Received: 6 May 2021 | Accepted: 11 August 2022

DOI: 10.1111/rssc.12594

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series C Applied Statistics

Leveraging network structure to improve pooled testing efficiency

Daniel K. Sewell

Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA

Correspondence
Daniel K. Sewell, Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA.
Email: daniel-sewell@uiowa.edu

Funding information
Centers for Disease Control and Prevention, Grant/Award Numbers: 1 U01 CK000594-01-00, 5 U01 CK000531-02

Abstract

Screening is a powerful tool for infection control, allowing for infectious individuals, whether they be symptomatic or asymptomatic, to be identified and isolated. The resource burden of regular and comprehensive screening can often be prohibitive, however. One such measure to address this is pooled testing, whereby groups of individuals are each given a composite test; should a group receive a positive diagnostic test result, those comprising the group are then tested individually. Infectious disease is spread through a transmission network, and this paper shows how assigning individuals to pools based on this underlying network can improve the efficiency of the pooled testing strategy, thereby reducing the resource burden. We designed a simulated annealing algorithm to improve the pooled testing efficiency as measured by the ratio of the expected number of correct classifications to the expected number of tests performed. We then evaluated our approach using an agent-based model designed to simulate the spread of SARS-CoV-2 in a school setting. Our results suggest that our approach can decrease the number of tests required to regularly screen the student body, and that these reductions are quite robust to assigning pools based on partially observed or noisy versions of the network.

KEYWORDS

epidemiology, group testing, infectious disease, optimisation, transmission networks

1 | INTRODUCTION

Screening and subsequent isolation of identified infectious individuals is a powerful tool in infectious disease control in a given population. For example, in a modelling study Paltiel et al. (2020) showed that screening every other day on a small college campus would be necessary to control the spread of SARS-CoV-2.

Frequent screening is especially important in the context of a disease with high degree of 'silent spreading,' that is, transmission occurring from pre-symptomatic or asymptomatic infectives, as is the case with Coronavirus Disease 2019 (COVID-19). In certain populations, individuals with COVID-19 were found to be asymptomatic at rates from 50% to nearly 80% (Denny et al., 2020; Oran & Topol, 2020; Sutton et al., 2020). One study estimated that over half of all transmission events were through silent spreading, and hence even comprehensive and immediate isolation of symptomatic cases is insufficient to achieve infection control (Moghadas et al., 2020). Regular comprehensive screening can lead to identifying all infectious individuals regardless of the presentation of symptoms, thereby pre-empting potential transmission events by isolating infectives.

In addition, relying on contact tracing and quarantining can be less effective when contacts are difficult to trace. For example, in a 2017 mumps outbreak at Penn State University, only around 17% of reported contacts were reached and verified to have followed quarantining protocol (Bharti et al., 2020); the true percentage of contacts effectively quarantined was undoubtedly smaller, as this percentage does not account for those contacts unreported by the index case.

Screening, however, requires a large amount of testing and is therefore very often cost prohibitive. To illustrate this fact, during the fall semester of 2020, in the midst of the global COVID-19 pandemic, only 6% of small colleges and universities routinely tested their students for SARS-CoV-2 (Nadworny & McMinn, 2020). Further compounding the resources required to effectively use screening for infection control is the finding that the frequency of testing is critical and even more important than, for example, the diagnostic test's sensitivity (Larremore et al., 2021).

One method of reducing the resource burden for frequent and comprehensive screening is pooled testing. Dorfman (1943) first formalised the strategy of taking batches of samples and pooling them in order to perform a single joint test. Should a pooled test result in a positive diagnosis then each sample in the pool is subsequently tested individually in order to identify the infected individual(s). This approach to testing is referred to as a two-stage Dorfman procedure, and, perhaps due to its simplicity, has seen widespread use (Hughes-Oliver, 2006). There has since been numerous variations on this idea, including different strategies for following up a positive test result (e.g., Sterrett, 1957), and having more than two layers of the hierarchical testing strategy (e.g., Malinovsky et al., 2020) (non-hierarchical strategies in which each sample may appear in multiple pools also has received much attention, but this paper will not focus on these approaches). See Hughes-Oliver (2006) or Bilder (2022) for more information on pooled testing.

By reducing the resource requirements, pooled testing allows for screening to be used as infection control in settings where comprehensive individual-level screening would be infeasible. Anthony Fauci, the Director of the National Institute of Allergy and Infectious Diseases, stated at a U.S. Senate hearing, '[Pooled testing] is a really good tool. It can be used in any of a number of circumstances, including at the community level or even in schools' (Associated Press, 2020) and others have noted its potential application at schools, offices, religious organisations and factories (Lee et al., 2020). In low prevalence areas, pooled testing can be particularly effective. Wacharapluesadee et al. (2020) determined that in low prevalence settings the required laboratory resources could be decreased by up to 80%; Pilcher et al. (2020) estimated that when compared

with individual testing, pooled testing could screen 2 to 20 times as many individuals for the same cost; and Abdalhamid et al. (2020) estimated that so long as the incidence rate of COVID-19 is 10% or less pooled testing could save 69% of reagents and personnel time.

In recent practice, China, Germany, Israel, and Thailand have implemented pooled testing for SARS-CoV-2 (Mandavilli, 2020). In the United States, the Nebraska Public Health Laboratory performed pooled testing until prevalence became too high (Stone, 2020), and Duke University effectively implemented twice-per-week screening of its student population via a pooled testing strategy (Denny et al., 2020). While pooled testing is widely used beyond infectious disease settings (Bilder, 2019, cites its usage in determining virus transmission from an insect to a plant, bacteria screening for food, discovery of new pharmaceuticals, and verification of computer network security), the focus of this paper is the use of pooled testing in screening a population for an infectious disease; the terminology used henceforth will reflect this focus.

