Group Testing on General Set-Systems

Mira Gonen, Michael Langberg, and Alex Sprintson

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834789

***Abstract*—Group testing is one of the fundamental problems in coding theory and combinatorics in which one is to identify a subset of contaminated items from a given ground set. There has been renewed interest in group testing recently due to its applications in diagnostic virology, including pool testing for the novel coronavirus. The majority of existing works on group testing focus on the *uniform* setting in which any subset of size d from a ground set V of size n is potentially contaminated.**

In this work, we consider a *generalized* version of group testing with an arbitrary set-system of potentially contaminated sets. The generalized problem is characterized by a hypergraph $H = (V, E)$, where V represents the ground set and edges $e \in E$ represent potentially contaminated sets. The problem of generalized group testing is motivated by practical settings in which not all subsets of a given size d may be potentially contaminated, rather, due to social dynamics, geographical limitations, or other considerations, there exist subsets that can be readily ruled out. For example, in the context of pool testing, the edge set E may consist of families, work teams, or students in a classroom, i.e., subsets likely to be mutually contaminated. The goal in studying the generalized setting is to leverage the additional knowledge characterized by $H = (V, E)$ to reduce the number of tests.

The paper considers both adaptive and non-adaptive group testing and makes the following contributions. First, for the non-adaptive setting, we show that finding an optimal solution for the generalized version of group testing is NP-hard. For this setting, we present a solution that requires $O(d \log |E|)$ tests, where d is the maximum size of a set $e \in E$. Our solutions generalize those given for the traditional setting and are shown to be of order-optimal size $O(\log |E|)$ for hypergraphs with edges that have "large" symmetric differences. For the adaptive setting, when edges in E are of size exactly d, we present a solution of size $O(\log |E| + d \log^2 d)$ that comes close to the lower bound of $\Omega(\log |E| + d)$.

I. Introduction

Group testing is one of the fundamental problems in coding theory, statistical inference, and combinatorics due to its practical importance in a broad range of applications, such as multi-access communication [1], pattern matching [2], molecular biology, and others. The problem has deep connection to other fundamental problems in combinatorics and coding theory [3].

The group testing problem was subject to a large number of studies since its introduction more than 75 years ago (refer to, e.g., [4], [5] and references therein). An instance of the group testing problem includes a ground set V of items of size n, a subset of which may be *contaminated* (we refer to the latter set as a *contaminated set*). In the traditional setting, for a given parameter d, any subset of V of size d may potentially be a contaminated set. The contaminated set can be detected through a pooling process which includes a series of tests, where each test reveals the existence of a contaminated item in the tested subset of items. The goal of the group testing problem is to design a minimum set of tests that can identify the contaminated set of items in V. The testing algorithms can be constructed in a non-adaptive manner, i.e., fixed in advance, or in an *adaptive* manner, i.e., each test can depend on the outcome of previous tests.

In this work, we study a more general version of group testing, termed here *generalized group testing*, in which the set-system of potentially contaminated sets is characterized by a hypergraph $H = (V, E)$, in which the vertex set V represents the ground set and the hyperedges in E represent potentially contaminated sets. Traditional group testing thus corresponds to the hypergraph H with edge set E consisting of *all* subsets of V of size d. Our study of arbitrary edge sets E grants the flexibility required for a broad range of settings, including those in which the set-system E of potentially contaminated sets does not have any uniformity or regularity properties. For example, in the context of pool testing, E can represent potentially contaminated sets that correspond to families, work teams, or groups of friends or students in a classroom [6]. We consider settings in which $H = (V, E)$ is known in advance, e.g., E can capture groups of individuals that are likely get infected. Our goal in studying the generalized setting is to leverage the additional knowledge characterized by $H = (V, E)$ to minimize the number of required tests.

Contribution. The paper considers both adaptive and non-adaptive group testing and makes the following contributions. First, for the non-adaptive setting, we show that finding an optimal solution for the generalized version of the group testing problem is NP-hard, and approximating the optimal solution within a factor of $1 + \varepsilon$ (for sufficiently small ε) is as hard as coloring a 3-colorable graph with n^{ε} colors. The latter is a well known open problem, e.g., [7], [8]. For the non-adaptive setting, we present a solution that requires $O(d \log |E|)$ tests, where d is the maximum size of a set $e \in E$. Our solutions generalize those given in the classical setting and are shown to be of order-optimal size $O(\log |E|)$ for hypergraphs with edges that have "large" symmetric differences. For the adaptive setting, in which all edges in E are of size exactly d, we present a solution of size $O(\log |E| + d \log^2 d)$ that comes close to the lower bound of $\Omega(\log |E| + d)$. For the adaptive setting in which d is the maximum size of a set $e \in E$, we obtain an upper-bound of $O(\log |E| + d^2)$.

