Sparsity-Constrained Community-Based Group Testing

Sarthak Jain, Martina Cardone, Soheil Mohajer
University of Minnesota, Minneapolis, MN 55455, USA, Email: {jain0122, mcardone, soheil}@umn.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619156

***Abstract*—In this work, we consider the sparsity-constrained community-based group testing problem, where the population follows a community structure. In particular, the community consists of F families, each with M members. A number k_f out of the F families are infected, and a family is said to be infected if k_m out of its M members are infected. Furthermore, the sparsity constraint allows at most ρ_T individuals to be grouped in each test. For this sparsity-constrained community model, we propose a probabilistic group testing algorithm that can identify the infected population with a vanishing probability of error and we provide an upper-bound on the number of tests. When $k_m = \Theta(M)$ and $M = \omega(\log(FM))$, our bound outperforms the existing sparsity-constrained group testing results trivially applied to the community model. If the sparsity constraint is relaxed, our achievable bound reduces to existing bounds for community-based group testing. Moreover, our scheme can also be applied to the classical dilution model, where it outperforms existing noise-level-independent schemes in the literature.**

I. Introduction

Group testing (GT), first introduced in 1943 in [1], is an umbrella term for the methods used to identify k defective items among n items, with as few tests as possible. The main idea consists of performing tests on pools/groups of items rather than testing each item individually. GT has many applications, ranging from medicine [2] to engineering [3], and is broadly classified into *combinatorial* GT and *probabilistic* GT. In combinatorial GT, the goal is to identify the defective items with a zero error probability [4]. In probabilistic GT, instead, it suffices that the error probability goes to zero asymptotically, as $n \to \infty$; moreover, for finite n, the error probability can be made arbitrarily small by appropriately scaling the number of tests [5]–[11]. GT can be *noiseless* or *noisy* [5], [10], [11]. In noiseless GT, each test is error free (no false positives or false negatives); whereas, in noisy GT, the test results may be erroneous [5], [9]–[13]. Examples of noise models include the binary symmetric noise [5], the dilution noise [9], [10], [13], and the threshold noise [11].

Most GT problems assume a *combinatorial prior* on the set of defective items. This means that the k defective items are equally likely to be any of the $\binom{n}{k}$ items. In this case, the counting bound [5] states that the number of tests required to identify the k defective items is at least $\Theta\big(k \log\big(\frac{n}{k}\big)\big)$. When k follows a sparse regime, that is, $k = \Theta\big(n^{\delta_k}\big)$ for some constant $\delta_k \in [0, 1)$, this is significantly less than individual testing, which requires $\Theta(n)$ tests. For sparse k, the counting bound indeed becomes $\Theta(k \log(n))$. Several GT schemes achieve the counting bound for noiseless GT with a combinatorial prior and hence, are order optimal [14]. Recent works have considered variants of GT with additional information on the set of defective items [15]–[18]. In [16], the authors introduced one such model, referred to as the *community* model, and analyzed the symmetric and general variants of it. The symmetric community model considers a community of F families, each with M members. A number k_f out of the F families are *infected*. If a family is healthy, none of its members are infected; if a family is infected, k_m out of its M members are infected. Ignoring the community structure, this model reduces to identifying $k_f k_m$ infected members out of $n = FM$ members, which in the sparse regime requires $\Theta(k_f k_m \log(n))$ tests. However, it was observed that leveraging the community structure can greatly reduce the number of tests [16], [18]. Note that the community-based GT exhibits similarities to other problems, such as GT with blocks of positives [19] and one-bit group-sparse signal reconstruction [20].

In this work, we consider the symmetric community model of [16], but we additionally impose a *sparsity constraint*. This constraint allows at most ρ_T individuals to be pooled in each test. This model has practical significance. Many infections, such as COVID-19, are indeed governed by community spread, and a community model is suitable to capture such scenarios. This model can be helpful also in bio-security applications, e.g., to test consignments of seeds/flowers [21]. Moreover, in many real world applications, there is often a constraint on the number of items that can be pooled in each test. This constraint may depend on several factors, e.g., test equipment capacity and test efficacy. For example, in swab pooling methods for COVID-19 testing, it is recommended to pool up to 16 swabs in each test [22]; and some HIV testing schemes allow 80 individual samples per test [23], [24].

For the sparsity-constrained community model, we propose a probabilistic GT scheme to identify the infected population with a probability greater than $1 - n^{-\lambda}$, for any constant $\lambda > 0$, and show that the number of tests required is at most $\Theta\Big(\frac{F \log(n)}{f(\widehat{\rho})}\Big)$, where $\widehat{\rho} = \min\Big\{\rho_T, \Big\lfloor \frac{F}{2k_f} \Big\rfloor\Big\}$ and $f(\rho) = \rho\Bigg(1 - \Big(1 - \frac{k_m}{M}\Big)^{\frac{\rho_T}{2\rho}}\Bigg)$. For $k_m = \Theta(M)$ and $M = \omega(\log(FM))$, our scheme requires much fewer tests than applying existing sparsity-constrained GT schemes [24], [25] to the community model. Moreover, without the sparsity constraint, our bound reduces to existing bounds in community-

This research was supported in part by the U.S. National Science Foundation under Grant CCF-1907785.

