number of individuals (vaccination problem - SIR scenario), and quarantining a limited number of individuals (quarantining problem) are NP-hard. Also, given a network and limited resources, finding the optimal strategy for placement of a limited number of sensors for monitoring the course of an epidemic is NP-hard [21].

The Active Screening problem as defined in Section 3 is a generalized (harder) case of the above problems where we try to treat infected people without removing them from the graph since there is no permanent immunity and re-infection is possible (SEIS scenario). Based on Prop. 1, we also observe that a disease is unlikely to die out on its own in low-resource countries (c is low) with highly contagious diseases (high α), thus necessitating active screening.

4.2 Budgetary Threshold for Random Intervention

We can gain insight into how uncertainty in individuals' health states affects our problem by examining the fully-naive random screening strategy. We focus on the budget k, the number of nodes that can be screened and treated in one period. Intuitively, increasing k will lead to faster reduction of disease prevalence with random screening.

Lemma 1. *Assume that we know the infected patients belong to a set I_t in every round t such that $|I_t| \leq m$, where m is an arbitrary constant corresponding to the size of the network. Then, the epidemic dies out using k random interventions every round if $k > m(\lambda_A^* \alpha - c)$.*

Proof. The k random interventions among I_t nodes increase c by at least k/m and α is unchanged. Thus, the disease will die out if $\frac{\alpha}{c+k/m} < 1/\lambda_A^*$. □

Besides providing a threshold for k for which a naive intervention can achieve disease eradication, the above result can be understood as the price of limited information. Lower values of m, meaning more information (better estimate of the true health state), requires fewer random interventions to eradicate the disease. This underscores how uncertainty in the health states is an additional challenge when the number of interventions are limited.

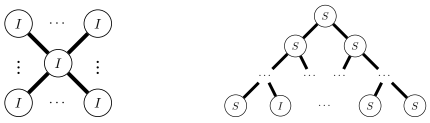
4.3 Eigenvalue and Max Belief

Given the importance of information revealed above, a simple alternative to the eigenvalue approach could be to select k nodes with the top belief of being infected $b_{i,I}^t$ at every time step (denoted further as *Max Belief*). Unfortunately, both the eigenvalue method and Max Belief method have shortcomings in our dynamic problem. We demonstrate this through some observations for different classes of networks. In all the observations, $(\alpha, \beta, c) = (1, 1, 0)$. Also, for the sake of comparison, we assume all beliefs are close to the true states.

Observation 1. *There exists a class of graphs where the Max Belief method with a budget of $k \sim O(1)$ requires an expected $O(n!)$*

rounds to completely eradicate the disease whereas an eigenvalue-based method can eradicate the disease in an expected $O(n^2)$ rounds just with a budget of $k = 2$.

Justification. Consider a star graph (Figure 1a), where all the nodes are initially in I state. With a budget of 2, the eigenvalue method will choose the star center and one arbitrary node among non-central nodes to treat in every round. The disease will thus die out in an expected $\sum_{i=1}^{n-1} \frac{n-1}{i} \sim O(n^2)$ rounds. On the other hand, the Max Belief method will choose k nodes randomly among the nodes in state I. If the center node is not picked in every two rounds ($S \xrightarrow{\text{1 round}} E \xrightarrow{\text{1 round}} I$) before the disease dies out, the center will become infected, and after two more rounds the non-central nodes will be I except $2k$ nodes which can be either in S, E or I state (we ignore this w.l.o.g.). The probability of the center node being chosen every second round (because it takes two rounds to move from S to I state) is $\frac{k}{|I|}$ where $|I|$ is the total number of infected nodes in the round with the center being in I state. The probability of the center node being chosen every second round until the disease dies out is $\prod_{i=0}^{\frac{n}{2k-1}-1} \frac{k}{n-(2k-1)i}$. This gives the desired result.



