Information-Theoretic and Algorithmic Thresholds for Group Testing

Amin Coja-Oghlan
Goethe University, Frankfurt, Germany
acoghlan@math.uni-frankfurt.de

Oliver Gebhard
Goethe University, Frankfurt, Germany
gebhard@math.uni-frankfurt.de

Max Hahn-Klimroth
Goethe University, Frankfurt, Germany
hahnklim@math.uni-frankfurt.de

Philipp Loick
Goethe University, Frankfurt, Germany
loick@math.uni-frankfurt.de

Abstract

In the group testing problem we aim to identify a small number of infected individuals within a large population. We avail ourselves to a procedure that can test a group of multiple individuals, with the test result coming out positive iff at least one individual in the group is infected. With all tests conducted in parallel, what is the least number of tests required to identify the status of all individuals? In a recent test design [Aldridge et al. 2016] the individuals are assigned to test groups randomly, with every individual joining an equal number of groups. We pinpoint the sharp threshold for the number of tests required in this randomised design so that it is information-theoretically possible to infer the infection status of every individual. Moreover, we analyse two efficient inference algorithms. These results settle conjectures from [Aldridge et al. 2014, Johnson et al. 2019].

2012 ACM Subject Classification Theory of computation → Theory and algorithms for application domains; Theory of computation → Bayesian analysis; Theory of computation → Machine learning theory

Keywords and phrases Group testing problem, phase transitions, information theory, efficient algorithms, sharp threshold, Bayesian inference

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.43

Category Track A: Algorithms, Complexity and Games

Related Version A full version of the paper is available at `https://arxiv.org/pdf/1902.02202.pdf`.

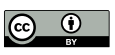
Funding *Amin Coja-Oghlan*: Supported of DFG 646/3.
Max Hahn-Klimroth: Supported by Stiftung Polytechnische Gesellschaft.
Philipp Loick: Supported of DFG 646/3.

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

1 Introduction

1.1 Background and motivation

The group testing problem goes back to the work of Dorfman from the 1940s [19]. Among a large population a few individuals are infected with a rare disease. The objective is to identify the infected individuals effectively. At our disposal we have a testing procedure capable of



46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 43; pp. 43:1–43:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

not merely testing one individual, but several. The test result will be positive if any one individual in the test group is infected, and negative otherwise; all tests are conducted in parallel. We are at liberty to assign a single individual to several test groups. The aim is to devise a test design that identifies the status of every single individual correctly while requiring as small a number of tests as possible.

A recently proposed test design allocates the individuals to tests randomly [7, 10, 11, 28, 32]. To be precise, given integers $n, m, \Delta > 0$ we create a random bipartite multi-graph by choosing independently for each of the n vertices $x_1, \ldots, x_n$ "on the left" Δ neighbours among the m vertices $a_1, \ldots, a_m$ "on the right" uniformly at random with replacement. The vertices $x_1, \ldots, x_n$ represent the individuals, the $a_1, \ldots, a_m$ represent the test groups and an individual joins a test group iff the corresponding vertices are adjacent. The wisdom behind this construction is that the expansion properties of the random bipartite graph precipitate virtuous correlations, facilitating inference.

Given n and (an estimate of) the number k of infected individuals, what is the least m for which, with a suitable choice of Δ, the status of every individual can be inferred correctly from the test results with high probability? Like in many other inference problems the answer comes in two instalments. First, we might ask for what m it is *information-theoretically* possible to detect the infected individuals. In other words, regardless of computational resources, do the test results contain enough information in principle to identify the infection status of every individual? Second, for what m does this problem admit *efficient algorithms*?

The first main result of this paper resolves the information-theoretic question completely. Specifically, Aldridge, Johnson and Scarlett [11] obtained a function $m_{\inf} = m_{\inf}(n, k)$ such that for any fixed $\varepsilon > 0$ the inference problem is information-theoretically infeasible if $m < (1-\varepsilon)m_{\inf}$. They conjectured that this bound is tight, i.e., that for $m > (1+\varepsilon)m_{\inf}(n, k)$ there is an (exponential) algorithm that correctly identifies the infected individuals with high probability. We prove this conjecture.

Furthermore, concerning the algorithmic question, Johnson, Aldridge and Scarlett [28] obtained a function $m_{\mathrm{alg}} = m_{\mathrm{alg}}(n, k)$ that exceeds $m_{\inf}$ by a modest constant factor such that for $m > (1 + \varepsilon)m_{\mathrm{alg}}$ certain efficient algorithms successfully identify the infected individuals with high probability. They conjectured that `SCOMP`, their most sophisticated algorithm, actually succeeds for smaller values of m. We refute this conjecture and show that `SCOMP` fails to outperform a much simpler algorithm called `DD`.

A technical novelty of the present work is that we investigate the group testing problem from a new perspective. While most prior contributions rely either on elementary calculations and/or information-theoretic arguments [10, 11, 28, 38, 39], here we bring to bear techniques from the theory of random constraint satisfaction problems [5, 31]. Indeed, group testing can be viewed naturally as a constraint satisfaction problem: the tests provide the constraints and the task is to find all possible ways of assigning a status ("infected" or "not infected") to the n individuals in a way consistent with the given test results. Since the allocation of individuals to tests is random, this question is similar in nature to, e.g., the random k-SAT problem that asks for a Boolean assignment that satisfies a random collection of clauses [4, 6, 16, 18]. Apart from obtaining the aforementioned new results, this novel perspective allows for short proofs of results that were established more laboriously in prior work. It also puts the group testing problem in the same framework as the considerable body of recent work on other inference problems on random graphs such as the stochastic block model (e.g., [1, 15, 17, 34, 35, 41]).

We proceed to state the main results of the paper precisely, followed by a detailed discussion of the prior literature on group testing. An outline of the proof strategy follows in Section 2.

