Adaptive Group Testing on Networks with Community Structure: The Stochastic Block Model

Surin Ahn, *Student Member, IEEE,* Wei-Ning Chen, and Ayfer Özgür *Senior Member, IEEE*

Abstract—Group testing was conceived during World War II to identify soldiers infected with syphilis using as few tests as possible, and it has attracted renewed interest during the COVID-19 pandemic. A long-standing assumption in the probabilistic variant of the group testing problem is that individuals are infected by the disease *independently*. However, this assumption rarely holds in practice, as diseases often spread through interactions between individuals and therefore cause infections to be correlated. Inspired by characteristics of COVID-19 and other infectious diseases, we introduce an infection model over networks which generalizes the traditional i.i.d. model from probabilistic group testing. Under this 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 a simple community-aware algorithm outperforms the baseline binary splitting algorithm when the model parameters are conducive to "strong community structure." Moreover, our novel lower bounds imply that the community-aware algorithm is order-optimal in certain parameter regimes. We extend our bounds to the noisy setting and support our results with numerical experiments.

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]–[17].

Dorfman's seminal work and many subsequent works by other authors [18]–[24] assume that the disease infects individuals in a statistically independent fashion. The simplest and most widely studied case, known as the *i.i.d. model* or *binomial model*, assumes individuals are infected independently with some common probability p.[1] 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. 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 between individuals in close contact with each other.

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 introduce a novel community-oriented infection model, called the *stochastic block infection model* (SBIM), which generalizes the standard i.i.d. model to a setting in which the disease can be transmitted between individuals. Our model is equivalent to a certain graph-based infection spread

[1]In a related, commonly studied probabilistic model—often called the *combinatorial prior model* or the *hypergeometric model* —it is assumed that a random set of d individuals out of n are infected according to some distribution (typically uniform) over all $\binom{n}{d}$ possibilities [25]–[34]. While this is slightly different from an i.i.d. assumption, it is still somewhat simplistic and fails to capture any dependencies that may exist between individuals.

mechanism operating upon the well-known *stochastic block model* (SBM) for random graphs. For decades, the SBM has been utilized across the social, biological, and information sciences as a very simple yet natural way to model community structure in probabilistic networks.

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 [35] which is oblivious to the underlying network structure, and a simple *community-aware* algorithm which essentially performs two stages of binary splitting: the first stage identifies the communities containing at least one infected member, and the second stage performs more fine-grained testing within the infected communities. We give precise upper bounds on the expected number of tests performed by each algorithm. Crucially, we show that when the model parameters yield "strong community structure" (in which case the disease is much more likely to be transmitted within a community than between communities), the community-aware algorithm's average complexity is asymptotically strictly better than that of binary splitting. Furthermore, we derive novel information-theoretic lower bounds that apply to all adaptive strategies and imply the order-optimality of the community-aware algorithm in certain parameter regimes. We then extend our algorithms and bounds to the noisy setting—in which the test outcomes are passed through a binary symmetric channel—and find that the presence of noise does not affect the relative gains of using a community-oriented approach. Finally, we corroborate our results with numerical experiments. 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 epidemiology. In the past, group testing has been successfully applied to diverse domains including wireless communications [19], [22], [36]–[41], machine learning [42]–[44], signal processing [45], [46], and the analysis of data streams [47], [48]. 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 [49]–[52], 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 packets, then each test must correspond to a valid path in the network. By contrast, our problem setup permits arbitrary tests, but we ask whether *knowledge* of the interaction network can help reduce the number of tests.

There is a rich literature on adaptive group testing dating back to the early work of Dorfman [1] and others [20], [53]–[59], with several important results having emerged in recent years, e.g., [60], [61], [31], [62], [63], [64]. These works focus on relatively simple combinatorial or probabilistic models. A few prior works have departed from these standard models by assuming that infections occur independently with non-identical prior probabilities [18], [23], [24]. 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.[2]

The idea of community-aware group testing was first explored in [13], which assumed the population is partitioned into disjoint "families" and that the disease spreads in two stages with independent infections at each stage. Our work considers an infection mechanism which similarly operates in two stages but is designed to model the interaction-based transmissions by which diseases often spread in reality. Finally, we would like to acknowledge a number of independent and concurrent works related to community-aware group testing [16], [65]–[68].

b) Notation: Let $[n] \triangleq \{1, 2, \ldots, n\}$. 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. $X \triangleq (X_1, \ldots, X_n) \in \{0, 1\}^n$ is the infection status vector, where $X_i = 1$ iff the ith individual is infected. With a slight abuse of notation, let $X_{\mathcal{C}_i} \in \{0, 1\}$, $i \in [m]$, be the infection status of community $\mathcal{C}_i$, where $X_{\mathcal{C}_i} = 1$ iff $\exists i \in \mathcal{C}_i : X_i = 1$. The indicator function for an event $\mathcal{A}$ is given by $\mathbb{1}_{\mathcal{A}}$. The entropy of a discrete random variable and the binary entropy function (both in bits) are $H(\cdot)$ and $\mathsf{h}_{\mathsf{b}}(\cdot)$, respectively. We write $f(x) \prec g(x)$ to denote $f(x) = o(g(x))$, and $f(x) \preceq g(x)$ to denote $f(x) = O(g(x))$.

c) Paper Organization: The rest of this paper is organized as follows. In Section II, we provide background and preliminary results. In Section III, we introduce the stochastic block infection model (SBIM) and discuss its equivalence to a certain graph-based infection spread mechanism acting upon the stochastic block model. In Section IV, we discuss the main algorithms studied in this paper: binary splitting and our proposed community-aware algorithm. Section V gives upper and lower

---

[2]For the sake of obtaining comprehensive results, we focus on a symmetric model in which infections are correlated and identically distributed. However, the general infection model that we propose in Section III-C is fully non-i.i.d.

bounds for adaptive group testing over the general SBIM, and Section VI provides an in-depth treatment of the disjoint k-cliques model, which is a special case of the SBIM. We then extend our algorithms and bounds to the noisy case in Section VII. Finally, we present the results of our numerical experiments in Section VIII, and conclude in Section IX. All omitted proofs are given in the Appendices.

II. BACKGROUND AND PRELIMINARY RESULTS

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 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 describes how to select subsets $\mathcal{S}_1, \ldots, \mathcal{S}_T$ and—given the corresponding test outcomes $Y_1, \ldots, Y_T$—how to generate an estimate $\hat{X}$ of X. In *adaptive* schemes, the subsets $\mathcal{S}_t$ are chosen sequentially and are allowed to depend on the previous test outcomes. In the first part of this paper, we assume that test outcomes are *noiseless* (meaning the algorithm gets to observe Y as given in (1)), and we require *exact recovery* of $X_1, \ldots, X_n$ (i.e., zero error).

Subsequently, we consider a *noisy* variant of the problem in which the test outcomes are given by

$$Y = \Big(\bigvee_{i \in \mathcal{S}} X_i \Big) \oplus \xi, \tag{2}$$

where $\xi \sim \text{Bernoulli}(\rho)$ for some $\rho \in (0, \frac{1}{2})$, and $\oplus$ denotes modulo-2 addition. This is the widely-adopted *symmetric noise model* [69]–[71], and it is equivalent to passing each noiseless test outcome through a binary symmetric channel with crossover probability ρ. It is assumed that tests are subject to independent noise. Due to the uncertainty in the test outcomes, we can no longer guarantee exact recovery of X. Instead, we seek to ensure a vanishing error probability $P_e \triangleq \Pr(\hat{X} \neq X)$, where the randomness is due to the infection statuses and the noisy test outcomes.

In our setting, the number of tests T performed by an adaptive scheme is a random variable because it depends on the X_i, which are generated by our probabilistic SBIM model, as well as the (possibly noisy) test results. Our goal is to characterize the average complexity of adaptive schemes under the SBIM by providing both upper and lower bounds on $\mathbb{E}[T]$.

B. Information-Theoretic Lower Bounds

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

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

This bound highlights the intimate connection between adaptive group testing and source coding. Indeed, to summarize a discussion from [19], 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., [72, Eqn. 5.38]). Equation (3) will serve as the point of departure for the lower bounds on $\mathbb{E}[T]$ that we derive under the SBIM in the noiseless case. The key challenge will be to obtain good approximations to $H(X)$ in the presence of correlations induced by the underlying network.

For the noisy setting, we prove the following counterpart to (3). This lower bound holds for any adaptive scheme and any underlying stochastic infection model, including those with correlations. We provide the proof in Section VII.

Theorem 1. *Assume* $H(X_1, X_2, \ldots, X_n) \to \infty$ *as* $n \to \infty$. *Under the symmetric noise model* (2), *any adaptive algorithm achieving* $P_e \to 0$ *must use an average number of tests lower bounded as*

$$\mathbb{E}[T] \geq \frac{H(X_1, \ldots, X_n)}{I(\rho)}, \tag{4}$$

where $I(\rho) = 1 - \mathsf{h}_{\mathsf{b}}(\rho) = 1 - \rho \log_2 \frac{1}{\rho} - (1-\rho) \log \frac{1}{1-\rho}$ *is the capacity of the binary symmetric channel with crossover probability* $\rho \in (0, \frac{1}{2})$.

Note that our bound recovers the noiseless lower bound (3) when $\rho = 0$. Moreover, in the special case of the combinatorial prior model where the number of infections d is fixed and the set of infected members is uniformly distributed over the $\binom{n}{d}$ possibilities, our bound reduces to $\mathbb{E}[T] \geq \frac{\log \binom{n}{d}}{I(\rho)}$. A version of this bound appears in [71], which does not prove it directly but argues it can be shown using an existing result from [61] along with the variable-length coding capacity of the binary symmetric channel. In Section VII, we provide a stand-alone proof of the more general lower bound in Theorem 1 which encompasses all adaptive schemes and probabilistic infection models with symmetric testing noise. Though we will primarily focus on the implications of Theorem 1 in the context of the SBIM, we again emphasize that this result is independent of any particular infection model and thus can be of interest in its own right.

III. STOCHASTIC BLOCK INFECTION MODEL (SBIM)

In this section, we introduce the *stochastic block infection model* (SBIM), which extends the traditional i.i.d. group testing model to a community-oriented setting. Here, individuals infect their fellow community members with a higher probability than those in other communities, giving rise to strongly correlated clusters of infections. We then describe a special case of the SBIM—the *disjoint k-cliques model*—in which the communities are "disconnected." Finally, we discuss the relationship between the SBIM and the stochastic block model (SBM), and conclude the section with some practical considerations.

A. General SBIM

Assume we are given a partition of the population of size n into $m \triangleq n/k$ communities $\mathcal{C}_1, \ldots, \mathcal{C}_m$ of size $|\mathcal{C}_i| = k$, $\forall i \in [m]$. The SBIM comprises the following two stages (each executed once):

1) **Seed Selection:** Individuals in the population are infected independently with probability $p \in (0, 1]$. These initial infected members 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 infects its neighbors within the same community independently with probability $q_1 \in [0, 1]$ and those outside its community independently with probability $q_2 \in [0, 1]$, where $q_1 > q_2$. This models how the disease spreads through the population via interactions between carriers and nearby individuals. Members of the same community are more likely to interact with each other within a given time frame (e.g., by interacting socially or professionally, or by being in the same physical space, e.g., a supermarket or a restaurant) and therefore more likely to infect each other than members of different communities.

We denote this model by $\mathsf{SBIM}(n, k, p, q_1, q_2)$. Note that $\mathsf{SBIM}(n, k, p, 0, 0)$, for any value of k, is equivalent to the i.i.d. group testing model with prior probability p. We assume the communities are known to the group testing algorithms in advance, but that nothing more is known about the specific interactions between individuals.

