A Markov Decision Process Framework for Efficient and Implementable Contact Tracing and Isolation

George Z. Li[2,*] Arash Haddadan[1] Ann Li[1] Madhav Marathe[1,3]
Aravind Srinivasan[2] Anil Vulikanti[1,3] Zeyu Zhao[2]

Abstract

Efficient contact tracing and isolation is an effective strategy to control epidemics. It was used effectively during the Ebola epidemic and successfully implemented in several parts of the world during the ongoing COVID-19 pandemic. An important consideration in contact tracing is the budget on the number of individuals asked to quarantine—the budget is limited for socioeconomic reasons. In this paper, we present a Markov Decision Process (MDP) framework to formulate the problem of using contact tracing to reduce the size of an outbreak while asking a limited number of people to quarantine. We formulate each step of the MDP as a combinatorial problem, MinExposed, which we demonstrate is NP-Hard; as a result, we develop an LP-based approximation algorithm. Though this algorithm directly solves MinExposed, it is often impractical in the real world due to information constraints. To this end, we develop a greedy approach based on insights from the analysis of the previous algorithm, which we show is more interpretable. A key feature of the greedy algorithm is that it does not need complete information of the underlying social contact network. This makes the heuristic implementable in practice and is an important consideration. Finally, we carry out experiments on simulations of the MDP run on real-world networks, and show how the algorithms can help in *bending* the epidemic curve while limiting the number of isolated individuals. Our experimental results demonstrate that the greedy algorithm and its variants are especially effective, robust, and practical in a variety of realistic scenarios, such as when the contact graph and specific transmission probabilities are not known. All code can be found in our GitHub repository: https://github.com/gzli929/ContactTracing.

1 Introduction

Contact tracing followed by isolation is one of the most effective ways to control epidemics caused by infectious diseases. In this intervention strategy, contact tracers ask infected individuals to report their recent contacts; they then trace these contacts, requesting them to isolate for a certain period of time [Kretzschmar *et al.*, 2020]. The role of contact tracing during the Ebola, measles, and COVID-19 outbreaks is well-documented [Keeling *et al.*, 2020; Kretzschmar *et al.*, 2020; Liu *et al.*, 2015]. However, its effectiveness is dependent on the accuracy and quantity of information

[1]Biocomplexity Institute and Initiative, University of Virginia
[2]Department of Computer Science, University of Maryland
[3]Department of Computer Science, University of Virginia
*Correspondence to George Li at gzli929@gmail.com

on the contacts, the speed at which tracing is conducted, and the compliance of individuals in self-isolating. Recently, technologies such as the Google-Apple app [Ahmed *et al.*, 2020] have provided a solution to augment human contact tracers. When contact tracing apps are used, the strategy is called *digital* contact tracing; otherwise, it is called *manual* contact tracing. Our algorithms and heuristics will be applicable to both manual contact tracing and digital contact tracing.

A main limitation of contact tracing is the number of individuals who can be asked to isolate; this number is constrained since isolation imposes a significant economic and social burden to the population. For manual contact tracing, the budget is also dependent on the economic cost of hiring contact tracers. From these constraints, we can see a clear trade-off between reducing infection spread and minimizing socioeconomic costs. This brings forth a natural question that we study: which individuals should we ask to isolate in order to make the most effective use of the budget for contact tracing?

In addition to constraints on the number of individuals who are isolated, we also address the practical challenges of contact tracing. Most notably, contact tracing graphs and their associated transmission probabilities are noisy, sparse, and dynamic [Liu *et al.*, 2020; Sayampanathan *et al.*, 2021]. Motivated by this, we seek robust algorithms that can deal with such uncertainties. Additionally, we need to consider the simplicity and utility of such algorithms to encourage widespread use. These factors motivate us to develop simple but effective heuristics for our problem [Russell and Norvig, 2002; Yadav *et al.*, 2016].

Our contributions. We use a Markov Decision Process (MDP) framework to formulate the problem of efficient contact tracing that reduces the size of the outbreak using a limited number of contact tracers (see Section 3). The basic setup is as follows: let $G = (V, E)$ be the social contact network and let the disease spread on G by an SIR type diffusion process [Marathe and Vullikanti, 2013]. At each timestep t, we assume the policymaker knows the infected set. Constrained by the number of contact tracers, the policymaker wants to choose a set of nodes that minimizes the total number of infections at the end of the epidemic when asked to quarantine. We call this problem MinTotalInf. Since the disease dynamics are constantly changing (due to fluctuating attitudes and behavior), we will only consider finite time horizons of the MDP by solving the problem, MinExposed, which focuses on the second neighborhood of the infected set.

- We prove that MinExposed is NP-Hard (see Section 4). Given the hardness result, we develop an LP-based algorithm for MinExposed, proving rigorous approximation guarantees. Using insights from the analysis of the LP-based algorithm, we introduce a greedy approximation algorithm, which is interpretable and practical (see Section 5).
- While maintaining the theoretical properties of our algorithms, we show that we can incorporate *fairness guarantees*, ensuring no demographic group is disproportionately affected by contact tracing or the disease. Furthermore, we experimentally verify that incorporating these fairness constraints is possible while *not degrading solution quality much* (Sections 6 and 8.3).
- Our provable approximation algorithms require knowledge of the (local) contact graph, transmission rates, and compliance rates, which is unrealistic in practice. We draw on the intuition gained from our theoretical results to devise heuristics which requires minimal information of the contact graph or disease model—and includes differential privacy for user privacy—and thus can be made operational in the real world (see Section 7).
- We run simulations of an epidemic with realistic contact network and parameter values to assess the performance of our algorithms and heuristics. The results suggest that the heuristics perform well even under the limited information model (see Section 8).

2 Related Work

Manual contact tracing is a widely used strategy that has been successful in controlling past outbreaks; see Armbruster and Brandeau [2007a,b]; Eames [2007]; Kiss *et al.* [2005, 2008] for a discussion on contact tracing, its effectiveness, and mathematical models to study contact tracing in networks. Recently, digital contact tracing has emerged as another powerful technology to control outbreaks, especially evident in the COVID-19 pandemic [Ahmed *et al.*, 2020; Salathé *et al.*, 2020; Lorch *et al.*, 2020]. Though the importance of contact tracing is well studied, we are the first to view it as an algorithmic problem and give provable guarantees to our methods. Moreover, we are the first to address fairness concerns in quarantine decisions.

Our paper adds to a line of work developing theoretical models for intervention problems in networked epidemic processes; however, prior works have only considered these problems in idealized settings. For instance, [Eubank *et al.*, 2006; Hayrapetyan *et al.*, 2005; Sambaturu *et al.*, 2020; Minutoli *et al.*, 2020] consider problems of optimizing interventions such as vaccination and social distancing in a non-adaptive and complete information setting, where the intervention is only done at the start of the epidemic. In contrast, our paper focuses on the more realistic contact tracing problem, in which decisions need to be made at each timestep. Moreover, we only assume knowledge of a local neighborhood of the currently infected nodes, which is more realistically available.

Concurrent to our work, Meister and Kleinberg [2021] introduce a model of manual contact tracing and design provably optimal algorithms. They focus on developing algorithms to mitigate the disease's health detriments after the outbreak stops while we focus on quarantining to minimize the disease spread during the outbreak. Though they have a more realistic model of discovering contacts, we claim our contact tracing model is more realistic since it operates in real time. Furthermore, their model applies to manual contact tracing and remains primarily a theoretical contribution while our model applies to both and additionally yields improved practical heuristics. Finally, we note that Meister and Kleinberg [2021] mention the dynamic setting of contact tracing as important future work; our paper takes a first step in addressing this complex problem.

3 Preliminaries

Recall that the epidemic spreads on G by an SIR-type process; let $I(t)$ be the set of infected nodes and let each node $u \in I(t)$ transmit the disease to each of their neighbors v independently with probability q_{uv}. Denote the current set of infected people $I = I(t)$. We assume I is known to the policymaker and will begin to self-isolate at the next timestep, remaining quarantined until recovery. Although all previously infected nodes self-isolate, neighbors of I have been exposed to the disease and may be infected in the next timestep. Let $V_1 = N_G(I) - I$ be the first neighborhood of I. Because V_1 can continue to spread the disease to the rest of the graph, policymakers must contact trace these individuals and ask them to isolate. Since this process is expensive and time-intensive for both contact tracers and quarantined individuals, we denote B to be the budget on the number of nodes which contact tracers can reach. Further complicating the costs of contact tracing, some of the individuals contacted may be noncompliant, refusing to quarantine. For model generality, we assume node u complies with probability c_u. Given these parameters and constraints, the objective of policymakers can be formulated as MinTotalInf, which seeks to minimize the total number of infections in G at the end of the epidemic. MinTotalInf is a highly idealized problem to solve since the contact graph, transmission rates, and compliance rates are all constantly changing due

to various forms of social distancing. As a result, we focus on locally optimal solutions with the objective of minimizing the expected number of infections in the second neighborhood of I. We denote this neighborhood as $V_2 = N_G(V_1) - I - V_1$ and formalize the problem next.