(a) (b)

Figure 1: Comparing Eigenvalue and Max Belief

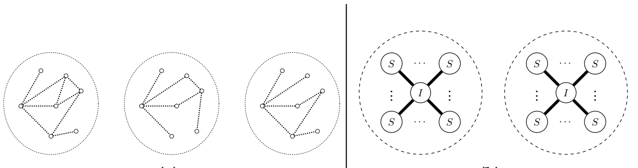
Observation 2. *There exists a class of graphs where an eigenvalue-based method can never eradicate the disease with a budget of $k < \frac{n}{2}$ whereas the Max Belief method can eradicate the disease in one round with a budget of $k \sim O(1)$.*

Justification. Consider a binary tree (Figure 1b), with $\Theta(k)$ leaf nodes in I state and others in S state. An eigenvalue-based method chooses the nodes that equally partitions the graph, and thus in this case it will start choosing from the root and go down the tree in breadth-first order, and reach the leaf nodes only after it has chosen all the $\frac{n-1}{2}$ parent nodes. Max Belief however can eradicate the disease in the first round by simply choosing k nodes which have the highest probability of being in I state, which are the infected leafs.

4.4 Community Based Approach

Infectious diseases such as TB are transmitted via close contact with an infected person, usually within communities [10]. Curing whole communities may potentially be an efficient way to reduce infection (can be interpreted as *graph shattering* [27]), since infection propagation is stopped for large sections of the graph. Also in our case, given the lack of additional information about the network like patient attributes, it is natural to utilize this approach. We also note that forming communities might enable us to reduce the largest eigenvalue, i.e. apply Algorithm **??**, in a scalable fashion.

However, we show in the following Observations that using communities alone can be both better or worse than Greedy or eigenvalue based approaches for different classes of graphs, further motivating the need for our algorithm, TRACE, which identifies communities in addition to considering beliefs and reducing the largest eigenvalue. The exact method of achieving scalability using communities is elucidated in the next section.



(a) (b)

Figure 2: Comparing Eigenvalue, Community and Greedy

Observation 3. *There exists a class of graphs where an eigenvalue-based method can never eradicate the disease with a budget of k and the Greedy method requires an expected $O(|V|^k)$ rounds to completely eradicate the disease, but a community-based method can eradicate the disease in an expected $O(|V|^2)$ rounds.*

Justification. Let us consider a graph where there exists M disjoint clusters (Figure 2a), each of size less than or equal to the budget k with $k << M$, where M is the number of communities. All the nodes are in I state and are arranged in each cluster such that the the top k nodes, removal of which causes the most decrease in the largest eigenvalue, all lie in different clusters. In such graphs, it is evident that community-based algorithms can cure one community at a time and can achieve full eradication after an expected $M^2 \sim (|V|/k)^2 \sim O(|V|^2)$ rounds because a cured community cannot infect other communities. However, an eigenvalue-based technique may not choose communities as a whole and therefore, an eradication cannot be guaranteed unless the budget is increased to $|V|$ which is equal to the size of the graph. Similarly, the Greedy method may not choose communities as a whole and therefore takes an expected $\binom{|V|-1}{k-1}$ rounds to cure the first community, $\binom{|V|-k-1}{k-1}$ rounds to cure the second community, and so on, thus taking approximately $O(|V|^k)$ rounds to cure all the infected nodes.

Observation 4. *There exists a class of graphs where a community-based method can never eradicate the disease whereas the Greedy or eigenvalue-based method either can eradicate the disease in one round with a budget of k.*

Justification. Consider M disconnected star graphs (Figure 2b), where $M - 1$ stars are of size less than k and one star is of size k, and $k \leq M$. All the center nodes of the stars are in I state, and all the other nodes are in S state. With a budget of k, community-based algorithms will keep choosing the same star with k nodes thus never eradicating the disease. However, either the Greedy or eigenvalue-based method can directly choose the k center nodes in the first round and completely eradicate the disease in one shot.

5 TRACE ALGORITHM FOR ACTIVE SCREENING

We introduce a structured algorithm to generate an online POMDP policy—Targeted Resolution of Active diseases using Communities and Eigenvalues (TRACE)—that combines elements of the three approaches (Max Belief, and eigenvalue based, and community based methods) to identify the k individuals to actively screen at every time-step. The complete TRACE algorithm is shown in Algorithm 1. There are two distinct parts to this algorithm.

