Optimizing Pool Testing for Epidemic Surveillance

Jack Heavey
University of Virginia
Charlottesville, United States of America
jch7jm@virginia.edu

Abhijin Adiga
University of Virginia
Charlottesville, United States of America
abhijin@virginia.edu

Anil Vullikanti
University of Virginia
Charlottesville, United States of America
vsakumar@virginia.edu

ABSTRACT

Testing is one of the key tools in public health surveillance. Often testing resources are limited, and pooled testing emerged as a viable strategy during the COVID outbreak, for early detection of the outbreak or clearing the most number of individuals (maximum "welfare"). Here, we study the problem of selecting pools for testing which maximizes welfare. However, this problem is a very challenging optimization problem because the infection status of individuals can be correlated. Prior work on choosing pools has ignored network correlations.

We design an efficient approximation algorithm for this problem, using techniques from stochastic and combinatorial optimization: sample average approximation, linear programming and randomized rounding. We further speed up our algorithms using techniques for combinatorially solving the linear program. We evaluate our method on multiple networked datasets, including one derived from a hospital, and show significant benefit in modeling network correlations.

KEYWORDS

Networks, optimization, approximation algorithms, SOPS

ACM Reference Format:
Jack Heavey, Abhijin Adiga, and Anil Vullikanti. 2026. Optimizing Pool Testing for Epidemic Surveillance. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/KTGS1394

1 INTRODUCTION

Testing has always been recognized as one of the key tools in public health surveillance [1, 4, 13, 20, 28], and many other areas, such as infrastructure networks and social networks which involve spreading processes [14, 18, 21]. During the COVID-19 pandemic, rapid testing was one of the main strategies in controlling infections, e.g., [19, 23, 25]. This was particularly crucial because of asymptomatic infectiousness of COVID-19.

Many types of objectives have been studied in the context of surveillance, such as (1) maximizing the probability of detection [21], (2) minimizing the delay in detection [12, 21], and (3) clearing the most number of uninfected patients (referred to as welfare maximization) [10, 22]—this was motivated by the need for a negative test result to get back to normal activities during the pandemic, e.g., [10, 22, 29]. In the early part of the pandemic, there was a severe shortage of resources for testing, such as personnel, lab equipment and chemical reagents (e.g., for the RT-PCR test) [9, 17, 24]. This motivated the classical "pool testing" strategy, first proposed in 1943 [5] (also referred to as group testing): samples from multiple individuals are combined into a single test for a specific pathogen [9, 24]. If the test returns negative, then every individual within the pool is cleared of the disease, otherwise, it is an indicator that at least one individual within the pool is infected.

Designing testing strategies which optimize these objectives within given resource constraints are challenging problems. Efficient algorithms are known for the first two objectives (detection probability and delay) [12, 21], but the welfare maximization problem remains poorly understood. [10] design algorithms for welfare maximization in the simplified setting where there are no correlations between infections among individuals. As we show here, ignoring network correlations can significantly impact the maximum welfare.

In this paper, we study the problems of selecting pools for maximizing the welfare of a population (the MaxWelfare problem). In this settings, we consider the diseases spreading across a known contact network contact network. Our contributions are summarized below.

- We show that ignoring network correlations, as in [10], can have a significant impact on both the maximum welfare and minimum detection delay. We also show that errors in the model of epidemic spread can also have a significant impact on the optimal strategies (Section 4).
- We develop bicriteria approximation algorithms for these problems (Section 5), using stochastic optimization techniques, with rigorous worst case performance bounds. While these are polynomial time algorithms, they require solving large linear programs (LPs), and do not scale for large networks. Using multiplicative weights based combinatorial techniques for solving LPs [34], we show that our algorithms can be scaled by multiple orders of magnitude, without degrading the performance.
- We provide experimental results on three real world networks (including a contact network constructed using Electronic Health Record data from a Virginia academic hospital) for both our traditional and ϵ-approximation algorithm. We show that the practical performance of our algorithms is better than the worst case bounds we show in our theoretical analysis. While our sample average approximation technique based analysis requires $\Omega(n^2 \log n)$ simulated disease cascades over a graph of size n, we show in practice that a linear number of sampled cascades is sufficient, allowing us to scale to much larger problem instances. Our multiplicative weights based algorithm also has good practical

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/KTGS1394

performance, and scales to large instances. Our results show that in real instances also, ignoring network correlations has a very significant impact on the welfare, matching our theoretical bounds. We also find that errors in the disease model decrease calculated welfare at a linear rate, and that there is a diminishing returns effect of increasing the pool size.

2 RELATED WORK

The idea of pooling multiple samples together within a single test was first proposed in Dorfman [6]. In this initial work, Dorfman finds the percentage cost savings in ascertaining the exact infection status of all individuals based on the disease prevalence within the population when compared to using individual tests for each member of the population of interest. These savings in reduced total test numbers require an accurate estimation of disease prevalence, however, which is often an unknown quantity in real world epidemic settings.

[22] and [15, 16] examine a different problem, proposing a dual-objective optimization problem of minimizing the spread of a pathogen throughout a population while also minimizing the number of uninfected individuals placed in quarantine, all while given a fixed testing budget. These works considered the heterogeneity within the population, assigning both factors that correspond to the likelihood of an individual being infected as well as a cost to self isolation that differs between different professions within a population.

[10] is the first work that limits the maximum group size based on practical lab limitations and rather looks to clear individuals from quarantine, maximizing the sum population utility of individuals that are cleared given a set of pooled tests. This work was borne from a real-world experiment performed at the Potosinian Institute for Scientific and Technological Research in San Luis Potosí, Mexico. Lock et al. also explore the theoretical improvement of overlapping pool tests compared to the practical limitations of using a single sample across multiple tests, bounding the maximum improvement in expected welfare.

Pool testing has been applied in other domains such as industrial testing, experiment design, coding theory [7] with several variants considered [2]. There has been renewed interest in this topic in the context of COVID-19 [31]. [27] consider the objective of minimizing the pooled testing efficiency, which is measured by the ratio of the expected number of correct classifications to the expected number of tests performed and use simulated annealing to solve this problem. [3] propose two-step sampled group testing algorithms with provable guarantees for an infection model on random connection graphs. Our work, unlike the above, provides guarantees for pooled testing on arbitrary graphs.

Our algorithm assumes that a potential network over which an epidemic might spread is known or learnable. This assumption may not always hold, where active learning has shown to be effective on only partially observed graphs[11, 30].

3 PRELIMINARIES

We assume a discrete time SIR model on a graph $G = (V, E)$, in which the disease spreads from an infectious node u to a susceptible neighbor $v \in N(u)$ with probability p_{uv}, independently, for exactly one time step. Each infected node u then enters the recovered state on the next time step. For a node that enters the recovered state at time τ, the node will still test positive until time $\tau + h$ - this reflects the waning immunity to a disease that occurs, with the duration of h depending on the specific pathogen that has been recovered from. Let $\text{SRC} \subset V$ denote a set of sources for an outbreak; the sources could also be random. A cascade H (also referred to as an outbreak) is a random subgraph of G on which the disease spreads, starting at the subset SRC. Figure 1 shows examples of the SIR process and cascades in the graph.