The SBIM can be viewed as a model for the initial spread of an epidemic. It is motivated in part by diseases such as COVID-19, which are introduced into a population from an external source and subsequently transmitted between individuals in close contact. We also believe the SBIM can be a natural model for other application areas where group testing has played a role.

For example, in the context of coding for multiple access sensor networks [38]–[40] it can capture the fact that sensors in close proximity can have correlated activity patterns and measurements.

B. Special Case: Disjoint k-Cliques Model

After analyzing the SBIM in full generality in Section V, we thoroughly investigate the special case of $\mathsf{SBIM}(n, k, p, q, 0)$, which we call the *disjoint k-cliques model*, in Section VI. Here, we have $m \triangleq n/k$ communities of size k, with seed selection probability p, intra-community transmission rate q, and an inter-community transmission rate of zero. Thus, the communities can be treated as independent, as no transmissions between communities are possible. We note that in this special case our model becomes similar (but not equivalent) to the disjoint families model introduced in [13]. We comment further on this in Section VI. Figure 1 illustrates the $\mathsf{SBIM}(n, k, p, q_1, q_2)$ and contrasts the disjoint k-cliques model ($q_2 = 0$) with the general SBIM ($q_2 > 0$).

C. Relationship to the Stochastic Block Model (SBM)

The SBIM is equivalent to a certain graph-based infection spread model operating upon the *stochastic block model* (SBM) [73]—a well-known random graph model with the tendency to produce graphs with community structure. The standard SBM produces a random undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as follows. (In our context, the vertices $\mathcal{V}$ represent members of the population, and the edges $\mathcal{E}$ can be thought of as representing an interaction (e.g., a social or professional interaction, being in proximity, etc.) between the two members of the population in a time frame of interest.) First, it is assumed the n vertices are partitioned into m communities, $\mathcal{C}_1, \ldots, \mathcal{C}_m$, where $\bigcup_{i \in [m]} \mathcal{C}_i = \mathcal{V}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$, $\forall i \neq j$. In addition, we are given a symmetric matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ of edge probabilities. The graph 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 $\mathbf{P}_{ij}$.

Consider a special case of the SBM in which the communities are all of size k (where k is a factor of n), the edge probabilities within each community are constant (p_1), and the edge probabilities between communities are also constant (p_2, where $p_2 < p_1$, which models the assumption that members of the same community are more likely to have an interaction). That is, the diagonal entries of $\mathbf{P}$ are p_1, and the off-diagonal entries are p_2. Additionally, consider the following probabilistic infection model with parameters $p, q \in [0, 1]$ which operates upon an arbitrary graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. First, the vertices are infected independently with probability p, producing the seeds $\mathcal{V}_s \subseteq \mathcal{V}$. Next, every seed $v \in \mathcal{V}_s$
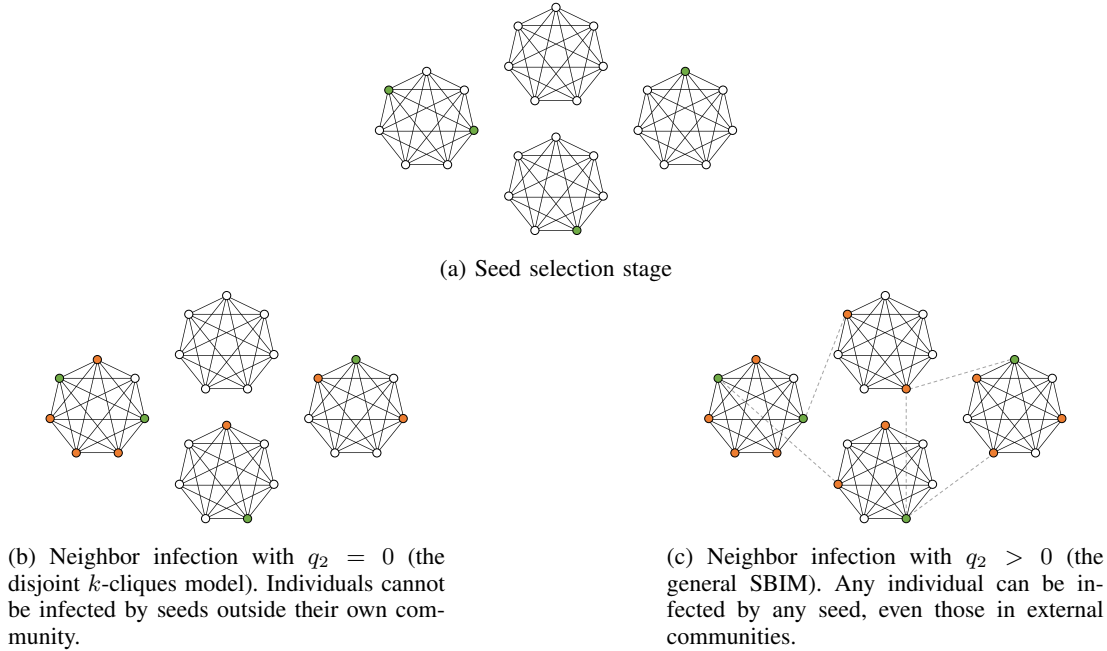
(a) Seed selection stage

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

(c) Neighbor infection with $q_2 > 0$ (the general SBIM). Any individual 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 individuals infected by seeds during the neighbor infection stage are colored orange.

infects its neighbors $\mathcal{N}(v) = \{u \in \mathcal{V} : \{u, v\} \in \mathcal{E}\}$ independently with probability q. This models the fact that if two members have an interaction, the disease is transmitted between them with a certain probability q. [3] Note that this infection model reduces to the i.i.d. group testing model with prior p when (i) $q = 0$, or (ii) $\mathcal{G}$ is the empty graph ($\mathcal{E} = \emptyset$). Moreover, by setting $q_1 = p_1 \cdot q$ and $q_2 = p_2 \cdot q$, we see that the $\mathsf{SBIM}(n, k, p, q_1, q_2)$ is equivalent to this infection model operating upon the SBM.

D. Practical Considerations

The "communities" within the SBIM can represent populations at different scales: counties, cities, schools, companies, etc. 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 of the disease). Lastly, when the communities are not known in advance, one might first estimate the network from data (e.g., contact tracing, mobile phone, or social network data), then run a graph clustering algorithm to identify communities in the network.

At the same time, we acknowledge the practical limitations of the SBIM. First, the symmetry of the model (e.g., the assumptions that every community has the same

[3]This infection model forms the "first time step" of the independent cascade model [74] from the study of influence maximization in social networks.

probability of containing a seed and that a given individual can be infected by a seed from any community) does not capture the reality that the transmissibility of a disease can vary from person to person depending on their habits (e.g., whether they practice social distancing or mask wearing). However, we still believe the SBIM is an important and natural "first-order" extension of the traditional i.i.d. group testing model (which has been studied for decades) to models of greater complexity and practical relevance, while still being analytically tractable. We note that some of the aforementioned issues can be incorporated through the study of the general graph infection model we introduced in Section III-C, by assuming the matrix $\mathbf{P}$ has a more general structure, e.g., by allowing different edge probabilities within and/or between different blocks, and/or different block sizes. Studying the group testing problem in these more general settings is an exciting direction to pursue in future work.

IV. ALGORITHMS

A. Binary Splitting Algorithm

Most adaptive group testing algorithms are based on the idea, first introduced by Sobel and Groll [20], of recursively splitting the population until all infected members are found. The most fundamental adaptive procedure is *binary splitting*, which finds a single infected member at a time by repeatedly halving the population. It works even when the number of infected members d is

unknown [57], and is most effective in the sparse regime, $d = \Theta(n^{\beta})$, where $\beta \in [0, 1)$. We make extensive use of the following performance guarantee throughout this paper:[4]

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

Proof. The first step of binary splitting is to perform a single test on the entire population to check for the presence of an infected member. If the test is positive, an infected member is identified in a recursive fashion using at most $\lceil \log_2 n \rceil$ adaptive tests (see [61], [35, p.24-25], or [75, Theorem 1.2] for details of the proof). The infected individual is then removed from the population and the aforementioned steps are repeated until either no individuals remain, or a negative test is obtained in the first step. It is straightforward to see that $d + 1$ tests are performed due to the first step (once per infected member, and again when no infections remain), and $d\lceil \log_2 n \rceil$ total tests are used to recursively identify all d infected members. $\square$

We treat binary splitting as the baseline algorithm in this paper due to its simplicity and its role as a key subroutine in many other adaptive procedures.

B. Community-Aware Algorithm

As an alternative to standard adaptive procedures such as binary splitting, we consider a simple two-stage scheme which leverages the community structure of the graph. Our scheme first treats the communities as "meta-individuals" by mixing the samples within each community and applying binary splitting to quickly identify 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 Community-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.

---

[4]It is well-known that Lemma 1 can be improved via Hwang's *generalized binary splitting* algorithm [56] or Allemann's *split and overlap* algorithm [60]. However, in contrast to binary splitting, these methods require the number of infected individuals (or an upper bound on this quantity) to be known *a priori*.

3) For each positive test from Step 2, perform binary splitting within the corresponding community to identify the infected members.

Under what circumstances would we expect the community-aware algorithm to outperform binary splitting? Suppose the underlying model is $\mathsf{SBIM}(n, k, p, q_1, q_2)$. If the seed selection probability p is small, then we expect only a few of the $m \triangleq n/k$ communities to contain a seed. Thus, 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 community-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 community-aware algorithm under these circumstances. We will rigorously verify this intuition in the upcoming sections.

V. BOUNDS FOR THE SBIM

In this section, we derive general lower and upper bounds on the average complexity of adaptive group testing over the $\mathsf{SBIM}(n, k, p, q_1, q_2)$. As we saw in the previous section, the community-aware algorithm is a simple extension of the binary splitting algorithm to a community-oriented setting. From a technical perspective, our main contribution is a careful evaluation of the performance of these schemes for the SBIM model as well as the system entropy $H(X)$, which is required to obtain meaningful lower bounds. We start with the lower bound.

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. The following lemma gives both a general lower bound on $H(X)$ as well as an easier-to-compute bound in terms of two independent binomial random variables. The proof is in Appendix A.

Lemma 2. *Let $X_1, ..., X_n$ be the infection statuses generated from $\mathsf{SBIM}(n, k, p, q_1, q_2)$, as defined in Section III, and let S_i be the indicator variable of whether the ith individual is a seed. Then the number of tests*

T required to identify the infected individuals is lower bounded as

$$
\begin{aligned}
\mathbb{E}[T] &\geq H(X_1, ..., X_n) \\
&\geq n \cdot I(X_1; S_1) + H(X_1, \ldots, X_n \mid S_1, \ldots, S_n) \quad (5) \\
&\geq m \cdot \mathbb{E}_{Z,Z'} \left[(k - Z) \, \mathsf{h_b} \left(1 - (1 - q_1)^Z (1 - q_2)^{Z'} \right) \right], \\
&\hspace{8cm} (6)
\end{aligned}
$$

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

By leveraging the concentration of Z and Z' around their means, we obtain our first main result, which characterizes the asymptotic behavior of (6). The proof is in Appendix B.

Theorem 2 (SBIM Lower Bound). *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) $q_1 \leq \frac{1}{\sqrt{2k\left(\log\left(\frac{1}{kp}\right)+1\right)}}$.