5.1 Community Formation and Intervention

As we do not know the true health state of all nodes in the network, we form communities using beliefs. The two step process is described below and is a part of Algorithm 1.

Node Type Estimation: We assign an attractiveness score to reflect the effectiveness of intervening on the node. If we knew the true health state of every node, then we would intervene only on the infected nodes as only these nodes spread infection. However, in the absence of such precise information, at every time-step the nodes are sorted according to a measure of possible benefit, defined as $R_i^t = \sigma b_{i,E}^t + b_{i,I}^t$ for each node i (line 2), where σ is an arbitrary parameter that controls the relative importance of E nodes relative to I nodes. The nodes with the highest one-third of R^t values are labeled g_1 (group 1), the next one-third to be g_2 (group 2), and the rest to be g_3 (group 3) (line 3).

Super-Node Creation: After labeling all nodes, locally similar nodes (nodes of the same label that share an edge) are clustered into a super-node iteratively. This process generates a set of super-nodes, each of which is labeled as g_1, g_2 or g_3 based on the labeling of its component nodes. There can be multiple super-nodes with the same label in the network. The $size_{\mathbf{u}}$ of a super-node $\mathbf{u}$ is the number of component nodes in the super-node. The weights of edges between nodes in different super-nodes are added to produce new inter-super-node edges. This super-nodes formation uses the known method of *graph coarsening* [11] (line 4). As an example, in Figure 3 we combine the two g_1, two g_2 and three g_3 nodes to form three super nodes with size two (and another with size one). These super-nodes emulate the communities of I, E and S in real-world networks. We refer to the resultant graph of super-nodes as the community graph, where the belief of each node $b_{\mathbf{u},S}^t$ is the average of $b_{v,S}^t$ of all component nodes v in super-node $\mathbf{u}$.

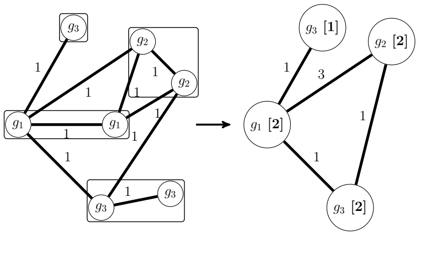


Figure 3: 4 super-nodes formed from 7 nodes

Next, we call the DynamicEigen sub-procedure to choose nodes to screen in the weighted community graph using *size* as weights on each super-node (line 5, where $\overline{\mathbf{A}}$ is the adjacency matrix of the community graph). The procedure returns a set of super-nodes where the total size (weight) is not lower than the budget k. If the total size is higher (line 6), we remove a super-node (line 7), compute left-over budget κ (line 8), modify the original graph by removing all nodes from the left-over super-nodes (line 9), and call the sub-procedure again to select κ nodes from the modified original graph with weights 1 on each node (line 10). It must be noted that our proposed DynamicEigen procedure is also one of the novel aspects of TRACE.

Algorithm 1 TRACE Algorithm

```
Input: Adjacency Matrix $A$ of graph, Belief $b^t$, Budget $k$
1: for all $i \in \{1, \dots, n\}$ do
2:     $R_i^t = \sigma b_{i,E}^t + b_{i,I}^t$
3: Sort $R^t$ and label each node as $g_1, g_2$, or $g_3$
4: $\overline{\mathbf{A}}, \overline{\mathbf{b}}^t, size \leftarrow Coarsen(A, g_1, g_2, g_3, b^t)$
5: $\mathbf{U} \leftarrow$ DynamicEigen($\overline{\mathbf{A}}, \overline{\mathbf{b}}^t, size, k$)
6: if $\sum_{\mathbf{u} \in \mathbf{U}} size_{\mathbf{u}} > k$ then
7:     $\mathbf{u}' \leftarrow$ the last selected super-node from $\mathbf{U}$
8:     $\kappa = k - \sum_{\mathbf{u} \in \mathbf{U} \backslash \mathbf{u}'} size_{\mathbf{u}}$
9:     $\underline{A}, \underline{b}^t \leftarrow$ remove all nodes in $\mathbf{U} \backslash \mathbf{u}'$ from $A, b^t$
10:    $a \leftarrow$ DynamicEigen($\underline{A}, \underline{b}^t, \mathbf{1}, \kappa$)
11: Active screen nodes $\{v \mid v \in a$ or $v \in \mathbf{u}$ for $\mathbf{u} \in \mathbf{U} \backslash \mathbf{u}'\}$
```

