Heterogeneity Aware Two-stage Group Testing

Mohamed A. Attia Wei-Ting Chang Ravi Tandon

Department of Electrical and Computer Engineering
University of Arizona, Tucson, Arizona 85721
Email: {*madel, wchang, tandonr*}@email.arizona.edu

***Abstract*—Group testing refers to the process of testing pooled samples to reduce the total number of tests. Given the current pandemic, and the shortage of test supplies for COVID-19, group testing can play a critical role in time and cost efficient diagnostics. In many scenarios, samples collected from users are also accompanied with auxiliary information (such as demographics, history of exposure, onset of symptoms). Such auxiliary information may differ across patients, and is typically not considered while designing group testing algorithms. In this paper, we abstract such heterogeneity using a model where the population can be categorized into clusters with different prevalence rates. The main result of this work is to show that exploiting knowledge heterogeneity can further improve the efficiency of group testing. Motivated by the practical constraints and diagnostic considerations, we focus on two-stage group testing algorithms, where in the first stage, the goal is to detect as many negative samples by pooling, whereas the second stage involves individual testing to detect any remaining samples. For this class of algorithms, we prove that the gain in efficiency is related to the concavity of the number of tests as a function of the prevalence. We also show how one can choose the optimal pooling parameters for one of the algorithms in this class, namely, doubly constant pooling. We present lower bounds on the average number of tests as a function of the population heterogeneity profile, and also provide numerical results and comparisons.**

I. Introduction

Group testing was first studied by Dorfman [1] who introduced the idea of testing groups (or pools) of subsets of the population as opposed to individual testing with aims of reducing cost and time by reducing the amount of tests required. The goal is to identify all positive samples (for some disease or defection) out of a large population [1]–[10]. For an ideal (noiseless) pooled test, the outcome is negative when all samples in the pooled test are negative, and positive otherwise. Therefore, carefully designed group testings are needed to identify all positive samples with minimum number of tests.

Group testing has also been used for screening for diseases such as HIV [11], [12], Zika Virus [13] and more [11], [14]–[16]. Due to the recent COVID-19 outbreak, group testing has gained a lot of interest. The predominant method for detecting COVID-19 Coronavirus is real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) based diagnostic test. Interestingly, recent works have shown that using real time RT-PCR for detecting COVID-19 on pools of around 32 samples or fewer with at least one positive sample will reliably give a positive outcome [17]–[23]. The idea of group testing has also been widely applied in other fields such as communications [24], compressed sensing [25] and machine learning [26].

Group testing algorithms can be broadly categorized as either non-adaptive [2], [3], [27], [28] or adaptive [29]–[32]. Non-adaptive group testing algorithms refer to algorithms where all tests are designed beforehand. Non-adaptive algorithms are efficient due to the parallelizability of tests, however, they can lack flexibility since tests are designed in advance, which causes excessive/insufficient tests in most cases. On the other hand, adaptive group testing refers to algorithms where the design of some later tests can depend on the results of previous tests. Adaptive group testings are not fully parallelizable as its non-adaptive counterpart. A good compromise is to use a hybrid algorithm with multiple stages, where the tests within each stage is non-adaptive, but adaptive across stages [4], [5], [33].

Group testing can also be categorized based on the underlying assumptions made about the population disease prevalence. Algorithms can be designed for two different settings: (a) combinatorial and (b) probabilistic. In combinatorial test designs, one assumes that the number of positive samples is fixed among the population and a bound on the number of positives may be known. In probabilistic testing, one assumes that each sample is positive with some fixed probability, referred to as the prevalence rate p. In this work, we argue that in practice both of these assumptions do not utilize the auxiliary information that can be collected about the population of samples.

In most cases, screenings and responses to questionnaires are collected before the actual testing to learn additional information from the patients, such as demographics, history of exposure, onset of symptoms, medical and travel history [34]–[36]. Such auxiliary information can be linked to a finer understanding of the heterogeneity among the patients. One can potentially use this information in two different ways: (a) estimate the individual likelihood of the positivity of each patient and subsequently divide the population into clusters, with different local prevalence rates, according to some thresholds on the likelihood. The local prevalence of each cluster can then be found as the average likelihood of the patients in the cluster; or (b) categorizing the population into groups according to the potential risk, e.g., low risk (asymptomatic, no known exposure), medium risk (asymptomatic, known exposure), and high risk (symptomatic). The local prevalence of each cluster can be estimated using one of the prevalence estimation techniques in the literature [37]–[39]. Estimating prevalence rate often requires accessing past data. It is worth

This work was supported in part by a University of Arizona RII Reopening Award, and NSF Grants CAREER 1651492, CNS 1715947.

noting that, while clusters can be formed more intelligently with individual likelihood, past data on the individual-level (Approach (a)) may not be available due to privacy concern in practice, however, past data on the group-level (Approch (b)) is often public, can be obtained easily and estimation of individual prevalence is not required.

In this work, we study the benefits of leveraging the heterogeneity knowledge of the patient groups/clusters in order to reduce the total number of required tests. We refer to this approach as heterogeneity aware group testing. Specifically, we focus on Approach (b), where we do not rely on individual prevalence. We formalize this by assuming that the population is divided into C clusters. For each cluster $c \in [1, \ldots, C]$, we make the assumption that the probability that a patient is positive is equal to p_c, and the samples are i.i.d. within each cluster.

Main contributions: We consider heterogeneity-aware two-stage group testing under a probabilistic model, where in the first stage, the goal is to detect as many negative samples by pooling, whereas the second stage involves individual testing to detect any remaining samples. For the pooling strategy, we adapt doubly constant method [5], among each population cluster, where constant numbers of samples per pool and tests per sample are assumed. Doubly constant provide practical constraints on pooling parameters. In addition, as shown in [4], doubly constant outperforms other existing two-stage group testing algorithms. However, finding the optimal pooling parameters is non-trivial. We provide an approach to obtain the optimal pool size through a series of approximations for a given prevalence rate and total number of tests per sample. We present the empirically derived optimal pooling parameters for doubly constant pooling. We show that, by exploiting heterogeneity, the efficiency of group testing is improved due to the concavity of the number of tests as a function of the prevalence rate p. In particular, we show numerically that the average total tests per sample is concave for optimized pooling constants, i.e., pool size and tests per sample. We also provide analytical results for the concavity property for small values of p. We present a lower bound on the number of tests as a function of the heterogeneity profile (characterized by the prevalence rates of the clusters, namely $p_1, p_2, \ldots, p_C$). Moreover, we conduct experiments through simulations showing that heterogeneity aware performance is superior to heterogeneity unaware, and the performance enhances for higher levels of heterogeneity.

Related work: Heterogeneity aware group testing has been studied previously in several works [40]–[44]. In [40], the authors proposed an algorithm that first orders the samples in a positive pool by its individual estimated prevalence rate, and samples are re-tested individually until the first positive sample is found. The remaining samples are pooled again for subsequent testing. The individual prevalence rates are estimated using gender, age, race, symptoms, etc, either from past data or initial test results. The authors of [41] proposed ordered halving algorithm which creates two pools of equal size. By ordering the samples based on the prevalence rates and using the median as the threshold, the probability of one of the pools being positive is maximized, while the probability of the other pool being negative is minimized. In [42], the authors proposed an algorithm that first design tests that focus on minimizing the expected number of false negatives. Then, samples are carefully placed into each test based on their individual risk factors. The optimal placement is obtained through three-stage optimization problem with budget constraints. In addition to the test design, [42] also considered the effects of imperfect tests and dilution. The authors of [43] formulate the problem of finding the optimal pool size as a partitioning problem, which can then be converted to a constrained shortest path problem. The partitioning problem is formulated in a way that it contains some beneficial structure properties, which allow an originally NP-hard problem to be approximately solved in an efficient manner. The above works assumed that each sample is positive with an individual prevalence rate. To incorporate heterogeneity knowledge, the population was either split using threshold-based schemes into two clusters or through computationally intractable optimization problems.

One of the works that is closely related to our work is [44]. A threshold-based design is proposed in [44], with the flexibility of optimizing the threshold. Samples are grouped into low and high risks groups based on their estimated prevalence rates. Samples in the low risk pools are further divided and tested using Dorfman style testing, whereas samples in the high risk pools are tested individually. The authors in [44] proposed to estimate the individual prevalence rates using similar method of that in [40] with past data. While our work is similar to [44], we point out that there are subtle distinctions between the two. First, we consider doubly constant pooling strategy in this paper, for which obtaining the optimal pooling parameters is more complex. In addition, we are interested in understanding how much reduction can be obtained in a scenario where we use coarser knowledge of prevalence rates. Furthermore, we derive lower bounds that take heterogeneity into account. However, we note that the algorithm of [44] can also be applied when only coarser knowledge is available, in which case, the algorithm of [44] is equivalent to the heterogeneity aware Dorfman's algorithm that will be discussed and compared later.

Recently, [45] considered community aware group testing for a heterogeneous population model where the total population is divided into F families. It is assumed in [45] that each family is equally likely to have at least one infected member. For infected families, members within the family are infected with a different probability. The main idea is to mix the samples within each family, and then perform group testing on samples across families. We can view this model as accounting for a micro (family) level heterogeneity.

In this paper, we model the heterogeneity differently by grouping the population according to the risk of infection into C clusters with different local prevalence rates. In contrast to [45], we can view our approach as taking into account a macro-level view of heterogeneity. We show that significant gains are still achievable by using coarser knowledge, i.e., without assuming sample-level individual prevalence rates. We also present a lower bound (converse result) for the two-stage group testing problem with prevalence heterogeneity.

II. System Model

In this work, we consider a population of N heterogeneous samples, which can be categorized into C clusters. Let $v_n \in \{0,1\}$ denote the true status of sample n for all $n \in [1:N]$, where $v_n = 0$ when the true status of sample n is negative and $v_n = 1$ otherwise. In addition, we assume that each cluster c consists of $\alpha_c N$ i.i.d. samples, where α_c is the fraction of population in cluster $c \in [1:C]$. A sample in cluster c is positive with local prevalence rate p_c. We can compactly represent $\overline{p} = [p_1 \ \dots \ p_C]$ and $\overline{\alpha} = [\alpha_1 \ \dots \ \alpha_C]$ as the prevalence rate and population fraction vectors, respectively. The samples are categorized into clusters based on the auxiliary information obtained through screenings and questionnaires. For instance, we may consider the following scenario with $C = 3$ categories: (a) low risk (asymptomatic, no known exposure); (b) medium risk (asymptomatic, known exposure); and (c) high risk (symptomatic), with local prevalence rates p_L, p_M and p_H, where $p_L < p_M < p_H$. The average prevalence for the entire population can be written as the convex combination of p_c's with ratios α_c's, i.e.,

$$p = \overline{\alpha} \cdot \overline{p} = \sum_{c=1}^{C} \alpha_c p_c. \tag{1}$$

Given an average prevalence rate p, samples are pooled according to some pooling strategy ψ_p, which can be adaptive or non-adaptive. In this work, we focus on two-stage group testing algorithms, where in the first stage, the goal is to detect as many negative samples by pooling, whereas the second stage involves individual testing to detect any remaining samples. Let $\phi(\psi)$ denote the decoding algorithm for pooling strategy ψ. The pooling strategy ψ_p is feasible if it allows decodability. Let $\phi(\psi) = [\hat{v}_1, \ \dots, \ \hat{v}_N]$ denote the estimates for the n samples, then, we require

$$\Pr(\hat{v}_n = v_n) = 1, \ \forall n, \tag{2}$$

for exact recovery of the unknown status of the samples.