The MinExposed Problem: Given contact graph $G = (V, E)$, a subset $I \subseteq V$ of infected nodes, compliance probabilities c_u for $u \in V_1$, transmission probabilities q_{uv} for $(u, v) \in E$, and a budget B, the objective is to find a subset $Q \subseteq V_1$ satisfying $|Q| \leq B$ to quarantine which minimizes the expected number of infections in V_2. We let $F(Q)$ denote the objective value of MinExposed given that set Q is asked to quarantine. See Figure 1 for an example.

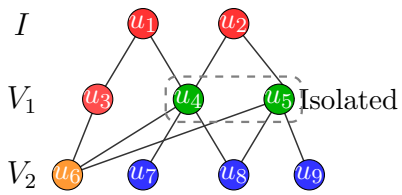

Figure 1: Example of MinExposed when transmission and compliance rates are 1: set $I = \{u_1, u_2\}$, $V_1 = \{u_3, u_4, u_5\}$, $V_2 = \{u_6, u_7, u_8, u_9\}$. Suppose $B = 2$; then set $Q = \{u_4, u_5\}$ is an optimal isolated set. Node u_6 is exposed, and the objective value for this solution is 1.

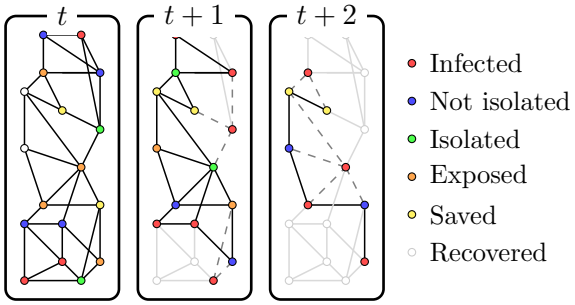

Figure 2: Our MDP model: In each timestep, the infected nodes are shown in red. The set of nodes in V_1 is colored by blue and green, where green represents that the solution to MinExposed suggests them to isolate. The nodes in orange and yellow are in V_2, where yellow nodes are not exposed since the green nodes self isolate. Since this figure considers the simple deterministic case, all nodes in the neighborhood of (non-isolated) infected nodes get infected in the next time step. Recovered nodes are depicted in grey.

In addition to minimizing infections, the policymaker also wants to ensure no demographic group is affected disproportionately by our contact tracing algorithms. In particular, the number of people quarantined in V_1 and infected in V_2 should be fair with respect to demographic groups. To account for this, we abstract away the attributes of a node $v \in V$ by a label $\ell(v) \in L$. We assume $\bigcup_{\ell \in L} \mathcal{R}_\ell = V$, where $\mathcal{R}_\ell \subseteq V$ denote the set of nodes with label $\ell \in L$. We also assume we are given constraints B_ℓ for $\ell \in L$ on the number of people in $V_1 \cap \mathcal{R}_\ell$ to quarantine and constraints a_ℓ for $\ell \in L$ on the expected number of infections in $V_2 \cap \mathcal{R}_\ell$. (Note that this is for full generality. One useful example to think about is where the budgets for each demographic group is proportional to their size, i.e., B_ℓ is proportional to $|V_1 \cap \mathcal{R}_\ell|$ and a_ℓ is proportional to $|V_2 \cap \mathcal{R}_\ell|$.) We will show how to extend our algorithms to satisfy these constraints while maintaining their utility.

4 MinExposed is NP-Hard

Theorem 1. *Even when all transmission and compliance probabilities are 1 and there are no fairness constraints,* MinExposed *is NP-Hard.*

Proof. Consider the Maximum Clique Problem: given a graph $G = (V, E)$ find a subset S of V such for $u, v \in S$ we have $(u, v) \in E$. Maximum clique problem is a well-known NP-hard problem

[Garey and Johnson, 1979]. It is also well-known that the Maximum clique problem can be reduced in polynomial time to the problem of deciding whether G contains a clique of size k. We reduce this problem to an instance of MinExposed where the transmission probability along all edges is 1 and compliance probabilities are all 1.

Define I to be a single node: $I = \{i\}$. For each node in G, we add a node in our MinExposed instance and connect it to i (So V is $N(I) - I$ in our MinExposed instance). Now for each edge $(u, v) \in E$ we add a node and connect it to u and v in $N(I) - I$. Now, consider the MinExposed on this instance with budget k. Let $V_1 = N(I) - I$ and $V_2 = N(V_1) - V_1 - I$. Let $Q^* \subseteq V_1$ be the optimal solution to this instance of MinExposed. We claim that G has a clique of size k if and only if $F(Q^*) = |E| - \binom{k}{2}$.

First, we show $F(Q^*) = |E| - \binom{k}{2}$ whenever G has a clique of size k. Let $U \subseteq V$ be a clique of size k in G. Then U corresponds to some set $S \in V_1$ (in our MinExposed instance) of size k. We have $F(S) = |E| - \binom{k}{2}$: let u, v be distinct nodes in U. For $(u, v) \in E$, edge (u, v) is exposed in solution S if either $u \notin S$ or $v \notin S$. This implies $F(S) = |E| - \binom{k}{2}$. By optimality of Q^*, we have $F(Q^*) \leq |E| - \binom{k}{2}$. Clearly, we also have $F(Q^*) \geq |E| - \binom{k}{2}$.

Conversely, suppose $F(Q^*) = |E| - \binom{k}{2}$. Let U be the set in V corresponding to Q^*. We claim that for distinct $u, v \in U$, we have $(u, v) \in E$. Suppose for contradiction that there are $u, v \in U$ (hence in Q^*) such that $(u, v) \notin E$. Then, there are less than $\binom{k}{2}$ nodes in V_2 that are covered by Q^*. This implies that $F(Q^*) > |E| - \binom{k}{2}$, which is a contradiction. □

5 Approximation Algorithms for MinExposed

In the previous section, we showed that even when all compliance and transmission probabilities are 1, MinExposed is NP-Hard. As a result, we focus on developing approximation algorithms. For ease of notation in the next sections, we let $p_u = 1 - \prod_{v \in I:(u,v) \in E} (1 - q_{uv})$ be the probability $u \in V_1$ gets infected in the next timestep.

Let D_v denote the number of neighbors in V_1 node $v \in V_2$ has and let $D = \max_{v \in V_2} D_v$. We first present a mixed integer linear program (MILP) to formulate MinExposed and show that applying DepRound gives a D-approximation (i.e., it provides a solution with objective value at most D times optimal). Using insight from the analysis of DepRound, we present a simple greedy algorithm, DegGreedy, which still guarantees a D-approximation. Furthermore, DegGreedy offers better interpretability and is easier to implement under noisy/incomplete information.

Note that we don't make an assumption of independence in our proofs, which is an advantage of our methods. This means our results apply even when transmissions are correlated (e.g., when the transmission events $v \to u$ and $w \to u$ are positively correlated, for neighbors $v, w \in I$ of $u \in V_1$): we just need to update the formula above for p_u to take correlations into account, in constraint (2) and later. Such correlations are common due to meetings in groups/crowds, making our methods especially desirable.

5.1 DepRound

Let $E' = E \cap (V_1 \times V_2)$ be the edges which can potentially transmit the disease in the next timestep. We can write MinExposed as an MILP:

$$\begin{aligned}
&\min \textstyle\sum_{v \in V_2} z_v && \text{s.t.} \\
&x_u + y_u = 1 && \text{for } u \in V_1 \\
&\textstyle\sum_{u \in V_1} x_u \leq B && (1) \\
&z_v \geq p_u \cdot [1 - c_u \cdot x_u] \cdot q_{uv} && \text{for } (u,v) \in E' \quad (2) \\
&x_u, y_u \in \{0,1\} && \text{for } u \in V_1 \\
&z_v \in [0,1] && \text{for } v \in V_2
\end{aligned}$$

We have x_u, y_u for $u \in V_1$ as indicators representing u being asked to quarantined and u potentially spreading the disease, respectively. We allow at most B nodes to be quarantined, as indicated by Constraint 1. Note that for $v \in V_2$, we have the following for every $u \in V_1$ with $(u,v) \in E$: the probability that v gets infected is lower bounded by the probability u is infected, u is not selected for quarantine or u does not comply, and u transmits the disease to v. Thus z_v for $v \in V_2$ represents a lower bound on the probability on v getting infected, as conveyed through Constraint 2.

Algorithm 1 DepRound

1: Relax the integer constraints of the MILP to obtain its LP relaxation
2: Solve the LP to get vectors $x, y \in \mathbb{R}^{V_1}$
3: Apply dependent rounding as in Srinivasan [2001] to vector x to obtain X_u for $u \in V_1$
4: $Q \leftarrow \{u \in V_1 : X_u = 1\}$

Based on this MILP, we give our algorithm for MinExposed. First, we relax the binary vector constraints on x_u, y_u, to get a computationally-feasible linear program (LP). The output of the LP will be vectors $x, y \in \mathbb{R}^{V_1}$ and $z \in \mathbb{R}^{V_2}$ with $x_u, y_u, z_v \in [0,1]$, with objective-function value *at most* as large as our optimal solution. However, x_u may be a fractional value, which does not directly imply a decision for our contact-tracing problem. Srinivasan [2001] presented a linear time randomized algorithm which given a vector $x \in [0,1]^n$ with $\sum_{i=1}^n x_i \leq k$ outputs a vector $X \in \{0,1\}^n$ satisfying:

(P1) For $i = 1, \ldots, n$, $\Pr[X_i = 1] = x_i$;

(P2) $\sum_{i=1}^n X_i \leq k$ with probability one.