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*

$$
\begin{aligned}
\mathbb{E}[T] \succeq\; &m^2 k^2 p q_2 \cdot \log \left(\frac{1}{npq_2} \right) \\
&+ m k^2 p q_1 \cdot \log \left(\frac{1}{q_1 + npq_2} \right).
\end{aligned}
$$

Remark 1. *The upper bounds on p, q_1 and q_2 in Theorem 2 allow us to evaluate the lower bound in Lemma 2 in a regime where the infected population is sparse enough. This is the relevant regime since group testing is known to improve upon individual testing when infections are sparse. However, the specific upper bounds we impose may be artifacts of our lower bounding technique and could potentially be loosened.*

As we will see in Section VI-A, a secondary lower bound under $\mathsf{SBIM}(n, k, p, q, 0, 0)$ (i.e., the disjoint k-cliques model) is given by $H(X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_m}) = \sum_{i \in [m]} H(X_{\mathcal{C}_i})$, which leverages the fact that the community-level infection statuses $\{X_{\mathcal{C}_1}, ..., X_{\mathcal{C}_m}\}$ are mutually independent in this setting. This bound turns out to dominate when $kp \preceq m^{-\beta}$ for some fixed $\beta \in (0, 1)$. It is difficult to obtain an analogous lower bound under the general SBIM since the $\{X_{\mathcal{C}_1}, ..., X_{\mathcal{C}_m}\}$ are no longer mutually independent when $q_2 > 0$. Therefore, we suspect that the lower bound given in Theorem 2 is not tight when kp is small. Obtaining a tighter bound in this regime is an open problem.

B. Algorithm Analysis

To analyze binary splitting and the community-aware algorithm over the SBIM, we begin by characterizing the marginal probability that a given individual will be infected. The proof is in Appendix C.

Lemma 3. *The marginal probability of infection for every individual 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: The following result bounds the expected number of tests used by the binary splitting algorithm under the SBIM.

Theorem 3 (Binary Splitting Bound). *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$, *the binary splitting algorithm identifies all infected individuals using T tests, where*

$$
\begin{aligned}
\mathbb{E}[T] \leq\; &n \cdot (\log_2 n + 2) \\
&\cdot \left(1 - (1 - p)(1 - p \cdot q_1)^{k-1}(1 - p \cdot q_2)^{n-k} \right) \\
&+ 1.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}[K] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1) \\
&= n \cdot \left(1 - (1 - p)(1 - p \cdot q_1)^{k-1}(1 - p \cdot q_2)^{n-k} \right)
\end{aligned}
$$

where the last equality follows from Lemma 3. Invoking Lemma 1 yields the result. $\qquad\square$

Corollary 1. *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$, *the average complexity of binary splitting satisfies*

$$
\begin{aligned}
\mathbb{E}[T] \preceq\; &m k^2 p \cdot (\log m + \log k) \\
&\cdot \left(\frac{1}{k} + q_1 + m q_2 + m k p^2 q_1 q_2 \right).
\end{aligned}
$$

Proof. Using the fact that $(1 + x)^k \geq 1 + kx$ for $x \geq -1$, $k \geq 1$, we have

$$
\begin{aligned}
\mathbb{E}[T] \preceq\; &n \cdot \log n \\
&\cdot \left(1 - (1 - p)(1 - kpq_1)(1 - (n - k) \cdot pq_2) \right) \\
\leq\; &n \cdot \log n \\
&\cdot \left((n - k)pq_2 + kpq_1 + p + k(n - k)p^3 q_1 q_2 \right) \\
\leq\; &m k^2 p \cdot (\log m + \log k) \\
&\cdot \left(\frac{1}{k} + q_1 + m q_2 + m k p^2 q_1 q_2 \right). \quad (7)
\end{aligned}
$$
$\qquad\square$

2) Community-Aware Algorithm: First, we provide a lemma needed to prove the upper bound for the community-aware algorithm. The proof is in Appendix D.

Lemma 4. *Let $X_{\mathcal{C}_i} \in \{0, 1\}$, $i \in [m]$, be the infection status of community $\mathcal{C}_i$, where $X_{\mathcal{C}_i} = 1$ iff*

there exists at least one infected member in $\mathcal{C}_i$. Under $\mathsf{SBIM}(n, k, p, q_1, q_2)$,

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

In Theorem 4 below (which is proved in Appendix F), the two terms in the sum correspond, respectively, to the expected number of tests in Steps 2 and 3 of the community-aware algorithm.

Theorem 4 (Community-Aware Bound)**.** *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$, *the community-aware algorithm identifies all infected individuals using T tests, where*

$$
\begin{aligned}
\mathbb{E}[T] \leq {} & \frac{n}{k} \cdot \Big(\log_2(n/k) + 3 \Big) \\
& \cdot \left(1 - (1-p)^k \left(1 - p \Big(1 - (1-q_2)^k \Big) \right)^{n-k} \right) \\
& + n \cdot \Big(\log_2 k + 2 \Big) \\
& \qquad \cdot \Big(1 - (1-p)(1 - pq_1)^{k-1}(1 - pq_2)^{n-k} \Big) \\
& + 1.
\end{aligned}
$$

Corollary 2. *Under* $\mathsf{SBIM}(n, k, p, q_1, q_2)$, *the average complexity of the community-aware algorithm satisfies*

$$
\begin{aligned}
\mathbb{E}[T] \preceq {} & mkp \cdot \log m \cdot \Big(1 + mkq_2 \Big) \\
& + mk^2 p \cdot \log k \cdot \left(\frac{1}{k} + q_1 + mq_2 + mkp^2 q_1 q_2 \right)
\end{aligned}
\tag{8}
$$

Proof. Let T_1 and T_2 be the first and second terms in the Theorem 4 bound, respectively. Using the fact that $(1 - q_2)^k \geq 1 - kq_2$, we have

$$1 - p \left(1 - (1 - q_2)^k \right) \geq 1 - pkq_2,$$

so

$$
\begin{aligned}
\mathbb{E}[T_1] \preceq {} & m \log m \\
& \cdot \left(1 - (1-p)^k \left(1 - p \left(1 - (1-q_2)^k \right) \right)^{n-k} \right) \\
\preceq {} & m \log m \cdot \left(1 - (1-p)^k \left(1 - pkq_2 \right)^{n-k} \right) \\
\preceq {} & m \log m \cdot \left(1 - (1 - kp)(1 - (n-k) \cdot pkq_2) \right) \\
\preceq {} & m \log m \cdot \left(kp + npkq_2 \right).
\end{aligned}
$$

We can then bound $\mathbb{E}[T_2]$ by following the previous asymptotic analysis for binary splitting:

$$\mathbb{E}[T_2] \preceq mk^2 p \log k \left(\frac{1}{k} + q_1 + mq_2 + mkp^2 q_1 q_2 \right).$$

$\square$

C. Discussion

Comparing (7) and (8) term-by-term, we see that the binary splitting bound has an extra additive factor of $mk^2 pq_1 \log m(1 + mkp^2 q_2)$ compared to the community-aware bound, implying that the community-aware algorithm is never worse (order-wise) than binary splitting. Furthermore, one can verify that when $q_1 \preceq 1/k$ and $q_2 = 0$ (which includes the i.i.d. setting, where community structure has no bearing on the infection spread), the bounds are order-wise equivalent. This supports our intuition that knowledge of the community structure may not help when q_1 and q_2 are small, as the infection statuses of the individuals are "mostly independent" in this regime.

In other regimes, the community-aware algorithm is asymptotically strictly better than binary splitting. The main takeaway from the following corollary is that the community-aware algorithm can potentially improve upon binary splitting when there are several moderately sized communities in the network, and the transmission rate within each community is significant.

Corollary 3. *If* $\log m \succ \log k$, $kq_1 \succ 1$, *and* $1 \succeq mkq_2$, *then the community-aware algorithm's average complexity is asymptotically strictly better than binary splitting's by a factor of* $\min \left\{ kq_1, \frac{\log m}{\log k} \right\}$.

If $\log m \succ \log k$, $kq_1 \succ 1$, $mkq_2 \succeq 1$, *and* $mkq_2 \prec kq_1 \preceq \frac{1}{p^2}$, *then the improvement is a factor of* $\min \left\{ \frac{q_1}{m \cdot q_2}, \frac{\log m}{\log k} \right\}$.

Proof. Suppose $\log m \succ \log k$, $kq_1 \succ 1$, and $1 \succeq mkq_2$. Binary splitting's average complexity (7) becomes

$$mk^2 pq_1 \cdot \log m$$

whereas the community-aware algorithm's average complexity is

$$\max \left\{ \underbrace{mkp \cdot \log m}_{\text{(a)}}, \quad \underbrace{mk^2 pq_1 \cdot \log k}_{\text{(b)}} \right\}.$$

Both (a) and (b) are strictly smaller than the binary splitting bound. We see that (a) saves a factor of $kq_1 \succ 1$, while (b) saves a factor of $\frac{\log m}{\log k} \succ 1$. Thus, the overall improvement is a factor of $\min \left\{ kq_1, \frac{\log m}{\log k} \right\}$.

Next, suppose $\log m \succ \log k$, $kq_1 \succ 1$, $mkq_2 \succeq 1$, and $mkq_2 \prec kq_1 \preceq \frac{1}{p^2}$. Binary splitting's average complexity is $m \cdot k^2 \cdot p \cdot q_1 \cdot \log m$ (same as before), and the community-aware algorithm's complexity becomes

$$\max \left\{ m^2 k^2 pq_2 \cdot \log m, \quad mk^2 pq_1 \cdot \log k \right\},$$

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

$\square$

In general, the lower bound given in Theorem 2 exhibits a gap to the upper bounds in Theorem 3 and Theorem 4. However, we will see in the next section that this gap can be eliminated under certain assumptions.

VI. Bounds for the Disjoint k-Cliques Model

Having studied the SBIM in full generality, we now focus on the special case of $\mathsf{SBIM}(n, k, p, q, 0)$. Here, the transmission rate within a community is $q \in [0, 1]$, and no transmissions are possible between communities. This simplifying assumption allows us to obtain a tighter lower bound than in the general SBIM, and to further show that the community-aware algorithm is order-optimal in certain parameter regimes.

This setting is conceptually similar to the disjoint families model from [13]. 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 (probabilistically) 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

We obtain the following lower bounds for adaptive group testing over the the disjoint k-cliques model.

Lemma 5. *Under the disjoint k-cliques model, the number of tests T required to identify the infected individuals is lower bounded as*

$$\mathbb{E}[T] \geq H(X_1, ..., X_n)$$
$$\geq m \cdot \mathbb{E}_Z \left[(k - Z) \cdot \mathsf{h_b} \left(1 - (1 - q)^Z \right) \right],$$

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

Proof. Direct consequence of Lemma 2. □

Next is a technical lemma which characterizes the asymptotic behavior of Lemma 5 by leveraging the concentration of Z around its mean, using similar techniques as in Theorem 2. The proof is in Appendix E.

Lemma 6. *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*

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

Upon combining Lemma 5 and Lemma 6, 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

$$\mathbb{E}[T] \succeq mk^2 pq \cdot \left(\log k + \log \log \left(\frac{1}{kp} \right) \right). \quad (9)$$

Recall that $X_{\mathcal{C}_i}$ is the indicator variable of whether community $\mathcal{C}_i$ contains at least one infected member. A different lower bound is given by

$$\mathbb{E}[T] \geq H(X_1, \ldots, X_n) \overset{(a)}{\geq} H(X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_m})$$
$$= m \cdot \mathsf{h_b} \left(1 - (1 - p)^k \right) \overset{(b)}{\succeq} mkp \cdot \log_2(1/kp) \quad (10)$$

where (a) uses the fact that $X_{\mathcal{C}_1}, \ldots, X_{\mathcal{C}_m}$ are a function of $X_1, \ldots, X_n$, and (b) uses the fact that $\mathsf{h_b} \left(1 - (1 - p)^k \right) \succeq k \cdot p \cdot \log_2(1/kp)$ since $kp \preceq 1$.

