SIAM J. MATH. DATA SCI.
Vol. 1, No. 1, pp. 161–188

Decoding from Pooled Data: Sharp Information-Theoretic Bounds*

Ahmed El Alaoui[†], Aaditya Ramdas[†], Florent Krzakala[‡], Lenka Zdeborová[§], and
Michael I. Jordan[†]

Abstract. Consider a population consisting of n individuals, each of whom has one of d types (e.g., blood types, in which case $d = 4$). We are allowed to query this population by specifying a subset of it, and in response we observe a noiseless histogram (a d-dimensional vector of counts) of types of the pooled individuals. This measurement model arises in practical situations such as pooling of genetic data and may also be motivated by privacy considerations. We are interested in the number of queries one needs to unambiguously determine the type of each individual. We study this information-theoretic question under the random, dense setting where in each query, a random subset of individuals of size proportional to n is chosen. This makes the problem a particular example of a random constraint satisfaction problem (CSP) with a "planted" solution. We establish upper and lower bounds on the minimum number of queries m such that there is no solution other than the planted one with probability tending to one as $n \to \infty$. The bounds are nearly matching. Our proof relies on the computation of the exact "annealed free energy" of this model in the thermodynamic limit, which corresponds to an exponential rate of decay of the expected number of solutions to this planted CSP. As a by-product of the analysis, we derive an identity of independent interest relating the Gaussian integral over the space of Eulerian flows of a graph to its spanning tree polynomial.

Key words. compressed sensing, categorical data, planted models, random constraint satisfaction, satisfiability threshold

AMS subject classifications. 60C05, 60F10, 60K35

DOI. 10.1137/18M1183339

1. Introduction.

The theory of compressed sensing, where one is interested in recovering a high-dimensional signal from a small number of measurements, has grown into a rich field of investigation and found many applications [24]. From the inception of this theory, it has been understood that the structure of the signal, typically sparsity, plays a key role in the sample complexity, or number of measurements needed for reconstruction [25, 11, 9]. Here one usually considers a signal that is real-valued and which is compressed by taking random

*Received by the editors April 30, 2018; accepted for publication (in revised form) October 25, 2018; published electronically February 12, 2019.

http://www.siam.org/journals/simods/1-1/M118333.html

Funding: The third author's research was supported by the EU (FP/2007-2013/ERC grant agreement 307087-SPARCS). The fifth author's research was supported by the Mathematical Data Science program of the Office of Naval Research under grant N00014-15-1-2670.

[†]Department of Electrical Engineering and Computer Sciences, and Department of Statistics, UC Berkeley, Berkeley, CA 94720-1776 (elalaoui@berkeley.edu, aramdas@berkeley.edu, jordan@berkeley.edu).

[‡]Laboratoire de Physique Statistique, Département de Physique de l'École Normale Supérieure, PSL Research University, Université Paris Diderot, Sorbonne Paris Cité, Sorbonne Universités, UPMC Université Paris 06, CNRS, 75005 Paris, France (florent.krzakala@ens.fr).

[§]Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France (lenka.zdeborova@gmail.com).

linear combinations of its entries. For a broader range of applications, it is important to consider mathematical structures other than sparsity and mechanisms of compression suitable to signals other than real-valued vectors. A notable example is matrix completion [29], relevant to recommendation systems, where the task is to recover a matrix from the observation of a subset of it entries. It is known that this task is possible if this matrix has a *low rank* [10, 35]. In this paper, we consider another example—a combinatorial setting in which each entry of the signal vector takes a value from a finite alphabet. In this discrete problem, a natural model of compression is to count the occurrence of each symbol in a few chosen subsets of the signal's entries. In other words, we consider measurements that are histograms of pooled subsets of the signal. This model is not only natural mathematically, but it is also directly motivated by applications in genetics and privacy, as we discuss later.