5.2 DynamicEigen Procedure

Next, we describe the DynamicEigen procedure, which is shown in Algorithm 2. Prior methods to minimize the largest eigenvalue greedily chose nodes to delete in order to generate a graph with lower maximal eigenvalue. Since we do not know which nodes are infected and can transmit infection with certainty, we augment this method by incorporating uncertainty. To motivate our approach, consider a *hypothetical scenario* where the state of each node is known for sure. We only wish to intervene on infected and exposed nodes, and S nodes do not effect neighboring nodes.

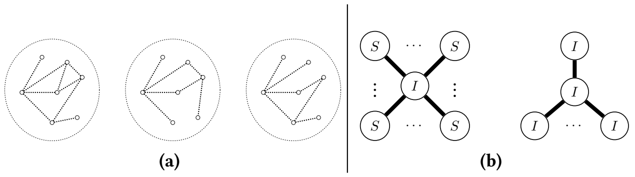
Using $A_{i,j} = A_{j,i} = 1$ to represent an edge from i to j in the adjacency matrix A of the input graph, we see that removing all edges from S nodes is same as multiplying the rows and columns of A corresponding to nodes in state S by zero. Then we can greedily choose among I and E nodes with the goal of reducing the largest eigenvalue of the adjacency matrix of the directed graph and return nodes that have total weights above the threshold k. While our intervention may be undone over time (treated nodes can be reinfected), repeated screenings may push the system towards lower disease prevalence.

Let us return to our problem setup, where we do not know the exact state of each node but rather have beliefs about each node. A natural extension of the hypothetical scenario above is to multiply the row of a node i in the adjacency matrix A by $1 - b_{i,S}^t$, the belief probability it is E or I (line 3). Algorithm 2 describes this approach. This is a softer version of making the row of all S nodes all zeros. Then, we perform greedy selection of nodes (lines 4-9) to reduce the largest eigenvalue of this matrix and to return nodes that have total weights above the threshold k.

Algorithm 2 DynamicEigen(A, b^t, w, k)

```
Input: Adjacency matrix $A$, belief $b$, function $w$ for weight of each node, min total weight of nodes to remove $k$
$V \leftarrow$ Number of vertex of input graph
for all $i \in \{1, \ldots, V\}$ do
    $A_{i,:} = A_{i,:} * (1 - b_{i,S})$                ▷ Multiply $i^{th}$ row
for all $i \in \{1, \ldots, V\}$ do
    $A' \leftarrow A$
    $\mathbf{A}'_{i,:} \leftarrow \mathbf{0}$ , $\mathbf{A}'_{:,i} \leftarrow \mathbf{0}$                ▷ Remove $i^{th}$ node
    $\lambda^i = LargestEigenvalue(A')$
Sort nodes $\langle v_1, \ldots, v_V \rangle$ corresponding to increasing $\lambda^i$
return first $h$ nodes such that $\sum_{i=1}^{h} w(v_i) \geq k$
```

Now that we have combined community structure with belief states (denoted *Comm* in Section 6), we compare it to the DynamicEigen procedure (without super-node formation).



(a) (b)

Figure 4: Comparing DynamicEigen and *Comm* approaches

Observation 5. *There exists a class of graphs where DynamicEigen without super-nodes can never completely eradicate the disease with a budget of* k *whereas the Comm algorithm can eradicate the disease in an expected* $O(n)$ *rounds.*

Justification. Consider a graph with M disjoint clusters (Figure 4a), each of size less than or equal to the budget k and $M > k$. All the nodes in all M communities are in I state. In such graphs, the *Comm* algorithm can treat one community at a time and achieve full eradication after $M \sim n/k \sim O(n)$ rounds as a community of S nodes cannot infect other communities. However, the DynamicEigen algorithm may not choose communities as a whole, therefore eradication cannot be guaranteed unless the budget is increased to n, which is equal to the size of the graph.