The goal of this work is to reduce the total number of expected tests required to screen infectious diseases by leveraging how the members of the population are connected without compromising the number of correctly classified individuals. By using the network and how one individual may transmit the disease to another, we can construct pools of individuals which can reduce the total number of expected tests. There is a long history of optimising pooled testing strategies when there is available additional information on the heterogeneity of probabilities of being infected (Bilder et al., 2010; Black et al., 2015; Hwang, 1975; Malinovsky et al., 2020; Yao & Hwang, 1988). This paper also relates to obtaining an efficient pooling strategy, yet the direction is orthogonal to previous work. While previous literature assumes subjects are independent with perhaps differing conditional probabilities of being infected given some exogenous information, our focus is on minimising the total number of expected tests based on the interrelatedness of the individuals of the population. Work by Lendle et al. (2012) showed that when those being tested belong to clusters with inter-cluster variation in disease prevalence, the association between individuals within clusters can be leveraged to improve the testing strategy. However, this work assumed (1) units in different clusters were independent, and (2) units within a cluster were exchangeable. Both these assumptions are incompatible with complex networks connecting individuals.

The remainder of the paper is structured as follows. Section 2 describes the proposed methods for minimising the expected number of tests. Section 3 outlines an agent-based model and describes its use for evaluating the efficacy of our proposed approach. Finally, we provide a brief discussion in Section 4.

2 | METHODOLOGY

2.1 | Objective

Let n denote the total number of individuals in the population. In order to screen all n individuals using a two-stage Dorfman procedure, we first partition the n individuals into P pools of $K = n/P$ individuals each. Let $\mathcal{I}_p \subset \{1, 2, \ldots, n\}$, $|\mathcal{I}_p| = K$, denote the set of individuals assigned to the pth pool. (Later we will relax $|\mathcal{I}_p| = K \; \forall p$, allowing for K not necessarily being a factor of n, but for the purposes of exposition we will currently make this simplifying assumption.) If any pool results in a positive test, we subsequently test the K individuals in the corresponding pool. Finally, let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denote the $n \times 1$ vector such that y_i equals 1 if the ith individual is infected and zero otherwise, and let $\boldsymbol{y}_{\mathcal{I}_p}$ denote the $K \times 1$ sub-vector of $\boldsymbol{y}$ corresponding to the individuals in the pth pool.

One of the primary outcomes of interest is the total number of tests, denoted T, and we wish to minimise the expected number of tests, $\mathbb{E}(T)$. Let S_e denote the sensitivity of the pooled test if at least one individual in the pool is positive, and let S_p denote the specificity of the pooled test. If individuals are independent and are infected according to the prevalence rate ρ, then the expected total number of tests in a pool of size K was given by Dorfman (1943) as

$$1 + KS_e - K(S_p + S_e - 1)(1 - \rho)^K. \tag{1}$$

However, simply minimising the total number of tests is insufficient when the sensitivity and specificity of the test are not both equal to 1. Rather, one should also account for the number of correctly classified individuals being tested, denoted C (see, e.g., Aprahamian et al., 2019; Graff & Roeloffs, 1972; Litvak et al., 2020). Malinovsky et al. (2016) suggested using the ratio of the expected number of correct classifications to the expected number of tests, $\mathbb{E}(C)/\mathbb{E}(T)$, which we will adopt here as our objective. In a similar way as $\mathbb{E}(T)$, Malinovsky et al. showed that for independent individuals the expected number of correctly classified individuals in a pool of size K can be computed as

$$K\left[S_e^2 + (1-\rho)(S_eS_p + 1 - S_e - S_e^2) + (1-\rho)^K(1-S_p)(S_p + S_e - 1)\right]. \tag{2}$$

The objective function we which to maximise then is the ratio of (1) to (2).

In the setting of infectious disease, individuals within the population are not independent, but rather a network describing the contact patterns, or more generally transmission opportunities, between these individuals induces dependency amongst them. A network, consisting of a set of individuals and a set of edges which act to connect pairs of individuals, can be represented as a $n \times n$ symmetric adjacency matrix A where $A_{ij} = A_{ji}$ equals 1 if there is an edge between the ith and jth individuals and zero otherwise; note that the diagonal elements all equal zero. A related matrix we will utilise later is D^{-1}, the matrix whose *(i,j)*th element, D_{ij}^{-1}, is the reciprocal of the geodesic distance between the ith and jth individuals, where the geodesic distance between two individuals in the network is defined to be the length of the shortest path between them. That is, $D_{ij} = \min\{L : A_{i\ell_1}A_{\ell_1\ell_2}\cdots A_{\ell_{L-1}j} = 1\}$, and D^{-1} contains the element-wise inverses (and does not represent the matrix inverse of D).When there does not exist a path between every pair of individuals, that is, we have a disconnected graph, the distance between unreachable pairs is defined by convention to be ∞, and hence $D_{ij}^{-1} = 0$. For our purposes (see Section 2.3) we set the diagonal elements of D^{-1} to be zero.

In this context, individuals are no longer exchangeable, and both the expected number of correct classifications and the correct number of tests are inherently functions of the pool assignments. Let Z denote the $n \times P$ matrix of pool assignments, so that Z_{ip} equals 1 if the ith individual belongs to the pth pool and zero otherwise. This matrix then induces the sets $\mathcal{I}_p$, $p = 1, \ldots, P$. The expected number of tests given the pool assignments can be derived as follows:

$$\begin{aligned}
\mathbb{E}(T|Z) &= \sum_{p=1}^{P}\left(1 + K_p\mathbb{P}(p\text{th pool tests }+)\right)\\
&= P + \sum_{p=1}^{P} K_p\left[S_e(1 - \mathbb{P}(\mathbf{y}'_{I_p}\mathbb{1}_{K_p} = 0)) + (1-S_p)\mathbb{P}(\mathbf{y}'_{I_p}\mathbb{1}_{K_p} = 0)\right]\\
&= P + nS_e - (S_p + S_e - 1)\sum_{p=1}^{P} K_p\mathbb{P}(\mathbf{y}'_{I_p}\mathbb{1}_{K_p} = 0),
\end{aligned} \tag{3}$$

where $\mathbb{1}_m$ is the $m \times 1$ vector of ones. Note that we have added a subscript to K to denote the size of the pth pool, since K will not in general be a factor of the total number of individuals to be tested. For example, in the analyses presented in this paper we have taken the remainder and distributed them across the other pools, so that pools will have either K or $K+1$ members.