Related works. The overwhelming majority of works on group testing fall under the traditional setting in which $H = (V, E)$ includes all edges of size d (or, alternatively, of size at most d). Known upper and lower bounds in this context

Mira Gonen is with the Department of Computer Science, Ariel University, Ariel, 40700, Israel (e-mail: mirag@ariel.ac.il).

Michael Langberg is with the Department of Electrical Engineering, State University of New-York at Buffalo, Buffalo, NY 14260, USA (e-mail: mikel@buffalo.edu). Work supported in part by NSF grant 1909451.

Alex Sprintson is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA (e-mail: spalex@tamu.edu).

are reviewed in Section II-A. For hypergraphs H that differ from the traditional setting, less is known. Using our notation, Nikolopoulos et al. [6], [9] study hypergraphs $H = (V, E)$ that have a certain *community structure*. Specifically, [6] assumes that V consists of F disjoint groups (referred to as *families*) and studies the special setting in which E includes all edges that intersect a bounded number of families. The paper leverages the structure of $H = (V, E)$ to maximize the efficiency of the group testing algorithms in the adaptive, non-adaptive, and probabilistic settings; [9] further studies the problem for the case of families that are not necessarily disjoint. Ahn et al. and Arasli and Ulukus [10], [11] also discuss group testing under community constraints using different infection spread models. Other recent works related to community aware group testing include [12], [13]. A few related papers [14], [15] focus on leveraging side-information (e.g., that can be obtained from contact tracing) to make the decoding algorithm faster. Finally, *graph-constrained group testing*, a variant of the group testing problem where the *tests* must conform to constraints imposed by a graph, is considered, e.g., in [16]–[18]. The results of this work differ significantly from those above as we study general set-systems of potentially contaminated sets, and do not place any constraints on the tests used.

II. Problem Formulation

An instance of the group testing problem includes a ground set V of n items, a subset of which may be contaminated, with the goal of designing a minimum set of tests that can identify the contaminated items. We first define traditional non-adaptive group testing.

Definition 1 (traditional group testing (non-adaptive)) *For a ground set V of size n and a parameter d, find a minimum size family $\mathcal{T}$ of subsets of V such that for any $A, B \subseteq V$ of size d there exists $T \in \mathcal{T}$ for which $A \cap T = \emptyset$ if and only if $B \cap T \neq \emptyset$.*

Equivalently to Definition 1, given a ground set $V = \{1, \ldots, n\}$, a family of tests $\mathcal{T} = \{T_1, \ldots, T_k\}$ corresponds to a $k \times n$ matrix T with $T_{ij} = 1$ if $j \in T_i$, and $T_{ij} = 0$ otherwise. The outcome $y_{i,A}$ of the test T_i on a subset A is 1 if there exists $j \in A$ with $T_{ij} = 1$, and 0 otherwise. Namely, $y_{i,A} = \vee_{j \in A} T_{ij}$. With this notation, we seek a family of tests $\mathcal{T} = \{T_1, \ldots, T_k\}$ of minimum cardinality with corresponding matrix T such that for any $A, B \subseteq V$ of size d there exists $T_i \in \mathcal{T}$ for which $y_{i,A} \neq y_{i,B}$. Such T_i is said to separate A and B.

In traditional group testing, given a subset $S \subseteq V$ of contaminated items of cardinality d, the outcomes $y_{1,S}, \ldots, y_{k,S}$ of tests $\mathcal{T} = \{T_1, \ldots, T_k\}$ can be used to reliably recover the contaminated subset S.

We now turn to define the object studied in this work - *generalized group testing* - in which one requires $\mathcal{T} = \{T_1, \ldots, T_k\}$ to separate not any two subsets A and B of size d, but rather any two subsets A and B in a known family E.

Definition 2 (generalized group testing (non-adaptive)) *Given a ground set V of size n and a family E of subsets of V, find a minimum size family $\mathcal{T}$ of subsets of V such that for any $A, B \in E$ there exists $T \in \mathcal{T}$ for which $A \cap T = \emptyset$ if and only if $B \cap T \neq \emptyset$.*

Notice that the ground set V and the family E can be represented by a hypergraph $H = (V, E)$ whose vertices are the elements of the ground set and whose hyperedges are the sets in E. As before, equivalently to Definition 2, a family of tests $\mathcal{T} = \{T_1, \ldots, T_k\}$ corresponds to a $k \times n$ matrix T, and in the generalized group testing problem we seek a minimum sized family of tests $\mathcal{T} = \{T_1, \ldots, T_k\}$ with corresponding matrix T such that for any $A, B \in E$ there exists $T_i \in \mathcal{T}$ for which $y_{i,A} \neq y_{i,B}$ (i.e., T_i separates A and B). In the general group testing problem, for any possible subset $S \in E$ of contaminated items, the outcomes $y_{1,S}, \ldots, y_{k,S}$ of $\mathcal{T} = \{T_1, \ldots, T_k\}$ can be used to reliably recover S.

