Optimal group testing

Amin Coja-Oghlan ACOGHLAN@MATH.UNI-FRANKFURT.DE
Oliver Gebhard GEBHARD@MATH.UNI-FRANKFURT.DE
Max Hahn-Klimroth HAHNKLIM@MATH.UNI-FRANKFURT.DE
Philipp Loick LOICK@MATH.UNI-FRANKFURT.DE
Goethe University, Mathematics Institute, 10 Robert Mayer St, Frankfurt 60325, Germany.

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

In the group testing problem, which goes back to the work of Dorfman (1943), we aim to identify a small set of $k \sim n^\theta$ infected individuals out of a population size n, $0 < \theta < 1$. We avail ourselves to a test procedure that can test a group of individuals, with the test returning a positive result iff at least one individual in the group is infected. All tests are conducted in parallel. The aim is to devise a test design with as few tests as possible so that the infected individuals can be identified with high probability. We establish an explicit sharp information-theoretic/algorithmic phase transition m_{inf}, showing that with more than m_{inf} tests the infected individuals can be identified in polynomial time, while this is impossible with fewer tests. In addition, we obtain an optimal two-stage adaptive group testing scheme. These results resolve problems prominently posed in [Aldridge et al. 2019, Johnson et al. 2018, Mézard and Toninelli 2011].

Keywords: Group testing, Bayesian inference, information theory, efficient algorithms

1. Introduction

The group testing problem. Various intriguing computational challenges come as inference problems 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, first, a positive algorithmic result showing that a certain number of queries suffice to learn the ground truth efficiently. Second, a matching information-theoretic lower bound showing that with fewer queries the problem is insoluble, regardless the amount of computational resources we are willing to throw at it.

Group testing is a prime example of such an inference problem. The problem has been receiving a great deal of attention recently; Aldridge et al. (2019) provide an up-to-date survey. The task is to identify within a large population those individuals infected with a rare disease. At our disposal we have a test procedure capable of not merely testing one individual but an entire group. The test will come back positive if any one individual in the group is infected and negative otherwise. All tests are conducted in parallel, i.e., there is one round of testing only. We are free to allocate individuals to test groups as we please. In particular, we may allocate each individual to an arbitrary number of groups, with no restriction on the group sizes. Randomisation is allowed. What is the least number of tests required to infer the set of infected individuals from the test results with high probability?

The two main results of this paper furnish matching algorithmic and information-theoretic bounds. These results close the considerable gap that the best prior bounds left. To elaborate, a

key feature of group testing is that the test design is at our discretion, which we exercise by equipping the new inference algorithm with a tailor-made test design. While the best previous algorithms relied on a test design based on a plain random graph, we instead harness a blend of a geometric and a random construction. This test design, reminiscent of recent advances in coding theory known as spatially coupled codes Felstrom and Zigangirov (1999); Kudekar et al. (2011), enables an optimal combinatorial inference algorithm that is easy to comprehend, implement and run. With respect to the lower bound, we improve over an argument from Mézard and Toninelli (2011); Aldridge (2019) based on the FKG inequality by introducing a subtle application of the probabilistic method. Let us proceed to state the main results formally.

A sharp algorithmic/information-theoretic threshold. Within a population of size n we aim to identify a set of $k \sim n^\theta$ infected individuals for a fixed parameter $0 < \theta < 1$. Let $\boldsymbol{\sigma} \in \{0,1\}^n$ be the vector whose 1-entries mark the infected individuals. Permuting the indices, we may assume that $\boldsymbol{\sigma}$ is a random vector of Hamming weight k. Let

$$m_{\inf} = m_{\inf}(n,\theta) = \max\left\{\frac{\theta}{\ln^2 2}, \frac{1-\theta}{\ln 2}\right\} n^\theta \ln n. \tag{1.1}$$

Theorem 1 *For any $0 < \theta < 1$, $\varepsilon > 0$ there exists a randomised test design comprising no more than $(1+\varepsilon)m_{\inf}$ tests and a polynomial time algorithm that given the test results outputs $\boldsymbol{\sigma}$ w.h.p.*

Theorem 2 *For any $0 < \theta < 1$, $\varepsilon > 0$ no test design with fewer than $(1-\varepsilon)m_{\inf}$ tests exists so that given the tests results any algorithm, efficient or not, outputs $\boldsymbol{\sigma}$ with a non-vanishing probability.*

Theorems 1 and 2 show that there occurs an algorithmic/information-theoretic phase transition at $m_{\inf}$. Indeed, if we allow for a number of tests greater than $m_{\inf}$, then there exist a test design and an efficient algorithm that solve the group testing problem w.h.p. By sharp contrast, once the number of tests drops below $m_{\inf}$, identifying the set of infected individuals is information-theoretically impossible. Theorem 1 significantly improves over the best previous positive results. Indeed, the best previous efficient algorithm, a greedy algorithm called DD, requires

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

tests Johnson et al. (2018). The DD algorithm comes with the simple random bipartite test design where every individual independently joins an equal number of test groups, chosen uniform at random. While $m_{\mathtt{DD}}$ matches the optimal bound $m_{\inf}$ from Theorem 1 for densities $\theta \geq 1/2$, the bounds diverge for $\theta < 1/2$.

Turning to the information-theoretic lower bound, the best prior result derived from the folklore observation the total number 2^m of conceivable tests results must asymptotically exceed the number $\binom{n}{k}$ of possible sets of infected individuals to answer correctly w.h.p. Hence, $2^m \geq (1+o(1))\binom{n}{k}$. Applying Stirling's formula, we obtain the lower bound

$$m_{\mathrm{ad}} = \frac{1-\theta}{\ln 2} n^\theta \ln n. \tag{1.3}$$

This bound matches $m_{\inf}$ for $\theta \leq \ln(2)/(1+\ln 2) \approx 0.41$, but the bounds differ for larger θ.

