Combinatorics, Probability and Computing (2021), **30**, pp. 811–848
doi:10.1017/S096354832100002X

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Optimal group testing

Amin Coja-Oghlan*, Oliver Gebhard, Max Hahn-Klimroth and Philipp Loick†

Goethe University Frankfurt, Robert-Mayer-Strasse 6–10, 60325 Frankfurt, Germany
*Corresponding author. Email: acoghlan@math.uni-frankfurt.de

(Received 17 April 2020; accepted 15 September 2020; revised 28 October 2020; first published online 28 January 2021)

Abstract

In the group testing problem the aim is to identify a small set of $k \sim n^\theta$ infected individuals out of a population size n, $0 < \theta < 1$. We avail ourselves of a test procedure capable of testing groups of individuals, with the test returning a positive result if and only if at least one individual in the group is infected. The aim is to devise a test design with as few tests as possible so that the set of infected individuals can be identified correctly with high probability. We establish an explicit sharp information-theoretic/algorithmic phase transition m_{inf} for non-adaptive group testing, where all tests are conducted in parallel. Thus with more than m_{inf} tests the infected individuals can be identified in polynomial time with high probability, while learning the set of infected individuals is information-theoretically impossible with fewer tests. In addition, we develop an optimal adaptive scheme where the tests are conducted in two stages.

2020 MSC Codes: 05C80, 60B20, 68P30

1. Introduction

1.1 Background and motivation

Various intriguing combinatorial problems come as inference tasks where we are to learn a hidden ground truth by means of indirect queries. The goal is to get by with as small a number of queries as possible. The ultimate solution to such a problem should consist of a positive algorithmic result showing that a certain number of queries suffice to learn the ground truth efficiently, complemented by a matching information-theoretic lower bound showing that with fewer queries the problem is insoluble, regardless of computational resources. Group testing is a prime example of such an inference problem [6]. The objective is to identify within a large population of size n a subset of k individuals infected with a rare disease. We presume that the number of infected individuals scales as a power $k = \lceil n^\theta \rceil$ of the population size with an exponent $\theta \in (0, 1)$, a parametrization suited to modelling the pivotal early stages of an epidemic [39]. Indeed, since early on in an epidemic test kits might be in short supply, it is vital to get the most diagnostic power out of the least number of tests. To this end we assume that the test gear is capable of not merely testing a single individual but an entire group. The test comes back positive if any one individual in the group is infected and negative otherwise. While in *non-adaptive* group testing all tests are conducted in parallel, in *adaptive* group testing test are conducted in several stages. In either case we

†Supported by DFG CO 646/3 and Stiftung Polytechnische Gesellschaft. An extended abstract version of this work appeared in the proceedings of the COLT 2020 conference (*Proc. Mach. Learning Res.* **125** (2020) 1374–1388).

are free to allocate individuals to test groups as we please. Randomization is allowed. What is the least number of tests required so that the set of infected individuals can be inferred from the test results with high probability? Furthermore, in adaptive group testing, what is the smallest depth of test stages required? Closing the considerable gaps that the best prior bounds left, the main results of this paper furnish matching algorithmic and information-theoretic bounds for both adaptive and non-adaptive group testing. Specifically, the best prior information-theoretic lower bound derives from the following folklore observation. Suppose that we conduct m tests that each return either 'positive' or 'negative'. Then, to correctly identify the set of infected individuals, we need the total number 2^m of conceivable test results to asymptotically exceed the number $\binom{n}{k}$ of possible sets of infected individuals. Hence $2^m \geqslant (1+o(1))\binom{n}{k}$. Thus Stirling's formula yields the lower bound

$$m_{\mathrm{ad}} = \frac{1-\theta}{\ln 2} n^{\theta} \ln n, \tag{1.1}$$

which applies to both adaptive and non-adaptive testing. On the positive side, a randomized non-adaptive test design with

$$m_{\mathrm{DD}} \sim \frac{\max\{\theta, 1-\theta\}}{\ln^2 2} n^{\theta} \ln n \tag{1.2}$$

tests exists from which a greedy algorithm called DD correctly infers the set of infected individuals with high probability [24]. Clearly $m_{\mathrm{ad}} < m_{\mathrm{DD}}$ for all infection densities θ and $m_{\mathrm{DD}}/m_{\mathrm{ad}} \to \infty$ as $\theta \to 1$. In addition, there is an efficient adaptive three-stage group testing scheme that asymptotically matches the lower bound m_{ad} [35]. We proceed to state the main results of the paper. First, improving both the information-theoretic and the algorithmic bounds, we present optimal results for non-adaptive group testing. Subsequently we show how the non-adaptive result can be harnessed to perform adaptive group testing with the least possible number $(1+o(1))m_{\mathrm{ad}}$ of tests in only two stages.

1.2 Non-adaptive group testing

A *non-adaptive test design* is a bipartite graph $G = (V \cup F, E)$ with one vertex class $V = V_n = \{x_1, \ldots, x_n\}$ representing individuals and the other class $F = F_m = \{a_1, \ldots, a_m\}$ representing tests. For a vertex v of G, let $\partial v = \partial_G v$ denote the set of neighbours of v. Thus an individual x_j takes part in a test a_i if and only if $x_j \in \partial a_i$. Since we can shuffle the individuals randomly, we may safely assume that the vector $\boldsymbol{\sigma} \in \{0,1\}^V$ whose 1-entries mark the infected individuals is a uniformly random vector of Hamming weight k. Furthermore, the test results induced by $\boldsymbol{\sigma}$ read

$$\hat{\boldsymbol{\sigma}}_{a_i} = \hat{\boldsymbol{\sigma}}_{G,a_i} = \max_{x \in \partial a_i} \boldsymbol{\sigma}_x.$$

Hence, given $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_G = (\hat{\boldsymbol{\sigma}}_{G,a})_{a \in F}$ and G, we aim to infer $\boldsymbol{\sigma}$. Thus we can represent an inference procedure by a function $\mathcal{A}_G \colon \{0,1\}^m \to \{0,1\}^n$. The following theorem improves the lower bound on the number of tests required for successful inference. Let

$$m_{\mathrm{inf}} = m_{\mathrm{inf}}(n,\theta) = \max\left\{\frac{\theta}{\ln^2 2}, \frac{1-\theta}{\ln 2}\right\} n^{\theta} \ln n. \tag{1.3}$$

Theorem 1.1. *For any $0 < \theta < 1$, $\varepsilon > 0$ there exists $n_0 = n_0(\theta, \varepsilon)$ such that for all $n > n_0$, all test designs G with $m \leqslant (1-\varepsilon)m_{\mathrm{inf}}$ tests and for every function $\mathcal{A}_G \colon \{0,1\}^m \to \{0,1\}^n$, we have*

$$\mathbb{P}[\mathcal{A}_G(\hat{\boldsymbol{\sigma}}_G) = \boldsymbol{\sigma}] < \varepsilon. \tag{1.4}$$

Theorem 1.1 rules out both deterministic and randomized test designs and inference procedures because (1.4) holds uniformly for all G and all $\mathcal{A}_G$. Thus no test design, randomized or not,

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

with fewer than m_{inf} tests allows us to infer the set of infected individuals with a non-vanishing probability. Since m_{inf} matches m_{DD} from (1.2) for $\theta \geqslant 1/2$, Theorem 1.1 shows that the positive result from [24] is optimal in this regime. The following theorem closes the remaining gap by furnishing an optimal positive result for all θ.

Theorem 1.2. *For any $0 < \theta < 1$, $\varepsilon > 0$ there is $n_0 = n_0(\theta, \varepsilon)$ such that for every $n > n_0$ there exist a randomized test design $\mathbf{G}$ comprising $m \leqslant (1 + \varepsilon)m_{\text{inf}}$ tests and a polynomial-time algorithm* `SPIV` *that, given $\mathbf{G}$ and the test results $\hat{\boldsymbol{\sigma}}_{\mathbf{G}}$ outputs $\boldsymbol{\sigma}$ with high probability.*

An obvious candidate for an optimal test design appears be a plain random bipartite graph. In fact, prior to the present work the best known test design consisted of a uniformly random bipartite graph where all vertices in V_n have the same degree Δ. In other words, every individual independently joins Δ random test groups. Applied to this random Δ-out test design the `DD` algorithm correctly recovers the set of infected individuals in polynomial time provided that the number of tests exceeds m_{DD} from (1.2). However, m_{DD} strictly exceeds m_{inf} for $\theta < 1/2$. While the random Δ-out test design with $(1 + o(1))m_{\text{inf}}$ tests is known to admit an exponential-time algorithm that successfully infers the set of infected individuals with high probability [12], we do not know of a polynomial-time algorithm that solves this inference problem. Instead, to facilitate the new efficient inference algorithm `SPIV`, the test design for Theorem 1.2 relies on a blend of a geometric and a random construction that is inspired by recent advances in coding theory known as spatially coupled low-density parity-check codes [19, 28]. Finally, for

$$\theta \leqslant \frac{\ln 2}{1 + \ln 2} \approx 0.41 \tag{1.5}$$

the number m_{inf} of tests required by Theorem 1.2 matches the folklore lower bound m_{ad} from (1.2) that applies to both adaptive and non-adaptive group testing. Hence in this regime adaptivity confers no advantage. By contrast, for $\theta > \ln(2)/(1 + \ln 2)$ the adaptive bound m_{ad} is strictly smaller than m_{inf}. Consequently, in this regime at least two test stages are necessary to match the lower bound. Indeed, the next theorem shows that two stages suffice.

1.3 Adaptive group testing

A *two-stage test design* consists of a bipartite graph $G = (V, F)$ along with a second bipartite graph $G' = G'(G, \hat{\boldsymbol{\sigma}}_G) = (V', F')$ with $V' \subset V$ that may depend on the test results $\hat{\boldsymbol{\sigma}}_G$ of the first test design G. Hence the task is to learn $\boldsymbol{\sigma}$ correctly with high probability from $G, \hat{\boldsymbol{\sigma}}_G, G'$ and the test results $\hat{\boldsymbol{\sigma}}_{G'}$ from the second stage while minimizing the total number $|F| + |F'|$ of tests. The following theorem shows that a two-stage test design and an efficient inference algorithm exist that meet the multi-stage adaptive lower bound (1.1).

Theorem 1.3. *For any $0 < \theta < 1$, $\varepsilon > 0$ there is $n_0 = n_0(\theta, \varepsilon)$ such that for every $n > n_0$ there exist a two-stage test design with no more than $(1 + \varepsilon)m_{\text{ad}}$ tests in total and a polynomial-time inference algorithm that outputs $\boldsymbol{\sigma}$ with high probability.*

Theorem 1.3 improves over [35] by reducing the number of stages from three to two, thus potentially significantly reducing the overall time required to complete the test procedure [11, 30]. The proof of Theorem 1.3 combines the test design and efficient algorithm from Theorem 1.2 with ideas from [34]. The question of whether an ‘adaptivity gap’ exists for group testing, *i.e.* if the number of tests can be reduced by allowing multiple stages, has been raised prominently [6]. Theorems 1.1–1.3 answer this question comprehensively. While for $\theta \leqslant \ln(2)/(1 + \ln(2)) \approx 0.41$ adaptivity confers no advantage, Theorem 1.1 shows that for $\theta > \ln(2)/(1 + \ln(2))$ there is

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

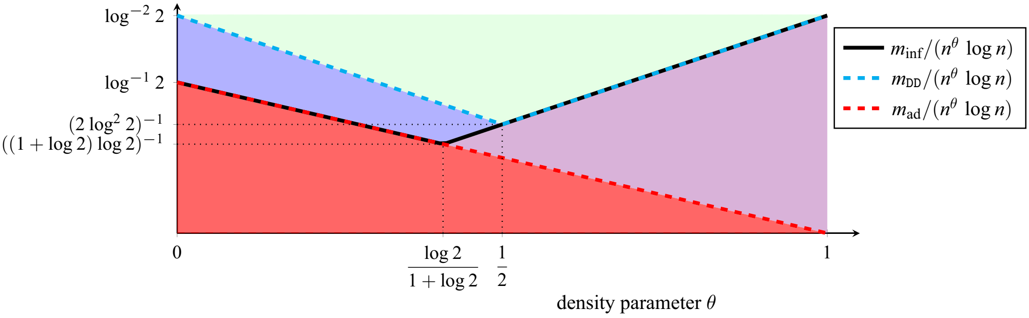

Figure 1. The phase transitions in group testing. The best previously known algorithm DD succeeds in the green region but not in the blue region. The new algorithm SPIV succeeds in both the green and blue regions. The black line indicates the non-adaptive information-theoretic threshold m_{inf}, below which non-adaptive group testing is impossible. In the red area even (multi-stage) adaptive inference is impossible. Finally, the two-stage adaptive group testing algorithm from Theorem 1.3 succeeds in the purple region.

a widening gap between m_{ad} and the number m_{inf} of tests required by the optimal non-adaptive test design. Further, Theorem 1.3 demonstrates that this gap can be closed by allowing merely two stages. Figure 1 illustrates the thresholds from Theorems 1.1–1.3.

1.4 Discussion

The group testing problem was first raised in 1943, when Dorfman [16] proposed a two-stage adaptive test design to test the US Army for syphilis. In a first stage disjoint groups of equal size are tested. All members of negative test groups are definitely uninfected. Then in the second stage the members of positive test groups get tested individually. Of course, this test design is far from optimal for low infection rates, but Dorfman's contribution triggered attempts to devise improved test schemes. An initial combinatorial group testing, where the aim is to construct a test design that is guaranteed to succeed on *all* vectors $\boldsymbol{\sigma}$, attracted significant attention. This version of the problem was studied, among others, by Erdős and Rényi [18], D'yachkov and Rykov [17] and Kautz and Singleton [25]. Hwang [22] was the first to propose an adaptive test design that asymptotically meets the information-theoretic lower bound m_{ad} from (1.1) for all $\theta \in [0, 1]$. However, this test design requires an unbounded number of stages. Conversely, D'yachkov and Rykov [17] showed that m_{ad} tests do not suffice for non-adaptive group testing. Indeed, $m \geqslant \min\{\Omega(k^2), n\}$ tests are required non-adaptively, making individual testing optimal for $\theta > 1/2$. For an excellent survey of combinatorial group testing, see [6].

Since the early 2000s attention has shifted to probabilistic group testing, which we study here as well. Thus, instead of asking for test designs and algorithms that are guaranteed to work for *all* $\boldsymbol{\sigma}$, we are content to recover $\boldsymbol{\sigma}$ with high probability. Berger and Levenshtein [9] presented a two-stage probabilistic group testing design and algorithm requiring

$$m_{\mathrm{BL,ad}} \sim 4n^\theta \ln n$$

tests in expectation. Their test design, known as the Bernoulli design, is based on a random bipartite graph where each individual joins every test independently with a carefully chosen edge probability. For a fixed θ the number $m_{\mathrm{BL,ad}}$ of tests is within a bounded factor of the information-theoretic lower bound m_{ad} from (1.1), although the gap $m_{\mathrm{ad}}/m_{\mathrm{BL,ad}}$ diverges as $\theta \to 1$. Unsurprisingly, the work of Berger and Levenshtein spurred efforts at closing the gap. Mézard, Tarzia and Toninelli proposed a different two-stage test design whose first stage consists

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

of a random bipartite graph called the constant weight design [31]. Here each individual independently joins an equal number of random tests. For their two-stage design they obtained an inference algorithm that gets by with about

$$m_{\mathrm{MTT,ad}} \sim \frac{1-\theta}{\ln^2 2} n^{\theta} \ln n \tag{1.6}$$

tests, a factor of $1/\ln 2$ above the elementary bound m_{ad}. Conversely, Mézard, Tarzia and Toninelli showed by means of the FKG inequality and positive correlation arguments that two-stage test algorithms from a certain restricted class cannot beat the bound (1.6). Furthermore, Aldridge, Johnson and Scarlett analysed non-adaptive test designs and inference algorithms [5, 24]. For the Bernoulli test design their best efficient algorithm DD requires

$$m_{\mathrm{DD,Be}} \sim \mathrm{e} \cdot \max\{\theta, 1-\theta\} n^{\theta} \ln n$$

tests. For the constant weight design they obtained the bound m_{DD} from (1.2). In addition, in a previous article [12] we showed that on the constant weight design an exponential-time algorithm correctly identifies the set of infected individuals with high probability if the number of tests exceeds m_{inf} from (1.3). Furthermore, Scarlett [35] discovered the aforementioned three-stage test design and polynomial-time algorithm that matches the universal lower bound m_{ad} from (1.1). Finally, concerning lower bounds, in the case of a linear number $k = \Theta(n)$ of infected individuals, Aldridge [4] showed via arguments similar to [31] that individual testing is optimal in the non-adaptive case, while Ungar [38] proved that individual testing is optimal, even adaptively, once $k \geqslant (3-\sqrt{5})n/2$.

A further variant of group testing is known as quantitative group testing or the coin weighing problem. In this problem tests are assumed to not merely indicate the presence of at least one infected individual but to return the number of infected individuals. Thus the tests are significantly more powerful. For quantitative group testing with k infected individuals, Alaoui, Ramdas, Krzakala, Zdeborová and Jordan [3] presented a test design with

$$m_{\mathrm{QGT}} \sim 2\left(1 + \frac{(n-k)\ln(1-k/n)}{k\ln(k/n)}\right)\frac{k\ln(n/k)}{\ln(k)} \quad \text{for } k = \Theta(n)$$

tests from which the set of infected individuals can be inferred in exponential time; the paper actually deals with the slightly more general pooled data problem. However, no efficient algorithm is known to come within a constant factor of m_{QGT}. Indeed, the best efficient algorithm, due to the same authors [2], requires $\Omega(k\ln(n/k))$ tests.

More broadly, the idea of harnessing random graphs to tackle inference problems has been gaining momentum. One important success has been the development of capacity achieving linear codes called spatially coupled low-density parity-check ('LDPC') codes [28, 29]. The Tanner graphs of these codes, which represent their check matrices, consist of a linear sequence of sparse random bipartite graphs with one class of vertices corresponding to the bits of the codeword and the other class corresponding to the parity checks. The bits and the checks are divided equitably into a number of compartments, which are arranged along a line. Each bit of the codeword takes part in random checks in a small number of preceding and subsequent compartments of checks along the line. This combination of a spatial arrangement and randomness facilitates efficient decoding by means of the belief propagation message passing algorithm. Furthermore, the general design idea of combining a linear spatial structure with a random graph has been extended to other inference problems. Perhaps the most prominent example is compressed sensing, that is, solving an underdetermined linear system subject to a sparsity constraint [14, 15, 26, 27], where a variant of belief propagation called approximate message passing matches an information-theoretic lower bound from [40].