In the following theorem, we summarize the refined lower bound obtained by combining (9) and (10):

Theorem 5 (Disjoint k-Cliques Lower Bound). *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*

$$\mathbb{E}[T] \succeq \max \left\{ mk^2 pq \cdot \left(\log k + \log \log \left(\frac{1}{kp} \right) \right), \right.$$
$$\left. mkp \cdot \log \left(\frac{1}{kp} \right), \quad 1 \right\}.$$

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 5 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 5 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.

Binary splitting	$m \cdot k^2 \cdot p \cdot \left(\frac{1}{k} + q\right) \cdot \log m + m \cdot k^2 \cdot p \cdot \left(\frac{1}{k} + q\right) \cdot \log k$
Community-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.

B. Algorithm Analysis

1) Binary Splitting: As a direct consequence of Theorem 3 and Corollary 1, we obtain the following non-asymptotic and asymptotic upper bounds on the expected number of tests used by binary splitting under the disjoint k-cliques model.

Corollary 4. *Under the disjoint k-cliques model, the binary splitting algorithm identifies all infected individuals using T tests, where*

$$
\mathbb{E}[T] \leq mk \cdot \Big(\log_2 m + \log_2 k + 2 \Big)
$$
$$
\cdot \Big(1 - (1-p)(1-pq)^{k-1} \Big) + 1
$$
$$
\preceq \big(mk^2 p \cdot (\log_2 m + \log_2 k) \cdot (1/k + q) \big).
$$

2) Community-Aware Algorithm: The following bounds are obtained as direct consequences of Theorem 4 and Corollary 2.

Corollary 5. *Under the disjoint k-cliques model, the community-aware algorithm identifies all infected individuals using T tests, where*

$$
\mathbb{E}[T] \leq m \cdot \Big(\log_2 m + 3 \Big)\Big(1 - (1-p)^k \Big)
$$
$$
+ n \cdot \Big(\log_2 k + 2 \Big)\Big(1 - (1-p)(1-pq)^{k-1} \Big) + 1
$$
$$
\preceq mkp \cdot \log m + mk^2 p \cdot \left(\frac{1}{k} + q \right) \cdot \log k.
$$

C. Discussion

We summarize the expected number of tests of binary splitting and the community-aware algorithm, as well as the information-theoretic lower bound, in Table I. If we compare the bounds for binary splitting and the community-aware algorithm term-by-term, we observe that the binary splitting bound has an extra additive factor of $mk^2 pq \log m$. Thus, as with the general SBIM, the community-aware algorithm is never worse (order-wise) than binary splitting.

Next, we discuss different parameter regimes where 1) the lower bound holds, 2) the community-aware algorithm is order-optimal (i.e., the lower bound is tight), and 3) the community-aware algorithm's average complexity is strictly better than binary splitting's. As stated in Theorem 5, 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 community-aware algorithm is order-optimal:

Corollary 6. *Under the disjoint k-cliques model, the community-aware algorithm is order-optimal under the following conditions:*

1) $kp \preceq m^{-\beta}$ for some fixed $\beta \in (0,1)$,
2) $\frac{1}{k} \preceq q \preceq \frac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$.

Proof. Plugging $\log \left(\frac{1}{kp}\right) \succeq \beta \log m$ 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

$$
\mathbb{E}[T] \succeq mkp \cdot \log m + mk^2 pq \cdot (\log k + \log \log m) + 1
$$
$$
\succeq mkp \cdot \log m + mk^2 pq \cdot \log k,
$$

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

$$
\mathbb{E}[T] \preceq mkp \cdot \log m + mk^2 pq \cdot \log k.
$$

$\square$

Finally, using Corollary 3 from our discussion on the general SBIM, we specify the regime where the community-aware algorithm outperforms binary splitting:

Corollary 7. *Under the disjoint k-cliques model, if $\log m \succ \log k$ and $kq \succ 1$, then the community-aware algorithm's average complexity is asymptotically strictly better than binary splitting's by a factor of $\min\left\{ kq, \frac{\log m}{\log k} \right\}$.*

Remark 2. *Recall from Corollary 3 that when $\log m \succ \log k$, $kq_1 \succ 1$, and $mkq_2 \preceq 1$ under the general SBIM, the improvement factor is also $\min\left\{ kq_1, \frac{\log m}{\log k} \right\}$, matching Corollary 7 above. Intuitively, this is because the SBIM resembles the disjoint k-cliques model when q_2 is very small.*

In Table II, we summarize the different parameter regimes discussed so far. As with the general SBIM, we find that the community-aware algorithm potentially improves 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 community-aware algorithm is order-optimal when the seeds are sparse.

Lower bound's conditions	$kp \preceq 1$ and $q \preceq \dfrac{1}{\sqrt{k \log\left(\frac{1}{kp}\right)}}$
Tightness conditions	$kp \preceq m^{-\beta}$ 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.

VII. NOISY SETTING

In this section, we develop noise-resilient analogues to our previously discussed algorithms, which we call *noisy binary splitting* (NBS) and the *noisy community-aware* (NCA) algorithm. We obtain bounds on the algorithms' average complexity under a high-probability recovery criterion and find that the NCA algorithm offers the same improvement in testing efficiency compared to NBS as observed between the corresponding algorithms in the noiseless case. We also provide the proof of our lower bound (Theorem 1) and discuss how this result implies the order-optimality of the NCA algorithm in the same parameter regimes as the noiseless community-aware algorithm. Thus, broadly speaking, the presence of testing noise does not affect the relative gains of using a community-oriented approach.

A. Noisy Binary Splitting

In [71], an algorithm called *modified noisy binary search* (MNBS)[5] was used as a sub-routine of an adaptive procedure for the symmetric noise model (2). The MNBS algorithm is said to succeed if it identifies an infected member of the population when one exists, and otherwise outputs a special symbol ϕ when no infected members remain in the population. We will employ the MNBS algorithm in a black-box manner and utilize the following performance guarantee [71, Lemma 2]:

Lemma 7 (MNBS Guarantee). *Under the symmetric noise model* (2), *given any* $\delta \in (0,1)$, *the MNBS algorithm succeeds with probability at least* $1-\delta$ *while satisfying*

$$\mathbb{E}[\# \text{ tests}] \leq \frac{\log n}{I(\rho)} + O\left(\log \frac{1}{\delta}\right) + O\left(\log\log n\right),$$

where $I(\rho) = 1 - \mathsf{h_b}(\rho) = 1 - \rho\log_2\frac{1}{\rho} - (1-\rho)\log\frac{1}{1-\rho}$ *is the capacity of the binary symmetric channel with crossover probability* $\rho \in (0, \frac{1}{2})$.

Proof. See [71, Appendix A]. $\square$

We obtain a very simple *noisy binary splitting* (NBS) algorithm for the symmetric model via repeated applications of MNBS. Note that NBS does not take

[5]This is essentially the noisy binary search algorithm from [76] adapted to the group testing framework.

into account the community structure of the population inherent to $\mathsf{SBIM}(n,k,p,q_1,q_2)$.

Noisy Binary Splitting (NBS) Algorithm
  0) Initialize $\hat{X} = \emptyset$; $\mathcal{P} = [n]$.
  1) Run MNBS with error parameter $\delta \in (0,1)$ on $\mathcal{P}$. If the result is ϕ, terminate and return $\hat{X}$. Otherwise, add the result to $\hat{X}$ and remove it from $\mathcal{P}$. Repeat Step 1.

Remark 3. *If binary search algorithms are developed for other noise models (e.g., dilution noise or Z-channel noise), then NBS (and consequently the NCA algorithm discussed later) can easily be adapted to these models by replacing the MNBS algorithm with the channel-specific algorithm. Likewise, to analyze the error probability and average number of tests, we need only replace the performance guarantees of the MNBS algorithm with those of the channel-specific algorithm.*

Recall that our objective is to ensure a vanishing probability of error, defined as $P_e \triangleq \Pr(\hat{X} \neq X)$. We obtain the following bound on the average number of tests used by NBS under this recovery guarantee. The proof closely follows that of [71, Theorem 1], the main difference being the random nature of infections in our setting, in contrast to their assumption that the number of infections is fixed.

Theorem 6 (NBS Bound). *Suppose the infections in the population are distributed according to the* $\mathsf{SBIM}(n,k,p,q_1,q_2)$, *and let*

$$\mu \triangleq n \cdot \left(1 - (1-p)(1-pq_1)^{k-1}(1-pq_2)^{n-k}\right)$$

denote the expected number of infected individuals. Under the symmetric noise model (2), *for any* $\delta \in (0,1)$ *such that* $\delta \prec \frac{1}{\mu}$, *the NBS algorithm achieves* $P_e \leq \delta\mu \to 0$ *and uses an expected number of tests satisfying*

$$\mathbb{E}[T] \preceq \frac{\mu\log n}{I(\rho)} + \mu\log\frac{1}{\delta}. \qquad (11)$$

Proof. We first analyze the probability of error. Note that, as long as the MNBS algorithm always succeeds in Step 1, the NBS algorithm will produce the correct output (i.e., achieve $\hat{X} = X$) and do so using $D+1$ calls to the MNBS algorithm, where D is the total number

of infected individuals in the population. (The $+1$ is due to the final call to the algorithm when no infected members remain.) Let $\mathcal{E}_i$ denote the event that the ith call to the MNBS algorithm fails. Then, by conditioning on the number of infections and performing a union bound over the error events, we have

$$
\begin{aligned}
P_e &\leq \Pr\left(\bigcup_{i=1}^{D}\mathcal{E}_i\right) \\
&= \sum_{d=0}^{n}\Pr(D=d)\cdot\Pr\left(\bigcup_{i=1}^{D}\mathcal{E}_i \,\Big|\, D=d\right) \\
&\leq \sum_{d=0}^{n}\Pr(D=d)\cdot d\delta \\
&= \delta\mu \to 0 \quad\text{(by assumption).}
\end{aligned}
$$

Next, we bound the average number of tests used by the algorithm. By Lemma 7, the average number of tests performed during the first $D+1$ calls to the MNBS algorithm is

$$
\begin{aligned}
\left(\mu+1\right)&\left(\frac{\log n}{I(\rho)}+O\left(\log\frac{1}{\delta}\right)+O(\log\log n)\right) \\
&\asymp \frac{\mu\log n}{I(\rho)}+\mu\log\frac{1}{\delta}. \quad (12)
\end{aligned}
$$

Note that the algorithm may make more than $D+1$ calls to MNBS if at least one of the calls fails. However, these additional tests do not affect the overall scaling in (12) for the following reasons. Since the tests are subject to independent noise, the calls to MNBS fail independently with probability $O(\delta)$. Hence, the number of failures encountered before the first success is distributed as a Geometric(θ) random variable where $\theta = 1 - O(\delta)$. The mean of this variable is $\frac{1-\theta}{\theta} = O(\delta)$, and it follows that $\mathbb{E}[T]$ is within a multiplicative $1+O(\delta)$ factor of (12). $\qquad\square$

Remark 4. *Note that the bound in Corollary 1 (the average number of tests used by binary splitting in the noiseless case) can be expressed as $\mu\log n$. The bound in Theorem 6 scales this inversely by $I(\rho)$ (the capacity of the binary symmetric channel) and includes an additive factor, $\mu\log\frac{1}{\delta}$, that depends on the desired error probability.*

We further remark that Theorem 6 applies to any underlying probabilistic infection model, not just the SBIM. One just needs to replace μ in the theorem statement with the corresponding model-specific quantities.

B. Noisy Community-Aware Algorithm

Following the same template as the noiseless community-aware algorithm, our *noisy community-aware* (NCA) algorithm applies NBS in two stages:

first to identify communities with at least one infected member, and then to identify the infected individuals within each positive community from the first stage.

Noisy Community-Aware (NCA) Algorithm

1) Mix the samples within each of the communities and run the NBS algorithm with error parameter $\delta' \in (0,1)$ on the mixed samples.
2) Run the NBS algorithm with error parameter $\delta \in (0,1)$ within each community identified as positive in Step 1.