The reconstruction problem in this discrete, combinatorial setting can be treated as a *constraint satisfaction problem* (CSP). CSPs have been the object of intense study in recent years in probability theory, computer science, information theory, and statistical physics. For certain families of CSPs, a deep understanding has begun to emerge regarding the number of solutions as a function of problem size, as well as the algorithmic feasibility of finding solutions when they exist (see, e.g., [32, 17, 15, 16, 18, 22, 23, 40]). Consider in particular a *planted* random CSP with n variables that take their values in the discrete set $\{1, \dots, d\}$, with $d \geq 2$. A number of m clauses is drawn uniformly at random under the constraint that they are all satisfied by a prespecified assignment, which is referred to as *the planted solution*. In our case, the signal is n-dimensional, d is the size of the alphabet, and there are m compressed observations (histograms) of the signal. The signal is the planted solution that satisfies all the constraints.

Two questions are of particular importance: (1) *How large should m be so that the planted solution is the unique solution?* (2) *Given that it is unique, how large should m be so that it is recoverable by a "tractable" algorithm?* Significant progress has been made on these questions for other CSPs, often initiated by insights from statistical physics and followed by a growing body of rigorous mathematical investigation. The emerging picture is that in many planted CSPs, when n is sufficiently large, all solutions become highly correlated with the planted one when $m > \kappa_{\mathsf{IT}} \cdot n$ for some "information-theoretic" (IT) constant $\kappa_{\mathsf{IT}} > 0$. Furthermore, one of these highly correlated solutions typically becomes recoverable by a random walk or a belief propagation (BP)-inspired algorithm when $m > \kappa_{\mathsf{BP}} \cdot n$ for some $\kappa_{\mathsf{BP}} > \kappa_{\mathsf{IT}}$ [17, 33, 31, 15]. Interestingly, it is known, at least heuristically, that in many problems these algorithms fail when $\kappa_{\mathsf{IT}} < m/n < \kappa_{\mathsf{BP}}$. Moreover, a tractable algorithm that succeeds in this regime is still lacking [1, 13, 44, 16]. In other words, there is a nontrivial regime $m/n \in (\kappa_{\mathsf{IT}}, \kappa_{\mathsf{BP}})$ where an essentially unique solution exists, but it is hard to recover.

For the random CSP we consider in this paper, which we call the *histogram query problem* (HQP), we undertake a detailed information-theoretic analysis which shows that the planted solution becomes unique as soon as $m > \gamma^* n / \log n$, with high probability as $n \to \infty$ for an explicit constant $\gamma^* = \gamma^*(d) > 0$. In a companion paper [28], we consider the algorithmic aspect of the problem and provide a BP-based algorithm that recovers the planted assignment if $m \geq \kappa^* \cdot n$ for a specific threshold κ^* and fails otherwise. This leaves a logarithmic gap between the information-theoretic threshold and the point at which our algorithm succeeds.

1.1. Problem and motivation.

The setting. Let $\{\boldsymbol{h}_a\}_{1\leq a\leq m}$ be a collection of d-dimensional arrays with nonnegative integer entries. For an assignment $\tau : \{1,\dots,n\} \mapsto \{1,\dots,d\}$ of the n variables, and given a realization of m random subsets $S_a \subset \{1,\dots,n\}$, the constraints of the HQP are given by $\boldsymbol{h}_a = \boldsymbol{h}_a(\tau)$ for all $1 \leq a \leq m$, with

$$\boldsymbol{h}_a(\tau) := \left(\left|\tau^{-1}(1) \cap S_a\right|, \dots, \left|\tau^{-1}(d) \cap S_a\right|\right) \in \mathbb{Z}_+^d.$$

We let $\tau^* : \{1,\dots,n\} \mapsto \{1,\dots,d\}$ be a planted assignment; i.e., we set $\boldsymbol{h}_a := \boldsymbol{h}_a(\tau^*)$ for all a for some realization of the sets $\{S_a\}$ and consider the problem of recovering the map τ^* given the observation of the arrays $\{\boldsymbol{h}_a\}_{1\leq a\leq m}$.