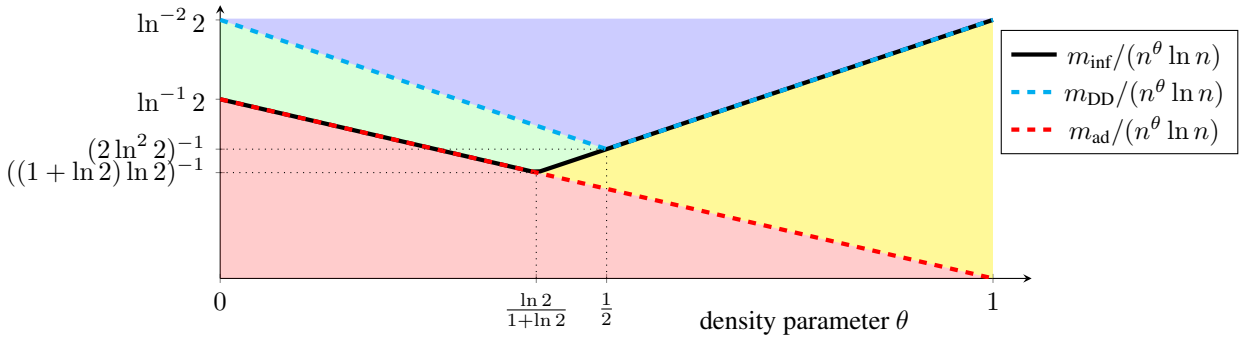

Figure 1: The best previously known algorithm DD succeeds in the blue but not in the green region. The new algorithm SPIV additionally succeeds in the green region. The black line indicates the non-adaptive information-theoretic threshold m_{inf}. In the red area even (multi-stage) adaptive inference is impossible. Finally, the two-stage adaptive group testing algorithm from Theorem 3 additionally succeeds in the yellow region.

Adaptive group testing. In the *adaptive* variant of the group testing problem, several stages of testing are allowed. In each stage the test design can take into account the outcomes of the tests conducted in the previous rounds. Apart from, naturally, minimising the total number of tests, in adaptive group testing we also aim to minimise the number of test stages Baldassini et al. (2013). A minimum number of stages is desirable because tests may be time-consuming Chen and Hwang (2008); Kwang-Ming and Ding-Zhu (2006). The elementary lower bound (1.3) shows that even adaptive test designs with an unlimited number of stages require at least m_{ad} tests. Conversely, the following theorem shows that actually just two stages suffice to reach the threshold m_{ad}.

Theorem 3 *For any $0 < \theta < 1$, $\varepsilon > 0$ 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 $\boldsymbol{\sigma}$ w.h.p.*

Theorem 3 improves over the three-stage test design from Scarlett (2019). The proof of Theorem 3 combines the test design and algorithm from Theorem 1 with ideas from Scarlett (2018).

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 Aldridge et al. (2014); Aldridge (2017); Aldridge et al. (2019); Johnson et al. (2018). Theorems 1–3 answer this question comprehensively. Namely, for infection densities $\theta < \ln(2)/(1+\ln(2)) \approx 0.41$ the non-adaptive test design from Theorems 1 matches the adaptive lower bound m_{ad}. Thus, in this regime adaptivity confers no advantage. By contrast, Theorem 2 shows that for $\theta > \ln(2)/(1+\ln(2))$ there is a widening gap between m_{ad} and the number of tests required by the optimal non-adaptive test design. Further, Theorem 3 demonstrates that this gap can be closed by allowing merely two stages of tests. Figure 1 illustrates the thresholds obtained in Theorems 1–3.

2. Overview: the new algorithm and the new lower bound

We describe the test design and algorithm for Theorem 1 and sketch the key ideas behind the proof of the information-theoretic lower bound for Theorem 2. We begin by discussing the random design harnessed in Aldridge et al. (2016); Johnson et al. (2018) to highlight several key concepts.

2.1. The random bipartite design

A natural first stab at a promising test design seizes upon a simple random bipartite (multi-)graph model. One vertex class $V = \{x_1, \ldots, x_n\}$ comprises the individuals. The other class $F = \{a_1, \ldots, a_m\}$ represents the tests. The edges are induced by having each individual independently join an equal number Δ of test, chosen uniformly at random with replacement. For an individual x_h let ∂x_h be the set of tests that it joins. Similarly, for a test a_i let ∂a_i be the set of test participants.

How should we choose Δ to extract the maximum amount of information? It seems natural to maximise the entropy of the vector of test results

$$\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}_a)_{a \in F} \qquad \text{with} \qquad \hat{\boldsymbol{\sigma}}_a = \max_{x \in \partial a} \boldsymbol{\sigma}_x. \tag{2.1}$$

Naturally, the average test degree equals $\Delta n/m$. Thus, the average number of infected individuals per test equals $\Delta k/m$. More precisely, since k is much smaller than n, the number of infected individuals in test a_i asymptotically has a Poisson distribution. Hence, choosing $\Delta \sim m \ln(2)/k$ effects that a test contains $\mathrm{Po}(\ln 2)$ infected individuals. Thus, about half the tests are positive w.h.p.

Given the test results, how do we best set about inferring the infected individuals? Clearly, every individual that occurs in a negative test is uninfected. Furthermore, each individual x_h that appears in a positive test a_i whose other participants all occur in negative tests must be infected; for x_h being infected is the only possible explanation of a_i being positive. Thus, we are left with two sets of individuals that it may be difficult to diagnose. First, the set V_{0+} of uninfected x that appear in positive tests only, i.e., *potential false positives*. Second, the set V_{1+} of infected individuals x that only appear in tests a that contain a second infected individual, i.e., *potential false negatives*. Clearly, if m is so small that both sets V_{0+}, V_{1+} are non-empty w.h.p., then inferring the set of infected individuals is impossible Coja-Oghlan et al. (2019b). This is because the test results remain unchanged if we declare any one individual $x \in V_{0+}$ infected and another $x \in V_{1+}$ uninfected.

But once m exceeds $m_{\inf}$, the set V_{1+} of false potential negatives is empty w.h.p. In effect, even though the set V_{0+} of potential false positives may still be non-empty, the set of infected individuals can be inferred by computing the assignment $\sigma \in \{0,1\}^V$ of minimum Hamming weight that 'explains' the test results Coja-Oghlan et al. (2019b). The problem of finding this σ can be expressed as a minimum hypergraph vertex cover problem. Thus, while the problem could be solved in exponential time, even on the random hypergraph no polynomial time vertex cover algorithm is known. Indeed, the problem is similar in flavour to the notorious planted clique problem (Alon et al. (1998)). In summary, the algorithmic challenge in group testing is to discriminate between the potential false positives V_{0+} and actual infected individuals.