1.2 The information-theoretic threshold

Throughout the paper we labour under the assumptions commonly made in the context of group testing; we will revisit their merit in Section 1.4. Specifically, we assume that the number k of infected individuals satisfies $k \sim n^\theta$ for a fixed $0 < \theta < 1$. Moreover, let $\boldsymbol{\sigma} \in \{0,1\}^{\{x_1,\ldots,x_n\}}$ be a vector of Hamming weight k chosen uniformly at random. The (one-)entries of $\boldsymbol{\sigma}$ indicate which of the n individuals are infected. Moreover, let $\boldsymbol{G} = \boldsymbol{G}(n, m, \Delta)$ signify the aforementioned random bipartite graph. Then $\boldsymbol{\sigma}$ induces a vector $\hat{\boldsymbol{\sigma}} \in \{0,1\}^{\{a_1,\ldots,a_m\}}$ that indicates which of the m tests come out positive. To be precise, $\hat{\boldsymbol{\sigma}}_i = 1$ iff test a_i is adjacent to an individual x_j with $\boldsymbol{\sigma}_{x_j} = 1$. For what m is it possible to recover $\boldsymbol{\sigma}$ from $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$? Here, we settle an important open question [28] on the sharpness of the information-theoretic threshold. (Throughout the paper all logarithms are base e.)

▶ **Theorem 1.** *Suppose that $0 < \theta < 1$, and $\varepsilon > 0$ and let*

$$m_{\inf} = m_{\inf}(n,\theta) = \frac{n^\theta(1-\theta)\log(n)}{\min\left\{1, \frac{1-\theta}{\theta}\log 2\right\}\log 2}.$$

- **(i)** *If $m < (1-\varepsilon)m_{\inf}(n,\theta)$, then there does not exist any algorithm that given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k$ outputs $\boldsymbol{\sigma}$ with a non-vanishing probability.*
- **(ii)** *If $m > (1+\varepsilon)m_{\inf}(n,\theta)$, then there exists an algorithm that given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ outputs $\boldsymbol{\sigma}$ with high probability.*

Since for $\theta \le \log(2)/(1+\log(2))$ the first part of Theorem 1 readily follows from a folklore argument [20], the interesting regime is $\theta > \log(2)/(1+\log(2)) \approx 0.41$. In this regime Theorem 1 strengthens a result from [11], who showed that for $m < (1-\varepsilon)m_{\inf}$ any inference algorithm has a strictly positive error probability. By comparison, Theorem 1 shows that any algorithm fails with *high* probability.

But the main contribution of Theorem 1 is the second, positive statement. While the case $\theta > 1/2$ is easy because a plain greedy algorithms succeeds [28], the case $\theta < 1/2$ proved more challenging and was so far only heuristically justified using techniques from statistical physics [32] and for $\theta < 1/3$ for a different test design [38, 39]. Indeed, Aldridge et al. [10] conjectured that in this case inferring $\boldsymbol{\sigma}$ from $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ is equivalent to solving a hypergraph minimum vertex cover problem. The proof of Theorem 1 vindicates this conjecture. Specifically, the vertex set of the hypergraph comprises all "potentially infected" individuals, i.e., those that do not appear in any negative test. The hyperedges are the neighbourhoods ∂a_i of the positive tests a_i in $\boldsymbol{G}$. Exhaustive search solves this vertex cover problem in time $\exp(O(n^\theta \log n))$. But how about efficient algorithms for general θ?

1.3 Efficient algorithms for group testing

Several polynomial time group testing algorithms have been proposed. A very simple greedy strategy called `DD` (for "definitive defectives") first labels all individuals that are members of negative test groups as healthy. Subsequently it checks for positive tests in which all individuals but one have been identified as healthy in the first step. Clearly, the single as yet unlabelled individual in such a test group must be infected. Up to this point all decisions made by `DD` are correct. But in the final step `DD` marks all as yet unclassified individuals as healthy, possibly causing false negatives. In fact, the output of `DD` may be inconsistent with the test results as possibly some positive tests may fail to spot an individual classified as "infected".

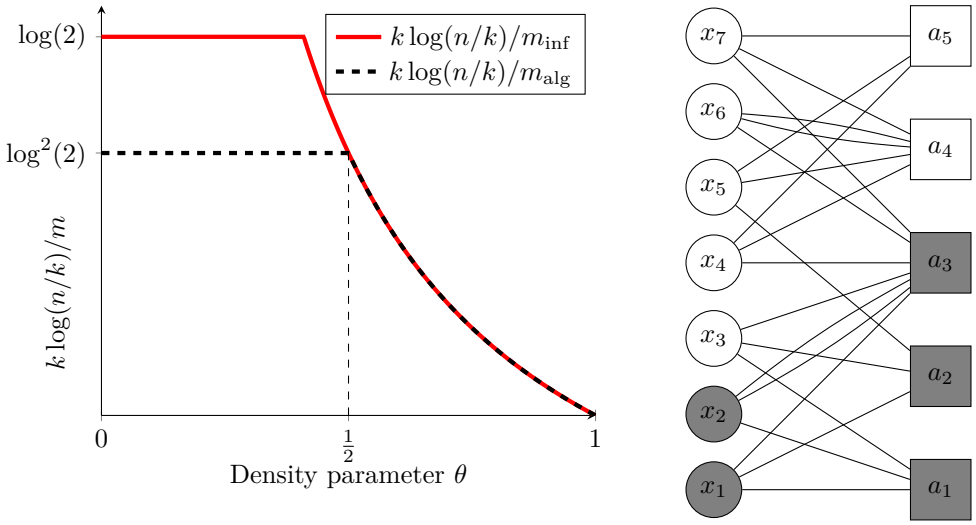


Figure 1 The left diagram displays $m_{\mathrm{inf}}, m_{\mathrm{alg}}$. The red line shows the information theoretic threshold m_{inf}, the dashed black line signifies the bound m_{alg} which is achieved by the both the SCOMP and the DD algorithm. The graph on the right illustrates a small example of a group testing instance, with the individuals $x_1, \dots, x_7$ on the left and the tests $a_1, \dots, a_5$ on the right. Infected individuals and positive tests are coloured in grey.

The more sophisticated SCOMP algorithm is roughly equivalent to the well-known greedy algorithm for the hypergraph vertex cover problem applied to the hypergraph from the previous paragraph. Specifically, in its first step SCOMP proceeds just like DD, classifying all individuals that occur in negative tests as healthy. Then SCOMP identifies as infected all unmarked individuals that appear in at least one test whose other participants are already known to be healthy. Subsequently the algorithm keeps picking an individual that appears in the largest number of as yet "unexplained" (viz. uncovered) positive tests and marks that individual as infected, with ties broken randomly, until every positive test contains an individual classified as infected. Clearly, SCOMP may produce false positives as well as false negatives. But at least the output is consistent with the test results.

