Group Testing with Correlation via Edge-Faulty Graphs

Hesam Nikpey, Jungyeol Kim, Xingran Chen, Saswati Sarkar, Shirin Saeedi Bidokhti
University of Pennsylvania
Email: {Hesam, Jungyeol, Xingranc, Swati, Saeedi}@seas.upenn.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834777

***Abstract*—In applications of group testing in networks, e.g. identifying individuals who are infected by a disease spread over a network, exploiting correlation among network nodes provides fundamental opportunities in reducing the number of tests needed. We model and analyze group testing on n correlated nodes whose interactions are specified by a graph G. We model correlation through an edge-faulty random graph formed from G in which each edge is dropped with probability $1-r$, and all nodes in the same component have the same state.**

We consider three classes of graphs: cycles and trees, d-regular graphs, and stochastic block models or SBM, and obtain lower and upper bounds on the number of tests needed to identify the defective nodes. Our results are expressed in terms of the number of tests needed when the nodes are independent and they are in terms of n, r, and the target error. In particular, we quantify the fundamental improvements that exploiting correlation offers by the ratio between the total number of nodes n and the equivalent number of independent nodes in a classic group testing algorithm.

The lower bounds are derived by illustrating a strong dependence of the number of tests needed on the expected number of components. In this regard, we establish a new approximation for the distribution of component sizes in "d-regular trees" which may be of independent interest and leads to a lower bound on the expected number of components in d-regular graphs.

The upper bounds are found by forming dense subgraphs in which nodes are more likely to be in the same state. When G is a cycle or tree, we show an improvement by a factor of $log(1/r)$. For grid, a graph with almost $2n$ edges, the improvement is by a factor of $(1-r)\log(1/r)$, indicating drastic improvement compared to trees. When G has a larger number of edges, as in SBM, the improvement can scale in n.

I. Introduction

Group testing [1] is a well studied problem at the intersection of many fields, including computer science [2]–[6], information theory [7]–[9] and computational biology [10], [11]. The goal is to find an unknown subset of n items that are different from the rest using the least number of tests. The target subset is often referred to as *defective*, corrupted or infected, in this work we use the term defective. To find the subset of defectives, items are tested in groups. The result of a test is positive if and only if at least one item in the group is defective. Group testing is beneficial when the number of defective items is $o(n)$, often assumed that the (expected) number of defective items is $n^\alpha, \alpha < 1$.

Over the years, this problem has been formulated via two approaches: the combinatorial approach and the information theoretic approach. In the "combinatorial" version of the problem, it is assumed that there are d defective items that are to be detected with zero error [1]. Using adaptive group testing (i.e., when who to test next depends on the results of the previous tests), there is a matching upper and lower bound on the number of tests in the form $d\log n + O(d)$ [1]. Using non-adaptive group testing (i.e., when the testing sequence is pre-determined), there is an upper bound of $O(d^2\log(n/d))$ and an almost matching lower bound of $\Omega(\frac{d^2\log n}{\log d})$. The "information theoretic" approach, on the other hand, assumes a prior statistic on the defectiveness of items, i.e., item i is assumed to be defective with probability p_i. The aim in this case is to identify the defective set with a high probability [12]. Roughly speaking, there is a lower bound in terms of the underlying entropy of the unknowns, and an almost matching upper bound up to a $\log n$ factor of the lower bound.

In most existing works, it is assumed that the state of the items are independent from each other, which is not realistic in many applications. Group testing for example can identify the infected individuals using fewer tests, and therefore in a more timely manner, than individual testing, during the spread of an infectious disease (eg, COVID-19) [13]–[17]. But the infection state of individuals are in general correlated, with correlation levels ranging from high to low, depending on how close they live: same household (high), same neighborhood, same city, same country (low). Correlation levels also depend on other factors such as frequency of contact, the number of paths between the individuals in the network of interactions. We elaborate on this further in Section I-A. With this motivation, we aim to model such correlation, design group testing techniques that exploit it, and quantify the gain they provide in reducing the number of tests needed.