Similarly, we can evaluate the expected number of correctly classified individuals given the pool assignments as

$$\begin{aligned}\mathbb{E}(C|Z) &= \sum_{p=1}^{P}\left[\sum_{k=1}^{K_p}\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=k)\left(S_e(kS_e+(K_p-k)S_p)+(1-S_e)(K_p-k)\right)\right.\\ &\qquad\left.+\,\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=0)\left(K_pS_p+(1-S_p)K_pS_p\right)\right]\\ &= nS_e^2+\sum_{p=1}^{P}\left[(K_p-\mu_p)\left(S_eS_p+1-S_e-S_e^2\right)\right.\\ &\qquad\left.+K_p(1-S_p)(S_p+S_e-1)\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=0)\right], \end{aligned} \tag{4}$$

where $\mu_p := \mathbb{E}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}|Z)$, the expected number of infected individuals in the pth pool. Taken together with (3), we have the following objective function we wish to maximise:

$$Q(Z) := \frac{\mathbb{E}(C|Z)}{\mathbb{E}(T|Z)}. \tag{5}$$

2.2 | Estimation of objective function

Unlike in the case of independent individuals, μ_p and $\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=k)$, $p=1,\ldots,P$, cannot be computed from the prevalence alone, yet prevalence and perhaps some disease characteristics are typically all that is known in practice. However, there is a wealth of well-studied transmission models available (see, e.g., Allen et al., 2008) which can be applied here.

Suppose we have a data generating function F. In very simple cases F may correspond to a closed-form likelihood, but in more realistic cases will be a simulator such as a network-based compartmental model, or even a highly complex agent-based model. We assume that F is parameterised by $\theta := (\theta_1, \theta_2)$, where θ_1 is the set of parameters of known disease characteristics (e.g., average recovery time), and θ_2 is the set of unknown parameters with prior $\pi(\theta_2)$. From F we can generate M *iid* samples of the n-dimensional vector of infection statuses, $y^{(m)}$, $m=1,2,\ldots,M$. The goal will be to generate such draws from F in order to obtain Monte Carlo estimates of $\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=0)$ and μ_p.

Since the prevalence ρ is typically the only thing known, we wish to generate $y^{(m)}$ from F conditional on $\frac{1}{n}y^{(m)'}\mathbb{1}_n=\rho$. That is, when writing $\mathbb{P}(\boldsymbol{y}'_{\mathcal{I}_p}\mathbb{1}_{K_p}=0)$ and μ_p above, we are implicitly conditioning on $y\sim F$ and $\frac{1}{n}y^{(m)'}\mathbb{1}_n=\rho$. (Note that in the pooled testing literature, the prevalence is always implicitly conditioned on, but the way the data is generated, F, is not.) This may be computationally infeasible to generate many draws of y with prevalence exactly equal to ρ and impossible in the likely case that $n\rho$ is not an integer. Hence we are interested in the approximate Bayesian computation (ABC) distribution

$$\pi_{\text{ABC}}(y, r, \theta|\rho) \propto 1_{\{|r-\rho|<h\}}\delta_r(r(y))f(y|\theta)\pi(\theta_2), \tag{6}$$

where $1_{\{a\}}$ equals one if condition a is true and zero otherwise, δ is the dirac delta function, $r(y)$ is the prevalence of y, and f is the probability mass function obtained from F (not necessarily known in closed form). As the tuning parameter $h \to 0$, the ABC marginal of y becomes $\pi(y|r(y) = \rho)$. Draws from π_{ABC} can easily be obtained in the following way:

1. `SET` $m = 1$
2. `GENERATE` $\theta_2^{(m)} \sim \pi(\theta_2)$
3. `GENERATE` $y^{(m)}$ from F parameterised by $(\theta_1, \theta_2^{(m)})$
4. `SET` $r(y^{(m)}) = y^{(m)'}\mathbb{1}_n$.
5. `IF` $|r(y^{(m)}) - \rho| < h$ `THEN` $m \leftarrow m + 1$.
6. `IF` $m = M + 1$ stop. `ELSE` Go back to step 2.
7. `RETURN` $\left\{y^{(1)}, \dots, y^{(M)}\right\}$

With M draws of y with the desired prevalence ρ (up to $\pm h$) and underlying data generating mechanism F, we can well approximate $Q(Z)$ by plugging in

$$\widehat{\mathbb{P}}\left(\boldsymbol{y}'_{I_p}\mathbb{1}_{K_p} = 0\right) = \frac{1}{M}\sum_{m=1}^{M} 1_{\left\{\boldsymbol{y}_{I_p}^{(m)'}\mathbb{1}_{K_p}=0\right\}},$$

$$\text{and } \widehat{\mu}_p = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{y}_{I_p}^{(m)'}\mathbb{1}_{K_p}.$$

Because the $y^{(m)}$'s are sparse, both of these estimates are reasonably fast to compute.

This approach is extremely flexible by allowing any data generating process to be used. Although this ABC algorithm may be somewhat time consuming depending on F, it only needs to be run once in order to provide the $y^{(m)}$'s necessary to optimise Z (see next section). It is also an embarrassingly parallel task.

2.3 | Optimising pool assignments

In the context of networked individuals, the absence of exchangeability implies that the specific pool assignments matter, and that the computation of the objective function depends on a specific set of assignations of n individuals into P pools of equal (or nearly equal) size. The constraint on equal sized pools leads us to formulate the problem in the following way. (For simplicity, we will assume that all pools have exactly K members; extending this to fixed K_p is trivial.) Let $Z := I_P \otimes \mathbb{1}_K$, where I_P is the $P \times P$ identity matrix and $\otimes$ denotes the Kronecker product; let σ be a permutation of $\{1, 2, \dots, n\}$ which will be used to control the pooling assignments; and let $Z_\sigma = (Z'_{\sigma,1}, \dots, Z'_{\sigma,n})'$ denote Z with its rows permuted according to σ. That is, the ith row of Z_σ, $Z_{\sigma,i}$, has a one in the place marking the pool to which i belongs and zeros everywhere else, and the pth column, $Z_{\sigma,\cdot p}$, has K_p ones marking those belonging to the pth pool and zeros everywhere else.