(P3) For all $S \subseteq [n]$, we have:

$$\begin{aligned}
\Pr[\textstyle\bigwedge_{i \in S}(X_i = 0)] &\leq \textstyle\prod_{i \in S} \Pr[X_i = 0]; \\
\Pr[\textstyle\bigwedge_{i \in S}(X_i = 1)] &\leq \textstyle\prod_{i \in S} \Pr[X_i = 1].
\end{aligned}$$

We use this to obtain $X \in \{0,1\}^{V_1}$ from vector x, giving our final solution of $Q = \{u \in V_1 : X_u = 1\}$. We call this algorithm DepRound and give its approximation guarantee next:

Theorem 2. *Applying Algorithm 1 to the above MILP yields a D-approximation for* MinExposed.

Proof. Let vectors x, y, z be the optimal solution the linear program relaxation and let $(X_u \in \{0,1\} : u \in V_1)$ be the output of dependent rounding. Recall that $Q = \{u \in V_1 : X_u = 1\}$. By **(P2)**, we have $|Q| \leq B$ with probability one, as desired.

We next analyze what happens to nodes $v \in V_2$. For ease of notation, define C_u to be the random variable that node $u \in V_1$ complies when asked to quarantine and Q_{uv} to be the random variable that node $u \in V_1$ transmits the disease to node $v \in V_2$. We have $\mathbb{E}[C_u] = c_u$ and $\mathbb{E}[Q_{uv}] = p_u \cdot q_{uv}$. The probability v gets infected is equal to the probability there exists a neighbor $u \in V_1$ of v which gets infected, does not get quarantined or gets quarantined and does not comply, and transmits to v:

$$\begin{aligned}\Pr[v \text{ gets infected}] &= \Pr[\exists u \in V_1 : (u,v) \in E \wedge [(C_u = 0) \vee (X_u = 0)] \wedge (Q_{uv} = 1)] \\ &\leq \textstyle\sum_{u:(u,v)\in E'} p_u \cdot [(1 - c_u) \cdot x_u + y_u] \cdot q_{uv} \\ &\leq \textstyle\sum_{u:(u,v)\in E'} z_v \leq D_v \cdot z_v.\end{aligned}$$

The first inequality holds by the union bound, the second inequality holds by Constraint 4, and the rest hold by definition. Using this, we analyze the MinExposed objective value.

$$\begin{aligned}F(Q) &= \textstyle\sum_{v\in V_2} \Pr[v \text{ gets infected}] \\ &\leq \textstyle\sum_{v\in V_2} D_v \cdot z_v \leq D \cdot \sum_{v\in V_2} z_v \leq D \cdot F(Q^*).\end{aligned}$$

Thus, our algorithm yields a D-approximation for MinExposed. □

5.2 DegGreedy

In the analysis of DepRound, we took advantage of the union bound as an upper bound to the MinExposed objective value in order to prove our approximation guarantee. Next, we present a simple greedy algorithm, DegGreedy, which directly optimizes the upper bound and thus still maintains a D-approximation. Recall that for a quarantine set Q, we have

$$\begin{aligned}F(Q) &\leq \textstyle\sum_{v\in V_2} \sum_{u:(u,v)\in E'} [(1 - c_u) \cdot x_u + y_u] \cdot p_u \cdot q_{uv} \\ &= \textstyle\sum_{v\in V_2} \sum_{u:(u,v)\in E'} [1 - c_u \cdot x_u] \cdot p_u \cdot q_{uv}.\end{aligned}$$

Ignoring the constant, we see that *minimizing* the upper bound on $F(Q)$ is equivalent to *maximizing*

$$\begin{aligned}&\textstyle\sum_{v\in V_2} \sum_{u:(u,v)\in E'} x_u \cdot c_u \cdot p_u \cdot q_{uv} \\ &= \textstyle\sum_{u\in V_1} \sum_{v\in V_2:(u,v)\in E} x_u \cdot c_u \cdot p_u \cdot q_{uv} \\ &= \textstyle\sum_{u\in V_1} x_u \cdot c_u \cdot p_u \cdot \sum_{v\in V_2:(u,v)\in E} q_{uv}\end{aligned}$$

subject to $\sum_{u\in V_1} x_u \leq B$. Since this is just a knapsack problem, it is clear that DegGreedy attains the optimal value and thus minimizes the *upper bound* on $F(Q)$.

Algorithm 2 DegGreedy

1: $w_u \leftarrow c_u \cdot p_u \cdot \sum_{v\in V_2,(u,v)\in E} q_{uv}$ for $u \in V_1$
2: pick B nodes with the highest w_u values in V_1 to be in Q, breaking ties arbitrarily

Theorem 3. *Algorithm 2 gives a D-approximation to* MinExposed.

Proof. Let x_u^*, y_u^*, z_v^* to be the optimal solution to the MILP in Section 5.1. Let Q be the set outputted by DegGreedy, $x_u = I\{u \in Q\}$ is the indicator for membership in Q, and $y_u = 1 - x_u$. Then we have

$$\begin{aligned} F(Q) &\leq \textstyle\sum_{v \in V_2} \sum_{u:(u,v) \in E'} \ [1 - c_u \cdot x_u] \cdot p_u \cdot q_{uv} \\ &\leq \textstyle\sum_{v \in V_2} \sum_{u:(u,v) \in E'} \ [1 - c_u \cdot x_u^*] \cdot p_u \cdot q_{uv} \\ &\leq \textstyle\sum_{v \in V_2} \sum_{u:(u,v) \in E'} \ z_v^* \\ &\leq \textstyle\sum_{v \in V_2} D_v \dot{z}_v^* \leq D \cdot \sum_{v \in V_2} z_v^* \leq D \cdot OPT \end{aligned}$$

where the first inequality holds by the union bound, the second holds because DegGreedy optimizes the upper bound, the third holds by Constraint 2, and the remaining hold by definition. □

6 Extension to Fairness Constraints

Recall that we want the following fairness guarantees: for V_1, we want the number of quarantined people with label ℓ to be at most B_ℓ and for V_2, we want the number of expected infected people with label ℓ to be at most a_ℓ (assuming there exists a feasible solution). We can extend both of our algorithms to satisfy the first constraint and we can extend DepRound to satisfy the second constraint approximately.

6.1 Fairness in V_1

Recall that the $\mathcal{R}_\ell$ are demographic groups and assume that $\sum_{\ell \in L} B_\ell = B$. Then we can guarantee that the number of quarantined nodes in $\mathcal{R}_\ell$ is at most some given budget B_ℓ. We can easily enforce this in our MILP formulations in Section 3.1 by adding the following constraint:

$$\textstyle\sum_{u \in \mathcal{R}_\ell} x_u \leq B_\ell \text{ for } \ell \in L. \tag{3}$$

For DepRound, this constraint guarantees fairness for the LP solutions, but the rounded solutions may still violate the constraints. To fix this, we modify step 3 of DepRound to rounding the vectors $[x_u : u \in V_1 \cap \mathcal{R}_\ell]$ representing each demographic group separately. We call this algorithm Fair DepRound and note that by **(P2)**, we have the fairness guarantee with probability 1. We can similarly we modify step 2 in DegGreedy to picking B_ℓ nodes with highest w_u to be in Q, for each $\ell \in L$. We call this algorithm Fair DegGreedy, and we have the fairness guarantee obviously.

Algorithm 3 Fair DepRound

1: Relax the integrality constraints of the MILP to obtain its LP relaxation
2: Solve the LP to get vectors $x, y \in \mathbb{R}^{V_1}$
3: Apply dependent rounding as in Srinivasan [2001] to vector $[x_u : u \in V_1 \cap \mathcal{R}_\ell]$ for $\ell \in L$ to obtain X_u for $u \in V_1$
4: $Q \leftarrow \{u \in V_1 : X_u = 1\}$

Algorithm 4 Fair DegGreedy

1: $w_u \leftarrow c_u \cdot p_u \cdot \sum_{v \in V_2, (u,v) \in E} q_{uv}$ for $u \in V_1$
2: **for** $\ell \in L$: pick B_ℓ nodes with the highest w_u values in $V_1 \cap R_\ell$ to be in Q (break ties arbitrary)

Theorem 4. *Algorithms 3 and 4 give a D-approximation for* MinExposed *under fairness constraints on V_1.*

Proof. The proofs are exactly the same as those of Theorems 1 and 2. □

The above guarantees only apply when demographic groups are disjoint, which is not always the case. To model overlapping demographic groups, we can either allow individuals to be (a) probabilistically assigned to demographic groups or (b) assigned to multiple demographic groups. We note that our results for (a) can also be useful when demographic-group classification is the output of some machine-learning model, and does not only apply to overlapping demographic groups. We also note that both of these extensions also maintain their D-approximation guarantee since those proofs only require the linear program optimality and **(P1)**: $\mathbb{E}[X_u] = x_u$.