Some recent papers have designed and analyzed group testing under specific correlation models, e.g., [18]–[20]. In [18], the authors consider correlation that is imposed by a one day spread of an infectious disease in a clustered network modeled by a stochastic block model (SBM). Each node is initially defective (infected) with some probability and its neighbors become defective probabilistically in the next day. The authors provide a simple adaptive algorithm and prove optimality in some regimes of operation and under some assumptions. In [19], the authors model correlation through a random edge-faulty graph and design novel near-optimal group testing techniques for a specific subset of the realizations of the correlation graph. We consider a similar correlation model. Through a different approach, we have been able to take a significant leap forward in that we can accommodate

all possible realizations of the graph.

In other related works, a graph could represent potential constraints on testing [21], [22]. In [21], the authors design optimal non-adaptive testing strategies where each group should be path connected. In particular, they use random walks to obtain the pool of tests. In a follow up work, [22] shows that either the constraints are too strong and no algorithm can do better than testing most of the nodes, or optimal algorithms can be designed matching with the unconstrained version of the problem. They attain sampling each edge with an optimized probability r . If a connected component is large enough, the algorithm tests the entire component. Our approach in this paper has similarities with [22] in aiming to find parts of the graph that are large and connected enough so that they remain connected with a decent probability after realizing the edges.

A. Our Model

We start by motivating the key attributes we capture in our model, and consider the interaction network for spread of COVID-19 in a network of people (nodes) towards that end. There is an edge between two nodes if the corresponding individuals are in physical proximity for a minimum amount of time each week. Such individuals are more likely to be in the same state than those who have been distant throughout. Thus, firstly, the probability of being in the same state decreases with increase in the length of paths (i.e., distance in interaction network) between nodes. Second, infection is more likely to spread from one node to another if there are many distinct paths between them. Thus, the probability that two nodes are in the same state increases with the increase in the number of distinct paths between them.

We capture correlation through a faulty-edge graph model. Consider a graph $G = (V, E)$ where the node set V represents the items and the edge set E represents connections/correlations between them. Suppose each edge is realized with probability $0 \leq r \leq 1$. After the sampling, we have a random graph that we denote by G_r. Each node is either defective or non-defective. All nodes in the same component of G_r are in the same state, rendering defectiveness a component property. We consider that each component is defective with probability (w.p.) p independent of others. As an example, consider graph G with five nodes and eight edges, and a sampled graph realization G_r as shown in Figure 1 (left) and Figure 1 (right) respectively. When $r = 1/3$, G_r is realized w.p. $(\frac{1}{3})^3(\frac{2}{3})^5$. There are two components in G_r, namely, v_1, v_4, v_5 and v_2, v_3; v_1, v_4, v_5 are in the same state, which is defective w.p. p, independent of the state of v_2, v_3.

This model importantly captures the two attributes we discussed: Clearly, a long path between two nodes in G has a smaller chance of survival in G_r, compared to a short path, making the end nodes less likely to be in the same state as the length of the path in G between them increases. Moreover, the probability that at least one path between two nodes survive in G_r increases with increase in the number of distinct paths between them in G, so having distinct paths between a pair of nodes in G makes them more likely to be in the same state.

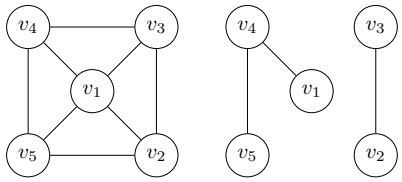

Fig. 1: Left: Graph G; Right: Graph G_r.

We aim to find the minimum expected number of tests needed to find the defective items with at most ϵn errors (or sometimes referred to as error of ϵn), where ϵ can potentially be $o(1)$. To be precise, let #ERR(H) be the number of nodes mispredicted by an algorithm on graph H. Then we require to have $\mathbb{E}_{H\sim G_r}[\#ERR(H)] \leq \epsilon n$ where the expectation is taken over G_r and possible randomization of the algorithm.