By optimising over permutations of the rows of Z, we are able to maintain the constraint over the pool sizes, that is, the row and column sums of Z_σ are constant regardless of the permutation. To perform this optimisation with respect to $Q(Z)$ as given in (5), we propose using a simulated annealing (SA) algorithm (Kirkpatrick et al., 1983). SA is a widely used stochastic

optimisation approach that is more effective than greedy algorithms at avoiding local modes. The algorithm requires a sequence of L decreasing temperatures $\{\mathcal{T}_\ell\}_{\ell=1}^{L}$ which controls the freedom of movement around the space of permutations. Higher (lower) temperatures imply that it is easier (harder) to transition to permutations with poorer values of the objective function. Much attention has been given to evaluating the theoretical properties of various cooling schedules, for example, Cohn and Fielding (1999). Speaking coarsely, the optimal solution is guaranteed for a cooling schedule which is sufficiently slow, such as the logarithmic schedule of Geman and Geman (1984). However, such schedules tend to be too slow to be practical, and instead faster schedules are implemented (Nourani & Andresen, 1998). In the analyses that follow, we used an exponential schedule of $\mathcal{T}_\ell = 2 \cdot (0.95)^\ell$ which worked well.

At the ℓth iteration, we randomly generate a candidate permutation $\tilde{\sigma}$. To determine if we move from the current permutation $\sigma^{(\ell-1)}$ to the candidate, we draw $u \sim Unif(0,1)$ and set $\sigma^{(\ell)}$ equal to $\tilde{\sigma}$ if

$$u < \exp\left\{\frac{\log Q(Z_{\tilde{\sigma}}) - \log Q(Z_{\sigma^{(\ell-1)}})}{\mathcal{T}_\ell}\right\},$$

and equal to $\sigma^{(\ell-1)}$ otherwise.

To generate the candidate permutation transitioning from some permutation σ, we propose the following. Let $S_\sigma := Z_\sigma' D^{-1} Z_\sigma$, the $P \times P$ matrix which sums the inverse geodesic distances between the individuals in each pair of pools as assigned by σ. The off-diagonal elements of S_σ provide a measure as to the closeness between individuals belonging to different pools. To generate a candidate permutation, we first choose a pair of pools with probability proportional to the upper triangular elements of S_σ. We then randomly choose one individual from each pool. These individuals' rows of Z_σ are swapped, corresponding to a new permutation $\tilde{\sigma}$. By basing our candidate permutation based on how close the two pools are, we avoid proposing unlikely candidates for a swap, thereby improving the acceptance rate. Should $\tilde{\sigma}$ be accepted, we can compute $S_{\sigma^{(\ell)}}$ by first assigning it to be equal to $S_{\sigma^{(\ell-1)}}$ and then updating those elements which involve the two pools whose members have changed. Convergence is reached when no swaps are accepted for a sufficiently long period of the chain.

We initialised our algorithm in the following way. We start by using the geodesic distances in a k-medoids clustering algorithm (Everitt et al., 2011), where the number of clusters is P. We then ordered each individual by the difference between the distance to the nearest cluster medoid and the median distance to the other medoids. We then in this order assigned individuals to their nearest non-full cluster until all individuals were assigned to a pool with $K-1$ others.

3 | EVALUATION VIA AN AGENT-BASED MODEL

3.1 | Agent-based model for COVID-19

In order to evaluate our approach to refining a pooled testing strategy, we built an agent-based model (ABM) to simulate the spread of COVID-19 in a high school setting. The goals were to (1) compare the objective function $Q(Z)$ as given in (5) when assigning individuals to testing pools randomly versus using our proposed network-based approach, and (2) determine how any potential improvement might be attenuated by adding noise to the network used to assign individuals to pools.

The agents in the model were students, and their infection status was recorded on a daily time scale. Students were set a pre-specified screening schedule of testing once per week. Test results were delayed by 1 day, and if a student obtained a positive test result, that student went into isolation. Isolation lasted for 10 days during which the student could not infect or be infected by others.

Each day, each susceptible student was infected according to a baseline importation probability of 0.0015, roughly corresponding to a 2% community prevalence rate. Upon becoming infected, each student would go through a cycle of up to 25 days of being infectious (1 to 12 pre-symptomatic days + day of symptom onset + 12 days after symptom onset), after which the student would no longer be infectious nor susceptible to becoming reinfected. (Note that for the sake of brevity we will simply refer to the time of symptom onset; for those who are asymptomatic this can be inferred to mean the day of peak viral load.) The length of infectivity varied due to a random incubation period. We randomly generated an incubation period in the range of 1 to 12 for each newly infected student according to He et al. (2020); after this period the student remained infectious for 13 additional days.

Excluding those students in isolation, the probability that an infectious individual who is t days away from their symptom onset infects a susceptible individual who is adjacent in the network was set to equal τc_t, $t = -12, -11, \ldots, 11, 12$, where c_t are values corresponding to the infectiousness profile of He et al. (2020), $t = 0$ corresponds to the date of symptom onset, and τ controls the overall transmission rate. We set τ to fix the basic reproduction number at 2.79, as was estimated in an early survey by Liu et al. (2020), accounting for the average degree in the network.

To estimate the sensitivity of tests according to how long an individual has been infected, we used data from Kucirka et al. (2020) to fit a logistic regression model using splines to fit a non-linear relationship between sensitivity and date relative to symptom onset. To account for dilution effects on sensitivity, we reduced the sensitivity of pooled tests by 16.4%, as was given in the COVID-19 example with pool size of 5 in Polage et al. (2020). The specificity was set at 0.995.

3.2 | Data generator

In computing the optimal pooling assignments, we did not use the ABM described above as our data generating function F. This was done intentionally for two reasons. First, while the ABM was not computationally onerous, it still required too much time to be used efficiently in the ABC algorithm described in Section 2.2 in our simulation study. A second and most important reason is that in practice our models will always be a simplification of reality with incorrect parameter estimates, and in our simulation study we wished to accurately reflect this disparity between how the data is actually generated (the ABM of Section 3.1) and how the data is posited to have been generated (F as described below).