Probabilistic Demographic Groups: we want to extend our fairness guarantees above to the case where the demographic characteristics are probabilistic. To formalize this, we assume that each person $u \in V_1$ is in demographic group $\ell \in L$ with probability $\ell_u \in [0, 1]$. Then we want the constraint

$$\sum_{u \in V_1} \ell_u X_u \leq B_\ell, \tag{4}$$

where X_u is the indicator variable for u being asked to quarantine. We claim that by adding this same constraint into the linear program in Section 5.1 (replacing X_u by x_u), DepRound achieves approximate fairness for V_1, as defined below.

Theorem 5. *Let $\epsilon > 0$ and ℓ_u for $u \in V_1, \ell \in L$ be given. If for each $\ell \in L$, we have $B_\ell \geq \frac{(2+\epsilon)\ln(|L|/\delta)}{\epsilon^2}$ for some parameter $\delta \in (0, 1)$, then* DepRound *guarantees that the probability there exists a fairness constraint broken by more than an $1 + \epsilon$ multiplicative factor is at most δ.*

Proof. We begin by noting that the outputs X_u are negatively correlated, as stated in **(P3)**, so we can invoke the results of Panconesi and Srinivasan [1997] to get Chernoff-Hoeffding-like bounds for linear combinations of X_u. In particular, we will have the following for each $\ell \in L$.

$$\Pr[\textstyle\sum_{u \in V_1} \ell_u X_u \geq (1 + \epsilon) B_\ell] \leq \exp(-\epsilon^2 B_\ell / (2 + \epsilon)). \tag{5}$$

By the union bound, we have

$$\Pr[\exists \ell \in L : \textstyle\sum_{u \in V_1} \ell_u X_u \geq (1 + \epsilon) B_\ell] \leq |L| \cdot \exp(-\epsilon^2 B_\ell / (2 + \epsilon)). \tag{6}$$

When B_ℓ is suitably large as in the theorem statement, this probability is at most δ. □

Overlapping Demographic Groups: another case we want to consider is when demographic groups aren't necessarily disjoint. Here, Fair DepRound is no longer well defined because the vectors which we want to apply dependent rounding to now overlap. To get around this, the idea is to split the demographic groups into $2^{|L|}$ new groups corresponding to the subsets of L. These groups are now disjoint, so we can solve the linear program as before and apply dependent rounding separately (and thus independently) to each group. We call this new algorithm Fair DepRound$'$ due to lack of creativity.

Theorem 6. *Fair* DepRound′ *gives the following fairness guarantees, even when demographic groups aren't necessarily disjoint:*

1. *the budget constraints are satisfied in expectation:* $\mathbb{E}[\sum_{u \in R_\ell} X_u] \leq B_\ell$ *for each* $\ell \in L$.
2. *Let* C_ℓ *denote the number of sets* $A \subseteq 2^L$ *such that the set of people with label* A *is nonempty and let* $C^* = \max_\ell C_\ell$. *Then for all* $t > 0$, *the probability that any demographic group's budget is violated by more than an additive* t *is at most* δ, *provided that* $C^* \leq \frac{2t^2}{\ln(|L|/\delta)}$.

Proof. The first part follows directly by **(P1)** and the linearity of expectation. For the second part, let X_A be the subset of nodes in V_1 which have labels $A \subseteq 2^L$. Since $\sum_{u \in X_A} x_u$ is not necessarily integral, **(P2)** doesn't apply. We use the following generalization proved in Srinivasan [2001]:

(P2′) given a vector $x \in \mathbb{R}^d$ with $S = \sum_{i=1}^d x_i$ not necessarily integral, dependent rounding outputs a vector $X \in \{0,1\}^d$ such that $\sum_{i=1}^d X_i \in \{\lfloor S \rfloor, \lceil S \rceil\}$.

In other words, the number of isolations chosen by dependent rounding differs from the budget allocated by the optimal linear program solution by at most 1 in each group $A \subseteq 2^L$. Since rounding is applied independently to the groups, the additive constraint violation can be bounded by Hoeffding's Theorem:

$$\Pr[\textstyle\sum_{u \in R_\ell} X_u - B_\ell \geq t] \leq \exp[-2t^2/C_\ell] \leq \exp[-2t^2/C^*]. \tag{7}$$

By the union bound

$$\Pr[\exists \ell : \textstyle\sum_{u \in R_\ell} X_u - B_\ell \geq t] \leq \exp[-2t^2/C_\ell] \leq |L| \exp[-2t^2/C^*]. \tag{8}$$

Thus, if t is sufficiently large as in the theorem statement, this probability is at most δ. In general, we have that $C^* \leq \min\{|V_1|, 2^{|L|-1}\}$ but this number can be much smaller in practice. □

6.2 Fairness in V_2

Suppose $\mathcal{R}_\ell$ are the (not necessarily disjoint) demographic groups. We want the expected number of infections in each $\mathcal{R}_\ell$ to be at most some given a_ℓ. Adding the following constraint for each $\ell \in [L]$ to the MILP formulation is sufficient to guarantee that the fairness constraint is satisfied approximately:

$$\textstyle\sum_{v \in \mathcal{R}_\ell \cap V_2} \sum_{u \in V_1 : (u,v) \in E} (1 - c_u \cdot x_u) \cdot p_u \cdot q_{uv} \leq a_\ell.$$

Theorem 7. *Let* a_ℓ *for* $\ell \in L$ *and* $\epsilon > 0$ *be given. Define* $w_{u\ell} = p_u \sum_{v \in \mathcal{R}_\ell \cap V_2 : (u,v) \in E} q_{uv}$ *and* $w^* = \max_{u \in V_1, \ell} w_{u\ell}$. *If for each* ℓ, *we have* $a_\ell \geq \frac{(2+\epsilon) w^* \ln(|L|/\delta)}{\epsilon^2}$ *for some parameter* $\delta \in (0,1)$, *then Fair* DepRound *guarantees that the probability that there exists a fairness constraint broken by more than a* $1 + \epsilon$ *multiplicative factor is at most* δ.

Proof. The proof is similar to that of Theorem 5. Let $\{X_u\}$ be the binary vector coming from rounding $\{x_u\}$, and let $Y_u = 1 - X_u$. Let I_ℓ denote the expected number of infections in R_ℓ, given the quarantine set output by the algorithm. First note that

$$\begin{aligned} I_\ell &\leq \textstyle\sum_{v \in \mathcal{R}_\ell} \sum_{u \in V_1 : (u,v) \in E} (1 - c_u \cdot X_u) \cdot p_u \cdot q_{uv} \\ &= \textstyle\sum_{u \in V_1} w_{u\ell} \cdot (1 - c_u \cdot X_u) \end{aligned}$$

for each $\ell \in L$, by the union bound. The random variables X_u are negatively associated [Dubhashi *et al.*, 2007], so the random variables $1 - c_u \cdot X_u$ are also negatively associated. Hence, by tail bounds [Schmidt *et al.*, 1995], we have

$$\Pr[\textstyle\sum_{u \in V_1} w_{u\ell} \cdot (1 - c_u \cdot X_u) \geq (1+\epsilon) \cdot a_\ell] \leq \exp(-\epsilon^2 a_\ell/(2+\epsilon)w^*)$$

for $\ell \in L$. As a result, we can also claim that

$$\Pr[I_\ell \geq (1+\epsilon)a_\ell] \leq \exp(-\epsilon^2 a_\ell/(2+\epsilon)w^*)$$

for $\ell \in L$. Now, by the union bound, we have

$$\Pr[\exists \ell \in L : I_\ell \geq (1+\epsilon)a_\ell] \leq |L| \cdot \exp(-\epsilon^2 a_\ell/(2+\epsilon)w^*).$$

Simple algebra shows that this is at most δ if a_ℓ is suitably large as in the theorem statement. □

7 Practical Implementation

Our MDP formulation of efficient contact tracing and isolation assumes knowledge of the contact graph, transmission rates, and compliance rates. In the real world, however, these values may not be known. While the average transmission rate can be estimated, the compliance rates are difficult to predict and the knowledge of the contact graph is limited (and dependent on the type of contact tracing). In this section, we develop heuristics based on DegGreedy which can be implemented for digital and manual contact tracing.

7.1 Digital Contact Tracing

Many digital contact tracing apps are implemented based on a proximity approach, where devices randomly generate encrypted keys and exchange those keys with devices in their proximity [Abueg *et al.*, 2020]. These exchanges are stored locally in each individual's device. When a person tests positive, they can choose to alert all their contacts through the keys from the list. Though there is no direct cost in alerting contacts, quarantining too many people incurs an economic deficit to society so we still need to limit the number of isolations. Hence, we can apply our framework to digital contact tracing.

When apps are implemented using the proximity approach, we can extract necessary quantities to apply DegGreedy. We assume there is a uniform transmission rate p between contacts and a uniform compliance rate which can be set to 1 without loss of generality. Under these assumptions, DegGreedy reduces to picking nodes u with highest weight w_u, where

$$w_u = |N(u) \cap V_2| \cdot \left[1 - (1-p)^{|N(u) \cap I|}\right].$$

To increase interpretability, when p is small as is the case in COVID-19, we can use a first-order approximation to the Binomial expansion to estimate:

$$w_u \approx p \cdot |N(u) \cap V_2| \cdot |N(u) \cap I|.$$

Finally, we add noise from a discrete Gaussian with $\varepsilon = 1$ to guarantee edge differential privacy for the contact graph [Hay *et al.*, 2009; Bun and Steinke, 2016; Canonne *et al.*, 2020] and pick the B nodes with the highest noisy weight. With this scheme, contact tracing apps can easily implement this variant of DegGreedy, which we denote as Private DegGreedy.