While in some inference problems, such as LDPC decoding or compressed sensing, the number of queries required to enable an efficient inference algorithm matches the information-theoretic

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

lower bound, in many other problems gaps remain. A prominent example is the stochastic block model [1, 13, 32], an extreme case of which is the notorious planted clique problem [7]. For both these models the existence of a genuine computationally intractable phase where the problem can be solved in exponential but not in polynomial time appears to be an intriguing possibility. Further examples include code division multiple access [37, 41], quantitative group testing [2], sparse principal component analysis [10] and sparse high-dimensional regression [33]. The problem of solving the group testing inference problem on the test design from [24] could be added to the list. Indeed, while an exponential-time algorithm (that reduces the problem to minimum hypergraph vertex cover) infers the set of infected individuals with high probability with only $(1+\varepsilon)m_{\inf}$ tests, the best known polynomial algorithm requires $(1+\varepsilon)m_{\mathrm{DD}}$ tests. Instead of developing a better algorithm for the test design from [24], here we exercise the discretion of constructing a different test design that the group testing problem affords. The new design is tailored to enable an efficient algorithm SPIV for Theorem 1.2 that gets by with $(1+\varepsilon)m_{\inf}$ tests. While prior applications of the idea of spatial coupling such as coding and compressed sensing required sophisticated message passing algorithms [19, 28, 29], the SPIV algorithm is purely combinatorial and extremely transparent. The main step of the algorithm merely computes a weighted sum to discriminate between infected individuals and ‘disguised’ healthy individuals. Furthermore, the analysis of the algorithm is based on a technically subtle but conceptually clean large deviations analysis. This technique of blending combinatorial ideas and large deviations methods with spatial coupling promises to be an exciting route for future research. Applications might include noisy versions of group testing, quantitative group testing or the coin weighing problem [2]. Beyond these immediate extensions, it would be most interesting to see if the SPIV strategy extends to other inference problems for sparse data. Clearly our algorithm is not currently practical for typically small problem sizes. But experience shows that such theoretical results often have meaningful practical bearing. A prominent case is that of spatially coupled LDPC and polar codes [8]. After being theoretically studied in the coding community, they today underlie the mobile communication standards 4G and 5G [36]. Another example is the ellipsoid method [20]. The opportunity to solve linear programs in polynomial time caused a lot of excitement within theoretical and practical research. We view the experimental study of group testing algorithms as an interesting next research step and are optimistic that (a variant of) the SPIV algorithm can lead to a corresponding success story for efficient and practical decoding in group testing.

1.5 Organization

After collecting some preliminaries and introducing notation in Section 2, we prove Theorem 1.1 in Section 3. Section 4 then deals with the test design and the inference algorithm for Theorem 1.2. Finally, in Section 5 we prove Theorem 1.3.

2. Preliminaries

As we saw in Section 1.2, a non-adaptive test design can be represented by a bipartite graph $G=(V\cup F, E)$, with one vertex class V representing the individuals and the other class F representing the tests. We refer to the number $|V|$ of individuals as the *order* of the test design and to the number $|F|$ of tests as its *size*. For a vertex v of G we let $\partial_G v$ denote the set of neighbours. Where G is apparent from the notation we just write ∂v. Furthermore, for an integer $k\leqslant |V|$ we let $\boldsymbol{\sigma}_{G,k}=(\boldsymbol{\sigma}_{G,k,x})_{x\in V}\in\{0,1\}^V$ denote a random vector of Hamming weight k. Additionally, we let

$$\hat{\boldsymbol{\sigma}}_{G,k}=(\hat{\boldsymbol{\sigma}}_{G,k,a})_{a\in F}\in\{0,1\}^F \quad \text{with } \hat{\boldsymbol{\sigma}}_{G,k,a}=\max_{x\in\partial_G a}\boldsymbol{\sigma}_{G,k,x} \tag{2.1}$$

be the associated vector of test results. Where G and/or k are apparent from the context, we drop them from the notation. More generally, for a given vector $\tau\in\{0,1\}^V$ we introduce a vector $\hat{\tau}_G=$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

$(\hat{\tau}_{G,a})_{a\in F}$ by letting $\hat{\tau}_{G,a}=\max_{x\in\partial_G a}\tau_x$, just as in (2.1). Furthermore, for a given $\tau\in\{0,1\}^V$ we let

$$V_0(G,\tau)=\{x\in V\colon \tau_x=0\},\qquad V_1(G,\tau)=\{x\in V\colon \tau_x=1\},$$
$$F_0(G,\tau)=\{a\in F\colon \hat{\tau}_{G,a}=0\},\quad F_1(G,\tau)=\{a\in F\colon \hat{\tau}_{G,a}=1\}$$

be the set of healthy and infected individuals, respectively the set of negative and positive tests. The *Kullback–Leibler divergence* of $p,q\in(0,1)$ is denoted by

$$D_{\mathrm{KL}}(q\|p)=q\ln\left(\frac{q}{p}\right)+(1-q)\ln\left(\frac{1-q}{1-p}\right).$$

We will occasionally apply the following Chernoff bound.

Lemma 2.1 (Theorem 2.1 of [23]). *Let $\mathbf{X}$ be a binomial random variable with parameters N,p. Then*

$$\mathbb{P}[X\geqslant qN]\leqslant\exp(-ND_{\mathrm{KL}}(q\|p))\quad \textit{for } p<q<1, \tag{2.2}$$
$$\mathbb{P}[X\leqslant qN]\leqslant\exp(-ND_{\mathrm{KL}}(q\|p))\quad \textit{for } 0<q<p. \tag{2.3}$$

In addition, we recall that the *hypergeometric distribution* $\mathrm{Hyp}(L,M,N)$ is defined by

$$\mathbb{P}[\mathrm{Hyp}(L,M,N)=k]=\binom{M}{k}\binom{L-M}{N-k}\binom{L}{N}^{-1}\qquad (k\in\{0,1,\ldots,M\wedge N\}).$$

Hence, out of a total of L items of which M are special, we draw N items without replacement and count the number of special items in the draw. The mean of the hypergeometric distribution equals MN/L. It is well known that the Chernoff bound extends to the hypergeometric distribution.

Lemma 2.2 ([21]). *For a hypergeometric variable $\mathbf{X}\sim\mathrm{Hyp}(L,M,N)$, the bounds (2.2)–(2.3) hold with $p=M/L$.*

Throughout the paper we use asymptotic notation $o(\,\cdot\,),\omega(\,\cdot\,),O(\,\cdot\,),\Omega(\,\cdot\,),\Theta(\,\cdot\,)$ to refer to limit $n\to\infty$. It is understood that the constants hidden in, for example, a $O(\,\cdot\,)$-term may depend on the density parameter θ or other parameters. Furthermore, we say that a statement holds with high probability if it holds with probability $1-o(1)$. Moreover, for two numbers a,b we abbreviate $\min\{a,b\}=a\wedge b$, whereas for two events A,B we let $A\wedge B$ denote the event of A and B occurring.

3. The information-theoretic lower bound

In this section we prove Theorem 1.1. Section 3.1 contains a proof outline stating the main steps towards the proof of Theorem 1.1. Subsequently, we proceed with the proofs of these steps in Sections 3.2 and 3.3. The proof combines techniques based on the FKG inequality and positive correlation that were developed in [6] and [31] with new combinatorial ideas. Throughout this section we fix a number $\theta\in(0,1)$ and we let $k=\lceil n^\theta\rceil$.

3.1 Outline

The starting point is a simple and well-known observation. Namely, for a test design $G=G_{n,m}=(V_n,F_m)$ and a vector $\tau\in\{0,1\}^{F_m}$ of test results, let

$$\mathcal{S}_k(G,\tau)=\left\{\sigma\in\{0,1\}^{V_n}\colon \sum_{x\in V_n}\sigma_x=k,\ \hat{\sigma}_G=\tau\right\}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

be the set of all possible vectors σ of Hamming weight k that give rise to the test results τ, *i.e.* the satisfying set [6]. Further, let $Z_k(G,\tau)=|\mathcal{S}_k(G,\tau)|$ be the number of such vectors σ. Also recall that $\boldsymbol{\sigma}=\boldsymbol{\sigma}_{G,k}\in\{0,1\}^{V_n}$ is a random vector of Hamming weight k and that $\hat{\boldsymbol{\sigma}}=\hat{\boldsymbol{\sigma}}_{G,k}$ comprises the test results that $\boldsymbol{\sigma}$ renders under the test design G. We observe that the posterior of $\boldsymbol{\sigma}$ given $\hat{\boldsymbol{\sigma}}$ is the uniform distribution on $\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})$.

Fact 3.1. For any $G,\sigma\in\{0,1\}^{V_n}$ we have

$$\mathbb{P}[\boldsymbol{\sigma}=\sigma\mid\hat{\boldsymbol{\sigma}}]=\mathbf{1}\{\sigma\in\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})\}/Z_k(G,\hat{\boldsymbol{\sigma}}).$$

As an immediate consequence of Fact 3.1, the success probability of any inference scheme $\mathcal{A}_G\colon\{0,1\}^{F_m}\to\{0,1\}^{V_n}$ is bounded by $1/Z_k(G,\hat{\boldsymbol{\sigma}})$. Indeed, an optimal inference algorithm is to simply return a uniform sample from $\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})$.

Fact 3.2. For any test design G and for any $\mathcal{A}_G\colon\{0,1\}^{F_m}\to\{0,1\}^{V_n}$, we have

$$\mathbb{P}[\mathcal{A}_G(\hat{\boldsymbol{\sigma}})=\boldsymbol{\sigma}\mid\hat{\boldsymbol{\sigma}}]\leqslant 1/Z_k(G,\hat{\boldsymbol{\sigma}}).$$

Hence, in order to prove Theorem 1.1, we just need to show that $Z_k(G,\hat{\boldsymbol{\sigma}})$ is large for any test design G with $m<(1-\varepsilon)m_{\inf}$ tests. In other words, we need to show that with high probability there are many vectors $\sigma\in\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})$ that give rise to the test results $\hat{\boldsymbol{\sigma}}$.

We obtain these σ by making diligent local changes to $\boldsymbol{\sigma}$. More precisely, we identify two sets $V_{0+}=V_{0+}(G,\boldsymbol{\sigma})$, $V_{1+}=V_{1+}(G,\boldsymbol{\sigma})$ of individuals whose infection status can be flipped without altering the test results. Specifically, following [4] we call an individual $x\in V_n$ *disguised* if every test $a\in\partial_G x$ contains another individual $y\in\partial_G a\setminus\{x\}$ with $\boldsymbol{\sigma}_y=1$. Let $V_+=V_+(G,\boldsymbol{\sigma})$ be the set of all disguised individuals. Moreover, let

$$V_{0+}=V_{0+}(G,\boldsymbol{\sigma})=\{x\in V_+\colon\boldsymbol{\sigma}_x=0\},\qquad V_{1+}=V_{1+}(G,\boldsymbol{\sigma})=\{x\in V_+\colon\boldsymbol{\sigma}_x=1\}.\tag{3.1}$$

Hence V_{0+} is the set of all healthy disguised individuals while V_{1+} contains all infected disguised individuals.

Fact 3.3. We have $Z_k(G,\hat{\boldsymbol{\sigma}})\geqslant|V_{0+}(G,\boldsymbol{\sigma})|\cdot|V_{1+}(G,\boldsymbol{\sigma})|$.

Proof. For a pair $(x,y)\in V_{0+}(G,\boldsymbol{\sigma})\times V_{1+}(G,\boldsymbol{\sigma})$, obtain $\boldsymbol{\tau}$ from $\boldsymbol{\sigma}$ by letting $\boldsymbol{\tau}_x=1$, $\boldsymbol{\tau}_y=0$ and $\boldsymbol{\tau}_z=\boldsymbol{\sigma}_z$ for all $z\neq x,y$. Then $\boldsymbol{\tau}$ has Hamming weight k and $\hat{\boldsymbol{\tau}}_G=\hat{\boldsymbol{\sigma}}$. Thus $\boldsymbol{\tau}\in\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})$. □

Hence an obvious proof strategy for Theorem 1.1 is to exhibit a large number of disguised individuals. A similar strategy has been pursued in the proof of the conditional lower bound of Mézard, Tarzia and Toninelli [31] and the proof of Aldridge's lower bound for the linear case $k=\Theta(n)$ [4]. Both [4] and [31] exhibit disguised individuals via positive correlation and the FKG inequality. However, while in [4] it sufficed to find one disguised individual to yield a constant error probability for any algorithm, the polynomially small prior in the sublinear regime requires us to find an enormous number of disguised individuals, $\omega(n/k)$ to be precise. We do not see how to stretch the arguments in [4] and [31] to obtain the desired lower bound for all $\theta\in(0,1)$. Yet for θ *extremely* close to one it is possible to combine the positive correlation argument with new combinatorial ideas to obtain the following.

Proposition 3.1. *For any $\varepsilon>0$ there exists $\theta_0=\theta_0(\varepsilon)<1$ such that for every $\theta\in(\theta_0,1)$ there exists $n_0=n_0(\theta,\varepsilon)$ such that, for all $n>n_0$ and all test designs $G=G_{n,m}$ with $m\leqslant(1-\varepsilon)m_{\inf}$, we have*

$$\mathbb{P}[|V_{0+}(G,\boldsymbol{\sigma})|\wedge|V_{1+}(G,\boldsymbol{\sigma})|\geqslant\ln n]>1-\varepsilon.$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

The proof of Proposition 3.1 can be found in Section 3.2. The second step towards Theorem 1.1 is a reduction from larger to smaller values of θ. Suppose we wish to apply a test scheme designed for an infection density $\theta \in (0, 1)$ to a larger infection density $\theta' \in (\theta, 1)$. Then we could dilute the larger infection density by adding a large number of healthy 'dummy' individuals. A careful analysis of this dilution process yields the following result. As the elementary lower bound (1.1) coincides with $m_{\inf}$ for $\theta \leqslant \ln(2)/(1 + \ln 2)$, we need not worry about this regime.

Proposition 3.2. *For any* $\ln(2)/(1 + \ln(2)) < \theta < \theta' < 1$, $t > 0$ *there exists* $n_0 = n_0(\theta, \theta', t) > 0$ *such that, for every* $n > n_0$ *and for every test design* G *of order* n*, there exist an integer* n' *such that*

$$k = \lceil n^{\theta} \rceil = \lceil n'^{\,\theta'} \rceil$$

and a test design G' *of order* n' *with the same number of tests as* G *such that the following is true. Let* $\boldsymbol{\tau} \in \{0, 1\}^{V_{n'}}$ *be a random vector of Hamming weight* k *and let* $\hat{\boldsymbol{\tau}}_a = \max_{x \in \partial_{G'} a} \boldsymbol{\tau}_x$ *comprise the test results of* G'*. Then*

$$\mathbb{P}[Z_k(G, \hat{\boldsymbol{\sigma}}) \leqslant t] \leqslant \mathbb{P}[Z_k(G', \hat{\boldsymbol{\tau}}) \leqslant t].$$

Hence, if a test design exists for $\theta < \theta'$ that beats $m_{\inf}(n, \theta)$, then there is a test design for infection density θ' that beats $m_{\inf}(n', \theta')$. We prove Proposition 3.1 in Section 3.2. Theorem 1.1 is an easy consequence of Propositions 3.1 and 3.2.

Proof of Theorem 1.1. For $\theta \leqslant \ln(2)/(1 + \ln(2))$ the assertion follows from the elementary lower bound (1.1). Hence, fix $\varepsilon > 0$ and assume for contradiction that some $\theta \in (\ln(2)/(1 + \ln(2)), 1)$ for infinitely many n admits a test design G of order n and size $m \leqslant (1 - \varepsilon) m_{\inf}(n, \theta)$ such that $\mathbb{P}[Z_k(G, \hat{\boldsymbol{\sigma}}_G) \leqslant 1/\varepsilon] \geqslant \varepsilon$. Then Proposition 3.2 shows that, for $\theta' > \theta$ arbitrarily close to one for an integer n' with $k = \lceil n'^{\,\theta'} \rceil$, a test design $G' = G_{n',m}$ exists such that

$$\mathbb{P}[Z_k(G', \hat{\boldsymbol{\tau}}) \leqslant 1/\varepsilon] \geqslant \varepsilon. \tag{3.2}$$

Furthermore, (1.3) shows that for large n

$$m_{\inf}(n', \theta') = \frac{\theta'}{\ln^2 2} n'^{\,\theta'} \ln n' = \frac{\theta + o(1)}{\ln^2 2} n^{\theta} \ln n = (1 + o(1)) m_{\inf}(n, \theta).$$

Hence the number m of tests of G' satisfies $m \leqslant (1 - \varepsilon + o(1)) m_{\inf}(n', \theta')$. Thus (3.2) contradicts Fact 3.3 and Proposition 3.1. □

3.2 Proof of Proposition 3.1

Given a small $\varepsilon > 0$ we choose $\theta_0 = \theta_0(\varepsilon) \in (0, 1)$ sufficiently close to one and fix $\theta \in (\theta_0, 1)$. Additionally, pick $\xi = \xi(\varepsilon, \theta) \in (0, 1)$ such that

$$2(1 - \theta) < \xi < \theta\varepsilon. \tag{3.3}$$

We fix ε, θ, ξ throughout this section. To avoid the (mild) stochastic dependences that result from the total number of infected individuals being fixed, instead of $\boldsymbol{\sigma}$ we will consider a vector $\boldsymbol{\chi} \in \{0, 1\}^{V_n}$ whose entries are stochastically independent. Specifically, every entry of $\boldsymbol{\chi}$ equals one with probability

$$p = \frac{k - \sqrt{k} \ln n}{n} \tag{3.4}$$

independently. Let $\hat{\boldsymbol{\chi}}_G \in \{0, 1\}^{F_m}$ be the corresponding vector of test results. The following lemma shows that it suffices to estimate $|V_{0+}(G, \boldsymbol{\chi})|, |V_{1+}(G, \boldsymbol{\chi})|$, thus the number of disguised uninfected and infected individuals. Let G denote an arbitrary test design with individuals $V_n = \{x_1, \ldots, x_n\}$ and tests $F_m = \{a_1, \ldots, a_m\}$.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Lemma 3.3. *There is $n_0 = n_0(\theta, \varepsilon)$ such that, for all $n > n_0$ and for all G with $m \leqslant m_{\inf}$, the following is true:*

$$\text{if } \mathbb{P}[|V_{0+}(G, \boldsymbol{\chi})| \wedge |V_{1+}(G, \boldsymbol{\chi})| \geqslant 2 \ln n] > 1 - \varepsilon/4,$$
$$\text{then } \mathbb{P}[|V_{0+}(G, \boldsymbol{\sigma})| \wedge |V_{1+}(G, \boldsymbol{\sigma})| \geqslant \ln n] > 1 - \varepsilon.$$

Proof. Let

$$\mathcal{X} = \left\{ k - 2\sqrt{k} \ln n \leqslant \sum_{x \in V_n} \boldsymbol{\chi}_x \leqslant k \right\}.$$

The Chernoff bound shows that for large enough n,

$$\mathbb{P}[\mathcal{X}] > 1 - \eta/4. \tag{3.5}$$

Further, given $\mathcal{X}$ we can couple $\boldsymbol{\chi}, \boldsymbol{\sigma}$ such that the latter is obtained by turning $k - \sum_{x \in V_n} \boldsymbol{\chi}_x$ random zero entries of the former into ones. Since turning zero entries into ones can only increase the number of disguised individuals, on $\mathcal{X}$ we have

$$V_{1+}(G, \boldsymbol{\sigma}) \geqslant V_{1+}(G, \boldsymbol{\chi}). \tag{3.6}$$

Of course, it is possible that $|V_{0+}(G, \boldsymbol{\sigma})| < |V_{0+}(G, \boldsymbol{\chi})|$. But since on $\mathcal{X}$ the two vectors $\boldsymbol{\sigma}, \boldsymbol{\chi}$ differ in no more than $2\sqrt{k} \ln n$ entries, we obtain the bound

$$\mathbb{E}[|V_{0+}(G, \boldsymbol{\chi})| - |V_{0+}(G, \boldsymbol{\sigma})| \mid \mathcal{X}] \leqslant \frac{2\sqrt{k} \ln n}{n - k} |V_{0+}(G, \boldsymbol{\chi})| < n^{-1/3} |V_{0+}(G, \boldsymbol{\chi})|,$$

provided n is sufficiently large. Hence Markov's inequality shows that for large enough n

$$\mathbb{P}[|V_{0+}(G, \boldsymbol{\chi})| - |V_{0+}(G, \boldsymbol{\sigma})| > |V_{0+}(G, \boldsymbol{\chi})|/2 \mid \mathcal{X}] < \varepsilon/4. \tag{3.7}$$

Combining (3.5), (3.6) and (3.7) completes the proof. □

As a next step we show that there is no point in having very big tests a that contain more than, say, $\Gamma = \Gamma(n, \theta) = n^{1-\theta} \ln n$ individuals. This is because in any case all such tests are positive with high probability, so there is little point in actually conducting them. Indeed, the following lemma shows that with high probability all tests of very high degree contain at least two infected individuals.