We used a susceptible-infected-susceptible (SIS) model as F when determining optimal pool assignments. The SIS discrete-time network-based simulator takes as input a transmission parameter β, a number of days T, and a $n \times n$ network adjacency matrix A and outputs a n-dimensional binary vector indicating the infection status of each individual on the Tth day. The SIS simulator begins by randomly infecting one individual who remains infected for 7 days. Each day, each infected individual with probability β infects her/his susceptible contacts. We set $T = 300$ to ensure we had reached a state of equilibrium, and used a uniform prior on β over the range 1.15 to 1.85 times the epidemic threshold divided by the length of the infectious

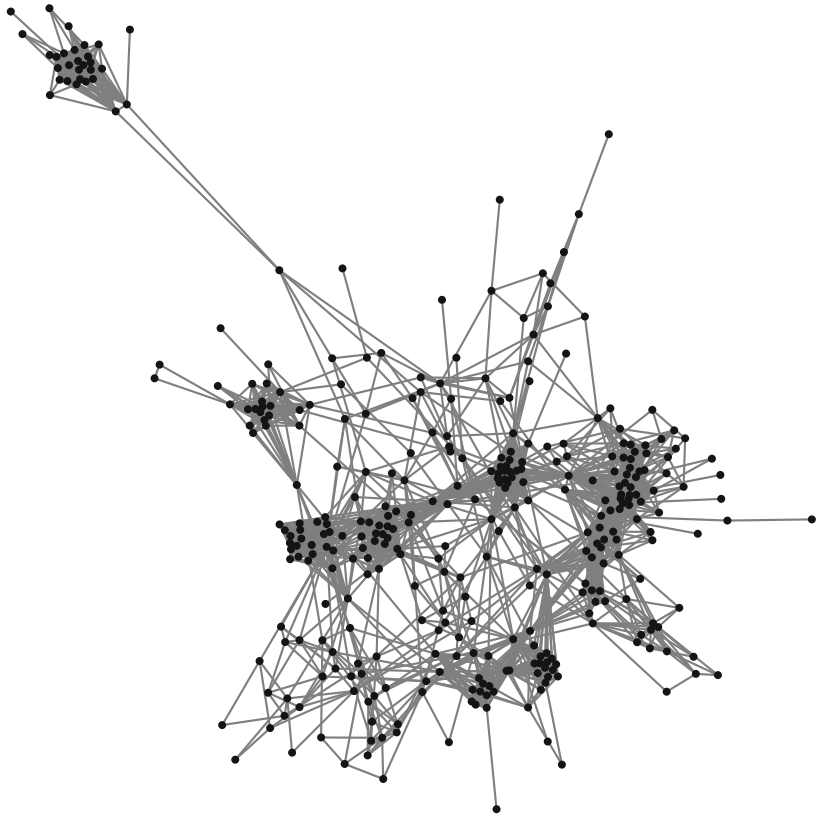

FIGURE 1 High school friendship network of 355 students

period[1] for an infectious period of 7 days. Values larger or smaller than this tended to yield prevalence rates too small and too large, respectively. As described in the next section, we altered the true network A in various ways when implementing our SIS simulator, which in turn was used to estimate the $\mathbb{P}(\boldsymbol{y}'_{I_p}\mathbb{1}_{K_p} = 0)$ and μ_p values in the simulated annealing algorithm.

3.3 | Simulation setup

We used a real friendship network of 355 high schoolers (McFarland, 2001), available in the R package `NetData` (Nowak et al., 2012). This network, shown in Figure 1, was treated as the ground truth network used in the ABM to generate secondary infections. We considered pool sizes to range from 2 to 25 and chose the value that yielded the highest value of $E(C|Z)/E(T|Z)$. We then ran the ABM to simulate a 10-week period.

Because in reality we usually do not have the true network, we considered the following six settings for altering the true network when assigning students to pools. In all settings, the transmission in the ABM corresponded to the ground truth network.

[1]This threshold relates β to the recovery rate and the largest eigenvalue of A. Specifically, when β divided by the recovery rate is larger than the reciprocal of the largest eigenvalue of A, the epidemic will not die out. See, for example, Newman (2010) for details.

- *Random*. Individuals were assigned randomly to pools.
- *Oracle*. Individuals were assigned to pools using our proposed approach on the true network.
- *Nomination*. We emulated a social survey in which students were asked to nominate five other students with whom the respondent spends time. Each student randomly selected up to five of their contacts, and we then applied our proposed approach on the resulting network to assign individuals to pools.
- *Partial recall*. We emulated a social survey in which students were asked to list those with whom the respondent spends time. Each respondent listed each of their contacts with probability 0.6; that is, on average only 60% of contacts were listed. Our approach was then used on the resulting network to assign individuals to pools.
- *Nomination + re-wiring*. To add additional noise, we implemented the *Nomination* approach and subsequently rewired each edge between two random individuals with probability 0.05.
- *Partial recall + re-wiring*. To add additional noise, we implemented the *Partial recall* approach and subsequently re-wired each edge between two random individuals with probability 0.05.

We ran 250 simulations for each setting, recording the total number of tests and total number of correct classifications from each.

3.4 | Results

Using the high school network of 355 students, we compared pool sizes K ranging from 2 to 25 individuals using both our proposed method and the method of Malinovsky et al. (2016) ignoring the underlying contact network. Figure 2 shows how the correct number of classifications per test varies over K. The objective function from the method ignoring the network is dominated by