7.2 Manual Contact Tracing

Now we turn our attention to manual contact tracing, which proceeds as follows: when a person tests positive for the disease, they are added to a queue of infected people. Contact tracers then arbitrarily pick and interview people from this queue to extract information about their neighbors. Finally, they contact these neighbors and ask them to quarantine. As a result, policymakers choose nodes in V_1 to contact without information about V_2. Though this restricts the applicability of our results, our algorithms still motivate useful heuristics for contact tracing in the above process.

Like before, we will assume the policymakers only know the average transmission and compliance rates. Recall from our digital contact tracing analysis that DegGreedy in the case where all transmission and compliance rates are assumed to be uniform already favors picking nodes with higher degree. In particular, when transmission rates are 1, DegGreedy is exactly equivalent to picking nodes in V_1 with highest degree in V_2. We emphasize that although one may claim this is a very intuitive result, our work is the first to motivate this theoretically.

The importance of selecting high degree nodes motivates a heuristic, which we call SegDegree (due to how we simulate it in experiments). The idea is to garner additional information during interviews with the infected nodes: when asking for their neighbors, we can also ask them to classify each neighbor into sets of *high* ($\mathcal{H}$) or *low* ($\mathcal{L}$) degree. Then we randomly/arbitrarily pick nodes from the set of high degree nodes $\mathcal{H}$ to contact trace. In our experiments, we simulate SegDegree by ranking the nodes in V_1 by their degree. We define $\mathcal{H}$ to be nodes with degree in the top 25%; the remaining nodes are in $\mathcal{L}$. In order to represent inaccurate judgement of high/low degrees, we sample $3B/4$ nodes from $\mathcal{H}$ and $B/4$ nodes from $\mathcal{L}$ to be our final quarantine set Q

We note that this should be viewed as practical contributions motivated by the simplicity of current manual contact tracing implementations: picking arbitrary nodes from the set of exposed individuals. This restricts the potential effectiveness of manual contact tracing, which we our results and recommendation here can improve. In the practical implementation, we acknowledge the importance of mitigating any personal biases given the subjective nature of this classification process. Such methods may include providing defined classification thresholds and clear category specifications, and is left to the practitioner.

8 Experiments

Disease model. Our setup for the epidemic simulation is modeled loosely based on COVID-19. We assume a simple SIR model, with infectious duration of two time steps. At each timestep, we have a susceptible set (S), infected set ($I = I_1 \dot{\cup} I_2$), and a recovered set (R). Nodes in I_1 got infected this timestep and nodes in I_2 have already been infected for one timestep. While both I_1 and I_2 transmit the disease, all quarantine decisions will be made based on I_2 only. This represents how policymakers have incomplete information about the infection status of individuals: I_1 is not yet known to be carrying the disease because there is a 4-5 day incubation period and a wait time for COVID-19 testing. By the next timestep, I_1 has undergone testing and becomes I_2, now known to the policymaker. We note that even though this model is slightly different from the one in our theoretical analysis, our algorithms and problem formulation are still applicable since only I_2 is known to the policymaker.

Model parameters. For each simulation of the MDP, intervention begins at an early timestep with constant budget and continues over the course of the epidemic. Transmission probabilities are set based on the length of contact time and compliance probabilities are set based on the age group

of the person in accordance with the relative order presented in [Lou *et al.*, 2020] and [Carlucci *et al.*, 2020] (see Appendix for details). For digital contact tracing, we set the compliance probabilities to be half that of manual contact tracing. Each quarantine recommendation will instruct the individual to isolate for 2 timesteps. Under this setup, the performances of the intervention algorithms are compared using two different metrics: total number of infections to assess the impacts of an epidemic and the number of known infections at each timestep to maintain a manageable number of cases with respect of hospital resources and infrastructure.

Social contact networks. We use synthetic social contact networks for two counties in Virginia (summarized in Table 2) constructed by a first principles approach by Barrett *et al.* [2009] and Eubank *et al.* [2004]. We enforce fairness constraints and simulate varying compliance rates using the demographic data on age groups given in our social networks (see Table 1). Because casual contacts (e.g., during commuting) are not represented in these networks, we augment each network by increasing the degree of each node by about 15% [Keeling *et al.*, 2020] and show the robustness of our results by experimenting on these networks as well.

Table 1: Age group demographic information

Age Group	Name	Range (years)	Compliance Rate	Montgomery Population (%)	Albemarle Population (%)
p	pre-school	0-4	0.75	5	3
s	school-aged	5-17	0.80	15	11
a	adults	18-49	0.60	43	49
o	older-adults	50-64	0.85	21	23
g	golden-aged	65+	0.80	16	15

Budget. For manual contact tracing, we set the budget based on the state of Virginia, which has a population of roughly 8 million people and currently employs around 2000 contact tracers [VDH, 2020]. We then estimate the number of contact tracers for our graphs to be proportional to the population. Since each interview with an individual that has contracted COVID-19 takes 30 to 60 minutes [VDH, 2020], a contact tracer can make 4 to 8 isolation suggestions per day (or around 28 to 56 per timestep). We use this information to estimate the budget for the number of isolations. For digital contact tracing, we let the budget range from 0% to 5% of the population in order to understand the tradeoff between economic costs and disease intervention.

Table 2: Description of datasets (* indicates the network is augmented)

Network name	$\|V\|$	$\|E\|$	Max degree	estimated # of contact tracers	Budget
Montgomery	75457	648667	105	18-19	500-1000
Montgomery*	75457	768383	120	18-19	500-1000
Albemarle	131219	1423151	176	32-33	900-1800
Albemarle*	131219	1687724	205	32-33	900-1800

8.1 Comparison of Methods

We first compare our practical heuristics and theoretical algorithms against corresponding baselines by running simulations of the MDP with the budget set according to Table 2. To demonstrate the quality of our full information algorithms, we compare it with EC, a baseline which selects the nodes in V_1 with the highest eigenvector centrality for quarantine. We chose EC as a baseline since its a centrality measure which uses information from the full network, and we want to see how our local methods compare. Furthermore, EC is related to a heuristic studied for minimizing a graph's spectral radius [Tong *et al.*, 2012], which controls the size of the disease spread [Wang *et al.*, 2003].

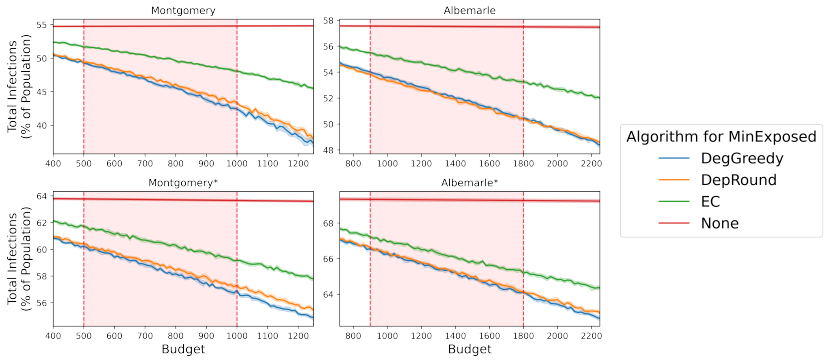

Figure 3: Algorithms for contact tracing under full information (estimated budget is shaded)

Despite requiring more information, EC performs significantly worse than DepRound and DegGreedy, as seen in Figure 3. Additionally, the sensitivity with respect to budget is about half that of DepRound and DegGreedy. Ultimately, the better performance and stronger sensitivity of our algorithms with respect to budget show that DegGreedy and DepRound may be useful in some places, such as China, where the second neighborhood's information is available.

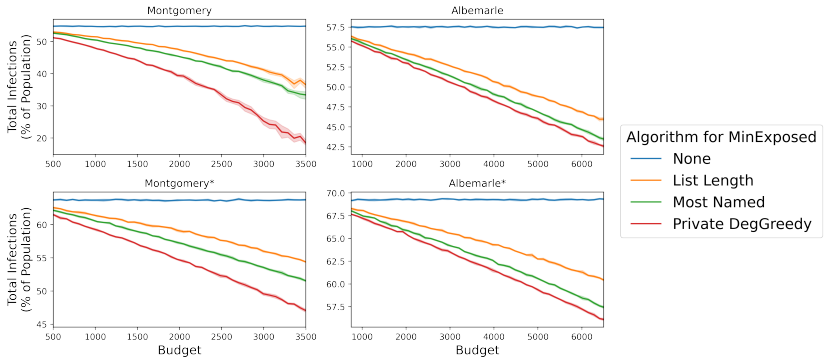

Figure 4: Comparison of digital contact tracing algorithms

Next, we compare Private DegGreedy with two intuitive baselines studied in Armbruster and Brandeau [2007b]: the MostNamed policy and ListLength policy. The MostNamed policy selects nodes in V_1 with the most infected neighbors and the ListLength policy is similar, but weighs each neighbor by the inverse of their degree. From Figure 4, we see that our heuristic improves upon the baseline without incurring more privacy loss. Interestingly, the margin of improvement is larger

for the Montgomery networks which have higher edge density. This is a result of adding discrete Gaussian noise: the noise added to w_u is $o(w_u)$, so the effect of the noise decreases as absolute degrees increase. We note that performing better in high density networks is a desirable quality here: diseases spread especially fast in such settings, making contact tracing even more crucial.