In this work, we focus on optimizing the performance of the testing algorithm, which is evaluated by finding the average number of tests per sample needed (or expected normalized number of tests normalized by the entire population N), defined as $T(p, \psi_p, \phi(\psi_p))$, to identify all positive samples. Tests are assumed to be perfect/noiseless, i.e., always return correct results. The optimal expected normalized number of tests can then be obtained by minimizing $T(p, \psi_p, \phi(\psi_p))$ over all feasible pooling and decoding methods, i.e., find the optimal pooling method ψ_p^{opt} and optimal decoding method $\phi^{\text{opt}}(\psi_p^{\text{opt}})$,

$$T^{\text{opt}}(p) = \min_{\Psi_p, \Phi(\Psi_p)} T(p, \psi_p, \phi(\psi_p)) = T(p, \psi_p^{\text{opt}}, \phi^{\text{opt}}(\psi_p^{\text{opt}})), \tag{3}$$

where Ψ_p is a set of all feasible two-stage pooling strategies for p, and $\Phi(\Psi_p)$ is a set of all algorithms that, when applied to Φ_p, allow decodability of the unknown status of samples, defined in (2).

A naive choice of ψ_p is to apply group testing on the entire population and ignore the heterogeneity knowledge, i.e., ignore the clustering. This approach is referred to as the *heterogeneity unaware* approach. Another possible choice of ψ_p is to perform individual group testing on each cluster c, referred to as the *heterogeneity aware* approach. Due to the independence of pooling design of each cluster, we can use the optimal pooling strategy and decoding algorithm for each cluster c, denoted as $\psi_{p_c}^{\text{opt}}$ and $\phi^{\text{opt}}(\psi_{p_c}^{\text{opt}})$, according to the local prevalence p_c. The total number of tests for this approach is the sum of numbers of tests required for all clusters. Let $\overline{\psi}_{\overline{p}} = [\psi_{p_1} \ \dots \ \psi_{p_C}]$ and $\overline{\phi}(\overline{\psi}_{\overline{p}}) = [\phi(\psi_{p_1}) \ \dots \ \phi(\psi_{p_C})]$. This gives the expected normalized number of tests, denoted as $T^{\text{Het.}}(\overline{p}, \overline{\alpha}, \overline{\psi}_{\overline{p}}, \overline{\phi}(\overline{\psi}_{\overline{p}}))$, as follows,

$$\begin{aligned} T^{\text{Het.}}(\overline{p}, \overline{\alpha}, \overline{\psi}_{\overline{p}}, \overline{\phi}(\overline{\psi}_{\overline{p}})) &= \min_{\substack{\Psi_{p_c}, \Phi_{p_c} \\ c \in [1:C]}} \sum_{c \in [1:C]} \alpha_c T(p_c, \psi_{p_c}, \phi(\psi_{p_c})) \\ &= \sum_{c \in [1:C]} \alpha_c T(p_c, \psi_{p_c}^{\text{opt}}, \phi^{\text{opt}}(\psi_{p_c}^{\text{opt}})) \\ &= \sum_{c \in [1:C]} \alpha_c T^{\text{opt}}(p_c). \end{aligned} \tag{4}$$

By comparing (3) and (4), we observe that the concavity of $T^{\text{opt}}(p)$ as a function of the prevalence p is a sufficient condition for the heterogeneity aware approach to acquire lower expected normalized number of tests, i.e.,

$$\sum_{c \in [1:C]} \alpha_c T^{\text{opt}}(p_c) \leq T^{\text{opt}}\Big(\sum_{c \in [1:C]} \alpha_c p_c\Big) = T^{\text{opt}}(p). \tag{5}$$

In other words, if $T^{\text{opt}}(\cdot)$ is concave, we can expect to have reduction on the expected normalized number of tests, which depends on the heterogeneity profile, i.e., p_c's. We note that proving concavity analytically could be non-trivial, thus, we show concavity of expected normalized number of tests of Doubly constant algorithm numerically and through approximation in Section IV-B.

In this paper, we focus on two-stage group testing, which is known practically [4], [5], [33] to achieve a good trade-off between parallelizability and low number of tests. Next, we give an overview on several two-stage group testing algorithms commonly considered in the literature.

III. Overview of Two-stage Group Testing

Two-stage (T.S.) group testing with an aim of exact recovery of the positive samples can be described as follows:

- Stage 1 (Pooled Testing): Samples are grouped into T_1 pools using some pooling strategy with parameters $\psi_p^{\text{T.S.}}$. The aim of this stage is to identify as many negative samples as possible from the pooled tests. If a pool is negative, all samples in the pool are declared as definite negatives (DNs). Equivalently, a sample is declared as negative if it appears in at least one negative pool. The number of pooled tests in Stage 1 is T_1.
- Stage 2 (Conservative Individual Testing): All samples that are not declared as DNs in the first stage are tested individually. The number of individual tests conducted in Stage 2 is denoted by T_2.

From here onwards, we omit the argument for the decoding algorithm from $T(p, \psi_p, \phi(\psi_p))$ since we use the same decoding algorithm for the pooling algorithms being discussed

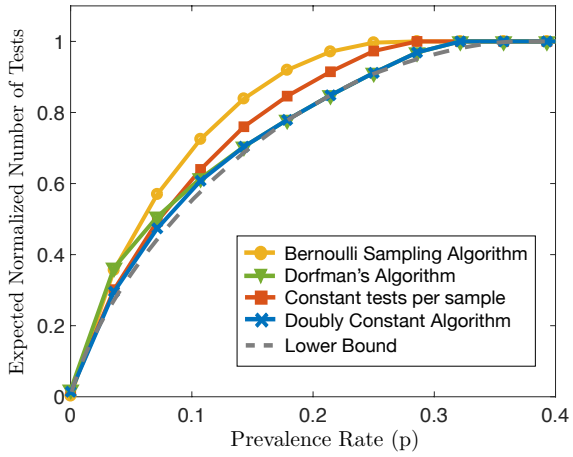

Fig. 1: Comparison of different Two-stage pooling testing algorithms. The optimal pooling parameters used in the generation of this figure are $s^* \approx 1/\sqrt{p}$ for Dorfman's and $\pi = \min\{1, 1/Np\}$ for Bernoulli. The parameters for constant tests per sample and doubly constant are numerically optimized to minimize the respective expected normalized numbers of tests. Doubly constant algorithm achieves the best performance, and is close to the lower bound in Lemma 1.

in this paper. We define $T(p, \psi_p^{\text{T.S.}}) = (T_1 + T_2)/N$ as the expected normalized number of tests using two-stage group testing. Clearly, T_2 critically depends on the effectiveness of the first stage at identifying negative samples. Therefore, we need to carefully design the pooling strategy $\psi_p^{\text{T.S.}}$ so that the total number of tests is minimized. We define the minimum expected normalized number of tests achieved by using optimal two-stage group testing $\psi_p^{\text{T.S. opt}}$ as

$$T^{\text{T.S. opt}}(p) = \min_{\text{all possible } \psi_p^{\text{T.S.}}} T(p, \psi_p^{\text{T.S.}}) = T(p, \psi_p^{\text{T.S. opt}}). \quad (6)$$

Different pooling algorithms have been considered in [4] for pooling samples in the first stage, namely, Dorfman's [1], Bernoulli sampling, constant tests per sample, and doubly constant [5]. We note that while two-stage decoding algorithm may also be referred as Dorfman's algorithm in the literature, we reserve the term *Dorfman's algorithm* for the pooling strategy in [1] throughout the paper. We next summarize the results of [4].

• **Dorfman's Algorithm**: All N samples are partitioned into N/s disjoint pools, each of size s. Each pool is then tested individually. As a function of the prevalence rate p and the pool size s, the expected normalized number of tests can be expressed as

$$T(p, \psi_p^{\text{Dorfman}} = s) = \frac{1}{s} + 1 - (1-p)^s, \quad (7)$$

where optimal s^* can be approximated as $s^* \approx 1/\sqrt{p}$.

• **Bernoulli Sampling Algorithm**: In Stage 1, each sample participates in each of the T_1 tests with probability $\pi = \min\{1,\ 1/(Np)\}$. The expected normalized number of tests can be expressed as

$$T(p, \psi_p^{\text{Bernoulli}} = \pi) = \begin{cases} p + ep\ln\left(\frac{1-p}{p}\right), & \pi = 1/(Np) \\ \frac{1}{N} + 1 - (1-p)^N, & \pi = 1. \end{cases} \quad (8)$$

For the case when $\pi = 1$, there is no need for multiple tests in Stage 1, i.e., $T_1 = 1$. The algorithm is then equivalent to Dorfman's algorithm with $s = N$.

• **Constant Tests per Sample (CTS)**: In this scheme, each sample is only allowed to be in at most a constant r number of tests. Since the second stage is individual testing, that leaves us with $r-1$ tests per sample in the first stage. Stage 1 is divided into $r-1$ rounds, with $T_1/(r-1)$ pooled tests in each round. In every round, each sample participates in one of the $T_1/(r-1)$ pools selected uniformly at random. This gives the average pool size as $\bar{s} = N(r-1)/T_1$. The expected normalized number of tests is given as

$$T(p, \psi_p^{\text{CTS}} = (r, \bar{s})) = \frac{r-1}{\bar{s}} + p + (1-p)(1 - e^{-p\bar{s}})^{r-1}, \quad (9)$$

where the optimal values of r and $\bar{s}$ can be solved numerically to minimize $T(p, \psi_p^{\text{CTS}} = (r, \bar{s}))$.

• **Doubly Constant Algorithm**: In addition to restricting each sample to participate in at most r tests, the pool size is limited to exactly s samples per pool. Similar to constant tests per sample algorithm, this can be done by dividing Stage 1 into $r-1$ rounds. However, in every round, the samples are equally partitioned into $T_1/(r-1)$ pools. The expected normalized number of tests is given as [5],

$$\begin{aligned} &T(p, \psi_p^{\text{Doubly}} = (r, s)) \\ =&\frac{r-1}{s} + p + (1-p)(1 - (1-p)^{s-1})^{r-1}, \end{aligned} \quad (10)$$

where the optimal value of r and s can be solved numerically to minimize $T(p, \psi_p^{\text{Doubly}} = (r, s))$. Intuitively, the first stage of Doubly constant is equivalent to performing the first stage of Dorfman's $r-1$ times with random permutations of samples. Thus, T_1 requires $(r-1)/s$ tests after normalization. In the second stage, individual tests are performed on all positive samples and negative samples that are not in any negative pools.

In addition to studying and comparing two-stage pooling strategies, a lower bound is derived in [4], stated in the following lemma.