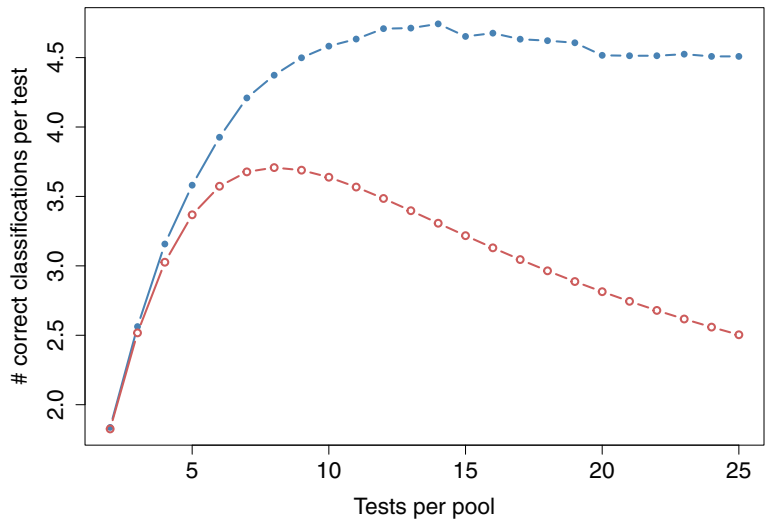

FIGURE 2 Expected number of correct classifications per test as a function of the number of individuals per test, K, for our proposed method (blue, filled circles) and for the approach ignoring the contact network (red, hollow circles) (colour online) [Colour figure can be viewed at wileyonlinelibrary.com]

our proposed approach. From this we obtain an optimal pool size of $K = 13$. It should be noted that while this will not be true in general, in this case the changes in the objective function are driven almost entirely by the changing number of tests required; the expected number of correct classifications for either approaches only ranged between 0.987 and 0.989 for all K we tried.

Figure 3 shows the value of the objective function over 100,000 iterations (500 temperatures, 200 iterations per temperature) of the proposed optimisation algorithm for $K = 13$, illustrating how the SA allows the solution to leave the initial local mode before achieving higher values. Figure 4 illustrates how our proposed approach can find pools of individuals who tend to either be densely intra-connected or all have low degree.

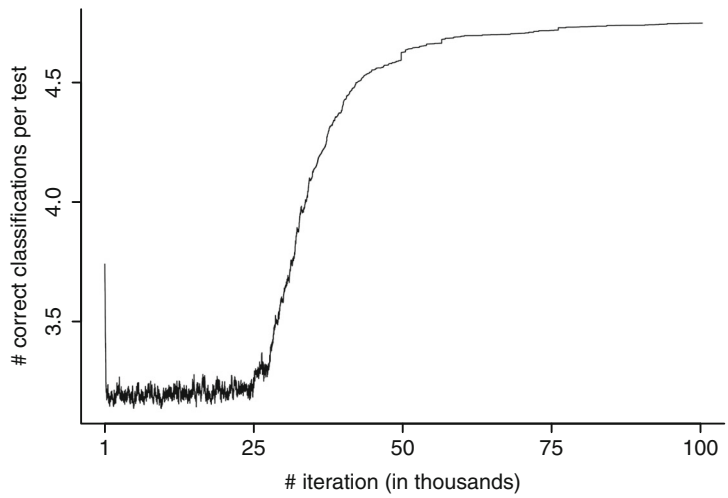

FIGURE 3 Objective function value over 100,000 iterations (horizontal axis) of the simulated annealing algorithm corresponding to the high school friendship network of 355 students and pool size equal to 13

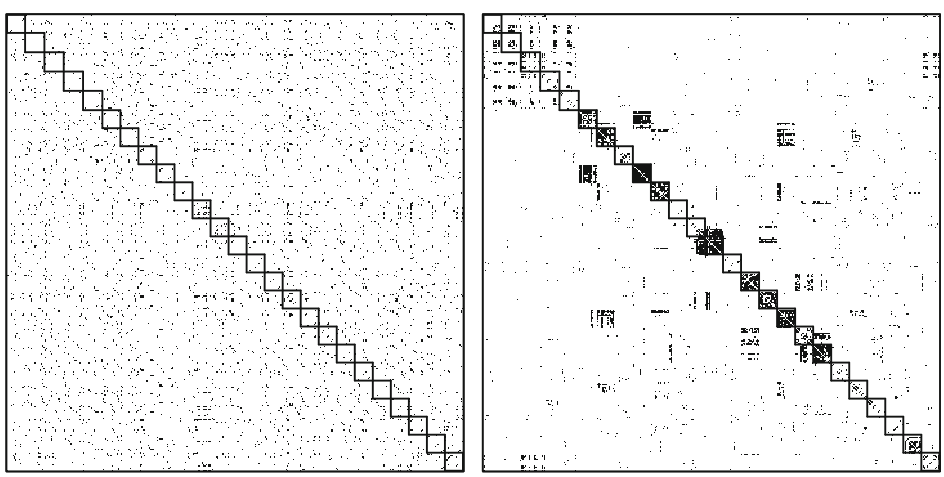

FIGURE 4 Adjacency matrix for the high school friendship network of 355 students. Boxes are drawn around edges connecting individuals in the same pool of size $K = 13$ students for randomly assigned pools (left) or pools assigned using our proposed approach (right).

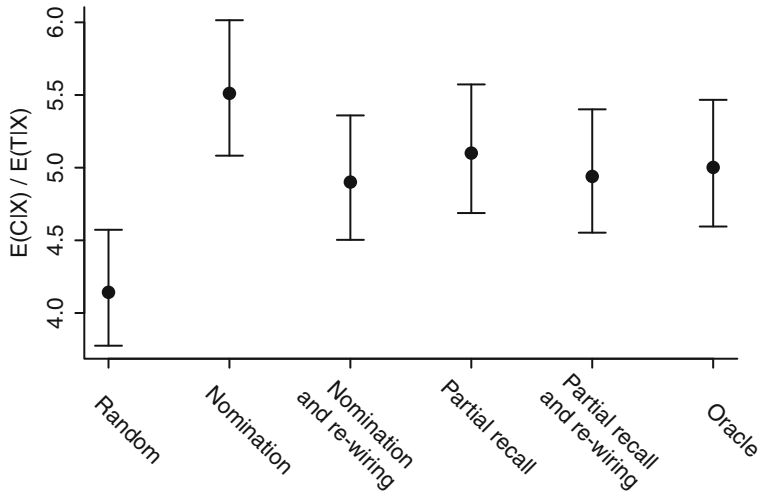

FIGURE 5 Simulation results showing the estimated objective function. 95% confidence intervals display the Monte Carlo error associated with the simulation results.

