Adaptive Group Testing on Networks with Community Structure

Surin Ahn, Wei-Ning Chen, and Ayfer Özgür

Abstract

Since the inception of the group testing problem in World War II, one of the prevailing assumptions in the probabilistic variant of the problem has been that individuals in the population are infected by a disease *independently*. However, this assumption rarely holds in practice, as diseases typically spread through interactions between individuals and therefore cause infections to be correlated. Inspired by characteristics of COVID-19 and similar diseases, we consider an infection model over networks which generalizes the traditional i.i.d. model from probabilistic group testing. Under this infection model, we ask whether knowledge of the network structure can be leveraged to perform group testing more efficiently, focusing specifically on community-structured graphs drawn from the stochastic block model. We prove that when the network and infection parameters are conducive to "strong community structure," our proposed adaptive, graph-aware algorithm outperforms the baseline binary splitting algorithm, and is even order-optimal in certain parameter regimes. We support our results with numerical simulations.

Index Terms

Group testing, infectious diseases, adaptive algorithms, stochastic block model, network community structure

I. Introduction

Identifying individuals who are infected by a disease is crucial for curbing epidemics and ensuring the well-being of society. However, due to high costs or limited resources, it is often infeasible to test every member of the population individually. During World War II, when the U.S. military sought to identify soldiers infected with syphilis, Dorfman introduced the breakthrough concept of *group testing* [1]. He showed that by testing *groups* or *pools* of samples, the infected people in a population of size n can be identified with far fewer than n tests. The key insight was that if the infected population is sparse, then each pooled test is likely to produce a negative result, in which case all individuals included in the test can simultaneously be deemed healthy. Today, group testing strategies are actively being used in the COVID-19 pandemic to identify infected individuals in an efficient and cost-effective manner [2]–[5]. There has also been a recent influx of papers which seek to improve or better understand group testing for COVID-19, e.g., [6]–[14].

Surin Ahn, Wei-Ning Chen, and Ayfer Özgür are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: surinahn@stanford.edu; wnchen@stanford.edu; aozgur@stanford.edu).

This work will be presented in part at the *2021 IEEE International Symposium on Information Theory (ISIT)*.

Dorfman's seminal work and many subsequent works by other authors [15]–[21] assume that the disease infects individuals in a statistically independent fashion. In the simplest case, known as i.i.d. group testing, it is assumed that every individual is infected i.i.d. with some common probability p. However, this assumption of independence rarely holds in practice. Diseases typically spread through *interactions* between individuals (e.g., familial, work-related, or other social interactions), thereby inducing correlated infections. It is thus natural to ask whether exploiting information about this connectivity structure can lead to more efficient group testing strategies [1]. This problem is especially timely given the critical role that group testing has played in the COVID-19 pandemic, and given that the disease is known to spread from close contact between individuals.

In this paper, we contribute to the nascent area of "group testing under correlations" by investigating whether knowledge of the *interaction network* dictating the spread of the disease can be leveraged to perform pooled testing more efficiently. We consider an infection model which generalizes the standard i.i.d. assumption to arbitrary graphs. Our primary focus is on networks drawn from the *stochastic block model* (SBM), which are likely to exhibit community structure: clusters of nodes with relatively dense internal connections.

On the algorithmic side, we consider *adaptive* group testing schemes, where the design of each test can be informed by the previous test results. We compare two different schemes: the standard *binary splitting* algorithm [29] which is oblivious to the underlying network structure, and a simple *graph-aware* algorithm that we propose. We give precise upper bounds on the expected number of tests performed by each algorithm. Crucially, we show that when the network and infection parameters yield strong community structure (in which case the disease is more likely to be transmitted within a community than between communities), the graph-aware algorithm's average complexity is asymptotically strictly better than that of binary splitting. We corroborate these results with numerical simulations. Finally, we derive novel information-theoretic lower bounds which apply universally to all adaptive strategies, and which asymptotically match the graph-aware algorithm's performance (up to constants) in certain parameter regimes. To the best of our knowledge, this is the first thorough characterization of the complexity of adaptive group testing in a networked setting.

We note that the underlying principles of this paper may be relevant to numerous settings beyond healthcare. In the past, group testing has been successfully applied to diverse domains including wireless communications [16], [19], [30]–[35], machine learning [36]–[38], signal processing [39], [40], and the analysis of data streams [41], [42]. In these settings and others, there may be a natural "clustering" of the population into different subgroups which can inform the design of better group testing strategies, i.e., be exploited as "side information." For example, devices which are closer together in a multiple access network may tend to be active or inactive at the same time. Exploring the potential applications of network-oriented group testing to these types of problems is of great interest.

a) Related Works: In *graph-constrained* group testing [43]–[46], the tests must conform to a given network topology. For example, if the objective is to identify faulty links in a communication network by sending diagnostic

[1]In a related, commonly studied probabilistic model, it is assumed that a random set of d individuals out of n are infected according to some probability distribution (typically the uniform distribution) over all such $\binom{n}{d}$ possibilities [22]–[28]. While this model does not quite assume that infections are independent, it is still somewhat simplistic and fails to capture any dependencies that may exist between individuals.

packets, then each test must correspond to a valid path in the network. By contrast, we allow the tests to be arbitrary, but ask whether *knowledge* of the interaction network can help to reduce the required number of tests.

A few prior works have assumed that infections occur independently and with non-identical prior probabilities [15], [20], [21]. However, our paper pertains to the fully non-i.i.d. case in which infections can be correlated with potentially different priors, depending on the network structure. The idea of *community-aware* group testing was first explored in [11]. This work assumes that the population is partitioned into disjoint "families," and that the disease spreads in the following two stages: 1) each family is infected i.i.d. with some fixed probability; 2) each member of an infected family is infected independently. Our work considers a different disease spread model which operates upon general graphs and is inspired by how diseases often spread in reality (via interactions). Finally, we would like to acknowledge a number of independent and concurrent works related to community-aware group testing [14], [47]–[50].

b) Notation: Let $[n] \triangleq \{1, 2, \dots, n\}$. We denote by n, k, and $m \triangleq \frac{n}{k}$ the size of the population, size of each community, and number of communities, respectively. $X \triangleq (X_1, \dots, X_n) \in \{0,1\}^n$ is the infection status vector, where $X_v = 1$ iff vertex v is infected; $X^\ell \triangleq (X_1, \dots, X_\ell)$, $\ell \in [n]$; $X_{\mathcal{C}_i} \in \{0,1\}$, $i \in [m]$, is the infection status of community $\mathcal{C}_i$, where $X_{\mathcal{C}_i} = 1$ iff $\exists v \in \mathcal{C}_i : X_v = 1$. The indicator function for event $\mathcal{A}$ is $\mathbb{1}_{\mathcal{A}}$. The entropy of a discrete random variable and the binary entropy function (both in bits) are $H(\cdot)$ and $\mathsf{h_b}(\cdot)$, respectively. We write $f(x) \prec g(x)$ to denote $f(x) = o(g(x))$, and $f(x) \preceq g(x)$ to denote $f(x) = O(g(x))$. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph with vertex set $\mathcal{V}$, edge set $\mathcal{E}$, and no self-loops. Let $\mathcal{N}(v) \triangleq \{u \in \mathcal{V} : (u, v) \in \mathcal{E}\}$ denote the neighbors of vertex v, and let $d(v) \triangleq |\mathcal{N}(v)|$ denote the degree of v.

c) Paper Organization: The rest of this paper is organized as follows. In Section II, we describe the infection and network models. In Section III, we provide background and preliminary ideas. In Section IV, we discuss the main algorithms studied in this paper: binary splitting and our proposed graph-aware algorithm. Section V gives upper and lower bounds for adaptive group testing on networks consisting of disjoint cliques, and Section VI generalizes these results to the stochastic block model. Finally, we present the results of our numerical experiments in Section VII, and conclude in Section VIII. All omitted proofs are given in the Appendix.

II. Infection and Network Models

A. Infection Model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph without self-loops, where the vertices represent people in the population and the edges represent interactions between them. We study the following probabilistic infection model with parameters $p, q \in [0, 1]$ which acts upon $\mathcal{G}$ in two stages (each executed once):

1) **Seed Selection:** Each vertex in $\mathcal{V}$ is infected i.i.d. with probability p. These initial infected vertices $\mathcal{V}_s \subseteq \mathcal{V}$ are called the *seeds*. They model the introduction of the disease into the population via some external entity (e.g., a traveler carrying the disease into a country).

2) **Neighbor Infection:** Every seed $v \in \mathcal{V}_s$ infects each of its neighbors $\mathcal{N}(v) = \{u \in \mathcal{V} : (u, v) \in \mathcal{E}\}$ i.i.d. with probability q. This models how the disease spreads through the population via interactions between carriers and nearby individuals.

Remark 1. *The above stages can be viewed as the initial spread of an epidemic. They also form the "first time step" of the independent cascade model [51] from the study of influence maximization in social networks. Our use of this model is motivated by diseases such as COVID-19, which are initially introduced into a population from an external source and subsequently transmitted between individuals in close contact. In practice, the specific values of p, q can be tailored to the disease in question (for example, by using contact tracing to estimate the infectiousness).*

Under a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with infection parameters p, q, we find that our setting reduces to the i.i.d. probabilistic group testing model with prior p when (i) $q = 0$, or (ii) $\mathcal{G}$ is the empty graph where $\mathcal{E} = \emptyset$. It follows that we cannot hope to do any better than classical group testing schemes in these settings.

B. Network Model

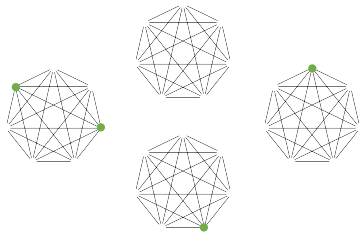
For the rest of this paper, we assume that the underlying network is drawn from the *stochastic block model* (SBM) [52] – a well-known random graph model with the tendency to produce community-structured graphs. The standard SBM has the following parameters:

- n vertices
- a partition of the vertex set $\mathcal{V} = \{1, 2, \ldots, n\}$ into m communities, $\mathcal{C}_1, \ldots, \mathcal{C}_m$, where $\bigcup_{i \in [m]} \mathcal{C}_i = \mathcal{V}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$, $\forall i \neq j$
- a symmetric matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ of edge probabilities.

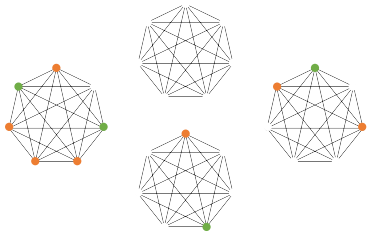
The random graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is then generated by first initializing $\mathcal{E} = \emptyset$, then adding an edge between each pair of vertices $u \in \mathcal{C}_i$, $v \in \mathcal{C}_j$, $u \neq v$, with probability $\mathbf{P}_{ij}$.