Lemma 3.4. *There exists $n_0 = n_0(\theta, \varepsilon) > 0$ such that, for all $n > n_0$ and all test designs G with $m \leqslant m_{\inf}$ tests,*

$$\mathbb{P}[\exists a \in F_m \colon |\partial_G a| > \Gamma \wedge |\partial_G a \cap V_1(G, \boldsymbol{\chi})| \leqslant 1] < \varepsilon/8.$$

Proof. Consider a test a of degree $\gamma = |\partial_G a| \geqslant \Gamma$. Because in $\boldsymbol{\chi}$ each of the γ individuals that take part in a is infected with probability p independently, we have

$$\begin{aligned} \mathbb{P}[|\partial_G a \cap V_1(G, \boldsymbol{\sigma})| \leqslant 1] &= \mathbb{P}[\mathrm{Bin}(\gamma, p) \leqslant 1] \\ &= (1-p)^\gamma + \gamma p (1-p)^{\gamma - 1} \\ &\leqslant (1 + \gamma p/(1-p)) \exp(-\gamma p) \\ &= n^{o(1) - 1}. \end{aligned} \tag{3.8}$$

Since $m \leqslant m_{\inf} = O(n^\theta)$ for a fixed $\theta < 1$, the assertion follows from (3.8) and the union bound. □

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Let G^* be test design obtained from $G = G_{n,m}$ by deleting all tests of degree larger than Γ. If indeed every test of degree at least Γ contains at least two infected individuals, then $V_{0+}(G^*, \boldsymbol{\chi}) = V_{0+}(G, \boldsymbol{\chi})$ and $V_{1+}(G^*, \boldsymbol{\chi}) = V_{1+}(G, \boldsymbol{\chi})$. Hence Lemma 3.4 shows that it suffices to bound $|V_{0+}(G^*, \boldsymbol{\chi})|, |V_{1+}(G^*, \boldsymbol{\chi})|$. To this end we observe that G^* contains few individuals of very high degree.

Lemma 3.5. *There is* $n_0 = n_0(\theta, \varepsilon) > 0$ *such that, for all* $n > n_0$ *and all test designs* G *with* $m \leqslant m_{\inf}$, *we have*

$$|\{x \in V_n : |\partial_{G^*} x| > \ln^3 n\}| \leqslant \frac{n \ln \ln n}{\ln n}.$$

Proof. Since $\max_{a \in F_m} |\partial_{G^*} a| \leqslant \Gamma = n^{1-\theta} \ln n$, double counting yields

$$\sum_{x \in V_n} |\partial_{G^*} x| = \sum_{a \in F_m} |\partial_{G^*} a| \leqslant m_{\inf} \Gamma = O(n \ln^2 n).$$

Consequently there are no more than $O(n/\ln n)$ individuals $x \in V_n$ with $|\partial_{G^*} x| > \ln^3 n$. □

Further, obtain $G^{(0)}$ from G^* by deleting all individuals of degree greater than $\ln^3 n$ (but keeping all tests). Then the degrees of $G^{(0)}$ satisfy

$$\max_{a \in F(G^{(0)})} |\partial_{G^{(0)}} a| \leqslant \Gamma, \qquad \max_{x \in V(G^{(0)})} |\partial_{G^{(0)}} x| \leqslant \ln^3 n. \tag{3.9}$$

Let $\boldsymbol{\chi}^{(0)} = (\boldsymbol{\chi}_x)_{x \in V(G^{(0)})}$ signify the restriction of $\boldsymbol{\chi}$ to the individuals that remain in $G^{(0)}$. With these preparations in place we are ready to commence the main step of the proof of Proposition 3.1. Given a test design G with $m \leqslant (1-\varepsilon) m_{\inf}$, we are going to construct a sequence $y_1, y_2, \ldots, y_N$, $N = \lceil n^{1-\xi} \rceil$, of individuals of $G^{(0)}$ such that each y_i individually has a moderately high probability of being disguised. Of course, to conclude that in the end a large number of disguised y_i actually materialize, we need to cope with stochastic dependences. To this end we will pick individuals y_i that have pairwise distance at least five in $G^{(0)}$. The degree bounds (3.9) guarantee a sufficient supply of such far apart individuals. To be precise, starting from $G^{(0)}$ we construct a sequence of test designs $G^{(1)}, G^{(2)}, \ldots, G^{(N)}$ inductively as follows. For each $i \geqslant 1$ select a variable $y_{i-1} \in V(G^{(i-1)})$ whose probability of being disguised is maximum; ties are broken arbitrarily. In formulas,

$$\mathbb{P}[y_{i-1} \in V_+(G^{(i-1)}, \boldsymbol{\chi}^{(i-1)})] = \max_{y \in V(G^{(i-1)})} \mathbb{P}[y \in V_+(G^{(i-1)}, \boldsymbol{\chi}^{(i-1)})],$$

where, of course, $\boldsymbol{\chi}^{(i-1)}$ is the only random object. Then obtain $G^{(i)}$ from $G^{(i-1)}$ by removing y_{i-1} along with all vertices (*i.e.* tests or individuals) at distance at most four from y_{i-1}. Moreover, let $\boldsymbol{\chi}^{(i)}$ denote the restriction $(\boldsymbol{\chi}_x)_{x \in V(G^{(i)})}$ of $\boldsymbol{\chi}$ to $G^{(i)}$. The following lemma estimates the probability of y_i being disguised. Let $m^* = |F(G^*)|$ be the total number of tests of G of degree at most Γ.

Lemma 3.6. *There exists* $n_0 = n_0(\varepsilon, \theta, \xi)$ *such that, for all* $n > n_0$ *and all* G *with* $m \leqslant (1-\varepsilon) m_{\inf}$, *we have*

$$\min_{1 \leqslant i \leqslant N} \mathbb{P}[y_i \in V_+(G^{(i)})] \geqslant \exp\left(-\frac{m \ln^2 2}{n^{\theta}} - 1\right).$$

The proof of Lemma 3.6 requires three intermediate steps. First we need a lower bound on the number of individuals in $G^{(i)}$. Recall that $N = \lceil n^{1-\xi} \rceil$.

Claim 1. *We have* $\min_{0 \leqslant i \leqslant N} |V(G^{(i)})| \geqslant n - N\Gamma^2 \ln^6 n$.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Proof. Since throughout the construction of the $G^{(i)}$ we only delete vertices, the degree bound (3.9) implies

$$\max_{a\in F(G^{(i)})} |\partial_{G^{(i)}} a| \leqslant \Gamma = n^{1-\theta}\ln n, \qquad \max_{x\in V(G^{(i)})} |\partial_{G^{(i)}} x| \leqslant \ln^3 n \quad \text{for all } i\leqslant N. \tag{3.10}$$

We now proceed by induction on i. For $i=0$ there is nothing to show. Going from i to $i+1\leqslant N$, we notice that because all individuals $x\in V(G^{(i)})\setminus V(G^{(i+1)})$ have distance at most four from y_{i+1}, (3.10) ensures that

$$|V(G^{(i)})\setminus V(G^{(i+1)})| \leqslant \Gamma^2 \ln^6 n. \tag{3.11}$$

Iterating (3.11), we obtain $|V(G^{(0)})\setminus V(G^{(i+1)})| \leqslant (i+1)\Gamma^2\ln^6 n$, whence $|V(G^{(i+1)})| \geqslant n-(i+1)\Gamma^2\ln^6 n$. □

The following claim resembles the proof of [4, Theorem 1] (where the case $k=\Omega(n)$ is considered).

Claim 2. *Let $\mathcal{D}^{(i)}(x)=\{x\in V_+(G^{(i)})\}$ and let*

$$L^{(i)} = \frac{1}{|V(G^{(i)})|}\sum_{x\in V(G^{(i)})} \ln \mathbb{P}[\mathcal{D}^{(i)}(x)]. \tag{3.12}$$

Then

$$L^{(i)} \geqslant \frac{|F(G^{(i)})|}{|V(G^{(i)})|} \min_{a\in F(G^{(i)})} |\partial_{G^{(i)}} a| \ln(1-(1-p)^{|\partial_{G^{(i)}} a|-1}). \tag{3.13}$$

Proof. For an individual $x\in V(G^{(i)})$ and a test $a\in\partial_{G^{(i)}} x$, let $\mathcal{D}^{(i)}(x,a)$ be the event that there is another individual $z\in\partial_{G^{(i)}} a\setminus\{x\}$ such that $\boldsymbol{\chi}_z=1$. Then for every $x\in V(G^{(i)})$ we have

$$\mathbb{P}[\mathcal{D}^{(i)}(x)] = \mathbb{P}\left[\bigcap_{a\in\partial_{G^{(i)}} x} \mathcal{D}^{(i)}(x,a)\right]. \tag{3.14}$$

Furthermore, the events $\mathcal{D}^{(i)}(x,a)$ are increasing with respect to $\boldsymbol{\chi}$ because having more infected individuals can only increase the probability of being disguised. As the FKG inequality shows that a family of increasing events are mutually positively correlated, we have

$$\mathbb{P}[\mathcal{D}^{(i)}(x)] \geqslant \prod_{a\in\partial_{G^{(i)}} x} \mathbb{P}[\mathcal{D}^{(i)}(x,a)]. \tag{3.15}$$

Moreover, because each entry of $\boldsymbol{\chi}$ is one with probability p independently, we obtain

$$\mathbb{P}[\mathcal{D}^{(i)}(x,a)] = 1-(1-p)^{|\partial_{G^{(i)}} a|-1}. \tag{3.16}$$

Finally, combining (3.14)–(3.16), we obtain

$$\begin{aligned}|V(G^{(i)})|L^{(i)} &\geqslant \sum_{x\in V(G^{(i)})}\sum_{a\in F(G^{(i)})} \mathbf{1}\{a\in\partial_{G^{(i)}} x\}\ln(1-(1-p)^{|\partial_{G^{(i)}} a|-1})\\ &= \sum_{a\in F(G^{(i)})} |\partial_{G^{(i)}} a|\ln(1-(1-p)^{|\partial_{G^{(i)}} a|-1})\\ &\geqslant |F(G^{(i)})| \min_{a\in F(G^{(i)})} |\partial_{G^{(i)}} a|\ln(1-(1-p)^{|\partial_{G^{(i)}} a|-1}),\end{aligned}$$

as claimed. □

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

As a final preparation for the proof of Lemma 3.6 we need the following estimate. Since it is suboptimal in a group testing scheme to test specific individuals separately (see *e.g.* the argument in [6]), we can assume the test size to be of size 2 or larger.

Claim 3. *Given $p=p(n)$ as in (3.4), let $z^*=z^*(p,n)$ be the unique minimum of the function $z\in[2,\infty)\mapsto z\ln(1-(1-p)^{z-1})$. Then we find $z^*=(1+O(n^{-\Omega(1)}))\ln(2)/p$.*

Proof. We consider three separate cases.

Case 1: $z=o(1/p)$. We obtain

$$\begin{aligned} z\ln(1-(1-p)^{z-1}) &= z\ln(1-\exp(-pz+O(p^2z))) \\ &= z\ln(1-(1-pz+O(p^2z^2))) \\ &= \frac{z}{\ln}(zp+O(zp)^2) \\ &= o(1/p). \end{aligned} \tag{3.17}$$

Case 2: $z=\omega(1/p)$. We find

$$\begin{aligned} z\ln(1-(1-p)^{z-1}) &= z\ln(1-\exp\,(-pz+O(p^2z))) \\ &= -z(\exp(-\,pz)+O(\exp(-\,2pz))) \\ &= -\frac{1}{p}pz(\exp(-pz)+\exp(-2pz)) \\ &= o(1/p). \end{aligned} \tag{3.18}$$

Case 3: $z=\Theta(1/p)$. Letting $d=zp$, we obtain

$$\begin{aligned} z\ln(1-(1-p)^{z-1}) &= \frac{d}{p}\ln(1-\exp(-d+O(p))) \\ &= \frac{d}{p}\ln(1-\exp(-d))+O(1). \end{aligned} \tag{3.19}$$

Since the function $d\in(0,\infty)\mapsto d\ln(1-\exp(-\,d))$ attains its unique minimum at $d=\ln 2$, (3.19) dominates (3.17) and (3.18). Thus the minimizer reads $z=\ln(2)/p+O(p^{-1/2})$. □

Proof of Lemma 3.6. Combining Claims 2 and 3, we see that for all test designs G with $m\leqslant(1-\varepsilon)m_{\inf}$ and for all $i\leqslant N$,

$$L^{(i)}\geqslant-(1+O(n^{-\Omega(1)}))\frac{|F(G^{(i)})|\ln^2 2}{|V(G^{(i)})|p}\geqslant-(1+O(n^{-\Omega(1)}))\frac{m\ln^2 2}{|V(G^{(i)})|p}.$$

Hence Claim 1, (3.3) and the choice $p=(k+\sqrt{k}\ln n)/n$ imply that for all $i\leqslant N$

$$L^{(i)}\geqslant-(1+O(n^{-\Omega(1)}))\frac{m\ln^2 2}{(n-N\Delta^2\ln^6 n)p}\geqslant-(1+O(n^{-\Omega(1)}))\frac{m\ln^2 2}{n^\theta}. \tag{3.20}$$

Further, combining the definition (3.12) of $L^{(i)}$ with (3.20), we conclude that for every $i\leqslant N$ there exists an individual $y_i\in V(G^{(i)})$ such that

$$\mathbb{P}[y_i\in V_+(G^{(i)})]=\mathbb{P}[\mathcal{D}^{(i)}(y_i)]\geqslant\exp\,(L^{(i)})\geqslant\exp\left(-(1+O(n^{-\Omega(1)}))\frac{m\ln^2(2)}{n^\theta}\right),$$

which implies the assertion. □

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Lemma 3.6 implies the following bound on $|V_{0+}(G^*, \boldsymbol{\chi})|, |V_{1+}(G^*, \boldsymbol{\chi})|$.

Corollary 3.7. *There exists $n_0 = n_0(\varepsilon, \theta, \xi)$ such that, for all $n > n_0$ and all $G = G_{n,m}$ with $m \leqslant (1-\varepsilon)m_{\mathrm{inf}}$, we have*

$$\mathbb{P}[|V_{0+}(G^*, \boldsymbol{\chi})| \wedge |V_{1+}(G^*, \boldsymbol{\chi})| < \ln^4 n] < \varepsilon/8.$$

Proof. We observe that $V_+(G^{(i)}, \boldsymbol{\chi}) \subset V_+(G^*, \boldsymbol{\chi})$ for all $i \leqslant N$ because, by construction, for any individual $x \in V(G^{(i)})$ every test $a \in \partial_{G^*} x$ of G^* that x belongs to is still present in $G^{(i)}$. Consequently we obtain the bound

$$\mathbb{P}[x \in V_+(G^*)] \geqslant \mathbb{P}[x \in V(G^{(i)})] \quad \text{for all } i \in [N], x \in V(G^*). \tag{3.21}$$

Combining (3.21) with Lemma 3.6, we obtain

$$\mathbb{P}[y^{(i)} \in V_+(G^*)] \geqslant \exp(-\ln^2(2)n^{-\theta}m - 1) \geqslant \exp(-(1-\varepsilon)\ln^2(2)n^{-\theta}m_{\mathrm{inf}} - 1) \quad \text{for all } i \in [N].$$

Hence, recalling the definition of m_{inf} from (1.3), we obtain

$$\mathbb{P}[y^{(i)} \in V_+(G^*)] \geqslant \exp(-(1-\varepsilon)\theta\ln(n) - 1) = n^{(\varepsilon-1)\theta}/\mathrm{e} \quad \text{for all } i \in [N]. \tag{3.22}$$

Since the entry $\boldsymbol{\chi}_{y^{(i)}}$ is independent of the event $\{y^{(i)} \in V_+(G^*)\}$, the definitions (3.1) of $V_{0+}(G^*, \boldsymbol{\chi})$ and $V_{1+}(G^*, \boldsymbol{\chi})$ and (3.22) yield

$$\mathbb{P}[y^{(i)} \in V_{0+}(G^*, \boldsymbol{\chi})] \geqslant (1-p) \cdot \frac{n^{(\varepsilon-1)\theta}}{\mathrm{e}} \geqslant \frac{n^{\varepsilon\theta-1}}{3},$$

$$\mathbb{P}[y^{(i)} \in V_{1+}(G^*, \boldsymbol{\chi})] \geqslant p \cdot \frac{n^{(\varepsilon-1)\theta}}{\mathrm{e}} \geqslant \frac{n^{\varepsilon\theta-1}}{3} \quad \text{for all } i \in [N],$$

provided n is sufficiently large. Therefore, recalling $N = \lceil n^{1-\xi} \rceil$, we obtain for large enough n,

$$\mathbb{E}|\{y^{(1)}, \ldots, y^{(N)}\} \cap V_{0+}(G^*, \boldsymbol{\chi})| \geqslant n^{\varepsilon\theta-\xi}/3, \quad \mathbb{E}|\{y^{(1)}, \ldots, y^{(N)}\} \cap V_{1+}(G^*, \boldsymbol{\chi})| \geqslant n^{\varepsilon\theta-\xi}/3. \tag{3.23}$$

Further, because the pairwise distances of $y^{(1)}, \ldots, y^{(N)}$ in G^* exceed four, the events $\{y^{(i)} \in V_{0+}(G^*, \boldsymbol{\chi})\}_{i \leqslant N}$ are mutually independent. So are the events $\{y^{(i)} \in V_{1+}(G^*, \boldsymbol{\chi})\}_{i \leqslant N}$. Finally, since (3.3) ensures that $\varepsilon\theta - \xi > 0$, where the assumption of θ being close to 1 comes in, (3.23) and the Chernoff bound yield

$$\mathbb{P}[|\{y^{(1)}, \ldots, y^{(N)}\} \cap V_{0+}(G^*, \boldsymbol{\chi})| \leqslant \ln^2 n] \leqslant \mathbb{P}[\mathrm{Bin}(N, n^{\varepsilon\theta-1}/3) \leqslant \ln^2 n] \leqslant \exp(-n^{\Omega(1)}),$$

$$\mathbb{P}[|\{y^{(1)}, \ldots, y^{(N)}\} \cap V_{1+}(G^*, \boldsymbol{\chi})| \leqslant \ln^2 n] \leqslant \mathbb{P}[\mathrm{Bin}(N, n^{\varepsilon\theta-1}/3) \leqslant \ln^2 n] \leqslant \exp(-n^{\Omega(1)}),$$

whence the assertion is immediate. □

Proof of Proposition 3.1. Suppose that $n > n_0(\varepsilon, \theta, \xi)$ is large enough and let $G = G_{n,m}$ be a test design with $m \leqslant (1-\varepsilon)m_{\mathrm{inf}}$ tests. If for every test $a \in F_m$ of degree $|\partial_G a| > \Gamma$ we have $|\partial_G a \cap V_1(G, \boldsymbol{\chi})| \geqslant 2$, then $V_{0+}(G, \boldsymbol{\chi}) = V_{0+}(G^*, \boldsymbol{\chi})$ and $V_{1+}(G, \boldsymbol{\chi}) = V_{1+}(G^*, \boldsymbol{\chi})$. Therefore the assertion is an immediate consequence of Lemma 3.3, Lemma 3.4 and Corollary 3.7. □

3.3 Proof of Proposition 3.2

Given $\varepsilon > 0$ and $\ln(2)/(1+\ln(2)) \leqslant \theta < \theta' < 1$, we choose a large enough $n_0 = n_0(\varepsilon, \theta, \theta')$ and assume that $n > n_0$. Furthermore, let G be a test design with $m \leqslant (1-\varepsilon)m_{\mathrm{inf}}(n, \theta)$ for the purpose of identifying $k = \lceil n^\theta \rceil$ infected individuals. Starting from the test design G infection for density θ, we are going to construct a random test design $\boldsymbol{G}'$ for infection density θ' with the same number m of tests as G. The following lemma fixes the order of $\boldsymbol{G}'$.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Lemma 3.8. *There exists an integer* $n^{\theta/\theta'}/2 \leqslant n' \leqslant 2n^{\theta/\theta'} \wedge n$ *such that* $k' = \lceil n'^{\,\theta'} \rceil = k$.