Finally, matters improve once the number m of tests exceeds the bound $m_{\mathtt{DD}}$ from (1.2). Then the set V_{0+} of potential false positives is much smaller than the set of infected individuals w.h.p. Therefore, the expansion properties of the random graph allow to identify the infected individuals. Indeed, Johnson et al. (2018) show that a simple greedy algorithm known as DD (for 'Definitive Defectives') succeeds. In its first step DD marks all individuals that appear in negative tests as uninfected. Then it labels as infected every individual that appear in a positive tests whose other individuals have all been marked uninfected by the first step. All remaining individuals are marked uninfected. No better algorithm was known previously.

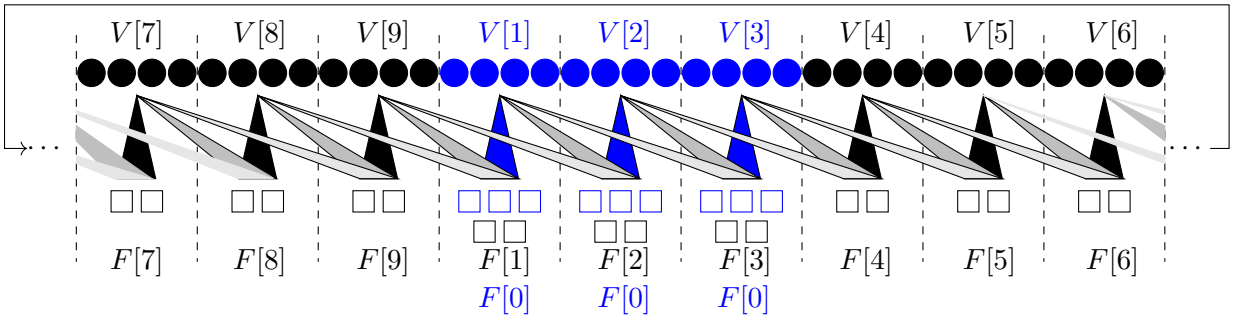

Figure 2: The spatially coupled test design with $n = 36, \ell = 9, s = 3$. The individuals in the seed groups $V[1] \cup \cdots \cup V[s]$ (blue) are equipped with additional test $F[0]$ (blue rectangles). The black rectangles represent the tests $F[1] \cup \cdots \cup F[\ell]$.

2.2. The new test design

To better discriminate between potential false positives and actual infected individuals we devise a new test design with a superimposed geometric structure. Specifically, we divide both the individuals and the tests into $\ell = \lceil \ln^{1/2} n \rceil$ compartments of equal size. The compartments are arranged in a ring and each individual joins an equal number of random tests in the $s = \lceil \ln \ln n \rceil = o(\ell)$ subsequent compartments along the ring. To get the algorithm started, we equip the first s compartments of individuals with additional tests so that they can be easily diagnosed via a greedy strategy. Then the algorithm will work its way along the ring, diagnosing one compartment after the other guided by the information gathered on the previous compartments. The construction of the test design is inspired by the recently discovered spatially coupled linear codes Felstrom and Zigangirov (1999); Kudekar et al. (2011, 2013).

To make this idea precise partition the set $V = \{x_1, \ldots, x_n\}$ of individuals into pairwise disjoint subsets $V[1], \ldots, V[\ell]$ of size $|V[j]| \sim n/\ell$ each. With these compartments we associate sets $F[1], \ldots, F[\ell]$ of tests of equal sizes $|F[i]| = m/\ell$, where we let $\varepsilon = \ln^{-1/3} n$ and where

$$m = (1+\varepsilon) \max\left\{\frac{\theta}{\ln^2 2}, \frac{1-\theta}{\ln 2}\right\} k \ln(n) + O(\ell) \tag{2.2}$$

is an integer divisible by ℓ. Additionally, we introduce a set $F[0]$ of $2ms/\ell$ extra tests to facilitate the algorithm for diagnosing the first s compartments. Thus, the total number of tests comes to

$$|F[0]| + \sum_{i=1}^{\ell} |F[i]| = (1 + 2s/\ell)m \sim m_{\inf}.$$

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 $\Delta = m \ln(2)/k + O(s)$ be an integer divisible by s For $i = 1, \ldots, \ell$ and $j = 1, \ldots, s$ every individual $x \in V[i]$ joins Δ/s tests in the set $F[i+j-1]$. These tests are chosen uniformly at random with replacement. All choices are mutually independent. Thus, 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. Additionally, each individual from the seed $V[1] \cup \cdots \cup V[s]$ joins 2Δ independently

chosen random tests from $F[0]$, drawn uniformly with replacement. Figure 2 illustrates the test design. We think of this random test design, denoted by $\boldsymbol{G}$, as a random bipartite (multi-)graph with vertex classes $V = \{x_1, \dots, x_n\}$ and $F = F[0] \cup F[1] \cup \dots \cup F[\ell]$. The set of neighbours of a vertex v of $\boldsymbol{G}$ is denoted by ∂v. Moreover, $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}_a)_{a \in F[0] \cup \dots \cup F[\ell]}$ signifies the vector of test results as defined in (2.1).

2.3. The Spatial Inference Vertex Cover ('SPIV') algorithm

We aim to infer the vector $\boldsymbol{\sigma}$ from the test results $\hat{\boldsymbol{\sigma}}$ and, of course, the test design $\boldsymbol{G}$. The algorithm for Theorem 1 proceeds in three phases.

Phase 1: the seed. 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. This is what the tests $F[0]$ comprising individuals from the seed compartments $V[1], \dots, V[s]$ are for. Thus, in its first phase the SPIV algorithm simply applies the DD algorithm of Aldridge et al. (2014) to identify the infected individuals among $V[1], \dots, V[s]$ from the tests $F[0]$. The vector τ signifies the algorithm's current estimate of $\boldsymbol{\sigma}$.