In this paper, we assume the communities are all of size k, where k is a factor of n (so that the number of communities is $m = n/k$), and that there is a constant edge probability p_1 within communities, and probability p_2 between communities, where $p_1 > p_2$. That is, $\mathbf{P}$ equals p_1 along the diagonal entries and p_2 on the off-diagonal entries. Finally, we assume that the communities are known to the group testing algorithms in advance, but that the graph itself may not be known. This is a natural assumption, as it may be easier to know which communities people belong to (e.g., families, schools, or workplaces) than whether specific individuals have interacted with each other.

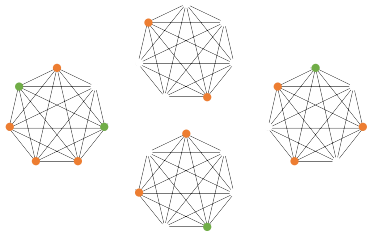
*a) Stochastic Block Infection Model (*SBIM*):* Our infection model acting upon the SBM can be equivalently stated in terms of a stochastic community infection model. Assume we are given a partition of the vertex set $\mathcal{V} = \{1, \ldots, n\}$ into m communities $\mathcal{C}_1, \ldots, \mathcal{C}_m$. Our modified model still begins by selecting each node i.i.d. with probability p to be a seed. However, in the neighbor infection phase, each seed infects its neighbors *within the same community* i.i.d. with probability q_1 and infects those *outside its community* i.i.d. with probability q_2, where $q_1 > q_2$. The equivalence of this model and the original model can be seen by setting $q_1 = p_1 \cdot q$ and $q_2 = p_2 \cdot q$,



(a) Seed selection stage



(b) Neighbor infection with $q_2 = 0$ (the disjoint k-cliques model). Nodes cannot be infected by seeds outside their own community.



(c) Neighbor infection with $q_2 > 0$. Any node can be infected by any seed, even those in external communities.

Fig. 1: Illustration of $\mathsf{SBIM}(n, k, p, q_1, q_2)$. In this example, there are $m = 4$ communities of size $k = 7$. Seeds are colored green, and nodes infected by seeds during the neighbor infection stage are colored orange.

where q is the neighbor infection probability in the original model. We call this the *Stochastic Block Infection Model*, denoted by $\mathsf{SBIM}(n, k, p, q_1, q_2)$. Note that $\mathsf{SBIM}(n, k, p, 0, 0)$, with k an arbitrary factor of n, is equivalent to the i.i.d. group testing model.

b) Disjoint k-Cliques Model: Before analyzing the SBIM in full generality in Section VI, we begin in Section V by investigating the special case of $\mathsf{SBIM}(n, k, p, q, 0)$, which we refer to as the *disjoint k-cliques model*. Here, we have $m = n/k$ communities of size k, with seed selection probability p, intra-community transmission rate q, and an inter-community transmission rate of zero. Figure 1 illustrates the SBIM and the difference between the disjoint k-cliques model ($q_2 = 0$) and the general SBIM with $q_2 > 0$.

III. Background and Preliminaries

A. The Group Testing Problem

In the group testing problem, a *test* corresponds to a subset of individuals $\mathcal{S} \subseteq [n]$. The test outcome is *positive* if $X_i = 1$ for some $i \in \mathcal{S}$; that is, if at least one member of $\mathcal{S}$ is infected. Otherwise, the test outcome is negative. Equivalently, the outcome is a binary variable $Y \in \{0, 1\}$ given by a Boolean OR operation over $\mathcal{S}$:

$$Y = \bigvee_{i \in \mathcal{S}} X_i. \tag{1}$$

A group testing algorithm or scheme describes how to select subsets $\mathcal{S}_1, \dots, \mathcal{S}_T$ such that the infection statuses $X_1, \dots, X_n$ can be determined from the corresponding outcomes $Y_1, \dots, Y_T$. In *adaptive* schemes, the choice of each $\mathcal{S}_t$ is allowed to depend on $\{\mathcal{S}_{t'} : t' < t\}$. Moreover, due to the underlying randomness in the X_i in our probabilistic setting, the total number of tests T performed by any adaptive scheme is a random variable. In this work, we assume that test outcomes are *noiseless* (meaning the algorithms get to observe Y_t as given in (1)), and we require a scheme to *exactly* recover $X_1, \dots, X_n$ (i.e., achieve zero error). Our goal is to characterize the complexity of adaptive schemes under the SBIM by providing both upper and lower bounds on $\mathbb{E}[T]$.

B. Marginal Infection Probability for General Graphs

Under a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the marginal infection probability of a given vertex v can be characterized in terms of its degree $d(v)$ as follows.

Lemma 1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a finite, undirected graph without self-loops. Under $\mathcal{G}$, the infection status of a vertex $v \in \mathcal{V}$ is $X_v \sim \mathsf{Bernoulli}(r_v)$, where*

$$r_v \triangleq \mathbb{P}(X_v = 1) = 1 - (1-p)(1-pq)^{d(v)}. \tag{2}$$

Under a general graph, different nodes may have different degrees and hence different marginal probabilities of infection. From (2), we see that r_v is monotonically non-decreasing with $d(v)$. Note also that the X_v can be correlated.

C. Information-Theoretic Lower Bound

A fundamental result in probabilistic group testing is that *any* adaptive algorithm which is guaranteed to identify all infected members of the population, assuming noiseless test results, requires a number of tests T satisfying

$$\mathbb{E}[T] \geq H(X_1, \dots, X_n). \tag{3}$$

This bound highlights the intimate connection between adaptive group testing and source coding. Indeed, to summarize a discussion from [16], the outcomes of the adaptive tests can be viewed as a binary, variable-length source code for X. The lower bound then follows directly from existing results in data compression (e.g., [53, Eqn. 5.38]). Equation (3) will serve as the point of departure for the lower bounds on $\mathbb{E}[T]$ that we derive in this paper. The key challenge will be to obtain good approximations to $H(X)$ in the presence of correlations induced by the underlying network.

IV. Algorithms

A. Binary Splitting Algorithm

Most adaptive group testing algorithms are based on the idea of recursively splitting the population until all infected members are found. The most standard one is *binary splitting*, which finds a single infected member at a time by repeatedly halving the population. This algorithm identifies all infected members using $\alpha \log_2 n + O(\alpha)$

adaptive tests (see [54], [29, p.24], or [55, Theorem 1.2]), where α is the number of infected members. It works even when α is unknown, and is most effective in the sparse regime, $\alpha = \Theta(n^\beta)$, where $\beta \in [0, 1)$. We treat binary splitting as our baseline in this paper, and we will utilize the following performance guarantee[2]:

Lemma 2. *In a population of size n with α infected members, where $\alpha \geq 1$, the binary splitting algorithm is guaranteed to identify all infected members using at most $\alpha\lceil \log_2 n \rceil \leq \alpha \log_2 n + \alpha$ tests.*

B. Graph-Aware Algorithm

As an alternative to standard adaptive procedures such as binary splitting, we consider a simple adaptive scheme which leverages the community structure of the graph. Our scheme first treats the communities as "meta-individuals" and applies binary splitting to quickly identify the infected communities (i.e., those with at least one infected member). Subsequently, we run binary splitting again – this time within each infected community – to identify the infected individuals. Note that this procedure will recover the infection statuses of all members of the population with zero error, which follows from the fact that binary splitting achieves exact recovery.

Adaptive Graph-Aware Algorithm

1) Mix the samples within each community.
2) Perform binary splitting on the mixed samples to determine which communities contain at least one infected member.
3) For each positive test from Step 2, perform binary splitting within the corresponding community to identify infected members.

Under what circumstances should we expect the graph-aware algorithm to outperform binary splitting? Suppose the underlying interaction network and infection model follow $\mathsf{SBIM}(n, k, p, q_1, q_2)$. If the seed selection probability p is small, then we expect only a few of the $m = n/k$ communities to contain a seed. This means that after the neighbor infection stage, several of the communities are likely to contain no infected members at all, especially if q_2 is small. In Step 2 of the graph-aware algorithm, we can efficiently rule out these uninfected communities from consideration. In Step 3, we need only perform group testing within each of the remaining communities (which contain at least one infected member). In contrast, the binary splitting algorithm ignores the community structure (specifically, the fact that entire communities are likely to be uninfected), and is therefore unlikely to enjoy the same benefits as the graph-aware algorithm under these circumstances. We will rigorously verify this intuition in the upcoming sections.

V. Disjoint k-Cliques Model

We first consider the disjoint k-cliques model with communities $\mathcal{C}_1, \ldots, \mathcal{C}_m$. The seed selection probability is $p \in (0, 1]$, the transmission rate within a community is $q \in [0, 1]$, and no transmissions are possible between

[2]It is well-known that Lemma 2 can be improved via Hwang's generalized binary splitting approach [56]. However, this method requires the number of infected individuals to be known *a priori*. As we do not make such assumptions, we restrict our attention to the standard binary splitting algorithm.

communities. This setting resembles the disjoint families model from [11]. However, in their model, each member of an "infected family" is infected independently with a fixed probability, whereas the infection rate within a given community in our model depends on the number of seeds in the community, which in turn depends on the size of the community. This models the realistic scenario where a larger community has a larger probability of being "infected," i.e., having some infected members. In addition, the state of a given member of an infected community is not independent of the states of the other members; an individual has a higher probability of being infected if there are more infected members in their community. This property of our model makes the derivation of lower bounds and the analysis of group testing schemes more intricate.

A. Information-Theoretic Lower Bound

Recall from (3) that $\mathbb{E}[T] \geq H(X)$ for any adaptive group testing algorithm which exactly identifies the infected individuals using T tests. Since the infection statuses across the m disjoint cliques are independent, we have $\mathbb{E}[T] \geq m \cdot H(X_1, \ldots, X_k)$, where without loss of generality we assume $\mathcal{C}_1 = [k]$. Thus, obtaining a lower bound on $\mathbb{E}[T]$ reduces to lower bounding $H(X_1, \ldots, X_k)$, i.e., the entropy corresponding to a single k-clique. The following lemma lower bounds $H(X_1, \ldots, X_k)$ in terms of a binomial random variable, which ultimately leads to the asymptotic lower bound given in Theorem 1.

Lemma 3. *Under the disjoint k-cliques model, the number of tests T required to identify the infected individuals is lower bounded as*

$$\mathbb{E}[T] \geq m \cdot \mathbb{E}_Z \left[(k - Z) \cdot \mathsf{h_b} \left(1 - (1-q)^Z \right) \right],$$

where $Z \sim \mathsf{Binom}\,(k, p)$.