Proof. Let $n'' = \lceil n^{\theta/\theta'}/2 \rceil$. Then $(4n'')^{\theta'} > k$ but $n''^{\,\theta'} < k$ because the function $z \in (1, \infty) \mapsto z^{\theta'}$ has derivative less than one. For the same reason, for any integer $n'' < N < 4n''$ we have $(N+1)^{\theta'} - N^{\theta'} \leqslant 1$ and thus

$$\lceil (N+1)^{\theta'} \rceil - \lceil N^{\theta'} \rceil \leqslant 1.$$

Consequently there exists an integer $n' \in (n'', 4n'')$ such that $\lceil n'^{\,\theta'} \rceil = k$. □

Given the test design G with individuals $V_n = \{x_1, \dots, x_n\}$ and tests $F_m = \{a_1, \dots, a_m\}$, we now construct the test design $\boldsymbol{G}'$ as follows. Choose a subset $V(\boldsymbol{G}') \subset V_n$ of n' individuals uniformly at random. Then $\boldsymbol{G}'$ is the subgraph that G induces on $V(\boldsymbol{G}') \cup F_m$. Thus $\boldsymbol{G}'$ has the same tests as G but we simply leave out from every test the individuals that do not belong to the random subset $V(\boldsymbol{G}')$. Let $\boldsymbol{\tau} \in \{0,1\}^{V(\boldsymbol{G}')}$ be a random vector of Hamming weight k and let $\hat{\boldsymbol{\tau}} \in \{0,1\}^{F_m}$ be the induced vector of test results

$$\hat{\boldsymbol{\tau}}_a = \max_{x \in \partial_{\boldsymbol{G}'} a} \boldsymbol{\tau}_x \qquad (a \in F_m).$$

Lemma 3.9. *For any integer* $t > 0$ *we have*

$$\mathbb{P}[Z_k(G, \hat{\boldsymbol{\sigma}}) \geqslant t] \geqslant \mathbb{P}[Z_k(\boldsymbol{G}', \hat{\boldsymbol{\tau}}) \geqslant t].$$

Proof. The choice of n' ensures that $k' = \lceil n'^{\,\theta'} \rceil = k$. Therefore the random sets $\{x \in V : \boldsymbol{\sigma}_x = 1\}$ and $\{x \in V(\boldsymbol{G}') : \boldsymbol{\tau}_x = 1\}$ are identically distributed. Indeed, we obtain the latter by first choosing the random subset $V(\boldsymbol{G}')$ of V_n and then choosing a random subset of $V(\boldsymbol{G}')$ size k. Clearly this two-step procedure is equivalent to just choosing a random subset of size k out of V_n. Hence we can couple $\boldsymbol{\sigma}, \boldsymbol{\tau}$ such that the sets $\{x \in V : \boldsymbol{\sigma}_x = 1\}$, $\{x \in V : \boldsymbol{\tau}_x = 1\}$ are identical. Then the construction of $\boldsymbol{G}'$ ensures that the vectors $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\tau}}$ coincide as well.

Now consider a vector $\sigma' \in \mathcal{S}_k(\boldsymbol{G}', \hat{\boldsymbol{\tau}})$ that explains the test results. Extend σ' to a vector $\sigma \in \{0,1\}^{V_n}$ by setting $\sigma_x = 0$ for all $x \in V_n \setminus V(\boldsymbol{G}')$. Then $\sigma \in \mathcal{S}_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$. Hence $Z_k(G, \hat{\boldsymbol{\sigma}}) \geqslant Z_k(\boldsymbol{G}', \hat{\boldsymbol{\tau}})$. □

Proof of Proposition 3.2. Lemma 3.9 shows that for any $t > 0$

$$\mathbb{P}[Z_k(G, \hat{\boldsymbol{\sigma}}) \geqslant t] \geqslant \mathbb{P}[Z_k(\boldsymbol{G}', \hat{\boldsymbol{\tau}}) \geqslant t] = \mathbb{E}[\mathbb{P}[Z_k(\boldsymbol{G}', \hat{\boldsymbol{\tau}}) \geqslant t \mid \boldsymbol{G}']].$$

Consequently there exists an outcome G' of $\boldsymbol{G}'$ such that $\mathbb{P}[Z_k(G, \hat{\boldsymbol{\sigma}}) \geqslant t] \geqslant \mathbb{P}[Z_k(G', \hat{\boldsymbol{\tau}}) \geqslant t]$. □

4. The non-adaptive group testing algorithm `SPIV`

In this section we describe the new test design and the associated inference algorithm `SPIV` for Theorem 1.2. Section 4.1 recaps the random bipartite testing scheme and state-of-the-art decoding algorithms using this scheme, while Section 4.2 introduces the new spatially coupled test design which comes with a polynomial-time decoding algorithm matching the information-theoretic lower bound. Subsequently, Section 4.3 introduces the decoding algorithm and gives an outline for proving its performance guarantees. Finally, Sections 4.4–4.10 contain the proofs of the assertions stated in the outline. Throughout we fix $\theta \in (0,1)$ and $\varepsilon > 0$, and we tacitly assume that $n > n_0(\varepsilon, \theta)$ is large enough for the various estimates to hold.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

4.1 The random bipartite graph and the DD algorithm

To motivate the new test design we begin with a brief discussion of the plain random design used in prior work and the best previously known inference algorithm DD [12, 24]. At first glance a promising candidate test design appears to be a random bipartite graph with one vertex class $V_n = \{x_1, \ldots, x_n\}$ representing individuals and the other class $F_m = \{a_1, \ldots, a_m\}$ representing tests. Indeed, two slightly different random graph models have been proposed [6]. First, in the *Bernoulli model* each V_n–F_m edge is present with a certain probability (the same for every pair) independently of all others. However, due to the relatively heavy lower tail of the degrees of the individuals, this test design turns out to be inferior to a second model where the degrees of the individuals are fixed. Specifically, in the Δ*-out model* every individual independently joins an equal number of Δ tests drawn uniformly at random without replacement [31]. Clearly, in order to extract the maximum amount of information, Δ should be chosen so as to maximize the entropy of the vector of test results. Specifically, since the average test degree equals $\Delta n/m$ and a total of k individuals are infected, the average number of infected individuals per test comes to $\Delta k/m$. Indeed, since $k \sim n^\theta$ for a fixed $\theta < 1$, the number of infected individuals in test a_i can be well approximated by a Poisson variable. Therefore, setting

$$\Delta \sim \frac{m}{k} \ln 2 \tag{4.1}$$

ensures that about half the tests are positive with high probability. With respect to the performance of the Δ-out model, [12, Theorem 1.1] implies together with Theorem 1.1 that this simple construction is information-theoretically optimal. Indeed, $m = (1 + \varepsilon + o(1))m_{\inf}$ tests suffice so that an exponential-time algorithm correctly infers the set of infected individuals. Specifically, the algorithm solves a minimum hypergraph vertex cover problem with the individuals as the vertex set and the positive test groups as the hyperedges. For $m = (1 + \varepsilon + o(1))m_{\inf}$ the unique optimal solution is precisely the correct set of infected individuals with high probability. While the worst-case NP-hardness of hypergraph vertex cover does not, of course, preclude the existence of an algorithm that is efficient on random hypergraphs, despite considerable efforts no such algorithm has been found. In fact, as we saw in Section 1.4, for a good number of broadly similar inference and optimization problems on random graphs no efficient information-theoretically optimal algorithms are known. But for m exceeding the threshold m_{DD} from (1.2), an efficient greedy algorithm DD correctly recovers $\boldsymbol{\sigma}$ with high probability. The algorithm proceeds in three steps.

DD1 Declare every individual that appears in a negative test uninfected and subsequently remove all negative tests and all individuals that they contain.

DD2 For every remaining (positive) test of degree one, declare the individual that appears in the test infected.

DD3 Declare all other individuals as uninfected.

The decisions made by the first two steps **DD1**–**DD2** are clearly correct, but **DD3** might produce false negatives. Prior to the present work DD was the best known polynomial-time group testing algorithm. While DD correctly identifies the set of infected individuals with high probability if $m > (1 + \varepsilon)m_{\mathrm{DD}}$ [24], the algorithm fails if $m < (1 - \varepsilon)m_{\mathrm{DD}}$ with high probability [12].

4.2 Spatial coupling

The new efficient algorithm SPIV for Theorem 1.2 that gets by with the optimal number $(1 + \varepsilon + o(1))m_{\inf}$ of tests comes with a tailor-made test design that, inspired by spatially coupled codes [19, 28, 29], combines randomization with a superimposed geometric structure. Specifically, we divide both the individuals and the tests into

$$\ell = \lceil \ln^{1/2} n \rceil \tag{4.2}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

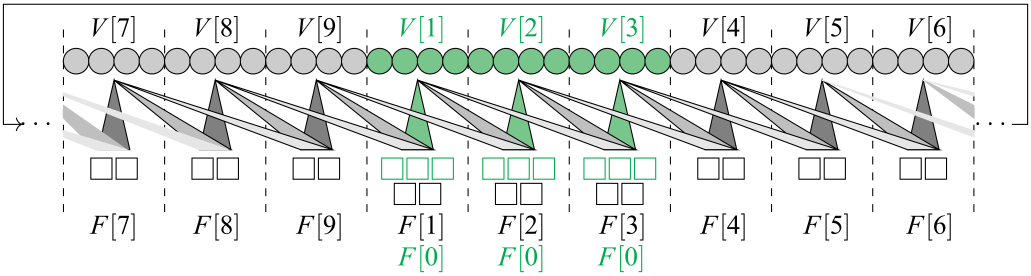

Figure 2. The spatially coupled test design with $n = 36, \ell = 9, s = 3$. The individuals in the seed groups $V[1] \cup \cdots \cup V[s]$ (green) are equipped with additional test $F[0]$ (green rectangles). The black rectangles represent the tests $F[1] \cup \cdots \cup F[\ell]$.

compartments of equal size. The compartments are arranged along a ring, and each individual joins an equal number of random tests in the

$$s = \lceil \ln \ln n \rceil = o(\ell) \tag{4.3}$$

topologically subsequent compartments. Additionally, to get the algorithm started we equip the first s compartments with extra tests so that they can be easily diagnosed via the DD algorithm. Then, having diagnosed the initial compartments correctly, SPIV will work its way along the ring, diagnosing one compartment after the other.

To implement this idea precisely we partition the set $V = V_n = \{x_1, \ldots, x_n\}$ of individuals into pairwise disjoint subsets $V[1], \ldots, V[\ell]$ of sizes $|V[j]| \in \{\lfloor n/\ell \rfloor, \lceil n/\ell \rceil\}$. With each compartment $V[i]$ of individuals we associate a compartment $F[i]$ of tests of size $|F[i]| = m/\ell$ for an integer m that is divisible by ℓ. Additionally, we introduce a set $F[0]$ of $10\lceil (ks/\ell) \ln n \rceil$ extra tests to facilitate the greedy algorithm for diagnosing the first s compartments. Thus the total number of tests comes to

$$|F[0]| + \sum_{i=1}^{\ell} |F[i]| = (1 + O(s/\ell))m = (1 + o(1))m. \tag{4.4}$$

Finally, for notational convenience we define $V[\ell + i] = V[i]$ and $F[\ell + i] = F[i]$ for $i = 1, \ldots, s$. The test groups are composed as follows: let

$$k = \lceil n^{\theta} \rceil \quad \text{and let } \Delta = \frac{m \ln 2}{k} + O(s) \tag{4.5}$$

be an integer divisible by s; see (4.1). Then we construct a random bipartite graph as follows.

SC1 For $i = 1, \ldots, \ell$ and $j = 1, \ldots, s$ every individual $x \in V[i]$ joins Δ/s tests from $F[i + j - 1]$ chosen uniformly at random without replacement. The choices are mutually independent for all individuals x and all j.

SC2 Additionally, each individual from $V[1] \cup \cdots \cup V[s]$ independently joins $\lceil 10 \ln(2) \ln n \rceil$ random tests from $F[0]$, drawn uniformly without replacement.

Thus **SC1** provides that the individuals in compartment $V[i]$ take part in the next s compartments $F[i], \ldots, F[i + s - 1]$ of tests along the ring. Furthermore, **SC2** supplies the tests required by the DD algorithm to diagnose the first s compartments. Figure 2 provides an illustration of the resulting random test design. From here on the test design produced by **SC1–SC2** is denoted by $\mathbf{G}$. Furthermore $\boldsymbol{\sigma} \in \{0, 1\}^V$ denotes a uniformly random vector of Hamming weight k, drawn independently of $\mathbf{G}$, and $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}_a)_{a \in F[0] \cup \cdots \cup F[\ell]}$ signifies the vector of test results

$$\hat{\boldsymbol{\sigma}}_a = \max_{x \in \partial a} \boldsymbol{\sigma}_x.$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

In addition, let $V_1 = \{x \in V : \boldsymbol{\sigma}_x = 1\}$ be the set of infected individuals and let $V_0 = V \setminus V_1$ be the set of healthy individuals. Moreover, let $F = F[0] \cup F[1] \cup \cdots \cup F[\ell]$ be the set of all tests, let $F_1 = \{a \in F : \hat{\boldsymbol{\sigma}}_a = 1\}$ be the set of all positive tests and let $F_0 = F \setminus F_1$ be the set of all negative tests. Finally, let

$$V_0[i] = V[i] \cap V_0, \quad V_1[i] = V[i] \cap V_1, \quad F_0[i] = F[i] \cap F_0, \quad F_1[i] = F[i] \cap F_1.$$

The following proposition summarizes a few basic properties of the test design $\mathbf{G}$.

Proposition 4.1. *If $m = \Theta(n^\theta \ln n)$, then $\mathbf{G}$ enjoys the following properties with probability $1 - o(n^{-2})$.*

(i) *The infected individual counts in the various compartments satisfy*

$$\frac{k}{\ell} - \sqrt{\frac{k}{\ell}} \ln n \leqslant \min_{i \in [\ell]} |V_1[i]| \leqslant \max_{i \in [\ell]} |V_1[i]| \leqslant \frac{k}{\ell} + \sqrt{\frac{k}{\ell}} \ln n.$$

(ii) *For all $i \in [\ell]$ and all $j \in [s]$, the test degrees satisfy*

$$\frac{\Delta n}{ms} - \sqrt{\frac{\Delta n}{ms}} \ln n \leqslant \min_{a \in F[i+j-1]} |V[i] \cap \partial a| \leqslant \max_{a \in F[i+j-1]} |V[i] \cap \partial a| \leqslant \frac{\Delta n}{ms} + \sqrt{\frac{\Delta n}{ms}} \ln n.$$

(iii) *For all $i \in [\ell]$, the number of negative tests in compartment $F[i]$ satisfies*

$$\frac{m}{2\ell} - \sqrt{m} \ln^3 n \leqslant |F_0[i]| \leqslant \frac{m}{2\ell} + \sqrt{m} \ln^3 n.$$

We prove Proposition 4.1 in Section 4.4. Finally, as a preparation for things to come, we point out that, for any specific individual $x \in V[i]$ and any particular test $a \in F[i+j]$, $j = 0, \ldots, s-1$, we have

$$\mathbb{P}[x \in \partial a] = 1 - \mathbb{P}[x \notin \partial a] = 1 - \binom{|F[i+j]| - 1}{\Delta/s}\binom{|F[i+j]|}{\Delta/s}^{-1} = \frac{\Delta \ell}{ms} + O\left(\left(\frac{\Delta \ell}{ms}\right)^2\right). \tag{4.6}$$

4.3 The spatial inference vertex cover (‘SPIV’) algorithm

The SPIV algorithm for Theorem 1.2 proceeds in three phases. The plan of attack is for the algorithm to work its way along the ring, diagnosing one compartment after the other aided by what has been learned about the preceding compartments. Of course, we need to start somewhere. Hence in its first phase SPIV diagnoses the seed compartments $V[1], \ldots, V[s]$.

4.3.1 Phase 1: the seed

Specifically, the first phase of SPIV applies the DD greedy algorithm from Section 4.1 to the subgraph of $\mathbf{G}$ induced on the individuals $V[1] \cup \cdots \cup V[s]$ and the tests $F[0]$. Throughout the vector $\tau \in \{0, 1\}^V$ signifies the algorithm’s current estimate of the ground truth $\boldsymbol{\sigma}$.

Algorithm 1: SPIV, phase 1.

```
Input: $\mathbf{G}$, $\hat{\boldsymbol{\sigma}}$
Output: an estimate of $\boldsymbol{\sigma}$
Let $(\tau_x)_{x \in V[1] \cup \cdots \cup V[s]} \in \{0, 1\}^{V[1] \cup \cdots \cup V[s]}$ be the result of applying DD to the tests $F[0]$;
Set $\tau_x = 0$ for all individuals $x \in V \setminus (V[1] \cup \cdots \cup V[s])$;
```

The following proposition, whose proof can be found in Section 4.5, summarizes the analysis of phase 1.

Proposition 4.2. *With high probability the output of* DD *satisfies* $\tau_x = \boldsymbol{\sigma}_x$ *for all* $x \in V[1] \cup \cdots \cup V[s]$.

4.3.2 Phase 2: enter the ring

This is the main phase of the algorithm. Thanks to Proposition 4.2 we may assume that the seed has been diagnosed correctly. Now, the program is to diagnose one compartment after the other, based on what the algorithm learned previously. Hence, assume that we managed to diagnose compartments $V[1], \ldots, V[i]$ correctly. How do we proceed to compartment $V[i+1]$? For a start, we can safely mark as uninfected all individuals in $V[i+1]$ that appear in a negative test. But a simple calculation reveals that this will still leave us with many more than k undiagnosed individuals with high probability. To be precise, consider the set of uninfected disguised individuals

$$V_{0+}[i+1] = \{x \in V_0[i+1] : \hat{\boldsymbol{\sigma}}_a = 1 \text{ for all } a \in \partial x\},$$

that is, uninfected individuals that fail to appear in a negative test. In Section 4.6 we prove the following.