Observation 6. *There exists a class of graphs where the Comm algorithm with a budget of* k *requires an expected* $O((n - n')!)$ *rounds, to completely eradicate the disease whereas DynamicEigen without super-nodes can eradicate the disease in an expected* $O(n')$ *rounds with a budget of* k*, where* n' *is the size of the smaller star.*

Justification. Consider a graph with two stars of different sizes (Figure 4b) where the smaller star is of size $n' \geq k$ and the larger star has a size of $n - n'$. Initially, the center node in the larger star is in state I and the other nodes are in state S. All the nodes in the smaller star are in state I. The dynamic eigenvalue algorithm can eradicate the disease with just a budget of k in an expected $O(n')$ rounds by choosing both the stars' center and then choosing one non-central node and the center, or two non-central nodes in each round based on if the center node is in I state. However, the *Comm* algorithm will cluster the smaller star and cure all of them before choosing the I node in the larger star, where by then all of the nodes in the larger star would have been infected. Based on an analysis similar to Observation 1, we can conclude that the disease will die out in an expected $O((n - n')!)$ rounds.

Observation 7. *Suppose the belief states equal the actual health states and* $(\alpha, \beta, c) = (1, 1, 0)$*. Then, TRACE is guaranteed to perform better than or at least as well as its individual components, in terms of both budget and time, in all the classes of graphs discussed in the Observations.*

Proof. For example, in Figure 1a, in case of exact beliefs, it is guaranteed that TRACE will choose the central node since that is the best choice by eigenvalue (all I nodes have equal belief of [0,0,1]) and thereby eradicate the disease in $O(|V|^2)$ rounds with a budget of $k = 2$. Similarly, in Figure 1b, TRACE is guaranteed to choose all the k infected nodes since all the other nodes have zero belief of being in I state, thus eradicating the disease in one round. Thus, following Algorithm 1, we can similarly show that TRACE will in fact perform at least as well as its individual components in all the discussed classes of graphs (variants of trees, stars, and clusters). We omit the details for brevity. □

Thus, TRACE is able to leverage the advantages of each approach. Although these special graphs do not by themselves represent real-world human contact graphs, real graphs are formed from a combination of these special graphs. Estimating that the belief space representation is a reasonably accurate embedding of the information we do have (there is no misinformation in observations while screening), we hypothesize that TRACE's superior performance in these skeleton graphs can be extended to interpret good performance in realistic graphs as well. This hypothesis is validated via experiments.

6 EXPERIMENTS

We consider three real-world datasets on which we perform experiments.

1. **India** network [6]: A human contact network with $n = 202$ nodes, collected from a rural village in India, a setting in which TB active screening may take place ($1/\lambda^*_A \sim 0.095$).
2. Infectious **Exhibition** network [14]: A real-world human contact network with $n = 410$ nodes, collected during an artificial simulation of contagion and containment at an exhibition ($1/\lambda^*_A \sim 0.043$).
3. **Irvine** network [15]: An online social network with $n = 1899$ nodes, constructed from sent messages between the users of an online community of students from UC Irvine ($1/\lambda^*_A \sim 0.021$).

As discussed in Section 3, we first attempt to solve our special POMDP using the state-of-the-art modified POMCP algorithm [19]. We show in Figure 5 that POMCP takes exponential time with increasing n and fails to scale up beyond 10 nodes (India network) for fixed values of k and T while TRACE is able to generate an online POMDP policy for the whole network without exponential increase in runtime. Factored POMDPs [29] and newer algorithms like DESPOT [22] also fail to scale up beyond a few nodes due to memory overflow. All results are averages over 20 simulation runs.