Analysing SCOMP has been prominently posed as an open problem in the group testing literature [8, 10, 28]. Indeed, Aldridge et al. [10] opined that "the complicated sequential nature of SCOMP makes it difficult to analyse mathematically". On the positive side, [10] proved that SCOMP succeeds in recovering $\boldsymbol{\sigma}$ correctly given $(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ if $m > (1+\varepsilon)m_{\mathrm{alg}}(n,\theta)$ w.h.p., where

$$m_{\mathrm{alg}} = m_{\mathrm{alg}}(n,\theta) = \frac{n^{\theta}(1-\theta)\log(n)}{\min\left\{1, \frac{1-\theta}{\theta}\right\}\log^2 2}. \tag{1.1}$$

However, the algorithm succeeds for a trivial reason; namely, for $m > (1+\varepsilon)m_{\mathrm{alg}}$ even DD suffices to recover $\boldsymbol{\sigma}$ w.h.p. Yet based on experimental evidence [10, 28] conjectured that SCOMP strictly outperforms DD. The following theorem refutes this conjecture.

▶ **Theorem 2.** *Suppose that $0 < \theta < 1$ and $\varepsilon > 0$. If $m < (1-\varepsilon)m_{\mathrm{alg}}(n,\theta)$, then given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ w.h.p. both SCOMP and DD fail to output $\boldsymbol{\sigma}$.*

For $\theta < 1/2$ the information-theoretic bound provided by Theorem 1 and the algorithmic bound m_{alg} supplied by Theorem 2 remain a modest constant factor apart; see Figure 1. In some other inference problems on random graphs such as the stochastic block model similar

gaps appear between the information-theoretic and the algorithmic bounds [1, 17, 34, 41]. There have been attempts at investigating to what extent these gaps are due to genuine computational barriers, i.e., [23, 24, 25, 26]. Whether there actually exists a computationally hard regime for group testing, or whether the gap can be closed by smarter algorithms, remains an exciting question for future research.

1.4 Discussion and related work

Dorfman's original group testing scheme, intended to test the American army for syphilis, was *adaptive*. In a first round of tests each soldier would be allocated to precisely one test group. If the test result came out negative, none of the soldiers in the group were infected. In a second round the soldiers whose group was tested positively would then be tested individually. Of course, Dorfman's scheme was not information-theoretically optimal. An optimal adaptive scheme that involves several test stages, with the tests conducted in the present stage governed by the results from the previous stages, is known [20, 12]. In the adaptive scenario the information-theoretic threshold works out to be

$$m_{\mathrm{inf}}^{\mathrm{adapt}}(n,k) = \frac{n^\theta(1-\theta)\log(n)}{\log 2}.$$

The lower bound, i.e., that no adaptive design gets by with $(1-\varepsilon)m_{\mathrm{inf}}^{\mathrm{adapt}}(n,k)$ tests, follows from a very simple information-theoretic consideration. Namely, with a total of m tests at our disposal there are merely 2^m possible test outcomes, and we need this number to exceed the count $\binom{n}{k}$ of possible vectors $\boldsymbol{\sigma}$.

More recently there has been a great deal of interest in non-adaptive group testing, where the infection status of each individual is to be determined after just one round of tests [7, 9, 10, 11, 14, 22, 28, 32, 38, 39]. This is the version of the problem that we deal with in the present paper. An important advantage of the non-adaptive scenario is that tests, which may be time-consuming, can be conducted in parallel. Indeed, some of today's most popular applications of group testing are non-adaptive such as DNA screening [14, 30, 37] or protein interaction experiments [36, 40] in computational molecular biology. The randomised test design that we deal with here is the best currently known non-adaptive design (in terms of the number of tests required).

The most interesting regime for the group testing problem is when the number k of infected individuals scales as a power n^θ of the entire population. Mathematically this is because in the linear regime $k = \Omega(n)$ the optimal strategy is to perform n individual tests [9]. Thus, for k linear in n there is nothing interesting to do. But the sublinear case is also of practical relevance, as witnessed by Heap's law in epidemiology [13] or biological applications [22].

Apart from the randomised test design $\boldsymbol{G}$ where each individual chooses precisely Δ tests (with replacement), the so-called Bernoulli design assigns each individual to every test with a certain probability independently. A considerable amount of attention has been devoted to this model, and its information-theoretic threshold as well as the thresholds for various algorithms have been determined [8, 7, 10, 38, 39]. However, the Bernoulli test design, while easier to analyse, is provably inferior to the test design $\boldsymbol{G}$ that we study here. This is because in the Bernoulli design there are likely quite a few individuals that participate in far fewer tests than expected due to random degree fluctuations.

1.5 Notation

Throughout the paper $\boldsymbol{G} = \boldsymbol{G}(n, m, \Delta)$ denotes the random bipartite graph that describes which individuals take part in which test groups, the vector $\boldsymbol{\sigma} \in \{0,1\}^{\{x_1,\dots,x_n\}}$ encodes which individuals are infected, and $\hat{\boldsymbol{\sigma}} \in \{0,1\}^{\{a_1,\dots,a_m\}}$ indicates the test results. Moreover, $k \sim n^\theta$ signifies the number of infected individuals. Additionally, we write $V = V_n = \{x_1, \dots, x_n\}$ for the set of all individuals and $V_0 = \{x_i \in V : \boldsymbol{\sigma}_{x_i} = 0\}$, $V_1 = V \setminus V_0$ for the set of healthy and infected individuals, respectively. For an individual $x \in V$ we write ∂x for the set of tests a_i adjacent to x. Analogously, for a test a_i we denote by ∂a_i the set of individuals that take part in the test. Finally, all asymptotic notation refers to the limit $n \to \infty$. Thus, $o(1)$ denotes a term that vanishes in the limit of large n, while $\omega(1)$ stands for function that diverges to ∞ as $n \to \infty$.

2 Outline

We give an overview of the main arguments upon which the proofs of Theorems 1 and 2 rest.