It is evident by our definitions that the traditional group testing problem with parameter d corresponds to the generalized problem with a hypergraph $H = (V, E)$ in which the edge set consists of all subsets of V of size d.

We now turn to define the adaptive version of group testing in which one can design the tests $\mathcal{T}$ adaptively, that is, test T_i may depend on the outcomes of tests T_j for $j < i$. We present the definition for the generalized case, with the definition for the traditional setting following as a special case.

Definition 3 (generalized adaptive group testing) *Given a ground set V of size n, a family E of subsets of V, and a fixed but unknown subset $S \in E$ of contaminated items, interactively design a family $\mathcal{T} = \{T_1, \ldots, T_k\}$ of subsets of V such that for any $2 \leq i \leq k$ the choice of T_i depends on $\{y_{j,S} | j < i\}$, where $y_{j,S}$ is 1 if $S \cap T_j \neq \emptyset$, and 0 otherwise. The outcomes $y_{1,S}, \ldots, y_{k,S}$ can be used to reliably recover the contaminated subset S, in the sense that for any other $A \in E$ there exists an index i such that T_i separates A and S, i.e., $y_{i,A} \neq y_{i,S}$. The governing adaptive algorithm is said to use at most k tests if for any $S \in E$, the interactively designed family of tests $\mathcal{T}$ is of size at most k. One seeks to find an adaptive scheme that minimizes the value of k.*

A. Prior upper and lower bounds on group testing

We start by stating the information-theoretic lower bound that follows from the fact that each test (in both the adaptive and non-adaptive setting) yields at most one bit of information regarding the contaminated set (see, e.g., [5], [19]–[21]).

Claim 1 *(Adaptive & non-adaptive information-theoretic lower bound) For any $0 < d < n$, the size of an optimal adaptive or non-adaptive solution for the traditional group testing problem with parameters n and d is at least $\Omega(d \log(n/d))$. The size of an optimal solution for the generalized group testing problem with corresponding hypergraph $H = (V, E)$ is at least $\lceil \log_2 |E| \rceil$.*

For non-adaptive traditional group-testing there is an improved lower bound [22]:

Claim 2 *(Non-adaptive lower bound) For any $0 < d < n$, the size of an optimal non-adaptive solution for the traditional group testing problem with parameters n and d is $\Omega(\min\{n, d^2 \log n / \log d\})$.*

Known upper-bounds for the traditional group testing problem in the non-adaptive and adaptive setting (for general n, d) are given below (see e.g., [5], [20], [23]):

Claim 3 *(Non-adaptive upper bound) For any $0 < d < n$, all d contaminated items in a given ground set V of size n can be found using at most $O(d^2 \log n)$ tests.*

Claim 4 *(Adaptive upper bound) For any $0 < d < n$, all d contaminated items in a given ground set of size n can be found using at most $d \log_2(n/d) + O(d)$ adaptive tests, even when d is unknown.*

The table below compares the upper and lower bounds presented in this work (for small values of d) with those surveyed above (we use the notation: Adaptive (A), Non-Adaptive (NA), Upper-bound (UB), Lower-bound (LB)). Note that our results for generalized group testing on $H = (V, E)$ match, or come close to matching, those of traditional group testing when E is taken to be all subsets of size d in V (and thus $\log |E| = \Theta(d \log(n/d))$).

	Traditional	Generalized
NA/UB	$O(d^2 \log n)$	$O(d \log \lvert E \rvert)$
NA/LB	$\Omega(d^2 \log n / \log d)$	$\Omega(d \log \lvert E \rvert / \log d)$
A/UB	$d \log_2(n/d) + O(d)$	$O(\log \lvert E \rvert + d \log^2 d)$
A/LB	$\Omega(d \log(n/d))$	$\Omega(\log \lvert E \rvert + d)$

III. Non-adaptive Generalized Group Testing

A. The Computational Complexity of finding an Optimal or Approximately-Optimal Solution

We first prove that the generalized group testing problem is NP-hard by showing a reduction from 3-colorability. [1]

Theorem 1 *Finding the size of an optimal solution for the (non-adaptive) generalized group testing problem is NP-hard.*

Proof: Given a graph $G = (V, E)$, which is an instance of the 3-colorability problem, define the following instance of the generalized group testing problem. Let $V = \{1, 2, \ldots, n\}$ and assume without loss of generality that $|E| = 2^\ell - 1$ for some integer $\ell \geq 2$ (otherwise one can modify G by adding an additional 2-colorable component). Define the hypergraph $H = (V_H, E_H)$ for the generalized group testing problem as follows: $V_H = E \cup V$ and

$$E_H = \{\{1, \ldots, n, e\} \mid e \in E\} \cup \{\{i, e\} | i \in V, e \in E, \exists j \in V \text{ s.t } e = (i, j)\}.$$

Note that $|V_H| = 2^\ell - 1 + n$ (recall that $|E| = 2^\ell - 1$) and that $|E_H| = 2^\ell - 1 + 2(2^\ell - 1) = 2^{\ell+1} + 2^\ell - 3 > 2^{\ell+1}$. This latter fact implies (using the bounds of Claim 1) that the optimal solution for the generalized group testing problem corresponding to H is of size at least $\ell + 2$.