Our approach is to relate the problem to an equivalent independent group testing problem with fewer nodes and provide a basis for comparison and quantification of the improvements that our methods offer by exploiting correlation. The tests can not be designed with the knowledge of G_r, only the value of r is known apriori. In the extreme case of $r = 0$, the problem is reduced to the classic group testing with $|V|$ independent nodes. If $r = 1$, all components of G remain connected and hence the problem is reduced to independent group testing with only components of G. When $0 < r < 1$, the problem is non-trivial, because there can be multiple components, some with more than one node, and the number and composition of the components is apriori unknown. Thus, it is not apriori known which nodes will be in the same state. Our group testing strategies will seek to circumvent this challenge by identifying parts of G that are connected enough so that they remain connected in G_r with a high probability.

We use the following **notations** for the rest of the paper. Let $\textsc{CrltOPT}(G, r, p, \epsilon)$ be the expected number of tests in an optimal algorithm on graph G with parameters r, probability of defectiveness p, and an error of ϵn. Note that we use expected number of tests here as even if our algorithm is deterministic, we will use classic group testing as a black-box which is a randomized algorithm. Let $\textsc{IndepOPT}(n, p, \epsilon)$ be the minimum expected number of tests needed for n items in order to find the defective set with the error probability at most ϵ, where each item is *independently* defective with probability p. It is noteworthy to mention that the definition of error in IndepOPT is different from CrltOPT. In their design of IndepOPT, they ensure that with probability $1 - \epsilon$ all nodes are predicted correctly, and with probability ϵ at least one node is mispredicted, which is also the error defined in IndepPOT in our notation. When clear from the context, we may drop p, r, ϵ from the notations.

B. Contributions

We obtain upper and lower bounds on the number of group tests needed to determine the states, defective or otherwise, of individual nodes in a large class of interaction graphs in presence of correlation among the states of nodes. We

progressively consider 1) cycles and trees (about n links), 2) $d-$regular graphs (about $dn/2$ links) and 3) stochastic block models or SBM ($\Theta(n^2)$ links). The correlation is captured by the factor r (see Section I-A). The bounds are obtained in terms of the number of tests needed when the states are independent, and help us quantify the efficiency brought forth by group testing in terms of r.

For trees and cycles, we prove an upper bound on the optimal number of tests in terms of the number of group tests when there are $n\log(1/r)$ independent nodes. Note that one can trivially determine the states of each node by disregarding correlation and testing among n nodes (e.g. using classic group testing techniques). Our upper bound therefore shows that group testing can reduce the tests by a factor of $\log(1/r)$, which is less than 1 when $r > 1/2$. As r approaches 1 the multiplicative factor reduces even further implying even greater benefits due to group testing. Our lower bound, on the other hand, shows an improvement factor $(1-r)$.

For $d-$regular graphs we prove new bounds for the distribution of components. This leads to a lower bound that is expressed as a sum series depending on r and n. We further prove an upper bound for a specific 4-regular graph, namely grid, in terms of the number of group tests when there are $n(1-r)\log(1/r)$ independent items. Thus, the improvement factor is $(1-r)\log(1/r)$, as opposed to only $\log(1/r)$ for trees; this hints us that group testing gets more efficient drastically for denser graphs.

SBM divides network into communities such that nodes in the same community are more connected than nodes in different communities. We show that the reduction in the test count due to group testing can be classified into three regimes: 1) strong intra-community connectivity but sparse inter-community connectivity, which reduces the effective number of independent nodes to the number of communities, 2) fully connected graph, thus, all nodes have the same state 3) most of the nodes are isolated, thus states of all nodes are independent. The number of tests in 1) and 3) can be determined from the characterizations of networks in which all nodes are independent, and only one test is necessary in 2).

II. A Lower Bound for Sparse Graphs

In this section, we give lower bounds for the number of tests needed when the underlying graph has $o(n^2)$ edges. Cycles (n edges) and trees ($n-1$ edges) belong to this category for example. We obtain these bounds by reducing the problem to the number of tests needed for the independent case.

