Efficient and Accurate Group Testing via Belief Propagation: An Empirical Study

Amin Coja-Oghlan
Faculty of Computer Science, TU Dortmund, Germany

Max Hahn-Klimroth
Faculty of Computer Science, TU Dortmund, Germany

Philipp Loick
Institute for Mathematics, Goethe Universität, Frankfurt am Main, Germany

Manuel Penschuck
Faculty of Computer Science, Goethe Universität, Frankfurt am Main, Germany

Abstract

The group testing problem asks for efficient pooling schemes and inference algorithms that allow to screen moderately large numbers of samples for rare infections. The goal is to accurately identify the infected individuals while minimizing the number of tests.

We propose the novel adaptive pooling scheme *adaptive Belief Propagation* (ABP) that acknowledges practical limitations such as limited pooling sizes and noisy tests that may give imperfect answers. We demonstrate that the accuracy of ABP surpasses that of individual testing despite using few overall tests. The new design comes with Belief Propagation as an efficient inference algorithm. While the development of ABP is guided by mathematical analyses and asymptotic insights, we conduct an experimental study to obtain results on practical population sizes.

2012 ACM Subject Classification Mathematics of computing → Probabilistic inference problems; Mathematics of computing → Random graphs; Mathematics of computing → Coding theory

Keywords and phrases Group testing, Probabilistic Construction, Belief Propagation, Simulation

Digital Object Identifier 10.4230/LIPIcs.SEA.2022.8

Related Version *Previous Version*: `https://arxiv.org/abs/2105.07882`

Supplementary Material *Software (Source Code)*: `https://github.com/manpen/group-testing`

Funding This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under DFG CO 646/3, DFG CO 646/5, and DFG ME 2088/5-1 (FOR 2975 – Algorithms, Dynamics, and Information Flow in Networks).

1 Introduction

Every day medical laboratories around the globe screen moderately large numbers of samples for rare pathogens. The vast majority of samples, anywhere between 90% and 99.9%, are actually uninfected [9, 25, 28, 40, 32, 37, 38, 39, 42]. Labs therefore test pools of samples rather than individual samples. The *group testing problem* asks for pooling strategies that minimise the total number of tests required while maximising the accuracy of the results. The latter is crucial because test results are generally not perfectly accurate.

Practical solutions are challenging precisely because the number of samples in a real-world scenario is in the hundreds or thousands. While the group testing problem has inspired a body of mathematical work for the asymptotical scenario [5, 13, 12], these results, where the number of samples grows to infinity, do not directly apply to practical problem sizes. They

also tend to construct excessively large test pools or distribute samples in very many tests [5, 13, 12]. Yet practical problem sizes are too large to exhaustively search for an optimal test. Thus, pooling schemes from the 1940s remain in practical use [9, 25, 28, 40, 32, 37, 38, 39, 42].

The aim of this paper is to investigate better test designs for practical problem sizes. We focus on improving the *accuracy* of the results, i.e., avoiding false positives and/or negatives while keeping the number of tests as small as possible. Indeed, group testing, originally invented to reduce the number of tests, actually excels at improving the accuracy of the results. This may seem surprising at first glance because one might deem individual testing optimal in terms of accuracy. It is not. Group testing does better in much the same way as error-correcting codes gain power from encoding entire blocks of data simultaneously.

1.1 Our contributions and outline

Given the moderate number of samples in real-world scenarios, we obtain practically meaningful results by conducting an extensive experimental study based on theoretical work on group testing as well as recent ideas from information theory and statistical physics. Our novel test design ABP improves the accuracy of the overall results while keeping the number of tests conducted low. Furthermore, the new test design requires only relatively small test pools and only assigns each sample to a small number of tests. Finally, the design comes with an efficient, easy-to-implement algorithm to infer the status of the individual samples from the test results, namely the Belief Propagation (BP) message passing algorithm.

We proceed to discuss the mathematical model we work with in Sec. 1.2. In Sec. 2, we discuss designs and algorithms that are in practical use or have been studied in the mathematical literature on group testing. In Sec. 3, we present the details behind our novel test design ABP and relate ABP to the theoretical work on group testing and asymptotic considerations in Sec. 5. Finally, in Sec. 6 we discuss the potential impact of the new results and future directions for both empirical and theoretical work.

1.2 The model

We work with a simple but standard model of group testing that allows for inaccurate test results [5]. Let $x_1, \dots, x_n$ be the samples to be tested and let $\lambda \in [0, 1]$ be the prior probability that any one sample is infected. The true infection status of each sample is indicated by $\boldsymbol{\sigma}(x_j) \in \{0, 1\}$, with 1 representing "infected". The $\boldsymbol{\sigma}(x_j)$ are assumed to be independent Bernoulli variables with mean λ. We refer to the vector $\boldsymbol{\sigma} = (\boldsymbol{\sigma}(x_j))_{j=1,\dots,n}$ as the *ground truth*. Let $k = \sum_{j=1}^{n} \mathbf{1}\{\boldsymbol{\sigma}(x_j) = 1\}$ signify the actual number of infected samples.

A *test design* is a bipartite graph G with one class $\mathcal{X} = \{x_1, \dots, x_n\}$ representing the n samples and the other class $\mathcal{A} = \{a_1, \dots, a_m\}$ representing the test pools. An edge between $\{x_j, a_i\}$ indicates that x_j is included in test pool a_i. For each x_j we let $\partial x_j = \partial_G x_j$ be the set of test pools that include x_j; analogously, ∂a_i denotes the samples x_j in pool a_i.

Let $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\sigma}}(a_i))_{i=1,\dots,m}$ denote the test results. Ideally, test a_i should report positive iff at least one sample $x_j \in \partial a_i$ is infected. But the actual result $\hat{\boldsymbol{\sigma}}(a_i)$ may include independent noise controlled via the *specificity* p and *sensitivity* q as follows:

$$\hat{\boldsymbol{\sigma}}(a_i) = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1-p \end{cases} \qquad \text{if } \boldsymbol{\sigma}(x_j) = 0 \text{ for all } x_j \in \partial a_i \tag{1}$$

$$\hat{\boldsymbol{\sigma}}(a_i) = \begin{cases} 0 & \text{with probability } 1-q \\ 1 & \text{with probability } q \end{cases} \qquad \text{if } \boldsymbol{\sigma}(x_j) = 1 \text{ for some } x_j \in \partial a_i \tag{2}$$

Unless $p = q = 1$ and every x_j is tested separately, the ground truth $\boldsymbol{\sigma}$ cannot be inferred perfectly form the test results $\hat{\boldsymbol{\sigma}}$ of a single "one-shot" test design [2]. Indeed, under the noise model in Eqs. (1) and (2) the posterior of the ground truth given the test results reads

$$\mu_G(\sigma) = \mathbb{P}\left[\boldsymbol{\sigma} = \sigma \mid \hat{\boldsymbol{\sigma}}\right] \propto \prod_{i=1}^{n} \lambda^{\sigma(x_i)}(1-\lambda)^{1-\sigma(x_i)} \prod_{i=1}^{m} \psi_{\hat{\boldsymbol{\sigma}}(a_i)}\left((\sigma(y))_{y\in\partial a_i}\right) \tag{3}$$

$$\text{where} \quad \psi_0(\sigma_1,\dots,\sigma_\ell) = p^{1-\bigvee_{i=1}^{\ell}\sigma_i}(1-q)^{\bigvee_{i=1}^{\ell}\sigma_i}, \quad \text{and}$$
$$\psi_1(\sigma_1,\dots,\sigma_\ell) = (1-p)^{1-\bigvee_{i=1}^{\ell}\sigma_i} q^{\bigvee_{i=1}^{\ell}\sigma_i},$$

and where the $\propto$-notation hides the normalisaton required to turn μ_G into a probability distribution. Hence, the information-theoretically optimal inference algorithm just draws a random sample from the distribution μ_G. In effect, the design's accuracy is governed by the entropy of the posterior μ_G: the smaller the entropy the better the results. Furthermore, depending on the specific design G there may or may not exist an *efficient* algorithm for sampling from μ_G.