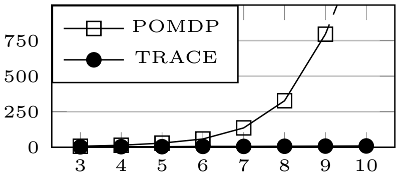


Figure 5: Runtime (s) v/s Number of nodes (n); $k = 3, T = 10$

Settings. Next, we analyze TRACE's performance under various α, β, c settings. α, β, c may depend on social contact patterns and biological factors which may vary across populations [23]. We explore a range of these parameters to show disease behavior under a variety of scenarios. Since eradication does not depend on β (by Proposition 1), we vary only α, c and fix $\beta = 0.25$ for the experiments. The passive treatment rate c may vary widely, as it depends on resource availability (clinic accessibility, outreach campaigns, etc.). In all simulations, the budget is $k = 5\%$ of the total population, and $\sigma = 0.5$.

Setup. In the real world, active screening is performed only after conducting initial surveys on the prevalence and incidence of the disease. To simulate this, we run our experiments in two stages.

(1) Stage 1 (**Survey Stage**), starts at $t = 0$ with equal number of S, E, I individuals and ends at $t = 10$. No active screening is done and the disease evolves naturally. The initial belief b^0 for all nodes is assumed to be $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ since we have no prior information. Beliefs are updated when individuals come to the clinic voluntarily (with probability c).

(2) Stage 2 (**Active Screening Stage**), we consider various screening algorithms. We perform active screening from $t = 11$ to $t = 30$ to represent 10 years of time (each round is 6 months [7]). We compare the benefit of these screening strategies over and above no intervention (**None**), where in **None** the evolution of the health states is based on disease dynamics with no active screening.

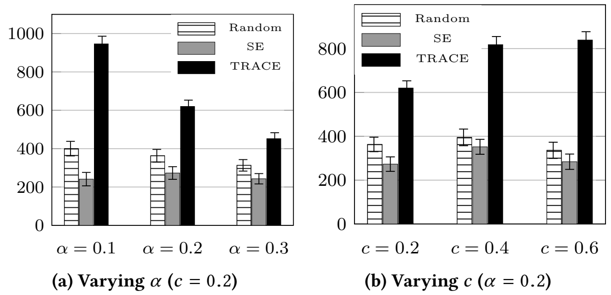
Comparison with baselines. Given the lack of previous algorithms, Figures 6 and 7 show the performance of TRACE against simple baselines:

(1a) **Random**: Randomly select nodes for active screening.

(1b) **Static Eigen (SE)**: Choose the nodes using Algorithm 2 after removing lines 2 & 3 (no belief information), on the network (no super-node formation). This baseline uses only the graph structure information.

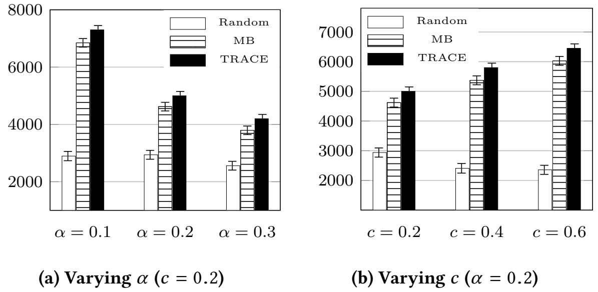
TRACE provides significant improvement over None compared to SE and Random ($p < 0.05$). The improvement is also practically significant (Cohen's $d > 1$: large effect).

Comparison with individual components. Figure 8 shows the performance of the three approaches that were combined to form TRACE, illustrating that no single approach is solely responsible for TRACE's performance. We compare the increase in $\sum_{t=0}^{t=30} |S|_t$ for each approach over None. TRACE's performance is both statistically and practically significant ($p < 0.05$ and Cohen's $d \sim 0.6$: medium effect) when compared to the three approaches:



(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

Figure 6: Increase in $\sum_{t=0}^{t=30} |S|_t$ for naive baselines and TRACE over None (India network)



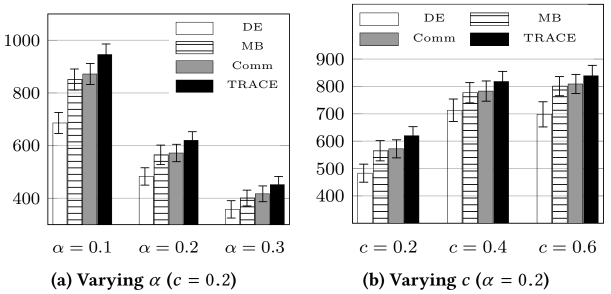
(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

Figure 7: Increase in $\sum_{t=0}^{t=30} |S|_t$ for naive baselines and TRACE over None (Irvine network)

(2a) **Dynamic Eigen (DE)**: Choose the nodes using just Algorithm 2 without any super-node formation.