Let $C(G_r)$ be the number of components of G_r. Then:

Lemma 1. *For each realization of G_r with $C(G_r)$ components, we have*

$$\textsc{IndepOPT}(C(G_r), p, \epsilon n) \leq \textsc{CrltOPT}(G, r, p, \epsilon).$$

Remark 1. *Note that the bound is non-trivial when $\epsilon < 1/n$, and this error or smaller errors can be satisfied in the algorithms in [12].*

We now obtain a concentration result on the random variable $C(G_r)$ in terms of the number of edges m.

Lemma 2. *Let $\delta > 0$. With probability $1-\delta$ we have*

$$|C(G_r) - \mathbb{E}[C(G_r)|] \leq O(\sqrt{m\log 1/\delta}).$$

Specifically, when $\mathbb{E}[C(G_r)] = cn$ for a constant c, and m is $o(n^2)$, with high probability the number of components is within $cn \pm o(n)\sqrt{\log 1/\delta}$.

The above concentration result has been obtained from an application of classical results on *Edge Exposure Martingale* and *Node Exposure Martingale* defined on Graphs (in [23]). Roughly speaking, the edges (respectively, nodes) of the graph are exposed sequentially and martingales are defined by applying a desired graph function to the exposed set of edges (respectively, nodes). Using these martingales, concentration results can be obtained for any function of the graph with a desired edge Lipschitz condition [23]. Considering the number of components as the graph function, the above lemma follows from these classical concentration results (see [23, Chapter 7]).

We can now obtain the following specialized result for a cycle or a tree:

Theorem 1. *For a cycle or a tree, G,*

$$\textsc{IndepOPT}((1-r)n - 10\sqrt{n\log n}, p, \epsilon n)$$
$$\leq \textsc{CrltOPT}(G, r) + O(1/n).$$

Proof. In a tree, by removing each edge we get one more component, so after removing k edges the tree and the cycle respectively has $k+1$ and k components.

G_r is obtained by removing each edge in G w.p. $1-r$, so $\mathbb{E}[C(G_r)|]$ is $1+(1-r)(n-1)$ for trees, and $1+(1-r)n$ for cycles. By Lemma 2 and $\delta = 1/n^2$, $C(G_r)$ is $(1-r)n \pm O(\sqrt{n\log n})$ with probability $1-1/n^2$. The difference in tests is at most n, and since the difference from $(1-r)n$ exceeds $O(\sqrt{n\log n})$ w.p. $1/n^2$, the expected difference is $O(1/n)$. Applying Lemma 1 thus completes the proof. □

III. An Upper Bound for Graphs with a Few Edges: Cycles and Trees

In this section, we provide algorithms to find the defective items and provide theoretical bounds. We start with a simple cycle, and subsequently generalize the method to devise algorithms for any tree. Note that an algorithm for arbitrary trees provides potentially suboptimal algorithms for general graphs, by just considering a spanning tree of it.

First, we provide an algorithm for cycles.

1) Let $l = \max\{\frac{\log[1/(1-\epsilon/2)]}{\log 1/r}, 1\}$. Partition the cycle into $\lceil n/l \rceil$ paths $P_1, P_2, \ldots, P_{\lceil n/l \rceil}$ of the same length l, except one path that may be shorter.
2) for each path, choose one of its nodes at random and let the corresponding nodes be $v_{P_1}, v_{P_2}, \ldots v_{P_{\lceil n/l \rceil}}$.
3) Use an $\textsc{IndepOPT}(\lceil n/l \rceil, p, \epsilon')$ algorithm (by [12, Theorem 2] for adaptive or [12, Theorem 4] for non-adaptive group testing) to find the defective items among

$v_{P_1}, v_{P_2}, \ldots v_{P_{\lceil n/l \rceil}}$ where $\epsilon' < \frac{\epsilon}{2}$ and the probability of being defective equals to p.

4) Assign the state of all the nodes in P_i as v_i for all i.