Input: $\boldsymbol{G}$, $\hat{\boldsymbol{\sigma}}$, k, $\varepsilon > 0$
Output: estimate of $\boldsymbol{\sigma}$
Initialise $\tau_x = 0$ for all individuals x Let $(\tau_x)_{x \in V[1] \cup \dots \cup V[s]} \in \{0,1\}^{V[1] \cup \dots \cup V[s]}$ be the result of applying DD to the tests $F[0]$
Algorithm 1: SPIV, phase 1

Crucially, we only apply the DD algorithm to the seed $V[1] \cup \dots \cup V[s]$ with merely $ns/\ell = o(n)$. Therefore, the number of dedicated seed tests $F[0]$, while exceeding the number of tests required by DD, is negligible by comparison to m. The following proposition, which follows from the analysis of DD from Johnson et al. (2018), summarises where we stand at the end of phase 1. Its proof can be found in Section 4.5 of the full version of this paper Coja-Oghlan et al. (2019a).

Proposition 4 *W.h.p. the output of* DD *satisfies* $\tau_x = \boldsymbol{\sigma}_x$ *for all* $x \in V[1] \cup \dots \cup V[s]$.

Phase 2, first attempt: enter the ring. This is the main phase of the algorithm. Thanks to Proposition 4 we may assume that the seed has been diagnosed correctly. Now, the grand strategy is to diagnose one compartment after the other, based on what the algorithm learned previously. Hence, assume that we managed to diagnose compartments $1, \dots, i$ correctly. How do we proceed to compartment $i+1$? For a start, we can safely mark as uninfected all individuals in $V[i+1]$ that appear in a negative test. Unfortunately, this will still leave us with $n^{1-\max\{(1-\theta)\ln 2, \theta\}+o(1)} \gg k$ undiagnosed individuals w.h.p. Thus, only a small fraction of the as yet undiagnosed individuals in $V[i+1]$ are actually infected. Hence, we need to discriminate between the set

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

of 'potential false positives', i.e., uninfected individuals that fail to appear in any negative test, 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 $1 \le j \le s$ let $\boldsymbol{W}_{x,j}$ be the number

of tests in compartment $F[i+j-1]$ that contain x but that do not contain an infected individual from the preceding compartments $V[1] \cup \cdots \cup V[i]$. In formulas,

$$\boldsymbol{W}_{x,j} = |\{a \in \partial x \cap F[i+j-1] : \partial a \cap (V_1[1] \cup \cdots \cup V_1[i]) = \emptyset\}|\,. \tag{2.3}$$

Crucially, the mean of $\boldsymbol{W}_{x,j}$ depends on whether x is infected or a potential false positive.

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 tests that individuals join are chosen independently, conditioning on x being infected does not skew the distribution of the individuals from the prior compartments $V[i+j-s+1], \ldots, V[i]$ that appear in a. Furthermore, we chose Δ, m so that for each of these compartments $V[h]$ the expected number of infected individuals that join a has mean $(\ln 2)/s$. Indeed, because the individuals choose their tests independently, $|V_1[h] \cap \partial a|$ is asymptotically Poisson. Consequently,

$$\mathbb{P}\left[V_1[h] \cap \partial a = \emptyset\right] \sim \exp(-(\ln 2)/s) = 2^{-1/s}. \tag{2.4}$$

Since, finally, the events $\{V_1[h] \cap \partial a = \emptyset\}_{h=i+j-s+1,\ldots,i}$ are mutually independent and x joins a total of Δ/s tests $a \in F[i+j]$, (2.4) implies

$$\mathbb{E}\left[\boldsymbol{W}_{x,j}\right] \sim 2^{-(s-j)/s}\Delta/s = 2^{j/s-1}\Delta/s. \tag{2.5}$$

Potential false positives ($x \in V_{0+}[i+1]$): Similarly 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 individual. In effect, the number of positives in a turns into a conditional Poisson $\mathrm{Po}_{\geq 1}(\ln 2)$. 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, an event that occurs with probability $(1+o(1))j/s$. Summing up the conditional Poisson and recalling that x appears in Δ/s tests $a \in F[j]$, we thus obtain

$$\mathbb{E}\left[\boldsymbol{W}_{x,j}\right] \sim \frac{\Delta}{s}\sum_{t\geq 1}\mathbb{P}\left[\mathrm{Po}_{\geq 1}(\ln 2) = t\right](j/s)^t = (2^{j/s}-1)\Delta/s. \tag{2.6}$$

Since $2^{j/s-1} > 2^{j/s} - 1$ for $j = 1, \ldots, s-1$, the mean (2.5) always exceeds (2.6). Therefore, we consider the sum $\boldsymbol{W}_x = \sum_{j=1}^{s-1}\boldsymbol{W}_{x,j}$, whose mean comes to

$$\mathbb{E}[\boldsymbol{W}_x] \sim \Delta \cdot \begin{cases} 1/(2\ln 2) & \text{if } x \in V_1[i+1], \\ (1-\ln 2)/\ln 2 & \text{if } x \in V_{0+}[i+1]. \end{cases}$$

Providing the algorithm made no mistake diagnosing the first i compartments, it can easily calculate $\boldsymbol{W}_x$ for every $x \in V[i+1]$ because the summands $\boldsymbol{W}_{x,j}$ depend on the test results and the infected individuals in $V[1] \cup \cdots \cup V[i]$ only. Thus, we could use the $\boldsymbol{W}_x$ to sieve out the potential false positives so long as no (or very few) $x \in V_{0+}[i+1]$ reach a value $\boldsymbol{W}_x$ as high as $\Delta/(2\ln 2)$, the mean for infected individuals.

Hence, we need to analyse the upper tail of $\boldsymbol{W}_x$ for $x \in V_{0+}[i+1]$. This large deviations analysis, though delicate, can be carried out precisely. But unfortunately the tail of $\boldsymbol{W}_x$ is too heavy. Even though for any specific $x \in V_{0+}[i+1]$ it is unlikely that $\boldsymbol{W}_x \geq \Delta/(2\ln 2)$, the outliers still exceed the number k of infected individuals w.h.p. Thus, it's back to the drawing board.