We now state and prove the performance guarantee of the NCA algorithm.

Theorem 7 (NCA Bound)**.** *Suppose the infections in the population are distributed according to* SBIM(n,k,p,q_1,q_2)*, and let*

$$
\mu \triangleq n\cdot\left(1-(1-p)(1-pq_1)^{k-1}(1-pq_2)^{n-k}\right)
$$

$$
\mu_M \triangleq m\cdot\left(1-(1-p)^k\left(1-p\cdot(1-(1-q_2)^k)\right)^{n-k}\right)
$$

denote the expected number of infected individuals and expected number of infected communities, respectively. Under the symmetric noise model (2), for any $\delta,\delta' \in (0,1)$ such that $\delta' \prec \frac{1}{\mu_M}$ and $\delta \prec \frac{1}{\mu}$, the NCA algorithm achieves $P_e \leq \delta'\mu_M + \delta\mu \to 0$ and uses an expected number of tests satisfying

$$
\mathbb{E}[T] \preceq \frac{\mu_M\log m + \mu\log k}{I(\rho)} + \mu\log\frac{1}{\delta}. \quad (13)
$$

Proof. Provided that each step of the algorithm succeeds, the algorithm's final output will be correct. Moreover, the NBS algorithm will be executed once in Step 1 and M times in Step 2, where M is the number of infected communities. Let $\mathcal{E}^{(1)}$ denote the event that the single call to NBS in Step 1 fails, and let $\mathcal{E}^{(2)}_i$ denote the event that the ith call to NBS in Step 2 fails. Additionally, let μ_K denote the expected number of infected members in a single community conditioned on the event that the community contains at least one infected member. We have

$$
P_e \leq \Pr\left(\mathcal{E}^{(1)}\cup\bigcup_{i=1}^{M}\mathcal{E}^{(2)}_i\right) \quad (14)
$$

$$
\leq \Pr(\mathcal{E}^{(1)}) + \sum_{j=0}^{m}\Pr(M=j)\cdot\Pr\left(\bigcup_{i=1}^{M}\mathcal{E}^{(2)}_i \,\Big|\, M=j\right) \quad (15)
$$

$$
\leq \delta'\cdot\mu_M + \sum_{j=0}^{m}\Pr(M=j)\cdot j\cdot\delta\cdot\mu_K \quad (16)
$$

$$
= \delta'\cdot\mu_M + \delta\cdot\mu_K\cdot\mu_M \quad (17)
$$

$$
= \delta'\cdot\mu_M + \delta\cdot\mu \to 0 \quad (18)
$$

where (15) applies a union bound and conditions on the number of infected communities, (16) applies another union bound along with the NBS error guarantee from Theorem 6, and (18) uses the identity $\mu = \mu_K \cdot \mu_M$.

To obtain a bound on the average number of tests, we will apply Theorem 6 twice and sum the results. First, we apply it with μ replaced by μ_M, n replaced by m, and δ replaced by δ' (corresponding to Step 1). Second, we apply it with μ replaced by μ_K and n replaced by k (corresponding to Step 2). The overall average number of tests performed during a single call to the NBS algorithm in Step 1 and M calls to the algorithm in Step 2 is thus given by

$$
\begin{aligned}
\mathbb{E}[T] &\preceq \frac{\mu_M \log m}{I(\rho)} + \mu_M \log \frac{1}{\delta'} \\
&\quad + \mu_M \left(\frac{\mu_K \log k}{I(\rho)} + \mu_K \log \frac{1}{\delta} \right) \\
&= \frac{\mu_M \log m}{I(\rho)} + \mu_M \log \frac{1}{\delta'} + \frac{\mu \log k}{I(\rho)} + \mu \log \frac{1}{\delta} \\
&\preceq \frac{\mu_M \log m + \mu \log k}{I(\rho)} + \mu \log \frac{1}{\delta}
\end{aligned}
$$

where the second line uses the identity $\mu = \mu_K \cdot \mu_M$ and the final line uses the fact that $\mu_M \log \frac{1}{\delta'} \preceq \mu \log \frac{1}{\delta}$. Again, we note that additional tests may be required if one or more of the calls to the NBS algorithm fails, but that the above scaling will remain intact. (See the argument in the proof of Theorem 6.) □

Remark 5. *The bound in Corollary 2 (the community-aware algorithm in the noiseless case) can be expressed as $\mu_M \log m + \mu \log k$. Hence, as similarly discussed in Remark 4, we find that Theorem 7 scales the noiseless bound inversely by $I(\rho)$ and includes an additive factor that depends on the desired error probability δ.*

It follows that the NCA algorithm is always at least as efficient (order-wise) as NBS (recall our discussion in Section V-C). Moreover, if $\delta \succeq k^{-\frac{1}{I(\rho)}}$ (which ensures that the $\mu \log \frac{1}{\delta}$ term does not dominate), then the NCA algorithm is strictly better than NBS under the same conditions we derived in Corollary 3. That is, the parameter regimes in which the community-aware algorithm improves upon binary splitting in the noiseless case are the same regimes in which the NCA algorithm improves upon NBS, provided that the desired error probability is not too stringent.

Lastly, we emphasize the generality of Theorem 7 beyond the SBIM. The result holds under any probabilistic infection model satisfying the following symmetry condition, which states that the expected number of infections in an infected community should be the same across all communities: $\forall i, j \in [m]$,

$$
\mathbb{E}\Big[\sum_{\ell \in \mathcal{C}_i} X_\ell \,|\, X_{\mathcal{C}_i} = 1 \Big] = \mathbb{E}\Big[\sum_{\ell \in \mathcal{C}_j} X_\ell \,|\, X_{\mathcal{C}_j} = 1 \Big].
$$

Also, note that our result holds even when there are no "communities" in the population. In this case, we simply set $m = 1$, $k = n$ in (13) and find that it reduces to the community-oblivious bound in Theorem 6.

C. Proof of the Lower Bound (Theorem 1)

Before delving into the proof of Theorem 1, we make a few remarks. Provided that $\delta \succeq k^{-1/I(\rho)}$, note that the NBS upper bound (Theorem 6) and the NCA upper bound (Theorem 7) each scale their noiseless counterpart (Theorem 3 and Theorem 4, respectively) by $I(\rho)$. Moreover, our lower bound in Theorem 1 also scales its noiseless counterpart (3) by $I(\rho)$. Taken together, these results imply that the NCA algorithm is order-optimal under the same conditions as the noiseless community-aware algorithm (recall Corollary 6), as long as δ is not too small. Thus, the benefits of using a community-aware scheme persist in the presence of symmetric testing noise.

To prove our main result, we will leverage a slightly more general version of Fano's inequality given in the following lemma.

Lemma 8. *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr(X \neq \hat{X})$, we have*

$$
H(X \,|\, Y) \leq 1 + P_e \cdot H(X).
$$

Proof. The proof is a small modification to the proof of Fano's inequality [72, Theorem 2.10.1]. First, we define the error random variable

$$
E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X. \end{cases}
$$

Then, we expand $H(X, E \,|\, \hat{X})$ using two applications of the chain rule for entropy:

$$
\begin{aligned}
H(X, E \,|\, \hat{X}) &= H(X \,|\, \hat{X}) + \underbrace{H(E \,|\, X, \hat{X})}_{=0} \\
&= H(E \,|\, \hat{X}) + H(X \,|\, E, \hat{X}).
\end{aligned}
$$

Since conditioning reduces entropy, we have $H(E \,|\, \hat{X}) \leq H(E) = \mathsf{h_b}(P_e) \leq 1$. Therefore, we currently have

$$
H(X \,|\, \hat{X}) \leq 1 + H(X \,|\, E, \hat{X})
$$

and we can bound $H(X \mid E, \hat{X})$ as

$$
H(X \mid E, \hat{X}) = \Pr(E = 0) \cdot \underbrace{H(X \mid \hat{X}, E = 0)}_{=0}
$$
$$
+ \Pr(E = 1) \cdot H(X \mid \hat{X}, E = 1)
$$
$$
\leq P_e \cdot H(X)
$$

where the inequality follows from the fact that conditioning reduces entropy. Finally, the data processing inequality yields

$$
H(X \mid Y) \leq H(X \mid \hat{X}) \leq 1 + P_e \cdot H(X)
$$

as desired. $\qquad\square$

Now we prove our main result, which we restate below for convenience.

Theorem 1. *Assume $H(X_1, X_2, \ldots, X_n) \to \infty$ as $n \to \infty$. Under the symmetric noise model (2), any adaptive algorithm achieving $P_e \to 0$ must use an average number of tests lower bounded as*

$$
\mathbb{E}[T] \geq \frac{H(X_1, \ldots, X_n)}{I(\rho)}. \tag{19}
$$

Proof. For $i = 1, 2, \ldots, M$, where M is the maximum number of tests allowed by the algorithm, let $\mathcal{S}_i \subseteq [n]$ denote the subset of individuals included in the ith test. Further, let

$$
Z_i = \bigvee_{j \in \mathcal{S}_i} X_j
$$

denote the *noiseless* outcome of the ith test, and let

$$
Y_i = Z_i \oplus \xi, \quad \xi \sim \mathsf{Bernoulli}(\rho), \, \rho \in \left(0, \frac{1}{2}\right)
$$

denote the corresponding noisy outcome. An adaptive algorithm applies a sequence of tests $\mathcal{S}_1, \mathcal{S}_2, \ldots$ and observes their outcomes $Y_1, Y_2, \ldots$, where $\mathcal{S}_i$ can be chosen as a function of the previous test outcomes $Y_1, \ldots, Y_{i-1}$. The algorithm terminates at a random time T which is determined by the test outcomes up to the Tth test, i.e., $Y_1, \ldots, Y_T$. For $i > T$, we will write $\mathcal{S}_i = \emptyset$, $Z_i = \emptyset$, and $Y_i = \emptyset$ to indicate that the algorithm has been terminated and no further tests will be performed. We will also use the shorthand notation $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_M)$.

We proceed by expanding the mutual information between the infection statuses $X_1, \ldots, X_n$ and the noisy test outcomes $Y_1, \ldots, Y_M$ as follows:

$$
I(X; Y) = H(Y) - H(Y \mid X)
$$
$$
= \sum_{i=1}^{M} H(Y_i \mid Y_1, \ldots, Y_{i-1})
$$
$$
- \sum_{i=1}^{M} H(Y_i \mid X, Y_1, \ldots, Y_{i-1}). \tag{20}
$$

For now, let us assume that the algorithm is deterministic and that the first test, $\mathcal{S}_1$, is fixed. Note that $\mathcal{S}_i$ can be recursively deduced from $Y_1, \ldots, Y_{i-1}$, and hence $\mathcal{S}_i$ is a function of $Y_1, \ldots, Y_{i-1}$. Combining this with the fact that conditioning reduces entropy, we obtain

$$
H(Y_i \mid Y_1, \ldots, Y_{i-1}) = H(Y_i \mid \mathcal{S}_i, Y_1, \ldots, Y_{i-1})
$$
$$
\leq H(Y_i \mid \mathcal{S}_i). \tag{21}
$$

Next, we have

$$
H(Y_i \mid X, Y_1, \ldots, Y_{i-1}) = H(Y_i \mid X, Y_1, \ldots, Y_{i-1}, Z_i) \tag{22}
$$
$$
= H(Y_i \mid Z_i) \tag{23}
$$

where (22) uses the fact that Z_i is a function of $\{X, \mathcal{S}_i\}$ and hence is a function of $\{X, Y_1 \ldots, Y_{i-1}\}$; and (23) uses the fact that Y_i is conditionally independent of X and $Y_1, \ldots, Y_{i-1}$ given Z_i.