Note that for each i, the defectiveness probability of v_i is p. The probability that P_i is actually connected after a realization is r^{l-1}. So the probability that P_i is not in the same state of v_i is $1-r^{l-1}$. Then assuming that we detect all v_i's correctly, the error in G is at most $\lceil n/l \rceil \cdot (1-r^{l-1}) \cdot l$. By replacing $l = \max\{\frac{\log[1/(1-\epsilon/2)]}{\log 1/r}, 1\}$, the error becomes less than $\epsilon n/2$. But we might also have ϵ' probability of error for the v_i's (given the criteria set in INDEPOPT), means with probability $1-\epsilon'$ all the nodes are predicted correctly, and with probability ϵ' we have at least one mispredicted node, and at most n mispredicted nodes. So the total error from this part is at most $\epsilon' n < \epsilon n/2$. Then we have the following upper bound:

$$\text{CRLTOPT}(Cycle, r, p, \epsilon) \leq \text{INDEPOPT}(\lceil n/l \rceil, p, \epsilon'). \quad (1)$$

We now generalize the ideas for trees. As before, we partition the graph into $\lceil n/l \rceil$ groups of l nodes, find the probability of them being connected in a random realization, and then optimize it over l. At a high level, we try to group nodes that have small paths to each other, as shorter paths remain in the graph with higher probabilities.

Definition 1. *Let $S \subseteq V$ of a graph G. The smallest connecting closure of S is a subset $S' \subseteq V$ such that the induced graph over $S \cup S'$ is connected.*

For example, in Figure 1, if $S = \{v_1, v_5\}$, then $S = \{v_4\}$, as S becomes connected once v_4 is added to it.

Lemma 3. *Let G be a tree with n nodes. Given $l \leq n$ there is a partition of the nodes into $\lceil n/l \rceil$ groups of size l (except one that may have fewer nodes) such that the size of the smallest connecting closure for each group is at most l.*

Theorem 2. *Consider a tree with n nodes and let $l = \max\{\frac{\log[1/(1-\epsilon/2)]}{2\log 1/r}, 1\}$. Let $\epsilon' < \epsilon/2$. Then there is an algorithm that uses* INDEPOPT$(\lceil n/l \rceil, p, \epsilon')$ *tests and finds the defective set with at most $\epsilon \cdot n$ errors. I.e.,*

$$\text{CRLTOPT}(G, r, p, \epsilon) \leq \text{INDEPOPT}(\lceil n/l \rceil, p, \epsilon').$$

Proof. Consider the following algorithm:

1) By Lemma 3, partition the tree into $\lceil n/l \rceil$ groups $g_1, g_2, \ldots, g_{\lceil n/l \rceil}$ of the same length l, one group might be shorter than the other ones.
2) For each group, choose one of its nodes at random and let them be $v_{P_1}, v_{P_2}, \ldots v_{P_{\lceil n/l \rceil}}$.
3) Use an INDEPOPT$(\lceil n/l \rceil, p, \epsilon')$ algorithm to find the defective set among $v_{P_1}, v_{P_2}, \ldots v_{P_{\lceil n/l \rceil}}$.
4) Assign the state of all the nodes in g_i as v_i, for all i.

First, we calculate the probability that g_i is connected. By Lemma 3, we know that each g_i has the property that its smallest connecting closure is less than or equal to l. This ensures that at most l edges (over the edges already in g_i) are needed to to make g_i connected. Therefore, the probability of g_i be connected is at least r^{2l}. So the probability that g_i is not in the same state as v_i is at most $1-r^{2l}$. The rest of the proof revolves around proving that the total error is less than ϵn as was done for cycle and this completes the proof. □

IV. GRAPHS WITH MORE EDGES: GRID AND SBM

In this section, we focus on graphs that potentially have many edges. As the number of edges increases, the correlation between nodes increases even when r is not large. We know that there is a threshold phenomenon in some edge-faulty graphs, meaning that when r is below a threshold, there are many isolated nodes (and hence many independent tests are needed) and when r is above that threshold, we have a giant component (and hence a single test suffices). Most famously, this threshold is $\frac{\log n}{n}$ for Erdős-Rényi graphs. When G is a random d-regular graph, that is, it is drawn uniformly from the set of all d-regular graphs with n nodes, $\frac{1}{d-1}$ is a threshold almost surely in G_r [24].