TABLE I: Quantities of interest used throughout the paper.

Quantity	Definition
F	Number of families
M	Number of members in each family
n	Total number of members, that is, $n = FM$
$\mathcal{D}$	Set of infected families
k_f	Number of infected families
$\mathcal{B}_f$	Set of infected members in family $f \in [F]$
k_m	Number of infected members in an infected family
ρ_T	Maximum number of members allowed in each test
T	Number of tests performed by a GT scheme

based GT [16]. Our scheme can also be applied to the classical dilution model [9], [11], [13], [26], [27], where there is no community structure or sparsity constraint. For this model, our scheme requires $\Theta\Big(\frac{k\log(n)}{\alpha}\Big)$ tests, where α is the dilution noise parameter. This bound provides a factor of α improvement over the best achievable bound [28] of $\Theta\Big(\frac{k\log(n)}{\alpha^2}\Big)$ among noise-level-independent (NLI) schemes [9] (i.e., when the test design is independent of α).

Notation. For any $k \in \mathbb{N}$, we define $[k] := \{1, 2, \ldots, k\}$. For a set $\mathcal{X}$, $|\mathcal{X}|$ denotes its cardinality. For a matrix M, we use $\mathsf{M}_{i,:}$ and $\mathsf{M}_{:,j}$ to represent its ith row and jth column, respectively. An empty set is denoted by $\varnothing$. For a vector $\boldsymbol{x}$, we let $\mathsf{supp}(\boldsymbol{x}) := \{i : \boldsymbol{x}_i \neq 0\}$. Finally, $\wedge$ and $\vee$ represent the Boolean AND and OR operations, respectively.

II. System Model

We consider F families, denoted by $[F]$, where each family consists of M members (this is the symmetric model in [16]). The total number of members is, therefore, $n := FM$. The members of family $f \in [F]$ are referred to as $\mathcal{M}_f := \{m_{(f-1)M+i} : i \in [M]\}$. An unknown subset $\mathcal{D} \subseteq [F]$, consisting of k_f families (that is, $|\mathcal{D}| = k_f$), is infected. We assume a combinatorial prior on this subset of infected families, that is, the defective set is chosen uniformly at random among all the $\binom{F}{k_f}$ sets of this size k_f. If a family f is not infected, none of its members are infected; whereas, if f is infected, an unknown subset $\mathcal{B}_f \subseteq \mathcal{M}_f$ of the M members of that family are infected. Again, for the symmetric model considered here, we assume that $|\mathcal{B}_f| = k_m$ for all $f \in \mathcal{D}$. Moreover, we assume that $\mathcal{B}_f$ is chosen uniformly at random among all the $\binom{M}{k_m}$ subsets of size k_m. Table I summarizes the quantities used for problem formulation.

Our goal is to design a GT scheme, which uses as few tests as possible, to identify the infected population with a vanishing error probability, i.e., a probability of error that goes to zero at a rate of $n^{-\lambda}$ for some constant $\lambda > 0$. Due to practical considerations [22], [29], we impose a *sparsity constraint*, which restricts the number of members that can participate in each test. In particular, in any given test, at most ρ_T out of the n members can be pooled together.

III. Preliminaries and Related Results

In this section, we first introduce the contact matrix, which is the mathematical model for GT. Then, we review some existing results and adapt them so as to ensure a probability of error that goes to zero at a rate of $n^{-\lambda}$ for some constant $\lambda > 0$. In particular, we establish two benchmarks for the performance of our algorithm, based on existing methods.

A. Combinatorial GT

Consider a general (with no sparsity constraint or community structure) GT problem with N items among which k are defective. Note that we use (N, k) to depict a general group testing setting and (n, k_f, k_m) to describe the community model. Let T be the number of tests performed by a GT algorithm. These tests can be described using the *contact* matrix $\mathsf{M}^{(c)} \in \{0,1\}^{T\times N}$, where each row corresponds to a test and each column corresponds to an item. If $\mathsf{M}^{(c)}_{t,i} = 1$, then item i is selected in test t. Let $\boldsymbol{x} \in \{0,1\}^N$ be the indicator vector for the defective items, that is, $\boldsymbol{x}_i = 1$ if and only if item i is defective. Then, the result of the tests can be represented by a vector $\boldsymbol{y}^{(c)} \in \{0,1\}^T$ as

$$\boldsymbol{y}^{(c)} = \mathsf{M}^{(c)} \odot \boldsymbol{x}, \tag{1}$$

where $\odot$ denotes the matrix-vector *logical* multiplication, in which the arithmetic multiplication and addition are replaced by logical AND and OR, respectively. More precisely, we have $\boldsymbol{y}^{(c)}_t = \bigvee_{i=1}^N (\mathsf{M}^{(c)}_{t,i} \wedge \boldsymbol{x}_i)$. It is known that using a proper selection of $\mathsf{M}^{(c)}$ and an appropriate decoder, with probability at least $1 - N^{-\lambda}$ for any $\lambda > 0$, the set of defective items can be identified using $T = \Theta(k\log(N/k))$ tests [4].