Plugging (21) and (22) into (20) yields

$$
I(X; Y) \leq \sum_{i=1}^{M} \Big[H(Y_i \mid \mathcal{S}_i) - H(Y_i \mid Z_i) \Big]
$$
$$
= \sum_{i=1}^{M} \Pr(\mathcal{S}_i \neq \emptyset) \cdot \Big[H(Y_i \mid \mathcal{S}_i, \mathcal{S}_i \neq \emptyset)
$$
$$
- H(Y_i \mid Z_i, \mathcal{S}_i \neq \emptyset) \Big]
$$
$$
\leq \sum_{i=1}^{M} \Pr(\mathcal{S}_i \neq \emptyset) \cdot (1 - \mathsf{h_b}(\rho)) \tag{24}
$$
$$
= \sum_{i=1}^{M} \Pr(T \geq i) \cdot (1 - \mathsf{h_b}(\rho))
$$
$$
= \mathbb{E}[T] \cdot (1 - \mathsf{h_b}(\rho)) \tag{25}
$$
$$
= \mathbb{E}[T] \cdot I(\rho) \tag{26}
$$

where (24) uses the fact that $H(Y_i \mid \mathcal{S}_i, \mathcal{S}_i \neq \emptyset) \leq 1$ since Y_i is a binary random variable conditioned on $\mathcal{S}_i \neq \emptyset$, and (25) follows from the tail-sum formula for expectation.

Now, combining (26) and our modified Fano's inequality (Lemma 8) with the fact that $H(X) = I(X; Y) + H(X \mid Y)$ gives us

$$
H(X) \leq \mathbb{E}[T] \cdot I(\rho) + P_e \cdot H(X) + 1
$$

which can be rearranged to

$$
P_e \geq 1 - \frac{\mathbb{E}[T] \cdot I(\rho) + 1}{H(X)}.
$$

Finally, let $\varepsilon > 0$, and observe that if $\mathbb{E}[T] \leq \frac{H(X)}{I(\rho)} \cdot (1 - \varepsilon)$, then $P_e \geq \varepsilon - \frac{1}{H(X)} \to \varepsilon$ by the assumption that $H(X) \to \infty$ as $n \to \infty$.

We can easily extend our analysis to allow for randomized tests by assuming $\mathcal{S}_i$ depends on some external randomness θ_i, independent of the noise and

the infection statuses. Thus, $\mathcal{S}_i$ is a function of $\{Y_1, \ldots, Y_{i-1}, \theta_1, \ldots, \theta_{i-1}\}$. Then, using the shorthand notation $\theta = (\theta_1, \ldots, \theta_M)$, we have

$$
\begin{aligned}
H(X) = H(X \mid \theta) &= I(X; Y \mid \theta) + H(X \mid Y, \theta) \\
&= H(Y \mid \theta) - H(Y \mid X, \theta) + H(X \mid Y, \theta).
\end{aligned}
$$

Note that

$$
H(X \mid Y, \theta) \le H(X \mid Y) \le 1 + P_e \cdot H(X).
$$

Additionally, using the fact that $\mathcal{S}_i$ is a function of $\{Y_1, \ldots, Y_{i-1}, \theta_1, \ldots, \theta_i\}$, it follows that

$$
\begin{aligned}
H(Y \mid \theta) &= \sum_{i=1}^{M} H(Y_i \mid Y_1, \ldots, Y_{i-1}, \theta) \\
&= \sum_{i=1}^{M} H(Y_i \mid Y_1, \ldots, Y_{i-1}, \theta, \mathcal{S}_i) \\
&\le \sum_{i=1}^{M} H(Y_i \mid \mathcal{S}_i).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
H(Y \mid X, \theta) &= \sum_{i=1}^{M} H(Y_i \mid X, Y_1, \ldots Y_{i-1}, \theta) \\
&= \sum_{i=1}^{M} H(Y_i \mid X, Y_1, \ldots, Y_{i-1}, \theta, Z_i) \\
&= \sum_{i=1}^{M} H(Y_i \mid Z_i)
\end{aligned}
$$

and the remainder of the proof is exactly the same as in the deterministic case. □

VIII. NUMERICAL EXPERIMENTS

We implemented the binary splitting and community-aware algorithms and evaluated their performance over random instances of the SBIM. The population size was set to $n = 1,000$, and p was varied over the interval $[0, 0.01]$, while q_1, q_2 were fixed at different values. We ran 500 trials for each value of p, where a trial consists of generating an instance from $\mathrm{SBIM}(n, k, p, q_1, q_2)$ and observing the number of tests used by binary splitting and the community-aware algorithm to identify the infected individuals. For the lower bound, we computed

the non-asymptotic expression given in (5). To do so, one can expand (5) as follows:

$$
\begin{aligned}
&n \cdot I(X_1; S_1) + H(X \mid S) \\
&= n \cdot \left(H(X_1) - H(X_1 \mid S_1) \right) + H(X \mid S) \\
&= n \cdot \left(H(X_1) - \mathbb{P}(S_1 = 0) \cdot H(X_1 \mid S_1 = 0) \right) \\
&\qquad + H(X \mid S) \\
&= n \cdot \left(\mathsf{h_b}\Big(\mathbb{P}(X_1 = 1) \Big) \right. \\
&\qquad\qquad \left. - (1 - p) \cdot \mathsf{h_b}\Big(\mathbb{P}(X_1 = 1 \mid S_1 = 0) \Big) \right) \\
&\qquad + H(X \mid S)
\end{aligned}
$$

where $X = (X_1, \ldots, X_n)$, $S = (S_1, \ldots, S_n)$. The terms $\mathbb{P}(X_1 = 1)$ and $\mathbb{P}(X_1 = 1 \mid S_1 = 0)$ are straightforward to compute (see Lemma 3 and its proof in Appendix C). The term $H(X \mid S)$ is lower-bounded by (6). To estimate (6), we took an average 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 $k = 50$ 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 3 and Theorem 4, and we find that these bounds closely match the empirical results. Additionally, the community-aware algorithm consistently outperforms binary splitting. For example, in Figure 2c, at $p \approx 0.009$, binary splitting has already exceeded the individual testing threshold with an average of 1,044 tests, whereas the community-aware algorithm uses an average of 708 tests; this represents a 32% reduction in testing. The community-aware algorithm's performance also seems to exhibit lower variance than binary splitting. In Figure 3, we fix $q_1 = 0.1$, $q_2 = 0.001$, and vary the community size $k \in \{10, 50, 100, 200\}$. The community-aware algorithm appears to perform most favorably (relative to binary splitting) for more intermediate values of k, i.e., when there are several moderately sized communities in the network. These findings are consistent with our earlier theoretical results.

The estimated lower bound is fairly close to the community-aware algorithm's bound in the regime where the seeds are very sparse (small p) and the network exhibits strong community structure (intermediate k, q_1; small q_2). This corroborates our analysis from Section VI-C. In other regimes—such as when p, k, or q_2 are larger—there is a sizable gap between the
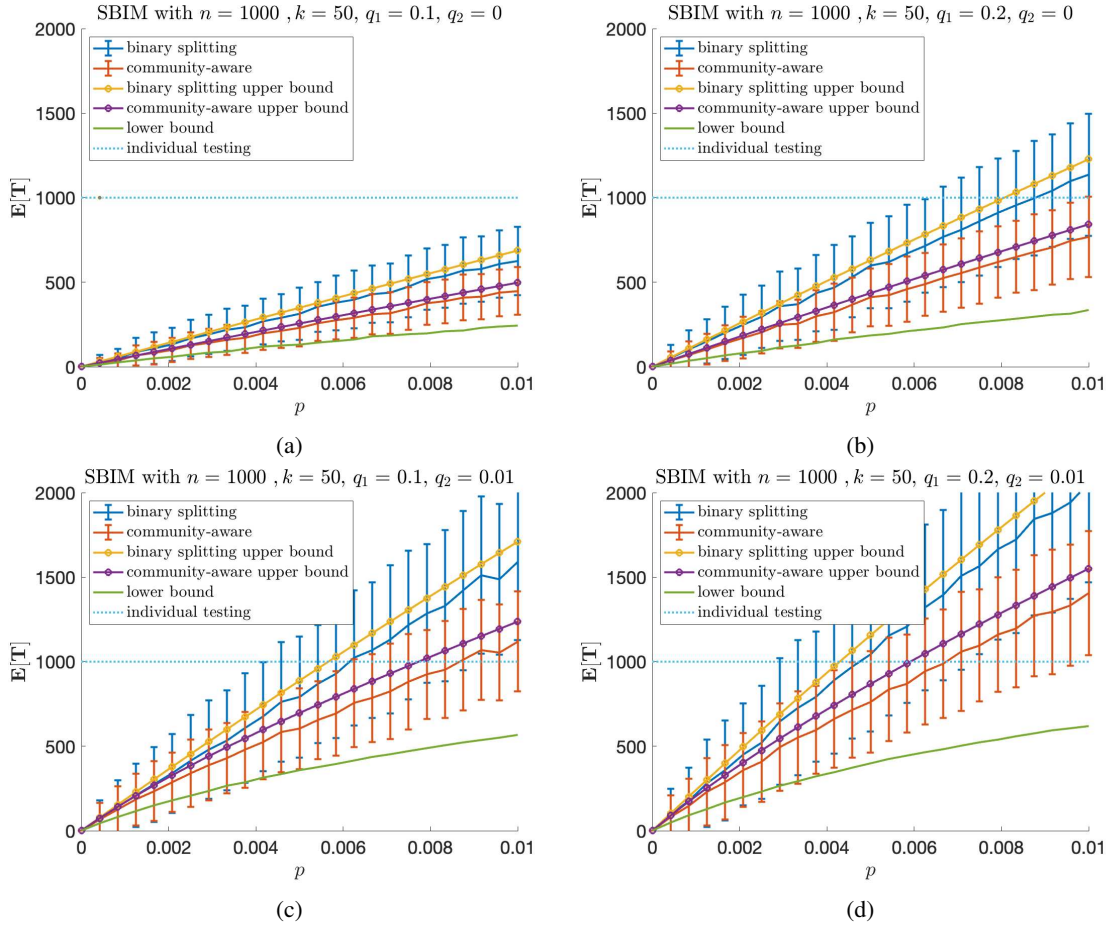
Fig. 2: Performance comparison between binary splitting and the community-aware algorithm under the SBIM with $n = 1000$, $k = 50$, and different values of p, q_1, q_2. Theoretical upper and lower bounds are also shown.

community-aware bound and the lower bound. However, this gap seems to be at most a constant factor in many cases, suggesting that the order-optimality of the community-aware algorithm still holds in broader regimes. Nevertheless, these results suggest the potential to further improve the non-asymptotic upper or lower bounds.

IX. CONCLUSION

In this paper, we investigated the group testing problem over networks with community structure. Motivated by infectious diseases such as COVID-19, we proposed a network-based infection model which generalizes the traditional i.i.d. group testing model to settings in which interactions between individuals dictate the disease spread. Our proposed adaptive algorithm, which exploits the known community structure of the underlying graph, provably outperforms the community-oblivious binary splitting algorithm and is order-optimal in certain parameter regimes, as implied by our novel

lower bounds based on the system entropy. Even in the presence of symmetric noise, our community-oriented approach offers the same gains in testing efficiency as it does in the noiseless case.