Phase 2, second attempt: optimal weights. The random variable $\boldsymbol{W}_x$ simply counts 'unexplained' positive tests that do not feature an infected individual from the known compartments $V[1], \dots, V[i]$. But not all of these tests reveal the same amount of information about x. For instance, we should really be paying more attention to 'early' unexplained tests $a \in F[i+1]$ than to 'late' tests $b \in F[i+s]$. Indeed, we already diagnosed $s-1$ out of the s compartments of individuals from which the participants of test a are drawn. Now, if $x \in V_{0+}[i+1]$ is a potential false positive, then a contains at least one infected individual, which thus belongs to $V[1] \cup \dots \cup V[i]$ with probability $(s-1)/s$. Hence, a large number of unexplained tests $a \in F[i+1]$ are quite a strong indication against x being a potential false positive. By contrast, only about a $1/s$ fraction of the individuals in a later test $b \in F[i+s]$ belong to the already recovered classes $V[1] \cup \dots \cup V[i]$. Such a positive test being 'unexplained' therefore does not have a very strong bearing on the status of x at all. Consequently, it seems promising to replace $\boldsymbol{W}_x$ by a weighted sum

$$\boldsymbol{W}^\star_x = \sum_{j=1}^{s-1} w_j \boldsymbol{W}_{x,j} \tag{2.7}$$

with suitably chosen weights $w_1, \dots, w_{s-1} \geq 0$. To choose $w_1, \dots, w_{s-1}$ optimally we investigate the tails of weighted sums of the form (2.7). From (2.5) we readily obtain the conditional mean of $\boldsymbol{W}^\star_x$ given x is infected:

$$\mathbb{E}\left[\boldsymbol{W}^\star_x \mid x \in V_1[i+1]\right] = \frac{\Delta}{s} \sum_{j=1}^{s-1} 2^{j/s-1} w_j. \tag{2.8}$$

Hence, we need to choose $w_1, \dots, w_{s-1}$ such that given $x \in V_{0+}[i+1]$ the probability of $\boldsymbol{W}^\star_x$ growing as large as (2.8) is minimised. A moderately intricate analysis reveals the large deviations rate function of $\boldsymbol{W}^\star_x$ given $x \in V_{0+}[i+1]$. We can therefore express this probability for *given* weights $w_1, \dots, w_s$ in terms of a convex optimisation problem $\mathcal{I}(w_1, \dots, w_{s-1})$:

$$\frac{1}{\Delta} \ln \mathbb{P}\left[\boldsymbol{W}^\star_x \geq \frac{\Delta}{s} \sum_{j=1}^{s-1} 2^{j/s-1} w_j \mid x \in V_{0+}\right] \sim -\mathcal{I}(w_1, \dots, w_{s-1}), \qquad \text{where}$$

$$\mathcal{I}(w_1, \dots, w_{s-1}) = \min_{0 \leq z_j \leq 1} \sum_{j=1}^{s-1} z_j \ln \frac{z_j}{2^{j/s}-1} + (1-z_j) \ln \frac{1-z_j}{2-2^{j/s}} \quad \text{s.t.} \quad \sum_{j=1}^{s-1} w_j (z_j - 2^{j/s-1}) = 0.$$

Thus, we need to maximise the objective function $\mathcal{I}(w_1, \dots, w_{s-1})$ on $w_1, \dots, w_{s-1}$. A delicate optimisation involving Lagrange multipliers leads to the optimal weights

$$w_j = \ln \frac{(1-2\varepsilon)2^{j/s-1}(2-2^{j/s})}{(1-(1-2\varepsilon)2^{j/s-1})(2^{j/s}-1)} \sim -\ln\left(1-2^{-j/s}\right) \tag{2.9}$$

The following lemma shows that with this optimal choice of weights the scores $\boldsymbol{W}^\star_x$ do indeed discriminate between the potential false positives and the infected individuals. The proof can be found in Section 4.8 of the full version Coja-Oghlan et al. (2019a).

Lemma 5 *With the weights* (2.9) *we have*

$$\mathbb{E}\left[\sum_{x \in V_{0+}[i+1]} \mathbf{1}\left\{\boldsymbol{W}^\star_x \geq (1-2\varepsilon)\frac{\Delta}{s} \sum_{j=1}^{s-1} 2^{j/s-1} w_j\right\}\right] \leq k n^{-\Omega(1)}.$$

Together with (2.8) and Markov's inequality, Lemma 5 shows that if we regard $x \in V[i+1]$ infected iff $\boldsymbol{W}^\star_x \geq (1-\varepsilon/2)\Delta$, then we misclassify no more than $kn^{-\Omega(1)} = o(k)$ individuals w.h.p.

Lemma 5 leaves us with two loose ends. First, calculating the scores $\boldsymbol{W}^\star_x$ involves the correct infection status $\boldsymbol{\sigma}_x$ of the individuals $x \in V[1] \cup \cdots \cup V[i]$ from the previous compartments. Naturally, while executing the algorithm we need to replace $\boldsymbol{\sigma}_x$ by the algorithm's estimate τ_x. Thus, the algorithm works with the approximate scores

$$W^\star_x(\tau) = \sum_{j=1}^{s-1} w_j \left|\{a \in \partial x \cap F[i+j-1] : \forall y \in \partial a : \tau_y = 0\}\right|. \tag{2.10}$$

To be precise, phase 2 of `SPIV` reads

for $i = s+1, \ldots, \ell$ **do**
  for $x \in V[i]$ **do**
    if $\exists a \in \partial x : \hat{\boldsymbol{\sigma}}_a = 0$ **then** $\tau_x = 0$ `// classify as healthy`;
    else if $W^\star_x(\tau) < (1-\varepsilon)\frac{\Delta}{s}\sum_{j=1}^{s-1} 2^{j/s-1} w_j$ **then** $\tau_x = 0$ `// tentative healthy`;
    else $\tau_x = 1$ `// tentative infected`;
  end
end

Algorithm 2: `SPIV`, phase 2.