B. Sparsity-Constrained Combinatorial GT

The result of [4] holds when the number of items to be tested together in each pool is arbitrary. In general, a larger number of tests is required if a sparsity constraint is imposed [24], [25], [30]. Let ρ_U be the maximum number of items allowed to participate in each test. From these results and classical GT [5], it can be argued that, to achieve a probability of error of $\widetilde{N}^{-\lambda}$ for some $\widetilde{N} \geq N$ and any constant $\lambda > 0$, the number of tests[1] required is at least equal to [5], [25],

$$\widehat{T}\Big(N, k, \rho_U, \widetilde{N}\Big) = \Theta\bigg(\max\Big\{\frac{N}{\rho_U}, k\log(N)\Big\}\log_N\big(\widetilde{N}\big)\bigg). \tag{2}$$

In our system model (Section II), we can ignore the community structure and directly identify all the $k_f k_m$ infected members out of the n members. With a sparsity constraint ρ_T, the number of required tests can be found from (2) as

$$T_{\mathsf{nC,S}} = \widehat{T}(n, k_f k_m, \rho_T, n) = \Theta\bigg(\max\Big\{\frac{n}{\rho_T}, k_f k_m \log(n)\Big\}\bigg). \tag{3}$$

C. Community-Based GT Without Sparsity Constraints

In the system model (Section II), if there is no sparsity constraint (that is, $\rho_T = \infty$), a two-stage algorithm, introduced in [16], can be utilized, where: (i) in the first stage, the k_f infected families are identified; and (ii) in the second stage,

[1]The additional term of $\Theta\Big(\log_N(\widetilde{N})\Big)$ in (2), compared to [5], [25], guarantees that the error probability vanishes at the desired rate of $\widetilde{N}^{-\lambda}$ instead of $N^{-\lambda}$.

TABLE II: Quantities of interest used in the GT scheme.

Quantity	Definition
T_{I}	Number of tests in the first stage
T_{II}	Number of tests in the second stage
ρ	Number of families selected in each test
r	Number of members sampled from each selected family
α	Probability that an infected family is active
$\mathcal{D}_t$	Set of active infected families during test t
d	Threshold for the d-threshold decoder
$\mathcal{R}_{f,t}$	Members of a selected family f that participate in test t

depending on the regime of (k_m, M), either individual testing or GT is performed only on the infected families (identified in the first stage) to identify their k_m infected members. For both stages, this algorithm leverages existing non-adaptive probabilistic GT schemes [14]. The numbers of tests in the first stage and second stage, respectively, are given by

$$\begin{aligned} T_{\mathsf{C,nS,I}} &= \Theta(k_f \log(n)), \\ T_{\mathsf{C,nS,II}} &= \begin{cases} k_f \Theta(M) & \text{if } k_m = \Theta(M), \\ k_f \Theta(k_m \log(n)) & \text{if } k_m = o(M). \end{cases} \end{aligned} \tag{4}$$

D. Incorporating Sparsity in the Two-Stage Algorithm

In the first stage of the algorithm of [16], initially a contact matrix is designed to identify the k_f infected families. However, since tests should be applied on the individual members (rather than on the families), once a family is selected to participate in a test, all of its members will be pooled to be tested. Therefore, since each family consists of M members, in order to satisfy the sparsity constraint of ρ_T (on the number of members allowed in each test), we can pool together at most $\frac{\rho_T}{M}$ families to be tested. In other words, the initial test matrix should be designed with a sparsity constraint of $\frac{\rho_T}{M}$. Hence, using (2), the number of tests required in the first stage of the algorithm is given by

$$\begin{aligned} T_{\mathsf{C,S,I}} &= \widehat{T}\Big(F, k_f, \frac{\rho_T}{M}, n\Big) \\ &= \Theta\bigg(\max\Big\{\frac{FM}{\rho_T}, k_f \log(F)\Big\} \log_F(n)\bigg). \end{aligned} \tag{5}$$

Note that since this scheme pools all the members of the families selected in a test, it only works when $\rho_T \geq M$.

IV. The Proposed GT Scheme

In this section, we propose a new sparse GT algorithm to identify the infected members in the community structured problem. Inspired by [16] (where there is no sparsity constraint), we adopt a two-stage GT procedure. In the first stage (see Section IV-A), the goal is to identify the k_f infected families, whereas in the second stage (see Section IV-B), we perform GT only on the infected families (identified in the first stage) to identify their k_m infected members. We denote by T_{I} and T_{II} the number of tests required in the first and second stages, respectively. Then, the total number of tests required by the proposed algorithm is given by $T = T_{\mathsf{I}} + T_{\mathsf{II}}$.

A. First Stage: Identifying Infected Families

We use a contact matrix $\mathsf{M}^{(c)} \in \{0,1\}^{T_{\mathsf{I}} \times F}$, initially designed for F families for the first stage of the algorithm (similar to Section III-A with $(N, k, T) = (F, k_f, T_{\mathsf{I}})$). For simplicity, we assume that $k_f \geq 2$ and $F \geq 2k_f$. Table II summarizes the parameters used in the proposed scheme.