Figure 5 shows the estimated objective function for each of the six settings from our simulation study run on the ABM, with 95% confidence intervals expressing the Monte Carlo error. Two things are striking. First, by leveraging the network, the ratio of the expected number of correct classifications to the expected number of tests markedly increases compared to random assignments (21% increase). Second, there was not a meaningful difference between the oracle assignments and those network-based assignments relying on incomplete or noisy network data. This is highly reassuring, as practitioners can be fairly certain that the true transmission network will be only partially or noisily observed.

4 | DISCUSSION

For an institution such as a workplace or school facing an infectious disease outbreak, comprehensive screening can be a powerful tool for infection control, particularly when a non-ignorable fraction of transmission is due to pre-symptomatic or asymptomatic infectives. The cost of such screening, however, is very often prohibitive, and pooled testing can alleviate this burden. We have proposed a method of assigning pools when implementing a pooled testing strategy which leverages the underlying transmission network to further reduce the resource burden. We have provided R code implementing our proposed approach in Appendix S1[2].

The results from our ABM suggest that our approach is robust to incomplete and noisy measurements of the true underlying transmission network. This is an important point, as the true transmission network is not often observed in practice. It should be noted that advancing technology such as sensor mote deployments (e.g., Jang et al., 2019) or contact tracing mobile phone apps (see Ahmed et al., 2020, for a survey) can more feasibly measure the contact network. Still, since this technology may be out of reach for many settings, we emulated common survey methods

[2]This script will automatically install from the web and load the R package `networkPooledTesting`, available through the author's website at https://myweb.uiowa.edu/dksewell/software/networkPooledTesting_1.0.tar.gz

for measuring social networks which are relatively inexpensive to administer and found negligible differences in performance when compared to using perfect knowledge of the underlying transmission network.

This work serves to introduce a network perspective into the rich group testing literature. It is, however, but a first step, and there are several important directions for future research. First, while we have focused on a two-stage Dorfman procedure, higher efficiencies may be achieved through deeper hierarchical procedures (Johnson et al., 1991) or Sterrett procedures (Malinovsky, 2019). Second, an important area of research is how best to optimise pooling assignations when there exists information on both network connections and differential prevalences (Black et al., 2015; Malinovsky et al., 2020). Third, our algorithms are still fairly computationally intensive. Scalability remains a future goal in order to ensure large organisations can reasonably apply our methods. Fourth, the constraint that all pools are the same size are typically imposed for logistical reasons or for simplicity. Consider Figure 4. While clearly a large part of the network structure is captured in the pooling, there still exists remaining block structure that is not accounted for in the pools. As an anonymous reviewer pointed out, optimal pooling assignments are in general unlikely to be formed from equally sized pools, and a better solution may be reached if pool size could fluctuate within the bounds of what is technically or logistically feasible.

ACKNOWLEDGEMENTS

This work was supported by the US Centers for Disease Control and Prevention (5 U01 CK000531-02, 1 U01 CK000594-01-00) as part of the MInD-Healthcare Program. Special thanks to Philip Polgreen of the MInD-Healthcare Program for helpful and informative discussions on the subject of pooled testing.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the R package 'NetData' at https://cran.r-project.org/web/packages/NetData/.

ORCID

Daniel K. Sewell https://orcid.org/0000-0002-9238-4026

REFERENCES

Abdalhamid, B., Bilder, C.R., McCutchen, E.L., Hinrichs, S.H., Koepsell, S.A. & Iwen, P.C. (2020) Assessment of specimen pooling to conserve SARS cov-2 testing resources. *American Journal of Clinical Pathology*, 153(6), 715–718.

Ahmed, N., Michelin, R.A., Xue, W., Ruj, S., Malaney, R., Kanhere, S.S. et al. (2020) A survey of COVID-19 contact tracing apps. *IEEE Access*, 8, 134577–134601.

Allen, L., Bauch, C., Castillo-Chavez, C., Earn, D., Feng, Z., Lewis, M. et al. (2008) *Mathematical epidemiology*. Berlin, Heidelberg: Springer.

Aprahamian, H., Bish, D.R. & Bish, E.K. (2019) Optimal risk-based group testing. *Management Science*, 65(9), 4365–4384.

Associated Press. (2020, July) FDA approves quest COVID-19 test for 'pooled' sample use. Available from: https://www.washingtonpost.com/politics/fda-approves-quest-covid-19-test-for-pooled-sample-use/2020/07/19/6457924c-c9eb-11ea-99b0-8426e26d203b_story.html [Accessed 5th January 2021].

Bharti, N., Exten, C. & Oliver-Veronesi, R.E. (2020) *Lessons from campus outbreak management using test, trace, and isolate efforts*.

Bilder, C.R. (2022) Group testing for identification. In: Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F. & Teugels, J.L. (Eds) *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat08227.

Bilder, C.R. (2019) Group testing for estimation. In: Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F. & Teugels, J.L. (Eds) *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat08231.

Bilder, C.R., Tebbs, J.M. & Chen, P. (2010) Informative retesting. *Journal of the American Statistical Association*, 105(491), 942–955.

Black, M.S., Bilder, C.R. & Tebbs, J.M. (2015) Optimal retesting configurations for hierarchical group testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(4), 693–710.

Cohn, H. & Fielding, M. (1999) Simulated annealing: searching for an optimal temperature schedule. *SIAM Journal on Optimization*, 9(3), 779–802.

Denny, T.N., Andrews, L., Bonsignori, M., Cavanaugh, K., Datto, M.B., Deckard, A. et al. (2020) Implementation of a pooled surveillance testing program for asymptomatic sars-cov-2 infections on a college campus — Duke University, Durham, North Carolina, August 2-October 11, 2020. *Morbidity and Mortality Weekly Report*, 69(46), 1743.

Dorfman, R. (1943) The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4), 436–440.

Everitt, B.S., Landau, S., Leese, M. & Stahl, D. (2011) *Cluster analysis*. Chichester, UK: Wiley.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.

Graff, L.E. & Roeloffs, R. (1972) Group testing in the presence of test error; an extension of the Dorfman procedure. *Technometrics*, 14(1), 113–122.

He, X., Lau, E.H., Wu, P., Deng, X., Wang, J., Hao, X. et al. (2020) Author correction: temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26, 1491–1493.