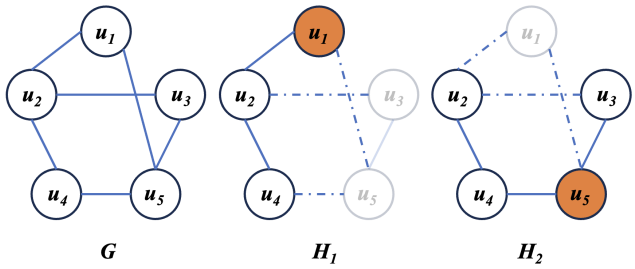

Figure 1: An example contact network and sampled cascades. H_1 has node u_1 as the source of the infection (infected at $\tau = 1$), spreading the infection to u_2 at time $\tau = 2$ and to u_4 at $\tau = 3$. u_3 and u_5 are uninfected. Similarly in H_2, we have u_5 as the source of our infection, u_3 and u_4 infected at time $\tau = 2$, u_2 infected at time $\tau = 3$. Node u_1 is uninfected in this cascade.

As in [10], a pool test t_S corresponds to a subset $S \subset V$ of nodes, which are tested in a pooled manner. If the test is given at time τ, the test will return positive if any node $v \in S$ entered the **Infected** state in the interval $[\tau - h, \tau]$. Otherwise, the test will return negative, meaning all nodes $v \in S$ are either in the **Susceptible** state or entered the **Recovered** state at time $\tau' < \tau - h$. We assume there is a bound n_p on the number of potential samples that can be combined into a single pool. A test t_S which results in a negative result clears all the nodes $v \in S$ simultaneously. Let m_S denote the set of all possible n_p-subsets of V–this corresponds to the potential set of pools that can be tested. Let $T \subset m_S$ be a set of pooled tests selected. Let T^- denote the subsets of pool tests which came out negative, i.e., each test $t_S \in T^-$ is reported as negative. Each test $t_S \in T^+ = T \setminus T^-$ had a positive individual, and the result of the pool test is positive. Let $W(T) = \cup_{t_S \in T^-} S$ denote the set of nodes cleared by T.

We assume each individual $i \in V$ has a utility w_i if they are cleared by a set of tests T, meaning $i \in W(T)$. Here, we study the welfare maximization problem (MaxWelfare) [10], where the goal is to maximize the sum of the utilities of individuals cleared by the pool tests. We make a strong assumption that the tests are all done within a short time scale, which is negligible compared to the disease spread. This is formally defined below.

The MaxWelfare problem. Given a graph $G = (V, E)$, a disease model $M = (\mathbf{p}, \text{SRC}, h)$ (which specifies the transmission probability and sources), pool size n_p, a budget B, and current time step τ, choose a set of pools $T = \{S_1, \ldots, S_B\}$ such that $\sum_{i \in W(T)} w_i$ is maximized.

Example. In Figure 1, we see an example contact network and two sampled cascades. If we are considering the MaxWelfareproblem, we assume $n_p = 2$ and $B = 2$, then we can test two pools with a maximum size of two. An example set of tests would be $T = \{u_1, u_2\}, \{u_3, u_5\}$. We can see that in the first cascade, the first pool tests positive and the second pool tests negative, while in the second cascade both pools test positive. Thus, our expected number of clearances is going to be $(2 + 0)/2 = 1$.

Note that we assume $n_p \cdot B \ll n$, the number of nodes in the graph, making us unable to test all potential patients within the graph.

4 SIGNIFICANCE OF NETWORK CORRELATIONS AND DISEASE MODEL

We show that information about correlations in disease states due to network correlations, and about the disease model have a significant impact on the performance of pool testing. Not considering these components, or uncertainty about them can have a significant impact on the performance.

Importance of network correlations. [10] assumes infection independence in maximizing the welfare of a given population. However, in specific spread settings, we make the following claim:

Theorem 4.1. *There exist instances where an optimal solution using the methodology in [10] ignoring network correlations provides welfare $o(n)$, while the optimum is $\Theta(n)$.*

Proof. To prove our result, we present an example (see Figure 2) to illustrate the effects of network structure on pool design. Let $G(V, E)$ be a graph where $V = X \uplus Y \uplus \{s\}$, where X induces a clique, Y induces an independent set and s is a *special node* that is adjacent to all nodes in Y and exactly one node v in X. All edges incident with s have a transmission probability p while the edges between pairs of nodes in X have transmission probability 1. There are no other edges in the graph. Also, only s has a non-zero probability of being a source.

The objective is to find the best pool S of maximum size $1 < n_p < |X|, |Y|$ (budget $B = 1$). Also, we will have a constraint that s cannot belong to the pool. In the above example, for any node in $X \cup Y$, the marginal probability of being infected is p. Under the assumption of independent infections as in [10], any n_p sized subset would be the best solution.

However, we observe that the network induces dependencies among nodes thus demanding a more non-trivial analysis that accounts for network structure. We consider three scenarios: (i) $S \subset X$, (ii) $S \subset Y$, and (iii) S has nodes from both sets. For maximizing welfare, we compute the probability of at least one node being infected in S conditioned on the fact that s is infected. Under Scenario (i), this probability for any S is p; any node in X will be infected if and only if s infects v. For Scenario (ii), it is $1 - (1 - p)^{n_p}$ for any S as every node in Y can be independently infected by s. For Scenario (iii), it is $1 - (1 - p)^{\ell+1}$ where $\ell = |S \cap Y| < n_p$. Any solution corresponding to Scenario (i) has the smallest probability, and is therefore the optimal solution.

Given that we select a pool from Scenario (i), we have an expected welfare of $(1 - p)n_p = O(|X|) = O(n)$ when taking into account network structure, while we have any n_p sized subset that doesn't include node s using the framework of [10]. Because any given subset is going to be equally likely, the expected welfare using this framework going to be $o(n)$ on this graph. □

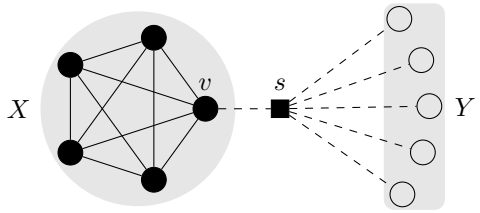

Figure 2: An example graph to illustrate the effect of networks on pool design. The solid edges correspond to transmission probability of 1 while the dashed edges correspond to a probability of p. Here, $|X| = |Y| = 5$.