Probabilistic design of the contact matrix: We first choose a $T_{\mathsf{I}} \times F$ contact matrix $\mathsf{M}^{(c)}$ with a *family-sparsity* parameter $\rho \in [F]$ (which will be determined later). To this end, each row of $\mathsf{M}^{(c)}$ is uniformly, randomly, and independently from other rows, selected from the $\binom{F}{\rho}$ possible rows that have Hamming weight equal to ρ.

Family representative sets: Unlike the scheme in Section III-D, where all the members of a selected family participate in a test, we choose a set of *representative* members for each selected family to participate in tests. In particular, for each test t, a subset $\mathcal{R}_{f,t} \subseteq \mathcal{M}_f$ of members participate in the test. More formally, the set of individuals that are pooled together in test t is given by $\bigcup_{f \in \mathsf{supp}\left(\mathsf{M}^{(c)}_{t,:}\right)} \mathcal{R}_{f,t}$. To this end, for each (t, f), we select $\mathcal{R}_{f,t}$ uniformly at random from all the $\binom{M}{r}$ possible subsets of $\mathcal{M}_f$ of size $r := |\mathcal{R}_{f,t}| = \left\lfloor \frac{\rho_T}{\rho} \right\rfloor$. With the above designs of $\mathsf{M}^{(c)}$ and $\mathcal{R}_{f,t}$, the number of members that participate in each test satisfies

$$\sum_{f=1}^{F} \mathsf{M}^{(c)}_{t,f} \, |\mathcal{R}_{f,t}| = \sum_{f \in \mathsf{supp}\left(\mathsf{M}^{(c)}_{t,f}\right)} \left\lfloor \frac{\rho_T}{\rho} \right\rfloor = \rho \left\lfloor \frac{\rho_T}{\rho} \right\rfloor \leq \rho_T, \tag{6}$$

and hence, the sparsity constraint is satisfied.

The sampling matrix: With the representative sets (instead of the entire family) participating in each test, the identity in (1) does not hold in general. To see this, consider a case where $\mathsf{M}^{(c)}_{t,f} = 1$, and $\mathcal{R}_{f,t} \cap \mathcal{B}_f = \varnothing$ for an infected family $f \in \mathcal{D}$. Then, even if f is infected and selected to participate in the test, it will not cause the test t to be positive, since no infected member of the family is in its representative set. In other words, such an infected family *pretends* to be healthy in the test. To capture this uncertainty, we define a *sampling* matrix $\mathsf{M}^{(s)} \in \{0,1\}^{T_{\mathsf{I}} \times F}$ obtained from $\mathsf{M}^{(c)}$. We call an infected family $f \in \mathcal{D}$ *active* in test t, if and only if, $\mathcal{R}_{f,t} \cap \mathcal{B}_f \neq \varnothing$. We denote the set of active infected families of test t by $\mathcal{D}_t \subseteq \mathcal{D}$. Then, the sampling matrix $\mathsf{M}^{(s)}$ is given by

$$\mathsf{M}^{(s)}_{t,f} = \begin{cases} \mathsf{M}^{(c)}_{t,f} & \text{if } f \in ([F] \setminus \mathcal{D}) \cup \mathcal{D}_t, \\ 0 & \text{if } f \in \mathcal{D} \setminus \mathcal{D}_t, \end{cases} \tag{7}$$

and the actual results of the tests (performed on the representatives of the families) are given by

$$\boldsymbol{y}^{(s)} = \mathsf{M}^{(s)} \odot \boldsymbol{x}. \tag{8}$$

To understand the sampling matrix in (7), let us consider an infected family $f \in \mathcal{D}$ that is selected in test t (i.e., $\mathsf{M}^{(c)}_{t,f} = 1$). Now, if $\mathcal{R}_{f,t} \cap \mathcal{B}_f = \varnothing$, although f is infected, none of its infected members participate in test t. In other words, family f hides its true identity in test t. Since $\mathsf{M}^{(c)}_{t,f} = 1$ and $\boldsymbol{x}_f = 1$, we

have $\boldsymbol{y}_t^{(c)} = 1$. However, the actual test result $\boldsymbol{y}_t^{(s)}$ should not be influenced by f. This can be ensured by setting $\mathsf{M}_{t,f}^{(s)} = 0$.

Let α be the probability that an infected family is active, that is,

$$\alpha = \mathbb{P}[f \in \mathcal{D}_t | f \in \mathcal{D}] = 1 - \frac{\binom{M-k_m}{r}}{\binom{M}{r}}. \tag{9}$$

In other words, $\mathsf{M}_{t,f}^{(c)} = 1$ is replaced by $\mathsf{M}_{t,f}^{(s)} = 0$ with probability $1-\alpha$. Moreover, if $\alpha = 1$, then $\mathsf{M}^{(c)} = \mathsf{M}^{(s)}$. Note that the behavior of $\mathsf{M}^{(s)}$ and $\mathsf{M}^{(c)}$ is similar to that of the dilution model, that we recently studied in [28], and has also been investigated in [11], [13], [31].

Given this construction of $\mathsf{M}^{(c)}$ and $\mathcal{R}_{f,t}$, and the probabilistic nature of $\mathsf{M}^{(s)}$ and $\boldsymbol{y}^{(s)}$, the families are classified as infected or healthy using the following d-threshold decoder.