Lemma 4.3. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \leqslant m = O(n^{\theta} \ln n)$. *Then with high probability for all* $s \leqslant i < \ell$ *we have*

$$|V_{0+}[i+1]| = (1 + O(n^{-\Omega(1)}))\frac{n}{\ell 2^{\Delta}}.$$

Hence by the definition (4.5) of Δ for m close to m_{inf} the set $V_{0+}[i+1]$ has size $k^{1+\Omega(1)} \gg k$ with high probability. Thus the challenge is to discriminate between $V_{0+}[i+1]$ and the set $V_1[i+1]$ of actual infected individuals in compartment $i+1$. The key observation is that we can tell these sets apart by counting currently 'unexplained' positive tests. To be precise, for an individual $x \in V[i+1]$ and $1 \leqslant j \leqslant s$, let $\boldsymbol{W}_{x,j}$ be the number of tests in compartment $F[i+j]$ that contain x but do not contain an infected individual from the preceding compartments $V[1] \cup \cdots \cup V[i]$. In formulas,

$$\boldsymbol{W}_{x,j} = |\{a \in \partial x \cap F[i+j] : \partial a \cap (V_1[1] \cup \cdots \cup V_1[i]) = \emptyset\}|. \tag{4.7}$$

Crucially, the following back-of-the-envelope calculation shows that the mean of this random variable depends on whether x is infected or healthy but disguised.

Infected individuals ($x \in V_1[i+1]$). Consider a test $a \in \partial x \cap F[i+j]$, $j = 1, \ldots, s$. Because the individuals join tests independently, conditioning on x being infected does not skew the distribution of the individuals from the $s-j$ prior compartments $V[i+j-s+1], \ldots, V[i]$ that appear in a. Furthermore, we chose Δ so that for each of these compartments $V[h]$ the expected number of infected individuals that join a has mean $(\ln 2)/s$. Indeed, due to independence it is not difficult to see that $|V_1[h] \cap \partial a|$ is approximately a Poisson variable. Consequently

$$\mathbb{P}[(V_1[i+j-s+1] \cup \cdots \cup V_1[i]) \cap \partial a = \emptyset] \sim 2^{-(s-j)/s}. \tag{4.8}$$

Hence, because x appears in Δ/s tests $a \in F[i+j]$, the linearity of expectation yields

$$\mathbb{E}[\boldsymbol{W}_{x,j} \mid x \in V_1[i+1]] \sim 2^{j/s-1}\frac{\Delta}{s}. \tag{4.9}$$

Disguised healthy individual ($x \in V_{0+}[i+1]$). As above, for any individual $x \in V[i+1]$ and any $a \in \partial x \cap F[i+j]$, the *unconditional* number of infected individuals in a is asymptotically $\mathrm{Po}(\ln 2)$. But given $x \in V_{0+}[i+1]$ we know that a is positive. Thus $\partial a \setminus \{x\}$ contains at least one infected

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

individual. In effect, the number of positives in a approximately turns into a conditional Poisson $\mathrm{Po}_{\geqslant 1}(\ln 2)$. Consequently, for test a not to include any infected individual from one of the known compartments $V[h]$, $h = i + j - s + 1, \ldots, i$, every infected individual in test a must stem from the j yet undiagnosed compartments. Summing up the conditional Poisson and recalling that x appears in Δ/s tests $a \in F[j]$, we thus obtain

$$\mathbb{E}[\boldsymbol{W}_{x,j} \mid x \in V_{0+}[i+1]] \sim \frac{\Delta}{s} \sum_{t \geqslant 1} \mathbb{P}[\mathrm{Po}_{\geqslant 1}(\ln 2) = t](j/s)^t = (2^{j/s} - 1)\frac{\Delta}{s}. \tag{4.10}$$

An initial idea to tell $V_{0+}[i+1]$ and $V_1[i+1]$ apart might thus be to simply calculate

$$\boldsymbol{W}_x = \sum_{j=1}^{s-1} \boldsymbol{W}_{x,j} \quad (x \in V[i+1]). \tag{4.11}$$

Indeed, (4.9) and (4.10) yield

$$\mathbb{E}[\boldsymbol{W}_x \mid x \in V_1[i+1]] \sim \frac{\Delta}{2\ln 2} = 0.721\ldots\ \Delta$$

whereas

$$\mathbb{E}[\boldsymbol{W}_x \mid x \in V_{0+}[i+1]] \sim \frac{\Delta(1-\ln 2)}{\ln 2} = 0.442\ldots\ \Delta.$$

But unfortunately a careful large deviations analysis reveals that $\boldsymbol{W}_x$ is not sufficiently concentrated. More precisely, even for $m = (1 + \varepsilon + o(1))m_{\inf}$ there are as many as $k^{1+\Omega(1)}$ 'outliers' $x \in V_{0+}[i+1]$ whose $\boldsymbol{W}_x$ grows as large as the mean $\Delta/(2\ln 2)$ of actual infected individuals with high probability. On consideration, the plain sum (4.11) does seem to leave something to be desired. While $\boldsymbol{W}_x$ counts all as yet unexplained positive tests equally, not all of these tests reveal the same amount of information. In fact, we should really be paying more attention to 'early' unexplained tests $a \in F[i+1]$ than to 'late' ones $b \in F[i+s]$, for we have already diagnosed $s-1$ out of the s compartments of individuals that a draws on, whereas only one of the s compartments that contribute to b has already been diagnosed. Thus the unexplained test a is a much stronger indication that x might be infected. Consequently, it seems promising to replace $\boldsymbol{W}_x$ with the weighted sum

$$\boldsymbol{W}^{\star}_x = \sum_{j=1}^{s-1} w_j \boldsymbol{W}_{x,j} \tag{4.12}$$

with $w_1, \ldots, w_{s-1} \geqslant 0$ chosen so as to gauge the amount of information carried by the different compartments.

To find the optimal weights $w_1, \ldots, w_{s-1}$, we need to investigate the rate function of $\boldsymbol{W}^{\star}_x$ given $x \in V_{0+}[i+1]$. More specifically, we should minimize the probability that $\boldsymbol{W}^{\star}_x$ given $x \in V_{0+}[i+1]$ grows as large as the mean of $\boldsymbol{W}^{\star}_x$ given $x \in V_1[i+1]$, which we read off (4.9) easily:

$$\mathbb{E}[\boldsymbol{W}^{\star}_x \mid x \in V_1[i+1]] \sim \frac{\Delta}{s} \sum_{j=1}^{s-1} 2^{j/s-1} w_j. \tag{4.13}$$

A careful large deviations analysis followed by a Lagrangian optimization leads to the optimal choice

$$w_j = \ln \frac{(1-2\zeta)2^{j/s-1}(2-2^{j/s})}{(1-(1-2\zeta)2^{j/s-1})(2^{j/s}-1)}, \quad \text{where } \zeta = 1/s^2. \tag{4.14}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

The following two lemmas show that with these weights the scores $\boldsymbol{W}^\star_x$ discriminate well between the potential false positives and the infected individuals. More precisely, thresholding $\boldsymbol{W}^\star_x$ we end up misclassifying no more than $o(k)$ individuals x with high probability. Recall that

$$\Delta = \frac{m \ln 2}{k} + O(s).$$

Lemma 4.4. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \leqslant m = O(n^\theta \ln n)$. *With high probability we have*

$$\sum_{s\leqslant i<\ell}\sum_{x\in V_1[i]} \mathbf{1}\left\{\boldsymbol{W}^\star_x < (1-\zeta/2)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j\right\} \leqslant k\exp\left(-\frac{\Omega(\ln n)}{(\ln\ln n)^4}\right). \tag{4.15}$$

Lemma 4.5. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \leqslant m = O(n^\theta \ln n)$. *With high probability we have*

$$\sum_{s\leqslant i<\ell}\sum_{x\in V_{0+}[i]} \mathbf{1}\left\{\boldsymbol{W}^\star_x > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j\right\} \leqslant k^{1-\Omega(1)}. \tag{4.16}$$

We prove these two lemmas in Sections 4.7 and 4.8. Lemmas 4.4–4.5 leave us with only one loose end. Namely, calculating the scores $\boldsymbol{W}^\star_x$ requires knowledge of the correct infection status $\boldsymbol{\sigma}_x$ of *all* the individuals $x \in V[1]\cup\cdots\cup V[i]$ from the previous compartments. But since the right-hand side expressions in (4.15) and (4.16) are non-zero, it is unrealistic to assume that the algorithm's estimates τ_x will consistently match the ground truth $\boldsymbol{\sigma}_x$ beyond the seed compartments. Hoping that the algorithm's estimate will not stray too far, we thus have to make do with the approximate scores

$$W^\star_x(\tau) = \sum_{j=1}^{s-1} w_j W_{x,j}(\tau), \qquad \text{where } W_{x,j}(\tau) = \left|\left\{a \in \partial x \cap F[i+j-1] : \max_{y\in\partial a\cap(V[1]\cup\cdots V[i])} \tau_y = 0\right\}\right|. \tag{4.17}$$

Hence phase 2 of SPIV reads as follows.

Algorithm 2: SPIV, phase 2.

```
3  for $i = s, \ldots, \ell - 1$ do
4      for $x \in V[i+1]$ do
5          if $\exists a \in \partial x \colon \hat{\boldsymbol{\sigma}}_a = 0$ then
6              $\tau_x = 0$ // classify as uninfected
7          else if $W^\star_x(\tau) < (1-\zeta)(\Delta/s)\sum_{j=1}^{s-1} 2^{j/s-1} w_j$ then
8              $\tau_x = 0$ // tentatively classify as uninfected
9          else
10             $\tau_x = 1$ // tentatively classify as infected
```

Since phase 2 of SPIV uses the approximations from (4.17), there seems to be a risk of errors amplifying as we move along. Fortunately it turns out that errors proliferate only moderately and the second phase of SPIV will misclassify only $o(k)$ individuals. The following proposition summarizes the analysis of phase 2.

Proposition 4.6. *Suppose that* $(1+\varepsilon)m_{\mathrm{ad}} \leqslant m = O(k \ln n)$. *With high probability the assignment* τ *obtained after steps 1–10 satisfies*

$$\sum_{x\in V} \mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leqslant k\exp\left(-\frac{\ln n}{(\ln\ln n)^6}\right).$$

The proof of Proposition 4.6 can be found in Section 4.9.

4.3.3 Phase 3: cleaning up

The final phase of the algorithm rectifies the errors incurred during phase 2. The combinatorial insight that makes this possible is that for $m \geqslant (1+\varepsilon)m_{\inf}$ every infected individual has at least $\Omega(\Delta)$ positive tests to itself with high probability. Thus these tests do not feature a second infected individual. Phase 3 of the algorithm exploits this observation by simply thresholding the number S_x of tests where there is no other infected individual besides potentially x. Thanks to the expansion properties of the graph $\mathbf{G}$, each iteration of the thresholding procedure reduces the number of misclassified individuals by at least a factor of three. In effect, after $\ln n$ iterations all individuals will be classified correctly with high probability. Of course, due to Proposition 4.2 we do not need to reconsider the seed $V[1] \cup \cdots \cup V[s]$.

Algorithm 3: SPIV, phase 3.

```
Let $\tau^{(1)} = \tau$;
for $i = 1, \ldots, \lceil \ln n \rceil$ do
    For all $x \in V[s+1] \cup \cdots \cup V[\ell]$ calculate
        $S_x(\tau^{(i)}) = \sum_{a \in \partial x:\, \hat{\boldsymbol{\sigma}}_a = 1} \mathbf{1}\{\forall y \in \partial a \setminus \{x\}: \tau_y^{(i)} = 0\}$;
    Let $\tau_x^{(i+1)} = \begin{cases} \tau_x^{(i)} & \text{if } x \in V[1] \cup \cdots \cup V[s], \\ \mathbf{1}\{S_x(\tau^{(i)}) > \ln^{1/4} n\} & \text{otherwise} \end{cases}$;
return $\tau^{(\lceil \ln n \rceil)}$
```

Proposition 4.7. *Suppose that* $(1+\varepsilon)m_{\inf} \leqslant m = O(n^{\theta} \ln n)$. *With high probability, for all* $1 \leqslant i \leqslant \lceil \ln n \rceil$ *we have*

$$\sum_{x \in V} \mathbf{1}\{\tau_x^{(i+1)} \neq \boldsymbol{\sigma}_x\} \leqslant \frac{1}{3} \sum_{x \in V} \mathbf{1}\{\tau_x^{(i)} \neq \boldsymbol{\sigma}_x\}.$$

We prove Proposition 4.7 in Section 4.10.

Proof of Theorem 1.2. The theorem is an immediate consequence of Propositions 4.2, 4.6 and 4.7. While Proposition 4.6 accounts for the second term in the maximum of (1.3), the first part is due to Proposition 4.7. □

4.4 Proof of Proposition 4.1

Recall that we need to establish the following three statements.

(i) The infected individual counts in the various compartments satisfy

$$\frac{k}{\ell} - \sqrt{\frac{k}{\ell}} \ln n \leqslant \min_{i \in [\ell]} |V_1[i]| \leqslant \max_{i \in [\ell]} |V_1[i]| \leqslant \frac{k}{\ell} + \sqrt{\frac{k}{\ell}} \ln n.$$

(ii) For all $i \in [\ell]$ and all $j \in [s]$, the test degrees satisfy

$$\frac{\Delta n}{ms} - \sqrt{\frac{\Delta n}{ms}} \ln n \leqslant \min_{a \in F[i+j-1]} |V[i] \cap \partial a| \leqslant \max_{a \in F[i+j-1]} |V[i] \cap \partial a| \leqslant \frac{\Delta n}{ms} + \sqrt{\frac{\Delta n}{ms}} \ln n.$$

(iii) For all $i \in [\ell]$, the number of negative tests in compartment $F[i]$ satisfies

$$\frac{m}{2\ell} - \sqrt{m}\ln^3 n \leqslant |F_0[i]| \leqslant \frac{m}{2\ell} + \sqrt{m}\ln^3 n.$$

The number $|V_1[i]|$ of infected individuals in compartment $V[i]$ has distribution $\mathrm{Hyp}(n, k, |V[i]|)$. Since $||V[i]| - n/\ell| \leqslant 1$, (i) is an immediate consequence of the Chernoff bound from Lemma 2.2. With respect to (ii), we recall from (4.6) that

$$\mathbb{P}[x \in \partial a] = \frac{\Delta\ell}{ms}\left(1 + O\left(\frac{\Delta\ell}{ms}\right)\right).$$

Hence, because the various individuals $x \in V[i]$ join tests independently, the number $|V[i] \cap \partial a|$ of test participants from $V[i]$ has distribution

$$|V[i] \cap \partial a| \sim \mathrm{Bin}(|V[i]|, \Delta\ell/(ms) + O((\Delta\ell/ms)^2)).$$

Since $|V[i]| = n/\ell + O(1)$, assertion (ii) follows from (4.5) and the Chernoff bound from Lemma 2.1.

Coming to (iii), due to part (i) we may condition on

$$\mathcal{E} = \{\forall i \in [\ell] \colon |V_1[i]| = k/\ell + O(\sqrt{k/\ell}\ln n)\}.$$

Hence, with h ranging over the s compartments whose individuals join tests in $F[i]$, (4.6) implies that for every test $a \in F[i]$ the number of infected individuals $|V_1 \cap \partial a|$ is distributed as a sum of independent binomial variables

$$|V_1 \cap \partial a| \sim \sum_h \boldsymbol{X}_h \quad \text{with } \boldsymbol{X}_h \sim \mathrm{Bin}\left(V_1[h], \frac{\Delta\ell}{ms} + O\left(\left(\frac{\Delta\ell}{ms}\right)^2\right)\right).$$

Consequently (4.5) ensures that the event $V_1 \cap \partial a = \emptyset$ has conditional probability

$$\begin{aligned}
\mathbb{P}[V_1 \cap \partial a = \emptyset \mid \mathcal{E}] &= \prod_h \mathbb{P}[\boldsymbol{X}_h = 0 \mid \mathcal{E}] \\
&= \exp\left[s\left(\frac{k}{\ell} + O\left(\sqrt{\frac{k}{\ell}}\ln n\right)\right)\ln\left(1 - \frac{\Delta\ell}{ms} + O\left(\left(\frac{\Delta\ell}{ms}\right)^2\right)\right)\right] \\
&= \exp\left[-\frac{sk}{\ell}\cdot\frac{\Delta\ell}{ms} + O\left(\sqrt{\frac{k}{\ell}}\cdot\frac{\Delta\ell}{m}\right) + O\left(\frac{sk}{\ell}\cdot\left(\frac{\Delta\ell}{ms}\right)^2\right)\right] \\
&= \frac{1}{2} + O(\sqrt{\ell/k}).
\end{aligned}$$

Therefore we obtain the estimate

$$\mathbb{E}[|F_0[i]| \mid \mathcal{E}] = \frac{m}{2\ell} + O(\sqrt{m}\ln n). \tag{4.18}$$

Finally, changing the set of tests that a specific infected individual $x \in V_1[h]$ joins shifts $|F_0[i]|$ by at most Δ (while tinkering with uninfected ones does not change $|F_0[i]|$ at all). Therefore the Azuma–Hoeffding inequality [23, Corollary 2.27] yields

$$\mathbb{P}[||F_0[i]| - \mathbb{E}[|F_0[i]| \mid \mathcal{E}]| \geqslant t \mid \mathcal{E}] \leqslant 2\exp\left(-\frac{t^2}{2k\Delta^2}\right) \quad \text{for any } t > 0. \tag{4.19}$$

Thus (iii) follows from (4.5), (4.18) and (4.19) with $t = \sqrt{m}\ln^3 n$.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

4.5 Proof of Proposition 4.2

Let $D=\lceil 10\ln(2)\ln n\rceil$ and recall that $|F[0]|=\lceil 10ks\ln(n)/\ell\rceil$. Since by **SC2** every individual from $\in V[1]\cup\cdots\cup V[s]$ joins D random tests from $F[0]$, in analogy to (4.6) for every $x\in V[1]\cup\cdots\cup V[s]$ and every test $a\in F[0]$ we obtain

$$\begin{aligned}\mathbb{P}[x\in\partial a]&=1-\mathbb{P}[x\notin\partial a]\\&=1-\binom{|F[0]|-1}{D}\binom{|F[0]|}{D}^{-1}\\&=\frac{D}{|F[0]|}\left(1+O\left(\frac{D}{|F[0]|}\right)\right)\\&=\frac{\ell\ln 2}{ks}(1+O(n^{-\Omega(1)})).\end{aligned}\tag{4.20}$$

Let $F_1[0]$ be the set of tests $a\in F[0]$ with $\hat{\boldsymbol{\sigma}}_a=1$.