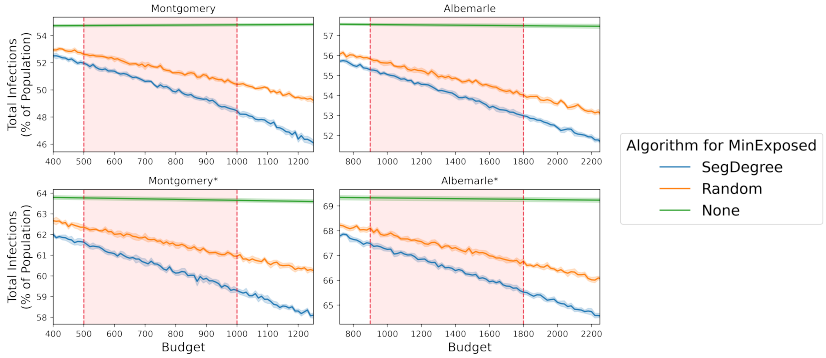

Figure 5: Comparison of manual contact tracing algorithms (estimated budget is shaded)

Finally, we compare SegDegree with the baseline adopted by many states in the United States: selecting nodes in V_1 at random [NAS, 2021; VDH, 2020]. Figure 5 shows that introducing a simple additional step in manual contact tracing decreases total infections by 50% more than Random when compared to no intervention. Furthermore, SegDegree has a larger sensitivity with respect to budget which makes investing in new contact tracers more effective.

8.2 Visualizing the Epicurve

In addition to decreasing the total infections, our methods reduce the peak of the curve and shift it to occur at later timesteps (see Figure 6). This is important in practice since a later peak enables time for developing of vaccines, which can potentially stop the infection before the peak. As mentioned before, having a lower peak is important as well since hospital capacity is limited; if the peak number of infections is too high, many people are unable to receive adequate treatment.

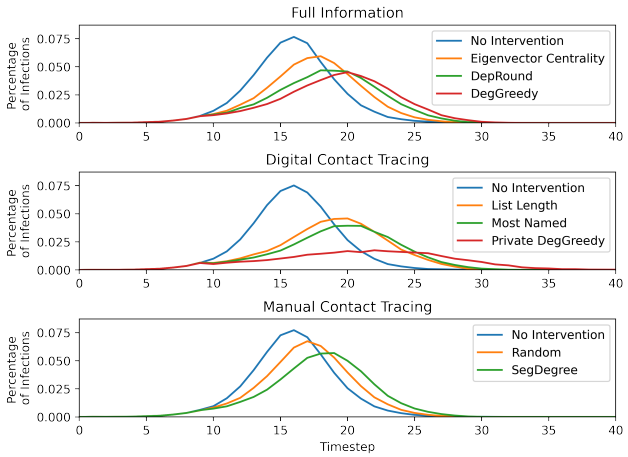

Figure 6: Montgomery Epicurve Visualizations (See Appendix B.1 for epicurves on other networks)

8.3 The Price of Fairness

Due to the economic and social costs of self-isolation, it is important that policymakers ensure demographics are not disproportionately impacted. In these experiments, we focus on age groups and consider four policies: **(A)** no fairness constraint **(B)** the budget is proportional to the population of each age group **(C)** more budget is allocated to the older population **(D)** less budget is allocated to the working age population (see Appendix for formal definitions). As seen in Table 3, Policy A (with no fairness constraints) leads to the lowest total infections, but the differences are not statistically significant. Thus, upholding (reasonable) fairness constraints does not significantly reduce the efficacy of our algorithms.

Table 3: Comparison of Fair DepRound and Fair DegGreedy under different policies

Algorithm	County	Policy	Original	Augmented
DepRound	Montgomery	A	45.31 ± 0.44	57.77 ± 0.23
		B	45.45 ± 0.40	57.82 ± 0.22
		C	45.58 ± 0.36	57.84 ± 0.21
		D	45.47 ± 0.34	57.85 ± 0.23
	Albermarle	A	51.42 ± 0.17	64.65 ± 0.15
		B	51.50 ± 0.21	64.74 ± 0.15
		C	51.59 ± 0.17	64.83 ± 0.14
		D	51.51 ± 0.13	64.77 ± 0.15
DegGreedy	Montgomery	A	44.72 ± 0.41	57.45 ± 0.22
		B	44.81 ± 0.41	57.54 ± 0.23
		C	44.94 ± 0.43	57.54 ± 0.20
		D	44.85 ± 0.41	57.50 ± 0.19
	Albemarle	A	51.69 ± 0.21	64.61 ± 0.17
		B	51.66 ± 0.20	64.66 ± 0.19
		C	51.72 ± 0.18	64.70 ± 0.17
		D	51.71 ± 0.18	64.65 ± 0.18

9 Conclusions

Here, we formulate the problem of efficient contact tracing as a MDP and each timestep of the MDP as a combinatorial problem, MinExposed. Since MinExposed is NP-Hard, we give an approximation algorithm for it by formulating it as a linear program and performing dependent rounding. Motivated by the analysis of DepRound, we devise a greedy algorithm which is more interpretable and extendable to cases where there is a realistic amount of information available. We modify DegGreedy to be implementable with limited knowledge in both digital and manual contact tracing. Though motivated by our theoretical guarantees, our devised practical heuristics (*i*) do not need network information, (*ii*) do not need disease model information, and (*iii*) only require the approximate degrees of nodes in V_1 (and V_2 for digital contact tracing). Our heuristics, which are simple and robust, can easily be deployed in practice. We then show that despite the

minimal knowledge required, they perform strongly in our experiments. Despite our heuristic, a limitation of our theoretical model is the assumption of contact graph knowledge. A natural next step is to combine our model with that of Meister and Kleinberg [2021] to include graph discovery as part of the contact tracing model.

Acknowledgements: George Li, Aravind Srinivasan, and Zeyu Zhao were supported in part by NSF award number CCF-1918749. Ann Li, Arash Haddadan, Madhav Marathe, and Anil Vullikanti were supported in part by NSF award number CCF-1918656.

References

Matthew Abueg, Robert Hinch, Neo Wu, Luyang Liu, William Probert, Austin Wu, Paul Eastham, Yusef Shafi, Matt Rosencrantz, Michael Dikovsky, Zhao Cheng, Anel Nurtay, Lucie Abeler-Dörner, David Bonsall, Michael V. McConnell, Shawn O'Banion, and Christophe Fraser. Modeling the combined effect of digital exposure notification and non-pharmaceutical interventions on the COVID-19 epidemic in Washington state. *medRxiv*, 2020.

Nadeem Ahmed, Regio A. Michelin, Wanli Xue, Sushmita Ruj, Robert Malaney, Salil S. Kanhere, Aruna Seneviratne, Wen Hu, Helge Janicke, and Sanjay K. Jha. A Survey of COVID-19 Contact Tracing Apps. *IEEE Access*, 8:134577–134601, 2020.

Benjamin Armbruster and Margaret L Brandeau. Contact tracing to control infectious disease: when enough is enough. *Health Care Managment Science*, 10(4):341–355, 2007.

Benjamin Armbruster and Margaret L Brandeau. Who do you know? a simulation study of infectious disease control through contact tracing. In *Proceedings of the 2007 Western Multiconference on Computer Simulation*, pages 79–85, 2007.

Christopher L. Barrett, Richard J. Beckman, Maleq Khan, V. S. Anil Kumar, Madhav V. Marathe, Paula E. Stretz, Tridib Dutta, and Bryan Lewis. Generation and analysis of large synthetic social contact networks. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1003–1014, 2009.

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. *CoRR*, abs/1605.02065, 2016.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The Discrete Gaussian for Differential Privacy. In *NeurIPS*, 2020.

Leonardo Carlucci, Ines D'Ambrosio, and Michela Balsamo. Demographic and Attitudinal Factors of Adherence to Quarantine Guidelines During COVID-19: The Italian Model. *Frontiers in psychology*, 11, October 2020.

Devdatt P. Dubhashi, Johan Jonasson, and Desh Ranjan. Positive Influence and Negative Dependence. *Comb. Probab. Comput.*, 16(1):29–41, 2007.

Ken TD Eames. Contact tracing strategies in heterogeneous populations. *Epidemiology & Infection*, 135(3):443–454, 2007.

Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180—184, May 2004.

S. Eubank, A. Vullikanti, M. V. Marathe, et al. Structure of Social Contact Networks and Their Impact on Epidemics. In *Discrete Methods in Epidemiology*, volume 70, pages 179–200. American Math. Soc., Providence, RI, 2006.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., USA, 1979.

Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate Estimation of the Degree Distribution of Private Networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178, 2009.

A. Hayrapetyan, D. Kempe, M. Pál, et al. Unbalanced Graph Cuts. In *ESA*, pages 191–202, 2005.

Matt J Keeling, T Déirdre Hollingsworth, and Jonathan M Read. The Efficacy of Contact Tracing for the Containment of the 2019 Novel Coronavirus (COVID-19). *medRxiv*, 2020.

Istvan Z Kiss, Darren M Green, and Rowland R Kao. Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1407–1414, 2005.

Istvan Z Kiss, Darren M Green, and Rowland R Kao. The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. *Journal of the Royal Society Interface*, 5(24):791–799, 2008.

Mirjam E Kretzschmar, Ganna Rozhnova, Martin C J Bootsma, Michiel van Boven, Janneke H H M van de Wijgert, and Marc J M Bonten. Impact of delays on effectiveness of contact tracing strategies for covid-19: a modelling study. *The Lancet Public Health*, 5(8):e452 – e459, 2020.

Fengchen Liu, Wayne T. A. Enanoria, Jennifer Zipprich, Seth Blumberg, Kathleen Harriman, Sarah F. Ackley, William D. Wheaton, Justine L. Allpress, and Travis C. Porco. The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for california. *BMC Public Health*, 15(1):447, May 2015. PMCID: PMC4438575.

Feng Liu, Xin Li, and Gaofeng Zhu. Using the contact network model and Metropolis-Hastings sampling to reconstruct the COVID-19 spread on the "Diamond Princess". *Science bulletin*, 65(15):1297–1305, 2020.

Lars Lorch, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. A Spatiotemporal Epidemic Model to Quantify the Effects of Contact Tracing, Testing, and Containment. *arXiv preprint arXiv:2004.07641*, 2020.

Qing Lou, De-Quan Su, Sun-Qin Wang, E Gao, Lian-Qiao Li, and Zhi-Qiang Zhuo. Home quarantine compliance is low in children with fever during COVID-19 epidemic. *World journal of clinical cases*, 8(16):3465–3473, August 2020.

Madhav Marathe and Anil Vullikanti. Computational Epidemiology. *Communications of the ACM*, 56(7):88–96, 2013.

Michela Meister and Jon Kleinberg. Optimizing the order of actions in contact tracing, 2021.

Marco Minutoli, Prathyush Sambaturu, Mahantesh Halappanavar, Antonino Tumeo, Ananth Kalyanaraman, and Anil Vullikanti. PREEMPT: Scalable Epidemic Interventions Using Submodular Optimization on Multi-GPU Systems. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2020.

State Approaches to Contact Tracing during the COVID-19 Pandemic, May 2021. `https://www.nashp.org/state-approaches-to-contact-tracing-covid-19/`.

A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the chernoff-hoeffding bounds. *SIAM J. Comput.*, 26:350–368, 1997.

Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

Marcel Salathé, Christian L Althaus, Richard Neher, Silvia Stringhini, Emma Hodcroft, Jacques Fellay, Marcel Zwahlen, Gabriela Senti, Manuel Battegay, Annelies Wilder-Smith, et al. COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. *Swiss medical weekly*, 150(1112), 2020.

Prathyush Sambaturu, Bijaya Adhikari, B Aditya Prakash, Srinivasan Venkatramanan, and Anil Vullikanti. Designing Effective and Practical Interventions to Contain Epidemics. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1187–1195, 2020.

Andrew A Sayampanathan, Cheryl S Heng, Phua Hwee Pin, Junxiong Pang, Teoh Yee Leong, and Vernon J Lee. Infectivity of asymptomatic versus symptomatic COVID-19. *The Lancet*, 397(10269):93–94, 2021.

J. P. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-Hoeffding Bounds for Applications with Limited Independence. *SIAM Journal on Discrete Mathematics*, 8:223–250, 1995.

Aravind Srinivasan. Distributions on level-sets with applications to approximation algorithms. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 588–597, 2001.

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

Virginia Department of Health, Jan 2020. `https://www.vdh.virginia.gov/coronavirus/prevention-tips/contact-tracing/`.

Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. IEEE Computer Society Press, 2003.

Amulya Yadav, Hau Chan, Albert Xin Jiang, Haifeng Xu, Eric Rice, and Milind Tambe. Using Social Networks to Aid Homeless Shelters: Dynamic Influence Maximization under Uncertainty. In *AAMAS*, volume 16, pages 740–748, 2016.

Appendix A: Experimental Details

A.1: Computational Setup

We run our simulations on Amazon EC2 c5a.24xlarge instances with 96 vCPUs and and 185GB of RAM. To solve LP and MILP problems, we used Google OR-Tools [Perron and Furnon] with a Gurobi (version 9.1) [Gurobi Optimization, 2021] backend. To simulate the disease spread on our networks, we used Epidemics on Networks [Miller and Ting, 2020]. The full list of software dependencies can be found in our code (https://github.com/gzli929/ContactTracing).

A.2: Experimental Parameters

We run most of our experiments on 4 networks: Montgomery, Albemarle, Montgomery*, and Albemarle*. The default budget is set as the center of the predicted range. All Montgomery graphs, unless otherwise stated, are run with 750 budget for manual contact tracing and 2700 for digital contact tracing. All Albemarle graphs are run with 1350 budget for manual contact tracing and 4700 for digital contact tracing. The demographic labels and contact duration times for the Montgomery graph are sampled from the distribution of the Albemarle graph.

Contact duration times are transformed into transmission rates by defining an exponential cumulative distribution function such that the average duration is equal to the average transmission. We set the average transmission parameter as 0.05 and held it constant across all our experiments. The compliance rates for each individual follow the age group averages but have added noise from the uniform distribution of [-0.05, 0.05]. Since individuals are less likely to comply to quarantine recommendations from digital apps, we scale the compliance rates for each network to average around 50% for our digital contact tracing algorithms. We also add discrete Gaussian noise with $\epsilon = 1$ to ensure differential privacy for our digital contact tracing baselines. Unless otherwise stated, we conducted our sensitivity experiments with these default values and plotted the 95% confidence interval for the average of 10 trials.

In the fairness experiments, the policies are defined formally as follows. We are given an infected set I and total budget B. Let $n(l)$ be the number of people in V_1 with labels l, for labels p, s, a, o, g. Let $n = \sum_{l \in L} n(l)$.

- **(A)** no age consideration: there is only one label with budget B.
- **(B)** the isolation budget is distributed proportional to the population of each age group, i.e. $B_l = B \cdot \frac{n(l)}{n}$ for each $l \in L$.
- **(C)** more budget is allocated to the older population (age group g), i.e. $B_l = B \cdot \frac{n(l)}{n+n(g)}$ for $l \neq g$ and $B_g = 2B \cdot \frac{n(g)}{n+n(g)}$.
- **(D)** less budget is allocated to the working age population (age groups a and o), i.e. $B_l = B \cdot \frac{n(l)}{n+n(g)+n(p)+n(s)}$ for $l \in \{a, o\}$ and $B_l = 2B \cdot \frac{n(l)}{n+n(g)+n(p)+n(s)}$ for $l \in \{p, s, g\}$.

Appendix B: Additional Experiments

B.1: Epicurve Visualizations

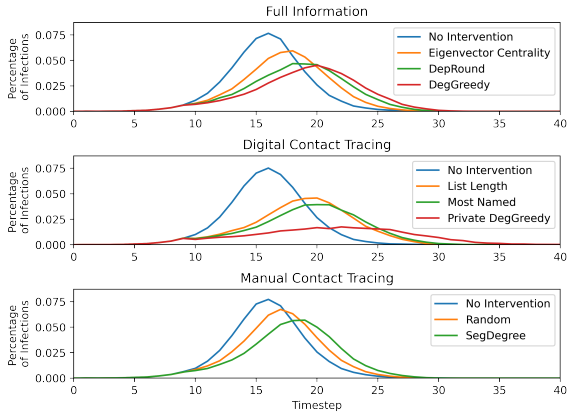

Figure 7: Epicurve Visualizations for Montgomery

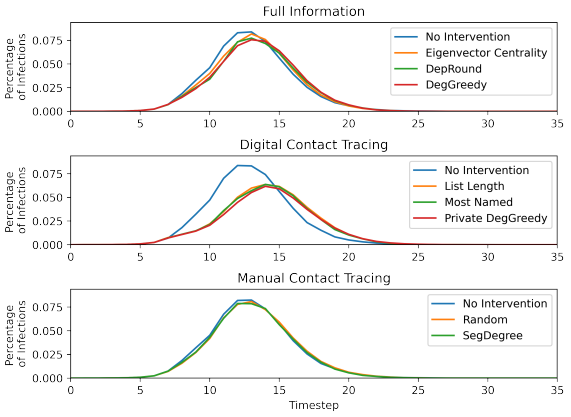

Figure 8: Epicurve Visualizations for Albemarle

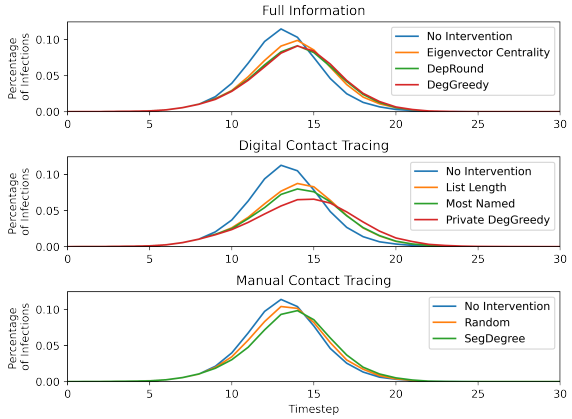

Figure 9: Epicurve Visualizations for Montgomery (augmented)