[1] A valid 3-coloring of a graph G is an assignment of at most 3 colors to its vertices so that the vertices of each edge are assigned to distinct colors.

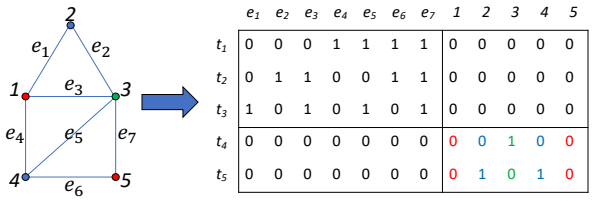

Fig. 1. Illustration of the reduction in Theorem 1. As $|E| = 7$, the parameter ℓ equals 3. Note that each edge e_m has a unique vector u_m encoded in rows t_1, t_2, t_3, and each vertex has a unique color encoded by rows t_3, t_4.

In what follows, we show that G is 3-colorable (but not 2-colorable) if and only if the generalized group testing problem corresponding to H can be solved using $\ell + 2$ tests.

For the first direction assume that G is 3-colorable (but not 2-colorable). Define a $(\ell + 2) \times (2^\ell - 1 + n)$ matrix T as follows. For every edge $e_m \in E$ let u_m be a distinct identifying binary non-zero encoding using ℓ bits. For $1 \leq m \leq 2^\ell - 1$ the m'th column in T is an $(\ell + 2)$-length vector whose first ℓ entries are u_m and the last 2 entries are 0. For $2^\ell \leq m \leq 2^\ell + n - 1$ the m'th column in T is a $(\ell + 2)$-length vector whose first ℓ entries are 0 and the last 2 entries are 00 if for $i = m - (2^\ell - 1)$, node $i \in V$ is colored with the first color, 01 if i is colored with the second color, and 10 if i is colored with the third color. The construction is illustrated in Figure 1 on a simple example graph G with 5 nodes.

To show that T is a feasible solution for the generalized group testing problem we prove that for any hyperedges $A, A' \in E_H$ there exists $1 \leq t \leq \ell + 2$ such that $y_{t,A} \neq y_{t,A'}$. First assume that A and A' satisfy one of the following cases for $m \neq m'$: $A = \{1, \ldots, n, e_m\}$, $A' = \{1, \ldots, n, e_{m'}\}$; or $A = \{i, e_m\}$, $A' = \{i', e_{m'}\}$; or $A = \{i, e_m\}$, $A' = \{1, \ldots, n, e_{m'}\}$. In these cases, since the encodings u_m of e_m and $u_{m'}$ of $e_{m'}$ are distinct, there exists an entry t, $1 \leq t \leq \ell$, for which $T_{t,m} \neq T_{t,m'}$. This, in turn, implies (given the construction of T) that $y_{t,A} = \vee_{j \in A} T_{t,j} \neq \vee_{j \in A'} T_{t,j} = y_{t,A'}$.

If $A = \{1, \ldots, n, e_m\}$ and $A' = \{i', e_m\}$, then, by our construction of T there exists $t \in \{\ell + 1, \ell + 2\}$ for which $T_{t, 2^\ell - 1 + i'} = 0$. For such t, there must be an $r \in V$ for which $T_{t, 2^\ell - 1 + r} = 1$, as otherwise the nodes in V are colored by two colors alone. Thus, it is not hard to verify that in this case as well $y_{t,A} \neq y_{t,A'}$.

Finally, if $A = \{i, e_m\}$, $A' = \{i', e_m\}$ where $e_m = (i, i')$, then since i and i' are assigned to distinct colors in G we get that there exists $t \in \{\ell + 1, \ell + 2\}$ such that $T_{t, 2^\ell - 1 + i} \neq T_{t, 2^\ell - 1 + i'}$. As above, in this case as well $y_{t,A} \neq y_{t,A'}$. Therefore any two subsets of E_H can be separated.

For the second direction, assume that $\mathcal{T}$ is a solution for the generalized group testing problem with $|\mathcal{T}| = \ell + 2$, and let T be the corresponding $(\ell + 2) \times (2^\ell - 1 + n)$ matrix. We show that G is 3-colorable. Denote by $T^{col}_1, \ldots, T^{col}_{2^\ell - 1 + n}$ the columns of T, and let w be the vector corresponding to the union of the last n columns in T. Namely $w = T^{col}_{2^\ell} \vee, \ldots, \vee T^{col}_{2^\ell - 1 + n}$ and $w_t = 1$ if and only if there exists $m \geq 2^\ell$ for which $T_{t,m} = 1$. Then, it holds that the support $|\mathrm{sup}(w)|$ of w is at most of size 2. Assume to the