Impact of Uncertainty in Disease Model

Lemma 4.2. *There exist instances where an error of $\theta(1/n)$ in the transmission probability can change the welfare $W(T)$ associated with a set of pooled tests T from $\Theta(n)$ to $o(n)$.*

Proof. Consider the graph $X \cup V$ where X and V are cliques of size $n/2$ and exactly one edge connects $x \in X$ to $v \in V$. Figure 3 shows this with $n = 10$. Assume we have an infection model $M = (\mathbf{p}, \text{SRC}, h)$ where $h > n$, but this infection model is inaccurate. Instead model $M' = (\mathbf{p}', \text{SRC}, h)$ with $0 < \mathbf{p}' - \mathbf{p} \leq 1/n$ would be accurate. Assume in both M and M', SRC $= \{v\}$, and we have $n_p < n/2$, $B = 1$, and $\tau = 2$.

Suppose, under assumed model M, we have an optimal set of tests $T^\star$. Based on the structure of the provided graph, $T^\star$ is going to consist of nodes in $X \setminus x$ with equal probability, and the expected welfare is going to be $n_p * (1 - p^2)^{n_p}$. However, if we have an error in our infection model, then we are going to have actual expected welfare $n_p * (1 - p'^2)^{n_p} \leq n_p * (1 - (p + \frac{1}{n})^2)^{n_p} = n_p * (1 - p^2 - 2/n - 1/n^2)^{n_p} = n_p * (\frac{n^2 - n^2p^2 - 2n - 1}{n^2})^{n_p}$. □

5 OUR APPROACH

5.1 The MaxWelfare problem

Let $H_i = (V, E_i)$ denote a cascade formed by sampling a subset of edges from G. We consider N cascades, where N will be specified in

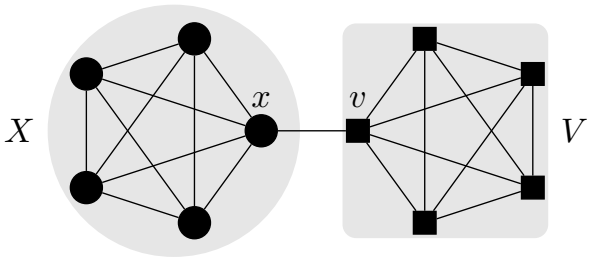

Figure 3: Example figure denoting how uncertainty in the disease model can affect the output expected welfare.

the analysis later. Let $C(v,i)$ denote the nodes in the connected component containing node v in H_i, and let $C(\text{SRC},i) = \cup_{v\in\text{SRC}} C(v,i)$; the nodes in $C(\text{SRC},i)$ will all be infected in the cascade H_i for a given set of infection sources $v \in V$ at some time step τ corresponding to the distance of a node from the source and will still test positive until time step $\tau + h$. In Figure 4, for example, we have $C(\text{SRC},1) = \{u_1,u_2,u_4\}$ and $C(\text{SRC},2) = \{u_2,u_3,u_4,u_5\}$. Let $F(v,i) = \{S \in \mathcal{S} : S \cap C(\text{SRC},i) = \emptyset\}$ be the set of possible pool tests which will clear node v in cascade i, meaning v and all potential nodes pooled with v are either (i): not in $C(\text{SRC},i)$ or (ii): are in $C(\text{SRC},i)$ but have entered the **Recovered** state more than h time steps back from when the test is administered. We denote the set of nodes that enter the **Recovered** state at time τ as $R(\tau)$. We define the set of nodes that can potentially be cleared in a cascade as $\mathbb{C}(i) = \{v \in (V \setminus C(\text{SRC},i)) \cup \bigcup_{j=h}^{\tau-1} R(\tau-j)\}$. For a pooled test set $T \in m_S$ administered at time τ, let $W(T,i) = |\{v \in (V \setminus C(\text{SRC},i)) \cup \bigcup_{j=h}^{\tau-1} R(\tau-j) : T \cap F(v,i) \neq \emptyset\}|$ denote the number of nodes which get cleared by the set T in cascade H_i, and let $W_{avg}(T) = \frac{1}{N}\sum_{i=1}^{N} W(T,i)$.

We use a linear programming (LP) relaxation plus rounding approach. Let $X(S)$ be an indicator for $S \in \mathcal{S}$, which is 1 if t_S is picked. Let $Y(v,i)$ be an indicator for each $v \in V, i = 1,\ldots,N$ which is 1 if node v is cleared in cascade H_i by some test. We have variables $x(S)$ and $y(v,i)$ corresponding to these indicators in the IP below.

$$\max \frac{1}{N}\sum_i \sum_{v\in\mathbb{C}(i))} w_v y(v,i) \tag{1}$$

$$\forall v \in \mathbb{C}(i) : y(v,i) \leq \sum_{S\in F(v,i)} x(S) \tag{2}$$

$$\sum_S x(S) \leq B \tag{3}$$

$$\forall S,v,i : x(S), y(v,i) \in \{0,1\} \tag{4}$$

Moving forward, we assume that each individual i has welfare $w_i = 1$. We therefore drop this coefficient for the rest of the paper and speak equivalently about the expected number of clearances and the expected welfare of the population.

Lemma 5.1 proves that the above IP will solve the MaxWelfare problem.

Lemma 5.1. *The above Integer Program is valid, and if we have solutions $x^\star, y^\star$ to the above program, then $\frac{1}{N}\sum_i \sum_{v\in\mathbb{C}<(i)} y^\star(v,i) = W_{avg}(T^\star)$ for an optimal set of pools $T^\star$.*

Proof. Consider the optimal set $T^\star$, and define $x^\star, y^\star$ as follows: $\forall S \in T^\star, x^\star(S) = 1$ and $S' \notin T^\star, x^\star(S') = 0$. Constraint 2 will hold as we have $|T^\star| = \sum_S x(S) \leq B$. By the definition of $F(v,i)$, we have $y(v,i) = 1$ if it is cleared by any $S \in T^\star$, which by the first constraint will hold if $x(S) = 1$. Thus, the first constraint also holds for the optimal. Thus, our program is feasible for $T^\star$, and by the definition of $W_{avg}(T^\star)$ is exactly equal to the definition of the objective value. Thus, the integer program is valid. □

However, IPs are NP-Complete in general. Therefore, we present Algorithm 1, which utilizes LP relaxation and randomized rounding as well as Sample Average Approximation (SAA). We go on to show that Algorithm 1 still provides good theoretical and practical performance.