Lemma 1 (Lower Bound for Two-stage Group Testing [4])**.** *For conservative two-stage group testing, the expected normalized number of tests is lower bounded as,*

1. $T^{\text{T.S. opt.}}(p) \geq 1$ *for* $p \geq 0.382$;
2. $T^{\text{T.S. opt.}}(p) \geq \frac{1}{g(p)}(\ln g(p) + 1)$ *for* $p < 0.171$;
3. $T^{\text{T.S. opt.}}(p) \geq p + \frac{1}{f(p)}(\ln((1-p)f(p)) + 1)$, *otherwise*,

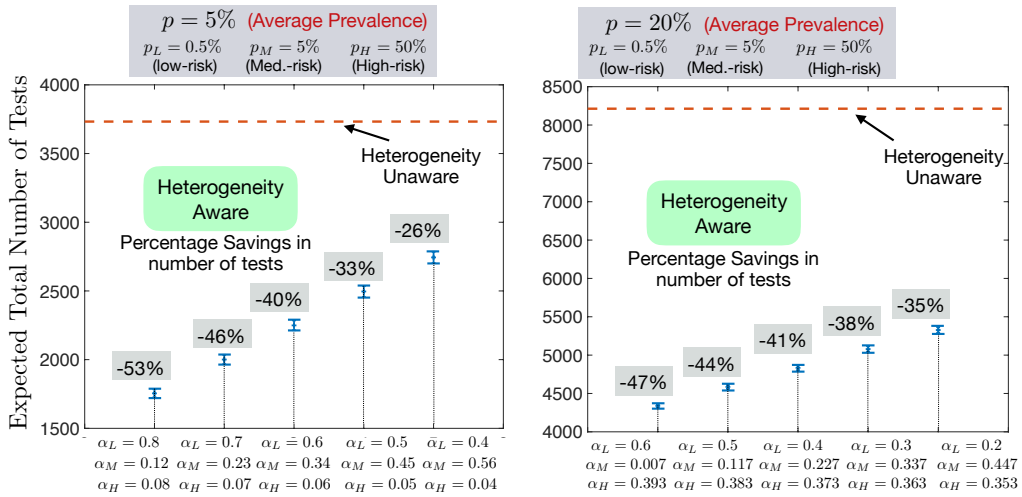

Fig. 2: Average number of tests using heterogeneity aware two-stage group testing for a total of $N = 10000$ samples with optimal pooling parameters, and two values of average prevalence rate $p = 0.05$ and $p = 0.2$. Figure 5 shows the optimal pooling parameters for each cluster found numerically. Reduction in number of tests is shown compared to heterogeneity unaware scheme for different heterogeneity profiles, determined by varying cluster ratios.

where

$$f(p) = \max_{w=2,3,\ldots} \{-w \ln(1-(1-p)^{w-1})\},$$
$$g(p) = \max_{w=2,3,\ldots} \{-w \ln(1-(1-p)^{w})\},$$

where w is the pool size.

Bound 1 is a universal bound from [46] that can be applied to any group testing algorithm. The idea behind Bound 2 and 3 is as follows. Consider any conservative two-stage group testing algorithm and assume that the number of tests in Stage 1 T_1 is fixed. The goal becomes to derive lower bounds on the expected number of tests in Stage 2 $E[T_2]$ as a function of T_1. For Bound 2, the authors in [4] count every sample that solely appears in positive pools. For bound 3, all positive samples and all negative samples that only appear in positive pools are counted. Essentially, one works with marginal probability that a sample cannot be classified and the other works with conditional probabilities that a sample cannot be classified given the events that the true status of that particular sample is positive or negative during Stage 1. Once the bounds for $E[T_2]$ are derived, one can then optimize each bound over T_1. From $f(p)$ and $g(p)$, we can see that they are determined by the pool size w and the prevalence rate p. However, for a population with heterogeneous samples, one needs to keep track of the number of samples from each cluster and the contributions of their respective local prevalence rates to the lower bounds. Therefore, the lower bounds in Lemma 1 cannot be directly applied to our setting.

The results in [4] showed (also demonstrated in Figure 1) that doubly constant pooling algorithm achieves the lowest expected normalized number of tests and is very close to the lower bound in Lemma 1. In addition, practically desirable aspects for cases when there are underlying pooling constraints in terms of pool size and maximum number of tests per each sample are incorporated in doubly constant algorithm. Henceforth, for the scope of this paper, we focus on the analysis of *Doubly constant algorithm* and showing the benefits of leveraging knowledge about heterogeneity across samples.

IV. Main Results and Discussions

In this work, we provide guarantees on the benefits of utilizing the heterogeneity knowledge using two-stage group testing. Figure 2 shows numerical results for the case of $C = 3$ clusters: low, medium and high risk, with local prevalence rates: $p_L = 0.005$, $p_M = 0.05$ and $p_H = 0.5$, respectively. We consider two cases for average prevalence rate $p = 0.05$ and $p = 0.2$. We compute the number of tests achieved by using a heterogeneity aware scheme, i.e., applying two-stage doubly constant group testing over the population clusters independently. We note that for the high-risk cluster, the optimal pooling parameters (r, s) are in fact both 1, i.e., equivalent to individual testing, as shown in the next Section.

Figure 2 shows the expected normalized number of tests as well as the error bar for various population fraction values, i.e., $(\alpha_L, \alpha_M, \alpha_H)$. We can see the improvement of applying the heterogeneity aware model compared to heterogeneity unaware for different heterogeneity profiles. We can also notice that larger size of low-risk patient cluster requires less number of total tests, i.e., larger α_L. For instance, for $p = 5\%$ and $(\alpha_L, \alpha_M, \alpha_H) = (0.8, 0.12, 0.08)$, the heterogeneity unaware scheme requires a total of 3733 tests for $N = 10000$ samples, and the heterogeneity aware scheme requires a total of 1754 tests, which is a 53.01% reduction. Moreover, we can see that the total number of tests is centered around the mean and that is because the number of undecided samples in stage 2, T_2, follows a Binomial distribution (see [4]).

In this section, we first analyze the doubly constant algorithm for the optimized choice of the pooling parameters (r, s) to obtain the optimized number of tests $T^*(p)$. Next, we

study the concavity of $T^*(p)$ as a function of the prevalence rate p which is a desired property for the heterogeneity aware approach to outperform the heterogeneity unaware model. Due to the complexity of obtaining $T^*(p)$ in a closed form, we present an approximation for low prevalence values which gives some theoretical insights on the concavity of $T^*(p)$.

We note that the choice of pooling algorithm is independent of how we utilize the heterogeneity knowledge in this work. We focus on Doubly constant algorithm for its superior performance.

A. Optimizing Two-stage Group Testing using Doubly Constant Algorithm

For simplicity, since we only consider doubly constant for the rest of this paper, we drop the notation for pooling strategy ψ_p^{Doubly} and directly express the expected normalized number of tests using the prevalence rate p, number of tests per sample r, and the pool size s, i.e., $T(p, \psi_p^{\text{Doubly}} = (r, s)) \triangleq T(p, r, s)$. Recall that $T(p, r, s)$ is defined as:

$$T(p,r,s) = \frac{r-1}{s} + p + (1-p)(1-(1-p)^{s-1})^{r-1}. \quad (11)$$

For practical constraints, we assume maximum $s_{\max}$ pool size and maximum $r_{\max}$ number of tests per sample over two stages. To satisfy the pooling constraints, we must have $s \leq s_{\max}$ and $r \leq r_{\max}$.

We can find the optimal pool size $s^*(p, r)$ as a function of the prevalence rate p and the number of tests per sample r by minimizing the expression in (11). Taking the derivative of (11) w.r.t. s and setting it to zero, we can obtain $s^*(p, r)$ by solving the following equation,

$$s^2(1-p)^s \ln(1-p)\big(1-(1-p)^{s-1}\big)^{r-2} + 1 = 0, \;\; s \leq s_{\max}. \quad (12)$$

However, the above equation is hard to solve in a closed form. The following Lemma (proved in Appendix A) provides an approximation for the expected normalized number of tests optimized over the parameter s.

Lemma 2. *For a small prevalence rate p, the expected normalized number of tests optimized over the pool size s, $T^*(p, r)$ can be approximated by,*

$$T^*(p,r) \approx \hat{T}(p,r) = T(p, r, \hat{s}(p,r)), \quad (13)$$

where

$$\hat{s}(p,r) = \min\left(s_{\max}, \;\left\lfloor 1 + p^{-(r-1)/r} \right\rceil\right), \quad (14)$$

is an approximation for the optimal s^*. The notation $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

In Figure 3, we compare the approximation $\hat{T}(p, r)$ in (13) to the exact $T^*(p, r)$ (obtained numerically) for different values of p and r with no constraint on the maximum pool size $s_{\max}$. We notice that $\hat{T}(p, r)$ is closer to $T^*(p, r)$ for smaller values of p. Moreover for $r = 2$, $\hat{T}(p, r)$ is a good approximation and as r increases the difference between exact and approximate values increases.

In Figure 4, we compare the optimized expected normalized number of tests $T^*(p, r)$, i.e., using the optimal value $s^*(p, r)$ computed numerically, for different values of r and no constraint on pool size $s_{\max}$. We also compare versus the lower bound on the optimal expected normalized number of tests for conservative two-stage group testing presented in [4, Theorem 5]. We notice that for different ranges of p, exactly one value of r will give the smallest expected normalized number of tests. The lower concave hull of the curves from all values of r gives the optimized expected normalized number of tests over all values of (r, s) for different values of p, denoted as $T^*(p)$. Note that $T^*(p)$ gives an upper bound on the optimal two-stage group testing, i.e., $T^{\text{T.S. opt}}(p) \leq T^*(p)$. In particular, the optimal r-values, $r^*(p)$ as well as the corresponding optimal pool size ranges, $s^*(p, r)$, for different ranges of p can be summarized in Figure 5. We present more numerical results on optimal parameters for constrained pool size with two values $s_{\max} = 16$ and 32 in Appendix C.

B. Heterogeneity Aware Two-stage Group testing

We propose the *heterogeneity aware* approach in which two-stage group testing is done separately over C different clusters. For every cluster c, we use the optimal pooling parameters given in Figure 5 according to the local prevalence p_c. The total number of tests for this scheme is the sum of numbers of tests of all C clusters. Thus, the expected normalized number of tests is $\sum_{c=1}^{C} \alpha_c T^*(p_c)$. Heterogeneity awareness helps in reducing expected normalized number of tests provided that the function $T^*(p)$ is concave, i.e.,

$$\sum_{c=1}^{C} \alpha_c T^*(p_c) \leq T^*\left(\sum_{c=1}^{C} \alpha_c p_c\right) = T^*(p).$$

Figure 4 suggests that the exact $T^*(p)$ is concave, which shows the potential of applying heterogeneity aware schemes to reduce expected normalized number of required tests. Since $T^*(p)$ is hard to solve in a closed form, we are unable to prove its concavity analytically. Lemma 2 suggests that $\hat{T}(p, r)$ is a good approximation for $T^*(p, r)$ when p is small. In the following Lemma (proved in Appendix B), we give some theoretical insights by proving that $\hat{T}(p, r)$ gives some concavity guarantees for small values of p.