In contrast, *adaptive group testing* uses multiple stages; an *ℓ-stage test design* is a sequence $G^{(0)}, G^{(1)}, \dots, G^{(\ell)}$ of test designs such that $G^{(i+1)}$ is obtained from $G^{(i)}$ by adding tests and edges based on the results from previous stages. The results of the new tests are assumed to be distributed independently according to Eqs. (1) and (2). The aim, of course, is to diligently add tests so as to maximally reduce the entropy of the posterior.

In summary, the group testing problem poses the following, partially conflicting, challenges:

- **(i)** We require an adaptive test design with high accuracy and a small number of tests.
- **(ii)** We require an *efficient* algorithm that infers the $\boldsymbol{\sigma}(x_j)$ from the observed $\hat{\boldsymbol{\sigma}}(a_i)$.
- **(iii)** Practical limitations require a small number of samples in a test and tests per sample.
- **(iv)** We aim for a small number of test stages to ensure a timely reporting of test outcomes – or at least ensure that most samples can be diagnosed after the first or second stage.

2 Established designs and algorithms

2.1 Individual testing

The most straightforward test strategy, of course, is to conduct $m = n$ individual tests for each of the n samples. Naturally, in the case $p = q = 1$, individual testing will register the status of each sample correctly. However, realistic values for p and q range between 0.95 and 0.99 [9, 10, 31, 40, 44]. Then individual testing will produce numbers of false positives/negatives distributed as $\mathrm{Bin}(n-k, 1-p)$ and $\mathrm{Bin}(k, 1-q)$, respectively.

The accuracy of the results could obviously be boosted by conducting two or three individual tests per sample. Indeed, if we test each x_j twice and report x_j as infected only if both tests come back positive, then we could reduce the expected number of false positives to $(n-k)(1-p)^2$. But we would now expect a slightly larger number of $2k(1-q)$ false negatives. To reduce the number of false positives and negatives simultaneously we could test each x_j thrice and report the majority of the three test results.

2.2 Dorfman

The test designs that appear to be currently most widely adapted in practice date back to the 1940s. Indeed, the idea of group testing was first brought up by Dorfman in 1943 [20]. He suggested a two-stage test procedure, we denote as Dorfman. In the first stage, every sample gets placed in precisely one pool. All pools are the same size, which depends on the prior λ only. Pools with a positive test result get tested separately in the second stage.

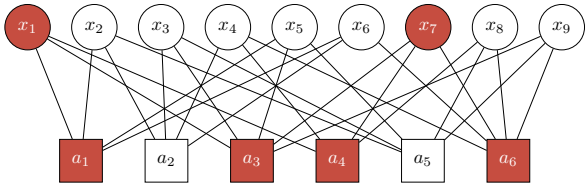

Figure 1 Illustration of a random biregular test design with $\Delta = 3$ and $\Gamma = 4$.

Depending on the prior, this scheme can significantly reduce the number of tests required. For example, with $\lambda = 0.05$ this scheme uses pools of size five and the expected overall number of tests conducted in both stages comes to about $0.426n$. At the same time, DORFMAN's two-stage procedure reduces the number of false positives because a sample is ultimately reported as positive only if both the tests are positive. But for the same reason, the expected number of false negatives increases. For instance, with $n = 10^4$ and $k = \lambda n = 500$, we expect 18.2 false positives and 9.95 false negatives.

A natural extension of the DORFMAN procedure employs three stages. In the first stage, relatively large pools are formed. The second stage then splits the positive pools into smaller sub-pools and the third stage resorts to individual testing. In effect, as with the two-stage procedure, the expected number of false positives decreases while the expected number of false negatives increases. For $n = 10^4$ and $k = \lambda n = 500$ the expected numbers of false positives/negatives work out to be 11.76 and 14.8, respectively.

2.3 Probabilistic constructions

More sophisticated test designs have been proposed in the mathematical theory of group testing. The currently best, and in certain asymptotic settings provably optimal, test designs harness randomisation [5, 13]. For instance, in the *random biregular test design* illustrated in Fig. 1 every test pool has an equal size Γ and every individual sample joins an equal number Δ of pools. In other words, the test design $G = G_{n,m}(\Gamma, \Delta)$ is chosen uniformly at random from the set of all (Δ, Γ)-regular bipartite graphs (e.g., see [41]).[1] To maximize information gained, the parameters Γ and Δ need to be chosen as to maximise the conditional entropy of the vector $\hat{\boldsymbol{\sigma}}$ of test results, i.e., so that about half the tests will be positive:[2]

$$\Delta = m\log(2)/(n\lambda) \qquad \Gamma = \log(2)/\lambda \tag{4}$$

Intuitively, the randomness of the test design minimizes dependencies between the different test results $\hat{\boldsymbol{\sigma}}(a_i)$. Thus, with the parameters as in Eq. (4) and for a number m of tests up to a threshold, we can hope to squeeze up to one bit of information from each test. Similar randomised constructions are used in coding theory and compressed sensing [17, 18, 26, 35].

Unlike previous discussions, the random biregular design has no obvious inference algorithm. For $p = q = 1$, a posteriori inference implies a minimum hypergraph vertex cover [12], which is an NP-hard problem and even on random instances no efficient algorithm is known.

[1] G is typically drawn from the pairing model [8, 34]. Then, in rare cases the same individual joins a test pool twice. In practice, such double occurrence could, of course, be reduced to single occurrences.

[2] Due to rounding issues, we cannot ensure that the expected number of positive tests is precisely $m/2$.

Definite defectives (DD) [4] is a blunt but efficient algorithm. The algorithm classifies every sample that is only included in positive test as infected under the condition that it appears in at least one positive test pool where all other samples appear in a negative test. All other samples are classified as uninfected. In symbols,

$$\boldsymbol{\sigma}_{\mathtt{DD}}(x_j) = \bigwedge_{a\in\partial x_j} \hat{\boldsymbol{\sigma}}(a) \quad \wedge \quad \bigvee_{a\in\partial x_j} \bigwedge_{y\in\partial a} \bigvee_{b\in\partial y} (1-\hat{\boldsymbol{\sigma}}(b)).$$

For $p = q = 1$ this algorithm will never produce false positives but may render false negatives. Several similarly-flavoured algorithms have been analysed mathematically. Aldridge analysed an adaptive test design whose different stages employ random biregular test designs with suitably chosen degrees [3]. This adaptive test design carried out over an unbounded number of stages (which may take too long in practice) achieves rates in excess of 0.95 bits per tests.

2.4 Glauber dynamics

While DD merely extracts binary information about each sample, we want a more fine-grained picture of the posterior distribution Eq. (3) of the random test design. Glauber dynamics (Glauber) is a Markov Chain Monte Carlo algorithm and starts at a random initial configuration $\sigma^{(0)} = (\sigma^{(0)}(x_i))_{i=1,\dots,n}$ drawn from the prior. Thus, the individual $\sigma^{(0)}(x_i)$ are independent $\mathrm{Be}(\lambda)$ variables. Glauber then proceeds to generate a random sequence $(\sigma^{(t)})_{t=0,\dots,T}$ of configurations by updating the status of a random sample at each time step according to Eq. (3); see [27] for details of the update rule. The hope is that for moderate T the empirical means of the sequence approximate the actual posteriors well, i.e.,

$$\mu_G(\{\boldsymbol{\sigma}(x_j) = s\}) \approx \frac{1}{T}\sum_{i=0}^{T} \mathbf{1}\left\{\sigma^{(t)}(x_j) = s\right\} \qquad (j = 1,\dots,n;\ s\in\{0,1\}). \tag{5}$$

We are unaware of a rigorous analysis of Glauber. Further, an exact empirical assessment appears difficult as the marginals of the posterior in Eq. (3) cannot be computed by exhaustive enumeration even for moderate values of n. Still, [16] studies Glauber experimentally.