The d-threshold decoder: Let $\boldsymbol{y}_t^{(s)} \in \{0,1\}$ be the result of test $t \in [T_{\mathsf{I}}]$, given by $\boldsymbol{y}_t^{(s)} = \mathsf{M}_{t,:}^{(s)} \odot \boldsymbol{x}$. We define the score $S_{f,t}$ of family f in test t as

$$S_{f,t} = \begin{cases} 1 & \text{if } \mathsf{M}_{t,f}^{(c)} = 1 \text{ and } \boldsymbol{y}_t^{(s)} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Then, for a given $d > 0$, family f is marked as infected if and only if $S_f = \sum_{t=1}^{T_{\mathsf{I}}} S_{f,t} \geq d$. □

The following theorem provides the number of tests required in the first stage of the algorithm to ensure that the construction above can decode $\boldsymbol{x}$ with an overwhelming probability.

Theorem 1. *There exists a choice of the parameters (ρ, d) such that the d-threshold decoder requires at most*

$$T_{\mathsf{I}} = \min_{\rho \in [\widehat{\rho}]} \frac{\zeta(1+\lambda)F\log(n)}{\rho\,\alpha} \leq \frac{\zeta(1+\lambda)F\log(n)}{f(\widehat{\rho})} \tag{11a}$$

tests to identify the k_f infected families with error probability $\mathsf{P}_e \leq n^{-\lambda}$, for any $\lambda > 0$, where α is given in (9) *and*

$$f(\rho) = \rho\left(1 - \left(1 - \frac{k_m}{M}\right)^{\frac{\rho_T}{2\rho}}\right),\ \zeta = 64\ \mathrm{e}^4, \tag{11b}$$

$$\widehat{\rho} = \min\left\{\rho_T, \left\lfloor \frac{F}{2k_f} \right\rfloor\right\}. \tag{11c}$$

Proof. The proof of Theorem 1 and the choice of the parameters (ρ, d) (see (22)) are provided in Section VI. □

Remark 1. *Our scheme is NLI [9] because the construction of $\mathsf{M}^{(c)}$ does not depend on the noise parameter α. With no sparsity constraint, i.e., $\rho_T = \infty$, we have that $\widehat{\rho} = \left\lfloor \frac{F}{2k_f} \right\rfloor$. Our proposed scheme can then be used with the classical dilution model [9]–[11], [13], [28], where: (i) the task is to identify k defective items out of n items; and (ii) the defective items exhibit a dilution effect with probability α, independent of ρ. This leads to $T_{\mathsf{I}} = \Theta\left(\frac{k_f \log(n)}{\alpha}\right)$. To the best of our knowledge, the best achievable bound in the literature for the dilution model using a NLI GT scheme is $\Theta\left(\frac{k\log(n)}{\alpha^2}\right)$ tests [28] and our scheme outperforms this by a factor of α.*

B. Second Stage: Identifying All the Infected Members

To identify all the $k_f k_m$ infected members, we can either perform individual testing or sparsity-constrained GT, for each of the k_f families identified in the first stage. For the *linear* regime of k_m (i.e., $k_m = \Theta(M)$), individual testing (which has sparsity of 1) is preferred. In this case, we would require

$$T_{\mathsf{II,L}} = k_f \Theta(M) \tag{12}$$

tests. Otherwise, if k_m follows a sub-linear regime (i.e., $k_m = o(M)$), performing sparsity-constrained GT (see Section III-B) in each of the k_f infected families would be preferred. This would require a number of tests equal to

$$T_{\mathsf{II,sL}} = \begin{cases} k_f \Theta\left(\frac{M}{\rho_T} \frac{\log(n)}{\log(M)}\right) & \text{if } \rho_T = o\left(\frac{M}{k_m}\right), \\ k_f \Theta(k_m \log(n)) & \text{if } \rho_T = \Theta\left(\frac{M}{k_m}\right). \end{cases} \tag{13}$$

Hence, depending on the regime of M, the number of tests T_{II} for the second stage, can be obtained from (12) or (13).

V. Analysis and Comparison

In this section, we further analyze the performance (in terms of number of tests required) of the GT scheme proposed in Section IV. Note that all the comparisons are order-wise, and the multiplicative constants behind the Θ notation are ignored. In particular, from Theorem 1 we have the following corollary.

Corollary 1. *It holds that*

$$T_{\mathsf{I}} \leq \Theta\left(\max\left\{\frac{FM}{\rho_T k_m}, k_f\right\}\log(n)\right). \tag{14}$$

Proof. The proof can be found in [32, Appendix A]. □

We now compare our scheme with existing results. Note that, due to the structure of the problem, the primary interest is on a specific regime of parameters, namely: (i) the total number of infected members falls within a sparse regime, i.e., $k_f k_m = o(n)$ (otherwise individual testing would be optimum); (ii) once a family is infected, a significant number of its members get infected, i.e., $k_m = \Theta(M)$; and (iii) the size of the families is not very small, i.e., $M = \omega(\log(n))$ (otherwise each family can be thought as an individual).