Lemma 3. *For small values of prevalence rate p, we can further approximate $\hat{T}(p, r)$ as follows,*

$$\hat{T}(p,r) \approx \frac{(r-1)p^{(r-1)/r}}{1+p^{(r-1)/r}} + p + (1-p)p^{(r-1)/r},$$

where the new approximation is concave.

To understand the performance of our proposed scheme, we derive a lower bound for conservative two-stage heterogeneity aware group testing. The lower bound is obtained by first bounding the expected number of tests in Stage 2, T_2, which is a function of the number of tests in Stage 1, T_1. The bound for the minimum normalized number of tests is then minimized over T_1. Our result is summarized in the following Theorem (proved in Appendix D).

Theorem 1. *For a two-stage group testing problem with a population of N samples that are categorized into C*

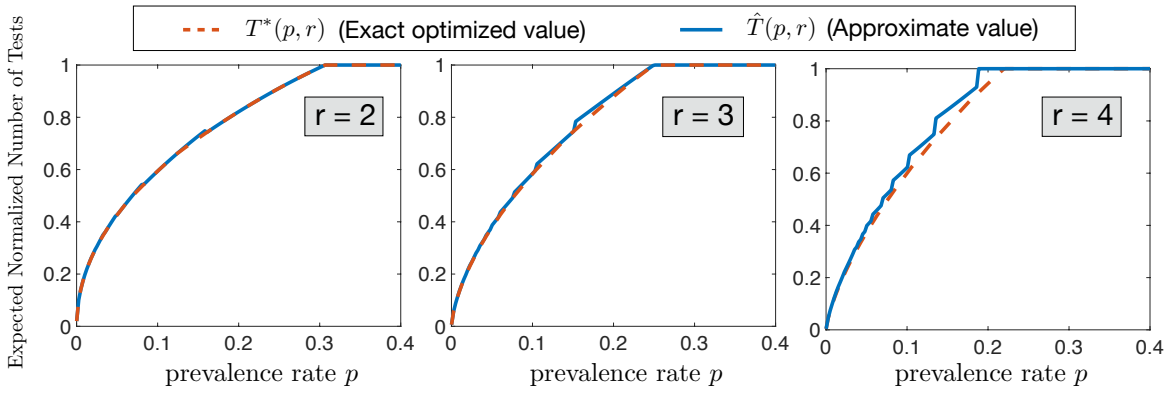

Fig. 3: Exact vs approximate value of $\bar{T}^*(p,r)$ versus p for different values of r.

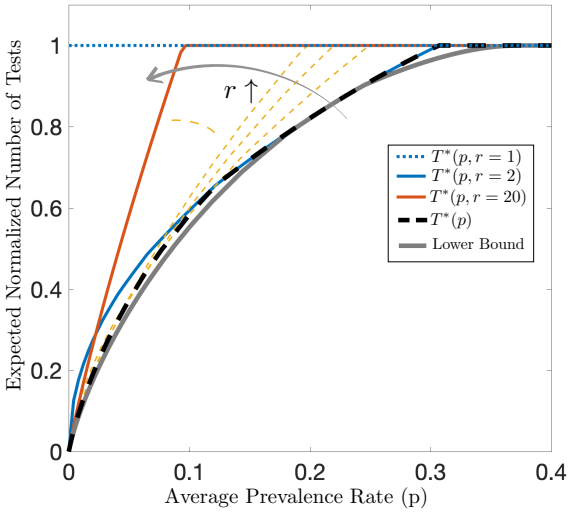

Fig. 4: $T^*(p,r)$ versus the lower bound as a function of the prevalence p for different values of r.

clusters, with population fraction $\overline{\alpha}$ and local prevalence $\overline{p}$, the minimum normalized tests needed using conservative model is lower bounded by the following two bounds,

$$1.\ T^{T.S.\ opt}(\overline{p},\overline{\alpha}) \geq \frac{1}{f(s_{\max},\overline{p},\overline{\alpha})}\left(1+\ln f(s_{\max},\overline{p},\overline{\alpha})\right),$$

$$2.\ T^{T.S.\ opt}(\overline{p},\overline{\alpha}) \geq \sum_{c=1}^{C} p_c\alpha_c + \frac{1}{g(s_{\max},\overline{p},\overline{\alpha})}\left(\sum_{c=1}^{C}\alpha_c\ln(1-p_c)+\ln g(s_{\max},\overline{p},\overline{\alpha})+1\right),$$

where $f(s_{\max},\overline{p},\overline{\alpha})$ and $g(s_{\max},\overline{p},\overline{\alpha})$ are defined in (15) *and* (16)*. $s_{\max}$ denotes the maximum allowed pool size. The ranges of average prevalence rate p that these bounds dominate depend on $\overline{\alpha}$ and $\overline{p}$. The final lower bound can be obtained by taking the maximum of the two bounds above.*

The idea behind the bounds in Theorem 1 is similar to that of Lemma 1. However, as mentioned above, we need to keep track of the number of samples from each cluster in each pool since the probability of a pool being positive requires us to know the exact composition of the pool. This is particular important for Bound 2 in Theorem 1 due to the fact that we are working with the conditional probability. We need to know which cluster that the sample we conditioned on comes from. We remark here that the lower bounds in Lemma 1 are obtained under the assumption of a homogeneous population with i.i.d. samples. Therefore, Lemma 1 is not directly applicable to our heterogeneous population model, where samples are only i.i.d. within each cluster. In fact, the second and third bounds in Lemma 1 are special cases of Bounds 1 and 2 in Theorem 1 when $p_c = p$ for $c \in [1:C]$ and $s_{\max} = N$. Moreover, to add practicality, our lower bound in Theorem 1 imposes the pool size constraint. However, the effect of having constrained maximum number of tests per sample $r_{\max}$ is not reflected in our lower bound and is a subject for future research.

V. Numerical Simulations and Comparisons

In this section, we present numerical simulation results to show (a) the benefit of exploiting the heterogeneity knowledge; (b) comparison between different pooling methods with/without heterogeneity knowledge; (c) comparison between heterogeneity aware and unaware schemes with the corresponding lower bounds; (d) the impact of constraining the pooling parameters, i.e., maximum pool size and maximum tests per sample $r_{\max}$; (e) impact of the accuracy of prevalence estimation; and (f) impact of the number of clusters and fineness of heterogeneity knowledge.

For our simulations, we assume a heterogeneity model such that the population is divided into $C = 3$ clusters. For ease of presentation, we assume the prevalence in clusters (p_L, p_M, p_H) scales with the average prevalence p, e.g., $(\frac{p_L}{p}, \frac{p_M}{p}, \frac{p_H}{p}) = (c_L, c_M, c_H)$. The constant vector (c_L, c_M, c_H) indicates the level of heterogeneity. We also pick values $(\alpha_L, \alpha_M, \alpha_H)$ that satisfies $\alpha_L + \alpha_M + \alpha_H = 1$, $0 \leq \alpha_L, \alpha_M, \alpha_H \leq 1$ and $\alpha_L c_L + \alpha_M c_M + \alpha_H c_H = 1$ where the latter constraint follows directly from (1). In particular, for low (medium and high, respectively) heterogeneity, we use population fraction values $(\alpha_L, \alpha_M, \alpha_H) = (0.6, 0.22, 0.18)$ $((0.6, 0.22, 0.18)$ and $(0.6, 0.29, 0.11)$, respectively). We assume that the population fraction values follow the trend

Prevalence range p	Tests per sample $r^*(p)$	Pool size $s^*(p,r)$	Required tests $T^*(p)$
$[0.307, 1]$	1	1	1
$[0.121, 0.307)$	2	$[3:4]$	$[0.654, 1]$
$[0.064, 0.121)$	3	$[6:8]$	$[0.440, 0.654)$
$[0.036, 0.064)$	4	$[11:16]$	$[0.294, 0.440)$
$[0.019, 0.036)$	5	$[21:33]$	$[0.181, 0.294)$
$[0.01, 0.019)$	6	$[40:64]$	$[0.110, 0.181)$
$[0.005, 0.01)$	7	$[75:126]$	$[0.064, 0.110)$
$[0, 0.005)$	> 8	> 145	< 0.064

Fig. 5: Optimal parameters (r^*, s^*) for doubly constant pooling and corresponding normalized tests per sample, for different prevalence rates p.

$\alpha_L \geq \alpha_M \geq \alpha_H$. This assumption can be justified, for instance, in a pandemic scenario as follows: we expect a small group of the population who are front-line workers such as healthcare workers to have a higher chance of being infected. There may also be other essential workers who need to be in contact with other people, forming the medium-risk group, while a large fraction of population (low-risk) is able to socially distance more effectively at home, and may have a lower chance of being infected. We note that one way to quantify the level of heterogeneity is by evaluating the following function,

$$H_{\text{Level}} = |c_M - c_L| + |c_H - c_M|.$$

In the following comparisons, we refer the vector $(c_L, c_M, c_H) = (0.7, 1.0, 2.0)$ as low heterogeneity, the vector $(c_L, c_M, c_H) = (0.1, 1.0, 4.0)$ as medium heterogeneity and $(c_L, c_M, c_H) = (0.05, 1.0, 6.0)$ as high heterogeneity.

• **Comparison between pooling algorithms:** In Figure 6, we compare four different pooling algorithms, namely Bernoulli, constant tests per sample, Dorfman's, and doubly constant. We assume there are no pooling constraints. We further assume 3 values for average prevalence rate, $p = 5\%$, $p = 10\%$, and $p = 15\%$. It can be seen that, except doubly constant, the performances of other pooling algorithms changes for different heterogeneity profiles. For instance, for low heterogeneity profile, Dorfman always outperforms Bernoulli, and for high heterogeneity, the opposite is true. Also, constant tests per sample always outperforms Bernoulli algorithm. However, the performance of doubly constant is consistently the best among all four pooling algorithms for different heterogeneity levels and different prevalence rates.

• **Achieved performance versus lower bounds:** Next, we show how far the achieved performance is from optimality for different population prevalence rates. In Figure 7, we compare the average number of tests per sample between heterogeneity unaware scheme and low, medium and high heterogeneity aware schemes versus the corresponding lower bounds obtained from Theorem 1 and [4, Theorem 5]. We use the optimal value of pool size $s^*(p,r)$ computed numerically using (12). Moreover, we use the optimal value of number of tests per sample, $r^*(p)$, using Figure 5. We assume there are no pooling constraints.

It can be seen that the heterogeneity knowledge reduces the average number of tests per sample even when the heterogeneity is low. As the heterogeneity level increases, we see a larger reduction in the number of tests. Note that the curves for medium and high heterogeneity are cut off due to the fact that the local prevalence rates will not be a valid probability when the average prevalence rates exceed certain values for their respective heterogeneity profile. For example, high heterogeneity has a $c_H = 6$ for the high-risk cluster. Therefore, the average prevalence rate p cannot exceed $p_H/c_H \leq 1/6$. We also note that for the curve for low heterogeneity, Bound 2 dominates for $p < 0.351$, and Bound 1 dominates for $0.351 \leq p \leq 0.5$.