Algorithm 1 RoundPool

Inputs: $G = (V,E)$, B, potential pools $\boldsymbol{S}$, Infection Model M
Outputs: Set of Pools $\boldsymbol{T}$

1: Create cascades $H_1 \ldots H_N$ by sampling from M
2: Solve the LP obtained by relaxing constraint 4 to $x(S), y(v,i) \in [0,1] \forall S,v,i$.
3: Let $x^\star, y^\star$ denote the solutions to step 1.
4: **for** $S \in \boldsymbol{S}$ **do**
5: Add S to $\boldsymbol{T}$ with probability $x^\star(S)$
6: **end for**
7: return $\boldsymbol{T}$

Lemma 5.2. *The solution $T = \{S_1,\ldots,S_K\}$ computed by the above algorithm satisfies: (1)$\mathbb{E}[|T|] \leq B$, and (2)$E[W_{avg}(T)] \geq \frac{1}{2}W_{avg}(T^\star)$.*

Proof. For the first statement, we have $Pr[S \in T] = x(S)$. Because our randomized rounding is independent, we have $E[|T|] = \sum_S Pr[S \in T] = \sum_S x(S) \leq B$ by equation 3.

For the second part, we have $W_{avg}(T) = \frac{1}{N}\sum_i W(T,i)$. As before, define $Y(v,i)$ be the indicator variable that node v was cleared by some test $S \in T$ in cascade i, so $\sum_v Y(v,i) = W(T,i)$. We therefore have the following:

$$\begin{aligned}
\Pr[Y(v,i)=1] &= 1 - \prod_{S\in F(v,i)} (1 - x^*(S)) \\
&\geq 1 - \prod_{S\in F(v,i)} e^{-x^*(S)} \\
&= 1 - e^{-\sum_{S\in F(v,i)} x^*(S)} \\
&\geq 1 - e^{-y^*(v,i)} \\
&\geq y^*(v,i)/2,
\end{aligned}$$

The first equality is defined by our definition of $F(v,i)$ and the results from our linear program. The first inequality follows because $1 - z \leq e^{-z}$ for $z \in [0,1]$. The second inequality follows because $\sum_{S\in F(v,i)} x^*(S) \geq y^*(v,i)$, The third inequality follows because $1 - z/2 \geq e^{-z}$ for $z \in [0,1]$. By linearity of expectation, we have $\mathbb{E}[W_{avg}(T)] = \frac{1}{N}\sum_{v,i} Y(v,i)$ and $\frac{1}{2}\sum_{v,i} y^\star(v,i) \geq W(T^\star)$ due to LPs being an upper bound on IPs, it therefore directly follows that $\mathbb{E}[W_{avg}(T)] \geq \frac{1}{2}W_{avg}(T^\star)$. □

Lemma 5.2 shows that Algorithm 1 gives a solution whose expected size is less than or equal to the budget and has $\mathbb{E}[W_{avg}(T)]$ close to $W_{avg}(T^\star)$ with this expectation being over the stochasticity of the algorithm. However, maximizing $W_{avg}(T)$ is not directly equal to maximizing $\mathbb{E}[W(T)]$ directly. We will go on to show, however, that given N that is sufficiently large, that $W_{avg}(T)$ provides a reasonable estimate of $\mathbb{E}[W(T)]$.

Theorem 5.3. *(Theorem 1.1 of [8]) Let $Z = \sum_{i=1}^n Z_i$, where Z_i are independently distributed random variables in $[0,1]$. Then, for any $\epsilon \in (0,1)$, we have $\Pr[Z \notin [(1-\epsilon)E[Z],(1+\epsilon)E[Z]]] \leq 2exp(-\epsilon^2 E[Z]/3)$. Also, for any $t > 2eE[Z]$, $\Pr[Z > t] \leq 2^{-t}$.*

This theorem shows that a bound can be placed on the sum of random variables with high probability. With some transformations and a specific choice of ϵ, we can produce bounds on the output of our algorithm with high probability.

Lemma 5.4. *Let $N \geq \frac{3}{\epsilon^2} n^2 \log n$ for $\epsilon \in (0,1)$ and ${m_S}^B \subseteq m_S$ consist of all possible sets of pools $\mathbb{T}$ such that $W(\mathbb{T}) \geq 1$, that is the set of pool tests that will clear at least one individual. Then* $\Pr[$*there exists $T \in {m_S}^B$ such that $W_{avg}(T) \notin [(1-\epsilon)W(T), (1+\epsilon)W(T)] \leq 2/n^2$.*

Proof. For any fixed $T \in m_S$, we have $N\frac{W_{avg}(T)}{n} = \sum_{i=1}^{N} \frac{W(T,i)}{n}$. Let $Z_i = \frac{W(T,i)}{n}$, and let $Z = \sum_{i=1}^{N} Z_i$. Note that by the definition of $W(T,i)$, we have $Z_i \in [0,1]$. Also, $\mathbb{E}[W(T,i)] = W(T)$ (where the expectation is over random cascades H_i), which therefore implies $\mathbb{E}[NW_{avg}(T)] = NW(T)$. Applying Theorem 5.3 to $Z = N\frac{W_{avg}(T)}{n}$, we have:

$$\begin{aligned} &\Pr\left[W_{avg}(T) \notin [(1-\epsilon)W(T), (1+\epsilon)W(T)]\right] \\ &= \Pr\left[N\frac{W_{avg}(T)}{n} \notin [(1-\epsilon)N\frac{W(T)}{n}, (1+\epsilon)N\frac{W(T)}{n}]\right] \quad (5) \\ &\leq 2\exp(-\epsilon^2 NE[Z]/(3n)) \\ &= 2\exp(-\epsilon^2 NW(T)/(3n)) \\ &\leq \exp(-\epsilon^2 N/(3n)) \end{aligned}$$

If we have $N \geq \frac{3}{\epsilon^2} n^2 \log n$, then plugging in we have:

$$2\exp(-\epsilon^2 NW(T)/(3n)) \leq 2\exp(-n\log n)$$

Next we have $|{m_S}^B| \leq (n^{n_P})^B$, so by a union bound, we have $\Pr[$there exists $T \in {m_S}^B$ such that $W_{avg}(T) \notin [(1-\epsilon)W(T), (1+\epsilon)W(T)] \leq 2\exp(-n\log n)n^{n_P B}$.

When we assume $n_p B$ is smaller than $n-2$, which fits our assumption that $n_p B \ll n$ stated in the problem definitions, we have the following:

$2\exp(-n\log n)n^{n_P B} \leq 2/n^2$

□

Here we show that for any problem instance where it is possible to clear at least one individual in any given infection instance for a set of pools, that $W_{avg}(T)$ is going to be close to $W(T)$ with high probability given that the average is performed over sufficiently large number of cascades N. This holds true even for $T^\star$. Therefore, we present Theorem 5.5, which combines Lemmas 5.2 and 5.4 to show that the solution provided by Algorithm 1 will be close to the optimal solution given sufficiently large N.

Theorem 5.5. *Let $\mathbb{E}[W(T^\star)]$ be the optimal number of clearances for a given graph G, maximum pool size n_p, and budget B and assume $B \geq \frac{3}{\epsilon^2}\log\frac{n^2}{2}$. Then we will have:*

(1) $|T| \leq (1+\epsilon)B$ *and*
(2) $\mathbb{E}[W(T)] \geq (1-\epsilon)^2\mathbb{E}[W(T^\star)]/2]$

hold simultaneously with probability $1-\frac{3}{n^2}$.