2.1 The Nishimori identity

The very first item on the agenda is to get a handle on the posterior distribution of $\boldsymbol{\sigma}$ given $\boldsymbol{G}$ and $\hat{\boldsymbol{\sigma}}$. To this end, let $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ be the set of all vectors $\sigma \in \{0,1\}^V$ of Hamming weight k such that

$$\hat{\boldsymbol{\sigma}}_{a_i} = \mathbf{1}\{\exists x \in \partial a_i : \sigma_x = 1\} \qquad \text{for all } i \in [m].$$

In words, $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ contains the set of all vectors σ with k ones that label the individuals infected/healthy in a way consistent with the test results. Let $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = |S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})|$. The following proposition shows that the posterior of $\boldsymbol{\sigma}$ given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ is uniform on $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$.

▶ **Proposition 3** ([7]).
For all $\tau \in \{0,1\}^{\{x_1,\dots,x_n\}}$ *we have* $\mathbb{P}[\boldsymbol{\sigma} = \tau \mid \boldsymbol{G}, \hat{\boldsymbol{\sigma}}] = \dfrac{\mathbf{1}\{\tau \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})\}}{Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})}$.

Adopting the jargon of the recent literature on inference problems on random graphs, we refer to Proposition 3 as the *Nishimori identity* [15, 41]. The proposition shows that apart from the actual test results, there is no further "hidden information" about $\boldsymbol{\sigma}$ encoded in $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$. In particular, the information-theoretically optimal inference algorithm just outputs a uniform sample from $S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$. In effect, we obtain the following.

▶ **Corollary 4.**
1. *If* $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ *w.h.p., then for any algorithm* $\mathcal{A}$ *we have*

$$\mathbb{P}[\mathcal{A}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}] = o(1).$$

2. *If* $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 1$ *w.h.p., then there is an algorithm* $\mathcal{A}$ *such that*

$$\mathbb{P}[\mathcal{A}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}, k) = \boldsymbol{\sigma}] = 1 - o(1).$$

Both the positive and the negative part of Corollary 4 assume that the precise number k of infected individuals is known to the algorithm. This assumption makes the negative part stronger, but weakens the positive part. Yet we will see in due course how in the positive scenario the assumption that k be known can be removed. The upshot is that we need to get a handle on $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$.

2.2 The information-theoretic threshold

We proceed to discuss the proof of Theorem 1. The proofs of the first, negative statement and of the second, positive statement hinge on two separate arguments. We begin with the negative statement that w.h.p. $\boldsymbol{\sigma}$ cannot be inferred if $m < (1-\varepsilon)m_{\inf}$.

2.2.1 The information-theoretic lower bound

In light of Corollary 4 in order to prove the first part of Theorem 1 we need to show that the number $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ of assignments consistent with the test results $\hat{\boldsymbol{\sigma}}$ is unbounded w.h.p. The proof of this fact is based on a very simple idea: we just identify a biggish number of individuals whose infection status could be flipped without affecting the test results. To be precise, let $V_0^+ = V_0^+(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ be the set of all healthy individuals x_i such that every test in which x_i occurs is positive; in symbols,

$$V_0^+ = \{x_i \in V_0 : \forall a \in \partial x_i \exists y \in \partial a : \boldsymbol{\sigma}_y = 1\}. \tag{2.1}$$

Similarly, let V_1^+ be the set of all infected individuals x_i such that every test in which x_i occurs features another infected individual; in symbols,

$$V_1^+ = \{x_i \in V_1 : \forall a \in \partial x_i \exists y \in \partial a \setminus \{x_i\} : \boldsymbol{\sigma}_y = 1\}.$$

We think of the individuals in V_0^+ as the "potential false positives". Indeed, if for any $x_i \in V_0^+$ we obtain $\boldsymbol{\sigma}'$ from $\boldsymbol{\sigma}$ by setting x_i to one, then $\boldsymbol{\sigma}'$ will render the same test results as $\boldsymbol{\sigma}$. Similarly, the individuals in V_1^+ are potential false negatives.

The following lemma yields a bound on m below which potential false positives and negatives abound. A simple (omitted) calculation also yields the value of Δ that is optimal to facilitate inference, namely $\Delta = \lceil \frac{m}{k} \log 2 \rceil$.

▶ **Lemma 5.** *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that*

$$m < \frac{(1-\varepsilon)\theta}{(1-\theta)\log^2 2} n^\theta (1-\theta) \log n$$

Then even with the optimal choice $\Delta = \lceil \frac{m}{k} \log 2 \rceil$ we have $|V_0^+|, |V_1^+| = n^{\Omega(1)}$ w.h.p.

The proof of Lemma 5 relies on a basic random graphs argument. As an immediate application we obtain the following information-theoretic lower bound.

▶ **Corollary 6.** *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that*

$$m < \frac{(1-\varepsilon)\theta}{(1-\theta)\log^2 2} n^\theta (1-\theta) \log n \tag{2.2}$$

Then $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.

Proof. We need to exhibit alternative vectors $\boldsymbol{\sigma}' \in \{0,1\}^V$ with Hamming weight k that render the same test results as $\boldsymbol{\sigma}$. Thus, pick any $x_i \in V_0^+$ and any $x_j \in V_1^+$ and obtain $\boldsymbol{\sigma}'$ from $\boldsymbol{\sigma}$ by setting $\boldsymbol{\sigma}'_{x_i} = 1$ and $\boldsymbol{\sigma}'_{x_j} = 0$. By construction, $\boldsymbol{\sigma}'$ has Hamming weight k and renders the same test results. Hence, Lemma 5 shows that $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) \geq |V_0^+ \times V_1^+| = \Omega(n^{2\theta}) \gg 1$ w.h.p. ◀

The bound (2.2) matches $m_{\inf}$ for $\theta \gtrapprox 0.41$. A simpler, purely information-theoretic argument covers the remaining θ.

▶ **Lemma 7.** *Let $\varepsilon > 0$, $0 < \theta < 1$. If $m < \frac{1-\varepsilon}{\log 2} n^\theta (1-\theta) \log n$, then $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ w.h.p.*

We thus conclude that for all $0 < \theta < 1$, w.h.p. $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = \omega(1)$ if $m < (1-\varepsilon)m_{\inf}$. Therefore, the desired information-theoretic lower bound follows from Corollary 4.