Next is a technical lemma which characterizes the asymptotic behavior of Lemma 3 by leveraging the concentration of Z around its mean.

Lemma 4. *Let $Z \sim \mathsf{Binom}\,(k, p)$ and assume $kp \preceq 1$ and $q \preceq \frac{1}{\sqrt{k} \cdot \sqrt{\log\left(\frac{1}{k \cdot p}\right)}}$. Then*

$$\mathbb{E}_Z \left[(k - Z) \cdot \mathsf{h_b} \left(1 - (1-q)^Z \right) \right] \succeq k^2 \cdot p \cdot q \cdot \left(\log k + \log \log \left(\frac{1}{k \cdot p} \right) \right).$$

Remark 2. *The upper bounds on p and q in Lemma 4 are somewhat necessary, as they ensure that the infected population is sparse enough for group testing to improve upon individual testing. However, the specific upper bound on q is an artifact of our lower bounding technique and could potentially be loosened.*

Upon combining Lemma 3 and Lemma 4, we see that the number of tests T needed to recover all infected members in the disjoint k-cliques graph (in the specified parameter regime) is lower bounded as

$$\mathbb{E}[T] \succeq m \cdot k^2 \cdot p \cdot q \cdot \left(\log k + \log \log \left(\frac{1}{kp} \right) \right).$$

Note that another lower bound is given by

$$\mathbb{E}[T] \geq H(X_1, \ldots, X_n) \overset{(a)}{\geq} H(X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_m}) = m \cdot \mathsf{h_b} \Big(1 - (1-p)^k \Big) \tag{4}$$

where (a) uses the fact that $X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_m}$ are a function of $X_1, \ldots, X_n$. Furthermore, since $kp \preceq 1$, we have $\mathsf{h_b}\Big(1-(1-p)^k\Big) \succeq k \cdot p \cdot \log_2(1/kp)$. We summarize the refined lower bound in the following theorem:

Theorem 1. *Assume $kp \preceq 1$ and $q \preceq \frac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$. Then under the disjoint k-cliques model, the expected number of tests required to identify the infected individuals is lower bounded as*

$$\mathbb{E}[T] \succeq \max\left\{ m \cdot k^2 \cdot p \cdot q \cdot \left(\log k + \log\log\left(\frac{1}{k \cdot p}\right)\right), \quad m \cdot k \cdot p \cdot \log\left(\frac{1}{k \cdot p}\right), \quad 1 \right\}.$$

Recall that $q = 0$ corresponds to i.i.d. group testing, in which case (3) gives the lower bound $\mathbb{E}[T] \geq n \cdot \mathsf{h_b}(p) \geq np\log(1/p)$. On the other hand, substituting $q = 0$ into Theorem 1 yields $np\log(1/kp)$, which differs from the i.i.d. case by an additive factor of $np\log(1/k)$. In this special case, our bound can be seen as slightly suboptimal. However, observe that when $q = 0$, the disjoint k-cliques models are equivalent for *all* values of k. This is because the community structure plays no role in the i.i.d. setting. Therefore, Theorem 1 holds for any value of k when $q = 0$, and can thus be maximized over k to obtain the best-possible bound. The maximum occurs at $k = 1$ (i.e., when every vertex is its own community), which recovers the i.i.d. lower bound of $np\log(1/p)$ as desired.

B. Algorithm Analysis

1) Binary Splitting: The following result bounds the expected number of tests used by the binary splitting algorithm under the disjoint k-cliques model.

Theorem 2. *Under the disjoint k-cliques model, the binary splitting algorithm identifies all infected individuals using T tests, where*

$$\mathbb{E}[T] \leq m \cdot k \cdot \Big(\log_2 m + \log_2 k + 1\Big) \cdot \Big(1 - (1-p)(1-pq)^{k-1}\Big).$$

Proof. Let K be the number of infected nodes (which is a random variable in our setting). Then

$$\mathbb{E}[K] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = n \cdot r$$

where $r = 1 - (1-p)(1-pq)^{k-1}$ by Lemma 1. Invoking Lemma 2 yields the result. □

a) Asymptotic Analysis: Using Theorem 2, we find that the average complexity of binary splitting is $O\left(m \cdot k^2 \cdot p \cdot (\log_2 m + \log_2 k) \cdot (1/k + q)\right)$ since

$$\begin{aligned}
\mathbb{E}[T] &\preceq m \cdot k \cdot (\log m + \log k) \cdot \Big(1 - (1-p)(1-pq)^{k-1}\Big) \\
&\overset{(a)}{\leq} m \cdot k \cdot (\log m + \log k) \cdot \Big(1 - (1-p)(1-kpq)\Big) \\
&= m \cdot k \cdot (\log m + \log k) \cdot (p + kpq - kp^2q) \\
&\leq m \cdot k \cdot (\log m + \log k) \cdot (p + kpq) \\
&= m \cdot k^2 \cdot p \cdot (\log m + \log k) \cdot \Big(\frac{1}{k} + q\Big) \qquad (5)
\end{aligned}$$

where in (a) we use the fact that $(1+x)^k \geq 1 + kx$ for $x \geq -1$, $k \geq 1$.

2) Graph-Aware Algorithm: Next, we provide an upper bound on the expected number of tests performed by the graph-aware algorithm. The two terms in the sum below correspond, respectively, to the expected number of tests in Steps 2 and 3 of the algorithm.

Theorem 3. *Under the disjoint k-cliques model, the graph-aware algorithm identifies all infected individuals using T tests, where*

$$\mathbb{E}[T] \leq m \cdot \Big(\log_2 m + 1\Big) \cdot \Big(1 - (1-p)^k\Big) + n \cdot \Big(\log_2 k + 1\Big) \cdot \Big(1 - (1-p)(1-pq)^{k-1}\Big)$$

a) Asymptotic Analysis: Using Theorem 3 and the fact that $(1+x)^k \geq 1 + kx$ for $x \geq -1$, $k \geq 1$, we find that the average complexity of the graph-aware algorithm is given by

$$\mathbb{E}[T] \preceq m \cdot k \cdot p \cdot \log m + m \cdot k^2 \cdot p \cdot \Big(\frac{1}{k} + q\Big) \cdot \log k. \tag{6}$$

C. Discussion

We summarize the expected number of tests of binary splitting and the graph-aware algorithm, as well as the information-theoretic lower bound, in Table I.

Binary splitting	$m \cdot k^2 \cdot p \cdot \left(\frac{1}{k} + q\right) \cdot \log m + m \cdot k^2 \cdot p \cdot \left(\frac{1}{k} + q\right) \cdot \log k$
Graph-aware	$m \cdot k \cdot p \cdot \log m + m \cdot k^2 \cdot p \cdot \left(\frac{1}{k} + q\right) \cdot \log k$
Lower bound	$m \cdot k \cdot p \cdot \log\left(\frac{1}{kp}\right) + m \cdot k^2 \cdot p \cdot q \cdot \left(\log k + \log\log\left(\frac{1}{kp}\right)\right) + 1$

TABLE I: Upper and lower bounds on the expected number of tests in the disjoint k-cliques model.

If we compare the bounds for binary splitting and the graph-aware algorithm term-by-term, we observe that the binary splitting bound has an extra additive factor of $mk^2pq \log m$. Thus, the graph-aware algorithm is never worse (order-wise) than binary splitting. Furthermore, when $q = 0$ (the i.i.d. setting, where community structure has no bearing on the infection spread), the bounds are order-wise equivalent. As we will see next, the graph-aware algorithm is in fact strictly better than binary splitting in some cases.

Next, we discuss different parameter regimes where 1) the lower bound holds, 2) the graph-aware algorithm is order-optimal (i.e., the lower bound is tight), and 3) the graph-aware algorithm's average complexity is strictly better than binary splitting's. As stated in Corollary 1, the lower bound holds when $kp \preceq 1$ and $q \preceq \frac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$. The next corollary specifies the regime where the graph-aware algorithm is tight:

Corollary 1. *If the following conditions hold:*

1) $kp \preceq m^{-\alpha}$ for some fixed $\alpha \in (0,1)$,

2) $\frac{1}{k} \preceq q \preceq \frac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$,

then the lower bound is tight, and moreover the graph-aware algorithm is order-optimal.

Proof. Plugging $\log\left(\frac{1}{kp}\right) \succeq \alpha \log m$ into the lower bound and using the fact that $k \succeq \log\left(\frac{1}{kp}\right)$ from the second condition (which implies $\log k \succeq \log\log m$) yields

$$\begin{aligned}\mathbb{E}[T] &\succeq m \cdot k \cdot p \cdot \log m + m \cdot k^2 \cdot p \cdot q \cdot (\log k + \log\log m) + 1 \\ &\succeq m \cdot k \cdot p \cdot \log m + m \cdot k^2 \cdot p \cdot q \cdot \log k,\end{aligned}$$

and applying $q \succeq 1/k$ to the bound for the graph-aware algorithm yields

$$\mathbb{E}\left[T\right] \preceq m \cdot k \cdot p \cdot \log m + m \cdot k^2 \cdot p \cdot q \cdot \log k.$$

□

Finally, we specify the regime where the graph-aware algorithm outperforms binary splitting:

Corollary 2. *If* $\log m \succ \log k$ *and* $kq \succ 1$*, then the graph-aware algorithm's average complexity is asymptotically strictly better than binary splitting's by a factor of* $\min\left\{kq, \frac{\log m}{\log k}\right\}$.

Proof. Under the above conditions, binary splitting's average complexity is

$$m \cdot k^2 \cdot p \cdot q \cdot \log m$$

whereas the graph aware algorithm's average complexity is

$$\max\left\{\underbrace{m \cdot k \cdot p \cdot \log m}_{\text{(a)}}, \quad \underbrace{m \cdot k^2 \cdot p \cdot q \cdot \log k}_{\text{(b)}}\right\}.$$

Both (a) and (b) are strictly smaller than the binary splitting bound. We see that (a) saves a factor of $kq \succ 1$, while (b) saves a factor of $\frac{\log m}{\log k} \succ 1$. □

In Table II, we summarize the different parameter regimes discussed so far.

Lower bound's conditions	$kp \preceq 1$ and $q \preceq \frac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$
Tightness conditions	$kp \preceq m^{-\alpha}$ and $1 \preceq kq \preceq \sqrt{k/\log\left(\frac{1}{kp}\right)}$
Improvement conditions	$\log m \succ \log k$ and $kq \succ 1$

TABLE II: Parameter regimes of interest for the disjoint k-cliques model.