This problem can be viewed informally as that of decoding a discrete high-dimensional signal consisting of categorical variables from a set of measurements formed by pooling together the variables belonging to a subset of the signal. It is useful to think of the n variables as each describing the type or category of an individual in a population of size n, where each individual has exactly one type among d. For instance, the categories may represent blood types or some other discrete feature such as ethnicity or age group. Then, the observation $\boldsymbol{h}_a$ is the histogram of types of a subpopulation S_a. We let $\boldsymbol{\pi} = \frac{1}{n}\left(\left|\tau^{*-1}(1)\right|, \dots, \left|\tau^{*-1}(d)\right|\right)$ denote the vector of proportions of assigned values, i.e., the empirical distribution of categories.

We consider here a model in which each variable participates in a given constraint independently and with probability $\alpha \in (0,1)$. Thus, the sets $\{S_a\}_{1\leq a\leq m}$ are independent draws of a random set S where $\Pr(i \in S) = \alpha$ independently for each $i \in \{1,\dots,n\}$. We are thus in the "dense regime" where $\mathbb{E}[|S|] = \alpha n$; i.e., the number of variables participating in each constraint (the degree of each factor in the CSP) is linear in n.

Motivation. This model is inspired by practical problems in which a data analyst can only assay certain summary statistics involving a moderate or large number of participants. This may be done for privacy reasons, or it may be inherent in the data-collection process (see, e.g., [39, 30]). For example, in DNA assays, the pooling of allele measurements across multiple strands of DNA is necessary given the impracticality of separately analyzing individual strands. Thus, the data consists of a frequency spectrum of alleles, a "histogram" in our language. In the privacy-related situation, one may take the viewpoint of an attacker whose goal is to gain a granular knowledge of the database from coarse measurements or that of a guard who wishes to prevent this scenario from happening. It is then natural to ask how many histogram queries it takes to exactly determine the category of each individual.

Related problems. Note that the case $d = 2$ of the HQP can be seen as a compressed sensing problem with a binary sensing matrix and binary signal. While the bulk of the literature in the field of compressed sensing is devoted to the case in which both the signal of interest and the sensing matrix are real-valued, the binary case has also been considered, notably in relation to code division multiple access (CDMA) [46, 41] and group testing [27, 34, 38]. In the case of categorical variables with $d \geq 3$ categories, it is natural to consider measurements consisting of histograms of the categories in the pooled subpopulation. In the literature on compressed sensing, one commonly considers the setting where the sensing matrices have i.i.d. entries with finite second moment, and the signal has an arbitrary empirical distribution of its entries. It has been established that, under the scaling $m = \kappa n$, whereas the success of message-passing algorithms requires $\kappa > \kappa_{\mathsf{BP}}$ [6], the information-theoretic threshold is

$\kappa_{\mathsf{IT}} = 0$ in the discrete signal case [43, 26], indicating that uniqueness of the solution happens at a finer scale $m = o(n)$. Here we consider the HQP with arbitrary d, for which the exact scaling for investigating uniqueness is $m = \gamma \frac{n}{\log n}$, with finite $\gamma > 0$, and provide tight bounds on the information-theoretic threshold.

Prior work on the HQP. The study of this problem for generic values of d was initiated in [42] in the two settings where the sets $\{S_a\}$ are deterministic or random. They showed in both these cases with a simple counting argument that under the condition that $\boldsymbol{\pi}$ is the uniform distribution, if $m < \frac{\log d}{d-1} \frac{n}{\log n}$ then the set of collected histograms does not uniquely determine the planted assignment τ^* (with high probability in the random case). On the other hand, for the deterministic setting, they provided a querying strategy that recovers τ^* provided that $m > c_0 \frac{n}{\log n}$, where c_0 is an absolute constant independent of d. For the random setting and under the condition that the sets S_a are of average size $n/2$, they proved via a first moment bound that $m > c_1 \frac{n}{\log n}$, with c_1 also constant and independent of d, suffices to uniquely determine τ^*, although no algorithm was proposed in this setting.