2.2.2 The information-theoretic upper bound

The proof of the information-theoretic upper bound is the principal achievement of the present work. The proof rests upon techniques that have come to play an important role in the theory of random constraint satisfaction problems. Specifically, we need to show that $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 1$ w.h.p., i.e., that $\boldsymbol{\sigma}$ is the only assignment compatible with the test results w.h.p. We establish this result by combining two separate arguments. First, we use a moment calculation to show that w.h.p. there are no other solutions that have a small "overlap" with $\boldsymbol{\sigma}$. Then we use an expansion argument to show that w.h.p. there are no alternative solutions with a big overlap. Both these arguments are variants of the arguments that have been used to study the solution space geometry of random constraint satisfaction problems such as random k-SAT or random k-XORSAT [3, 4, 21], as well as the freezing thresholds of random constraint satisfaction problems [2, 33]. Yet to our knowledge these methods have thus far not been applied to the group testing problem.

Formally, we define

$$Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = |\{\sigma \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) : \langle \boldsymbol{\sigma}, \sigma \rangle = \ell\}|$$

as the number of assignments $\sigma \in S_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ whose *overlap* $\langle \boldsymbol{\sigma}, \sigma \rangle = \sum_{i=1}^{n} \mathbf{1}\{\sigma_{x_i} = \boldsymbol{\sigma}_{x_i} = 1\}$ with $\boldsymbol{\sigma}$ is equal to ℓ. The following two propositions rule out assignments with a small and a big overlap, respectively. In either case we choose $\Delta = \lceil \frac{m}{k} \log 2 \rceil$ to take its optimal value.

▶ **Proposition 8.** *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that $m > (1+\varepsilon) m_{\inf}(k, \theta)$. W.h.p. we have $Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 0$ for all $\ell < (1 - 1/\log n)k$.*

▶ **Proposition 9.** *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that $m > (1+\varepsilon) m_{\inf}(k, \theta)$. W.h.p. we have $Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 0$ for all $(1 - 1/\log n)k \leq \ell < k$.*

We defer the proofs of Propositions 8 and 9 to Sections 3 and 4, respectively.

Propositions 8 and 9 readily imply that $Z_k(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 1$ w.h.p. if $m > (1+\varepsilon) m_{\inf}(k, \theta)$. Hence, Corollary 4 shows that there exists an inference algorithm that given $\boldsymbol{G}, \hat{\boldsymbol{\sigma}}$ *and* k outputs $\boldsymbol{\sigma}$ w.h.p. However, up to now this algorithm relies on exactly knowing the number of infected individuals k, which in practice could be rather difficult to learn.

Fortunately this assumption can be removed. Namely, the following proposition shows that w.h.p. there is no assignment σ that is compatible with the test results and that has Hamming weight less than k.

▶ **Proposition 10.** *Let $\varepsilon > 0$ and $0 < \theta < 1$ and assume that $m > (1+\varepsilon) m_{\inf}(k, \theta)$. W.h.p. we have $\sum_{k' < k} Z_{k'}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) = 0$.*

As an immediate consequence of Proposition 10 we conclude that for $m > (1+\varepsilon) m_{\inf}(k, \theta)$ the problem of inferring $\boldsymbol{\sigma}$ boils down to a minimum vertex cover problem, as previously conjectured by Aldridge, Baldassini and Johnson [10]. Namely, let $\mathcal{P}$ be the set of all positive tests, i.e., all tests a_i, $i \in [m]$, with $\hat{\boldsymbol{\sigma}}_{a_i} = 1$. Moreover, let V^+ be the set of all variables $x_i \in V$ such that $\partial x_i \subseteq \mathcal{P}$; in words, x_i takes part in positive tests only. We set up a hypergraph $\boldsymbol{H}$ with vertex set V^+ and hyperedges $\partial a_i \cap V^+$, $a_i \in \mathcal{P}$. Clearly, the set of all individuals x_i with $\boldsymbol{\sigma}_{x_i} = 1$ provides a valid vertex cover of $\boldsymbol{H}$ (as any positive test must feature an infected individual). Conversely, Propositions 8 and 9 show that w.h.p. this is the unique vertex cover of size k, and Proposition 10 shows that there is no strictly smaller vertex cover w.h.p. Therefore, w.h.p. we can infer $\boldsymbol{\sigma}$ even without prior knowledge of k by way of solving this minimum vertex cover instance.

2.3 The SCOMP algorithm

For $\theta \geq 1/2$ we have $m_{\mathrm{alg}} = m_{\mathrm{inf}}$ and thus Theorem 1 implies that SCOMP fails to infer $\boldsymbol{\sigma}$ w.h.p. for $m < (1-\varepsilon)m_{\mathrm{alg}}$. Therefore, we are left to establish Theorem 2 for $\theta < 1/2$, in which case

$$m_{\mathrm{alg}} = \frac{n^{\theta}(1-\theta)\log(n)}{\log^2 2}. \tag{2.3}$$

The proof of Theorem 2 for $\theta < 1/2$ hinges on two lemmas. First we show that below m_{alg}, the set V_1^{--} of infected individuals that the second step of SCOMP identifies correctly is empty. Formally, with V_0^+ from (2.1),

$$V_1^{--} = \left\{x \in V_1 : \exists a \in \partial x : \partial a \setminus \{x\} \subseteq V_0 \setminus V_0^+\right\}.$$

▶ **Lemma 11.** *Suppose that $0 < \theta < 1/2$ and $\varepsilon > 0$. If $m < (1-\varepsilon)m_{\mathrm{alg}}$, then for all $\Delta > 0$ we have $V_1^{--}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}^*) = \emptyset$ w.h.p.*

With the second step of SCOMP failing to 'explain' (viz. cover) any positive tests, the greedy vertex cover algorithm takes over. This algorithm is applied to the hypergraph whose vertices are the as yet unclassified individuals and whose edges are the neighbourhoods of the positive tests. Our second lemma shows that the set $V^{+,\Delta}$ of potententially false positive individuals $x \in V_0^+$ that participate in the maximum number Δ of different test is far greater than the actual number k of infected individuals. Formally, let

$$V_0^{+,\Delta} = \left\{x \in V_0^+ : |\partial x| = \Delta\right\}.$$

▶ **Lemma 12.** *Suppose that $0 < \theta < 1/2$ and $\varepsilon > 0$. If $m < (1-\varepsilon)m_{\mathrm{alg}}$, then for all $\Delta > 0$ we have $V_0^{+,\Delta} \geq k \log n$ w.h.p.*

The proofs of Lemmas 11 and 12 are based on moment calculations that turn out to be mildly subtle due to the potentially very large degrees of the underlying graph $\boldsymbol{G}$. We complete the proof of Theorem 2 as follows.