(2b) **Max Belief (MB)**: Choose the nodes with the higher *belief* of being infected in that time-step, i.e. $b_{i,I}^t$.

(2c) **Community (Comm)**: Choose the nodes by a 0-1 knapsack algorithm (knapsack weight = budget k) after super-node formation.



(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

Figure 8: Performance by TRACE component (India network)

Further, we analyze the minimum additional budget required to achieve performance comparable to TRACE in Figure 9, revealing the budgetary savings from using TRACE. TRACE with all its components produces significant savings over attempting to use each component alone.

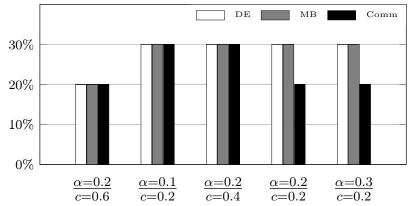
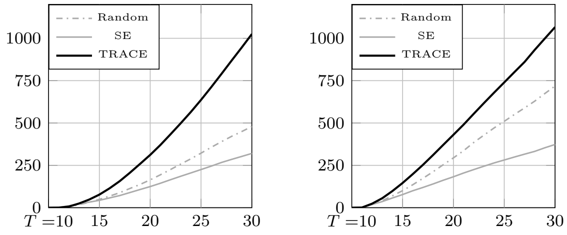


Figure 9: Minimum extra budget (in %) required to match performance of TRACE (India network)

The synergy of belief states, eigenvalues and community gives TRACE a clear advantage on both the datasets (Figure 10), where we see an increasing divergence over time in the performance of TRACE compared to **Random** and **SE**.



(a) India network **(b) Exhibition network**

Figure 10: Increase in $\sum_{t=0}^{t=T} |S|_t$ over None for varying T ($\alpha = 0.1, \beta = 0.25, c = 0.2$)

7 CONCLUSION

We proposed a novel active screening model and an algorithm (TRACE) to facilitate multi-round active screening for recurrent diseases. Unlike existing works in AI literature, the Active Screening model incorporates uncertainty of health states as well as the SEIS disease complexities of no permanent cure and a latent stage. TRACE performs significantly better, in a scalable fashion, than the baselines and each of its components individually in a variety of real-world inspired settings.

Future directions include incorporating more complex disease models (e.g. including maternal immunity, carrier states etc.), including birth and death processes, and introducing patient heterogeneity (age, gender, medical history and other features) and costs of treatment and screening into the model.

REFERENCES

[1] Benjamin Armbruster and Margaret L Brandeau. 2007. Optimal mix of screening and contact tracing for endemic diseases. *Mathematical biosciences* 209, 2 (2007), 386–402.

[2] James Aspnes, Kevin Chang, and Aleksandr Yampolskiy. 2006. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. *J. Comput. System Sci.* 72, 6 (2006), 1077–1093.

[3] National Tuberculosis Controllers Association et al. 2005. Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports* 54, RR-15 (2005), 1.

[4] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. 2012. OR ForumâĂŤA POMDP approach to personalize mammography screening decisions. *Operations Research* 60, 5 (2012), 1019–1034.

[5] Frank G Ball, Edward S Knock, and Philip D OâĂŹNeill. 2015. Stochastic epidemic models featuring contact tracing with delays. *Mathematical biosciences* 266 (2015), 23–35.

[6] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. The diffusion of microfinance. *Science* 341, 6144 (2013), 1236498.

[7] CDC. 2011. Tuberculosis: General Information. (2011). https://www.cdc.gov/tb/publications/factsheets/general/tb.pdf

[8] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)* 10, 4 (2008), 1.