2.5 Informative Dorfman

Informative Dorfman [29] is a multi-stage test design that uses the posterior marginals of a first stage (e.g., as approximated by Glauber) to determine the group sizes of a subsequent Dorfman test design. More precisely, it sorts the samples increasingly by their marginals and groups them in this order. The pools containing samples with small marginals are relatively large, while samples with marginals above 0.3 get tested individually. In the empirical study [16] of a combination of Glauber and InfDorfman, Cuturi et al. find that this procedure works decently well for a given number of tests but is still outperformed by quite a margin by more complicated multi-stage test designs and algorithms.

3 Adaptive Belief Propagation (aBP)

In this section we discuss our novel design aBP and its inference algorithm. The first stage employs the random biregular test design (Sec. 2.3). Given the results of the first stage, in the second and third stage we use a blend of the random biregular design and InfDorfman. For the inference algorithm we seize upon the BP message passing paradigm [33].

3.1 Belief Propagation

In recent years the Belief Propagation (BP) message passing paradigm has been applied in combination with randomised constructions with stunning success. Prominent examples include coding theory and other signal processing tasks such as compressed sensing [18, 26, 35]. The development of BP with randomised constructions has been inspired by ideas from the statistical mechanics of disordered systems [30]. More recently, substantial mathematical research has been devoted to BP (e.g., [6, 14, 21, 43]). Although most of this theoretical work is asymptotical, we let these ideas guide our quest for a practical group testing design.

BP is a generic message passing technique to approximate the marginals of Boltzmann distributions on factor graphs (e.g., Eq. (3)). The basic intuition behind BP is that under certain assumptions the posterior distribution admits a succinct representation in terms of *messages* [14, 15, 30, 45]. These assumptions are provably met in many Bayes-optimal inference problems on random factor graphs including the group testing problem as modelled in Sec. 1.2; at least asymptotically as the problem size tends to infinity [7, 11].

At first glance the posterior distribution Eq. (3) appears to be quite a difficult object to study; e.g., to estimate its entropy, we might have to inspect all 2^n possible vectors $\sigma \in \{0,1\}^n$. But according to the BP paradigm we can get a handle on the posterior distribution in terms of messages associated with the edges of the test design $G = G_{n,m}(\Gamma, \Delta)$. Formally, the *message space* of $\mathcal{M}(G)$ consists of vectors $(\mu_{x_j \to a_i}(s),\ \mu_{a_i \to x_j}(s))_{j=1,\ldots,n;\, i=1,\ldots,m;\, x_j \in \partial a_i;\, s \in \{0,1\}}$.

The idea is that there are two messages $\mu_{x_j \to a_i}(\,\cdot\,)$ and $\mu_{a_i \to x_j}(\,\cdot\,)$ associated with every edge of G, one directed from the sample x_j to the test a_i and one in the opposite direction. The messages themselves are probability distributions on $\{0,1\}$. Thus, we have $\mu_{x_j \to a_i}(0),\ \mu_{x_j \to a_i}(1) \in [0,1]$ and $\mu_{x_j \to a_i}(0) + \mu_{x_j \to a_i}(1) = 1$, and analogously for $\mu_{a_i \to x_j}(\,\cdot\,)$.

Roughly speaking, $\mu_{a_i \to x_j}(\,\cdot\,)$ represents the impact that a_i has on x_j in the absence of all other tests $b \in \partial x_j$. Moreover, $\mu_{x_j \to a_i}(\,\cdot\,)$ represents the status of x_j in the absence of test a_i. More formally, we define the *standard message* $\mu_{G, x_j \to a_i}(s)$ as the posterior probability that $\boldsymbol{\sigma}(x_j) = s$ given the test design $G - a_i$ obtained from G by omitting test a_i and given the test results $(\hat{\boldsymbol{\sigma}}(a_h))_{h \neq i}$. With the notation of Eq. (3), we can write this probability out as

$$\mu_{G, x_j \to a_i}(s) \propto \sum_{\sigma \in \{0,1\}^{\mathcal{X}},\ \sigma(x_j) = s} \prod_{i=1}^{n} \lambda^{\sigma(x_i)} (1-\lambda)^{1-\sigma(x_i)} \prod_{i=1}^{m} \psi_{\hat{\boldsymbol{\sigma}}(a_i)}\left((\sigma_y)_{y \in \partial a_i}\right)$$

with the $\propto$-sign hiding the normalisation to ensure that $\mu_{G, x_j \to a_i}(0) + \mu_{G, x_j \to a_i}(1) = 1$. Similarly, the standard message $\mu_{G, a_i \to x_j}(s)$ is defined as the posterior probability that $\boldsymbol{\sigma}(x_j) = s$ given the test design $G - (\partial x_j \setminus \{a_i\})$ obtained by removing all tests that x_j takes part in except for a_i and given the test results $\hat{\boldsymbol{\sigma}}(a_h)$ of all tests $a_h \notin \partial x_j \setminus \{a_i\}$.

Conceived wisdom, vindicated mathematically for a broad family of inference problems, predicts that asymptotically these messages satisfy the following BP equations [7, 11, 14, 45]:

$$\mu_{G, x \to a}(s) \propto \lambda^s (1-\lambda)^{1-s} \prod_{b \in \partial x \setminus \{a\}} \mu_{G, b \to x}(s), \tag{6}$$

$$\mu_{G, a \to x}(0) \propto 1 - q + (p + q - 1) \prod_{y \in \partial a \setminus \{x\}} \mu_{G, y \to a}(0), \quad \mu_{G, a \to x}(1) \propto 1 - q \qquad \text{if } \hat{\boldsymbol{\sigma}}(a) = 0, \tag{7}$$

$$\mu_{G, a \to x}(0) \propto q + (1 - p - q) \prod_{y \in \partial a \setminus \{x\}} \mu_{G, y \to a}(0), \qquad \mu_{G, a \to x}(1) \propto q \qquad \text{if } \hat{\boldsymbol{\sigma}}(a) = 1 \tag{8}$$

These equations express the notion that the random biregular design $G_{n,m}(\Gamma, \Delta)$ minimises dependencies between the test results. Furthermore, we expect that the marginals of the posterior distribution can be well approximated in terms of the messages:

$$\mu_G(\{\boldsymbol{\sigma}(x_i) = s\}) \propto \lambda^s (1-\lambda)^{1-s} \prod_{b \in \partial x_i} \mu_{G, b \to x_i}(s) \tag{9}$$

Apart from the marginals, asymptotic results also suggest that the entropy of the posterior distribution can be approximated in terms of the messages [11, 14, 30]. This approximation comes in terms of a functional called the *Bethe free energy*, defined as

$$\mathcal{B}_G = \sum_{x\in\mathcal{X}} \mathcal{B}_{G,x} + \sum_{a\in\mathcal{A}} \mathcal{B}_{G,a} - \sum_{x\in\mathcal{X},a\in\partial x} \mathcal{B}_{G,x,a} \qquad \text{with} \tag{10}$$

$$\mathcal{B}_{G,x} = \log \sum_{s\in\{0,1\}} \prod_{a\in\partial x} \mu_{G,a\to x}(s) \tag{11}$$

$$\mathcal{B}_{G,a} = \begin{cases} \log\left(1-q+(p+q-1)\prod_{x\in\partial a}\mu_{G,x\to a}(0)\right) & \text{if } \hat{\boldsymbol{\sigma}}(a)=0 \\ \log\left(q+(1-p-q)\prod_{x\in\partial a}\mu_{G,x\to a}(0)\right) & \text{if } \hat{\boldsymbol{\sigma}}(a)=1 \end{cases} \tag{12}$$

$$\mathcal{B}_{G,x,a} = \log \sum_{s\in\{0,1\}} \mu_{G,x\to a}(s)\mu_{G,a\to x}(s). \tag{13}$$