Hughes-Oliver, J.M. (2006) *Pooling experiments for blood screening and drug discovery*. New York, NY: Springer, pp. 48–68.

Hwang, F.K. (1975) A generalized binomial group testing problem. *Journal of the American Statistical Association*, 70(352), 923–926.

Jang, H., Justice, S., Polgreen, P.M., Segre, A.M., Sewell, D.K. & Pemmaraju, S.V. (2019) Evaluating architectural changes to reduce infection spread in a dialysis unit. *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining '19*.

Johnson, N.L., Kotz, S. & Wu, X.-Z. (1991) *Inspection errors for attributes in quality control*. London: CRC Press.

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. (1983) Optimization by simulated annealing. *Science*, 220(4598), 671–680.

Kucirka, L.M., Lauer, S.A., Laeyendecker, O., Boon, D. & Lessler, J. (2020) Variation in false-negative rate of reverse transcriptase polymerase chain reaction–based sars-cov-2 tests by time since exposure. *Annals of Internal Medicine*, 173(4), 262–267.

Larremore, D.B., Wilder, B., Lester, E., Shehata, S., Burke, J.M., Hay, J.A. et al. (2021) Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances*, 7(1), eabd5393.

Lee, A., Kothare, A., Dollar, F., Azeez, L. & Rowley, N. (2020, September) Evidence-based medicine infosheet: COVID-19 diagnostics and testing. Available from: https://www.uthscsa.edu/sites/default/files/2018/ebm_diagnostics_infosheet_9_21_20.pdf.

Lendle, S.D., Hudgens, M.G. & Qaqish, B.F. (2012) Group testing for case identification with correlated responses. *Biometrics*, 68(2), 532–540.

Litvak, E., Dentzer, S. & Pagano, M. (2020) The right kind of pooled testing for the novel coronavirus: first, do no harm. *American Journal of Public Health*, 110(12), 1772–1773.

Liu, Y., Gayle, A.A., Wilder-Smith, A. & Rocklöv, J. (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 27(2), taaa021.

Malinovsky, Y. (2019) Sterrett procedure for the generalized group testing problem. *Methodology and Computing in Applied Probability*, 21, 829–840.

Malinovsky, Y., Albert, P.S. & Roy, A. (2016) Reader reaction: a note on the evaluation of group testing algorithms in the presence of misclassification. *Biometrics*, 72(1), 299–302.

Malinovsky, Y., Haber, G. & Albert, P.S. (2020) An optimal design for hierarchical generalized group testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3), 607–621.

Mandavilli, A. (2020, July) *Federal officials turn to a new testing strategy as infections surge*. Available from: https://www.nytimes.com/2020/07/01/health/coronavirus-pooled-testing.html [Accessed 4th January 2021].

McFarland, D.A. (2001) Student resistance: how the formal and informal organization of classrooms facilitate everyday forms of student defiance. *American Journal of Sociology*, 107(3), 612–678.

Moghadas, S.M., Shoukat, A., Fitzpatrick, M.C., Wells, C.R., Sah, P., Pandey, A. et al. (2020) Projecting hospital utilization during the COVID-19 outbreaks in the united states. *Proceedings of the National Academy of Sciences*, 117(16), 9122–9126.

Nadworny, E. & S. McMinn (2020, October) *Even in COVID-19 hot spots, many colleges aren't aggressively testing students*. Available from: https://www.npr.org/2020/10/06/919159473/even-in-covid-hot-spots-many-colleges-arent-aggressively-testing-students [Accessed 23rd December 2020].

Newman, M. (2010) *Networks: an introduction*. Oxford: Oxford University Press.

Nourani, Y. & Andresen, B. (1998, October) A comparison of simulated annealing cooling strategies. *Journal of Physics A: Mathematical and General*, 31(41), 8373–8385.

Nowak, M., Westwood, S.J., Messing, S. & McFarland, D. (2012) *NetData: network data for McFarland's SNA R labs*. R package version 0.3.

Oran, D.P. & Topol, E.J. (2020) Prevalence of asymptomatic sars-COV-2 infection. *Annals of Internal Medicine*, 173(5), 362–367.

Paltiel, A.D., Zheng, A. & Walensky, R.P. (2020) Assessment of sars-cov-2 screening strategies to permit the safe reopening of college campuses in the United States. *JAMA Network Open*, 3(7), e2016818.

Pilcher, C.D., Westreich, D. & Hudgens, M.G. (2020) Group testing for severe acute respiratory syndrome- coronavirus 2 to enable rapid scale-up of testing and real-time surveillance of incidence. *The Journal of Infectious Diseases*, 222(6), 903–909.

Polage, C.R., Lee, M.J., Hubbard, C., Rehder, C., Cardona, D., Denny, T. et al. (2020) Assessment of an online tool to simulate the effect of pooled testing for sars-cov-2 detection in asymptomatic and symptomatic populations. *JAMA Network Open*, 3(12), e2031517.

Sterrett, A. (1957) On the detection of defective members of large populations. *The Annals of Mathematical Statistics*, 28(4), 1033–1036.

Stone, A. (2020, March) Nebraska public health lab begins pool testing COVID-19 samples. Available from: https://www.ketv.com/article/nebraska-public-health-lab-begins-pool-testing-covid-19-samples/31934880 [Accessed 5th January 2021].

Sutton, D., Fuchs, K., D'Alton, M. & Goffman, D. (2020) Universal screening for sars-cov-2 in women admitted for delivery. *New England Journal of Medicine*, 382(22), 2163–2164.

Wacharapluesadee, S., Kaewpom, T., Ampoot, W., Ghai, S., Khamhang, W., Worachotsueptrakun, K. et al. (2020) Evaluating the efficiency of specimen pooling for PCR-based detection of COVID-19. *Journal of Medical Virology*, 92(10), 2193–2199.

Yao, Y.C. & Hwang, F.K. (1988) Individual testing of independent items in optimal group testing. *Probability in the Engineering and Informational Sciences*, 2(1), 23–29.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Sewell, D.K. (2022) Leveraging network structure to improve pooled testing efficiency. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5), 1648–1662. Available from: https://doi.org/10.1111/rssc.12594