In the above results, there is a gap that is both information-theoretic and algorithmic depending on the dimension d between the upper and lower bounds. Intuitively, the upper bounds should also depend on d since the decoding problem becomes easier (or at least it is no harder) for large d, for the simple reason that if it is possible to determine the categories of the population for $d = 2$, then one can proceed by dichotomy for larger d by merging the d groups into two supergroups, identifying which individuals belong to each of the two supergroups, and then recurse. We attempt to fill the information-theoretic gap in the random setting by providing tighter upper and lower bounds on the number of queries m necessary and sufficient to uniquely determine the planted assignment τ^* with high probability, which depend on the dimension d and $\boldsymbol{\pi}$ along with explicit constants. In a sequel paper, we consider the algorithmic aspect of the problem and provide a BP-based algorithm that recovers the planted assignment if $m \geq \kappa^*(\boldsymbol{\pi}, d) \cdot n$ for a specific threshold $\kappa^*(\boldsymbol{\pi}, d)$ and fails otherwise, indicating the putative existence of a statistical-computational gap in the random setting.

1.2. Main result.

Let Δ^{d-1} be the $d-1$-dimensional simplex and $H(\boldsymbol{x}) = -\sum_{r=1}^{d} x_r \log x_r$ for $\boldsymbol{x} \in \Delta^{d-1}$ be the Shannon entropy function. We write $\tau \sim \boldsymbol{\pi}$ to indicate that τ is a random assignment drawn from the uniform distribution over maps $\tau : \{1, \ldots, n\} \mapsto \{1, \ldots, d\}$ such that $\frac{1}{n}\left(\left|\tau^{-1}(1)\right|, \ldots, \left|\tau^{-1}(d)\right|\right) = \boldsymbol{\pi}$.

Theorem 1.1. *For an integer $n \geq 2$, $m = \gamma \frac{n}{\log n}$, $\gamma > 0$, $\alpha \in (0,1)$, and $\boldsymbol{\pi} \in \Delta^{d-1}$ with entries bounded away from 0 and 1. Let $\mathcal{E}$ be the event that τ^* is* not *the unique satisfying assignment to the* HQP*:*

$$\mathcal{E} = \{\exists \tau \in \{1, \ldots, d\}^n \ : \ \tau \neq \tau^*, \ \boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*) \ \forall a \in \{1, \ldots, m\}\}.$$

(i) *If*

$$\gamma < \gamma_{low} := \frac{H(\boldsymbol{\pi})}{d-1},$$

then

$$\lim_{n \to \infty} \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \Pr(\mathcal{E}) = 1.$$

(ii) *On the other hand, let $\boldsymbol{\pi}_{[\cdot]}$ be the vector of order statistics of $\boldsymbol{\pi}$: $\pi_{[1]} \geq \pi_{[2]} \geq \cdots \geq \pi_{[d]}$. For $1 \leq k \leq d-1$, let $\boldsymbol{\pi}^{(k)} \in \Delta^{k-1}$ be defined as $\pi_1^{(k)} = \sum_{r=1}^{d-k+1} \pi_{[r]}$ and $\pi_l^{(k)} = \pi_{[d-k+l]}$ for all $2 \leq l \leq k$. If*

$$\gamma > \gamma_{up} := 2 \max_{1 \leq k \leq d-1} \frac{H(\boldsymbol{\pi}) - H(\boldsymbol{\pi}^{(k)})}{d-k},$$

then

$$\lim_{n \to \infty} \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \Pr(\mathcal{E}) = 0.$$

Remarks and special cases.