The resulting approximation of the entropy reads

$$\begin{aligned}\mathcal{H}_G = \mathcal{B}_G - n\log\lambda + \sum_{i=1}^{n}\mu_G(\{\boldsymbol{\sigma}_x=0\})\log\frac{\lambda}{1-\lambda} \\ - \sum_{\substack{i=1\\ \hat{\boldsymbol{\sigma}}(a_i)=0}}^{m}\left[\frac{p\log(p)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)}{1-q+(p+q-1)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)} + \frac{(1-q)\log(1-q)(1-\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0))}{1-q+(p+q-1)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)}\right] \\ - \sum_{\substack{i=1\\ \hat{\boldsymbol{\sigma}}(a_i)=1}}^{m}\left[\frac{(1-p)\log(1-p)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)}{q+(1-p-q)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)} + \frac{q\log(q)(1-\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0))}{q+(1-p-q)\prod_{x\in\partial a_i}\mu_{G,x\to a_i}(0)}\right].\end{aligned} \tag{14}$$

Hence, in order to estimate the marginals and the entropy of the posterior we need to calculate the BP messages. A natural idea is to perform a fixed point iteration using the BP Eqs. (6)–(8). These equation usually possess several solutions [11, 45]. Whether or not the fixed point iteration homes in on the correct solution then depends on the initialisation.

While there is no generic recipe for choosing an appropriate initialisation $\mu^{(0)} \in \mathcal{M}(G)$, two choices suggest themselves. First, we can initialise the messages based to the prior λ, i.e.,

$$\mu^{(0)}_{x_j\to a_i}(s) = \lambda^s(1-\lambda)^{1-s}. \tag{15}$$

We can also initialise the messages according to the ground truth, i.e. $\mu^{(0)}_{x_j\to a_i}(s) = \boldsymbol{\sigma}(x_j)$. While the latter is not practically useful for the obvious reason, the analogy with other inference problems suggests that *if* the fixed point iteration converges to the same solution from both initialisations, then this solution actually is a good approximation to the correct messages. Luckily, this can be tested using empirical simulations.

3.1.1 Preventing oscillations

The textbook method to perform the fixed point iteration is to update all messages in parallel. This means that, starting from the initialisation $(\mu^{(0)}_{x_j\to a_i})_{i,j}$, we compute all test-to-sample approximations $\mu^{(0)}_{a_i\to x_j}$ via Eqs. (6)–(8). Then we use these together with Eq. (6) to compute the next approximation $(\mu^{(1)}_{x_j\to a_i}(\,\cdot\,))_{i,j}$ to all sample-to-test messages, and so forth.

Cuturi et al. [16] demonstrated experimentally that such parallel updates do not converge. Instead, the messages oscillate between odd and even rounds. A similar observation was already made by Sejdinovic and Johnson [36]. Similar oscillations emerge in other applications

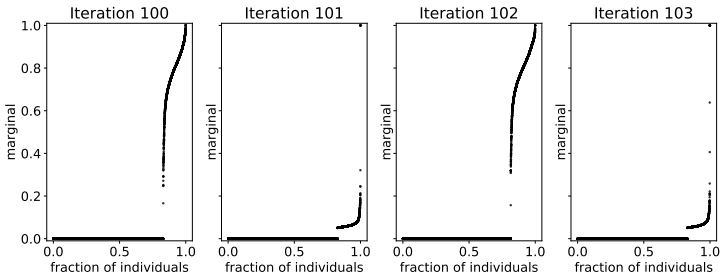

Figure 2 Oscillations in BP with parallel updates for $\lambda = 0.05$, 0.2 tests/n, and $p = q = 1$.

of BP and can also be observed in our simulations as illustrated in Fig. 2. They may result from an instability of the empirical mean of the messages. If in some particular iteration t the deviation from the prior

$$\sum_{j=1}^{n}\sum_{i=1}^{m} \mathbf{1}\{a_i \in \partial x_j\}\,(\mu^{(t)}_{x_j\to a_i}(1) - \lambda) \tag{16}$$

is positive, then we should expect a negative deviation in the next round. This is because due to Eq. (16) in the next iteration many tests will receive a relatively large indication that one of their samples may be infected. The test will therefore send out "less urgent" messages to the other samples. Conversely, if Eq. (16) is negative, then in iteration $t+1$ we expect to see a positive deviation. Due to the analytic nature of the update rules Eq. (6)–Eq. (8) these oscillations do not dampen down but actually amplify. This observation led the authors of [16] to turn to the computationally more intensive Glauber.

But actually oscillations of this type have been observed in other problems and several mitigations are known. We resort to a natural solution, namely to update the messages in a randomised order rather than in parallel to break the cycle of oscillations. Starting from the initialisation $(\mu^{(0)}_{x_j\to a_i}(\,\cdot\,))_{i,j}$, we apply Eq. (7)–Eq. (8) once to initialise the test-to-sample messages $\mu_{a_i\to x_j}(\,\cdot\,)$. Then at each time $t \geq 1$ we choose an edge $a_i x_j$ of G randomly and then randomly either process $\mu^{(t)}_{x_j\to a_i}(\,\cdot\,)$ or $\mu^{(t)}_{a_i\to x_j}(\,\cdot\,)$.

We stop the fixed point iteration after a fixed number T of steps. The precise choice of T is guided by experiments but T should be large enough so that every message will likely get updated several times. We note that this update scheme does not impede practical matters from using our algorithm in a laboratory setting since it purely pertains to the computations behind the scene and does not impact how samples are split and combined.

Beyond relying on asymptotic ideas and comparing the messages that result from the two aforementioned initialisations we take two additional steps to corroborate the results of BP. First, we compared the marginals obtained by BP with the empirical marginals of Glauber on a number of samples. They match. Second, we compared the marginals obtained via BP on moderately sized biregular test designs with the marginal distributions obtained via *population dynamics*, a heuristic intended to approximate the limiting distribution of the marginals as $n \to \infty$ [30]. They, too, align very well. Figure 3 displays the typical outcome of the BP along with the estimate Eq. (14) of the remaining entropy.