[9] Palanivel Chinnakali, Pruthu Thekkur, Gomathi Ramaswamy, Kalaiselvi Selvaraj, et al. 2016. Active screening for tuberculosis among slum dwellers in selected urban slums of Puducherry, South India. *Annals of Tropical Medicine and Public Health* 9, 4 (2016), 295.

[10] Collette N Classen, Robin Warren, Madeleine Richardson, John H Hauman, Robert P Gie, James HP Ellis, Paul D van Helden, and Nulda Beyers. 1999. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. *Thorax* 54, 2 (1999), 136–140.

[11] Bruce Hendrickson and Robert W Leland. 1995. A Multi-Level Algorithm For Partitioning Graphs. *SC* 95, 28 (1995).

[12] IHME. 2010. *The Global Burden of Disease: Generating Evidence, Guiding, Policy.* World Bank.

[13] Jing Kjeldsen. 2013. *The probability for infection in SIR and SIS networks.* Master's thesis.

[14] KONECT. 2017. Infectious network dataset – KONECT. http://konect.uni-koblenz.de/networks/sociopatterns-infectious

[15] KONECT. 2017. Uc irvine messages network dataset – KONECT. http://konect.uni-koblenz.de/networks/opsahl-ucsocial

[16] K Kranzer, H Afnan-Holmes, K Tomlin, Jonathan E Golub, AE Shapiro, A Schaap, EL Corbett, K Lönnroth, and JR Glynn. 2013. The benefits to communities and individuals of screening for active tuberculosis disease: a systematic review [State of the art series. Case finding/screening. Number 2 in the series]. *The international journal of tuberculosis and lung disease* 17, 4 (2013), 432–446.

[17] E Mitchell, Saskia den Boon, and K Lonnroth. 2013. Acceptability of household and community-based TB screening in high burden communities: a systematic literature review. WHO.

[18] B Aditya Prakash, Deepayan Chakrabarti, Nicholas C Valler, Michalis Faloutsos, and Christos Faloutsos. 2012. Threshold conditions for arbitrary cascade models on arbitrary networks. *Knowledge and information systems* 33, 3 (2012), 549–575.

[19] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems.* International Foundation for Autonomous Agents and Multiagent Systems, 123–131.

[20] Sudip Saha, Abhijin Adiga, B Aditya Prakash, and Anil Kumar S Vullikanti. 2015. Approximation algorithms for reducing the spectral radius to control epidemic spread. In *Proceedings of the 2015 SIAM International Conference on Data Mining.* SIAM, 568–576.

[21] Michael Shapiro and Edgar Delgado-Eckert. 2012. Finding the probability of infection in an SIR network is NP-Hard. *Mathematical biosciences* 240, 2 (2012), 77–84.

[22] Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. 2013. DESPOT: Online POMDP planning with regularization. In *Advances in neural information processing systems.* 1772–1780.

[23] Sze-chuan Suen, Eran Bendavid, and Jeremy D Goldhaber-Fiebert. 2014. Disease control implications of India's changing multi-drug resistant tuberculosis epidemic. *PloS one* 9, 3 (2014), e89822.

[24] Chengjun Sun and Ying-Hen Hsieh. 2010. Global analysis of an SEIR model with varying population size and vaccination. *Applied Mathematical Modelling* 34, 10 (2010), 2685–2697.

[25] Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, 245–254.

[26] P Van den Driessche, M Li, and J Muldowney. 1999. Global stability of SEIRS models in epidemiology. *Canadian Applied Mathematics Quarterly* 7 (1999), 409–425.

[27] Nan Wang. 2005. *Modeling and analysis of massive social networks.* Ph.D. Dissertation.

[28] WHO. 2017. Global tuberculosis report 2017. (2017).

[29] Jason D Williams, Pascal Poupart, and Steve Young. 2005. Factored partially observable Markov decision processes for dialogue management. In *Proc. IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems.* 76–82.

[30] Yao Zhang and B Aditya Prakash. 2015. Data-aware vaccine allocation over large networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 2 (2015), 20.