Proof. Let $Z_S = x^\star(S)$, the probability that we include set S into our set of pools T and define $Z = \sum_S Z_S$. We have $\mathbb{E}[|T|] = \sum_S x^\star(S) \leq B$ from Lemma 5.2. By Theorem 5.3, for any $\epsilon \in (0,1)$, we have $Pr[\sum_S Z_s \geq (1+\epsilon)B] \leq 2exp(-\epsilon^2 B/3)$. Assuming that we have $B \geq \frac{3}{\epsilon^2}\log\frac{n^2}{2}$, we can then substitute in to have $Pr[\sum_S Z_s \geq (1+\epsilon)B] \leq \frac{1}{n^2}$.

For the second part, we have $\mathbb{E}[W(T)] = W_{avg}(T)$. Therefore, by Lemma 5.4, we have $Pr[W(T) \leq (1-\epsilon)W_{avg}(T)] \leq \frac{1}{n^2}$. By Theorem 5.3 again, we are able to say that $Pr[W_{avg}(T) \leq (1-\epsilon)\mathbb{E}[W_{avg}(T)]] \leq \frac{1}{n^2}$, and then using Lemma 5.2, we have $\mathbb{E}[W_{avg}(T) \geq \frac{1}{2}W_{avg}(T^\star)$. Therefore, we have $Pr[W_{avg}(T) \leq (1-\epsilon)\frac{1}{2}W_{avg}(T^\star)]] \leq \frac{1}{n^2}$. Because $W_{avg}(T^\star) = \mathbb{E}[W(T^\star)]$ by definition, combining everything, we have $Pr[W(T) \leq (1-\epsilon)^2\mathbb{E}[W(T^\star)]/2] \leq \frac{2}{n^2}$.

□

Non-overlapping pools. The presented IP allows for overlapping sets, meaning a single node u can be in two different pools $S, S' \in T$. By changing (2) to $y(v,i) = \sum_{S \in F(v,i)} x(S)$, this would be disallowed. [10] analyze the theoretical benefit of allowing overlapping testing and find that the theoretical benefits were outweighed by the practical constraints within the lab.

Lemma 5.6. *Algorithm RoundPool will produce a set of pools in $O((n * N * |\mathbf{S}|)^{2.5})$.*

Proof. The above algorithm can be broken into two main steps: the linear program and the rounding step. The linear program is going to have $O(n*N)$ constraints and exactly $n*N+|\mathbf{S}|$ variables. Proven worst case complexities are shown to be a polynomial in the number of variables and constraints in a worst case scenario, which means that our runtime will be less than $O((n*N*|\mathbf{S}|)^{2.5}$. The second step of our algorithm is performed in linear time $O(|S|)$, and thus the linear scaling of the rounding step is dominated by the polynomial scaling of the LP. □

5.2 Speeding up RoundPool

RoundPool scales as $(Nn)^{2.5} = n^{7.5}$, which is not feasible for large instances, even using state-of-the-art LP solvers. We use the combinatorial approach of [34] for solving the LP, which allows us to significantly improve the scaling. Define $z(v,i) = 1 - y(v,i)$. We additionally define an additional variable λ. The below IP solves the equivalent problem after our definition of λ and the $z(\cdot)$ variables:

$$\begin{aligned} &\min \lambda \\ \forall v \in \mathbb{C}(i) : z(v,i) + \sum_{S \in F(v,i)} x(S) &\geq 1 \\ \frac{1}{N}\sum_i \sum_{v \in \mathbb{C}(i)} z(v,i) &\leq \lambda \\ \sum_S x(S) &\leq B \\ \forall S, v, i : x(S), z(v,i) &\in \{0,1\} \end{aligned}$$

If we consider λ as a fixed constant, this problem translates to finding a feasible solution given the mixed packing/covering constraints. [34] presents an $(1+\epsilon)$−approximation algorithm for

feasible mixed packing/covering problems, where covering constraints are met and packing constraints are violated by a factor of $O(\epsilon)$. Additionally, [33] provides a subroutine to check feasibility of the program. We present algorithm 3, which calls the algorithm MixedApproximation adapted from [34] as a subroutine.

Algorithm 2 MixedApproximation

Inputs: Covering matrix C, Packing Matrix P, ϵ
Outputs: Feasible Solution $\boldsymbol{x}$

1: Define $partial_j(M,x) = \sum_i M_{ij} e^{(Mx)_i} / \sum_i e^{(Mx)_i}$
2: $\lambda_0 \leftarrow |P|/|C|$
3: $U \leftarrow \log(|C|+|P|)/\epsilon^2$
4: $x \leftarrow \mathbf{0}$
5: $\hat{C}_i \leftarrow C_i x$
6: Define $\hat{c}_i \leftarrow (1-\epsilon)^{\hat{C}_i}$ if $\hat{C}_i \le U$ else $\hat{c}_i = 0$
7: $\hat{P}_i \leftarrow P_i x$
8: Define $\hat{p}_i \leftarrow (1+\epsilon)^{\hat{P}_i}$
9: **while** True **do**
10: **for** $j \in |x|$ **do**
11: $\hat{\lambda}_j \leftarrow P_j^T \hat{p} / C_j^T \hat{c}$
12: **while** $\hat{\lambda}_j \le (1+\epsilon)^2 \lambda_0/(1-\epsilon)$ **do**
13: Choose z such that $\max\{\max_i P_{ij} z, \max_{i:C_i x \le U} C_{ij} z\} = 1/2$
14: $x_j \leftarrow x_j + z$
15: $\hat{P}_i \leftarrow \hat{P}_i + z * P_{ij}$ for all $P_{ij} > 0$
16: $\hat{C}_i \leftarrow \hat{C}_i + z * C_{ij}$ for all $C_{ij} > 0$
17: **if** $\min_i \hat{C}_i \ge U$ **then**
18: return x/U
19: **end if**
20: Update $\hat{p}_i, \hat{c}_i$ as defined, recalculate $\lambda_j \leftarrow P_j^T \hat{p} / C_j^T \hat{c}$
21: **end while**
22: $ratio_j(x) \leftarrow partial_j(P,x)/partial_j(C,-x)$
23: **end for**
24: **if** $\min_j ratio_j(x) > 1$ **then**
25: return $infeasible$
26: **end if**
27: $\lambda_0 \leftarrow (1+\epsilon)\lambda_0$
28: **end while**

Algorithm 2 describes a sequential algorithm for finding a feasible solution to a mixed packing/covering problem that violates the normalized packing constraints by a factor of $1 + O(\epsilon)$. Within the inner-while loop starting on line 12, the current index is continuously iterated on by a small amount until the covering constraints are met or packing constraints will be violated by too much. The size of the iteration ensures that the packing constraints will not be violated by more than a factor of $O(\epsilon)$. We additionally check the feasibility of the problem, as algorithm 3 is searching for the minimum value of λ that is feasible, meaning that algorithm terminates after infeasibility of the subroutine is reached. After each index, we scale up by a factor of $(1+\epsilon)$ until all of the covering constraints are met, ensuring the allowable violation stays within $O(\epsilon)$.