Lemma 4.8. *With high probability the number of positive tests $a\in F[0]$ satisfies*

$$|F_1[0]|=|F[0]|\left(\frac{1}{2}+O(n^{-\Omega(1)})\right).$$

Proof. By Proposition 4.1 we may condition on the event $\mathcal{E}$ that

$$|V_1[1]\cup\cdots\cup V_1[s]|=\frac{ks}{\ell}(1+O(n^{-\Omega(1)})).$$

Hence (4.20) implies that, given $\mathcal{E}$, the expected number of infected individuals in a test $a\in F[0]$ comes to

$$\mathbb{E}[|\partial a\cap V_1|\mid\mathcal{E}]=\ln 2+O(n^{-\Omega(1)}).\tag{4.21}$$

Moreover, since individuals join tests independently, $|\partial a\cap V_1|$ is a binomial random variable. Hence (4.21) implies

$$\mathbb{P}[\partial a\cap V_1=\emptyset\mid\mathcal{E}]=\frac{1}{2}+O(n^{-\Omega(1)}).$$

Consequently, since $\mathbb{P}[\mathcal{E}]=1-o(n^{-2})$ by Proposition 4.1,

$$\mathbb{E}|F_1\cap F[0]|=\mathbb{E}|F_1[0]|=\frac{|F[0]|}{2}(1+O(n^{-\Omega(1)})).\tag{4.22}$$

Finally, changing the set ∂x of neighbours of an infected individual can shift $|F_1[0]|$ by at most D. Therefore the Azuma–Hoeffding inequality implies that

$$\mathbb{P}[||F_1[0]|-\mathbb{E}|F_1[0]||>t]\leqslant 2\exp\left(-\frac{t^2}{2D^2k}\right)\quad\text{for any } t>0.\tag{4.23}$$

Since $D=O(\ln n)$, combining (4.22) and (4.23) and setting, say, $t=k^{2/3}$ completes the proof. □

As an application of Lemma 4.8 we show that with high probability every seed individual x appears in a test $a\in F[0]$ whose other individuals are all healthy.

Corollary 4.9. *With high probability every individual $x\in V[1]\cup\cdots\cup V[s]$ appears in a test $a\in F[0]\cap\partial x$ such that $\partial a\setminus\{x\}\subset V_0$.*

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Proof. We expose the random bipartite graph induced on $V[1]\cup\cdots\cup V[s]$ and $F[0]$ in two rounds. In the first round we expose $\boldsymbol{\sigma}$ and all neighbourhoods $(\partial y)_{y\in(V[1]\cup\cdots\cup V[s])\setminus\{x\}}$. In the second round we expose ∂x. Let $\boldsymbol{X}$ be the number of negative tests $a\in F[0]$ after the first round. Since x has degree $D=O(\ln n)$, Lemma 4.8 implies that $\boldsymbol{X}=|F[0]|(\frac{1}{2}+O(n^{-\Omega(1)}))$ with high probability. Furthermore, given $\boldsymbol{X}$, the number of tests $a\in\partial x$ all of whose other individuals are uninfected has distribution $\mathrm{Hyp}(|F[0]|,\boldsymbol{X},D)$. Hence

$$\mathbb{P}[\forall a\in\partial x\colon V_1\cap\partial a\setminus\{x\}\neq\emptyset\mid\boldsymbol{X}]=\binom{|F[0]|-\boldsymbol{X}}{D}\binom{|F[0]|}{D}^{-1}\leqslant\exp(-D\boldsymbol{X}/|F[0]|).\tag{4.24}$$

Assuming

$$\boldsymbol{X}/|F[0]|=\frac{1}{2}+O(n^{-\Omega(1)})$$

and recalling that $D=\lceil 10\ln(2)\ln n\rceil$, we obtain $\exp(-D\boldsymbol{X}/|F[0]|)=o(1/n)$. Thus the assertion follows from (4.24) and the union bound. □

Proof of Proposition 4.2. Due to Corollary 4.9 we may assume that for every $x\in V[1]\cup\cdots\cup V[s]$ there is a test $a_x\in F[0]$ such that $\partial a_x\setminus\{x\}\subset V_0$. Hence, recalling the DD algorithm from Section 4.1, we see that the first step **DD1** will correctly identify all healthy individuals $x\in V_0[1]\cup\cdots\cup V_0[s]$. Moreover, the second step **DD2** will correctly classify all remaining individuals $V_1[1]\cup\cdots\cup V_1[s]$ as infected, and the last step **DD3** will be void. □

4.6 Proof of Lemma 4.3

Let $\mathcal{E}$ be the event that properties (i) and (iii) from Proposition 4.1 hold; then $\mathbb{P}[\mathcal{E}]=1-o(n^{-2})$. Moreover, let $\mathfrak{E}$ be the σ-algebra generated by $\boldsymbol{\sigma}$ and the neighbourhoods $(\partial x)_{x\in V_1}$. Then the event $\mathcal{E}$ is $\mathfrak{E}$-measurable while the neighbourhoods $(\partial x)_{x\in V_0}$ of the healthy individuals are independent of $\mathfrak{E}$. Recalling from **SC1** that the individuals $x\in V_0[i]$ choose Δ/s random tests in each of the compartments $F[i+j]$, $0\leqslant j\leqslant s-1$ independently, and remembering that $x\in V_{0+}[i]$ if and only if none of these tests is negative, on $\mathcal{E}$ we obtain

$$\begin{aligned}\mathbb{P}[x\in V_{0+}[i]\mid\mathfrak{E}]&=\binom{m/(2\ell)+O(\sqrt{m}\ln^3 n)}{\Delta/s}^{s}\binom{m/\ell}{\Delta/s}^{-s}\\&=\left(\frac{1+O(m^{-1/2}\ell\ln^3 n)}{2}\right)^{\Delta}\\&=2^{-\Delta}+O(m^{-1/2}\Delta\ell\ln^3 n)\\&=2^{-\Delta}(1+O(n^{-\theta/2}\ln^4 n))\quad\text{(due to (4.2) and (4.5)).}\end{aligned}\tag{4.25}$$

Because all $x\in V_0[i]$ choose their neighbourhoods independently, (4.25) implies that the conditional random variable $|V_{0+}[i]|$ given $\mathfrak{E}$ has distribution $\mathrm{Bin}(|V_0[i]|,2^{-\Delta}(1+O(n^{-\Omega(1)})))$. Therefore, since on $\mathcal{E}$ we have $|V_0[i]|=|V[i]|+O(n^{\theta})=n/\ell+O(n^{\theta})$, the assertion follows from the Chernoff bound from Lemma 2.1.

4.7 Proof of Lemma 4.4

The aim is to estimate the weighted sum $\boldsymbol{W}_x^\star$ for infected individuals $x\in V[i+1]$ with $s\leqslant i<\ell$. These individuals join tests in the s compartments $F[i+j]$, $j\in[s]$. Conversely, for each such j the tests $a\in F[i+j]$ recruit their individuals from the compartments $V[i+j-s+1],\ldots,V[i+j]$. Thus the compartments preceding $V[i+1]$ that the tests in $F[i+j]$ draw upon are $V[h]$ with

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

$i+j-s<h\leqslant i$. We begin by investigating the set $\mathcal{W}_{i,j}$ of tests $a\in F[i+j]$ without an infected individual from these compartments, that is,

$$\begin{aligned}\mathcal{W}_{i,j}&=\{a\in F[i+j]\colon (V_1[1]\cup\cdots\cup V_1[i])\cap\partial a=\emptyset\}\\&=\left\{a\in F[i+j]\colon \bigcup_{i+j-s+1<h\leqslant i} V_1[h]\cap\partial a=\emptyset\right\}.\end{aligned}$$

Claim 4. *With probability* $1-o(n^{-2})$, *for all* $s\leqslant i<\ell$, $j\in[s]$ *we have*

$$|\mathcal{W}_{i,j}|=2^{-(s-j)/s}\frac{m}{\ell}(1+O(n^{-\Omega(1)})).$$

Proof. We may condition on the event $\mathcal{E}$ that (i) from Proposition 4.1 occurs. To compute the mean of $|\mathcal{W}_{i,j}|$, fix a test $a\in F[i+j]$ and an index $i+j-s<h\leqslant i$. Then (4.6) shows that the probability that a fixed individual $x\in V[h]$ joins a equals

$$\mathbb{P}[x\in\partial a]=\frac{\Delta\ell}{ms}\left(1+O\left(\frac{\Delta\ell}{ms}\right)\right).$$

Hence the choices (4.2) and (4.5) of Δ and ℓ and the assumption $m=\Theta(k\ln n)$ ensure that

$$\begin{aligned}\mathbb{E}[|(V_1[i+j-s+1]\cup\cdots\cup V_1[i])\cap\partial a|\mid\mathcal{E}]&=(s-j)\left(\frac{\Delta\ell}{ms}\cdot\frac{k}{\ell}+O\left(\frac{\Delta^2k}{m^2s^2}\right)+O\left(\frac{\Delta\ell\sqrt{k}\ln n}{ms}\right)\right)\\&=\frac{s-j}{s}\ln 2+O(n^{-\Omega(1)}).\end{aligned}\tag{4.26}$$

Since by **SC1** the events $\{x\in\partial a\}_x$ are independent, $|V_1[h]\cap\partial a|$ is a binomial random variable for every h and all these random variables $(|V_1[h]\cap\partial a|)_h$ are mutually independent. Therefore (4.26) implies that

$$\mathbb{P}[(V_1[i+j-s+1]\cup\cdots V_1[i])\cap\partial a=\emptyset\mid\mathcal{E}]=2^{-(s-j)/s}+O(n^{-\Omega(1)}).\tag{4.27}$$

Hence

$$\begin{aligned}\mathbb{E}[|\mathcal{W}_{i,j}|\mid\mathcal{E}]&=\sum_{a\in F[i+j]}\mathbb{P}[(V_1[i+j-s+1]\cup\cdots\cup V_1[i])\cap\partial a=\emptyset\mid\mathcal{E}]\\&=\frac{m}{\ell}2^{-(s-j)/s}(1+O(n^{-\Omega(1)})).\end{aligned}\tag{4.28}$$

Finally, changing the neighbourhood ∂x of one infected individual $x\in V_1$ can alter $|\mathcal{W}_{i,j}|$ by at most Δ. Therefore the Azuma–Hoeffding inequality shows that for any $t>0$

$$\mathbb{P}[||\mathcal{W}_{i,j}|-\mathbb{E}[|\mathcal{W}_{i,j}|\mid\mathcal{E}]|>t\mid\mathcal{E}]\leqslant 2\exp\left(-\frac{t^2}{2k\Delta^2}\right).\tag{4.29}$$

Combining (4.28) and (4.29), applied with $t=\sqrt{m}\ln^2 n$, and taking a union bound on i,j completes the proof. □

As the next step we use Claim 4 to estimate the as yet unexplained test counts $\boldsymbol{W}_{x,j}$ from (4.7).

Claim 5. *For all* $s\leqslant i<\ell$, $x\in V_1[i+1]$ *and* $j\in[s]$ *we have*

$$\mathbb{P}[\boldsymbol{W}_{x,j}<(1-\varepsilon/2)2^{j/s-1}\Delta/s]\leqslant\exp\left(-\frac{\Omega(\ln n)}{(\ln\ln n)^4}\right).$$

Proof. Fix a pair of indices i, j and an individual $x \in V_1[i+1]$. We also condition on the event $\mathcal{E}$ that (i) from Proposition 4.1 occurs. Additionally, thanks to Claim 4 we may condition on the event

$$\mathcal{E}' = \left\{ |\mathcal{W}_{i,j}| = 2^{-(s-j)/s} \frac{m}{\ell} (1 + O(n^{-\Omega(1)})) \right\}.$$

Further, let $\mathfrak{E}$ be the σ-algebra generated by $\boldsymbol{\sigma}$ and by the neighbourhoods $(\partial y)_{y \in V[1] \cup \cdots \cup V[i]}$. Recall from **SC1** that x simply joins Δ/s random tests in compartment $F[i+j]$, independently of all other individuals, and remember from (4.7) that $\boldsymbol{W}_{x,j}$ counts tests $a \in \mathcal{W}_{i+j} \cap \partial x$. Therefore, since the events $\mathcal{E}, \mathcal{E}'$ and the random variable $|\mathcal{W}_{i,j}|$ are $\mathfrak{E}$-measurable while ∂x is independent of $\mathfrak{E}$, given $\mathfrak{E}$ the random variable $\boldsymbol{W}_{x,j}$ has a hypergeometric distribution Hyp$(m/\ell, |\mathcal{W}_{i,j}|, \Delta/s)$. Thus the assertion follows from the hypergeometric Chernoff bound from Lemma 2.2 and the choice (4.14) of ζ. □

Proof of Lemma 4.4. Since $\boldsymbol{W}_x^\star = \sum_{j=1}^s w_j \boldsymbol{W}_{x,j}$, the lemma is an immediate consequence of Markov's inequality and Claim 5. □

4.8 Proof of Lemma 4.5

We need to derive the rate functions of the random variable $\boldsymbol{W}_{x,j}$ that count as yet unexplained tests for $x \in V_{0+}[i+1]$. To this end we first investigate the set of positive tests in compartment $i+j$ that do not contain any infected individuals from the first i compartments. In symbols,

$$\mathcal{P}_{i+1,j} = \{a \in F_1[i+j] : \partial a \cap (V_1[1] \cup \cdots \cup V_1[i]) = \emptyset\} \quad (s \leqslant i < \ell,\ j \in [s]).$$

Claim 6. *With high probability, for all $s \leqslant i < \ell, j \in [s]$ we have*

$$|\mathcal{P}_{i+1,j}| = (1 + O(n^{-\Omega(1)}))(2^{j/s} - 1)\frac{m}{2\ell}.$$

Proof. We may condition on the event $\mathcal{E}$ that (i) from Proposition 4.1 occurs. As a first step we calculate the probability that $(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a \neq \emptyset$ for a specific test $a \in F[i+j]$. To this end we follow the steps of the proof of Claim 4. Since by (4.6) a specific individual $x \in V[h]$, $i < h \leqslant i+j$, joins a with probability $\mathbb{P}[x \in \partial a] = (\Delta\ell/(ms))(1 + O(\Delta\ell/(ms)))$, and since given $\mathcal{E}$ each compartment $V[h]$ contains $k/\ell + O(\sqrt{k/\ell} \ln n)$ infected individuals, we obtain, in perfect analogy to (4.26),

$$\mathbb{E}[|(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a| \mid \mathcal{E}] = \frac{j}{s} \ln 2 + O(n^{-\Omega(1)}). \tag{4.30}$$

Since the individuals $x \in V[i+1] \cup \cdots \cup V[i+j]$ join tests independently, (4.30) implies that

$$\mathbb{P}[(V_1[i+1] \cup \cdots \cup V_1[i+j]) \cap \partial a \neq \emptyset \mid \mathcal{E}] = 1 - 2^{-j/s} + O(n^{-\Omega(1)}). \tag{4.31}$$

Furthermore, we already verified in (4.27) that

$$\mathbb{P}[(V_1[i+j-s+1] \cup \cdots V_1[i]) \cap \partial a = \emptyset \mid \mathcal{E}] = 2^{-(s-j)/s} + O(n^{-\Omega(1)}). \tag{4.32}$$

Because the choices for the compartments $V[i+j-s+1] \cup \cdots \cup V[i+j]$ from which a draws its individuals are mutually independent, we can combine (4.31) with (4.32) to obtain

$$\mathbb{P}\left[\bigcup_{i+j-s<h\leqslant i} V_1[h] \cap \partial a = \emptyset \neq \bigcup_{i<h\leqslant i+j} V_1[h] \cap \partial a \mid \mathcal{E} \right] = \frac{2^{j/s} - 1}{2} + O(n^{-\Omega(1)}). \tag{4.33}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Further, (4.33) implies

$$\mathbb{E}[|\mathcal{P}_{i+1,j}| \mid \mathcal{E}] = \mathbb{E}\left[\left|\left\{a \in F_1[i+j] : \bigcup_{h\leqslant i} V_1[h] \cap \partial a = \emptyset \neq \bigcup_{i<h} V_1[h] \cap \partial a\right\}\right| \mid \mathcal{E}\right]$$
$$= (2^{j/s}-1)\frac{m}{2\ell}(1+O(n^{-\Omega(1)})). \tag{4.34}$$

Finally, altering the neighbourhood ∂x of any infected individual can shift $|\mathcal{P}_{i+1,j}|$ by at most Δ. Therefore the Azuma–Hoeffding inequality implies that

$$\mathbb{P}[||\mathcal{P}_{i+1,j}| - \mathbb{E}[|\mathcal{P}_{i+1,j}| \mid \mathcal{E}]| > t \mid \mathcal{E}] \leqslant 2\exp\left(-\frac{t^2}{2k\Delta^2}\right). \tag{4.35}$$

Thus the assertion follows from (4.5), (4.34) and (4.35) by setting $t = \sqrt{m}\ln^2 n$. □

Thanks to Proposition 4.1(iii) and Claim 6, in the following we may condition on the event

$$\mathcal{U} = \Bigg\{\forall s < i \leqslant \ell, j \in [s]:$$
$$|F_1[i+j]| = (1+O(n^{-\Omega(1)}))\frac{m}{2\ell} \wedge |\mathcal{P}_{i+1,j}| = (1+O(n^{-\Omega(1)}))(2^{j/s}-1)\frac{m}{2\ell}\Bigg\}. \tag{4.36}$$

As a next step we will determine the conditional distribution of $\boldsymbol{W}_{x,j}$ for $x \in V_{0+}[i+1]$ given $\mathcal{U}$.

Claim 7. *Let $s < i \leqslant \ell$ and $j \in [s]$. Given $\mathcal{U}$ for every $x \in V_{0+}[i+1]$, we have*

$$\boldsymbol{W}_{x,j} \sim \mathrm{Hyp}\left((1+O(n^{-\Omega(1)}))\frac{m}{2\ell}, (1+O(n^{-\Omega(1)}))(2^{j/s}-1)\frac{m}{2\ell}, \frac{\Delta}{s}\right). \tag{4.37}$$

Proof. By **SC1** each individual $x \in V_{0+}[i+1]$ joins Δ/s positive test from $F[i+j]$, drawn uniformly without replacement. Moreover, by (4.7) given $x \in V_{0+}[i+1]$ the random variable $\boldsymbol{W}_{x,j}$ counts the number of tests $a \in \mathcal{P}_{i+1,j} \cap \partial x$. Therefore $\boldsymbol{W}_{x,j} \sim \mathrm{Hyp}(|F_1[i+j], |\mathcal{P}_{i+1,j}|, \Delta/s)$. Hence, given $\mathcal{U}$ we obtain (4.37). □

The estimate (4.37) enables us to bound the probability that $\boldsymbol{W}_x^\star$ gets 'too large'.