• **Ignoring the community structure.** A naive algorithm that does not exploit the community structure of the problem was discussed in Section III-B. For the regime of interest on $(k_m, k_f k_m, M)$, the ratio of the total (both stages) number of tests required by the two algorithms can be bounded as

$$\frac{T_{\mathsf{I}} + T_{\mathsf{II,L}}}{T_{\mathsf{nC,S}}} \leq \frac{\Theta\left(\max\left\{\frac{n}{\rho_T k_m}, k_f\right\}\log(n) + k_f M\right)}{\Theta\left(\max\left\{\frac{n}{\rho_T}, k_f k_m \log(n)\right\}\right)} = \Theta\left(\frac{\log(n)}{M} + \frac{1}{\log(n)}\right). \tag{15}$$

From (15), we note that exploiting the community structure offers an order-wise reduction in the number of tests.

• **Enforcing sparsity for the community-based scheme.** As discussed in Section III-D, we can arrive at a sparse GT scheme that leverages the community structure. This requires

$T_{\mathsf{C,S,I}}$ tests (see (5)) in its first stage, while the number of tests required in the second stage is identical to that of our proposed algorithm (given in (12) or (13)). Since both schemes require the same number of tests in the second stage, we only compare their performance in the first stage. We have that

$$\frac{T_{\mathsf{I}}}{T_{\mathsf{C,S,I}}} \leq \frac{\Theta\Big(\max\Big\{\frac{n\log(F)}{\rho_T k_m}, k_f\log(F)\Big\}\Big)}{\Theta\Big(\max\Big\{\frac{n}{\rho_T}, k_f\log(F)\Big\}\Big)}$$
$$= \begin{cases} \Theta\Big(\frac{\log F}{M}\Big) & \text{if } 1 \leq \rho_T < \frac{n}{k_m k_f}, \\ \Theta\Big(\frac{\rho_T k_f \log(F)}{n}\Big) & \text{if } \frac{n}{k_f k_m} \leq \rho_T < \frac{n}{k_f \log(F)}, \\ \Theta(1) & \text{if } \rho_T \geq \frac{n}{k_f\log(F)}. \end{cases} \quad (16)$$

From equation (16), we note that our scheme outperforms (order-wise) the scheme proposed by the authors of [16] for a wide range of parameters. When no sparsity constraint is imposed (i.e., $\rho_T = \infty$), $T_{\mathsf{C,S,I}} = \Theta(k_f\log(n))$, which is identical to the number of tests required by our scheme (see (14)). Therefore, without any sparsity constraint, our scheme performs equivalent to the two-stage scheme of [16]. Also, note that the scheme of Section III-D is not feasible when $\rho_T < M$ whereas our scheme works for all $\rho_T \in [n]$.

VI. Proof of Theorem 1

In this section, we prove Theorem 1. We use two propositions that are stated next. Specifically, Proposition 1 bounds

$$\mu_p = \mathbb{E}\big[S_f \mid f \notin \mathcal{D}\big] \text{ and } \mu_m = \mathbb{E}\big[S_f \mid f \in \mathcal{D}\big], \quad (17)$$

where $S_f = \sum_{t=1}^{T_{\mathsf{I}}} S_{f,t}$ with $S_{f,t}$ defined in (10). We note that S_f of $f \in \mathcal{D}$ is expected to be higher than S_f of $f \notin \mathcal{D}$. This is formally shown by Proposition 1.

Proposition 1. *For $x \in [k_f]$, let h_x be defined as*

$$h_x := \sum_{\ell=0}^{x} \binom{x}{\ell}\alpha^\ell(1-\alpha)^{x-\ell}\left(1 - \frac{\binom{F-\ell-1}{\rho-1}}{\binom{F}{\rho}}\right), \quad (18)$$

for α given by (9)*. Then, for any ρ in the interval $\Big[\Big\lfloor\frac{F}{2k_f}\Big\rfloor\Big]$,*

(i) $h_x \leq \Big(1-\frac{\rho}{F}\Big) + \frac{\alpha\rho}{F}$,

(ii) $\mu_p = T_{\mathsf{I}}\Big(h_{k_f} - \Big(1-\frac{\rho}{F}\Big)\Big) \leq T_{\mathsf{I}}\frac{\alpha\rho}{F}$,

(iii) $\mu_m = T_{\mathsf{I}}\Big(\alpha + (1-\alpha)h_{k_f-1} - \Big(1-\frac{\rho}{F}\Big)\Big) \leq T_{\mathsf{I}}\frac{2\alpha\rho}{F}$,

(iv) $\mu_m - \mu_p \geq \frac{\alpha\rho T_{\mathsf{I}}}{2F}\mathrm{e}^{-2}$.

Proof. The proof can be found in [32, Appendix B]. □

The next proposition will be useful in the proof of Theorem 1 for choosing the family-sparsity parameter ρ.