Proof of Theorem 2. The first step of SCOMP (correctly) marks all individuals that appear in negative tests as healthy. Moreover, Lemma 11 implies that the second step of SCOMP is void w.h.p., because there is no single infected individual that appears in a test whose other individuals have already been identified as healthy by the first step. Consequently, SCOMP simply applies the greedy vertex cover algorithm. Now, thanks to Lemma 12 it suffices to prove that SCOMP will fail w.h.p. if $V_0^+ = \omega(k)$. Because they belong to positive tests only, all the individuals of V_0^+ are present in the vertex cover instance that SCOMP attempts to solve. Moreover, in the hypergraph no vertex has degree greater than Δ, because the degrees of $x_1, \ldots, x_n$ in $\boldsymbol{G}$ are equal to Δ. (Some of the hypergraph degrees may be strictly smaller than Δ because $\boldsymbol{G}$ is a multi-graph.) Therefore, since $|V_0^+| \geq k \log n$ while the actual set of infected individuals only has size k, w.h.p. the individual classified as infected by the very first step of the greedy set cover algorithm belongs to V_0^+. Hence, this individual is not actually infected, i.e., SCOMP errs w.h.p. ◀

Since the success probability of the SCOMP algorithm is at least as high as of the DD algorithm, we can prove the conjecture of [28] regarding the upper bound of the DD algorithm.

▶ **Corollary 13.** *If $m < (1-\varepsilon)m_{\mathrm{alg}}$, the DD algorithm will fail to retrieve the correct set of infected individuals w.h.p..*

3 Proof of Proposition 8

For $i \in [m]$ let Γ_i be the degree of a_i in $\boldsymbol{G}$, i.e., the number of edges incident with a_i; this number may exceed the number of different individuals that participate in test a_i as $\boldsymbol{G}$ may feature multi-edges. Let $\mathcal{G}$ be the σ-algebra generated by the random variables $(\Gamma_i)_{i\in[m]}$.

Given $\mathcal{G}$ we can generate $\boldsymbol{G}$ from the well-known *pairing model* [27]. Specifically, we create a set $\{x_i\} \times [\Delta]$ of Δ clones of each individual as well as sets $\{a_i\} \times [\Gamma_i]$ of clones of the tests. Then we draw a perfect matching of the complete bipartite graph on the vertex sets $\bigcup_{i=1}^{n} \{x_i\} \times [\Delta]$, $\bigcup_{i=1}^{m} \{a_i\} \times [\Gamma_i]$ uniformly at random. For each matching edge linking a clone of x_i with a clone of a_j we insert an i-j-edge. The resulting bipartite random multi-graph has the same distribution as $\boldsymbol{G}$ given $\mathcal{G}$. As an immediate application of this observation we obtain the following estimate.

▶ **Lemma 14.** *For every integer $0 \le \ell < k$ we have*

$$\mathbb{E}[Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) \mid \mathcal{G}] \le O(1) \cdot \binom{k}{\ell}\binom{n-k}{k-\ell} \prod_{i=1}^{m} 1 - 2(1-k/n)^{\Gamma_i} + 2\left(1 - 2k/n + \ell/n\right)^{\Gamma_i} \tag{3.1}$$

Proof. We use the linearity of expectation. The product of the two binomial coefficients simply accounts for the number of assignments σ that have overlap ℓ with $\boldsymbol{\sigma}$. Hence, with $\mathcal{S}$ the event that one specific $\sigma \in \{0,1\}^V$ that has overlap ℓ with $\boldsymbol{\sigma}$ belongs to $S_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$, we need to show that

$$\mathbb{P}\left[\mathcal{S} \mid \mathcal{G}\right] \le \prod_{i=1}^{m} 1 - 2(1-k/n)^{\Gamma_i} + 2\left(1 - 2k/n + \ell/n\right)^{\Gamma_i}. \tag{3.2}$$

By symmetry we may assume that $\boldsymbol{\sigma}_{x_i} = \mathbf{1}\{i \le k\}$ and that $\sigma_{x_i} = \mathbf{1}\{i \le \ell\} + \mathbf{1}\{k < i \le 2k - \ell\}$.

To establish (3.2) we harness the pairing model. Namely, given $\mathcal{G}$ we can think of each test a_i as a bin of capacity Γ_i. Moreover, we think of each clone (x_i, h), $h \in [\Delta]$, of an individual as a ball. The ball is labelled $(\boldsymbol{\sigma}_{x_i}, \sigma_{x_i}) \in \{0,1\}^2$. The random matching that creates $\boldsymbol{G}$ effectively tosses the Δn balls randomly into the bins. Hence, for $i \in [m]$ and for $j \in [\Gamma_i]$ let us write $\boldsymbol{A}_{i,j} = (\boldsymbol{A}_{i,j,1}, \boldsymbol{A}_{i,j,2}) \in \{0,1\}^2$ for the label of the jth ball that ends up in bin number i. Then we are left to calculate

$$\mathbb{P}\left[\mathcal{S} \mid \mathcal{G}\right] = \mathbb{P}\left[\forall i \in [m] : \max_{j\in[\Gamma_i]} \boldsymbol{A}_{i,j,1} = \max_{j\in[\Gamma_i]} \boldsymbol{A}_{i,j,2} \,\middle|\, \mathcal{G}\right], \tag{3.3}$$

i.e., the probability that a test a_i is positive with respect to first assignment $(\boldsymbol{A}_{i,j,1})_{j\in[\Gamma_i]}$ iff it is positive with respect to the second assignment $(\boldsymbol{A}_{i,j,2})_{j\in[\Gamma_i]}$.