The main takeaway is that the graph-aware algorithm can potentially improve upon binary splitting when (i) there are several moderately sized communities in the network, and (ii) the transmission rate within each clique is significant. Additionally, the graph-aware algorithm is order-optimal when the seeds are sparse. However, one can verify that when $q \preceq 1/k$, i.e., the intra-clique transmission rate is small, then the bounds for binary splitting and the graph-aware algorithm are order-wise equivalent. This suggests that knowledge of the community structure may not help in this regime. Intuitively, this makes sense because when q is small, the infection statuses of the vertices are "mostly independent."

VI. Stochastic Block Infection Model

Having studied the disjoint k-cliques model, we now turn to the fully general $\mathsf{SBIM}(n,k,p,q_1,q_2)$, where $p \in (0,1]$ and $q_1, q_2 \in [0,1]$.

A. Information-Theoretic Lower Bound

Similar to Section V-A, we obtain the following lower bounds for adaptive group testing over the SBIM. The proof of the following lemma is similar to that of Lemma 3, except that we must now take into account the transmission of the disease *between* communities. This is reflected in the additional binomial random variable Z' appearing within the binary entropy function.

Lemma 5. *Under $\mathsf{SBIM}(n,k,p,q_1,q_2)$, the number of tests T required to identify the infected individuals is lower bounded as*

$$\mathbb{E}[T] \geq m \cdot \mathbb{E}_{Z,Z'}\left[(k-Z)\cdot \mathsf{h_b}\left(1-(1-q_1)^{Z}(1-q_2)^{Z'}\right)\right],$$

where $Z \sim \mathsf{Binom}(k,p)$ and $Z' \sim \mathsf{Binom}(n-k,p)$ are independent.

Similar to Lemma 4, the following technical lemma leverages the concentration of Z and Z' around their means.

Lemma 6. *Let $Z \sim \mathsf{Binom}(k,p)$ and $Z' \sim \mathsf{Binom}(n-k,p)$ be independent, and assume*

1) $n \cdot p \cdot q_2 \preceq 1$,
2) $n \cdot p \succeq 1$,
3) $k \cdot p \cdot q_1 \preceq 1$,
4) $q_1 \leq \frac{1}{\sqrt{2k\left(\log\left(\frac{1}{kp}\right)+1\right)}}$.

Then the following lower bound holds:

$$\mathbb{E}_{Z,Z'}\left[(k-Z)\cdot \mathsf{h_b}\left(1-(1-q_1)^{Z}(1-q_2)^{Z'}\right)\right] \succeq mk^2pq_2\log\left(\frac{1}{npq_2}\right) + k^2pq_1\log\left(\frac{1}{q_1+npq_2}\right).$$

Theorem 4. *In the parameter regime specified in Lemma 6, the number of tests T needed to recover all infected members over $\mathsf{SBIM}(n,k,p,q_1,q_2)$ is lower bounded as*

$$\mathbb{E}[T] \succeq m^2\cdot k^2\cdot p\cdot q_2\cdot \log\left(\frac{1}{n\cdot p\cdot q_2}\right) + m\cdot k^2\cdot p\cdot q_1\cdot \log\left(\frac{1}{q_1+n\cdot p\cdot q_2}\right).$$

Remark 3. *Recall that in the disjoint k-cliques model, we obtained an additional lower bound in Equation (4) given by $H\left(X_{\mathcal{C}_1},...,X_{\mathcal{C}_m}\right)$, which dominates when $kp \preceq m^{-\alpha}$. However, under the general SBIM, the $\{X_{\mathcal{C}_1},...,X_{\mathcal{C}_m}\}$ are no longer mutually independent, rendering the analysis of $H\left(X_{\mathcal{C}_1},...,X_{\mathcal{C}_m}\right)$ difficult. Therefore, we suspect that the lower bound given in Theorem 4 is not tight when kp is small. Resolving this issue is an open problem.*

B. Algorithm Analysis

To analyze binary splitting and the graph-aware algorithm over the SBIM, we begin by extending Lemma 1.

Lemma 7. *The marginal probability of infection for every vertex v under $\mathsf{SBIM}(n,k,p,q_1,q_2)$ is given by*

$$\mathbb{P}(X_v=1) = 1-(1-p)\cdot(1-p\cdot q_1)^{k-1}\cdot(1-p\cdot q_2)^{n-k}.$$

1) Binary Splitting: Next, we generalize the bound in Theorem 2 to the SBIM. Notice that in both the Theorem 5 bound and the asymptotic bound derived below, we recover the corresponding bounds from the disjoint k-cliques setting when we set $q_1 = q$, $q_2 = 0$.

Theorem 5. *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$*, the binary splitting algorithm identifies all infected individuals using* T *tests, where*

$$\mathbb{E}[T] \leq n \cdot (\log_2 n + 1) \cdot \Big(1 - (1-p) \cdot (1 - p \cdot q_1)^{k-1} \cdot (1 - p \cdot q_2)^{n-k}\Big).$$

Proof. Let K be the number of infected nodes. Then

$$\mathbb{E}[K] = \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) = n \cdot r$$

where $r = 1 - (1-p) \cdot (1 - p \cdot q_1)^{k-1} \cdot (1 - p \cdot q_2)^{n-k}$ by Lemma 7. Invoking Lemma 2 yields the result. □

a) Asymptotic Analysis: Using the fact that $(1+x)^k \geq 1 + kx$ for $x \geq -1$, $k \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[T] &\preceq n \cdot \log n \cdot \Big(1 - (1-p)(1 - k \cdot p \cdot q_1) \cdot (1 - (n-k) \cdot p \cdot q_2)\Big) \\
&\leq n \cdot \log n \cdot \Big((n-k) \cdot p \cdot q_2 + k \cdot p \cdot q_1 + p + k \cdot (n-k) \cdot p^3 \cdot q_1 \cdot q_2\Big) \\
&\leq m \cdot k^2 \cdot p \cdot (\log m + \log k) \cdot \Big(\frac{1}{k} + q_1 + m \cdot q_2 + m \cdot k \cdot p^2 \cdot q_1 \cdot q_2\Big). \qquad (7)
\end{aligned}$$

2) Graph-Aware Algorithm: First, we provide a lemma needed to prove the upper bound for the graph-aware algorithm in Theorem 6. Again, note that by setting $q_1 = q$, $q_2 = 0$ in Theorem 6 and the resulting asymptotic bound, we recover the corresponding bounds from the disjoint k-cliques setting.

Lemma 8. *Let* $X_{\mathcal{C}_1}$ *be the indicator variable which equals 1 if at least one member of community* $\mathcal{C}_1$ *is infected. Then under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$*,*

$$\mathbb{P}(X_{\mathcal{C}_1} = 1) = 1 - (1-p)^k \cdot \left(1 - p \cdot \Big(1 - (1-q_2)^k\Big)\right)^{n-k}.$$

Theorem 6. *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$*, the graph-aware algorithm identifies all infected individuals using* T *tests, where*

$$\begin{aligned}
\mathbb{E}[T] \leq \frac{n}{k} \cdot \Big(\log_2(n/k) + 1\Big) \cdot \left(1 - (1-p)^k \cdot \left(1 - p \cdot \Big(1 - (1-q_2)^k\Big)\right)^{n-k}\right) \\
+ n \cdot \Big(\log_2 k + 1\Big) \cdot \Big(1 - (1-p) \cdot (1 - p \cdot q_1)^{k-1} \cdot (1 - p \cdot q_2)^{n-k}\Big).
\end{aligned}$$

Proof. Same steps as the proof of Theorem 3 (given in the Appendix), except using Lemma 7 and Lemma 8 wherever $\mathbb{P}(X_1 = 1)$ and $\mathbb{P}(X_{\mathcal{C}_1} = 1)$ are needed, respectively. □

a) Asymptotic Analysis: Let T_1 and T_2 be the first and second terms in the Theorem 6 bound, respectively. Using the fact that $(1-q_2)^k \geq 1 - kq_2$, we have

$$1 - p \cdot \big(1 - (1-q_2)^k\big) \geq 1 - p \cdot k \cdot q_2,$$

so

$$\begin{aligned}\mathbb{E}[T_1] &\preceq m\log m\cdot\left(1-(1-p)^k\cdot\left(1-p\left(1-(1-q_2)^k\right)\right)^{n-k}\right)\\ &\preceq m\log m\cdot\left(1-(1-p)^k\cdot(1-p\cdot k\cdot q_2)^{n-k}\right)\\ &\preceq m\log m\cdot(1-(1-k\cdot p)\cdot(1-(n-k)\cdot p\cdot k\cdot q_2))\\ &\preceq m\log m\cdot(k\cdot p+n\cdot p\cdot k\cdot q_2)\,.\end{aligned}$$

Following the previous asymptotic analysis for binary splitting,

$$\mathbb{E}[T_2]\preceq m\cdot k^2\cdot p\cdot\log k\cdot\left(\frac{1}{k}+q_1+m\cdot q_2+m\cdot k\cdot p^2\cdot q_1\cdot q_2\right).$$

Therefore,

$$\mathbb{E}[T]\preceq m\cdot k\cdot p\cdot\log m\cdot\left(1+m\cdot k\cdot q_2\right)+m\cdot k^2\cdot p\cdot\log k\cdot\left(\frac{1}{k}+q_1+m\cdot q_2+m\cdot k\cdot p^2\cdot q_1\cdot q_2\right). \quad (8)$$

C. Discussion

Comparing (7) and (8) term-by-term, we see that the binary splitting bound has an extra additive factor compared to the graph-aware bound, implying that the graph-aware algorithm is again never worse (order-wise) than binary splitting. In certain regimes, it is asymptotically strictly better than binary splitting. One such regime is given by

1) $\log m\succ\log k$
2) $kq_1\succ 1$
3) (i) $1\succeq mkq_2$

   or

   (ii) $mkq_2\succeq 1$ and $mkq_2\prec kq_1\preceq\frac{1}{p^2}$.

Suppose conditions 1, 2, and 3(i) hold. Binary splitting's average complexity (7) becomes

$$m\cdot k^2\cdot p\cdot q_1\cdot\log m$$

whereas the graph-aware algorithm's average complexity (8) becomes

$$\max\left\{m\cdot k\cdot p\cdot\log m,\quad m\cdot k^2\cdot p\cdot q_1\cdot\log k\right\}.$$

The first term in the graph-aware bound improves upon binary splitting's complexity by a factor of $kq_1\succ 1$, and the second term improves by a factor of $\frac{\log m}{\log k}\succ 1$. Thus, the overall improvement is a factor of $\min\left\{kq,\ \frac{\log m}{\log k}\right\}$, which matches Corollary 2. This is not very surprising because the SBIM asymptotically behaves like the disjoint k-cliques model under condition 3(i), i.e., when q_2 is very small.