We first study a (deterministic) 4-regular graph, known as the grid[1] and then provide near-optimal results for the stochastic block model. For deterministic graphs, the threshold results of random d-regular graphs (as in [24]) do not apply because the specific chosen graph may not be among the "good" graphs that constitute the almost sure result. We develop new statistical results on the number of components in the corresponding G_r.

A. The Grid

A grid with n nodes and side length $\sqrt{n}$ is a graph where nodes are in the form of $(a, b) : 1 \leq a, b \leq \sqrt{n}$. Node (a, b) is connected to its four close neighbors (if exist), namely $(a-1, b), (a+1, b), (a, b+1), (a, b-1)$. Border nodes (with $a \in \{1, \sqrt{n}\}$ or $b \in \{1, \sqrt{n}\}$) might have three or two neighbors.

The first step is to obtain a lower bound on the expected number of components in G_r. For this one first seeks the expected component size that nodes would belong to, see [23], [24]. Consider the following process. Pick a node $v \in V(G)$, mark it as processed, and let it be the root of a tree. For each $u \in V(G)$ that is not processed and is a neighbor of v, uv is realized w.p. r and added as a child of v. The same process is repeated for each realized u in a Breath First Search (BFS) order. When the process ends, there is a tree with root v, and the expected size of the tree is the expected size of the component that v ends up in.

An example is show in Figure 2. Node v_{11} is the root (colored in blue), and the children that are realized are in green, and the children that are not realized are in red. The component would be $\{v_{11}, v_{12}, v_7, v_2\}$.

By repeating the process for each node that is not processed yet, we get a spanning forest. The expected number of components in the forest is the expected number of components in G_r. Here, the challenge is that we don't know the number of available (unprocessed) neighbors of a node. It highly depends on the previously chosen nodes, especially when d is small, like in the grid. We circumvent this issue by analyzing an

[1]The degree regularity does not hold on the boundaries of the grid.

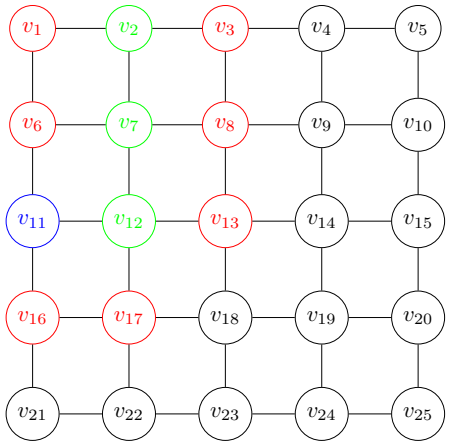

Fig. 2: An example of the procedure described in Section IV-A, starting with border node v_{11}.

infinite regular tree process that effectively corresponds to a more connected graph and therefore leads to the desired lower bound.

1) 3-regular Trees: Consider an infinite tree with root v such that each node in the tree has three children and each edge is realized w.p. r. Let $C(v)$ be the component that v ends up in. The following lemma approximates the distribution of $|C(v)|$.

Lemma 4. *Under the above process and for $t \in \mathbb{N}$,*

$$P(|C(v)| = t) = \frac{1}{2t+1}\binom{3t}{t} r^{t-1}(1-r)^{2t+1}.$$

Proof. Let T be an embedded tree with t nodes. In order for T to be realized, all the edges in T should be realized and the rest of the edges that have an end node in T should not be realized. There are $t-1$ edges in T, and each node has three potential edges, so there are $2t+1$ edges that are not realized. So T is realized w.p. $r^{t-1}(1-r)^{2t+1}$. We now only need C_t, the number of trees with t nodes and v as the root. Using a recursive argument we show that C_t has the same form as of second-order Catalan numbers with solution $C_t = \frac{1}{2t+1}\binom{3t}{t}$. Full proof is provided in the long version [25]. □

In the long version [25] we prove a threshold $1/3$ beyond which the probability of having infinitely large components is non-zero. For $r \le \frac{1}{3}$, the expected component size is:

$$\mathbb{E}(|C(v)|) = \sum_{t=1}^{\infty} \frac{t}{2t+1}\binom{3t}{t} r^{t-1}(1-r)^{2t+1} \qquad (2)$$

The proof generalizes to general $d-$regular tree processes.