Lemma 5.7. *Algorithm 3 runs in* $O((\log(|V|*N)(|V|*N*|S|)/\epsilon^2))$.

Proof. Algorithm 3 will run through the while loop a maximum of $O(\log \lambda)$ times, and because we set $\lambda = n * N$, we have the main algorithm running through it's loop $O(\log(n*N))$ times. Each iteration of the loop is dominated by the running time of the subroutine of 2, which runs in running time $O(\hat{L}/\epsilon^2)$ where $\hat{L}$ refers to the number of non-zero entries in the constraint matrix [34].

By our LPs definition, we will have $O(n*N) + |\mathbf{S}|$ non-zeros in our packing constraints and each covering constraint will have $1 + O(\mathbf{S})$ non-zero entries. Because we have $O(n*N)$ possible covering constraints, we will have $O(n*N*|\mathbf{S}|)$ non-zeros across all covering constraints. Thus, the runtime for one iteration is going to be a function of $O((n*N*|\mathbf{S}|+|V|*N)/\epsilon^2 = O((n*N*|\mathbf{S}|)/\epsilon^2)$ Thus, the overall runtime of this algorithm is going to be $O(\log(n*N)(n*N*|\mathbf{S}|)/\epsilon^2)$ □

Algorithm 3 PoolApproximation

Inputs: $G = (V,E)$, B, potential pools $\boldsymbol{S}$, Infection Model M, ϵ
Outputs: Set of Pools $\boldsymbol{T}$

1: Create cascades $H_1 \dots H_N$ by sampling from M.
2: $\lambda \leftarrow |V| * N$
3: Define C, P, p based on Linear Program for covering matrix C, Packing matrix P, and packing constraint p.
4: $feasability \leftarrow$ **True**
5: $\boldsymbol{X} \leftarrow \mathbf{0}$
6: **while** $feasibility$ **do**
7: $p = [B, \lambda]^T$
8: **for** $i \in P$ **do**
9: $P'_i = P_i / p_i$
10: **end for**
11: $z \leftarrow$ MixedApproximation(C, P', ϵ)
12: **if** $z = infeasible$ **then**
13: $feasibility \leftarrow$ **False**
14: **else**
15: $\lambda \leftarrow \lambda/2$
16: $\boldsymbol{X}(S) \leftarrow z(x(S)) \forall S \in \boldsymbol{S}$
17: **end if**
18: **end while**
19: **for** $S \in \boldsymbol{S}$ **do**
20: Add S to $\boldsymbol{T}$ with probability $\boldsymbol{X}(S)$
21: **end for**

6 EXPERIMENTS

Using our algorithm, we study the following questions:

- ***Efficiency of Sampled Cascades***: In 5.4, we say the number of sampled cascades $N = \Omega(n^2 \log n)$, but does our expectation asymptotically approach a solution in fewer samples?
- ***Effects of increasing pool size:*** How does increasing the size of the pools considered affect the expected welfare, and are there diminishing returns for increasing pool size?
- ***Number of potential pools considered***: Because the number of potential sets considered scales by a factor of $O(n^{n_p})$, it quickly becomes inefficient to examine all potential combinations of pools. How does reducing our search space from all potential pools affect our results?

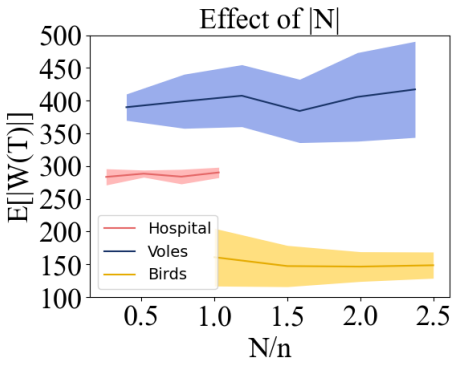

Figure 4: Graph illustrating the effect that increasing the number of sampled cascades has on expected welfare, with no significant gains in expected welfare occurring due to increasing the number of samples beyond a linear factor.

- ***Importance of Correlations in Practice***: We offer a toy example on the importance of considering infection correlations in the optimal selection of pools, but how important is this in real world graphs on realistic infection settings?
- ***Effects of Assumptions of the Infection Model***: How do errors in the assumed infection model M affect the true expected welfare, given that the disease model M' is more accurate?

For ease of computation, we assume $\tau = h = n$, meaning we are testing at the end of an epidemic. This will provide an absolute lower bound on expected welfare. We describe the data sets that we run these experiments on below:

Graph Name	Number of nodes	Number of edges
Virginia Academic Hospital Contacts	3885	61537
Aves Wildbird Network	202	4574
Vole Trapping Colocation	1263	3380

Table 1: Data sets used for experiments. The hospital network consists of patient and healthcare provider nodes, where edges denote a colocation of nodes. The graph spans four weeks across the hospital. The Aves Wildbird Network and Vole Trapping Colocation are publicly available networks sourced in [26], while the Virginia Academic Hospital Contacts can be found in our public GitHub repository

6.1 Efficiency of Sampled Cascades

On the **Virginia Academic Hospital Contacts** network, we examine the effect that the number of sampled cascades has on the expected welfare. We consider 20000 randomly selected subsets of nodes with $n_p = 4$ and $B = 100$, running each trial 5 times, averaging the the expected welfare. The results can be seen below:

We can see here that the number of sampled cascades trialed does not make a significant difference in the calculated expected welfare of our algorithm. Within our paper, we show that $N \geq O(n^2 \log(n))$.

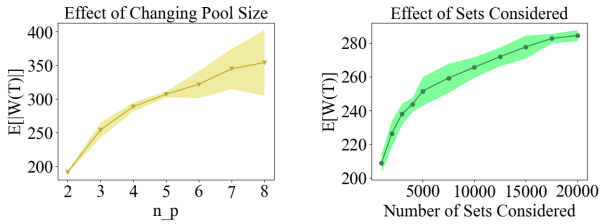

Figure 5: (Right) Graph showing the impact of changing n_p on the expected welfare. Higher values of n_p have sub-linear returns, as increasing the number of samples within a pool increases the probability that at least one of the samples will test positive. (Left) Plot showing the effect that the number of randomly selected sets has on the expected welfare. We can see that increasing the number of possible sets to consider increases the expected welfare, although we have diminishing returns

However, these experimental results show that only $N \geq O(n)$ are required in practice. We note that for less dense graphs, there is more uncertainty within the number of samples affecting the outcome, with the Virginia Academic Hospital data set having the most consistent expected welfare as the largest and most dense graph. The Vole graph is the least dense graph and we can see that we have a wide standard deviation.