However, the graph-aware algorithm still improves over binary splitting in a more intermediate regime for q_2. Under condition 3(ii), binary splitting's average complexity is the same as above, and the graph-aware algorithm's complexity becomes

$$\max\left\{m^2\cdot k^2\cdot p\cdot q_2\cdot\log m,\quad m\cdot k^2\cdot p\cdot q_1\cdot\log k\right\},$$

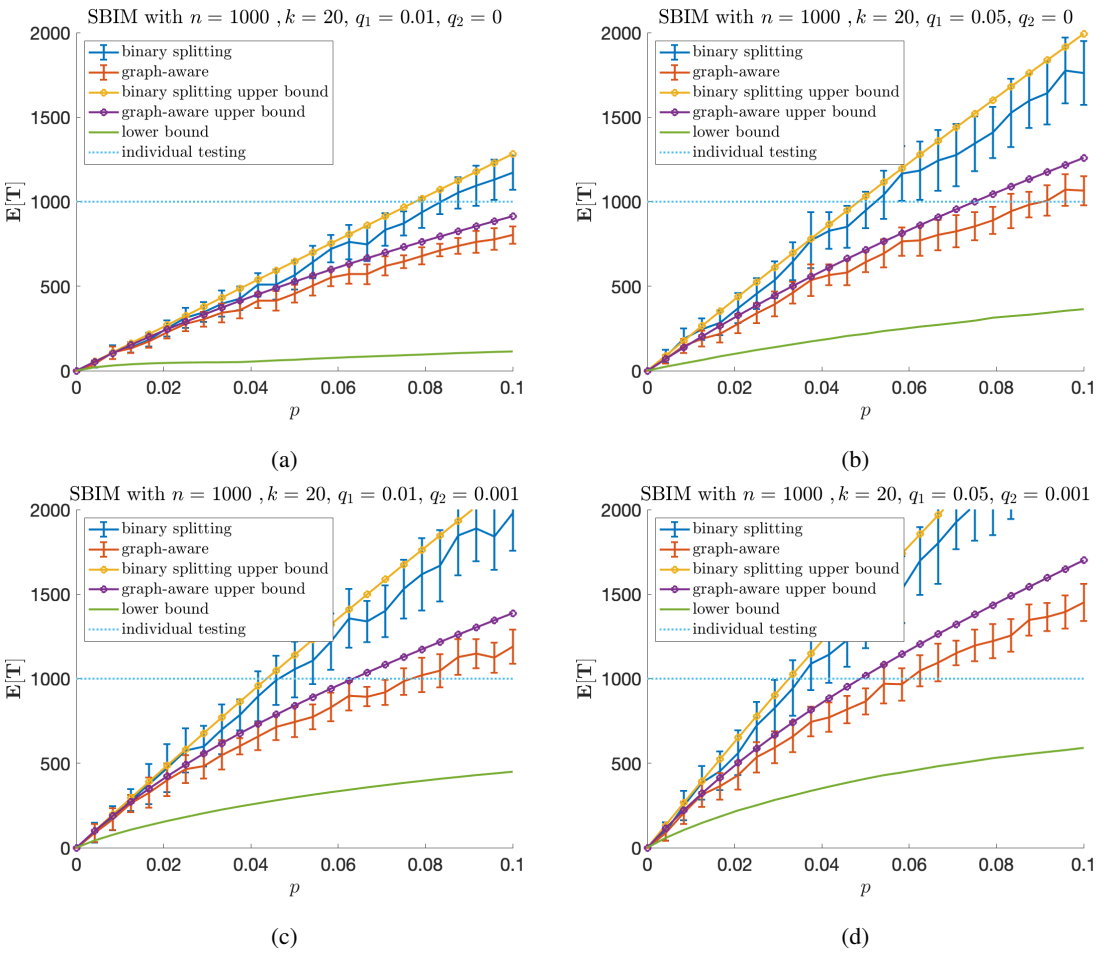


Fig. 2: Performance comparison between binary splitting and the graph-aware algorithm under the SBIM with $n = 1000$, $k = 20$, and different values of p, q_1, q_2. Theoretical upper and lower bounds are also shown.

which represents an improvement over binary splitting by a factor of $\min\left\{\frac{q_1}{m \cdot q_2}, \frac{\log m}{\log k}\right\} \succ 1$.

VII. Numerical Experiments

We implemented the binary splitting and graph-aware algorithms and evaluated their performance over random instances of the SBIM. The population size was set to $n = 1000$, and p was varied over the interval $[0,\, 0.1]$. We ran 20 trials for each value of p, where a trial consists of generating an instance from $\mathsf{SBIM}(n, k, p, q_1, q_2)$, then observing the number of tests used by binary splitting and the graph-aware algorithm to identify the infected nodes. We estimated the lower bound from Lemma 5 by averaging over many independent samples of $Z \sim \mathsf{Binom}(k, p)$ and $Z' \sim \mathsf{Binom}(n - k,\, p)$.

Figure 2 shows some representative plots of the estimated $\mathbb{E}[T]$ as a function of p, with $k = 20$ and different values of $q_1,\, q_2$. The error bars show $\pm$ one standard deviation of the values of T obtained for a particular value of p. For comparison, we also plot the theoretical upper bounds from Theorem 5 and Theorem 6; we find that these

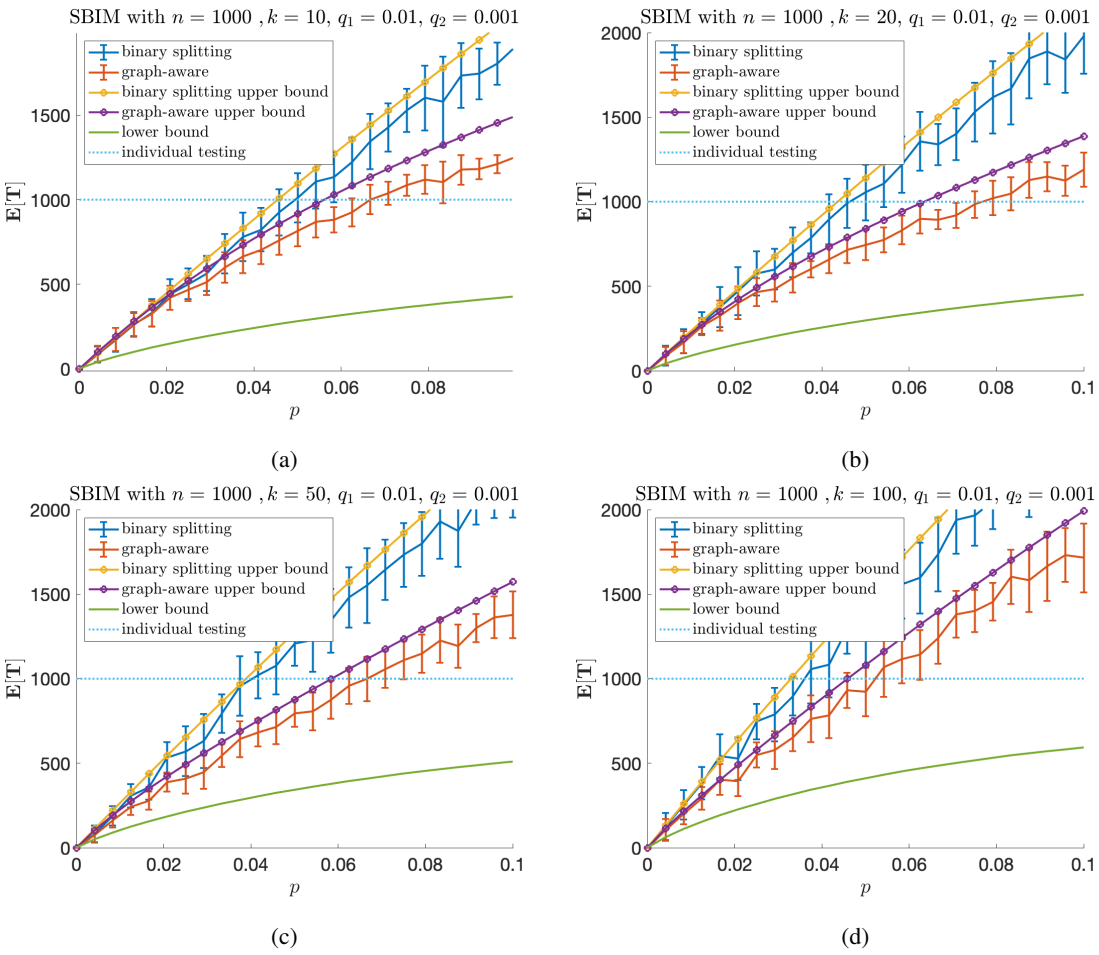


Fig. 3: Performance comparison between binary splitting and the graph-aware algorithm under the SBIM with $n = 1000$, $q_1 = 0.01$, $q_2 = 0.001$, and different values of p, k. Theoretical upper and lower bounds are also shown.

bounds closely match the empirical results. Additionally, the graph-aware algorithm consistently outperforms binary splitting. For example, in Figure 2b, at $p \approx 0.07$, binary splitting has surpassed the individual testing threshold with an average of 1271.5 tests, whereas the graph-aware algorithm uses an average of 813.8 tests; this represents a 36% reduction in testing. The graph-aware algorithm also seems to enjoy lower variance than binary splitting.

In Figure 3, we fix $q_1 = 0.01$, $q_2 = 0.001$, and vary the community size $k \in \{10, 20, 50, 100\}$. The graph-aware algorithm appears to perform most favorably for moderate values of k, such as $k = 20$ or $k = 50$, i.e., when there are several moderately sized communities in the network. These findings are consistent with our earlier theoretical results.

Although the graph-aware algorithm improves significantly upon binary splitting, there is still a sizable gap between the graph-aware bound and the lower bound shown in the plots. This suggests that in the non-asymptotic regime, either the lower bound is not tight or better algorithms exist.

VIII. Conclusion

In this paper, we investigated the group testing problem over networks with community structure. Motivated by diseases such as COVID-19, we proposed a network infection model to capture how certain diseases are introduced into a population and subsequently transmitted through close contact between individuals. Our proposed group testing algorithm, which exploits the structure of the underlying graph, provably outperforms the network-oblivious binary splitting algorithm, and is even order-optimal in certain parameter regimes.

We conclude with some practical considerations and future directions. First, we note that the community-structured networks studied in this paper can model populations at different scales: the "communities" can be schools, families, counties, etc. The insights from our work can also be extended to more general networks in the real world, where the communities may not be known in advance. In such instances, one might use the following pipeline to efficiently identify infected individuals in the population: 1) estimate the network from data (e.g., mobile phone data, Facebook social graph); 2) run a graph clustering algorithm to identify communities in the network; 3) perform graph-aware group testing using the previously identified communities. An interesting direction for future work is to explore the efficacy of such an approach. Other directions of interest include designing non-adaptive group testing schemes for our setting, studying the effect of noisy test outcomes, and extending our infection model to longer time horizons (e.g., SIR or SIS-type infection models from the epidemiology literature).