contrary that $|\mathtt{sup}(w)| \geq 3$, and without loss of generality that $w_\ell = w_{\ell+1} = w_{\ell+2} = 1$. As there are $2^\ell - 1$ edges e_m in E, we have that tests T_1 up to $T_{\ell-1}$ cannot separate at least one pair of subsets A, A' in E_H of the form $\{1,\ldots,n,e_m\}$ and $\{1,\ldots,n,e_{m'}\}$. This fact follows from the lower-bound of $\log_2|E|$ given in Claim 1. Moreover, these same subsets A and A' cannot be separated by tests $T_\ell, T_{\ell+1}$, and $T_{\ell+2}$ by our assumption that $w_\ell = w_{\ell+1} = w_{\ell+2} = 1$. Thus T is not a feasible solution to the generalized group testing problem corresponding to H, a contradiction.

We can now assume that $|\mathtt{sup}(w)| \leq 2$. Without loss of generality assume that entries 1 to ℓ of w are 0. Then for every $1 \leq i \leq n$ it holds that $|\mathtt{sup}(T^{col}_{2^\ell-1+i})| < 2$, as otherwise both $T_{\ell+1,2^\ell-1+i} = 1$ and $T_{\ell+2,2^\ell-1+i} = 1$ implying that for any $e = (i,j) \in E$ the subsets $A = \{1\ldots,n,e\}$ and $A' = \{i,e\}$ can not be separated. For a coloring of G, assign to node i in V the color corresponding to the values in $T_{\ell+1,2^\ell-1+i}, T_{\ell+2,2^\ell-1+i}$. This implies a legal coloring of G using the colors 00,01,10. Specifically, if $e = (i,j)$ then there exists a test $T_t \in \mathcal{T}$ that separates $A = \{i,e\}$ and $A' = \{j,e\}$ implying that $T_{t,2^\ell-1+i} \neq T_{t,2^\ell-1+j}$, which in turn guarantees that i and j are assigned to distinct colors. ■

We next extend Theorem 1 to establish the hardness of approximation of the size of an optimal group testing solution. Namely, we address the question of approximating the size of the optimal solution for the generalized group testing problem within a multiplicative factor of $1+\varepsilon$ (for small ε). We again reduce from 3-colorability, using the fact that coloring a 3-colorable graph with n^ε colors (for sufficiently small ε) is a well known open problem (see e.g. [7], [8] and references therein). Our proof appears in the full version of this work [24], and closely follows that of Theorem 1.

Theorem 2 *Let $\varepsilon > 0$ be sufficiently small. Approximating the size of an optimal solution for the (non-adaptive) generalized group testing problem within a multiplicative factor of $1+\varepsilon$ is as hard as coloring a 3-colorable graph with at most $n^{4\varepsilon}$ colors.*

B. Non-adaptive generalized group testing bounds

In the following theorem we present an upper bound for the non-adaptive generalized group testing problem. The stated bound is a function of the set-system size $|E|$, the maximum size d of any $e \in E$, and a parameter β which addresses the maximum pair-wise symmetric difference size of any two subsets in E. Loosely speaking, symmetric difference is an important and natural primitive, since if $|e \setminus e'|$ or $|e' \setminus e|$ is *large*, then it is *easier* for a test to separate e from e'. Implying that less tests may be needed when such sizes are large. Using a rough analog from coding theory, one can view the tests $\mathcal{T}$ as a syndrome based decoding process, and the (indicator vector of) edges e as codewords. In this analog, $\max\{|e \setminus e'|, |e' \setminus e|\}$ is related to distance, implying intuitively that a collection of codewords with large minimum distance is easier to decode than a collection with small minimum distance. In what follows, we set $\beta = \min_{e,e' \in E} \max\{|e \setminus e'|, |e' \setminus e|\}$. For $\beta = \Theta(d)$ our solution size matches the information-theoretic lower bound of $\log_2|E|$ stated in Claim 1 (up to a constant multiplicative factor). In the case in which no assumptions are made on β, our solution of size $O(d\log_2|E|)$ comes close to matching our (worst-case) lower bound of $\Omega(d\log|E|/\log d)$. Our solution is generated by constructing tests at random, refining the analysis appearing in, e.g., [25], addressing traditional group testing. All detailed proof appears in the full version of this work [24].