• **Constrained Pool Size and Maximum Number of Tests per Sample:** In Figure 8, we show the impact of introducing a constraint on maximum number of tests per sample, $r_{\max}$. We use the optimal value of number of tests per sample, $r^*(p)$, using Figure 5 such that r does not exceed a predetermined values, $r_{\max} = 2, 4, 6, 8$, and 10, i.e., $r = \min(r_{\max}, r^*(p))$. It can be seen that the reductions in average total tests diminish as the restriction on number of tests per sample loosens. This result suggests that our two-stage testing algorithm requires only small maximum number of tests per sample, $r_{\max}$. In particular, our analysis shows that increasing $r_{\max}$ beyond 4 does not have *a significant* improvement in reducing number of required tests.

In Figure 9, we show the impact of imposing the maximum pool size constraint. In particular, we consider three cases for maximum pooling constraint namely, unbounded, $s_{\max} = 32$ and $r_{\max} = 16$ and compare the required number of tests versus the corresponding lower bounds. For each population, we use the optimal pooling parameters given by the Figures 5, 15, and 16. This result shows that we still can achieve reduction in number of required tests using group testing while considering practical constraints on maximum pool size.

• **Accuracy of Prevalence Estimation:** In Figure 10, we show the impact of prevalence estimation error on the expected normalized number of tests for high heterogeneity. The

$$f(s_{\max}, \overline{p}, \overline{\alpha}) = - \min_{\substack{w \in [2:s_{\max}],\ w^{(c)} \in [0:\min\{w, \alpha_c N\}] \\ \sum_{c=1}^{C} w^{(c)} = w}} \left\{ w \ln \left(1 - \prod_{c \in [1:C]} (1 - p_c)^{w^{(c)}} \right) \right\}. \tag{15}$$

$$g(s_{\max}, \overline{p}, \overline{\alpha}) = - \min_{\substack{w \in [2:s_{\max}],\ w^{(c)} \in [0:\min\{w, \alpha_c N\}] \\ \sum_{c=1}^{C} w^{(c)} = w}} \left\{ \sum_{c=1}^{C} w^{(c)} \ln \left(1 - (1 - p_c)^{w^{(c)}-1} (1 - \min_{i \in [1:C]} p_i)^{w - w^{(c)}} \right) \right\}. \tag{16}$$

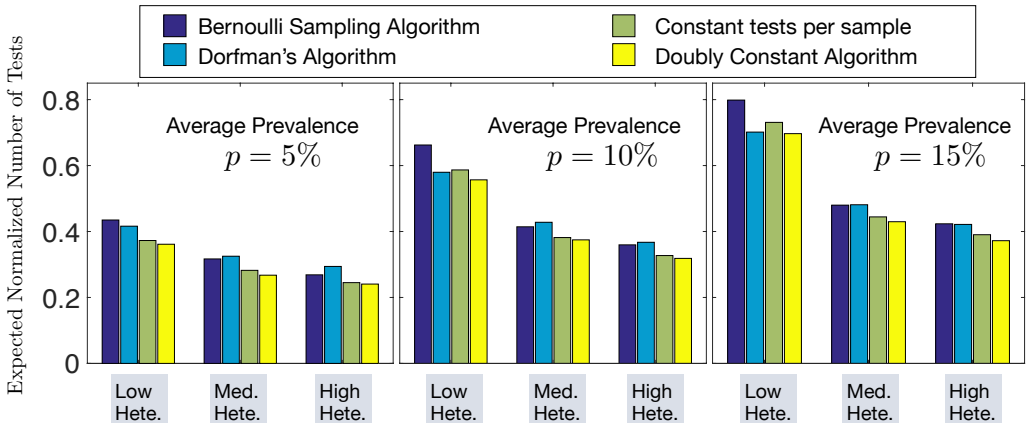

Fig. 6: Comparison of heterogeneity aware pooling algorithms used in stage 1, with different heterogeneity levels and average prevalence values, $p = 5\%$, $p = 10\%$, and $p = 15\%$. While different algorithms give varying performance for different heterogeneity profiles, doubly constant is consistently achieving best performance.

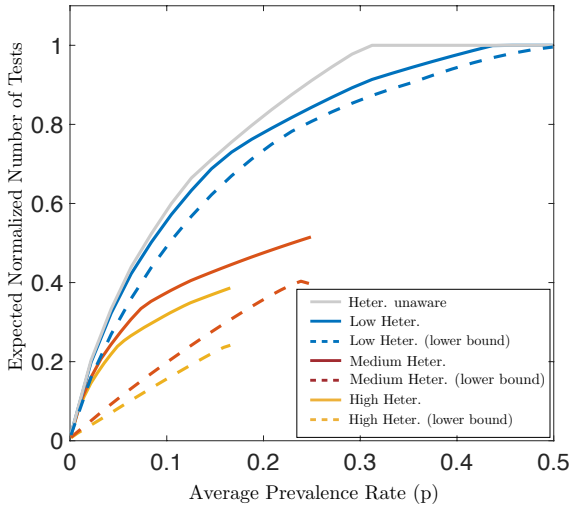

Fig. 7: Comparison between Heterogeneity unaware, low, medium and high heterogeneity schemes with optimal pool size $s^*(p, r)$ and tests per sample $r^*(p)$. Lower bounds are also shown as a function of the heterogeneity profile.

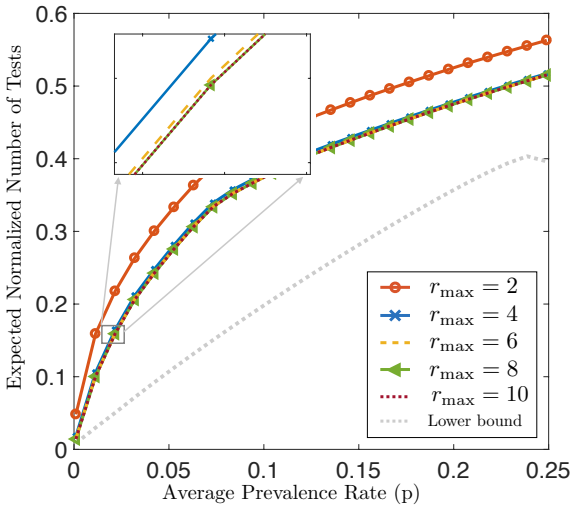

Fig. 8: Average number of tests per sample with medium heterogeneity with optimal pool size $s^*(p, r)$ and tests per sample $r^*(p)$ such that the maximum number of tests per sample, $r_{\max}$, is restricted to three values $2, 4, 6, 8$ and 10.

pooling parameters for each cluster can be found in Figure 5. It can be seen that the expected normalized number of tests is minimized whenever the estimation is correct (i.e., along the diagonal in the matrix). In addition, we can see that underestimation of prevalence rate is more severe than overestimation in terms of the total number of tests. This is due to the fact that when we underestimate p, we will group more samples in a pool during Stage 1. However, the fact that more positive samples are presented than expected results in more positive pools. Hence, we need to perform more individual testing in Stage 2. For these values of true p, estimating them to be 0.07 and 0.08 does not change the optimal pooling parameters since the local prevalence rates for each cluster when $p = 0.07$ and 0.08 fall in the same range in Figure 5, e.g., $p_M = 0.07$ for $p = 0.07$ and 0.08 for $p = 0.08$, and both results in $r = 3$ and $s \in [6 : 8]$.

- **Impact of the number of clusters C and fineness of heterogeneity knowledge:** With finer heterogeneity knowledge, one can form clusters more intelligently. For example, one can form more clusters, or have various choices of prevalence threshold for the clusters for a fixed C. In Figure 12, we show the impact of varying C. Specifically, we look at cases where $C = 1, 3$ and 6, with $C = 1$ being heterogeneity unaware and knowing only the prevalence rate of the entire population, $C = 3$ being heterogeneity aware and having coarser knowledge of prevalence rates, and $C = 6$ being heterogeneity aware and having finer knowledge of prevalence rates. For $C = 6$, the population fraction values are $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = (0.3, 0.25, 0.2, 0.15, 0.07, 0.03)$ and the prevalence rates for the clusters are $(p_1, p_2, p_3, p_4, p_5, p_6) = (0.005, 0.04, 0.1, 0.16, 0.24, 0.42)$. For $C = 3$, we form the low (medium/high) risk cluster using the populations of cluster $1, 2$ and 3 (cluster 4 and

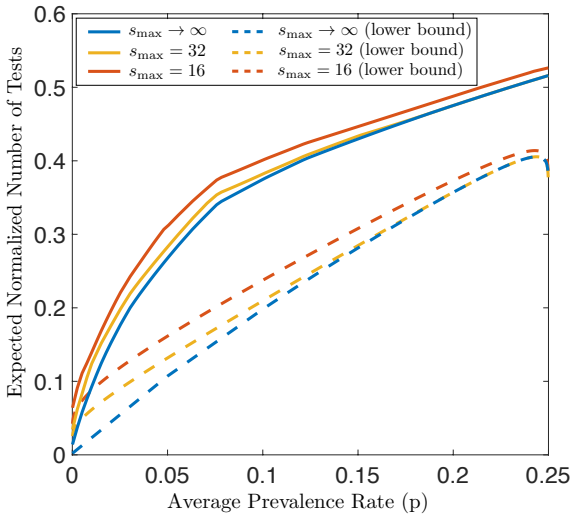

Fig. 9: Average number of tests per sample compared to the lower bounds with medium heterogeneity and optimal pooling parameters with the constraint on maximum pool size, $s_{\max}$ taking the values, unbounded, $s_{\max} = 32$ and $s_{\max} = 16$.

		True p			
		0.05	0.06	0.07	0.08
Estimated	0.05	0.2406	0.2676	0.2921	0.3139
	0.06	0.2417	0.2595	0.2765	0.2937
	0.07	0.2455	0.2604	0.2756	0.2906
	0.08	0.2455	0.2604	0.2756	0.2906

Fig. 10: A confusion matrix for the effect of prevalence estimation error on the expected normalized number of tests with high heterogeneity, i.e., $(c_L, c_M, c_H) = (0.05, 1.0, 6.0)$.

5/cluster 6) from $C = 6$ case, thus, we have $(\alpha_L, \alpha_M, \alpha_H) = (\alpha_1 + \alpha_2 + \alpha_3, \alpha_4 + \alpha_5, \alpha_6) = (0.75, 0.22, 0.03)$ and $(p_L, p_M, p_H) = (0.042, 0.1855, 0.42)$ by taking the average of the weighted prevalence rates of the corresponding clusters from $C = 6$ case. Finally, for $C = 1$, we have a prevalence rate of 0.0849 for the entire population (see the visualization of the cluster compositions in Figure 11). It can be seen that heterogeneity aware doubly constant with $C = 6$ outperforms the cases with $C = 1$ and 3. However, we can also see that at the other extreme where $C = N$, one recovers individual testing. This indicates that C needs to be chosen carefully to minimize the expected normalized number of tests. How to pick C and form each cluster is an interesting future work.