3.2 The first stage

As the first stage we use the random biregular design $G = G_{n,m}(\Delta, \Gamma)$ with the optimal parameters from Eq. (4). Thus, the only free parameter is the total number m of tests conducted in the first stage. Its choice is informed by BP. Specifically, we choose the

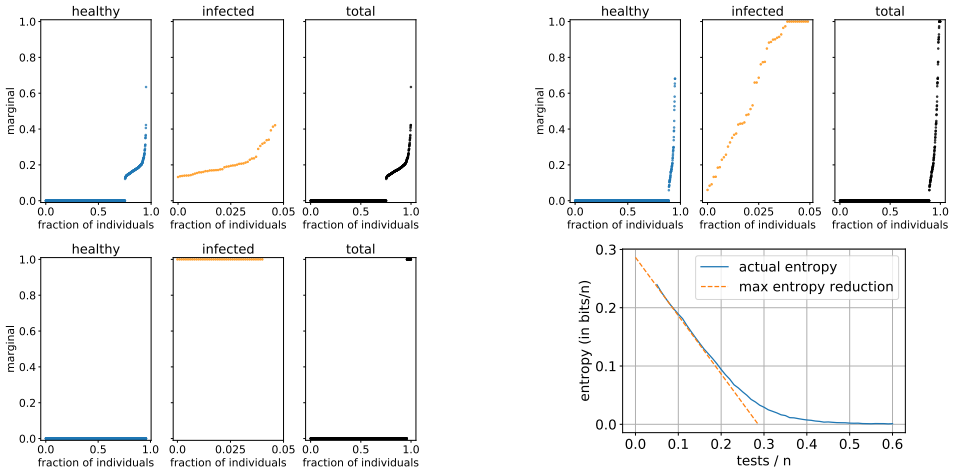

Figure 3 Posteriors of BP on a random biregular design with 0.15 (top left), 0.25 (top right) and 0.6 (bottom left) tests/n and remaining entropy (bottom right) for $\lambda = 0.05$ and $p = q = 1$.

largest number m of tests up to which each test yields the optimal entropy reduction of $\ln 2$. Practically, this means that we choose m to match the point at which the entropy plot for the corresponding parameter values flattens. The fourth graphic in Fig. 3 shows the approximation of the entropy as a function of the number of tests for $n = 1000$ and $\lambda = 0.05$ in the noiseless setting. For other priors and noise levels, the story turns out to be analogous.

3.3 The second and third stage

Given the approximation of the marginals from the first stage, how should we proceed? As we saw in Sections 2.3 and 2.5, two ideas for the subsequent stages proposed in the literature include InfDorfman as well as individual testing of all samples whose marginals are not entirely polarised after the first round. The former suffers from the same problem as the original Dorfman scheme, namely a potentially fairly large number of false positives and negatives. The latter strategy, known as DD, seems wasteful as it completely disregards any non-trivial information about the marginals resulting from the BP computation.

To remedy these issues, we propose a new design that blends the random biregular design with the InfDorfman scheme from Sec. 2.5. First, we directly report samples with marginals obtained from the first stage marginals less than 0.1% as healthy and those with marginals beyond 99.9% as infected. As illustrated in Fig. 4, the remaining samples are split into two groups, one comprising samples with marginals below 12.4% (*low risk*) and one with marginals above (*high risk*). The choice of 12.4% marks precisely the threshold beyond which the expressions Eq. (4) suggest that any sample should be placed in one test only.

For the high risk group, we set up an InfDorfman design G''. If such a pooled test turns out to be negative, we classify all samples in this pool as healthy. Otherwise, we conduct individual tests and classify samples solely based on these individual test results.

For the low risk group, we set up another random biregular test design on which we run BP where the priors are given by the posteriors of the first stage. The resulting marginals are again thresholded at 0.1% and 99.9%. Those samples whose marginals fall in between are subsequently retested individually with their classification being solely determined by the outcome of the individual test. To be more precise, let $\mathcal{X}'$ be the samples in the low risk

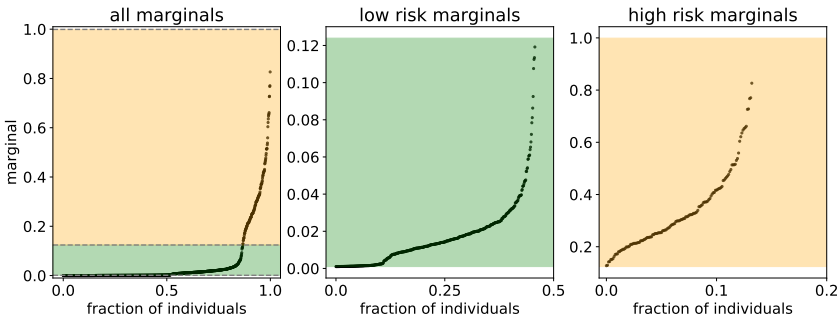

Figure 4 Low and high risk marginals for $\lambda = 0.05$ in the high noise setting with $m/n = 0.25$.

group, let $n' = |\mathcal{X}'|$ and let m' be the number of tests dedicated to this group. Based on the first round's BP results we approximate the average marginal $\lambda' = \frac{1}{n'} \sum_{x \in \mathcal{X}'} \mu_G(\{\boldsymbol{\sigma}(x) = 1\})$. Mimicking Eq. (4) we then choose the degrees

$$\Delta' = m' \log(2)/(n'\lambda') \qquad \Gamma' = \log(2)/\lambda' \tag{17}$$

and set up a random biregular test design $G' = G_{n',m'}(\Delta', \Gamma')$ on $\mathcal{X}'$. Furthermore, we modify the BP equations on this random biregular design to accommodate the marginals computed in the first stage. Hence, instead of using the universal prior λ' for all the samples, we substitute the separate marginals computed in the first stage:

$$\mu_{G',x\to a}(s) \propto \mu_G(\{\hat{\boldsymbol{\sigma}}(x) = 1\})^s (1 - \mu_G(\{\hat{\boldsymbol{\sigma}}(x) = 1\})^{1-s} \prod_{b \in \partial x \setminus \{a\}} \mu_{b\to x}(s) \tag{18}$$

3.4 Enhanced accuracy

In the following, we discuss a trade-off between accuracy and number of tests. The construction discussed so far is denoted as aBP-1, and the more accurate variants are aBP-2 and aBP-3. In aBP-1, almost all false results originate from the InfDorfman procedure in the second stage and neither the thresholding nor the second-stage random biregular design tend to produce a notable number of mistakes. Therefore, in aBP-2 and aBP-3 we perform the InfDorfman procedure twice or thrice independently in parallel.[3]

If we perform InfDorfman twice (aBP-2), we need to choose whether to reduce false negatives or false positives. Accordingly, we classify a sample as healthy (infected) if both InfDorfman procedures classify it as healthy (infected). In aBP-3, we classify according to the majority vote of the three InfDorfman schemes. We refer to Appendix A for a listing of the number of tests to be performed in the first and second stage.

4 Empirical investigation

In this section, we consider instance of $n = 1000$ samples and omit extensive simulations for $n = 100$ and $n = 10\,000$ since the results presented here, particularly the power of aBP carry over to those sizes. For smaller instance, rounding issues and few samples in the second stage necessitate slightly more tests; for larger n, we obtain a better performance.

[3] In case of aBP-3, we opted to keep the number of stages small. Instead, we may also run aBP-2 and only carry out a third round on samples where both runs yield different results.

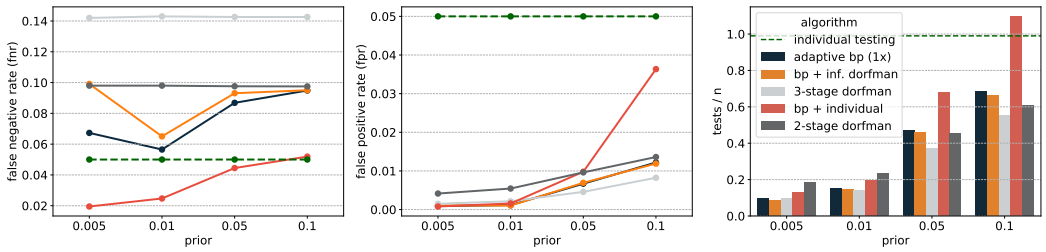

Figure 5 High noise scenario (sensitivity and specificity of $p = q = 95\%$).