Acknowledgement

This work was supported in part by NSF Grant #1817205, the Center for Science of Information (CSoI), an NSF Science and Technology Center under grant agreement CCF-0939370, a Cisco Systems Stanford Graduate Fellowship, and a National Semiconductor Corporation Stanford Graduate Fellowship.

References

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

[2] J. Ellenberg, "Five people. One test. This is how you get there." https://www.nytimes.com/2020/05/07/opinion/coronavirus-group-testing.html. Accessed: July 31, 2020.

[3] S. Mallapaty, "The mathematical strategy that could transform coronavirus testing." https://www.nature.com/articles/d41586-020-02053-6. Accessed: July 31, 2020.

[4] Centers for Disease Control and Prevention, "Interim guidance for use of pooling procedures in SARS-CoV-2 diagnostic, screening, and surveillance testing." https://www.cdc.gov/coronavirus/2019-ncov/lab/pooling-procedures.html. Accessed: December 21, 2020.

[5] C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, "Sample pooling as a strategy to detect community transmission of SARS-CoV-2," *JAMA*, vol. 323, no. 19, pp. 1967–1969, 2020.

[6] L. N. Theagarajan, "Group testing for COVID-19: How to stop worrying and test more," *arXiv preprint arXiv:2004.06306*, 2020.

[7] J. N. Eberhardt, N. P. Breuckmann, and C. S. Eberhardt, "Multi-stage group testing improves efficiency of large-scale COVID-19 screening," *Journal of Clinical Virology*, p. 104382, 2020.

[8] S. Ghosh, R. Agarwal, M. A. Rehan, S. Pathak, P. Agrawal, Y. Gupta, S. Consul, N. Gupta, R. Goyal, A. Rajwade, *et al.*, "A compressed sensing approach to group-testing for COVID-19 detection," *arXiv preprint arXiv:2005.07895*, 2020.

[9] A. Cohen, N. Shlezinger, A. Solomon, Y. C. Eldar, and M. Médard, "Multi-level group testing with application to one-shot pooled COVID-19 tests," *arXiv preprint arXiv:2010.06072*, 2020.

[10] L. Abraham, G. Becigneul, B. Coleman, B. Scholkopf, A. Shrivastava, and A. Smola, "Bloom origami assays: Practical group testing," *arXiv preprint arXiv:2008.02641*, 2020.

[11] P. Nikolopoulos, T. Guo, C. Fragouli, and S. Diggavi, "Community aware group testing," *arXiv preprint arXiv:2007.08111*, 2020.

[12] R. Gabrys, S. Pattabiraman, V. Rana, J. Ribeiro, M. Cheraghchi, V. Guruswami, and O. Milenkovic, "AC-DC: Amplification curve diagnostics for Covid-19 group testing," *arXiv preprint arXiv:2011.05223*, 2020.

[13] J. Zhu, K. Rivera, and D. Baron, "Noisy pooled PCR for virus testing," *arXiv preprint arXiv:2004.02689*, 2020.

[14] R. Goenka, S.-J. Cao, C.-W. Wong, A. Rajwade, and D. Baron, "Contact tracing enhances the efficiency of COVID-19 group testing," *arXiv preprint arXiv:2011.14186*, 2020.

[15] F. Hwang, "A generalized binomial group testing problem," *Journal of the American Statistical Association*, vol. 70, no. 352, pp. 923–926, 1975.

[16] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 185–191, 1985.

[17] M. Sobel and P. A. Groll, "Group testing to eliminate efficiently all defectives in a binomial sample," *Bell System Technical Journal*, vol. 38, no. 5, pp. 1179–1252, 1959.

[18] T. Berger and V. I. Levenshtein, "Asymptotic efficiency of two-stage disjunctive testing," *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 1741–1749, 2002.

[19] J. Luo and D. Guo, "Neighbor discovery in wireless ad hoc networks based on group testing," in *Allerton Conference on Communication, Control, and Computing*, pp. 791–797, 2008.

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

[21] T. Kealy, O. Johnson, and R. Piechocki, "The capacity of non-identical adaptive group testing," in *Allerton Conference on Communication, Control, and Computing*, pp. 101–108, 2014.

[22] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3671–3687, 2014.

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

[24] A. Mazumdar, "Nonadaptive group testing with random set of defectives," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7522–7531, 2016.

[25] K. Lee, K. Chandrasekher, R. Pedarsani, and K. Ramchandran, "Saffron: A fast, efficient, and robust framework for group testing based on sparse-graph codes," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4649–4664, 2019.

[26] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi, "Efficient algorithms for noisy group testing," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2113–2136, 2017.

[27] H. A. Inan, P. Kairouz, M. Wootters, and A. Özgür, "On the optimality of the Kautz-Singleton construction in probabilistic group testing," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5592–5603, 2019.

[28] H. A. Inan and A. Ozgur, "Strongly explicit and efficiently decodable probabilistic group testing," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 525–530, IEEE, 2020.

[29] D. Du, F. K. Hwang, and F. Hwang, *Combinatorial Group Testing and Its Applications*, vol. 12. World Scientific, 2000.

[30] 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.

[31] H. A. Inan, P. Kairouz, and A. Ozgur, "Sparse group testing codes for low-energy massive random access," in *Allerton Conference on Communication, Control, and Computing*, pp. 658–665, 2017.

[32] H. A. Inan, P. Kairouz, and A. Ozgur, "Energy-limited massive random access via noisy group testing," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1101–1105, 2018.

[33] H. A. Inan, S. Ahn, P. Kairouz, and A. Ozgur, "A group testing approach to random access for short-packet communication," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 96–100, 2019.

[34] H. A. Inan, P. Kairouz, and A. Özgür, "Sparse combinatorial group testing," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2729–2742, 2019.

[35] A. Cohen, A. Cohen, and O. Gurewitz, "Efficient data collection over multiple access wireless sensors network," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 491–504, 2020.

[36] S. Ubaru and A. Mazumdar, "Multilabel classification with group testing and codes," in *International Conference on Machine Learning*, pp. 3492–3501, 2017.

[37] Y. Zhou, U. Porwal, C. Zhang, H. Q. Ngo, X. Nguyen, C. Ré, and V. Govindaraju, "Parallel feature selection inspired by group testing," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3554–3562, 2014.

[38] D. Malioutov and K. Varshney, "Exact rule learning via boolean compressed sensing," in *International Conference on Machine Learning*, pp. 765–773, 2013.

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

[40] A. Cohen, N. Shlezinger, S. Salamatian, Y. C. Eldar, and M. Médard, "Serial quantization for sparse time sequences," *arXiv preprint arXiv:1907.01691*, 2019.

[41] G. Cormode and S. Muthukrishnan, "What's hot and what's not: tracking most frequent items dynamically," *ACM Transactions on Database Systems (TODS)*, vol. 30, no. 1, pp. 249–278, 2005.

[42] A. Emad and O. Milenkovic, "Poisson group testing: A probabilistic model for nonadaptive streaming boolean compressed sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3335–3339, 2014.

[43] N. J. Harvey, M. Patrascu, Y. Wen, S. Yekhanin, and V. W. Chan, "Non-adaptive fault diagnosis for all-optical networks via combinatorial group testing on graphs," in *IEEE International Conference on Computer Communications (INFOCOM)*, pp. 697–705, 2007.

[44] 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.

[45] A. Karbasi and M. Zadimoghaddam, "Sequential group testing with graph constraints," in *IEEE Information Theory Workshop (ITW)*, pp. 292–296, 2012.

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

[47] P. Nikolopoulos, S. R. Srinivasavaradhan, T. Guo, C. Fragouli, and S. Diggavi, "Group testing for overlapping communities," *arXiv preprint arXiv:2012.02804*, 2020.

[48] P. Bertolotti and A. Jadbabaie, "Network group testing," *arXiv preprint arXiv:2012.02847*, 2020.

[49] B. Arasli and S. Ulukus, "Group testing with a graph infection spread model," *arXiv preprint arXiv:2101.05792*, 2020.

[50] Y.-J. Lin, C.-H. Yu, T.-H. Liu, C.-S. Chang, and W.-T. Chen, "Positively correlated samples save pooled testing costs," *arXiv preprint arXiv:2011.09794*, 2020.

[51] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, 2003.

[52] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[53] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Wiley, 2006.

[54] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 2676–2680, 2013.

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

[56] F. K. Hwang, "A method for detecting all defective members in a population by group testing," *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 605–608, 1972.

APPENDIX

A. Proof of Lemma 1

Let Y_v be the indicator random variable of whether vertex v is a seed. First, we have

$$\begin{aligned}\mathbb{P}(X_v = 1) &= \underbrace{\mathbb{P}\left(X_v = 1 \,|\, Y_v = 1\right)}_{=1} \cdot \underbrace{\mathbb{P}(Y_v = 1)}_{=p} + \mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right) \cdot \mathbb{P}(Y_v = 0) \\ &= p + (1-p) \cdot \mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right).\end{aligned}$$

Given that v is not a seed, $X_v = 1$ if and only if v is infected by one of its neighbors. Hence,

$$\begin{aligned}\mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right) &= \mathbb{P}\left(\{v \text{ is infected by a neighbor}\}\right) \\ &= 1 - \mathbb{P}\left(\{v \text{ isn't infected by any neighbor}\}\right) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \mathbb{P}\left(\{v \text{ isn't infected by } u\}\right) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \left(1 - \mathbb{P}\left(\{v \text{ is infected by } u\}\right)\right) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \left(1 - \mathbb{P}\left(\{v \text{ is infected by } u\} \,|\, Y_u = 1\right) \cdot \mathbb{P}(Y_u = 1)\right) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} (1 - pq) \\ &= 1 - (1-pq)^{d(v)}.\end{aligned}$$

□

B. Proof of Lemma 3

Since $H(X_1, ..., X_n) = m \cdot H(X_1, ..., X_k)$, it suffices to lower bound $H(X_1, ..., X_k)$. By the fact that conditioning reduces entropy, we have

$$H\left(X_1, ..., X_k\right) \geq H\left(X_1, ..., X_k \,|\, Y_1, ..., Y_k\right) = \sum_{y^k \in \{0,1\}^k} \mathbb{P}\left(Y^k = y_k\right) \cdot H\left(X^k \,\middle|\, Y^k = y^k\right)$$

where Y_v is the indicator variable of whether vertex v is a seed. Observe that after conditioning on the locations of the seeds, $X_1, ..., X_k$ are mutually independent. Moreover, by symmetry, both $\mathbb{P}\left(Y^k = y^k\right)$ and $H\left(X^k \,\middle|\, Y^k = y^k\right)$ depend on $\sum_i y_i$, (i.e., the empirical distribution of y^k). Indeed, the marginal distribution of X_i can be specified as follows:

$$\mathbb{P}\left(X_i = 1 \,\middle|\, Y^k = y^k\right) = \begin{cases} 1, & \text{if } y_i = 1, \\ 1 - (1-q)^{\left(\sum_i y_i\right)}, & \text{if } y_i = 0, \end{cases}$$

and the conditional entropy is

$$H\left(X^k \,\middle|\, Y^k = y^k\right) = \left(k - \sum_i y_i\right) \cdot \mathsf{h}_\mathsf{b}\left(1 - (1-q)^{\left(\sum_i y_i\right)}\right),$$

where $\mathsf{h}_\mathsf{b}(\cdot)$ is the binary entropy function. Therefore, by writing $Z = \sum_i Y_i$, we have

$$H\left(X^k \,\middle|\, Y^k\right) = \mathbb{E}_Z\left[(k-Z)\cdot \mathsf{h}_\mathsf{b}\left(1-(1-q)^Z\right)\right], \tag{9}$$

where $Z \sim \mathsf{Binom}(k,p)$.