VI. Conclusions

In this paper, we consider the problem of group testing for a heterogeneous population, where the population can be divided into several clusters with local different prevalence rates. We showed the benefits of applying conservative two-stage group testing algorithm independently to each cluster. We showed that the efficiency of heterogeneity-aware group testing algorithm can be improved due to the concavity of the expected normalized number of tests as a function of the prevalence. A lower bound on the required number of tests was

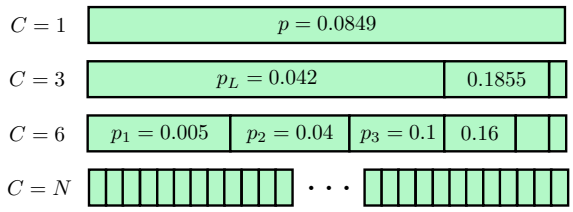

Fig. 11: The visualization of cluster compositions for $C = 1, 3, 6$ and N with the prevalence rate of each cluster.

No. of Clusters	$C = 1$	$C = 3$	$C = 6$	$C = N$
Exp. Normalized No. of Tests	0.5265	0.5177	0.4760	1

Fig. 12: Expected normalized number of tests for three different fineness of heterogeneity knowledge, where heterogeneity unaware doubly constant is used for $C = 1$ and heterogeneity aware doubly constant is used for $C = 3$ and 6.

derived based on the heterogeneity profile as well as a practical constraint on the pool size. Our numerical results confirmed that heterogeneity knowledge indeed provides significant improvement for group testing performance compared to the case where heterogeneity is unknown, i.e., assuming i.i.d. samples. Interesting future directions include studying the benefits of heterogeneity aware schemes for other model assumptions such as, noisy pooled testing and non-conservative testing models where pooled testing can be useful to detect both positive and negative samples.

Noisy group testing has been studied in [30], [31], [47]–[50]. In the noisy setting, the pooled tests could yield noisy outcomes, which subsequently lead to incorrect classification. In practice, noises can be results of dilution or inherently inaccurate testing. While noisy group testing is relatively well-understood with homogeneous population, there are few works that assume heterogeneous population. Understanding the impact of noise for pooled testing in a heterogeneous population is an interesting direction for future work.

In a recent interesting work [33], it was shown that by leveraging quantitative information from the test outcomes (as opposed to binary outcomes), one can further reduce the number of tests for two-stage adaptive testing algorithms. Another interesting direction would be to jointly use quantitative information from pooled tests (as in [33]) and prevalence heterogeneity across samples.

References

[1] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[2] M. Aldridge, "Individual testing is optimal for nonadaptive group testing in the linear regime," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2058–2061, 2018.

[3] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 3019–3035, 2014.

[4] M. Aldridge, "Conservative two-stage group testing," *CoRR*, vol. abs/2005.06617, 2020. [Online]. Available: http://arxiv.org/abs/2005.06617

[5] A. Z. Broder and R. Kumar, "A note on double pooling tests," *CoRR*, vol. abs/2004.01684, 2020. [Online]. Available: http://arxiv.org/abs/2004.01684

[6] R. Hanel and S. Thurner, "Boosting test-efficiency by pooled testing strategies for SARS-CoV-2," *CoRR*, vol. abs/2003.09944, 2020. [Online]. Available: http://arxiv.org/abs/2003.09944

[7] D. Aragón-Caqueo, J. Fernández-Salinas, and D. Laroze, "Optimization of group size in pool testing strategy for SARS-CoV-2: A simple mathematical model," *Journal of Medical Virology*, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25929

[8] M. B. Gongalsky, "Early detection of superspreaders by mass group pool testing can mitigate COVID-19 pandemic," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/27/2020.04.22.20076166

[9] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: an information theory perspective," *CoRR*, vol. abs/1902.06002, 2019. [Online]. Available: http://arxiv.org/abs/1902.06002

[10] C. Gollier and O. Gossner, "Group testing against COVID-19," *COVID Economics*, vol. 2, 2020.

[11] S. R. Mehta, V. T. Nguyen, G. Osorio, S. Little, and D. M. Smith, "Evaluation of pooled rapid HIV antibody screening of patients admitted to a San Diego Hospital," *Journal of virological methods*, vol. 174, no. 1-2, pp. 94–98, 2011.

[12] L. M. Wein and S. A. Zenios, "Pooled testing for HIV screening: capturing the dilution effect," *Operations Research*, vol. 44, no. 4, pp. 543–569, 1996.

[13] Food and Drug Administration, *Revised Recommendations for Reducing the Risk of Zika Virus Transmission by Blood and Blood Components*. U.S. Department of Health and Human Services, 2018.

[14] L. Novack, E. Shinar, J. Safi, H. Soliman, A. Yaari, N. Galai, J. S. Pliskin, and B. Sarov, "Evaluation of pooled screening for anti-HCV in two blood services set-ups," *Tropical Medicine & International Health*, vol. 12, no. 3, pp. 415–421, 2007.

[15] S. M. Taylor, J. J. Juliano, P. A. Trottman, J. B. Griffin, S. H. Landis, P. Kitsa, A. K. Tshefu, and S. R. Meshnick, "High-Throughput Pooling and Real-Time PCR-Based Strategy for Malaria Detection," *Journal of Clinical Microbiology*, vol. 48, no. 2, pp. 512–519, 2010. [Online]. Available: https://jcm.asm.org/content/48/2/512

[16] C. A. Muñoz-Zanzi, W. O. Johnson, M. C. Thurmond, and S. K. Hietala, "Pooled-Sample Testing as a Herd-Screening Tool for Detection of Bovine Viral Diarrhea Virus Persistently Infected Cattle," *Journal of Veterinary Diagnostic Investigation*, vol. 12, no. 3, pp. 195–203, 2000, pMID: 10826831. [Online]. Available: https://doi.org/10.1177/104063870001200301

[17] S. Ghosh, R. Agarwal, M. A. Rehan, S. Pathak, P. Agrawal, Y. Gupta, S. Consul, N. Gupta, R. Goyal, A. Rajwade *et al.*, "A Compressed Sensing Approach to Group-testing for COVID-19 Detection," *CoRR*, vol. abs/2005.07895, 2020. [Online]. Available: http://arxiv.org/abs/2005.07895

[18] C. M. Verdun, T. Fuchs, P. Harar, D. Elbrächter, D. S. Fischer, J. Berner, P. Grohs, F. J. Theis, and F. Krahmer, "Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/13/2020.04.30.20085290

[19] J.-T. Seong, "Group Testing-Based Robust Algorithm for Diagnosis of COVID-19," *Diagnostics*, vol. 10, no. 6, 2020. [Online]. Available: https://www.mdpi.com/2075-4418/10/6/396

[20] B. Abdalhamid, C. R. Bilder, E. L. McCutchen, S. H. Hinrichs, S. A. Koepsell, and P. C. Iwen, "Assessment of specimen pooling to conserve SARS CoV-2 testing resources," *American journal of clinical pathology*, vol. 153, no. 6, pp. 715–718, 2020.

[21] R. Ben-Ami, A. Klochendler, M. Seidel, T. Sido, O. Gurel-Gurevich, M. Yassour, E. Meshorer, G. Benedek, I. Fogel, E. Oiknine-Djian *et al.*, "Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection," *Clinical Microbiology and Infection*, 2020.

[22] N. Shental, S. Levy, S. Skorniakov, V. Wuvshet, Y. Shemer-Avni, A. Porgador, and T. Hertz, "Efficient high throughput SARS-CoV-2 testing to detect asymptomatic carriers," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/20/2020.04.14.20064618

[23] I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony, "Evaluation of COVID-19 RT-qPCR test in multi-sample pools," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/27/2020.03.26.20039438

[24] T. Berger, N. Mehravari, D. Towsley, and J. Wolf, "Random multiple-access communication and group testing," *IEEE Transactions on Communications*, vol. 32, no. 7, pp. 769–779, 1984.

[25] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, "Group testing and sparse signal recovery," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, 2008, pp. 1059–1063.

[26] Y. Zhou, U. Porwal, C. Zhang, H. Q. Ngo, X. Nguyen, C. Ré, and V. Govindaraju, "Parallel feature selection inspired by group testing," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 3554–3562. [Online]. Available: http://papers.nips.cc/paper/5296-parallel-feature-selection-inspired-by-group-testing.pdf

[27] B. W. Heng and J. Scarlett, "Non-adaptive group testing in the linear regime: Strong converse and approximate recovery," *CoRR*, vol. abs/2006.01325, 2020. [Online]. Available: http://arxiv.org/abs/2006.01325

[28] E. Price and J. Scarlett, "A fast binary splitting approach to non-adaptive group testing," *CoRR*, vol. abs/2006.10268, 2020. [Online]. Available: http://arxiv.org/abs/2006.10268

[29] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," *CoRR*, vol. abs/1301.7023, 2013. [Online]. Available: http://arxiv.org/abs/1301.7023

[30] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3646–3661, 2019.

[31] M. Cuturi, O. Teboul, and J.-P. Vert, "Noisy Adaptive Group Testing using Bayesian Sequential Experimental Design," *CoRR*, vol. abs/2004.12508, 2020. [Online]. Available: http://arxiv.org/abs/2004.12508

[32] M. Aldridge, "Rates of adaptive group testing in the linear regime," *CoRR*, vol. abs/1901.09687, 2019. [Online]. Available: http://arxiv.org/abs/1901.09687

[33] A. Heidarzadeh and K. R. Narayanan, "Two-Stage Adaptive Pooling with RT-qPCR for COVID-19 Screening," *CoRR*, vol. abs/2007.02695, 2020. [Online]. Available: http://arxiv.org/abs/2007.02695

[34] South Dakota Department of Health, *COVID-19: Employee Screening Questions and Guidelines*. South Dakota Department of Health, 2020.

[35] Ontario Ministry of Health, *COVID-19 Screening Tool for Long-Term Care Homes and Retirement Homes*. Ontario Ministry of Health, 2020.

[36] American Dental Hygienists' Association, *ADHA COVID-19 Patient Screening Questionnaire*. American Dental Hygienists' Association, 2020.

[37] I. Bergel, "Variable pool testing for infection spread estimation," *CoRR*, vol. abs/2004.03322, 2020. [Online]. Available: http://arxiv.org/abs/2004.03322

[38] X. M. Tu, E. Litvak, and M. Pagano, "On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening," *Biometrika*, vol. 82, no. 2, pp. 287–297, 06 1995. [Online]. Available: https://doi.org/10.1093/biomet/82.2.287

[39] W. Zhang, A. Liu, Q. Li, and P. S. Albert, "Incorporating retesting outcomes for estimation of disease prevalence," *Statistics in Medicine*, vol. 39, no. 6, pp. 687–697, 2019.

[40] C. R. Bilder, J. M. Tebbs, and P. Chen, "Informative retesting," *Journal of the American Statistical Association*, vol. 105, no. 491, pp. 942–955, 2010.

[41] M. S. Black, C. R. Bilder, and J. M. Tebbs, "Group testing in heterogeneous populations by using halving algorithms," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 61, no. 2, pp. 277–290, 2012.