We conclude with some future directions. As discussed in Section V and further suggested by our simulations, there remains a gap between our upper and lower bounds in certain regimes of the general SBIM, due to the difficulty of bounding $H\left(X_{\mathcal{C}_1}, ..., X_{\mathcal{C}_m}\right)$ when $\{X_{\mathcal{C}_1}, ..., X_{\mathcal{C}_m}\}$ are not mutually independent. Other directions of interest include designing non-adaptive group testing schemes for our setting, deriving bounds under other noise models such as dilution and Z-channel models, and extending our infection model to longer time horizons (e.g., SIR or SIS-type infection models from the epidemiology literature). Finally, characterizing the complexity of group testing under the general graph-based infection model described in Section III-C (beyond the SBIM studied in this paper) is fertile ground for future work.
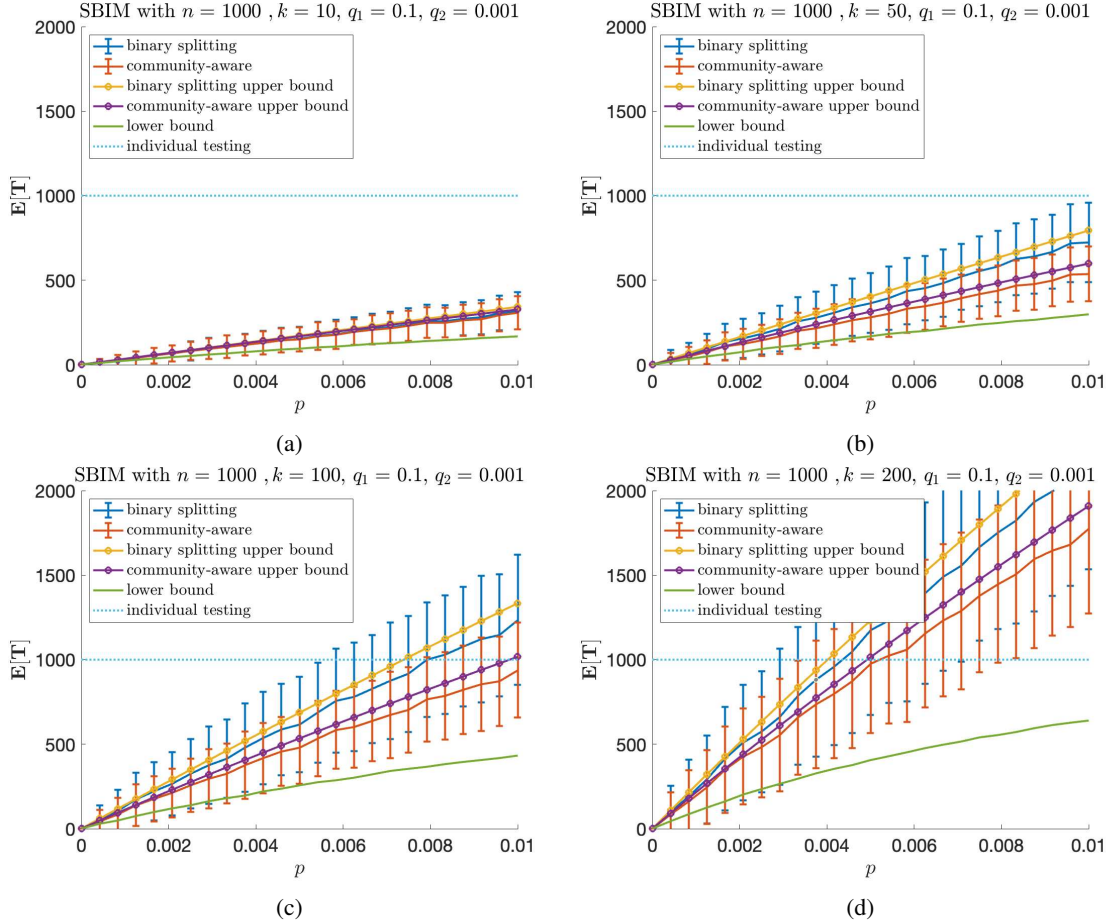
Fig. 3: Performance comparison between binary splitting and the community-aware algorithm under the SBIM with $n = 1000$, $q_1 = 0.1$, $q_2 = 0.001$, and different values of p, k. Theoretical upper and lower bounds are also shown.

APPENDIX A
PROOF OF LEMMA 2

Let S_i be the indicator variable of whether the ith individual is a seed. We will use the shorthand notation $X = (X_1, \ldots, X_n)$, $S = (S_1, \ldots, S_n)$. By standard information-theoretic arguments, we have

$$
\begin{aligned}
H(X) &= I(X;S) + H(X \mid S) \\
&= H(S) - H(S \mid X) + H(X \mid S) \\
&= n \cdot H(S_1) - \sum_{i=1}^{n} H(S_i \mid X, S_1, \ldots, S_{i-1}) \\
&\qquad + H(X \mid S) \\
&\geq n \cdot H(S_1) - n \cdot H(S_1 \mid X_1) + H(X \mid S) \\
&= n \cdot I(X_1; S_1) + H(X \mid S).
\end{aligned}
$$

This proves (5) in Lemma 2.

Next, we prove (6). First, note that $H(X) \geq H(X \mid S)$ since $I(X;S) \geq 0$. (This also follows directly from the fact that conditioning reduces entropy.) We have

$$
H(X \mid S) = \sum_{s \in \{0,1\}^n} \mathbb{P}(S = s) \cdot H(X \mid S = s).
$$

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

$$
\mathbb{P}(X_i = 1 \mid S = s)
$$
$$
= \begin{cases}
1, & s_i = 1, \\
1 - (1-q_1)^{\sum_{j \in \mathcal{C}_\ell} s_j} (1-q_2)^{\sum_{j \notin \mathcal{C}_\ell} s_j}, & s_i = 0.
\end{cases}
$$

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

$$
H(X \mid S = s) = \sum_{\ell=1}^{m} (k - z_\ell)
$$
$$
\cdot \mathsf{h_b} \left(1 - (1-q_1)^{z_\ell} (1-q_2)^{\sum_{\ell' \neq \ell} z_{\ell'}} \right),
$$

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

$$
H(X \mid S) = \mathbb{E}_{Z,Z'}[m \cdot (k - Z) \\
\cdot h_b(1 - (1 - q_1)^Z (1 - q_2)^{Z'})], \tag{27}
$$

where $Z \sim$ Binom(k, p) and $Z' \sim$ Binom$(n - k, p)$. $\qquad\square$

APPENDIX B
PROOF OF THEOREM 2

In light of Lemma 2, it suffices to show that the following lower bound holds under the assumptions in the theorem statement:

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

where $Z \sim$ Binom(k, p) and $Z' \sim$ Binom$(n - k, p)$. Our proof leverages the concentration of Z and Z' around their means.

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

$$
\frac{1}{2} \geq Z \cdot q_1 + Z' \cdot q_2 \geq 1 - (1 - q_1)^Z (1 - q_2)^{Z'}. \tag{28}
$$

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

$$
\mathbb{E}_{Z,Z'}\left[(k - Z)h_b\left(1 - (1 - q_1)^Z (1 - q_2)^{Z'}\right)\right] \\
\geq \mathbb{E}_{Z,Z'}\left[(k - Z)h_b\left(1 - (1 - q_1)^Z (1 - q_2)^{Z'}\right) \\
\cdot \mathbb{1}_{\{0 \leq Z \leq z^*\}} \cdot \mathbb{1}_{\{np(1-\epsilon) \leq Z' \leq np(1+\epsilon)\}}\right] \\
\geq \mathbb{E}_{Z,Z'}\left[(k - Z)h_b\left(1 - (1 - q_2)^{Z'}\right) \cdot \mathbb{1}_{\{Z = 0\}} \\
\cdot \mathbb{1}_{\{np(1-\epsilon) \leq Z' \leq np(1+\epsilon)\}}\right] \tag{29} \\
+ \mathbb{E}_{Z,Z'}\left[(k - Z)h_b\left(1 - (1 - q_1)^Z (1 - q_2)^{Z'}\right) \\
\cdot \mathbb{1}_{\{1 \leq Z \leq z^*\}} \cdot \mathbb{1}_{\{np(1-\epsilon) \leq Z' \leq np(1+\epsilon)\}}\right]. \tag{30}
$$

We will pick $\epsilon = \frac{1}{2}$. Then (29) can be bounded above by

$$
k \cdot h_b\left(q_2 \cdot np(1 - \epsilon) - (q_2 \cdot np(1 - \epsilon))^2\right) \\
\cdot \left(1 - 2 \cdot \exp\left(-\frac{n\epsilon^2 p}{3}\right)\right) \\
\succeq k \cdot \left(npq_2(1 - \epsilon) \log\left(\frac{1}{npq_2(1 - \epsilon)}\right) \\
\cdot \left(1 - 2 \cdot \exp\left(-\frac{n\epsilon^2 p}{3}\right)\right)\right) \\
\succeq k \cdot \left(npq_2 \log\left(\frac{1}{npq_2}\right)\right)
$$

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, (30) can be bounded above by

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

We will now lower bound $\mathbb{E}_Z\left[(k - Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right]$ as in Theorem 5. Observe that

$$
\mathbb{E}_Z\left[(k - Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right] \\
\geq \mathbb{E}_Z[k - Z] - k \cdot \mathbb{P}\{Z = 0\} - k \cdot \mathbb{P}\{Z \geq z^*\} \\
\geq k\left(1 - p - (1 - p)^k - \mathbb{P}\{Z \geq z^*\}\right) \\
\succeq k\left(kp - \mathbb{P}\{Z \geq z^*\}\right). \tag{31}
$$

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

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

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 (31) yields

$$\mathbb{E}_Z\left[(k - Z)\mathbb{1}_{\{1 \leq Z \leq z^*\}}\right] \succeq k^2 p, \tag{32}$$

and thus by putting together our bounds on (29), (30), we arrive at

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

$\square$

APPENDIX C
PROOF OF LEMMA 3

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

$$\mathbb{P}(X_v = 1) = \underbrace{\mathbb{P}(X_v = 1 \mid S_v = 1)}_{=1} \cdot \underbrace{\mathbb{P}(S_v = 1)}_{=p}$$
$$\quad + \mathbb{P}(X_v = 1 \mid S_v = 0)) \cdot \mathbb{P}(S_v = 0)$$
$$= p + (1 - p) \cdot \mathbb{P}(X_v = 1 \mid S_v = 0).$$

Given that v is not a seed, $X_v = 1$ if and only if v is infected by another seed. Hence,

$$\mathbb{P}(X_v = 1 \mid S_v = 0)$$
$$= \mathbb{P}(\{v \text{ is infected by another individual}\})$$
$$= 1 - \prod_{u \in [n]} \mathbb{P}(\{v \text{ isn't infected by } u\})$$
$$= 1 - \prod_{u \in [n]} \left(1 - \mathbb{P}(\{v \text{ is infected by } u\})\right)$$
$$= 1 - \prod_{u \in [n]} \left(1 - \mathbb{P}(\{v \text{ is infected by } u\} \mid S_u = 1)\right.$$
$$\left. \cdot \mathbb{P}(S_u = 1)\right)$$
$$= 1 - \left(\prod_{u \in \mathcal{C}_1 \setminus \{v\}} (1 - pq_1)\right) \cdot \left(\prod_{w \notin \mathcal{C}_1} (1 - pq_2)\right)$$
$$= 1 - (1 - pq_1)^{k-1} (1 - pq_2)^{n-k}.$$