The second issue is that phase 2 of `SPIV` is not going to classify *all* individuals correctly. Hence, there is risk of errors amplifying as we move from compartment to compartment. Fortunately, it turns out that errors proliferate only moderately. In effect, the second phase of `SPIV` will merely misclassify $kn^{-\Omega(1)} = o(k)$ individuals. The following proposition whose proof can be found in Section 4.9 of the full version Coja-Oghlan et al. (2019a) summarises the analysis of phase 2.

Proposition 6 *W.h.p. the assignment τ after phases 1 and 2 satisfies $\sum_{x \in V} \mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leq kn^{-\Omega(1)}$.*

Phase 3: clean-up. How do we correct the errors incurred during phase 2? A key insight is that w.h.p. every infected individual has at least $\Omega(\Delta)$ positive tests 'to itself', i.e., they do not feature a second infected individual. Phase 3 exploits this observation by simply thresholding the number U_x of tests where x is the unique supposedly infected individual. Thanks to the expansion properties of the graph $\boldsymbol{G}$, 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 w.h.p. Of course, by Proposition 4 we do not need to reconsider the individuals in the seed $V[1] \cup \cdots \cup V[s]$. Details can be found in Section 4.10 of the full version Coja-Oghlan et al. (2019a).

Proposition 7 *W.h.p. for all i we have $\sum_{x \in V} \mathbf{1}\{\tau_x^{(i+1)} \neq \boldsymbol{\sigma}_x\} \leq \frac{1}{3}\sum_{x \in V} \mathbf{1}\{\tau_x^{(i)} \neq \boldsymbol{\sigma}_x\}$.*

Proof of Theorem 1. The theorem is an immediate consequence of Propositions 4, 6 and 7. ■

2.4. The information-theoretic lower bound

The proof of Theorem 2 begins with an elementary (and well known) but crucial observation. Suppose that any G is a test design with a set $V(G)$ of n individuals and a set $F(G)$ of tests. As before let $\boldsymbol{\sigma}$ be the random $\{0,1\}$-vector with precisely k ones that indicates which individuals are infected

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

Algorithm 3: `SPIV`, phase 3.

and let $\hat{\boldsymbol{\sigma}}$ be the vector of test results. Further, let $\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})$ be the set of all vectors $\sigma \in \{0,1\}^{V(G)}$ of Hamming weight k that render the same test results $\hat{\boldsymbol{\sigma}}$, i.e., for every test $a \in F(G)$ we have $\max_{x \in \partial_G a} \boldsymbol{\sigma}_x = \max_{x \in \partial_G a} \sigma_x$. Then Bayes' rule immediately yields the following.

Fact 8 *The posterior of $\boldsymbol{\sigma}$ given $\hat{\boldsymbol{\sigma}}$ is the uniform distribution on $\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})$.*

Consequently, for any test design the information-theoretically optimal (albeit not generally efficient) inference algorithm is to simply output a uniform sample from $\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})$. Hence, $\boldsymbol{\sigma}$ can be inferred correctly w.h.p. from $\hat{\boldsymbol{\sigma}}$ iff $|\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})| = 1$ w.h.p. Thus, in order to prove an information-theoretic lower bound it suffices to prove that $\mathbb{P}\left[|\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})| > 1\right] = \Omega(1)$ for all test designs G. To prove Theorem 2 we proceed in two steps. First, we establish a lower bound for θ close to one.

Proposition 9 *For any $\eta > 0$ there exists $\theta_0 = \theta_0(\eta) < 1$ such that for all $\theta \in (\theta_0, 1)$ uniformly for all test designs G with $|F(G)| \leq (1-\eta) m_{\inf}(n, \theta)$ tests we have $|\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})| = n^{\Omega(1)}$ w.h.p.*

The somewhat subtle proof of Proposition 9 whose details are included in Section 3.2 of the full version Coja-Oghlan et al. (2019a) relies on two ingredients. First, we notice that there is no point in G having very big tests $a \in F(G)$ that contain more than, say, $n^{1-\theta} \ln(n)$ individuals. This is because w.h.p. all such tests are positive; they could therefore simply be replaced by constants. As a consequence, double counting shows that very few individuals occur in, say, more than $\ln^3 n$ tests. Thus, the bipartite graph representation of G is relatively sparse, the sparser the closer θ approaches one. Second, we adapt an argument from Aldridge's proof Aldridge (2019) of the information-theoretic lower bound for $k = \Theta(n)$. That proof does not extend directly to the sublinear scaling $k = n^\theta$ as the argument only shows that the event $|\mathcal{S}_k(G, \hat{\boldsymbol{\sigma}})| > 1$ occurs with a probability that tends to zero. However, one key step of the proof based on the FKG inequality can be used to show that *there exists* an individual y that is either a potential false positive or a potential false negative with a small but not extremely small probability. In formulas,

$$\mathbb{P}\left[y \in V_{0+}(G, \hat{\boldsymbol{\sigma}})\right] \geq (1 + o(1)) \exp\left(-\ln^2(2) |F(G)|/k\right), \tag{2.11}$$

$$\mathbb{P}\left[y \in V_{1+}(G, \hat{\boldsymbol{\sigma}})\right] \geq (1 + o(1)) n^{\theta - 1} \exp\left(-\ln^2(2) |F(G)|/k\right). \tag{2.12}$$

Unlike in the case that Aldridge considered, the probabilities on the r.h.s. tend to zero. However, because the graph G is quite sparse, we can construct a relatively big set Y of at least $n^{1-4(1-\theta)}$ variables y for which the bounds (2.11)–(2.12) hold such that the events $\{y \in V_{1+}(G, \hat{\boldsymbol{\sigma}})\}_{y \in Y}$ are only weakly correlated, and similarly for $\{y \in V_{0+}(G, \hat{\boldsymbol{\sigma}})\}_{y \in Y}$. This enables us to conclude that

both $|V_{0+}(G,\hat{\boldsymbol{\sigma}})|, |V_{1+}(G,\hat{\boldsymbol{\sigma}})| = n^{\Omega(1)}$ w.h.p., provided that θ is sufficiently close to one. Finally, as we saw in Section 2.1 already, if $|V_{0+}(G,\hat{\boldsymbol{\sigma}})|, |V_{1+}(G,\hat{\boldsymbol{\sigma}})| = n^{\Omega(1)}$, then $|\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})| = n^{\Omega(1)}$.