[42] H. Aprahamian, E. K. Bish, and D. R. Bish, "Adaptive risk-based pooling in public health screening," *IISE Transactions*, vol. 50, no. 9, pp. 753–766, 2018.

[43] H. Aprahamian, D. R. Bish, and E. K. Bish, "Optimal risk-based group testing," *Management Science*, vol. 65, no. 9, pp. 4365–4384, 2019. [Online]. Available: https://doi.org/10.1287/mnsc.2018.3138

[44] C. S. McMahan, J. M. Tebbs, and C. R. Bilder, "Informative Dorfman Screening," *Biometrics*, vol. 68, no. 1, pp. 287–296, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01644.x

[45] P. Nikolopoulos, T. Guo, C. Fragouli, and S. Diggavi, "Community aware group testing," *CoRR*, vol. abs/2007.08111, 2020. [Online]. Available: http://arxiv.org/abs/2007.08111

[46] P. Fischer, N. Klasner, and I. Wegener, "On the cut-off point for combinatorial group testing," *Discrete Appl. Math.*, vol. 91, no. 1–3, p. 83–92, Jan. 1999. [Online]. Available: https://doi.org/10.1016/S0166-218X(98)00119-X

[47] O. Gebhard, O. Johnson, P. Loick, and M. Rolvien, "Improved bounds for noisy group testing with constant tests per item," *CoRR*, vol. abs/2007.01376, 2020. [Online]. Available: http://arxiv.org/abs/2007.01376

[48] G. Atia and V. Saligrama, "Noisy group testing: An information theoretic perspective," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2009, pp. 355–362.

[49] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2011, pp. 1832–1839.

[50] J. Scarlett and V. Cevher, "Near-optimal noisy group testing via separate decoding of items," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2311–2315.

APPENDIX A
PROOF OF LEMMA 2

In order to find an approximation for $s^*(p,r)$, we approximate the derivative of $T(p,r,s)$ in (11) as follows,

$$\begin{aligned}
&\frac{\partial T(p,r,s)}{\partial s}\\
&= -\frac{(r-1)}{s^2} - (r-1)(1-p)^s \ln(1-p)\big(1-(1-p)^{s-1}\big)^{r-2}\\
&\overset{(a)}{\approx} -\frac{r-1}{s^2} - (r-1)(1-p)^s \ln(1-p)\big(p(s-1)\big)^{r-2}\\
&\overset{(b)}{\approx} -\frac{r-1}{s^2} + (r-1)p^{r-1}(1-p)^s(s-1)^{r-2}\\
&\overset{(c)}{\approx} -\frac{r-1}{s^2} + (r-1)p^{r-1}(s-1)^{r-2} = 0, \qquad (17)
\end{aligned}$$

where the approximation in the three steps assumes that $p \ll 1$. More specifically, (a) follows from approximating $1-(1-p)^{s-1}$ with $p(s-1)$ through binomial approximation; (b) follows from the fact that $\ln(1-p) \approx -p$ for small p; and (c) follows from $(1-p)^s \approx 1$ for small p. Hence for small prevalence rate p, the optimal pool size is approximately the solution of $s^2(s-1)^{r-2} = \frac{1}{p^{r-1}}$ while satisfying the constraint $s \le s_{\max}$, which can be further approximated in a closed form as $\hat{s}(p,r) = \min(s_{\max},\ \lfloor 1+p^{-(r-1)/r} \rceil)$ by approximating s^2 as $(s-1)^2$. Using this value for the pool size, we obtain the expression in (13) for the approximate optimized average number of tests per sample, where the bound $\hat{T}(p,r) < 1$ corresponds to individual testing, i.e., $s = r = 1$.

APPENDIX B
PROOF OF LEMMA 3

$\hat{T}(p,r)$ can be approximated as follows by plugging in the $\hat{s}(p,r)$ in (14) to $\hat{T}(p,r)$,

$$\begin{aligned}
&\hat{T}(p,r)\\
&= \frac{r-1}{\lfloor 1+p^{-(r-1)/r} \rceil} + p + (1-p)\left(1-(1-p)^{\lfloor p^{-(r-1)/r} \rceil}\right)^{r-1}\\
&\overset{(a)}{\approx} \frac{r-1}{1+p^{-(r-1)/r}} + p + (1-p)\left(1-(1-p)^{p^{-(r-1)/r}}\right)^{r-1}\\
&\overset{(b)}{\approx} \frac{(r-1)p^{(r-1)/r}}{1+p^{(r-1)/r}} + p + (1-p)p^{(r-1)/r}. \qquad (18)
\end{aligned}$$

where (a) follows from removing the rounding; and (b) follows from another binomial approximation for the term $(1-p)^{p^{-(r-1)/r}}$ when $p < 1$ and $p^{1/r} \ll 1$. Therefore, the second derivative of $\hat{T}(p,r)$ is approximated as

$$\begin{aligned}
\frac{\partial^2 \hat{T}(p,r)}{\partial p^2} \approx &-\frac{\frac{r-1}{r}p^{-\frac{2}{r}}}{\left(1+p^{\frac{r-1}{r}}\right)^2}\left(2+\frac{p^{-\frac{r-1}{r}}}{r}((2r-1)p+1)\right)\\
&-\frac{r-1}{r^2}p^{-\frac{r+1}{r}}((2r-1)p+1), \qquad (19)
\end{aligned}$$

which is negative for all values of p and r, hence proving concavity.

APPENDIX C
OPTIMAL PARAMETERS WITH CONSTRAINED POOL SIZES

In Figures 13 and 14, we illustrate the effect of adding the pooling size constraint, i.e., bounded $s_{\max}$, for $s_{\max} = 16$ and $s_{\max} = 32$, respectively. We plot $T^*(p,r)$ for different values of r with the corresponding lower bounds given by [4, Theorem 5]. We also plot $T^*(p)$ by optimizing over r. Figures 15 and 16 for $s_{\max} = 16$ and $s_{\max} = 32$, respectively, shows the optimal pooling parameters $r^*(p)$ and $s^*(p,r)$, for different ranges of p.

APPENDIX D
PROOF OF THEOREM 1

In this Section, we present the proof for the lower bounds on the expected number of tests for the heterogeneous model. We start by lower bounding the expected number of tests in Stage 2, T_2. In the conservative setting, all positive samples are tested again in Stage 2 regardless of the results of Stage 1. Therefore, Stage 2 consists of all positive samples and samples that are not declared as DNs in the first stage. For simplicity, we call samples that cannot be declared as DNs hidden samples. The main difference between our bound and the bound derived in [4] is that, since the statistics of each pool now depends on the local prevalence p_c, we need to keep track of the number of samples from each cluster in each pool. In addition, we assume that each cluster c has exactly $\alpha_c N$ samples, which is used during the minimization process.

Suppose that the sample n belongs to the cluster γ_n and p_{γ_n} denotes the local prevalence of cluster γ_n. We define $\mathcal{N}_c$ as the set of all samples in cluster c, i.e., $\mathcal{N}_c = \{n | \gamma_n = c, n = 1,\dots,N\}$, and $|\mathcal{N}_c| = \alpha_c N$ for all $c \in [1:C]$. Let us also define w_t as the number of samples in pool t, and $w_t^{(c)}$ as the number of samples from cluster c that participate in pool t. Clearly, $w_t = \sum_{c=1}^{C} w_t^{(c)}$. The expected number of tests in Stage 2 can be bounded as,

$$\begin{aligned}
E(T_2) &\ge E(\text{\# of hidden samples}) = \sum_{n=1}^{N} P(H_n) \qquad (20)\\
&= \sum_{n=1}^{N} [p_{\gamma_n} P(H_n | n \text{ is positive})\\
&\qquad + (1-p_{\gamma_n}) P(H_n | n \text{ is negative})]\\
&= \sum_{n=1}^{N} p_{\gamma_n} + \sum_{n=1}^{N} (1-p_{\gamma_n}) P(H_n | n \text{ is negative}) \qquad (21)
\end{aligned}$$

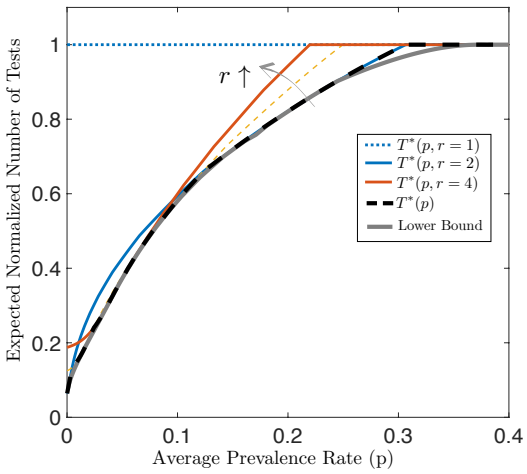

Fig. 13: Expected normalized number of tests $T^*(p, r)$ compared to the lower bound (Lemma 1) for different values of r and maximum pool size $s_{\max} = 16$.

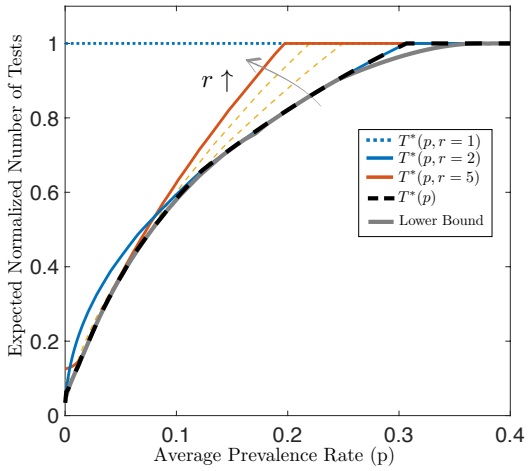

Fig. 14: Expected normalized number of tests $T^*(p, r)$ compared to the lower bound (Lemma 1) for different values of r and maximum pool size $s_{\max} = 32$.

Prevalence range p	Tests/sample $r^*(p)$	Pool size $s^*(p, r)$	Required tests $T^*(p)$
$[0.307, 1]$	1	1	1
$[0.121, 0.307)$	2	$[3:4]$	$[0.654, 1]$
$[0.064, 0.121)$	3	$[6:8]$	$[0.440, 0.654)$
$[0.025, 0.064)$	4	$[11:16]$	$[0.241, 0.440)$
$[0.005, 0.024)$	3	$[14:16]$	$[0.134, 0.241)$
$[0, 0.001)$	2	$[15:16]$	$[0.064, 0.134)$

Fig. 15: Optimal pooling parameters, (r, s) values and corresponding required tests, for different ranges of prevalence rate p and maximum pool size $s_{\max} = 16$.