The number of sampled cascades directly affects the size of the linear program, with the number of constraints being a factor of $O(Nn)$. When running these experiments, we were unable to solve the algorithm with more than 4000 sampled cascades in a computationally tractable way. Thus, we utilize the approximation algorithm presented above adopted from [34] in order to solve larger instances of our program. These results are presented in the supplementary material.

6.2 Effects of Changing Size of Pools n_p

[6] calculates the optimal size of pools based on the prevalence of a disease within a given population. In a true epidemic setting, however, this prevalence is often unknown, and thus an optimal value of n_p cannot be selected beforehand. Therefore, we analyze the effects that increasing n_p will have on the expected welfare of a fixed epidemic setting. We test on the Virginia Academic Hospital data set with 2000 fixed cascades and 20000 potential pools considered. The results are shown in Figure 5 (Right).

6.3 Number of Potential Pools Considered

On the **Virginia Academic Hospital Contacts** network using 1000 sampled infection cascades with an average of 1289.958 infections per cascade, we consider varying numbers of randomly selected subsets of nodes with $n_p = 4$ and $B = 100$. We run these trials across 5 randomized set selections and average the expected welfare output by our algorithm. The results can be summarized in Figure 5 (Left).

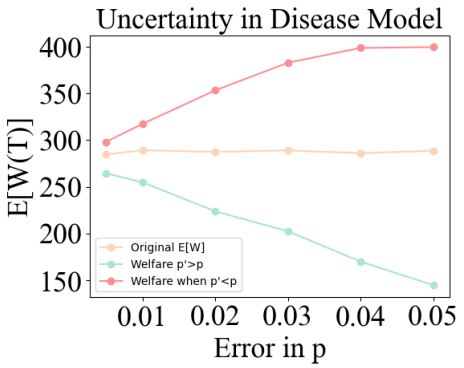

Figure 6: Plot displaying the impact of errors in estimated transmission probability p affect the overall expected welfare after rounding.

We can see from these trials that expanding the number of randomly selected subsets increases the expected welfare our algorithm's output, with these increases falling outside of the 95% confidence intervals in many instances. With this given experimental instance giving a maximum of approximately $9.77e12$ possible subsets, not all of these can be feasibly be considered. We see relatively diminishing returns to the expected welfare, however, as the number of considered subsets increases.

6.4 Importance of Considering Network Correlation

In [10], the authors consider selecting pools to maximize the expected welfare assuming independent infection status. We analyze how important the graph structure is compared to selecting random subsets, which equates to uncorrelated infection status if all individuals' utilitis and initial probabilities of infection are equivalent.

On the **Virginia Academic Hospital Contacts** data set, a random selection of subsets produces an average expected welfare of **183.511** for $n_p = 4, B = 100$. This stands in contrast to our expected welfare of **288.891** for the same set of cascades tested against with correlations from our experiments on 6.2 (where 20000 potential pools are considered in both settings).

6.5 Effects of Uncertainty within the Disease Model

A major assumption of our work is the assumption of an accurate disease model M that describes how a disease spreads across a network. On the **Virginia Academic Hospital Contacts** network, we analyze how errors both upward and downward in our transmission probability p affect the expected welfare calculated by our algorithm. Our initial probability of infection is $p = 0.055$ produces an average number of infections of 1289.958, as stated above. We then compute cascades with $p' < p$ and $p' > p$. These results can be seen in Figure 6.

We see a nearly linear decrease in expected welfare as p' increases above our assumption p. At the highest level, a value of $p' = .105$ produces approximately 2196 infections across the network, and our expected welfare for the selected pools decreases by a factor of nearly 50%.

On the other side, we see that the pools selected when $p' < p$ perform very well, reaching the upper bound of $n_p * b$ if infectiousness is low enough. Thus, when using our methodology for selecting pools, we should be careful not to underestimate p within our assumed disease model.

7 CONCLUSION

In this paper, we present a methodology for selecting subsets of individuals to test in a pooled manner for epidemic surveillance and clearance. Our bicritera algorithms take into account network structures within communities, which to our knowledge, has not been taken into account previously. Our algorithms show novel performance on real world data sets while requiring less computation than the shown worst case bounds.

Our work does make a number of simplifying assumptions that may not hold in real life. For example, we assume perfect accuracy of pool tests regardless of the size of pool n_p. In reality, works such as [32] have shown that pooled tests become less accurate as n_p increases. We additionally assume fully known networks for spread. Future works in this field could study the optimization problem when these assumptions don't hold, closer mirroring real world settings.

ACKNOWLEDGMENTS

This research is partially supported by NSF grants CCF-1918656 and CNS-2317193, and DTRA award HDTRA1-24-R-0028, Cooperative Agreement number 6NU50CK000555-03-01 from the Centers for Disease Control and Prevention (CDC) and DCLS, Network Models of Food Systems and their Application to Invasive Species Spread, grant no. 2019-67021-29933 from the USDA National Institute of Food and Agriculture, Agricultural AI for Transforming Workforce and Decision Support (AgAID) grant no. 2021-67021-35344 from the USDA National Institute of Food and Agriculture.

REFERENCES

[1] Bijaya Adhikari, Bryan Lewis, Anil Vullikanti, Jose Mauricio Jimenez, and B. Aditya Prakash. 2019. Fast and Near-Optimal Monitoring for Healthcare Acquired Infection Outbreaks. *PLoS Computational Biology* (2019).
[2] Matthew Aldridge, Oliver Johnson, Jonathan Scarlett, et al. 2019. Group testing: an information theory perspective. *Foundations and Trends® in Communications and Information Theory* 15, 3-4 (2019), 196–392.
[3] Batuhan Arasli and Sennur Ulukus. 2023. Group testing with a graph infection spread model. *Information* 14, 1 (2023), 48.
[4] Nicholas A Christakis and James H Fowler. 2010. Social network sensors for early detection of contagious outbreaks. *PloS one* 5, 9 (2010), e12948.
[5] Robert Dorfman. 1943. The detection of defective members of large populations. *The Annals of mathematical statistics* 14, 4 (1943), 436–440.
[6] Robert Dorfman. 1943. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics* 14, 4 (1943), 436–440. http://www.jstor.org/stable/2235930
[7] Dingzhu Du, Frank K Hwang, and Frank Hwang. 2000. *Combinatorial group testing and its applications*. Vol. 12. World Scientific.
[8] Devdatt P. Dubhashi and Alessandro Panconesi. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press. I–XIV, 1–196 pages.
[9] Jens Niklas Eberhardt, Nikolas Peter Breuckmann, and Christiane Sigrid Eberhardt. 2020. Multi-stage group testing improves efficiency of large-scale COVID-19 screening. *Journal of Clinical Virology* 128 (2020), 104382.
[10] Simon Finster, Michelle González Amador, Edwin Lock, Francisco Marmolejo-Cossío, Evi Micha, and Ariel D. Procaccia. 2024. Welfare-Maximizing Pooled

Testing. *SIGecom Exch.* 22, 1 (Oct. 2024), 66–73. https://doi.org/10.1145/3699824.3699829

[11] Simon Finster, Michelle González Amador, Edwin Lock, Francisco Marmolejo-Cossío, Evi Micha, and Ariel D. Procaccia. 2024. Welfare-Maximizing Pooled Testing. *SIGecom Exch.* 22, 1 (Oct. 2024), 66–73. https://doi.org/10.1145/3699824.3699829

[12] Jack Heavey, Jiaming Cui, Chen Chen, B Aditya Prakash, and Anil Vullikanti. 2022. Provable Sensor Sets for Epidemic Detection over Networks with Minimum Delay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10202–10209.