Claim 8. *Let*

$$\mathcal{M} = \min \frac{1}{s}\sum_{j=1}^{s-1} \mathbf{1}\{z_j \geqslant 2^{j/s}-1\} D_{\mathrm{KL}}(z_j \| 2^{j/s}-1)$$
$$\text{subject to} \sum_{j=1}^{s-1}(z_j - (1-2\zeta)2^{j/s-1})w_j = 0, \qquad z_1, \ldots, z_{s-1} \in [0,1].$$

Then, for all $s \leqslant i < \ell$ and all $x \in V[i+1]$, we have

$$\mathbb{P}\left[\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1}w_j \mid \mathcal{U},\ x \in V_{0+}[i+1]\right] \leqslant \exp(-(1+o(1))\mathcal{M}\Delta).$$

Proof. Let $s \leqslant i < \ell$ and $x \in V_{0+}[i+1]$. Step **SC1** of the construction of $\boldsymbol{G}$ ensures that the random variables $(\boldsymbol{W}_{x,j})_{j\in[s]}$ are independent because the tests in the various compartments $F[i+j], j \in [s]$, that x joins are drawn independently. Therefore the definition (4.12) of $\boldsymbol{W}_x^\star$ and Lemma 7 yield

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{W}^{\star}_{x} > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \,|\, \mathcal{U},\ x \in V_{0+}[i+1]\right]\\
&\quad= \mathbb{P}\left[\sum_{j=1}^{s-1} w_j \boldsymbol{W}_{x,j} \geqslant \frac{1-2\zeta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \,|\, \mathcal{U},\ x \in V_{0+}[i+1]\right]\\
&\quad\leqslant \sum_{y_1,\dots,y_s=0}^{\Delta} \mathbf{1}\left\{\sum_{j=1}^{s-1} w_j y_j \geqslant \frac{1-2\zeta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j\right\} \prod_{j=1}^{s-1} \mathbb{P}[\boldsymbol{W}_{x,j} \geqslant y_j \,|\, \mathcal{U},\ x \in V_{0+}[i+1]]. \qquad (4.38)
\end{aligned}$$

Further, let

$$\mathcal{Z} = \left\{(z_1,\dots,z_{s-1}) \in [0,1]^{s-1} : \sum_{j=1}^{s-1} (z_j - (1-2\zeta)2^{j/s-1})w_j = 0\right\}.$$

Substituting $y_j = \Delta z_j/s$ in (4.38) and bounding the total number of summands by $(\Delta+1)^s$, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{W}^{\star}_{x} > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j \,|\, \mathcal{U},\ x \in V_{0+}[i+1]\right]\\
&\quad\leqslant (\Delta+1)^s \max_{(z_1,\dots,z_s)\in\mathcal{Z}} \prod_{j=1}^{s-1} \mathbb{P}[\boldsymbol{W}_{x,j} \geqslant \Delta z_j/s \,|\, \mathcal{U},\ x \in V_{0+}[i+1]]. \qquad (4.39)
\end{aligned}$$

Moreover, Claim 7 and the Chernoff bound from Lemma 2.2 yield

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{W}_{x,j} \geqslant \Delta z_j/s \,|\, \mathcal{U},\ x \in V_{0+}[i+1]]\\
&\quad\leqslant \exp\left(-\mathbf{1}\{z_j \geqslant p_j\}\frac{\Delta}{s} D_{\mathrm{KL}}(z_j \| p_j)\right), \qquad \text{where } p_j = 2^{j/s} - 1 + O(n^{-\Omega(1)}).
\end{aligned}$$

Consequently, since (4.5) and the assumption $m = \Theta(k \ln n)$ ensure that $\Delta = \Theta(\ln n)$, we obtain

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{W}_{x,j} \geqslant \Delta z_j/s \,|\, \mathcal{U},\ x \in V_{0+}[i+1]]\\
&\quad\leqslant \exp\left(-\mathbf{1}\{z_j \geqslant 2^{j/s} - 1\}\frac{\Delta}{s} D_{\mathrm{KL}}(z_j \| 2^{j/s} - 1) + O(n^{-\Omega(1)})\right). \qquad (4.40)
\end{aligned}$$

Finally, the assertion follows from (4.39) and (4.40). □

As a next step we solve the optimization problem $\mathcal{M}$ from Claim 8.

Claim 9. *We have $\mathcal{M} = 1 - \ln 2 + O(\ln(s)/s)$.*

Proof. Fixing an auxiliary parameter $\delta \geqslant 0$, we set up the Lagrangian

$$\begin{aligned}
&\mathcal{L}_{\delta}(z_1,\dots,z_s,\lambda)\\
&\quad= \sum_{j=1}^{s-1} (\mathbf{1}\{z_j \geqslant 2^{j/s} - 1\} + \delta\mathbf{1}\{z_j < 2^{j/s} - 1\}) D_{\mathrm{KL}}(z_j \| 2^{j/s} - 1)\\
&\qquad + \frac{\lambda}{s}\sum_{j=1}^{s-1} w_j (z_j - (1-2\zeta)2^{j/s-1}).
\end{aligned}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

The partial derivatives come out as

$$\frac{\partial \mathcal{L}_\delta}{\partial \lambda} = -\frac{1}{s}\sum_{j=1}^{s-1}((1-2\zeta)2^{j/s-1} - z_j)w_j,$$

$$\frac{\partial \mathcal{L}_\delta}{\partial z_j} = -\lambda w_j + (\mathbf{1}\{z_j \geqslant 2^{j/s}-1\} + \delta\mathbf{1}\{z_j < 2^{j/s}-1\}) \ln \frac{z_j(2-2^{j/s})}{(1-z_j)(2^{j/s}-1)}.$$

Set $z_j^* = (1-2\zeta)2^{j/s-1}$ and $\lambda^* = 1$. Then clearly

$$\left.\frac{\partial \mathcal{L}_\delta}{\partial \lambda}\right|_{\lambda^*, z_1^*,\ldots,z_{s-1}^*} = 0. \tag{4.41}$$

Moreover, the choice (4.14) of ζ guarantees that $z_j^* \geqslant 2^{j/s}-1$. Hence, by the choice (4.14) of the weights w_j,

$$\left.\frac{\partial \mathcal{L}_\delta}{\partial z_j}\right|_{\lambda^*, z_1^*,\ldots,z_{s-1}^*} = 0. \tag{4.42}$$

Since $\mathcal{L}_\delta(y_1,\ldots,y_s,\lambda)$ is strictly convex in $z_1,\ldots,z_s$ for every $\delta > 0$, (4.41)–(4.42) imply that $\lambda^*, z_1^*,\ldots,z_{s-1}^*$ is a global minimizer. Furthermore, since this is true for any $\delta > 0$ and since $z_j^* \geqslant 2^{j/s}-1$, we conclude that $(z_1^*,\ldots,z_{s-1}^*)$ is an optimal solution to the minimization problem $\mathcal{M}$. Hence

$$\mathcal{M} = \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}(z_j^* \| 2^{j/s}-1) = \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}((1-2\zeta)2^{j/s-1} \| 2^{j/s}-1). \tag{4.43}$$

Since

$$\frac{\partial}{\partial \alpha} D_{\mathrm{KL}}((1-2\alpha)2^{z-1} \| 2^z - 1)$$
$$= 2^z[-z\ln(2) + \ln(1-2^{z-1}+\alpha 2^z) - \ln(1-2^{z-1}) - \ln(1-2\alpha) + \ln(2^z-1)],$$

we obtain

$$\frac{\partial}{\partial \alpha} D_{\mathrm{KL}}((1-2\alpha)2^{z-1} \| 2^z - 1) = O(\ln s) \quad \text{for all } z = 1/s,\ldots,(s-1)/s \text{ and } \alpha \in [0, 2\zeta].$$

Combining this bound with (4.43), we arrive at the estimate

$$\mathcal{M} = O(\zeta \ln s) + \frac{1}{s}\sum_{j=1}^{s-1} D_{\mathrm{KL}}(2^{j/s-1} \| 2^{j/s}-1). \tag{4.44}$$

Additionally, the function $f\colon z \in [0,1] \mapsto D_{\mathrm{KL}}(2^{z-1} \| 2^z - 1)$ is strictly decreasing and convex. Indeed,

$$f'(z) = \frac{2^{z-1}\ln 2}{2^z-1}\left((2^z-1)\ln\left(\frac{2^z}{2^z-1}\right) - 1\right),$$

$$f''(z) = (2^{z-1}\ln^2 2)\left(\ln\left(\frac{2^z}{2^z-1}\right) + \frac{2-2^z}{(2^z-1)^2}\right).$$

The first derivative is negative because $2^{z-1}/(2^z-1) > 0$ while $(2^z-1)\ln(2^z/(2^z-1)) < 1$ for all $z \in (0,1)$. Moreover, since evidently $f''(z) > 0$ for all $z \in (0,1)$, we obtain convexity. Further, l'Hôpital's rule yields

$$D_{\mathrm{KL}}(2^{1/s-1} \| 2^{1/s}-1) = O(\ln s).$$

As a consequence, we can approximate the sum (4.44) by an integral and obtain

$$\begin{aligned}\mathcal{M} &= O(\ln(s)/s) + \int_0^1 D_{\mathrm{KL}}(2^{z-1}\|2^z - 1)\mathrm{d}z\\ &= O(\ln(s)/s) + \frac{2(1-z)\ln^2(2) + 2^z\ln 2^z + (1-2^z)\ln(2^z-1)}{2\ln 2}\Bigg|_{z=0}^{z=1}\\ &= 1 - \ln(2) + O(\ln(s)/s),\end{aligned}$$

as claimed. □

Proof of Lemma 4.5. Fix $s \leqslant i < \ell$ and let $\boldsymbol{X}_i$ be the number of $x \in V_{0+}[i]$ such that

$$\boldsymbol{W}_x^\star > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j.$$

Also recall that Proposition 4.1(iii) and Claim 6 imply that $\mathbb{P}[\mathcal{U}] = 1 - o(1)$. Combining Lemma 4.3 with Claims 8 and 9, we conclude that

$$\mathbb{E}[\boldsymbol{X}_i \mid \mathcal{U}] \leqslant (1 + O(n^{-\Omega(1)}))2^{-\Delta} n \exp(-(1-\ln(2)+o(1))\Delta) = \exp(\ln n - (1+o(1))\Delta). \tag{4.45}$$

Recalling the definition (4.5) of Δ and using the assumption that $m \geqslant (1+\varepsilon)m_{\mathrm{ad}}$ for a fixed $\varepsilon > 0$, we obtain $\Delta \geqslant (1-\theta+\Omega(1))\ln n$. Combining this estimate with (4.45), we find

$$\mathbb{E}[\boldsymbol{X}_i \mid \mathcal{U}] \leqslant n^{\theta - \Omega(1)}. \tag{4.46}$$

Finally, the assertion follows from (4.46) and Markov's inequality. □

4.9 Proof of Proposition 4.6

The following lemma establishes an expansion property of $\boldsymbol{G}$. Specifically, if T is a small set of individuals, then there are few individuals x that share many tests with another individual from T.

Lemma 4.10. *Suppose that $m = \Theta(n^\theta \ln n)$. With high probability, for any set $T \subset V$ of size at most $\exp(-\ln^{7/8} n)k$ we have*

$$\left|\left\{x \in V : \sum_{a \in \partial x \setminus F[0]} \mathbf{1}\{T \cap \partial a \setminus \{x\} \neq \emptyset\} \geqslant \ln^{1/4} n\right\}\right| \leqslant \frac{|T|}{3}.$$

Proof. Fix a set $T \subset V$ of size $t = |T| \leqslant \exp(-\ln^{7/8} n)k$, a set $R \subset V$ of size $r = \lceil t/3 \rceil$ and let $\gamma = \lceil \ln^{1/4} n \rceil$. Furthermore, let $U \subset F[1] \cup \cdots \cup F[\ell]$ be a set of tests of size $\gamma r \leqslant u \leqslant \Delta t$. Additionally, let $\mathcal{E}(R, T, U)$ be the event that every test $a \in U$ contains two individuals from $R \cup T$. Then

$$\mathbb{P}\left[R \subset \left\{x \in V : \sum_{a \in \partial x \setminus F[0]} \mathbf{1}\{T \cap \partial a \setminus \{x\} \neq \emptyset\} \geqslant \gamma\right\}\right] \leqslant \mathbb{P}[\mathcal{E}(R, T, U)]. \tag{4.47}$$

Hence it suffices to estimate $\mathbb{P}[\mathcal{E}(R, T, U)]$.

Given a test $a \in U$, there are at most $\binom{r+t}{2}$ ways to choose two individuals $x_a, x_a' \in R \cup T$. Moreover, (4.6) shows that the probability of the event $\{x_a, x_a' \in \partial a\}$ is bounded by $(1 + o(1))(\Delta\ell/(ms))^2$. Therefore

$$\mathbb{P}[\mathcal{E}(R, T, U)] \leqslant \left[\binom{r+t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u.$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Consequently, the event $\mathcal{E}(t, u)$ that there exist sets R, T, U of sizes $|R| = r = \lceil t/3 \rceil$, $|T| = t$, $|U| = u$ such that $\mathcal{E}(R, T, U)$ occurs has probability

$$\mathbb{P}[\mathcal{E}(t,u)] \leqslant \binom{n}{r}\binom{n}{t}\binom{m}{u}\left[\binom{r+t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u.$$

Hence the bounds $\gamma t/3 \leqslant \gamma r \leqslant u \leqslant \Delta t$ yield

$$\begin{aligned}\mathbb{P}[\mathcal{E}(t,u)] &\leqslant \binom{n}{t}^2\binom{m}{u}\left[\binom{2t}{2}\left(\frac{(1+o(1))\Delta\ell}{ms}\right)^2\right]^u \\ &\leqslant \left(\frac{\mathrm{e}n}{t}\right)^{2t}\left(\frac{2\mathrm{e}\Delta^2\ell^2t^2}{ms^2u}\right)^u \\ &\leqslant \left[\left(\frac{\mathrm{e}n}{t}\right)^{3/\gamma}\frac{6\mathrm{e}\Delta^2\ell^2t}{\gamma ms^2}\right]^u \\ &\leqslant \left[\left(\frac{\mathrm{e}n}{t}\right)^{3/\gamma}\cdot\frac{t\ln^4 n}{m}\right]^u \qquad \text{(due to (4.2), (4.5)).}\end{aligned}$$

Further, since $\gamma = \Omega(\ln^{1/4} n)$ and $m = \Omega(k \ln n)$ while $t \leqslant \exp(-\ln^{7/8} n)k$, we obtain $\mathbb{P}[\mathcal{E}(t, u)] \leqslant \exp(-u\sqrt{\ln n})$. Thus

$$\sum_{\substack{1\leqslant t\leqslant k^{1-\alpha}\\ \gamma t/3\leqslant u\leqslant\Delta t}} \mathbb{P}[\mathcal{E}(t,u)] \leqslant \sum_{1\leqslant u\leqslant\Delta t} u\exp(-u\sqrt{\ln n}) = o(1). \tag{4.48}$$

Finally, the assertion follows from (4.47) and (4.48). □

Proof of Proposition 4.6. With τ the result of steps 1–10 of SPIV, let $\mathcal{M}[i] = \{x \in V[i] : \tau_x \neq \boldsymbol{\sigma}_x\}$ be the set of misclassified individuals in compartment $V[i]$. Proposition 4.2 shows that with high probability $\mathcal{M}[i] = \emptyset$ for all $i \leqslant s$. Further, we claim that for every $s \leqslant i < \ell$ and any individual $x \in \mathcal{M}[i + 1]$, one of the following three statements is true.

M1 $x \in V_1[i + 1]$ and

$$\boldsymbol{W}^\star_x < \left(1-\frac{\zeta}{2}\right)\frac{\Delta}{s}\sum_{j=1}^{s-1}2^{j/s-1}w_j,$$

M2 $x \in V_{0+}[i + 1]$ and

$$\boldsymbol{W}^\star_x > (1-2\zeta)\frac{\Delta}{s}\sum_{j=1}^{s-1}2^{j/s-1}w_j,$$

or

M3 $x \in V[i + 1]$ and

$$\sum_{a\in\partial x}\mathbf{1}\{\partial a\cap(\mathcal{M}[1]\cup\cdots\cup\mathcal{M}[i])\neq\emptyset\}\geqslant\ln^{1/4} n.$$

To see this, assume that $x \in \mathcal{M}[i + 1]$ while **M3** does not hold. Then, comparing (4.7) and (4.17), we obtain

$$|W_{x,j}(\tau) - \boldsymbol{W}_{x,j}| \leqslant \ln^{1/4} n \quad \text{for all } 1 \leqslant j < s. \tag{4.49}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Moreover, the definition (4.14) of the weights, the choice (4.3) of s and the choices (4.14) of ζ and the weights w_j ensure that $0 \leqslant w_j \leqslant O(s) = O(\ln\ln n)$. This bound implies together with the definition (4.12) of the scores $\boldsymbol{W}_x^\star$ and (4.49) that

$$|\boldsymbol{W}_x^\star - W_x^\star(\tau)| = o(\zeta\Delta). \tag{4.50}$$

Thus, combining (4.50) with the definition of τ_x in steps 5–10 of SPIV, we conclude that either **M1** or **M2** occurs.

Finally, to bound $\mathcal{M}[i+1]$ let $\mathcal{M}_1[i+1]$, $\mathcal{M}_2[i+1]$ and $\mathcal{M}_3[i+1]$ be the sets of individuals $x \in V[i+1]$ for which **M1**, **M2** or **M3** occurs, respectively. Then Lemmas 4.4 and 4.5 imply that with high probability

$$|\mathcal{M}_1[i+1]|, |\mathcal{M}_2[i+1]| \leqslant k\exp\left(-\frac{\ln n}{(\ln\ln n)^5}\right).$$

Furthermore, Lemma 4.10 shows that $|\mathcal{M}_3[i+1]| \leqslant \sum_{h=1}^{i} |\mathcal{M}[h]|$ with high probability. Hence we obtain the relation

$$|\mathcal{M}[i+1]| \leqslant k\exp\left(-\frac{\ln n}{(\ln\ln n)^5}\right) + \sum_{h=1}^{i} |\mathcal{M}[h]|. \tag{4.51}$$

Because (4.2) ensures that the total number of compartments is $\ell = O(\ln^{1/2} n)$, the bound (4.51) implies that $|\mathcal{M}[i+1]| \leqslant O(\ell^2 k \exp(-(\ln n)/(\ln\ln n)^5)$ for all $i \in [\ell]$ with high probability. Summing on i completes the proof. □

4.10 Proof of Proposition 4.7

For an infected individual $x \in V$ let

$$\boldsymbol{S}_x[j] = |\{a \in F[j] \cap \partial x \colon V_1 \cap \partial a = \{x\}\}| \quad \text{and} \quad \boldsymbol{S}_x = \sum_{j=1}^{\ell} \boldsymbol{S}_x[j].$$

Thus $\boldsymbol{S}_x[j]$ is the number of positive sets $a \in F[j]$ that x has to itself, *i.e.* tests that do not contain a second infected individual, and $\boldsymbol{S}_x$ is the total number of such tests.

Lemma 4.11. *Assume that $m \geqslant (1+\varepsilon)m_{\text{inf}}$. With high probability we have* $\min_{x \in V_1} \boldsymbol{S}_x \geqslant \sqrt{\Delta}$.