Prevalence range p	Tests/sample $r^*(p)$	Pool size $s^*(p, r)$	Required tests $T^*(p)$
$[0.307, 1]$	1	1	1
$[0.121, 0.307)$	2	$[3:4]$	$[0.654, 1]$
$[0.064, 0.121)$	3	$[6:8]$	$[0.440, 0.654)$
$[0.036, 0.064)$	4	$[11:16]$	$[0.294, 0.440)$
$[0.015, 0.036)$	5	$[21:32]$	$[0.160, 0.294)$
$[0.007, 0.015)$	4	$[29:32]$	$[0.108, 0.160)$
$[0.001, 0.007)$	3	$[30:32]$	$[0.064, 0.108)$
$[0, 0.001)$	2	32	$< [0.034, 0.064)$

Fig. 16: Optimal pooling parameters, (r, s) values and corresponding required tests, for different ranges of prevalence rate p and maximum pool size $s_{\max} = 32$.

where H_n is a Bernoulli random variable denoting the event that a given sample n is hidden, and (21) follows from the fact that for conservative testing, a positive sample is always hidden, thus, $P(H_n | n \text{ is positive}) = 1$.

We note that, while (20) and (21) are equivalent, we can take different steps later on so that they produce slightly different bounds that dominate in different regime of p like the bounds in Lemma 1.

A. *Lower Bound* 1

We first start with (20). A sample is hidden when all the pools it is in are positive. Let H_{tn} denote the event a given sample n is in a positive pool t. The probability of H_{tn} is given as,

$$P(H_{tn}) = 1 - \prod_{c \in [1:C]} q_c^{w_t^{(c)}}, \tag{22}$$

where $q_c \triangleq 1 - p_c$. Therefore, we know,

$$P(H_n) = \prod_{t: x_{tn}=1} \left(1 - \prod_{c \in [1:C]} q_c^{w_t^{(c)}} \right), \tag{23}$$

where $x_{tn} = 1$ whenever sample n is in pool t and $x_{tn} = 0$ otherwise. We next define the variable $L(n, \overline{p})$ as follows:

$$L(n, \overline{p}) = \ln P(H_n) \tag{24}$$

$$= \sum_{t=1}^{T_1} x_{tn} \ln \left(1 - \prod_{c \in [1:C]} q_c^{w_t^{(c)}} \right) \tag{25}$$

Using AM-GM inequality, we can now bound (20) as follows:

$$\sum_{n=1}^{N} P(H_n) = \sum_{n=1}^{N} e^{L(n, \overline{p})}$$

$$\geq N \exp \left(\frac{1}{N} \sum_{n=1}^{N} L(n, \overline{p}) \right). \tag{26}$$

Thus, the goal is to bound the term inside the exponential. Plugging the definition of $L(n, \overline{p})$ into (26), we have the following inequalities,

$$\frac{1}{N} \sum_{n=1}^{N} L(n, \overline{p}) \tag{27}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_1} x_{tn} \ln \left(1 - \prod_{c \in [1:C]} q_c^{w_t^{(c)}} \right) \tag{28}$$

$$\stackrel{(a)}{=}\frac{1}{N}\sum_{t=1}^{T_1}\left(\sum_{n=1}^{N}x_{tn}\right)\ln\left(1-\prod_{c\in[1:C]}q_c^{w_t^{(c)}}\right) \tag{29}$$

$$\stackrel{(b)}{=}\frac{1}{N}\sum_{t=1}^{T_1}w_t\ln\left(1-\prod_{c\in[1:C]}q_c^{w_t^{(c)}}\right) \tag{30}$$

$$\stackrel{(c)}{\geq}\frac{T_1}{N}\min_{\substack{w\in[2:s_{\max}]\\ w^{(c)}\in[0:\min\{w,\alpha_cN\}]\\ \sum_{c=1}^{C}w^{(c)}=w}}\left\{w\ln\left(1-\prod_{c\in[1:C]}q_c^{w^{(c)}}\right)\right\} \tag{31}$$

$$=-\lambda f(s_{\max},\overline{p},\overline{\alpha}), \tag{32}$$

where (a) follows from switching the order of summations; (b) follows from the definition of w_t; and (c) follows from bounding each term in the summation with the minimum value; $f(s_{\max},\overline{p},\overline{\alpha})$ is defined in (15) and $\lambda = T_1/N$. Therefore, the expected normalized number of tests can be bounded as,

$$\begin{aligned}T^{\text{T.S. opt}}(\overline{p},\overline{\alpha}) &\geq \frac{T_1}{N}+\frac{E\left(T_2\right)}{N}\\ &\geq \lambda+\exp\left(-\lambda f(s_{\max},\overline{p},\overline{\alpha})\right).\end{aligned} \tag{33}$$

Optimizing the R.H.S. of (33) with respect to λ, we obtain the optimal λ_{opt} that maximizes the R.H.S as follows:

$$\lambda_{\text{opt}}=\frac{-1}{f(s_{\max},\overline{p},\overline{\alpha})}\left(\ln\frac{1}{f(s_{\max},\overline{p},\overline{\alpha})}\right). \tag{34}$$

Plugging (34) back in (33), we arrive at the first lower bound stated in Theorem 1.

B. Lower Bound 2

We can next obtain the second lower bound by bounding the second term in (21). Note that we are now working with conditional probability. We need to be careful about which cluster sample n is from. To find the probability that a negative sample n is hidden, we first find the probability that it is hidden in pool t in the first stage. Alternatively, we can find the probability that a given negative sample n is *not* hidden in pool t. A sample n is not hidden in pool t when every sample in that pool is negative. We have,

$$\begin{aligned}P(H_{tn}|n\text{ is negative}) &= 1-P(\overline{H}_{tn}|n\text{ is negative})\\ &= 1-(1-p_{\gamma_n})^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}(1-p_c)^{w_t^{(c)}}\\ &= 1-q_{\gamma_n}^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}q_c^{w_t^{(c)}},\end{aligned}$$

where we have defined $q_c \triangleq 1-p_c$. From the above, we now compute the probability that sample n is hidden in all T_1 tests:

$$\begin{aligned}&P(H_n|n\text{ is negative})\\ &\quad=\prod_{t:x_{tn}=1}\left(1-q_{\gamma_n}^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}q_c^{w_t^{(c)}}\right),\end{aligned}$$

where $x_{tn}=1$ whenever sample n is in pool t and $x_{tn}=0$ otherwise. We next define the variable $L(n,\overline{p})$ as follows:

$$\begin{aligned}L(n,\overline{p}) &= \ln(1-p_{\gamma_n})P(H_n|n\text{ is negative})\\ &= \ln q_{\gamma_n}+\sum_{t=1}^{T_1}x_{tn}\ln\left(1-q_{\gamma_n}^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}q_c^{w_t^{(c)}}\right)\end{aligned}$$

Using AM-GM inequality, we can now bound the second term in (21) as follows:

$$\begin{aligned}&\sum_{n=1}^{N}(1-p_{\gamma_n})P(H_n|n\text{ is negative})\\ &\qquad=\sum_{n=1}^{N}e^{L(n,\overline{p})}\geq N\exp\left(\frac{1}{N}\sum_{n=1}^{N}L(n,\overline{p})\right)\end{aligned} \tag{35}$$

We need to bound the term inside the exponential appearing in (35). However, we need to bound it differently from [4] due to having an extra term, and different constraints. We now have a constraint on the number of samples from individual cluster for all cluster, instead of having one constraint on the entire population. In addition, since we are working with conditional probability, we need to know exactly which cluster the sample we conditioned on comes from. Plugging the definition of $L(n,\overline{p})$ into (35), we have the following equality,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}L(n,\overline{p}) &= \frac{1}{N}\sum_{n=1}^{N}\ln q_{\gamma_n}\\ &+\frac{1}{N}\sum_{t=1}^{T_1}\left(\sum_{n=1}^{N}x_{tn}\ln\left(1-q_{\gamma_n}^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}q_c^{w_t^{(c)}}\right)\right).\end{aligned} \tag{36}$$

We can rewrite the first term on the right as follows,

$$\frac{1}{N}\sum_{n=1}^{N}\ln q_{\gamma_n}=\sum_{c=1}^{C}\alpha_c\ln q_c, \tag{37}$$

from the observation that the summation consists of α_cN copies of $\ln q_c$'s for all c. We then bound the second term on the right of (36) as follows,

$$\begin{aligned}&\frac{1}{N}\sum_{t=1}^{T_1}\left(\sum_{n=1}^{N}x_{tn}\ln\left(1-q_{\gamma_n}^{w_t^{(\gamma_n)}-1}\prod_{c\in[1:C]\setminus\{\gamma_n\}}q_c^{w_t^{(c)}}\right)\right)\\ &\stackrel{(a)}{\geq}\frac{1}{N}\sum_{t=1}^{T_1}\left(\sum_{c=1}^{C}w_t^{(c)}\ln\left(1-q_c^{w_t^{(c)}-1}q_{\max}^{w_t-w_t^{(c)}}\right)\right)\\ &\stackrel{(b)}{\geq}\frac{T_1}{N}\min_{\substack{w\in[2:s_{\max}]\\ w^{(c)}\in[0:\min\{w,\alpha_cN\}]\\ \sum_{c=1}^{C}w^{(c)}=w}}\left\{\sum_{c=1}^{C}w^{(c)}\ln\left(1-q_c^{w^{(c)}-1}q_{\max}^{w-w^{(c)}}\right)\right\}\\ &=-\lambda g(s_{\max},\overline{p},\overline{\alpha}),\end{aligned} \tag{38}$$

where (a) follows from the fact that we can bound individual q_c using $q_{\max}=(1-\min_{i\in[1:C]}p_i)$ and rewrite the summation over n using the summation over c since there are $w_t^{(c)}$ copies of the same logarithm term for all c; and (b) follows by bounding the summation using T_1 times the minimum value; $g(s_{\max},\overline{p},\overline{\alpha})$ is defined in (16), and $\lambda=T_1/N$. By combining

(36), (37) and (38), we have,

$$\frac{1}{N}\sum_{n=1}^{N} L(n,\overline{p}) \geq \sum_{c=1}^{C} \alpha_c \ln q_c - \lambda g(s_{\max},\overline{p},\overline{\alpha}). \tag{39}$$

By rewriting $\sum_{n=1}^{N} p_{\gamma_n}$ in (21) as $\sum_{c=1}^{C} p_c \alpha_c N$, the expected average number of tests per sample can be bounded as,

$$\begin{aligned} T^{\text{T.S. opt}}(\overline{p},\overline{\alpha}) &\geq \frac{T_1}{N} + \frac{E\left(T_2\right)}{N} \\ &\geq \lambda + \sum_{c=1}^{C} p_c\alpha_c + \exp\left(\sum_{c=1}^{C} \alpha_c \ln q_c - \lambda g(s_{\max},\overline{p},\overline{\alpha})\right). \end{aligned} \tag{40}$$

Optimizing the R.H.S. of (40) with respect to λ, we obtain the optimal λ_{opt} that maximizes the R.H.S as follows:

$$\lambda_{\text{opt}} = \frac{1}{g(s_{\max},\overline{p},\overline{\alpha})}\left(\sum_{c=1}^{C} \alpha_c \ln q_c - \ln \frac{1}{g(s_{\max},\overline{p},\overline{\alpha})}\right). \tag{41}$$

Plugging (41) back in (40), we arrive at the second lower bound stated in Theorem 1.