[13] Penny Hitchcock, Allison Chamberlain, Megan Van Wagoner, Thomas V Inglesby, and Tara O'Toole. 2007. Challenges to global surveillance and response to infectious disease outbreaks of international importance. *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 5, 3 (2007), 206–227.

[14] Hsun-Ping Hsieh, Shou-De Lin, and Yu Zheng. 2015. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 437–446. https://doi.org/10.1145/2783258.2783344

[15] Jakob Jonnerby, Philip Lazos, Edwin Lock, Francisco Marmolejo-Cossío, Christopher Bronk Ramsey, and Divya Sridhar. 2020b. Test and Contain: A Resource-Optimal Testing Strategy for COVID-19. In *AI for Social Good Workshop*.

[16] Jakob Jonnerby, Philip Lazos, Edwin Lock, Francisco Marmolejo-Cossío, C. Bronk Ramsey, Meghana Shukla, and Divya Sridhar. 2020a. Maximising the Benefits of an Acutely Limited Number of COVID-19 Tests. arXiv:2004.13650 [q-bio.PE]

[17] Matthew M Kavanagh, Ngozi A Erondu, Oyewale Tomori, Victor J Dzau, Emelda A Okiro, Allan Maleche, Ifeyinwa C Aniebo, Umunya Rugege, Charles B Holmes, and Lawrence O Gostin. 2020. Access to lifesaving medical resources for African countries: COVID-19 testing and response, ethics, and politics. *The Lancet* 395, 10238 (2020), 1735–1738.

[18] Andreas Krause, H. McMahan, Carlos Guestrin, and Anupam Gupta. 2008. Robust Submodular Observation Selection. *Journal of Machine Learning Research* 9 (12 2008), 2761–2801.

[19] Daniel B Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M Burke, James A Hay, Milind Tambe, Michael J Mina, and Roy Parker. 2021. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science advances* 7, 1 (2021), eabd5393.

[20] Brice Leclère, David L Buckeridge, Pierre-Yves Boëlle, Pascal Astagneau, and Didier Lepelletier. 2017. Automated detection of hospital outbreaks: A systematic review of methods. *PloS one* 12, 4 (April 2017), e0176438. https://doi.org/10.1371/journal.pone.0176438

[21] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 420–429.

[22] Edwin Lock, Francisco Javier Marmolejo-Cossío, Jakob Jonnerby, Ninad Rajgopal, Héctor Alonso Guzmán-Gutiérrez, Luis Alejandro Benavides-Vázquez, José Roberto Tello-Ayala, and Philip Lazos. 2021. Optimal Testing and Containment Strategies for Universities in Mexico amid COVID-19. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (–, NY, USA) *(EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. https://doi.org/10.1145/3465416.3483300

[23] Michael J Mina and Kristian G Andersen. 2021. COVID-19 testing: One size does not fit all. *Science* 371, 6525 (2021), 126–127.

[24] Leon Mutesa, Pacifique Ndishimye, Yvan Butera, Jacob Souopgui, Annette Uwineza, Robert Rutayisire, Ella Larissa Ndoricimpaye, Emile Musoni, Nadine Rujeni, Thierry Nyatanyi, et al. 2021. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* 589, 7841 (2021), 276–280.

[25] Laura C Rosella, Ajay Agrawal, Joshua Gans, Avi Goldfarb, Sonia Sennik, and Janice Stein. 2022. Large-scale implementation of rapid antigen testing system for COVID-19 in workplaces. *Science Advances* 8, 8 (2022), eabm3608.

[26] Ryan A. Rossi and Nesreen K. Ahmed. 2015. The Network Data Repository with Interactive Graph Analytics and Visualization. In *AAAI*. https://networkrepository.com

[27] Daniel K Sewell. 2022. Leveraging network structure to improve pooled testing efficiency. *Journal of the Royal Statistical Society Series C: Applied Statistics* 71, 5 (2022), 1648–1662.

[28] Huijuan Shao, KSM Hossain, Hao Wu, Maleq Khan, Anil Vullikanti, B Aditya Prakash, Madhav Marathe, and Naren Ramakrishnan. 2018. Forecasting the Flu: designing social network sensors for epidemics. *SIGKDD epiDAMIK Workshop* (2018).

[29] Robby Sikka, Anne L Wyllie, Prem Premsrirut, and Ethan M Berke. 2022. COVID testing in the workplace: return to work testing in an occupational cohort. *medRxiv* (2022), 2022–02.

[30] Haley Stone, Jing Du, Yang Yang, Ashna Desai, Rebecca Dawson, Hao Xue, David Heslop, Matthew Scotch, Andreas Züfle, C. Raina MacIntyre, and Flora Salim. 2026. From Ecological Connectivity to Outbreak Risk: A Heterogeneous Graph Network for Epidemiological Reasoning under Sparse Spatiotemporal Data. arXiv:2601.09738 [q-bio.PE] https://arxiv.org/abs/2601.09738

[31] Angela Felicia Sunjaya and Anthony Paulo Sunjaya. 2020. Pooled testing for expanding COVID-19 mass surveillance. *Disaster Medicine and Public Health Preparedness* 14, 3 (2020), e42–e43.

[32] Idan Yelin, Noga Aharony, Einat Shaer Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagham Gandali, Omer Shkedi, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, and Roy Kishony. 2020. Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools. *Clinical Infectious Diseases* 71, 16 (05 2020), 2073–2078. https://doi.org/10.1093/cid/ciaa531 arXiv:https://academic.oup.com/cid/article-pdf/71/16/2073/34393030/ciaa531.pdf

[33] N.E. Young. 2001. Sequential and parallel algorithms for mixed packing and covering. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE. https://doi.org/10.1109/sfcs.2001.959930

[34] Neal E. Young. 2014. Nearly Linear-Time Approximation Schemes for Mixed Packing/Covering and Facility-Location Linear Programs. *CoRR* abs/1407.3015 (2014). arXiv:1407.3015 http://arxiv.org/abs/1407.3015