□

C. Proof of Theorem 4

Let $f(q) = \frac{\log(q)}{\log(1-q)}$, so that $f(q)$ solves $1-(1-q)^Z = 1-q$. Then we bound (9) by

$$\begin{aligned}
\mathbb{E}_Z\left[(k-Z)\cdot \mathsf{h}_\mathsf{b}\left(1-(1-q)^Z\right)\right] &\geq \mathbb{E}_Z\left[(k-Z)\cdot \mathsf{h}_\mathsf{b}\left(1-(1-q)^Z\right)\cdot \mathbb{1}_{\{1\leq Z\leq f(q)\}}\right] \\
&\overset{\text{(a)}}{\geq} \mathsf{h}_\mathsf{b}(q)\cdot \mathbb{E}_Z\left[(k-Z)\cdot \mathbb{1}_{\{1\leq Z\leq f(q)\}}\right] \\
&\geq \mathsf{h}_\mathsf{b}(q)\left(\mathbb{E}_Z[k-Z] - k\cdot\mathbb{P}\{Z=0\} - k\cdot\mathbb{P}\{Z>f(q)\}\right) \\
&= k\cdot \mathsf{h}_\mathsf{b}(q)\left((1-p)\left(1-(1-p)^{k-1}\right) - \mathbb{P}\{Z>f(q)\}\right) \\
&\overset{\text{(b)}}{\geq} k\cdot \mathsf{h}_\mathsf{b}(q)\left((1-p)\left((k-1)p-(k-1)^2p^2\right) - \mathbb{P}\{Z>f(q)\}\right) \\
&\overset{\text{(c)}}{\succeq} \frac{k}{2}\cdot \mathsf{h}_\mathsf{b}(q)\left(k\cdot p - \mathbb{P}\{Z>f(q)\}\right),
\end{aligned} \tag{10}$$

where (a) is due to the fact that $\mathsf{h}_\mathsf{b}(x) \geq \mathsf{h}_\mathsf{b}(q)$ for all $q \leq x \leq 1-q$, (b) holds since $(1-p)^r \leq e^{-pr}$ and $e^x \leq 1+x+x^2$ for $x \leq 1$, and (c) is due to the assumption $p \preceq 1/k$.

We then upper bound $\mathbb{P}\{Z > f(q)\}$ by Hoeffding's inequality:

$$\mathbb{P}\{Z>f(q)\} \leq \exp\left(-2k\left(p-\frac{f(q)}{k}\right)^2\right) \overset{\text{(a)}}{\preceq} \exp\left(-2k\left(\frac{f(q)}{2k}\right)^2\right) \leq \exp\left(-\frac{f(q)^2}{2k}\right) \overset{\text{(b)}}{\preceq} \frac{kp}{2},$$

where (a) holds by the assumption $k\cdot p\cdot q \preceq 1$, so that

$$k\cdot p \preceq \frac{q}{2}\log\left(\frac{1}{q}\right) \leq \frac{q}{1-q}\log\left(\frac{1}{q}\right) \leq f(q),$$

and (b) holds due to the assumption $q \preceq \frac{1}{\sqrt{k}\cdot\sqrt{\log\left(\frac{1}{k\cdot p}\right)}}$. Plugging into (10) yields

$$\mathbb{E}_Z\left[(k-Z)\cdot \mathsf{h}_\mathsf{b}\left(1-(1-q)^Z\right)\right] \succeq k^2\cdot p\cdot q\cdot \log\left(\frac{1}{q}\right) \succeq k^2\cdot p\cdot q\cdot\left(\log k + \log\log\left(\frac{1}{kp}\right)\right),$$

where in the last inequality we use the assumption $q \preceq \frac{1}{\sqrt{k}\cdot\sqrt{\log\left(\frac{1}{k\cdot p}\right)}}$ again.

□

D. Proof of Theorem 3

Let T_1 and T_2 be the number of tests performed, respectively, in Step 2 and Step 3 of the graph-aware algorithm. Specifically, T_1 is equal to the number of tests used by binary splitting to identify the infected k-cliques, and T_2 is the number of tests to identify infected individuals within each infected clique. Note that $T = T_1 + T_2$. We will bound $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$ separately.

Let Y be the number of infected k-cliques. We have

$$\mathbb{E}[Y] = \frac{n}{k} \cdot \mathbb{P}(X_{\mathcal{C}_1} = 1) = \frac{n}{k} \cdot \Big(1 - (1-p)^k\Big).$$

Taking Lemma 2 with $n = n/k$ and $\alpha = Y$ gives

$$T_1 \leq (\log_2(n/k) + 1) \cdot Y$$

so that

$$\mathbb{E}[T_1] \leq \frac{n}{k} \cdot \Big(\log_2(n/k) + 1\Big) \cdot \Big(1 - (1-p)^k\Big).$$

For the second stage of the algorithm, let Z_i denote the number of tests used by binary splitting to identify all infected members of the i^{th} clique. Since $T_2 = \sum_{i=1}^{n/k} Z_i \cdot \mathbb{1}_{\{X_{\mathcal{C}_i}=1\}}$, we have

$$\begin{aligned}
\mathbb{E}[T_2] &= \sum_{i=1}^{n/k} \mathbb{E}\left[Z_i \cdot \mathbb{1}_{\{X_{\mathcal{C}_i}=1\}}\right] \\
&= \frac{n}{k} \cdot \mathbb{E}\left[Z_1 \cdot \mathbb{1}_{\{X_{\mathcal{C}_1}=1\}}\right] \\
&= \frac{n}{k} \cdot \mathbb{P}(X_{\mathcal{C}_1} = 1) \cdot \mathbb{E}\left[Z_1 \,|\, X_{\mathcal{C}_1} = 1\right] \\
&= \frac{n}{k} \cdot \Big(1 - (1-p)^k\Big) \cdot \mathbb{E}\left[Z_1 \,|\, X_{\mathcal{C}_1} = 1\right].
\end{aligned}$$

Let M denote the number of infected members of $\mathcal{C}_1$. Then by Lemma 2,

$$\mathbb{E}\left[Z_1 \,|\, X_{\mathcal{C}_1} = 1\right] \leq (\log_2 k + 1) \cdot \mathbb{E}\left[M \,|\, X_{\mathcal{C}_1} = 1\right]$$

and, assuming without loss of generality that $\mathcal{C}_1 = [k]$,

$$\begin{aligned}
\mathbb{E}\left[M \,|\, X_{\mathcal{C}_1} = 1\right] &= \sum_{j=1}^{k} \mathbb{P}\left(X_j = 1 \,|\, X_{\mathcal{C}_1} = 1\right) \\
&= k \cdot \mathbb{P}\left(X_1 = 1 \,|\, X_{\mathcal{C}_1} = 1\right) \\
&= k \cdot \frac{\mathbb{P}(X_1 = 1,\, X_{\mathcal{C}_1} = 1)}{\mathbb{P}(X_{\mathcal{C}_1} = 1)} \\
&= k \cdot \frac{\mathbb{P}(X_1 = 1)}{\mathbb{P}(X_{\mathcal{C}_1} = 1)} \\
&= k \cdot \frac{1 - (1-p)(1-pq)^{k-1}}{1 - (1-p)^k}
\end{aligned}$$

where in the last line we invoke Lemma 1. Putting everything together gives

$$\mathbb{E}[T_2] \leq n \cdot (\log_2 k + 1) \cdot \Big(1 - (1-p)(1-pq)^{k-1}\Big)$$

and therefore

$$\mathbb{E}[T] \leq \frac{n}{k} \cdot \Big(\log_2(n/k) + 1\Big) \cdot \Big(1 - (1-p)^k\Big) + n \cdot \Big(\log_2 k + 1\Big) \cdot \Big(1 - (1-p)(1-pq)^{k-1}\Big).$$

□

E. Proof of Lemma 5

Notice that

$$H\left(X_1,...,X_n\right) \geq H\left(X_1,...,X_n\,|Y_1,...,Y_n\right) = \sum_{y^n\in\{0,1\}^n} \mathbb{P}\left(Y^n=y_n\right)\cdot H\left(X^n\,|\,Y^n=y^n\right).$$

Observe that after conditioning on the locations of the seeds, $X_1,...,X_n$ are mutually independent. Moreover, for $i\in\mathcal{C}_\ell$, the marginal distribution of X_i can be specified as follows:

$$\mathbb{P}\left(X_i=1\,|\,Y^n=y^n\right) = \begin{cases} 1, & \text{if } y_i=1,\\ 1-(1-q_1)^{\sum_{j\in\mathcal{C}_\ell} y_j}\,(1-q_2)^{\sum_{j\notin\mathcal{C}_\ell} y_j}, & \text{if } y_i=0. \end{cases}$$

Writing $z_\ell \triangleq \sum_{j\in\mathcal{C}_\ell} y_j$, the conditional entropy is

$$H\left(X^n\,|\,Y^n=y^n\right) = \sum_{\ell=1}^{m}(k-z_\ell)\cdot \mathsf{h_b}\left(1-(1-q_1)^{z_\ell}\,(1-q_2)^{\sum_{\ell'\neq\ell} z_{\ell'}}\right),$$

where $\mathsf{h_b}\left(\cdot\right)$ is the binary entropy function. Since $Y_i \overset{\text{i.i.d.}}{\sim} \mathsf{Ber}(p)$, we have $Z_\ell \overset{\text{i.i.d.}}{\sim} \mathsf{Binom}(k,p)$ and hence

$$H\left(X^n\,|\,Y^n\right) = \mathbb{E}_{Z,Z'}\left[m\cdot(k-Z)\cdot\mathsf{h_b}\left(1-(1-q_1)^Z(1-q_2)^{Z'}\right)\right], \tag{11}$$

where $Z\sim\mathsf{Binom}\left(k,p\right)$ and $Z'\sim\mathsf{Binom}\left(n-k,p\right)$.