The second step towards Theorem 2 is a reduction from larger to smaller values of θ. Due to the elementary lower bound (1.3) we may confine ourselves to $\theta > \ln(2)/(1+\ln 2)$.

Proposition 10 *Let $\eta > 0$ and $\ln(2)/(1+\ln(2)) < \theta < \theta' < 1$. If there exists a test design G with $|F(G)| \leq (1-2\eta)m_{\inf}(n,\theta)$ tests such that $|\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})| = t$ for $t \in \mathbb{N}$ w.h.p., then there exists a test design G' with $n' \sim n^{\theta/\theta'}$ individuals and $|F(G')| \leq (1-\eta)m_{\inf}(n',\theta')$ tests such that $|\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})| = t$ w.h.p.*

The idea behind the proof of Proposition 10 is to add to the $n' \sim n^{\theta/\theta'} = o(n)$ individuals for G' another $n-n'$ uninfected dummies, thereby bringing the infection density down from θ' to θ. Then the test design G can be applied to identify the infected individuals, and the dummies can just be disregarded. The proof can be found in Section 3.3 of the full version Coja-Oghlan et al. (2019a).

Proof of Theorem 2. Assume that for a $\theta > \ln(2)/(1+\ln(2))$ a test design G with $(1-2\eta)m_{\inf}(n,\theta)$ tests and $|\mathcal{S}_k(G,\hat{\boldsymbol{\sigma}})| = 1$ w.h.p. exists. Then Proposition 10 shows that for θ' arbitrarily close to one for infinitely many n' a successful test design with $(1-\eta)m_{\inf}(n',\theta')$ tests exists, in contradiction to Proposition 9. Moreover, Proposition 9 evinces that there are infinitely many indistinguishable configurations, which by Proposition 10 and the generalized pigeonhole principle also exist for all $\ln(2)/(1+\ln(2)) < \theta < 1$. Thus, choosing a configuration uniformly at random will not return the correct configuration w.h.p. ■

2.5. Optimal adaptive group testing

We finally come to the proof of Theorem 3. To obtain the optimal two-stage algorithm we combine the first two phases of the SPIV algorithm with an idea from Scarlett (2018). Specifically, the test design for the first stage is identical to the one from Section 2.2 with m in (2.2) replaced by

$$m = (1+\varepsilon)\frac{1-\theta}{\ln 2}k\ln(n) + O(\ell).$$

We continue to let τ signify the result of phases 1 and 2 of SPIV. Going over the analysis of the first two phases with the value of m above, we obtain a bound on the number of misclassified individuals.

Lemma 11 *For any $\theta \in (0,1)$ we have $\sum_{x\in V}\mathbf{1}\{\tau_x \neq \boldsymbol{\sigma}_x\} \leq kn^{-\Omega(1)}$.*

We prove Lemma 11 in Section 4.9 of the full version Coja-Oghlan et al. (2019a). Scarlett (2018) observed that a single-stage group testing scheme that correctly diagnoses all but $o(k)$ individuals with $(1+o(1))m_{\mathrm{ad}}$ tests can be turned into a two-stage design with $(1+o(1))m_{\mathrm{ad}}$ tests in total that diagnoses all individuals correctly w.h.p. What he was missing at the time was such an optimal single-stage test design (and algorithm). We combine Scarlett's construction with Lemma 11 to obtain Theorem 3. Specifically, in the second stage we test all individuals x that SPIV diagnosed as infected separately. Lemma 11 implies that the total number of tests required comes to $(1+o(1))k = o(m_{\mathrm{ad}})$. Additionally, we set up a new group testing instance with individuals $V_0(\tau) = \{x \in V : \tau_x = 0\}$ and $m_0(\tau) \sim (1+\varepsilon)k$ tests. Specifically, the test design for this instance is simply the purely random test design from Section 2.1. Because Lemma 11 implies that the number of infected indivduals in $V_0(\tau)$ is bounded by $kn^{-\Omega(1)}$ w.h.p., the number $m_0(\tau)$ of tests suffices to ensure that the DD algorithm will correctly diagnose all individuals in $V_0(\tau)$ w.h.p. Thus, the second stage requires a total of $(1+o(1))k = o(m_{\mathrm{ad}})$ tests to output $\boldsymbol{\sigma}$ w.h.p.

3. Discussion

Group Testing. The group testing problem was first raised in Dorfman (1943), where Dorfman proposed a two-stage adaptive test design. In a first round disjoint groups of equal size are tested. All members of negative test groups are uninfected. Then, in the second round the members of positive test groups would be tested individually. Of course, this test design is far from optimal. The first multi-round test design that meets the adaptive information-theoretic bound mad was proposed in Allemann (2013), building upon Hwang (1972). Scarlett (2019) improved this result by proposing a three-stage test design. Finally, Theorem 3 achieves the optimal result, namely a non-adaptive (i.e., single stage) algorithm for $\theta < ln(2)/(1+ln(2))$ and a two-stage algorithm for larger θ. Regarding non-adaptive group testing, Aldridge proved that in the case $k = \Theta(n)$ where a constant fraction of individuals are infected, the design that tests each individual separately is information-theoretically optimal Aldridge et al. (2019). As a consequence, recent research has focused on the sub-linear case $k \sim n^{\theta}$ for $\theta \in (\theta, 1)$ e.g., Aldridge et al. (2016); Aldridge (2017); Coja-Oghlan et al. (2019b); Mézard et al. (2008); Scarlett and Cevher (2016), which this paper also considers. This scaling is practically relevant because Heap's law in epidemiology predicts that certain infections spread sublinearly in the total population size Benz et al. (2008). The best previous test design was the plain random bipartite one as described in Section 2.1. Several inference algorithms were proposed for this test design, with the simple DD algorithm achieving the best previously known algorithmic bound Aldridge et al. (2014); Chan et al. (2011); Mézard and Toninelli (2011) showed that a specific class of algorithms to which DD belongs is not able of reach the universal information-theoretic lower bound in two stages, let alone non-adaptive group testing.