Theorem 3 *Consider an instance of the generalized group testing problem with corresponding hypergraph $H = (V,E)$ with edges of size at most d. In addition, assume that for all $e, e' \in E$ it holds that $\max\{|e \setminus e'|, |e' \setminus e|\} \geq \beta$, for some parameter $1 \leq \beta < d$. Then there exists a solution $\mathcal{T}$ to the generalized group testing problem corresponding to H of size $O(\frac{d}{\beta}\log|E|)$. Moreover, for any constant α, there exists an efficient randomized construction of $\mathcal{T}$ of size $O(\frac{d}{\beta}(\log|E| + \alpha))$ that is a valid solution with probability at least $1 - e^{-\alpha}$.*

Corollary 1 *Consider an instance of the generalized group testing problem with corresponding hypergraph $H = (V,E)$ for which all edges are of size at most d. Then there exists a solution $\mathcal{T}$ of size $O(d\log|E|)$.*

Claim 5 *Let $d \leq n' \leq n$ be integers. There exist hypergraphs $H = (V,E)$ with $|V| = n$, edges of size d, and $|E| = \binom{n'}{d}$ such that any non-adaptive solution for the generalized group testing problem corresponding to H requires $\Omega(\min\{n', d\log|E|/\log d\}) = \Omega(\min\{n', d^2\log n'/\log d\})$ tests.*

IV. Adaptive generalized group testing

We now address the adaptive setting. Adaptive schemes improve on non-adaptive constructions as the adaptive iterative process allows to *rule-out* potential contaminated items or subsets of the population in each iteration. While, in traditional group testing, the adaptive analysis is commonly governed by the *ground-set* size of potentially contaminated individuals, which shrinks with each iteration of the iterative process (see, e.g. [5], [26]), the analysis in our generalized setting must be governed by the number of remaining *edges* that are potentially contaminated. This difference causes a number of challenges that are addressed in the proof below.

Theorem 4 *Consider an instance $H = (V,E)$ of the generalized group testing problem in which edges in E are of size d. There is an adaptive algorithm that interactively designs a collection of tests $\mathcal{T}$ of size at most $O(\log|E| + d\log^2 d)$ that can reliably recover the contaminated set in E. Moreover, there exists instances $H = (V,E)$ with corresponding bound d for which any adaptive solution is of size at least $\Omega(\log|E| + d)$.*

Proof: The proof of the lower bound is given in the full version of this work [24]. For the upper bound, let $V = \{1,2,\ldots,n\}$ (that is, $|V| = n$), and for $1 \leq i \leq n$, let d_i be the degree of node i in V. Notice that $\sum_{i=1}^n d_i = d \cdot |E|$. In what follows, we present an adaptive algorithm using $O(\log|E| + d\log^2 d)$ tests. Roughly speaking, the algorithm

proceeds in rounds in which we search for subsets $T \subseteq V$ that intersect a constant fraction of the edge set. In round j, the constant fraction is required to be in the range $[\varepsilon_j, 1-\varepsilon_j]$, for $\varepsilon_j = 1/2^{j+1}$. Once such a subset T is found and used as a test, we are able to reduce the edge size of the instance at hand by a factor if $(1-\varepsilon_j)$ and thus to *make progress* towards finding the contaminated $e \in E$.

Each round j consists of several sub-rounds in which one consecutively finds subsets T corresponding to the same parameter ε_j, thus reducing the size of the edge set in each sub-round by a factor of $(1-\varepsilon_j)$. We move from round j to round $j+1$ once no such T corresponding to ε_j is found (and thus increase j by 1 allowing for additional flexibility in the requirements on T). Finally, once ε reaches $1/d$, we stop the iterative process and solve the remaining instance in a non-adaptive manner. The algorithm is presented below:

Algorithm 1 (Adaptive algorithm)

1) $E' \longleftarrow E$, $\varepsilon \longleftarrow 1/4$.
2) *If* $|E'| = 1$, *return* E'.
3) *Label the nodes in* V *such that* $d_1 \leq d_2 \leq \ldots \leq d_n$ *where* d_i *be the degree of node* i *in* V *w.r.t.* E'.
4) *If there exists a subset* $T \subseteq V$ *of the form* $T = \{1, \ldots t\}$ *(i.e.,* T *takes nodes in increasing order of degree) such that the number of edges of* E' *intersecting* T *is in the range* $[\varepsilon|E'|, (1-\varepsilon)|E'|]$ *then:*
   a) *Perform test with subset* T.
      i) *If the test is negative, set* $E' \longleftarrow \{e \in E' | e \cap T = \emptyset\}$ *(i.e., update* E' *to the subset of edges that do not intersect* T*), and return to (2).*
      ii) *If the test is positive, then set* $E' \longleftarrow \{e \in E' | e \cap T \neq \emptyset\}$ *(i.e., update* E' *to be the subset of edges that do intersect* T*) and return to (2).*
   b) *Else, if no such* T *is found, if* $\varepsilon > 1/d$ *then return to (2) with* $\varepsilon \longleftarrow \varepsilon/2$. *Otherwise, continue with* E'*, using the non-adaptive testing described in Theorem 3 to find the contaminated* $e \in E$.

We first address the correctness of the proposed algorithm. Let $e \in E$ be the contaminated subset. It suffices to show that throughout the execution of our algorithm, the subset e is in E'. This holds initially as $E' = E$, and continues to hold in Step 4a by the fact that T is positive if and only if $e \cap T \neq \emptyset$.