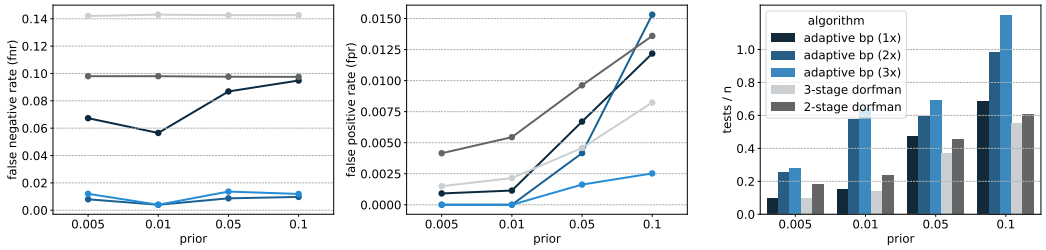

Figure 6 Reliability-enhanced ABP for high noise scenario (sensitivity/specificity of $p = q = 95\%$).

We study infection rates $\lambda \in \{0.5\%, 1\%, 5\%, 10\%\}$ and the following specificity/sensitivity scenarios (details on the parametrizations of the test designs for these settings are reported in Appendices A and B):

(a) perfectly reliable tests, i.e. $p = q = 1$,

(b) moderately values $p = 0.99$ and $q = 0.98$ (e.g., certain Covid-19 tests [9, 10, 31, 40, 44])

(c) a noisy scenario with $p = q = 0.95$.

Our experiments show that ABP improves the accuracy by an order of magnitude compared to known test designs while keeping the number of tests at a reasonable level. In the following, let the *false positive rate (fpr)* be the number of healthy samples falsely classified as infected over all healthy samples; define the *false negative rate (fnr)* analogously.

In the **High-noise scenario** with $p = q = 0.95$, ABP reaps the greatest gains. Figure 5 displays the results of ABP-1 in comparison to several previously known approaches. These include the widely used two- and three-stage DORFMAN designs (Sec. 2.2), the INFDORFMAN design (Sec. 2.5) as well as BP followed by individual testing advocated in the theoretical literature[4]. The figure shows that with about the same number of tests as 2-stage DORFMAN, ABP achieves up to 78% reduction in the number of false positives and an up to 42% reduction in the number of false negatives. The gains are particularly high for small priors.

Still, the absolute error rates in Fig. 5, particularly for large priors, may still be prohibitive for many real-world applications. Here our two designs ABP-2 and ABP-3 (Sec. 3.4) come to the rescue. As Fig. 6 shows, these designs, particularly ABP-3, dramatically reduce the number of false positives and negatives. Of course, these improvements come at the expense of a larger number of tests. But for priors $\lambda \leq 0.05$ the number of extra tests is moderate, and for the largest prior $\lambda = 0.1$ ABP-2 and ABP-3 require not many more tests than individual testing while being the only designs that deliver decent accuracy.

[4] With perfectly reliable tests, this approach is equivalent to DD algorithm followed by individual testing.

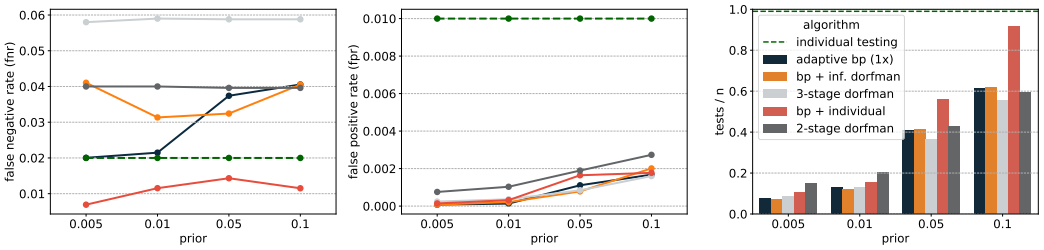

Figure 7 Sensitivity for moderate noise scenario ($p = 98\%$, $q = 99\%$).

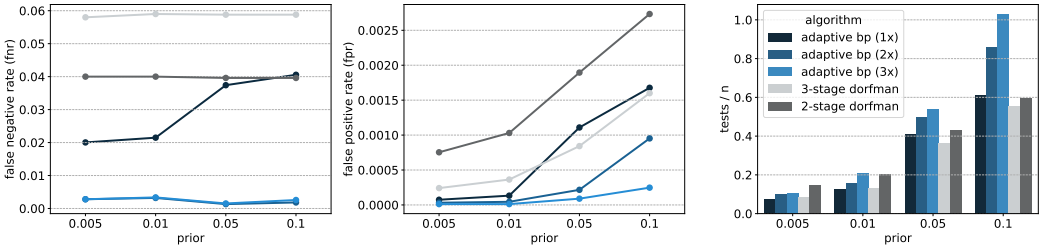

Figure 8 Reliability-enhanced ABP for moderate noise ($p = 98\%$, $q = 99\%$).

Figure 7 indicates a similar behaviour for **moderately high noise** with $p = 0.99$, $q = 0.98$. In comparison to the classical two- and three-stage DORFMAN, ABP requires at most 11% more tests for high priors of $\lambda = 0.1$ and even fewer for small priors. The benefit is that ABP boosts accuracy compared to all the previously known designs, particularly so for low priors. We point out that the gains vis-a-vis INFDORFMAN for moderately high priors are modest. The key benefit in ABP lies in its versatility to meaningfully enhance the accuracy at the expense of somewhat more tests as shown in Fig. 8. A similar extension of INFDORFMAN would yield a similar accuracy but require significantly more tests than ABP.

Even with **perfectly reliable** tests, the conventional DD approach (Sec. 2.3) is improved upon by ABP or the INFDORFMAN approach. Both schemes are able to reduce the number of tests compared to the former by up to 18% and comes within 19% to 32% of the information-theoretic lower bound. The gains vis-a-vis two-stage DORFMAN with up to 57% and individual testing with up to 94% are even more pronounced. We do not need to consider the accuracy in the noiseless case since all test designs recover the entire ground truth by construction.

In Fig. 10, we consider the fraction of samples that are identified by in each stage. It highlights that despite a total of three stages needed for ABP the majority of samples are identified already in the first and second stage, depending on the prior and noise level.

All examined algorithms require reasonable pool sizes and splits of the individual sample that are in line with common pooling procedures [22, 24, 29]. The maximum pool size is between 8 and 170 depending on noise level and prior, while the splits of the individual sample range between 3 and 19. It should be noted that the proposed algorithms and test designs can readily be adjusted to accommodate smaller pool sizes or individual sample splits – at the expense of somewhat more tests.

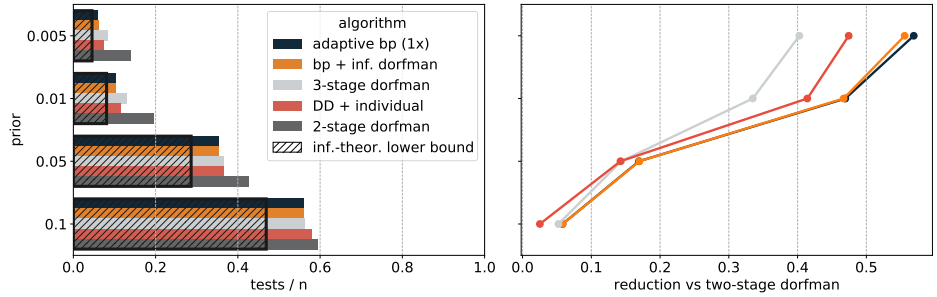

Figure 9 Simulation results for the noiseless setting. The black area represents a plausible information-theoretic lower bound for the number of tests. The left plot displays the numbers of tests required by the different designs; the right plot shows the reduction achieved by comparison to the 2-stage DORFMAN procedure, a classical and widely used test design.