Spatial coupling. The new test design for the `SPIV` algorithm is inspired by recent advances in coding theory known as spatially coupled low-density parity check codes Felstrom and Zigangirov (1999); Kudekar et al. (2011, 2013). The Tanner graphs (or parity check matrices) upon which such codes are based exhibit a spatial structure similar to our test design, with the bits of the code word partitioned into compartments arranged along a line segment. The Tanner graph is a random graph with a bounded average degree. Spatially coupled LDPC codes are known to asymptotically achieve capacity on binary memoryless channels Kumar et al. (2014). These codes come with an efficient decoding algorithm based on the Belief Propagation message passing scheme. The idea of spatial coupling has been extended to a few other inference problems, with compressed sensing possibly being the best known example Donoho (2006); Donoho et al. (2013); Krzakala et al. (2012); Kudekar and Pfister (2010). The inference algorithm in this case is based an approximate version of Belief Propagation known as Approximate Message Passing. The algorithm, which runs on a dense graph, meets the information-theoretic bound for compressed sensing.

Outlook. By comparison to prior versions of spatial coupling, a novelty here is that we obtain a simple combinatorial inference algorithm based merely on computing the weighted sum (2.7). This weighted sum incorporates a natural random variable that discriminates between positives and false positives and the analysis is based on a subtle but conceptually transparent large deviations analysis.

This technique of blending combinatorial ideas with the application of spatial coupling promises to be an exciting route for future research. Potential applications include the noisy versions of group testing, the quantitative group testing problem or the coin weighing problem Alaoui et al. (2019). Above and beyond these immediate extensions, it would be most interesting to see if the strategy behind `SPIV` extends to other inference problems that aim at learning sparse data.

Acknowledgments

We thank Arya Mazumdar for bringing the group testing problem to our attention. Amin Coja-Oghlan, Oliver Gebhard and Philipp Loick are supported by DFG CO 646/3.

References

A. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. Jordan. Decoding from pooled data: Phase transitions of message passing. *IEEE Transactions on Information Theory*, 65:572–585, 2019.

M. Aldridge. The capacity of bernoulli nonadaptive group testing. *IEEE Transactions on Information Theory*, 63:7142–7148, 2017.

M. Aldridge. Individual testing is optimal for nonadaptive group testing in the linear regime. *IEEE Transactions on Information Theory*, 65:2058–2061, 2019.

M. Aldridge, L. Baldassini, and O. Johnson. Group testing algorithms: bounds and simulations. *IEEE Transactions on Information Theory*, 60:3671–3687, 2014.

M. Aldridge, O. Johnson, and J. Scarlett. Improved group testing rates with constant column weight designs. *IEEE Transactions on Information Theory*, pages 1381–1385, 2016.

M. Aldridge, O. Johnson, and J. Scarlett. *Group testing: an information theory perspective*. Foundations and Trends in Communications and Information Theory, 2019.

A. Allemann. An efficient algorithm for combinatorial group testing. In H. Aydinian, F. Cicalese, and C. Deppe, editors, *Information Theory, Combinatorics, and Search Theory: In Memory of Rudolf Ahlswede*, pages 569–596. Springer, 2013.

N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures & Algorithms*, 13:457–466, 1998.

L. Baldassini, O. Johnson, and M. Aldridge. The capacity of adaptive group testing. *Proc. ISIT*, pages 2676–2680, 2013.

R. Benz, S. Swamidass, and P. Baldi. Discovery of power-laws in chemical space. *Journal of Chemical Information and Modeling*, 48:1138–1151, 2008.

C. Chan, P. Che, S. Jaggi, and V. Saligrama. Non-adaptive probabilistic group testing with noisy measurements: near-optimal bounds with efficient algorithms. *49th Annual Allerton Conference on Communication, Control, and Computing*, pages 1832–1839, 2011.

H. Chen and F. Hwang. A survey on nonadaptive group testing algorithms through the angle of decoding. *Journal of Combinatorial Optimization*, 15:49–59, 2008.

A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal group testing. *arXiv preprint arXiv:1911.02287*, 2019a.

A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Information-theoretic and algorithmic thresholds for group testing. *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, 132:43:1–43:14, 2019b.

D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

D. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59:7434–7464, 2013.

R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.

A. Felstrom and K. Zigangirov. Time-varying periodic convolutional codes with low-density parity-check matrix. *IEEE Transactions on Information Theory*, 45:2181–2191, 1999.

F. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67:605–608, 1972.

O. Johnson, M. Aldridge, and J. Scarlett. Performance of group testing algorithms with near-constant tests per item. *IEEE Transactions on Information Theory*, 65:707–723, 2018.

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2:021005, 2012.

S. Kudekar and H. Pfister. The effect of spatial coupling on compressive sensing. *48th Annual Allerton Conference on Communication, Control, and Computing*, pages 347–353, 2010.

S. Kudekar, T. Richardson, and R. Urbanke. Threshold saturation via spatial coupling: Why convolutional ldpc ensembles perform so well over the bec. *IEEE Transactions on Information Theory*, 57:803–834, 2011.

S. Kudekar, T. Richardson, and R. Urbanke. Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Transactions on Information Theory*, 59:7761–7813, 2013.

S. Kumar, A. Young, N. Macris, and H. Pfister. Threshold saturation for spatially coupled ldpc and ldgm codes on bms channels. *IEEE Transactions on Information Theory*, 60:7389–7415, 2014.

H. Kwang-Ming and D. Ding-Zhu. Pooling designs and nonadaptive group testing: important tools for dna sequencing. *World Scientific*, 2006.

M. Mézard and C. Toninelli. Group testing with random pools: Optimal two-stage algorithms. *IEEE Transactions on Information Theory*, 57:1736–1745, 2011.

M. Mézard, M. Tarzia, and C. Toninelli. Group testing with random pools: phase transitions and optimal strategy. *Journal of Statistical Physics*, 131:783–801, 2008.

J. Scarlett. Noisy adaptive group testing: Bounds and algorithms. *IEEE Transactions on Information Theory*, 65:3646–3661, 2018.

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

J. Scarlett and V. Cevher. Phase transitions in group testing. *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA 2016:40–53, 2016.