We now compute the number of tests performed by the algorithm. Consider round j in which $\varepsilon = \varepsilon_j = 1/2^{j+1}$. We first show that if T is not found in Step 4 then it must be the case that E' is of size at most $(1+2\varepsilon_j)2^{d(1+4\varepsilon_j)H(2\varepsilon_j)}$. Specifically, if there is no test T as described in Step 4, then it must be the case that for some node i the number of edges intersecting $\{1,2,\ldots i-1\}$ is less than $\varepsilon_j|E'|$ and the number of edges intersecting $\{1,2,\ldots i\}$ is more than $(1-\varepsilon_j)|E'|$. This implies that $d_i \geq (1-\varepsilon_j)|E'| - \varepsilon_j|E'| = (1-2\varepsilon_j)|E'|$. As $d_i \leq d_{i+1} \leq \ldots \leq d_n$, and $\sum_{\ell=1}^{n} d_\ell = d|E'|$ it holds that $(n-i+1)(1-2\varepsilon_j)|E'| \leq \sum_{\ell=i}^{n} d_\ell \leq d|E'|$. Therefore, $(n-i+1)(1-2\varepsilon_j) \leq d$, so $n-i+1 \leq d/(1-2\varepsilon_j)$.

Recall that less than $\varepsilon_j|E'|$ edges are adjacent to nodes $\{1,\ldots,i-1\}$. This implies that at least $(1-\varepsilon_j)|E'|$ edges are induced by nodes $\{i,\ldots,n\}$. As $|\{i,\ldots,n\}| = n-i+1 \leq d/(1-2\varepsilon_j)$, the number of edges of size d induced by $\{i,\ldots,n\}$ is at most $\binom{d/(1-2\varepsilon_j)}{d}$. This implies that $(1-\varepsilon_j)|E'| \leq \binom{d/(1-2\varepsilon_j)}{d} = \binom{d/(1-2\varepsilon_j)}{2\varepsilon_j \cdot d/(1-2\varepsilon_j)}$, so $|E'| \leq \frac{2^{d\cdot H(2\varepsilon_j)/(1-2\varepsilon_j)}}{1-\varepsilon_j} \leq (1+2\varepsilon_j)2^{d(1+4\varepsilon_j)H(2\varepsilon_j)}$.

To bound the number of tests in the adaptive algorithm, it suffices to analyze the total number of sub-rounds executed. Let m_j be the size of E' in the beginning of round j. From the above, we have that $m_1 = |E|$, and $m_j \leq (1+2\varepsilon_j)2^{d(1+4\varepsilon_j)H(2\varepsilon_j)}$, for any $j = 1,2,\ldots,\log d$. Moreover, as the size of E' in any sub-round of round j reduces its size by a factor of $1-\varepsilon_j \leq \frac{1}{1+\varepsilon_j}$. The total number of sub-rounds is thus bounded by

$$\begin{aligned}
\sum_{j=1}^{\log d} \frac{\log(m_j) - \log(m_{j+1})}{\log(1+\varepsilon_j)} &\leq \sum_{j=1}^{\log d} (\log(m_j) - \log(m_{j+1}))\frac{1}{\varepsilon_j} \\
&\leq 4\log|E| + \sum_{j=2}^{\log d} \log(m_j)\cdot 2^{j-1} \\
&\leq O\left(\log|E| + \sum_{j=2}^{\log d} dH(2\varepsilon_j)\cdot 2^j\right) \\
&= O\left(\log|E| + \sum_{j=2}^{\log d} d\cdot 2\varepsilon_j \log(1/2\varepsilon_j)\cdot 2^j\right) \\
&= O\left(\log|E| + \sum_{j=2}^{\log d} d\cdot j\right) = O\left(\log|E| + d\log^2 d\right).
\end{aligned}$$

Finally, to bound the total number of tests in the suggested algorithm, we add the number of tests performed in step 4b for the set E' once $\varepsilon_j \leq 1/d$. At this stage, the set E' is of size at most $(1+2\varepsilon_j)2^{d(1+4\varepsilon_j)H(2\varepsilon_j)} \leq 2^{O(d\cdot\frac{1}{d}\log d)} = d^{O(1)}$. Thus, the non-adaptive testing described in Theorem 3 when applied to E' will use at most $O(d\log|E'|) = O(d\log d)$ tests. We conclude that, all in all, the suggested adaptive algorithm uses $O\left(\log|E| + d\log^2 d\right)$ tests. ■

We remark that an adaptive upper-bound of $O(\log|E| + d^2)$ holds, using a similar proof, for the modified setting in which edges of E are of size at most d (as apposed to exactly d).

V. Concluding remarks

In this paper we consider a generalization of the traditional group testing problem in which the contaminated set is one of a collection of subsets characterized by the edge set of a hypergraph $H = (V,E)$. Leveraging this additional knowledge, we address both adaptive and non-adaptive group testing, in terms of upper and lower bounds, and for the latter, analyze the complexity of determining the size of optimal (or approximately optimal) solutions.