$\square$

APPENDIX D
PROOF OF LEMMA 4

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

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

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

$$\mathbb{P}(\mathcal{B}) = 1 - \prod_{u \notin \mathcal{C}_1} \mathbb{P}(\mathcal{B}_u^c)$$
$$= 1 - \prod_{u \notin \mathcal{C}_1} \left(\mathbb{P}(\mathcal{B}_u^c \mid S_u = 1) \cdot \underbrace{\mathbb{P}(S_u = 1)}_{=p}\right.$$
$$\left. + \underbrace{\mathbb{P}(\mathcal{B}_u^c \mid S_u = 0)}_{=1} \cdot \underbrace{\mathbb{P}(S_u = 0)}_{=1-p}\right)$$
$$= 1 - \prod_{u \notin \mathcal{C}_1} \left(1 - p + p \cdot \mathbb{P}(\mathcal{B}_u^c \mid S_u = 1)\right)$$
$$= 1 - \prod_{u \notin \mathcal{C}_1} \left(1 - p + p \cdot (1 - q_2)^k\right)$$
$$= 1 - \left(1 - p \cdot \left(1 - (1 - q_2)^k\right)\right)^{n-k}.$$

$\square$

APPENDIX E
PROOF OF LEMMA 6

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

$$\mathbb{E}_Z[(k - Z) \cdot \mathsf{h_b}(1 - (1 - q)^Z)]$$
$$\geq \mathbb{E}_Z[(k - Z) \cdot \mathsf{h_b}(1 - (1 - q)^Z) \cdot \mathbb{1}_{\{1 \leq Z \leq f(q)\}}]$$
$$\overset{(a)}{\geq} \mathsf{h_b}(q) \cdot \mathbb{E}_Z[(k - Z) \cdot \mathbb{1}_{\{1 \leq Z \leq f(q)\}}]$$
$$\geq \mathsf{h_b}(q)\left(\mathbb{E}_Z[k - Z] - k \cdot \mathbb{P}\{Z = 0\}\right.$$
$$\left. - k \cdot \mathbb{P}\{Z > f(q)\}\right)$$
$$= k \cdot \mathsf{h_b}(q)\left((1 - p)(1 - (1 - p)^{k-1}) - \mathbb{P}\{Z > f(q)\}\right)$$
$$\overset{(b)}{\geq} k \cdot \mathsf{h_b}(q)\left((1 - p)\left((k - 1)p - (k - 1)^2 p^2\right)\right.$$
$$\left. - \mathbb{P}\{Z > f(q)\}\right)$$
$$\overset{(c)}{\succeq} \frac{k}{2} \cdot \mathsf{h_b}(q)\left(k \cdot p - \mathbb{P}\{Z > f(q)\}\right), \tag{33}$$

where (a) is due to the fact that $h_b(x) \geq h_b(q)$ 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:

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

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 (33) yields

$$
\begin{aligned}
\mathbb{E}_Z\left[(k - Z)\cdot h_b\left(1 - (1 - q)^Z\right)\right] &\succeq k^2 \cdot p \cdot q \cdot \log\left(\frac{1}{q}\right) \\
&\succeq k^2 \cdot p \cdot q \cdot \left(\log k + \log\log\left(\frac{1}{kp}\right)\right),
\end{aligned}
$$

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.

$\square$

APPENDIX F
PROOF OF THEOREM 4

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

Let N be the number of infected communities. By Lemma 4, we have

$$
\begin{aligned}
\mathbb{E}[N] &= \frac{n}{k} \cdot \mathbb{P}(X_{\mathcal{C}_1} = 1) \\
&= \frac{n}{k} \cdot \left(1 - (1-p)^k\left(1 - p \cdot \left(1 - (1 - q_2)^k\right)\right)^{n-k}\right).
\end{aligned}
$$

Taking Lemma 1 with $n = n/k$ and $\alpha = N$ gives

$$
T_1 \leq (\log_2(n/k) + 2) \cdot N + 1.
$$

so that

$$
\begin{aligned}
\mathbb{E}[T_1] &\leq \frac{n}{k} \cdot \left(\log_2(n/k) + 1\right) \\
&\quad \cdot \left(1 - (1-p)^k\left(1 - p\left(1 - (1 - q_2)^k\right)\right)^{n-k}\right) \\
&\quad + 1.
\end{aligned}
$$

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 ith community. Since $T_2 = \sum_{i=1}^{n/k} Z_i \cdot \mathbb{1}_{\{X_{\mathcal{C}_i}=1\}}$, we have

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

Let M denote the number of infected members of $\mathcal{C}_1$. Then by Lemma 1,

$$
\mathbb{E}\left[Z_1 \mid X_{\mathcal{C}_1} = 1\right] \leq (\log_2 k + 2) \cdot \mathbb{E}\left[M \mid X_{\mathcal{C}_1} = 1\right] + 1
$$

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

$$
\begin{aligned}
\mathbb{E}\left[M \mid X_{\mathcal{C}_1} = 1\right] &= \sum_{j=1}^{k} \mathbb{P}\left(X_j = 1 \mid X_{\mathcal{C}_1} = 1\right) \\
&= k \cdot \mathbb{P}\left(X_1 = 1 \mid X_{\mathcal{C}_1} = 1\right) \\
&= k \cdot \frac{\mathbb{P}(X_1 = 1, X_{\mathcal{C}_1} = 1)}{\mathbb{P}(X_{\mathcal{C}_1} = 1)} \\
&= k \cdot \frac{\mathbb{P}(X_1 = 1)}{\mathbb{P}(X_{\mathcal{C}_1} = 1)} \\
&= k \cdot \frac{1 - (1-p)(1 - pq_1)^{k-1}(1 - pq_2)^{n-k}}{\mathbb{P}(X_{\mathcal{C}_1} = 1)}
\end{aligned}
$$

where in the last line we invoke Lemma 3. Putting everything together gives

$$
\begin{aligned}
\mathbb{E}[T_2] &\leq n \cdot (\log_2 k + 2) \\
&\quad \cdot \left(1 - (1-p)(1 - pq_1)^{k-1}(1 - pq_2)^{n-k}\right) \\
&\quad + \frac{n}{k} \cdot \left(1 - (1-p)^k\left(1 - p\left(1 - (1 - q_2)^k\right)\right)^{n-k}\right)
\end{aligned}
$$

and therefore

$$
\begin{aligned}
\mathbb{E}[T] &\leq \frac{n}{k} \cdot \left(\log_2(n/k) + 3\right) \\
&\quad \cdot \left(1 - (1-p)^k\left(1 - p\left(1 - (1 - q_2)^k\right)\right)^{n-k}\right) \\
&\quad + n \cdot \left(\log_2 k + 1\right) \\
&\quad \cdot \left(1 - (1-p)(1 - pq_1)^{k-1}(1 - pq_2)^{n-k}\right) \\
&\quad + 1.
\end{aligned}
$$

☐

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] C. Mentus, M. Romeo, and C. DiPaola, "Analysis and applications of adaptive group testing methods for covid-19," *MedRxiv*, 2020.

[7] 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.

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

[9] 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.

[10] S. Ghosh, R. Agarwal, M. A. Rehan, S. Pathak, P. Agarwal, Y. Gupta, S. Consul, N. Gupta, R. Goenka, A. Rajwade, *et al.*, "A compressed sensing approach to pooled RT-PCR testing for COVID-19 detection," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 248–264, 2021.

[11] 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," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1030–1034, IEEE, 2021.

[12] 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.

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

[14] 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.

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

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] F. Hwang, T. T. Song, and D. Z. Du, "Hypergeometric and generalized hypergeometric group testing," *SIAM Journal on Algebraic Discrete Methods*, vol. 2, no. 4, pp. 426–428, 1981.

[26] 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.

[27] 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.

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

[29] J. Scarlett and V. Cevher, "Converse bounds for noisy group testing with arbitrary measurement matrices," in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 2868–2872, Ieee, 2016.

[30] 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.

[31] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3646–3661, 2018.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

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

[45] 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.

[46] A. Cohen, N. Shlezinger, S. Salamatian, Y. C. Eldar, and M. Médard, "Serial quantization for sparse time sequences," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3299–3314, 2021.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] B. Spang and M. Wootters, "Unconstraining graph-constrained group testing," in *International Conference on Randomization and Computation (RANDOM)*, 2019.

[53] A. Sterrett, "On the detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 1033–1036, 1957.

[54] C. H. Li, "A sequential method for screening experimental variables," *Journal of the American Statistical Association*, vol. 57, no. 298, pp. 455–477, 1962.

[55] H. Finucan, "The blood testing problem," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 13, no. 1, pp. 43–50, 1964.

[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.

[57] M. Sobel and P. A. Groll, "Binomial group-testing with an unknown proportion of defectives," *Technometrics*, vol. 8, no. 4, pp. 631–656, 1966.

[58] M. Hu, F. Hwang, and J. K. Wang, "A boundary problem for group testing," *SIAM Journal on Algebraic Discrete Methods*, vol. 2, no. 2, pp. 81–87, 1981.

[59] L. Riccio and C. J. Colbourn, "Sharper bounds in adaptive group testing," *Taiwanese Journal of Mathematics*, pp. 669–673, 2000.

[60] A. Allemann, "An efficient algorithm for combinatorial group testing," in *Information Theory, Combinatorics, and Search Theory*, pp. 569–596, Springer, 2013.

[61] 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.

[62] J. Scarlett, "An efficient algorithm for capacity-approaching noisy adaptive group testing," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2679–2683, IEEE, 2019.

[63] M. Aldridge, "Rates of adaptive group testing in the linear regime," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 236–240, IEEE, 2019.

[64] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, "Optimal group testing," *Combinatorics, Probability and Computing*, vol. 30, no. 6, pp. 811–848, 2021.

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

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

[67] B. Arasli and S. Ulukus, "Group testing with a graph infection spread model," *Information*, vol. 14, no. 1, p. 48, 2023.

[68] Y.-J. Lin, C.-H. Yu, T.-H. Liu, C.-S. Chang, and W.-T. Chen, "Positively correlated samples save pooled testing costs," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2170–2182, 2021.

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

[70] J. Scarlett and V. Cevher, "Phase transitions in group testing," in *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pp. 40–53, SIAM, 2016.

[71] B. Teo and J. Scarlett, "Noisy adaptive group testing via noisy binary search," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3340–3353, 2022.

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

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

[74] 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.

[75] M. Aldridge, O. Johnson, J. Scarlett, *et al.*, "Group testing: an information theory perspective," *Foundations and Trends® in Communications and Information Theory*, vol. 15, no. 3-4, pp. 196–392, 2019.

[76] M. Ben-Or and A. Hassidim, "The Bayesian learner is optimal for noisy binary search (and pretty good for quantum as well)," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 221–230, IEEE, 2008.

Surin Ahn (Student Member, IEEE) is currently a Ph.D. candidate in the Department of Electrical Engineering at Stanford University. Her research interests include information theory and machine learning. She received the B.S.E. degree (with Honors) in Electrical Engineering, with certificates in Applications of Computing and Robotics & Intelligent Systems, from Princeton University in 2018, and the M.S. degree in Electrical Engineering from Stanford University in 2020. She is a recipient of the Stanford Graduate Fellowship (SGF) and the Stanford Institute for Human-Centered Artificial Intelligence (HAI) Graduate Fellowship.

Wei-Ning Chen received a B.Sc. degree in Electrical Engineering and Mathematics in 2016, and an M.S. degree in Communication Engineering in 2019, both from National Taiwan University. He is currently working towards a Ph.D. degree in the Department of Electrical Engineering at Stanford University. He is a recipient of the Stanford Graduate Fellowship (SGF). His research interests include information theory, statistics, and theoretical machine learning, with applications in differential privacy and federated learning.

Ayfer Özgür (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical Engineering, Stanford University, where she is the Chambers Faculty Scholar with the School of Engineering. Her research interests include information theory, wireless communication, statistics, and machine learning. She received the EPFL Best Ph.D. Thesis Award in 2010, the NSF CAREER Award in 2013, the Okawa Foundation Research Grant, Faculty Research Awards from Google and Meta, and the IEEE Communication Theory Technical Committee (CTTC) Early Achievement Award in 2018. She was selected as the Inaugural Goldsmith Lecturer of the IEEE ITSoc in 2020.