To calculate this probability we borrow a trick from the analysis of the random k-SAT model [16]. Namely, we consider a new set $\{0,1\}^2$-valued random variables $\boldsymbol{A}'_{i,j} = (\boldsymbol{A}'_{i,j,1}, \boldsymbol{A}''_{i,j,2})$ such that $(\boldsymbol{A}'_{i,j})_{i\in[m],j\in[\Gamma_i]}$ are mutually independent and such that

$$\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,1)\right] = \ell/n, \qquad \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,1)\right] = \mathbb{P}\left[\boldsymbol{A}'_{i,j} = (1,0)\right] = (k-\ell)/n,$$
$$\mathbb{P}\left[\boldsymbol{A}'_{i,j} = (0,0)\right] = (n - 2k + \ell)/n$$

for all i, j. Now, let $\mathcal{R}$ be the event that

$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,1)\right\}=\ell\Delta, \qquad \sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,0)\right\}=(n-2k+\ell)\Delta,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(1,0)\right\}=\sum_{i=1}^{m}\sum_{j=1}^{\Gamma_i}\mathbf{1}\left\{\boldsymbol{A}'_{i,j}=(0,1)\right\}=(k-\ell)\Delta,$$

i..e, that all of the sums on the l.h.s. are *precisely* equal to their expected values. Then $\boldsymbol{A}'=(\boldsymbol{A}'_{i,j})_{i,j}$ given $\mathcal{R}$ is distributed precisely as $\boldsymbol{A}=(\boldsymbol{A}_{i,j})_{i,j}$. Hence, (3.3) yields

$$\mathbb{P}\left[\mathcal{S}\mid\mathcal{G}\right]=\mathbb{P}\left[\forall i\in[m]:\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,1}=\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,2}\mid\mathcal{G},\mathcal{R}\right]. \tag{3.4}$$

Thus, let $\mathcal{A}=\left\{\forall i\in[m]:\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,1}=\max_{j\in[\Gamma_i]}\boldsymbol{A}'_{i,j,2}\right\}$. Because the $(\boldsymbol{A}'_{i,j})_{i,j}$ are mutually independent, we can easily compute the unconditional probability $\mathcal{A}$: by inclusion/exclusion,

$$\mathbb{P}\left[\mathcal{A}\mid\mathcal{G}\right]=\prod_{i=1}^{m}1-2(1-k/n)^{\Gamma_i}+2(1-2k/n+\ell/n)^{\Gamma_i} \tag{3.5}$$

(the probability that $\max\boldsymbol{A}'_{i,j,1}=\max\boldsymbol{A}'_{i,j,2}=1$, i.e., both tests positive, equals one minus the probability that $\max\boldsymbol{A}'_{i,j,1}=0$ minus the probability that $\max\boldsymbol{A}'_{i,j,2}=0$ plus the probability that $\max\boldsymbol{A}'_{i,j,1}=\max\boldsymbol{A}'_{i,j,2}=0$; then add the probability that $\max\boldsymbol{A}'_{i,j,1}=\max\boldsymbol{A}'_{i,j,2}=0$, i.e., both tests negative).

Finally, to deal with the conditioning we use Bayes' rule:

$$\mathbb{P}\left[\mathcal{A}\mid\mathcal{R},\mathcal{G}\right]=\frac{\mathbb{P}\left[\mathcal{A}\mid\mathcal{G}\right]\mathbb{P}\left[\mathcal{R}\mid\mathcal{A},\mathcal{G}\right]}{\mathbb{P}\left[\mathcal{R}\mid\mathcal{G}\right]}. \tag{3.6}$$

Since the $(\boldsymbol{A}'_{i,j})_{i,j}$ are independent, the Local Limit Theorem for sums of independent variables [29] yields $\mathbb{P}\left[\mathcal{R}\mid\mathcal{G}\right]=\Theta(\Delta n)^{-3/2}$, $\mathbb{P}\left[\mathcal{R}\mid\mathcal{A},\mathcal{G}\right]=\Theta(\Delta n)^{-3/2}$. Hence, (3.2) follows from (3.4)–(3.6). ◀

Proof of Proposition 8. The Chernoff bound implies that $\Gamma_i\geq\Gamma_{\min}=\Delta n/m-\sqrt{\Delta n/m}\log n$ for all $i\in[m]$ w.h.p. Further, assuming that the Γ_i satisfy this bound, we perform an elementary calculation to check that

$$\sum_{0\leq\ell\leq(1-1/\log n)k}\binom{k}{\ell}\binom{n-k}{k-\ell}\prod_{i=1}^{m}1-2(1-k/n)^{\Gamma_i}+2\left(1-2k/n+\ell/n\right)^{\Gamma_i}=o(1). \tag{3.7}$$

Therefore, the proposition follows from Lemma 14 and Markov's inequality. ◀

4 Proof of Proposition 9

The argument from Section 3 does not extend large overlaps (close to k) because the expression on the r.h.s. of (3.1) gets too large. In other words, merely just computing the expected number of solutions with a given overlap does not do the trick. This "lottery phenomenon" is ubiquitous in random constraint satisfaction problems: for big overlap values rare solution-rich instances drive up the expected number of solutions [4, 5]. In order to cope with this issue we take another leaf out of the random CSP literature [2, 33]. Namely, we show that the solution $\boldsymbol{\sigma}$ is locally rigid. That is, the expansion properties of the random bipartite graph $\boldsymbol{G}$ preclude the existence of other solutions that have a big overlap with $\boldsymbol{\sigma}$. The following lemma holds the key to this effect.

▶ **Lemma 15.** *For any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for all $m > (1+\varepsilon)m_{\inf}$ the following is true. Let $\mathcal{R}$ be the event that for every x_i with $\boldsymbol{\sigma}_{x_i} = 1$ there are at least $\delta\Delta$ tests $a \in \partial x_i$ such that $\partial a \setminus \{x_i\} \subseteq V_0$. Then $\mathbb{P}[\mathcal{R}] = 1 - o(1)$.*

Hence, w.h.p. any infected individual appears in plenty of tests where all the other individuals are healthy. This property causes $\boldsymbol{\sigma}$ to be locally rigid. To see why, consider the repercussions of just changing the status of a single individual x_i from infected to healthy. Because given $\mathcal{R}$ the individual x_i appears as the only infected individual in at least $\delta\Delta$ tests, in order to maintain the same tests results we will also need to flip at least one individual in each of these tests from healthy to infected. Since tests typically have relatively few individuals in common, the necessary number of flips from 0 to 1 will be $\Omega(\Delta) = \Omega(\log n)$. But then in order to keep the total number of infected individuals constant k, we will need to perform another $\Omega(\Delta)$ flips from 1 to 0. Yet given $\mathcal{R}$ each of these "second generation" individuals that we flip from infected to healthy is itself the only infected individual in many tests. Thus, the single flip that we started from triggers a veritable avalanche of flips, which will stop only after the overlap has dropped significantly. The next lemma formalises this intuition. The lemma shows that while the unconditional expectation of $Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ is "too big", the conditional expectation of $Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ given $\mathcal{R}$ is much smaller. Let $\boldsymbol{m}_0 = \boldsymbol{m}_0(\boldsymbol{G}, \hat{\boldsymbol{\sigma}})$ be the total number of negative tests.