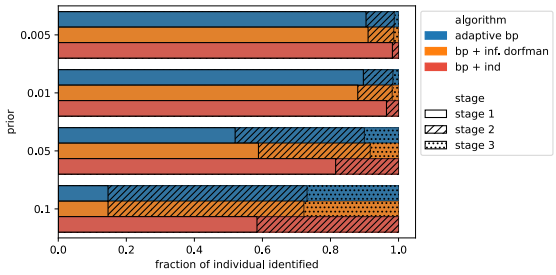

Figure 10 Fraction of samples identified in each stage by (i) ABP, (ii) BP followed by individual testing, and (iii) BP followed by INFDORFMAN.

5 Asymptotic considerations

Clearly, ABP relies on heuristics and is not asymptotically optimal. This begs the question of how we would adapt the design and algorithm if we decide to live unburdened by practical considerations and consider the case $n \to \infty$?

5.1 Variations on aBP

The optimal drop in entropy in Fig. 3 encourages running BP on a random biregular test design in the first. The discrete partition into three groups in the second stage, however, gives something away. Indeed, in the asymptotic regime infinitesimal intervals of posterior marginals contain an unbounded number of samples.[5] Thus, it seems information-theoretically optimal to construct a random biregular design for every single small marginal interval and repeat this procedure over a few stages. However, such an approach is impractical as, for moderate n, each random biregular design would only contain very few samples.

A simpler alternative that we considered is to still include all samples in one single second-stage test design, in which we choose the number of tests in which each sample takes part according to the posterior marginal from the first stage. Specifically, we chose these numbers so that in expectation half the tests should be positive. However, this design turned out to be unstable for small values of n because of random fluctuations.

[5] Of course, depending on the prior and the noise setting the distribution of the posterior marginals need not be supported on the entire unit interval.

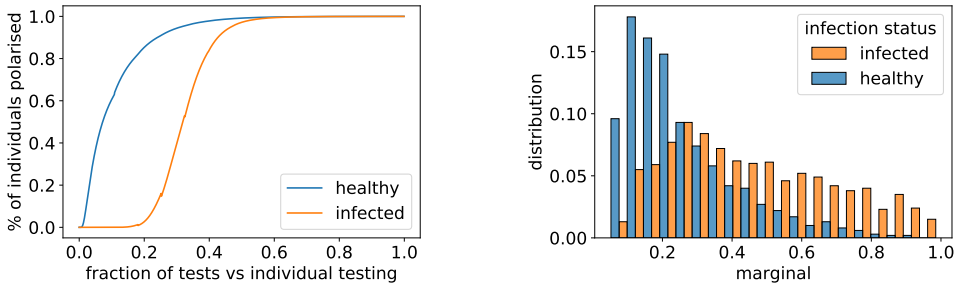

Figure 11 Asymptotic fraction of polarised marginals and the posteriors for non-polarised samples obtained with population dynamics on the distribution by [23] for $\lambda = 0.05$ and $p = q = 1$.

5.2 Plain Belief Propagation

Thus far we disregarded what might seem at first glance the most straightforward scheme: just run BP on a random biregular design and then simply threshold the marginals at, say, 50%. An obvious advantage of this approach is that it requires one stage only. Indeed, when we simulated this scheme for large group testing instances such as $n = 10\,000$, this approach turned out to work extremely well. Particularly for small priors such as 0.5% and 1% the plain BP plus thresholding design is on par or even outperforms ABP in terms of both efficiency and reliability. However, for smaller values of n plain BP plus threshold is extremely vulnerable to fluctuations of the number k of infected samples. This is because such fluctuations might cause the fraction of positive tests to significantly deviate from half.

5.3 Population dynamics

As already discussed, the *population dynamics* heuristics allows us to get a glimpse of the marginal distribution resulting from running BP as $n \to \infty$ [30]. To this end, we require as input the distribution of infected and healthy samples in the local neighbourhood of a sample which is provided in [23]. Subsequently, we iteratively sample the local neighbourhood for infected and healthy samples and perform one-step BP updates to model the marginal distribution of those samples whose marginal is not completely polarised. As shown in Fig. 11, the resulting distribution closely resembles the marginal distribution that we observe from running BP in our simulation in the first stage. As a side product, we obtain the proportion of polarised healthy and infected samples which lines up nicely with our simulation results. It should be noted that the population dynamics heuristic is nowhere near a complete analysis of BP on random biregular graphs. Given the gains in efficiency and reliability that we observe in this empirical work for moderately-sized instances, a formal analysis of BP seems to be an important next step in group testing research.

6 Discussion

Group testing is a powerful method to efficiently and accurately detect infected samples. Since the mathematical work on group testing deals with the asymptotic $n \to \infty$ scenario, practical adoption of methods proposed in this literature has been limited. Instead practitioners tend to apply very simple test designs dating back to the 1940s. In this paper we therefore conducted an experimental study that shows how a mildly more sophisticated test design can significantly improve the accuracy of the overall test results by comparison to classical methods without asking for many more tests. The new test design comes with an efficient,

easy-to-run and easy-to-implement algorithm that determines the status of each sample from the test results. Since the new design employs randomisation, its adoption is probably feasible only in a practical setting that employs a degree of automation in preparing test pools. But on the plus side the new ABP design keeps the pool sizes and the number of pools that each sample has to be placed in fairly low.

Apart from the group testing model studied in the present paper, there are complicated models; e.g., in *quantitative group testing* each test returns the *number* of infected samples rather than a binary positive or negative result. Further variants include the pooled data problem, the generalised coin weighing problem or the compressed sensing problem [1, 19].

What are the loose ends of the present work? On the one hand, it seems worthwhile to consider alternative noise models. A candidate might be one where the specificity decreases in the test size. Both the fixed noise model considered in this work and this diluted model have value from a practical perspective and it would be interesting to see whether our results carry over. On the other hand, the success of BP in practical group testing leaves us wondering whether it is guaranteed to converge to a fixpoint reminiscent of the ground truth. Hence, a mathematical analysis of BP remains as an outstanding open problem.

References

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

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

3 M. Aldridge. Conservative two-stage group testing. *arXiv*, 2020. `arXiv:2005.06617`.

4 M. Aldridge, L. Baldassini, and O. Johnson. Group testing algorithms: Bounds and simulations. *IEEE Transactions on Information Theory*, 60:3671–6687, 2014.

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

6 V. Bapst and A. Coja-Oghlan. Harnessing the bethe free energy. *Random Structures and Algorithms*, 49:694–741, 2016.

7 J. Barbier and D. Panchenko. Strong replica symmetry in high-dimensional optimal bayesian inference. *arXiv*, 2020. `arXiv:2005.03115`.

8 E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory and Series A*, 24, 1978.

9 V. Brault, B. Mallein, and JF Rupprecht. Group testing as a strategy for covid-19 epidemiological monitoring and community surveillance. *PLOS Computational Biology*, 17:e1008726, 2021.

10 A. Cohen and B. Kessel. False positives in reverse transcription pcr testing for sars-cov-2. *medRxiv*, page 10.1101/2020.04.26.20080911, 2020.

11 A. Coja-Oghlan, C. Efthymiou, N. Jaafari, M. Kang, and T. Kapetanopoulos. Charting the replica symmetric phase. *Communications in Mathematical Physics*, 359:603–698, 2018.

12 A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Information-theoretic and algorithmic thresholds for group testing. *Proc. 46th ICALP*, page #43, 2019.

13 A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal group testing. *Proc. 33rd COLT*, pages 1374–1388, 2020.

14 A. Coja-Oghlan and W. Perkins. Belief propagation on replica symmetric random factor graph models. *Annales de l'institut Henri Poincare D*, 5:211–249, 2018.

15 A. Coja-Oghlan and W. Perkins. Bethe states of random factor graphs. *Communications in Mathematical Physics*, 366:273–201, 2019.

16 M. Cuturi, O. Teboul, O. Berthet, A. Doucet, and J. Vert. Noisy adaptive group testing using bayesian sequential experimental design. *arXiv*, 2020. arXiv:2004.12508.