Beyond adaptive and non-adaptive group testing, several additional models have been studied for traditional group testing. These include, noisy group testing, probabilistic group testing, partial recovery, non-binary outcomes, an unknown bound on d, and more. Addressing those models in our generalized setting is subject to future research.

REFERENCES

[1] J. Wolf. Born again group testing: Multiaccess communications. *IEEE Transactions on Information Theory*, 31(2):185–191, 1985.

[2] R. Clifford, K. Efremenko, E. Porat, and A. Rothschild. Pattern matching with don't cares and few errors. *Journal of Computer and System Sciences*, 76(2):115–124, 2010.

[3] E. Porat and A. Rothschild. Explicit nonadaptive combinatorial group testing schemes. *IEEE Transactions on Information Theory*, 57(12):7982–7989, 2011.

[4] M. Aldridge, O. Johnson, and J. Scarlett. Group testing: an information theory perspective. *CoRR*, abs/1902.06002, 2019.

[5] D.Z Du and F.K. Hwang. *Combinatorial Group Testing and Its Applications*, volume 12 of Series on Applied Mathematics. Singapore: World Scientific Publishing Co. Inc., River Edge, NJ, 2nd edition, 2000.

[6] P. Nikolopoulos, S.R 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*, volume 130, pages 2341–2349. PMLR, 2021.

[7] S. Khanna, N. Linial, and S. Safra. On the Hardness of Approximating the Chromatic Number. *Combinatorica*, 20:393–415, 2000.

[8] M. Langberg. Graph coloring. In *Encyclopedia of Algorithms*, pages 368–371. Springer, 2008.

[9] P. Nikolopoulos, S.R Srinivasavaradhan, T. Guo, C. Fragouli, and S. Diggavi. Group testing for overlapping communities. In *Proceedings of the IEEE International Conference on Communications*, 2021.

[10] S. Ahn, W.N. Chen, and A. Ozgur. Adaptive group testing on networks with community structure. In *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021.

[11] B. Arasli and S. Ulukus. Graph and cluster formation based group testing. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1236–1241, 2021.

[12] P. Bertolotti and A. Jadbabaie. Network group testing. *Manuscript; available on https://arxiv.org/abs/2012.02847*, 2020.

[13] Y.J. Lin, C.H. Yu, T.H. Liu, C.S. Chang, and W.T. Chen. Positively correlated samples save pooled testing costs. *Manuscript; available on https://arxiv.org/abs/2011.09794*, 2020.

[14] J. Zhu, K. Rivera, and D. Baron. Noisy pooled pcr for virus testing, 2020.

[15] R. Goenka, S.J Cao, C.W Wong, A. Rajwade, and D. Baron. Contact tracing enhances the efficiency of covid-19 group testing. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8168–8172, 2021.

[16] M. Cheraghchi, A. Karbasi, S. Mohajerzefreh, and V. Saligrama. Graph-constrained group testing. *IEEE Transactions on Information Theory*, 58(1):248–262, 2010.

[17] A. Karbasi and M. Zadimoghaddam. Sequential group testing with graph constraints. In *2012 IEEE Information Theory Workshop*, pages 292–296, 2012.

[18] S. Luo, Y. Matsuura, Y. Miao, and M. Shigeno. Non-adaptive group testing on graphs with connectivity. *Journal of Combinatorial Optimization*, 38(1):278–291, 2019.

[19] M. Hahn-Klimroth and P. Loick. Optimal adaptive group testing. *Manuscript; available on https://arxiv.org/abs/1911.06647*, 2019.

[20] M. Aldridge, O. Johnson, and J. Scarlett. Group testing: An information theory perspective. *Found. Trend. Comms. Inf. Theory*, 15(3-4):196–392, 2019.

[21] O.T Johnson. Strong converses for group testing from finite block length results. *IEEE Trans. Inf. Theory*, 63(9):5923–5933, 2017.

[22] P. Erdos, P. Frankl, and Z. Furedi. Families of finite sets in which no set is covered by the union of *r* others. *Israel J. Math*, 51:79–89, 1985.

[23] E. Porat and A. Rothschild. Explicit non-adaptive combinatorial group testing schemes. In *Automata, Languages and Programming, 35th International Colloquium, ICALP*, pages 748–759, 2008.

[24] M. Gonen, M. Langberg, and A Sprintson. Group Testing on General Set-Systems. *Manuscript, available on arXiv.org*, 2022.

[25] N. Alon, D. Moshkovitz, and S. Safra. Algorithmic construction of sets for k-restrictions. *ACM Transactions on Algorithms*, 2(2):153—-177, 2006.

[26] L. Baldassini, O. Johnson, and M. Aldridge. The capacity of adaptive group testing. In *IEEE International Symposium on Information Theory, ISIT*, pages 2676–2680, 2013.