2) A Lower Bound for the Grid: In the BFS Spanning Forest of grid, any node in the tree, besides the root, has at most three children. If we choose the root from the border of the grid at each step, the root also has at most three children. Then, a random 3-regular graph is more connected than the trees appear in the process. Therefore the expected component size that we found in (2) provides an upper bound on the expected component size in the grid. So a lower bound on the number of components for 3-regular graphs is a lower bound for the grid. Let NC be the number of connected components as captured by the 3-regular tree process. This process is symmetric over all the nodes, so $\mathbb{E}[NC] = |V(G)|/\mathbb{E}[C(v)]$. Then immediately we have the following result.

Theorem 3. *For a grid with n nodes and $r \le 1/3$, we have*

$$\mathbb{E}(NC) = \frac{n}{\sum_{t=1}^{\infty} \frac{t}{2t+1}\binom{3t}{t} r^{t-1}(1-r)^{2t+1}}$$
$$\simeq \frac{n}{\frac{1-r}{r}\sqrt{\frac{3}{4\pi}}\sum_{t=1}^{\infty}\frac{\sqrt{t}}{(2t+1)}(\frac{27}{4}r(1-r)^2)^t}.$$

Corollary 1. *Similar to Theorem 1, by using Lemma 1 in conjunction with Theorem 3, we get*

$$\textsc{IndepOPT}(\frac{n}{\frac{1-r}{r}\sqrt{\frac{3}{4\pi}}\sum_{t=1}^{\infty}\frac{\sqrt{t}}{(2t+1)}(\frac{27}{4}r(1-r)^2)^t}, p, \epsilon n)$$
$$\le \textsc{CrltOPT}(Grid, r) + O(1/n).$$

3) An Upper Bound for the number of tests in a Grid: We partition the grid into subgrids of length k, where k is to be optimized, and compute the probability of error for each subgrid. We first estimate the probability that a subgrid becomes connected.

Theorem 4. *Let P_k be the probability that a grid of length $k > 1$ becomes connected when each of its edge is realized with probability r. Then we have:*

$$P_k \ge r^{\Theta((1-r)k^2)} = e^{\Theta(\log(r)(1-r)k^2))}.$$

Now similar to Theorem 2, by setting error probability of each group small enough, that is $1 - P_k < \epsilon/2$, we get $k < \sqrt{\frac{\log(1-\epsilon/2)}{(1-r)\log r}}$. Then the error is less than ϵn with at most n/k^2 independent node tests with error probability $\epsilon' < \epsilon/2$.

B. The Stochastic Block Model

A stochastic block model has g clusters of size $k = n/g$, where there exists an edge between any pair of nodes in the same and different cluster w.p. q_1, q_2 respectively, and $q_2 < q_1$. G_r can be described similar to G except that $r_1 = rq_1$ and $r_2 = rq_2$ replace q_1, q_2. Here, we assume that $k \gg \log n$. We prove in [25] the following characterization of the components and therefore the number of tests as follows:

Theorem 5.
- *If $r_1 \ge \frac{100\log n}{k}$ and $1-(1-r_2)^{k^2} \ge \frac{100\log g}{g}$, then with high probability G is connected. (first regime, one test needed)*
- *If $r_1 \ge \frac{100\log n}{k}$ and $1-(1-r_2)^{k^2} \le \frac{1}{100g}$, then with high probability each cluster is connected but most of the clusters are isolated. (second regime, g independent tests needed)*
- *If $r_1 \le \frac{1}{100k}$ and $r_2 \le \frac{1}{100n}$, then with high probability G has many isolated nodes. (third regime, $\Omega(n)$ independent tests needed)*
- *If $r_1 \le \frac{1}{100k}$ and $r_2 \ge \frac{100\log n}{n}$, and $g > 1$, then with high probability G is connected. (fourth regime, one test needed)*

REFERENCES

[1] D. Du, F. K. Hwang, and F. Hwang, *Combinatorial group testing and its applications*. World Scientific, 2000, vol. 12.