Proof. Due to Proposition 4.1 we may condition on the event

$$\mathcal{N} = \left\{\forall i \in [\ell] \colon \frac{m}{2\ell} - \sqrt{m}\ln n \leqslant |F_0[i]| \leqslant \frac{m}{2\ell} + \sqrt{m}\ln n\right\}.$$

We claim that, given $\mathcal{N}$ for each $x \in V_1[i]$, $i \in [\ell]$, the random variable $\boldsymbol{S}_x$ has distribution

$$\boldsymbol{S}_x[i+j-1] \sim \text{Hyp}\left(\frac{m}{\ell}, \frac{m}{2\ell} + O(\sqrt{m}\ln n), \frac{\Delta}{s}\right). \tag{4.52}$$

To see this, consider the set

$$F_x[i+j-1] = \{a \in F[i+j-1] \colon \partial a \cap V_1 \setminus \{x\} = \emptyset\}$$

of all tests in compartment $F[i+j-1]$ without an infected individual besides possibly x. Since x joins $\Delta/s = O(\ln n)$ tests in $F[i+j-1]$, given $\mathcal{N}$ we have

$$|F_{0,x}[i+j]| = |F_0[i+j]| + O(\ln n) = \frac{m}{2\ell} + O(\sqrt{m}\ln n). \tag{4.53}$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Furthermore, consider the experiment of first constructing the test design $\boldsymbol{G}$ and then resampling the set ∂x of neighbours of x; that is, independently of $\boldsymbol{G}$ we have x join Δ/s random tests in each compartment $F[i+j]$. Then the resulting test design $\boldsymbol{G}'$ has the same distribution as $\boldsymbol{G}$, and hence the random variable $\boldsymbol{S}'_x[i+j-1]$ that counts tests $a \in F[i+j-1] \cap \partial x$ that do not contain another infected individual has the same distribution as $\boldsymbol{S}_x[i+j-1]$. Moreover, the conditional distribution of $\boldsymbol{S}'_x[i+j-1]$ given $\boldsymbol{G}$ reads

$$\boldsymbol{S}'_x[i+j-1] \sim \mathrm{Hyp}\left(\frac{m}{\ell}, |F_{0,x}[i+j-1]|, \frac{\Delta}{s}\right). \tag{4.54}$$

Combining (4.53) and (4.54), we obtain (4.52).

To complete the proof we combine (4.52) with Lemma 2.2, which implies that

$$\begin{aligned}\mathbb{P}[\boldsymbol{S}_x[i+j-1] \leqslant \sqrt{\Delta} \mid x \in V_1] &\leqslant \exp\left(-\frac{\Delta}{s} D_{\mathrm{KL}}((1+o(1))s/\sqrt{\Delta} \| 1/2 + o(1))\right) \\ &= \exp\left(-(1+o(1))\frac{\Delta \ln 2}{s}\right).\end{aligned} \tag{4.55}$$

Since **SC1** ensures that the random variables $(\boldsymbol{S}_x[i+j-1])_{j\in[s]}$ are mutually independent, (4.55) yields

$$\mathbb{P}[\boldsymbol{S}_x \leqslant \sqrt{\Delta} \mid x \in V_1] \leqslant 2^{-(1+o(1))\Delta}. \tag{4.56}$$

Finally, the assumption $m \geqslant (1+\varepsilon)m_{\inf}$ for a fixed $\varepsilon > 0$ and the choice (4.5) of Δ ensure that $2^{-(1+o(1))\Delta} = o(1/k)$. Thus the assertion follows from (4.56) by taking a union bound on $x \in V_1$. □

Proof of Proposition 4.7. For $j = 1 \ldots \lceil \ln n \rceil$, let

$$\mathcal{M}_j = \{x \in V : \tau_x^{(j)} \neq \boldsymbol{\sigma}_x\}$$

contain all individuals that remain misclassified at the jth iteration of the clean-up step. Proposition 4.6 shows that with high probability

$$|\mathcal{M}_1| \leqslant k \exp\left(-\frac{\ln n}{(\ln \ln n)^6}\right). \tag{4.57}$$

Furthermore, in light of Lemma 4.11 we may condition on the event $\mathcal{A} = \{\min_{x\in V_1} \boldsymbol{S}_x \geqslant \sqrt{\Delta}\}$.

We now claim that, given $\mathcal{A}$, for every $j \geqslant 1$

$$\mathcal{M}_{j+1} \subset \left\{x \in V : \sum_{a \in \partial x \setminus F[0]} |\partial a \cap \mathcal{M}_j \setminus \{x\}| \geqslant \left\lceil \ln^{1/4} n \right\rceil\right\}. \tag{4.58}$$

To see this, suppose that $x \in \mathcal{M}_{j+1}$ and recall that the assumption $m \geqslant m_{\inf}$ and (4.5) ensure that $\Delta = \Omega(\ln n)$. Also recall that SPIV's step 15 thresholds the number

$$S_x(\tau^{(j)}) = \sum_{a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 1} \mathbf{1}\{\forall y \in \partial a \setminus \{x\} : \tau_y^{(j)} = 0\}$$

of positive tests containing x whose other individuals are deemed uninfected. There are two cases to consider.

Case 1: $x \in V_0$. In this case every positive tests $a \in \partial x$ contains an individual that is actually infected. Hence, if $\tau_y^{(j)} = 0$ for all $y \in \partial a \setminus \{x\}$, then $\partial a \cap \mathcal{M}_j \setminus \{x\} \neq \emptyset$. Consequently, since step 15 of SPIV applies the threshold of $S_x(\tau^{(j)}) \geqslant \ln^{1/4} n$, there are at least $\ln^{1/4} n$ tests $a \in \partial x$ such that $\partial a \cap \mathcal{M}_j \setminus \{x\} \neq \emptyset$.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Case 2: $x \in V_1$. Given $\mathcal{A}$, every infected x participates in at least $\boldsymbol{S}_x \geqslant \sqrt{\Delta} = \Omega(\ln^{1/2} n)$ tests that do not actually contain another infected individual. Hence, if $S_x(\tau^{(j)}) \leqslant \ln^{1/4} n$, then at least $\sqrt{\Delta} - \ln^{1/4} n \geqslant \ln^{1/4} n$ tests $a \in \partial x$ contain an individual from $\mathcal{M}_j \setminus \{x\}$.

Thus we obtain (4.58). Finally, (4.57), (4.58) and Lemma 4.10 show that with high probability $|\mathcal{M}_{j+1}| \leqslant |\mathcal{M}_j|/3$ for all $j \geqslant 1$. Consequently, $\mathcal{M}_{\lceil \ln n \rceil} = \emptyset$ with high probability. □

5. Optimal adaptive group testing

In this final section we show how the test design $\boldsymbol{G}$ from Section 4 can be extended into an optimal two-stage adaptive design. The key observation is that Proposition 4.6, which summarizes the analysis of the first two phases of SPIV (*i.e.* steps 1–10) only requires $m \geqslant (1+\varepsilon)m_{\mathrm{ad}}$ tests. In other words, the excess number $(1+\varepsilon)(m_{\inf} - m_{\mathrm{ad}})$ of tests required for non-adaptive group testing is necessary only to facilitate the clean-up step, namely phase 3 of SPIV. Replacing phase 3 of SPIV with a second test stage, we obtain an optimal adaptive test design. To this end we follow Scarlett [34], who observed that a single-stage group testing scheme that correctly diagnoses all but $o(k)$ individuals with $(1+o(1))m_{\mathrm{ad}}$ tests could be turned into a two-stage design that diagnoses all individuals correctly with high probability with $(1+o(1))m_{\mathrm{ad}}$ tests in total. (Of course, at the time no such optimal single-stage test design and algorithm were known.) The second test stage works as follows. Let τ denote the outcome of phases 1 and 2 of SPIV applied to $\boldsymbol{G}$ with $m = (1+\varepsilon)m_{\mathrm{ad}}$.

T1 Test every individual from the set $V_1(\tau) = \{x \in V : \tau_x = 1\}$ of individuals that SPIV diagnosed as infected separately.

T2 To the individuals $V_0(\tau) = \{x \in V : \tau_x = 0\}$ apply the random d-out design and the DD-algorithm from Section 4.1 with a total of $m = k$ tests and $d = \lceil 10 \ln n \rceil$.

Let $\tau' \in \{0,1\}^V$ be the result of **T1**–**T2**.

Proposition 5.1. *With high probability we have* $\tau'_x = \boldsymbol{\sigma}_x$ *for all* $x \in V$.

As a matter of course **T1** renders correct results, that is, for all individuals $x \in V_1(\tau)$ we have $\tau'_x = \boldsymbol{\sigma}_x$. Further, to analyse **T2** we use an argument similar to the analysis of the first phase of SPIV in Section 4.5; we include the analysis for the sake of completeness. We begin by investigating the number of negative tests. Let $\boldsymbol{G}'$ denote the test design set up by **T2**, let $F' = \{b_1, \ldots, b_k\}$ denote its set of tests and let $\hat{\boldsymbol{\sigma}}_{b_1}, \ldots, \hat{\boldsymbol{\sigma}}_{b_k}$ signify the corresponding test results. Further, let $F'_0 = \{b \in F' : \hat{\boldsymbol{\sigma}}_b = 0\}$ and $F'_1 = \{b \in F' : \hat{\boldsymbol{\sigma}}_b = 1\}$ be the set of negative and positive tests, respectively.

Lemma 5.2. *With high probability we have* $|F'_1| \leqslant k/2$.

Proof. Proposition 4.6 implies that with high probability

$$|V_0(\tau) \cap V_1| \leqslant \sum_{x \in V} \mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leqslant k \exp\left(-\frac{\ln n}{(\ln \ln n)^6}\right). \tag{5.1}$$

Moreover, since every individual $x \in V_0(\tau)$ joins d random tests, for any specific test $b \in F'$ we have

$$\mathbb{P}[x \in \partial_{\boldsymbol{G}'} b] = 1 - \mathbb{P}[x \notin \partial_{\boldsymbol{G}'} b] = 1 - \binom{k-1}{d}\binom{k}{d}^{-1} = \frac{d}{k}(1 + O(n^{-\Omega(1)})).$$

Hence, for every test $b \in F'$,

$$\mathbb{E}\left[|\partial b \cap V_1| \,\middle|\, |V_0(\tau) \cap V_1| \leqslant k \exp\left(-\frac{\ln n}{(\ln \ln n)^6}\right)\right] = O(1/\ln n).$$

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

Consequently,

$$\mathbb{E}[|F_1'| \mid |V_0(\tau) \cap V_1| \leqslant k/\ln n] = O(k/\ln n). \tag{5.2}$$

Finally, combining (5.1) and (5.2) and applying Markov's inequality, we conclude that $|F_1'| \leqslant k/2$ with high probability. □

Corollary 5.3. *With high probability, for every $x \in V_0(\tau)$ there is a test $b \in F'$ such that $\partial b \setminus \{x\} \subset V_0$.*

Proof. We construct the random graph $\boldsymbol{G}'$ in two rounds. In the first round we first expose the neighbourhoods $(\partial_{\boldsymbol{G}'} y)_{y \in V_0(\tau)\setminus\{x\}}$. Lemma 5.2 implies that after the first round the number $\boldsymbol{X}$ of tests that do not contain an infected individual $y \in V_0(\tau) \cap V_1$ exceeds $k/2$ with high probability. In the second round we expose $\partial_{\boldsymbol{G}'} x$. Because $\partial_{\boldsymbol{G}'} x$ is chosen independently of the neighbourhoods $(\partial_{\boldsymbol{G}'} y)_{y \in V_0(\tau)\setminus\{x\}}$, the number of tests $b \in \partial_{\boldsymbol{G}'} x$ that do not contain an infected individual $y \in V_0(\tau) \cap V_1$ has distribution $\mathrm{Hyp}(k, \boldsymbol{X}, d)$. Therefore, since $d \geqslant 10 \ln n$ we obtain

$$\mathbb{P}[\forall b \in \partial x \colon V_1 \cap \partial b \setminus \{x\} \neq \emptyset \mid \boldsymbol{X} \leqslant k/2] \leqslant \mathbb{P}[\mathrm{Hyp}(k, k/2, d) = 0] \leqslant 2^{-d} = o(1/n). \tag{5.3}$$

Finally, the assertion follows (5.3) and the union bound. □

Proof of Proposition 5.1. Corollary 5.3 shows that we may assume that for every $x \in V_0(\tau)$ there is a test $b_x \in F'$ with $\partial b_x \setminus \{x\} \subset V_0$. As a consequence, upon executing the first step **DD1** of the DD algorithm, **T2** will correctly diagnose all individuals $x \in V_0(\tau) \cap V_0$. Therefore, if $x \in V_0(\tau) \cap V_1$, then **DD2** will correctly identify x as infected because all other individuals $y \in \partial b_x$ were already identified as healthy by **DD1**. Thus $\tau_x' = \boldsymbol{\sigma}_x$ for all $x \in V$. □

Proof of Theorem 1.3. Proposition 5.1 already establishes that the output of the two-stage adaptive test is correct with high probability. Hence, to complete the proof we just observe that the total number of tests comes to $(1 + \varepsilon)m_{\mathrm{ad}}$ for the first stage plus $|V_1(\tau)| + k$ for the second stage. Furthermore, Proposition 4.6 implies that with high probability

$$|V_1(\tau)| \leqslant |V_1| + \sum_{x \in V} \mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leqslant k\left(1 + \exp\left(-\frac{\ln n}{(\ln \ln n)^6}\right)\right) = (1 + o(1))k.$$

Thus the second stage conducts $O(k) = o(m_{\mathrm{ad}})$ tests. □

Acknowledgement

We thank Arya Mazumdar for bringing the group testing problem to our attention.

References

[1] Abbe, E. (2017) Community detection and stochastic block models: recent developments. *J. Mach. Learning Res.* **18** 6446–6531.

[2] Alaoui, A., Ramdas, A., Krzakala, F., Zdeborová, L. and Jordan, M. (2019) Decoding from pooled data: phase transitions of message passing. *IEEE Trans. Inform. Theory* **65** 572–585.

[3] Alaoui, A., Ramdas, A., Krzakala, F., Zdeborová, L. and Jordan, M. (2019) Decoding from pooled data: sharp information-theoretic bounds. *SIAM J. Math. Data Sci.* **1** 161–188.

[4] Aldridge, M. (2019) Individual testing is optimal for nonadaptive group testing in the linear regime. *IEEE Trans. Inform. Theory* **65** 2058–2061.

[5] Aldridge, M., Baldassini, L. and Johnson, O. (2014) Group testing algorithms: bounds and simulations. *IEEE Trans. Inform. Theory* **60** 3671–3687.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

[6] Aldridge, M., Johnson, O. and Scarlett, J. (2019) Group testing: an information theory perspective. *Found. Trends Commun. Inform. Theory* **15** 196–392.
[7] Alon, N., Krivelevich, M. and Sudakov, B. (1998) Finding a large hidden clique in a random graph. In *Proceedings of the Ninth Annual ACM–SIAM Symposium on Discrete Algorithms (SODA '98)*, pp. 594–598.
[8] Arıkan, E. (2009) Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inform. Theory* **55** 3051–3073
[9] Berger, T. and Levenshtein, V. (2002) Asymptotic efficiency of two-stage disjunctive testing. *IEEE Trans. Inform. Theory* **48** 1741–1749.
[10] Brennan, M. and Bresler, G. (2019) Optimal average-case reductions to sparse PCA: from weak assumptions to strong hardness. *Proc. Mach. Learning Res.* **99** 469–470
[11] Chen, H. and Hwang, F. (2008) A survey on nonadaptive group testing algorithms through the angle of decoding. *J. Combin. Optim.* **15** 49–59.
[12] Coja-Oghlan, A., Gebhard, O., Hahn-Klimroth, M. and Loick, P. (2019) Information-theoretic and algorithmic thresholds for group testing. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, #43. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.
[13] Decelle, A., Krzakala, F., Moore, C. and Zdeborová, L. (2011) Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84** 066106.
[14] Donoho, D. (2006) Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306.
[15] Donoho, D., Javanmard, A. and Montanari, A. (2013) Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inform. Theory* **59** 7434–7464.
[16] Dorfman, R. (1943) The detection of defective members of large populations. *Ann. Math. Statist.* **14** 436–440.
[17] D'yachkov, A. and Rykov, V. (1982) Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii* **18** 166–171.
[18] Erdős, P. and Rényi, A. (1963) On two problems of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl* **8** 229–243.
[19] Felstrom, A. and Zigangirov, K. (1999) Time-varying periodic convolutional codes with low-density parity-check matrix. *IEEE Trans. Inform. Theory* **45** 2181–2191.
[20] Grötschel, M., Lovász, L. and Schrijver, A. (1988) *The Ellipsoid Method and Combinatorial Optimization.* Springer.
[21] Hoeffding, W. (1994) Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding* (N. Fisher and P. Sen, eds), Springer Series in Statistics (Perspectives in Statistics), pp. 409–426. Springer.
[22] Hwang, F. (1972) A method for detecting all defective members in a population by group testing. *J. Amer. Statist. Assoc.* **67** 605–608.
[23] Janson, S., Łuczak, T. and Ruciński, A. (2011) *Random Graphs.* Wiley.
[24] Johnson, O., Aldridge, M. and Scarlett, J. (2018) Performance of group testing algorithms with near-constant tests per item. *IEEE Trans. Inform. Theory* **65** 707–723.
[25] Kautz, W. and Singleton, R. (1964) Nonrandom binary superimposed codes. *IEEE Trans. Inform. Theory* **10** 363–377.
[26] Krzakala, F., Mézard, M., Sausset, F., Sun, Y. and Zdeborová, L. (2012) Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X* **2** 021005.
[27] Kudekar, S. and Pfister, H. D. (2010) The effect of spatial coupling on compressive sensing. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Allerton, IL, 2010*, pp. 347–353.
[28] Kudekar, S., Richardson, T. and Urbanke, R. (2011) Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC. *IEEE Trans. Inform. Theory* **57** 803–834.
[29] Kudekar, S., Richardson, T. and Urbanke, R. (2013) Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Trans. Inform. Theory* **59** 7761–7813.
[30] Kwang-Ming, H. and Ding-Zhu, D. (2006) *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing.* World Scientific.
[31] Mézard, M., Tarzia, M. and Toninelli, C. (2008) Group testing with random pools: phase transitions and optimal strategy. *J. Statist. Phys.* **131** 783–801.
[32] Moore, C. (2017) The computer science and physics of community detection: landscapes, phase transitions, and hardness. *Bull. EATCS* **121**.
[33] Reeves, G. and Pfister, H. (2019) Understanding phase transitions via mutual information and MMSE. arXiv:1907.02095
[34] Scarlett, J. (2018) Noisy adaptive group testing: bounds and algorithms. *IEEE Trans. Inform. Theory* **65** 3646–3661.
[35] Scarlett, J. (2019) An efficient algorithm for capacity-approaching noisy adaptive group testing. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2679–2683. IEEE.
[36] Sharma, A. and Salim, M. (2017) Polar code: the channel code contender for 5G scenarios. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, pp. 676–682. IEEE.

https://doi.org/10.1017/S096354832100002X Published online by Cambridge University Press

[37] Takeuchi, K., Tanaka, T. and Kawabata, T. (2011) Improvement of BP-based CDMA multiuser detection by spatial coupling. In *2011 IEEE International Symposium on Information Theory*, pp. 1489–1493. IEEE.

[38] Ungar, P. (1960) The cutoff point for group testing. *Commun. Pure Appl. Math.* **13** 49–54.

[39] Wang, L., Li, X., Zhang, Y. and Zhang, K. (2011) Evolution of scaling emergence in large-scale spatial epidemic spreading. *PLoS ONE* **6** e21197.

[40] Wu, Y. and Verdú, S. (2010) Rényi information dimension: fundamental limits of almost lossless analog compression. *IEEE Trans. Inform. Theory* **56** 3721–3748.

[41] Zdeborová, L. and Krzakala, F. (2016) Statistical physics of inference: thresholds and algorithms. *Adv. Phys.* **65** 453–552.

Cite this article: Coja-Oghlan A, Gebhard O, Hahn-Klimroth M and Loick P (2021). Optimal group testing. *Combinatorics, Probability and Computing* **30**, 811–848. https://doi.org/10.1017/S096354832100002X