Proposition 2. *Let $U \in \mathbb{N}$ and $\upsilon \in (0,1)$. Then,*

$$\arg\max_{\rho\in[U]} \rho\Big(1-\upsilon^{\frac{\rho_T}{\rho}}\Big) = U. \quad (19)$$

Proof. The proof can be found in [32, Appendix C]. □

We are ready to prove Theorem 1. Let P_+ and P_- be the probabilities of false positive and false negative errors of the d-threshold decoder for a given $f \in [F]$, respectively, i.e.,

$$\mathsf{P}_+ = \mathbb{P}[S_f \geq d | f \notin \mathcal{D}] \text{ and } \mathsf{P}_- = \mathbb{P}[S_f < d | f \in \mathcal{D}]. \quad (20)$$

By the union bound, the total error probability P_e can be upper bounded as

$$\mathsf{P}_e \leq (F-k_f)\mathsf{P}_+ + k_f\mathsf{P}_-. \quad (21)$$

We choose the following parameters,

$$\rho = \widehat{\rho},\ d = \frac{\mu_m+\mu_p}{2},\ T_{\mathsf{I}} = \frac{\zeta(1+\lambda)F\log(n)}{\rho\alpha}, \quad (22)$$

where $\widehat{\rho}$ and ζ are given in (11), α is given in (9), and $\lambda > 0$ is a constant. With these choices, we bound P_+ and P_- as

$$\mathrm{P}_+ = \mathbb{P}\Big[S_f \geq \frac{\mu_m+\mu_p}{2} \;\Big|\; f \notin \mathcal{D}\Big]$$
$$\overset{(a)}{=} \mathbb{P}\Big[S_f \geq \mu_p(1+\delta_p) \;\Big|\; f \notin \mathcal{D}\Big] \overset{(b)}{\leq} \exp\left(-\frac{\delta_p^2\mu_p}{2+\delta_p}\right)$$
$$= \exp\left(-\frac{(\mu_m-\mu_p)^2}{6\mu_p+2\mu_m}\right) \overset{(c)}{\leq} \exp\left(-\frac{\mathrm{e}^{-4}\alpha\rho T_{\mathsf{I}}}{40F}\right) \overset{(d)}{\leq} n^{-1-\lambda}, \quad (23)$$

where the labeled (in)equalities follow from: (a) letting $\delta_p = \frac{\mu_m-\mu_p}{2\mu_p} \geq 0$; (b) applying Chernoff's bound; (c) using Proposition 1; and (d) using T_{I} in (22).

The false negative error probability can be bounded as

$$\mathrm{P}_- \overset{(a)}{=} \mathbb{P}\Big[S_f < \mu_m(1-\delta_m) \;\Big|\; f \in \mathcal{D}\Big]$$
$$\overset{(b)}{\leq} \exp\left(-\frac{\delta_m^2\mu_m}{2}\right) \overset{(c)}{\leq} n^{-1-\lambda}, \quad (24)$$

where the labeled (in)equalities follow from: (a) letting $\delta_m = \frac{\mu_m-\mu_p}{2\mu_m} \in (0, 0.5]$; (b) using Chernoff's bound; and (c) using Proposition 1 and T_{I} in (22).

Combining (23) and (24) together with the union bound in (21), we get $\mathsf{P}_e \leq n^{-\lambda}$. Furthermore, the number of tests that suffice to achieve this probability of error is given by

$$T_{\mathsf{I}} \overset{(a)}{=} \frac{\zeta(1+\lambda)F\log(n)}{\rho\left(1-\frac{\binom{M-k_m}{r}}{\binom{M}{r}}\right)} = \frac{\zeta(1+\lambda)F\log(n)}{\rho\Big(1-\prod_{j=1}^{r}\Big(1-\frac{k_m}{M-j+1}\Big)\Big)}$$
$$\leq \frac{\zeta(1+\lambda)F\log(n)}{\rho\Big(1-\big(1-\frac{k_m}{M}\big)^r\Big)} \overset{(b)}{\leq} \frac{\zeta(1+\lambda)F\log(n)}{\rho\Big(1-\big(1-\frac{k_m}{M}\big)^{\frac{\rho_T}{2\rho}}\Big)}, \quad (25)$$

where (a) follows from using T_{I} in (22) and α in (9) and (b) is due to the fact that $\rho \leq \rho_T$ and hence, $r = \Big\lfloor\frac{\rho_T}{\rho}\Big\rfloor \geq \frac{\rho_T}{2\rho}$.

To conclude the proof, we find the value of ρ that minimizes (25). For this, we analyze the denominator of the right-hand side of (25), which is $f(\rho)$ defined in (11), where $\rho \leq \rho_T$. In the proof above, we also used Proposition 1, which requires $\rho \leq \Big\lfloor\frac{F}{2k_f}\Big\rfloor$. Thus, we need $\rho \leq \widehat{\rho}$, where $\widehat{\rho}$ is defined in (11). We now seek to maximize $f(\rho)$ over the set $\rho \in [\widehat{\rho}]$. Substituting $\upsilon = \big(1-\frac{k_m}{M}\big)^{\frac{1}{2}}$ in Proposition 2, it follows that the optimal choice of ρ is $\rho = \widehat{\rho}$. Using $\rho = \widehat{\rho}$ in (25) concludes the proof of Theorem 1.

REFERENCES

[1] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 12 1943.

[2] 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," *Frontiers in Public Health*, vol. 9, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.583377

[3] 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.

[4] D.-Z. Du and F. K. Hwang, *Combinatorial Group Testing and Its Applications*, 2nd ed. WORLD SCIENTIFIC, 1999. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/4252

[5] C. 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)*, 2011, pp. 1832–1839.