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

[3] 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.

[4] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, "Optimal group testing," in *Conference on Learning Theory*. PMLR, 2020, pp. 1374–1388.

[5] M. Cheraghchi, R. Gabrys, and O. Milenkovic, "Semiquantitative group testing in at most two rounds," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1973–1978.

[6] M. Cuturi, O. Teboul, Q. Berthet, A. Doucet, and J.-P. Vert, "Noisy adaptive group testing using bayesian sequential experimental design," *arXiv preprint arXiv:2004.12508*, 2020.

[7] M. Aldridge, O. Johnson, and J. Scarlett, "Group testing: an information theory perspective," *arXiv preprint arXiv:1902.06002*, 2019.

[8] P. Bertolotti and A. Jadbabaie, "Network group testing," 2021.

[9] M. T. Goodrich and D. S. Hirschberg, "Improved adaptive group testing algorithms with applications to multiple access channels and dead sensor diagnosis," *Journal of Combinatorial Optimization*, vol. 15, no. 1, pp. 95–121, 2008.

[10] E. Knill and S. Muthukrishnan, "Group testing problems in experimental molecular biology," *arXiv preprint math/9505211*, 1995.

[11] M. Farach, S. Kannan, E. Knill, and S. Muthukrishnan, "Group testing problems with sequences in experimental molecular biology," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997, pp. 357–367.

[12] T. Li, C. L. Chan, W. Huang, T. Kaced, and S. Jaggi, "Group testing with prior statistics," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 2346–2350.

[13] V. Brault, B. Mallein, and J.-F. Rupprecht, "Group testing as a strategy for covid-19 epidemiological monitoring and community surveillance," *PLoS computational biology*, vol. 17, no. 3, p. e1008726, 2021.

[14] 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*, p. 1205, 2021.

[15] L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L. Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi *et al.*, "A pooled testing strategy for identifying sars-cov-2 at low prevalence," *Nature*, vol. 589, no. 7841, pp. 276–280, 2021.

[16] M. Aldridge, "Conservative two-stage group testing," *arXiv preprint arXiv:2005.06617*, 2020.

[17] C. Gollier and O. Gossner, "Group testing against covid-19," EconPol Policy Brief, Tech. Rep., 2020.

[18] S. Ahn, W.-N. Chen, and A. Ozgur, "Adaptive group testing on networks with community structure," *arXiv preprint arXiv:2101.02405*, 2021.

[19] B. Arasli and S. Ulukus, "Group testing with a graph infection spread model," *arXiv preprint arXiv:2101.05792*, 2021.

[20] P. Nikolopoulos, S. R. Srinivasavaradhan, T. Guo, C. Fragouli, and S. Diggavi, "Group testing for connected communities," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2341–2349.

[21] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama, "Graph-constrained group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 1, pp. 248–262, 2012.

[22] B. Spang and M. Wootters, "Unconstraining graph-constrained group testing," *arXiv preprint arXiv:1809.03589*, 2018.

[23] N. Alon and J. H. Spencer, *The probabilistic method*. John Wiley & Sons, 2016.

[24] A. Goerdt, "The giant component threshold for random regular graphs with edge faults," in *International Symposium on Mathematical Foundations of Computer Science*. Springer, 1997, pp. 279–288.

[25] H. Nikpey, J. Kim, X. Chen, S. Sarkar, and S. Saeedi Bidokhti, "Group testing with correlation via edge-faultygraphs," *arXiv preprint*, 2022.