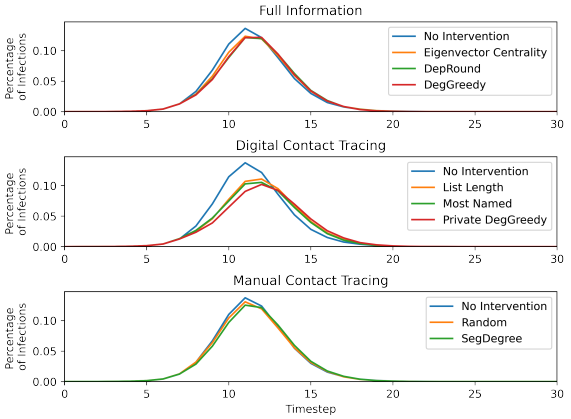

Figure 10: Epicurve Visualizations for Albemarle (augmented)

Here, we reproduce the epicurve plots shown in Section 7.2 for the remaining three counties. As seen in the above figures, each of our algorithms reduce and shift the peak of the epicurve in all of the social networks. In particular, Private DegGreedy consistently performs much better than the baselines on all four social networks. However, this improvement is less obvious when experimenting on Albemarle county. A similar phenomenon was also seen and explained in the main paper: when the degrees are far apart, then there is a larger difference between our algorithms (based on degree) and the baselines. Consider the extreme case where all degrees are equal; then any algorithm based on degree is arbitrary. We believe this is the reason algorithms such as SegDegree and Private DegGreedy perform well on Montgomery (where the edge density is very high) and less well on

Albemarle (where the edge density is much lower). Additionally, note that compliance rates are relatively low, adding more noise to the equation.

B.2: Peak Infections Comparison

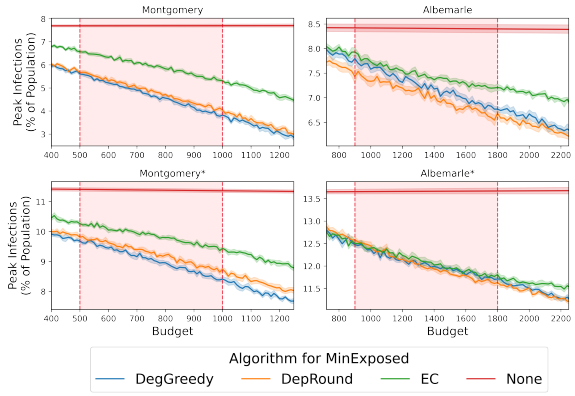

Figure 11: Budget Sensitivity for Peak Infections (Full Information Algorithms)

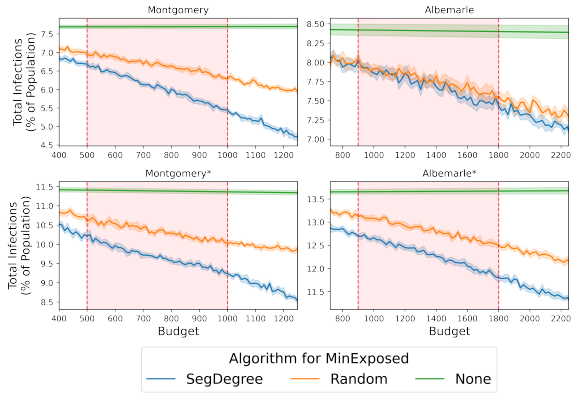

Figure 12: Budget Sensitivity for Peak Infections (Manual Contact Tracing Algorithms)

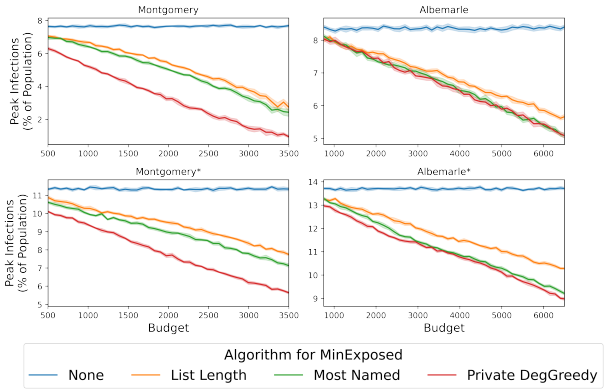

Figure 13: Budget Sensitivity for Peak Infections (Digital Contact Tracing Algorithms)

As seen in the sensitivity plots for each of the three contact tracing scenarios, our algorithms decrease the maximum number of people infected during any timestep, which we call the peak. While our algorithms perform similarly under the full information setting, DegGreedy exhibits stronger sensitivity to budget across all networks. Additionally, the stronger performance of DegGreedy on Montgomery (particularly with augmentation) suggests it may be especially effective on denser networks. In the setting of manual contact tracing, SegDegree consistently outperforms the Random baseline and the discrepancy increases as budget increases. In digital contact tracing, our algorithm Private DegGreedy outperforms MostNamed and ListLength on the Montgomery networks but has a similar performance with MostNamed Albemarle. Even so, Private DegGreedy generally results in a lower peak when the network is augmented, suggesting that it may be more

advantageous on denser networks. Across all the networks, Private DegGreedy exhibits stronger budget sensitivity than ListLength when lowering the peak number of infections.

B.3: Empirical Approximation Ratio

Here, we evaluate the empirical approximation factor of our algorithms and heuristics. We use the MILP optimal objective value to lower bound the true optimal when calculating the ratios.

	Albemarle				Montgomery			
Bucket	0	1	2	3	0	1	2	3
I ($\times 10^3$)	0.36	2.89	7.46	3.84	1.60	4.03	1.77	0.99
$\lvert V_1\rvert$ ($\times 10^3$)	2.06	13.72	35.01	32.79	6.30	20.13	17.21	13.97
$\lvert V_2\rvert$($\times 10^3$)	8.97	20.40	25.82	57.52	8.19	17.24	28.88	37.93
$\lvert(V_1 \times V_2) \cap E\rvert$ ($\times 10^3$)	11.82	45.68	90.73	298.70	12.76	44.52	91.63	123.52
D	7.37	17.20	32.27	72.79	12.85	27.69	37.96	41.06

Table 4: Summary of samples for which we calculate the empirical approximation ratio: Montgomery (490 instances) and Albemarle (461 instances). Samples come from simulating the MDP.

MinExposed Algorithms			bucket 0	bucket 1	bucket 2	bucket 3
DegGreedy	Approx. Factor	max	1.229	1.670	1.771	1.724
		mean	1.102	1.380	1.435	1.470
	Time Elapsed	max	1.887	6.654	4.172	1.768
		mean	0.865	4.270	1.525	0.666
DepRound	Approx. Factor	max	1.362	1.796	1.915	1.871
		mean	1.169	1.479	1.631	1.663
	Time Elapsed	max	5.337	18.495	14.893	27.754
		mean	1.383	7.093	7.161	8.858
SegDegree	Approx. Factor	max	1.777	1.918	2.039	1.994
		mean	1.484	1.656	1.762	1.793
	Time Elapsed	max	0.036	0.112	0.743	0.093
		mean	0.014	0.046	0.051	0.032
Random	Approx. Factor	max	2.055	2.033	2.084	2.052
		mean	1.631	1.779	1.896	1.879
	Time Elapsed	max	0.002	0.003	0.003	0.002
		mean	0.001	0.001	0.001	0.001

Table 5: Summary of performance of different algorithms for MinExposed on instances of Montgomery with budget of 750.

MinExposed Algorithms			bucket 0	bucket 1	bucket 2	bucket 3
DegGreedy	Approx. Factor	max	1.086	2.173	2.061	2.550
		mean	1.068	1.271	1.513	2.033
	Time Elapsed	max	0.106	11.560	19.914	18.356
		mean	0.101	3.861	9.996	6.607
DepRound	Approx. Factor	max	1.129	2.173	3.091	2.803
		mean	1.091	1.321	1.638	2.182
	Time Elapsed	max	6.610	153.246	296.449	1328.052
		mean	2.231	23.178	72.142	344.155
SegDegree	Approx. Factor	max	1.401	2.173	2.942	2.878
		mean	1.280	1.434	1.743	2.273
	Time Elapsed	max	0.008	0.093	0.184	0.167
		mean	0.005	0.030	0.084	0.077
Random	Approx. Factor	max	1.491	2.173	3.176	2.937
		mean	1.301	1.467	1.771	2.318
	Time Elapsed	max	0.001	0.003	0.005	0.004
		mean	0.001	0.001	0.003	0.002

Table 6: Summary of performance of different algorithms for MinExposed on instances of Albemarle with budget of 1350.