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

18 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.

19 D. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919, 2009.

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

21 C. Efthymiou, T. Hayes, D. Stefankovic, E. Vigoda, and Y. Yin. Convergence of mcmc and loopy bp in the tree uniqueness region for the hard-core model. *SIAM J. Comput.*, 48:581–643, 2019.

22 L. Garrison, J. Babigumira, A. Masaquel, B. Wang, D. Lalla, and M. Brammer. The lifetime economic burden of inaccurate her2 testing: Estimating the costs of false-positive and false-negative her2 test results in us patients with early-stage breast cancer. *Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 18:541–546, 2015.

23 O. Gebhard and P. Loick. Note on the offspring distribution for group testing in the linear regime. *arXiv*, 2021. arXiv:2103.13039.

24 E. Joly and B. Mallein. Group testing and pcr: a tale of charge value. *arXiv*, 2020. arXiv:2012.09096.

25 S. Kleinman, D. Strong, G. Tegtmeier, P. Holland, J. Gorlin, C. Cousins, R. Chiacchierini, and L. Pietrelli. Hepatitis b virus (hbv) dna screening of blood donations in minipools with the cobas ampliscreen hbv test. *Transfusion*, 45:1247–1257, 2005.

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

27 D. Levin, Y. Peres, and E. Wilmer. *Markov chains and mixing times*. AMS, 2 edition, 2017.

28 S. Mallapaty. The mathematical strategy that could transform coronavirus testing. *Nature*, 583:504–505, 2020.

29 C. McMahan, J. Tebbs, and C. Bilder. Informative dorfman screening. *Biometrics*, 68:287–296, 2012.

30 M. Mézard and A. Montanari. *Information and physics and computation*. Oxford University Pres, 2009.

31 M. Mueller, P. Derlet, C. Mudry, and G. Aeppli. Testing of asymptomatic individuals for fast feedback-control of covid-19 pandemic. *Physical biology*, 17:065007, 2020.

32 Y. Ohhashi, A. Pai, H. Halait, and R. Ziermann. Analytical and clinical performance evaluation of the cobas taqscreen mpx test for use on the cobas s201 system. *Journal of Virological Methods*, 165:246–253, 2010.

33 J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

34 M. Penschuck, U. Brandes, M. Hamann, S. Lamm, U. Meyer, I. Safro, P. Sanders, and C. Schulz. Recent advances in scalable network generation. *arXiv*, 2020. arXiv:2003.00736.

35 T. Richardson and R. Urbanke. *Modern coding theory*. Cambridge University Press, 2008.

36 D. Sejdinovic and O. Johnson. Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction. *48th Annual Allerton Conference on Communication and Control and and Computing*, pages 998–1003, 2010.

37 Noam Shental, Shlomia Levy, Vered Wuvshet, Shosh Skorniakov, Bar Shalem, Aner Ottolenghi, Yariv Greenshpan, Rachel Steinberg, Avishay Edri, Roni Gillis, Michal Goldhirsh, Khen Moscovici, Sinai Sachren, Lilach M. Friedman, Lior Nesher, Yonat Shemer-Avni, Angel Porgador, and Tomer Hertz. Efficient high-throughput sars-cov-2 testing to detect asymptomatic carriers. *Science Advances*, 6:eabc5961, 2020.

38 M. Sherlock, N. Zelota, and J. Klausner. Routine detection of acute hiv infection through rna pooling: Survey of current practice in the united states. *Sexually Transmitted Diseases*, 34:314–316, 2007.

39 J. Tebbs, C. McMahan, and C. Bilder. Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics*, 69:1064–1073, 2013.

40 L. Theagarajan. Group testing for covid-19: how to stop worrying and test more. *arXiv*, 2020. `arXiv:2004.06306`.

41 R. van der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics, 2016.

42 G. van Zyl, W. Preiser, S. Potschka, A. Lundershausen, R. Haubrich, and D. Smith. Pooling strategies to reduce the cost of hiv-1 rna load monitoring in a resource-limited setting. *Clinical Infectious Diseases*, 52:264–270, 2011.

43 P. Vontobel. Counting in graph covers: a combinatorial characterization of the bethe entropy function. *IEEE Transactions on Information Theory*, 59:6018–6048, 2013.

44 J. Watson, P. Whiting, and J. Brush. Interpreting a covid-19 test result. *BMJ*, page 369, 2020.

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

A Number of tests in first and second stage

Number of tests for the first and second stage found via optimization for various algorithms, priors and noise levels. The number of tests in the second stage in terms of the stated parameter c can be obtained as $c\lambda' n' \log(n')$ with λ' and n' defined as the average marginal and size of the low risk group, respectively.

		noiseless		moderate noise		high noise	
algorithm	prior	m1/n	c	m1/n	c	m1/n	c
BP + individual testing	0.5%	0.05	n/a	0.09	n/a	0.11	n/a
	1%	0.08	n/a	0.12	n/a	0.16	n/a
	5%	0.23	n/a	0.37	n/a	0.45	n/a
	10%	0.3	n/a	0.7	n/a	0.34	n/a
BP + InfDorfman	0.5%	0.045	n/a	0.05	n/a	0.045	n/a
	1%	0.075	n/a	0.075	n/a	0.1	n/a
	5%	0.28	n/a	0.24	n/a	0.16	n/a
	10%	0.125	n/a	0.1	n/a	0.1	n/a
aBP-1	0.5%	0.035	1.0	0.05	2.0	0.05	2.0
	1%	0.075	1.0	0.085	2.0	0.1	2.0
	5%	0.28	1.0	0.18	2.0	0.16	2.0
	10%	0.125	0.25	0.15	4.0	0.1	2.0
aBP-2	0.5%	n/a	n/a	0.075	8.0	0.02	8.0
	1%	n/a	n/a	0.12	8.0	0.03	8.0
	5%	n/a	n/a	0.4	2.0	0.36	2.0
	10%	n/a	n/a	0.5	2.0	0.325	2.0
aBP-3	0.5%	n/a	n/a	0.075	8.0	0.02	8.0
	1%	n/a	n/a	0.085	8.0	0.03	8.0
	5%	n/a	n/a	0.4	2.0	0.4	2.0
	10%	n/a	n/a	0.55	2.0	0.5	2.0

B Sample splits and test degree

The algorithms required the following number of maximum test degree and the following maximum and average split of samples. The algorithms can be readily adjusted to work with smaller test degrees or sample splits at the expense of slightly more tests.

		noiseless			moderate noise			high noise		
algorithm	prior	$\Gamma_{\max}$	$\Delta_{\max}$	Δ_{mean}	$\Gamma_{\max}$	$\Delta_{\max}$	Δ_{mean}	$\Gamma_{\max}$	$\Delta_{\max}$	Δ_{mean}
BP + individual testing	0.5%	140	8	7.0	134	13	12.0	137	16	15.0
	1%	75	7	6.0	67	9	8.0	69	12	11.0
	5%	14	4	3.1	14	6	5.2	14	7	6.2
	10%	7	3	2.3	8	6	5.2	6	3	2.8
BP + InfDorfman	0.5%	134	8	6.0	140	9	7.1	134	8	6.2
	1%	67	7	5.1	67	7	5.2	70	9	7.2
	5%	15	6	4.1	20	5	3.5	13	4	2.9
	10%	8	3	1.8	14	3	2.3	10	3	2.3
aBP	0.5%	143	8	5.2	140	11	7.6	140	13	8.0
	1%	67	8	5.1	71	12	6.7	70	13	8.0
	5%	15	7	4.1	147	12	6.2	66	12	6.5
	10%	8	3	1.8	172	19	10.3	50	10	4.9