□

F. Proof of Theorem 6

First we assume $n\cdot p\cdot q_2 \preceq 1$, and let $\epsilon\in(0,1)$ be a value to be specified. Define

$$z^* \triangleq \frac{1/2-np(1+\epsilon)q_2}{q_1}.$$

Then as long as Z and Z' satisfy the following two conditions

1) $\{np(1-\epsilon)\leq Z'\leq np(1+\epsilon)\}$,
2) $Z\leq z^*$,

we have

$$\frac{1}{2}\geq Z\cdot q_1 + Z'\cdot q_2 \geq 1-(1-q_1)^Z\,(1-q_2)^{Z'}. \tag{12}$$

Since $1-(1-q_1)^Z\,(1-q_2)^{Z'}$ is an increasing function of Z and Z', $\mathsf{h_b}\left(1-(1-q_1)^Z\,(1-q_2)^{Z'}\right)$ must increase with Z and Z' if they satisfy the above conditions. Therefore, we have

$$\begin{aligned} &\mathbb{E}_{Z,Z'}\left[(k-Z)\mathsf{h_b}\left(1-(1-q_1)^Z\,(1-q_2)^{Z'}\right)\right]\\ \geq&\mathbb{E}_{Z,Z'}\left[(k-Z)\mathsf{h_b}\left(1-(1-q_1)^Z\,(1-q_2)^{Z'}\right)\cdot\mathbb{1}_{\{0\leq Z\leq z^*\}}\cdot\mathbb{1}_{\{np(1-\epsilon)\leq Z'\leq np(1+\epsilon)\}}\right]\\ \geq&\underbrace{\mathbb{E}_{Z,Z'}\left[(k-Z)\mathsf{h_b}\left(1-(1-q_2)^{Z'}\right)\cdot\mathbb{1}_{\{Z=0\}}\cdot\mathbb{1}_{\{np(1-\epsilon)\leq Z'\leq np(1+\epsilon)\}}\right]}_{\text{(a)}}+\\ &\underbrace{\mathbb{E}_{Z,Z'}\left[(k-Z)\mathsf{h_b}\left(1-(1-q_1)^Z\,(1-q_2)^{Z'}\right)\cdot\mathbb{1}_{\{1\leq Z\leq z^*\}}\cdot\mathbb{1}_{\{np(1-\epsilon)\leq Z'\leq np(1+\epsilon)\}}\right]}_{\text{(b)}}. \end{aligned} \tag{13}$$

We will pick $\epsilon = \frac{1}{2}$. Then (a) can be bounded by

$$\begin{aligned}
\text{(a)} &\geq k \cdot \mathsf{h_b}\left(q_2 \cdot np(1-\epsilon) - (q_2 \cdot np(1-\epsilon))^2\right)\left(1 - 2 \cdot \exp\left(-\frac{n\epsilon^2 p}{3}\right)\right) \\
&\succeq k\left(npq_2(1-\epsilon)\log\left(\frac{1}{npq_2(1-\epsilon)}\right)\left(1 - 2 \cdot \exp\left(-\frac{n\epsilon^2 p}{3}\right)\right)\right) \\
&\succeq k\left(npq_2 \log\left(\frac{1}{npq_2}\right)\right)
\end{aligned}$$

where in the first inequality we use

1) $Z' \geq np(1-\epsilon)$
2) $(1-q_2)^{Z'} \leq e^{-q_2 \cdot Z'} \leq 1 - q_2 \cdot Z' + (q_2 \cdot Z')^2$
3) Chernoff bound on Z',

and in the third inequality we assume $np \succeq 1$. Next, (b) can be bounded by

$$\begin{aligned}
\text{(b)} &\geq \mathsf{h_b}\left(q_1 + npq_2(1-\epsilon) - (q_1 + npq_2(1-\epsilon))^2\right) \cdot \mathbb{E}_Z\left[(k-Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right] \cdot \left(1 - 2 \cdot \exp\left(-\frac{n\epsilon^2 p}{3}\right)\right) \\
&\succeq (q_1 + npq_2)\log\left(\frac{1}{q_1 + npq_2}\right) \cdot \mathbb{E}_Z\left[(k-Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right].
\end{aligned}$$

We will now lower bound $\mathbb{E}_Z\left[(k-Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right]$ as in Theorem 4. Observe that

$$\begin{aligned}
\mathbb{E}_Z\left[(k-Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right] &\geq \mathbb{E}_Z[k-Z] - k \cdot \mathbb{P}\{Z = 0\} - k \cdot \mathbb{P}\{Z \geq z^*\} \\
&\geq k\left(1 - p - (1-p)^k - \mathbb{P}\{Z \geq z^*\}\right) \\
&\succeq k\left(kp - \mathbb{P}\{Z \geq z^*\}\right). \qquad (14)
\end{aligned}$$

Finally, applying Hoeffding's inequality to $\mathbb{P}\{Z \geq z^*\}$ yields

$$\begin{aligned}
\mathbb{P}\{Z \geq z^*\} &\leq \exp\left(-2k\left(p - \frac{z^*}{k}\right)^2\right) = \exp\left(-2k\left(p - \frac{\frac{1}{2} - npq_2(1+\epsilon)}{q_1 k}\right)^2\right) \\
&\overset{(1)}{\preceq} \exp\left(-2k\left(\frac{1}{2q_1 k}\right)^2\right) = \exp\left(-\frac{1}{2kq_1^2}\right) \overset{(2)}{\leq} \frac{kp}{2},
\end{aligned}$$

where in (1) we use the facts that 1) $n \cdot p \cdot q_2 \preceq 1$ and 2) $p \preceq \frac{1}{q_1 k}$, and (2) holds when

$$q_1 \leq \frac{1}{\sqrt{2k \cdot \left(\log\left(\frac{1}{kp}\right) + 1\right)}}.$$

Plugging into (14) yields

$$\mathbb{E}_Z\left[(k-Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right] \succeq k^2 p, \qquad (15)$$

and thus by putting together our bounds on (a) and (b) in (13), we arrive at

$$\begin{aligned}
&\mathbb{E}_{Z,Z'}\left[(k-Z)\mathsf{h_b}\left(1 - (1-q_1)^Z (1-q_2)^{Z'}\right)\right] \qquad (16) \\
\geq &k\left(npq_2 \log\left(\frac{1}{npq_2}\right)\right) + k^2 p \cdot (q_1 + npq_2)\log\left(\frac{1}{q_1 + npq_2}\right) \qquad (17) \\
\geq &mk^2 pq_2 \log\left(\frac{1}{npq_2}\right) + k^2 p \cdot q_1 \log\left(\frac{1}{q_1 + npq_2}\right). \qquad (18)
\end{aligned}$$

□

G. Proof of Lemma 7

Let Y_v be the indicator random variable of whether vertex v is a seed, and assume without loss of generality that $v \in \mathcal{C}_1$. We have

$$\begin{aligned}\mathbb{P}(X_v = 1) &= \underbrace{\mathbb{P}(X_v = 1 \,|\, Y_v = 1)}_{=1} \cdot \underbrace{\mathbb{P}(Y_v = 1)}_{=p} + \mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right)) \cdot \mathbb{P}(Y_v = 0) \\ &= p + (1-p) \cdot \mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left(X_v = 1 \,|\, Y_v = 0\right) &= \mathbb{P}\big(\{v \text{ is infected by a neighbor}\}\big) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \mathbb{P}\big(\{v \text{ isn't infected by } u\}\big) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \Big(1 - \mathbb{P}\big(\{v \text{ is infected by } u\}\big)\Big) \\ &= 1 - \prod_{u \in \mathcal{N}(v)} \Big(1 - \mathbb{P}\left(\{v \text{ is infected by } u\} \,|\, Y_u = 1\right) \cdot \mathbb{P}(Y_u = 1)\Big) \\ &= 1 - \left(\prod_{u \in \mathcal{C}_1 \setminus \{v\}} (1 - p \cdot q_1)\right) \cdot \left(\prod_{w \notin \mathcal{C}_1} (1 - p \cdot q_2)\right) \\ &= 1 - (1 - p \cdot q_1)^{k-1} \cdot (1 - p \cdot q_2)^{n-k}.\end{aligned}$$

□

H. Proof of Lemma 8

Let $\mathcal{A}$ be the event that no member of community $\mathcal{C}_1$ is selected as a seed, and let $\mathcal{B}$ be the event that some member of $\mathcal{C}_1$ is infected by an individual outside $\mathcal{C}_1$. We further denote by $\mathcal{B}_u$ the event that vertex u infects some member of $\mathcal{C}_1$, where $u \notin \mathcal{C}_1$. Note that $X_{\mathcal{C}_1} = 1$ if and only if either $\mathcal{A}^c$ occurs or $\mathcal{A} \cap \mathcal{B}$ occurs. Moreover, $\mathcal{A}$ and $\mathcal{B}$ are independent events. We have that $\mathbb{P}(\mathcal{A}) = (1-p)^k$, and thus

$$\begin{aligned}\mathbb{P}(X_{\mathcal{C}_1} = 1) &= \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{A}) \cdot \mathbb{P}(\mathcal{B}) \\ &= 1 - (1-p)^k + (1-p)^k \cdot \mathbb{P}(\mathcal{B}) \\ &= 1 - (1-p)^k \cdot (1 - \mathbb{P}(\mathcal{B})).\end{aligned}$$

Finally, we compute $\mathbb{P}(\mathcal{B})$ as

$$\begin{aligned}
\mathbb{P}(\mathcal{B}) &= 1 - \prod_{u \notin \mathcal{C}_1} \mathbb{P}(\mathcal{B}_u^c) \\
&= 1 - \prod_{u \notin \mathcal{C}_1} \Big(\mathbb{P}(\mathcal{B}_u^c \,|\, Y_u = 1) \cdot \underbrace{\mathbb{P}(Y_u = 1)}_{=p} + \underbrace{\mathbb{P}(\mathcal{B}_u^c \,|\, Y_u = 0)}_{=1} \cdot \underbrace{\mathbb{P}(Y_u = 0)}_{=1-p} \Big) \\
&= 1 - \prod_{u \notin \mathcal{C}_1} \Big(1 - p + p \cdot \mathbb{P}(\mathcal{B}_u^c \,|\, Y_u = 1) \Big) \\
&= 1 - \prod_{u \notin \mathcal{C}_1} \Big(1 - p + p \cdot (1 - q_2)^k \Big) \\
&= 1 - \Big(1 - p \cdot \Big(1 - (1 - q_2)^k \Big) \Big)^{n-k} .
\end{aligned}$$

□