▶ **Lemma 16.** *Suppose that $(1 - 1/\log n)k \le \ell < k$ and let $\Gamma_{\min} = \min_{i\in[m]} \Gamma_i$, $\Gamma_{\max} = \max_{i\in[m]} \Gamma_i$. Then*

$$\mathbb{E}[Z_{k,\ell}(\boldsymbol{G}, \hat{\boldsymbol{\sigma}}) \mid \mathcal{G}, \mathcal{R}, \boldsymbol{m}_0] \le$$
$$O(1)\binom{k}{\ell}\binom{n-k}{k-\ell}\left(1 - \left(1 - \frac{k-\ell}{n}\right)^{\Gamma_{\max}}\right)^{\delta\Delta(k-\ell)}\left(\frac{n-2k+\ell}{n-k}\right)^{\Gamma_{\min}\boldsymbol{m}_0} = o(1) \qquad (4.1)$$

The proof of Lemma 16 requires some mildly delicate manoeuvres to cope with the stochastic dependences that are inherent in the random bipartite graph model.

Proof of Proposition 9. Standard tail bound arguments show that $\Gamma_{\max}, \Gamma_{\min} = \min_{i\in[m]} \Gamma_i = \Delta n/m + O(\sqrt{\Delta n/m \log n})$ w.h.p. Plugging these estimates into (4.1) and summing on $\ell > (1 - 1/\log n)k$ completes the proof. ◀

References

1 E. Abbe. Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research*, 18:1–86, 2018.

2 D. Achlioptas and A. Coja-Oghlan. Algorithmic barriers from phase transitions. *Proc. 49th FOCS*, pages 793–802, 2008.

3 D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi. On the solution space geometry of random formulas. *Random Structures and Algorithms*, 38:251–268, 2011.

4 D. Achlioptas and C. Moore. Random k-SAT: two moments suffice to cross a sharp threshold. *SIAM Journal on Computing*, 36:740–762, 2006.

5 D. Achlioptas, A. Naor, and Y. Peres. Rigorous location of phase transitions in hard optimization problems. *Nature*, 435:759–764, 2005.

6 D. Achlioptas and Y. Peres. The threshold for random k-SAT is $2^k \log 2 - O(k)$. *Journal of the AMS*, 17:947–973, 2004.

7 M. Aldridge. The capacity of Bernoulli nonadaptive group testing. *IEEE Transactions on Information Theory*, PP, 2015.

8 M. Aldridge. On the optimality of some group testing algorithms. *Proceedings of IEEE International Symposium on Information Theory*, pages 3085–3089, 2017.

9 M. Aldridge. Individual testing is optimal for nonadaptive group testing in the linear regime. *IEEE Transactions on Information Theory*, 65:2058–2061, 2018.

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

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

12 A. Alleman. An efficient algorithm for combinatorial group testing. *H. Aydinian, F. Cicalese, C. Deppe (eds) Information Theory, Combinatorics, and Search Theory. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg*, 7777:569–596, 2013.

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

14 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.

15 A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborová. Information-theoretic thresholds from the cavity method. *Advances in Mathematics*, 333:694–795, 2018.

16 A. Coja-Oghlan and K. Panagiotou. The asymptotic k-SAT threshold. *Advances in Mathematics*, 288:985–1068, 2016.

17 A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, 2011.

18 J. Ding, A. Sly, and N. Sun. Proof of the satisfiability conjecture for large k. *Proc. 47th STOC*, pages 59–68, 2015.

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

20 D. Du and F. Hwang. *Combinatorial group testing and its applications*. World Scientific, 2000.

21 O. Dubois and J. Mandler. The 3-XORSAT Threshold. *Proc. 43rd FOCS*, pages 769–778, 2002.

22 A. Emad and O. Milenkovic. Poisson group testing: a probabilistic model for nonadaptive streaming Boolean compressed sensing. *Proc. ICASSP*, pages 3335–3339, 2014.

23 V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted clique. *Proc. 45th STOC*, pages 655–664, 2013.

24 V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. *Proc. 48th STOC*, pages 77–86, 2015.

25 D. Gamarnik and M. Sudan. Performance of sequential local algorithms for the random NAE-K-SAT problem. *SIAM J. on Computing*, 46:590–619, 2017.

26 S. Hopkins, P. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer. The power of sum-of-squares for detecting hidden structures. *Proc. 58th FOCS*, pages 720–731, 2017.

27 S. Janson, T. Luczak, and A. Ruci'nski. *Random Graphs*. Wiley, 2000.

28 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, 2019.

29 A. Kolmogorov, A. Nikolaevich, and B. Gnedenko. Limit distributions for sums of independent random variables. *Addison-Wesley*, 1968.

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

31 M. Mézard and A. Montanari. *Information, physics and computation*. Oxford University Press, 2009.

32 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.

33 M. Molloy. The freezing threshold for k-colourings of a random graph. *Proc. 43rd STOC*, pages 921–930, 2012.

34 C. Moore. The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. *Bulletin of the EATCS*, 121, 2017.

35 E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, pages 1–31, 2014.

36 R. Mourad, Z. Dawy, and F. Morcos. Designing pooling systems for noisy high-throughput protein-protein interaction experiments using Boolean compressed sensing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:1478–1490, 2013.

37 H. Ngo and D. Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. *Discrete Mathematical Problems with Medical Applications*, 7:171–182, 2000.

38 J. Scarlett and V. Cevher. Phase transitions in group testing. *Proc. 27th SODA*, pages 40–53, 2016.

39 J. Scarlett and V. Cevher. Limits on support recovery with probabilistic models: an information-theoretic framework. *IEEE Transactions on Information Theory*, 63:593–620, 2017.

40 N. Thierry-Mieg. A new pooling strategy for high-throughput screening: the shifted transversal design. *BMC Bioinformatics*, 7:28, 2006.

41 L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65:453–552, 2016.