[6] A. Mazumdar, "Nonadaptive Group Testing With Random Set of Defectives," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7522–7531, 2016.

[7] A. Barg and A. Mazumdar, "Group testing schemes from codes and designs," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7131–7141, 2017.

[8] H. A. Inan, P. Kairouz, M. Wootters, and A. Ozgur, "On the Optimality of the Kautz-Singleton Construction in Probabilistic Group Testing," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 188–195.

[9] G. Arpino, N. Grometto, and A. S. Bandeira, "Group Testing in the High Dilution Regime," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 1955–1960.

[10] 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.

[11] X. Cheng, S. Jaggi, and Q. Zhou, "Generalized Group Testing," *IEEE Transactions on Information Theory*, vol. 69, no. 3, pp. 1413–1451, 2023.

[12] J. Scarlett, "Noisy Adaptive Group Testing: Bounds and Algorithms," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3646–3661, 2019.

[13] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group Testing With Probabilistic Tests: Theory, Design and Application," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 7057–7067, 2011.

[14] M. Aldridge, O. Johnson, and J. Scarlett, "Group Testing: An Information Theory Perspective," *Foundations and Trends® in Communications and Information Theory*, vol. 15, no. 3-4, pp. 196–392, 2019. [Online]. Available: http://dx.doi.org/10.1561/0100000099

[15] S.-J. Cao, R. Goenka, C.-W. Wong, A. Rajwade, and D. Baron, "Group Testing with Side Information via Generalized Approximate Message Passing," *IEEE Transactions on Signal Processing*, vol. 71, pp. 2366–2375, 2023.

[16] P. Nikolopoulos, S. Rajan Srinivasavaradhan, T. Guo, C. Fragouli, and S. Diggavi, "Group testing for connected communities," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 130. PMLR, 13–15 Apr 2021, pp. 2341–2349. [Online]. Available: https://proceedings.mlr.press/v130/nikolopoulos21a.html

[17] E. Karimi, A. Heidarzadeh, K. R. Narayanan, and A. Sprintson, "Noisy Group Testing with Side Information," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 2022, pp. 867–871.

[18] P. Nikolopoulos, S. R. Srinivasavaradhan, T. Guo, C. Fragouli, and S. Diggavi, "Group testing for overlapping communities," in *2021 IEEE International Conference on Communications (ICC)*, 2021, pp. 1–7.

[19] T. V. Bui, Y. Meng Chee, J. Scarlett, and V. K. Vu, "Group Testing with Blocks of Positives," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1082–1087.

[20] N. Koep, A. Behboodi, and R. Mathar, "Performance Analysis of One-bit Group-sparse Signal Reconstruction," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5272–5276.

[21] R. G. Clark, B. Barnes, and M. Parsa, "Clustered and Unclustered Group Testing for Biosecurity," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 29, pp. 193–211, 2023.

[22] A. P. Christoff, G. N. F. Cruz, A. F. R. Sereia, D. R. Boberg, D. C. de Bastiani, L. E. Yamanaka, G. Fongaro, P. H. Stoco, M. L. Bazzo, E. C. Grisard, C. Hernandes, and L. F. V. de Oliveira, "Swab pooling: A new method for large-scale RT-qPCR screening of SARS-CoV-2 avoiding sample dilution," *PLoS One*, vol. 16, no. 2, p. e0246544, 2021.

[23] 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. [Online]. Available: http://www.jstor.org/stable/171999

[24] O. Gebhard, M. Hahn-Klimroth, O. Parczyk, M. Penschuck, M. Rolvien, J. Scarlett, and N. Tan, "Near-Optimal Sparsity-Constrained Group Testing: Improved Bounds and Algorithms," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3253–3280, 2022.

[25] V. Gandikota, E. Grigorescu, S. Jaggi, and S. Zhou, "Nearly Optimal Sparse Group Testing," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2760–2773, 2019.

[26] A. Mazumdar and S. Mohajer, "Group testing with unreliable elements," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014, pp. 1–3.

[27] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Transactions on Information Theory*, vol. 58, pp. 1880–1901, 2009.

[28] S. Jain, M. Cardone, and S. Mohajer, "Identifying Reliable Machines for Distributed Matrix-Vector Multiplication," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 820–825.

[29] J. Fernández-Salinas, D. Aragón-Caqueo, G. Valdés, and D. Laroze, "Modelling pool testing for SARS-CoV-2: addressing heterogeneity in populations," *Epidemiol Infect*, vol. 149, p. e9, 2020.

[30] E. Price, J. Scarlett, and N. Tan, "Fast Splitting Algorithms for Sparsity-Constrained and Noisy Group Testing," *Information and Inference: A Journal of the IMA*, vol. 12, no. 2, pp. 1141–1171, 01 2023. [Online]. Available: https://doi.org/10.1093/imaiai/iaac031

[31] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1880–1901, 2012.

[32] S. Jain, M. Cardone, and S. Mohajer, "Sparsity-Constrained Community-Based Group Testing," *arXiv:2403.12419*, 2024.