- For $d = 2$, $\gamma_{\mathrm{up}} = 2H(\boldsymbol{\pi}) = 2\gamma_{\mathrm{low}}$.
- If $\boldsymbol{\pi} = (\frac{1}{d}, \ldots, \frac{1}{d})$, or, more generally, if $\boldsymbol{\pi}$ is such that $k = 1$ maximizes the expression defining γ_{up}, then $\gamma_{\mathrm{up}} = 2\frac{H(\boldsymbol{\pi})}{d-1} = 2\gamma_{\mathrm{low}}$.
- The resulting bounds do not depend on α as long as it is fixed and bounded away from 0 and 1. Its contribution in the problem is subdominant and vanishes as $n \to \infty$ under the scaling considered here.
- The number k in the expression of γ_{up} can be interpreted as the number of connected components of a graph on d vertices that depends on the overlap structure of the two assignments τ and τ^* and induces "maximum confusion" between them. This will become clear in later sections.
- We note that our result can be extended to the case where each individual i is assigned a type $r \in \{1, \ldots, d\}$ *independently* with probability π_r, instead of sampling the entire population to have the profile $\boldsymbol{\pi}$ exactly. This follows by standard concentration of measure arguments; see [37] for a proof sketch.

After a preliminary version of this manuscript appeared on the preprint server arXiv, Scarlett and Cevher [37] showed that the upper bound (ii) is actually tight, in the sense that one can now replace γ_{low} by γ_{up} in the statement of the lower bound (i). Consequently, the uniqueness of the planted assignment τ^* undergoes a sharp phase transition exactly at $\gamma = \gamma_{\mathrm{up}}$. Their result combined with ours shows that the HQP has a rather unusual feature; namely, it is an example of a planted CSP where a plain first moment method identifies the exact satisfiability threshold with no conditioning needed.

The proof of the above theorem occupies the rest of the manuscript.

1.3. Main ideas of the proof.

Our main contribution is the second part of Theorem 1.1, which establishes an upper bound on the uniqueness threshold of the random CSP with histogram constraints HQP. The proof uses the first moment method to upper bound the probability of existence of a nonplanted solution. Since we are in a planted model, the analysis of the first moment ends up bearing many similarities with a second moment computation in a purely random (nonplanted) model. Although second moment computations often require approximations, for the HQP it turns out that we are able to compute the exact annealed free energy of the model in the thermodynamic limit. That is, letting $\mathcal{Z}$ be the number of solutions of the CSP, we show that the limit

$$\mathfrak{F}(\gamma) := \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\left[\mathcal{Z} - 1\right]$$

exists and we compute its value exactly. Then the value of the threshold $\gamma_{\rm up}$ is obtained by locating the first point at which $\mathfrak{F}$ becomes negative:

$$\gamma_{\rm up} = \inf\left\{\gamma > 0 \ : \ \mathfrak{F}(\gamma) < 0\right\}.$$

Together with the fact that $\mathfrak{F}$ is a monotone function, which will become clear once $\mathfrak{F}$ is computed, it is clear that for any $\gamma > \gamma_{\rm up}$, $\mathbb{E}[\mathcal{Z} - 1]$ decays exponentially with n when the latter is sufficiently large.

This general strategy has been successfully pursued for a range of CSPs, such as K-SAT, NAE-SAT, and Independent Set, most of which are Boolean. For larger domain sizes, in order to carry out the second moment method one needs fine control of the overlap structure between the planted solution and a candidate solution. This control is at the core of the difficulty that arises in any second moment computation. To obtain such control, researchers have often imposed additional assumptions, at a cost of a weakening of the resulting bounds. For example, existing proofs for Graph Coloring and similar problems assume certain balancedness conditions (the overlap matrix needs to be close to doubly stochastic) without which the annealed free energy cannot be computed [2, 14, 5, 4]; this yields results that fall somewhat short of the bounds that the second moment method could achieve in principle [20]. In the present problem, due its rich combinatorial structure, we are able to obtain unconditional control of the overlap structure, for any domain size d, and compute the exact annealed free energy.

Concretely, computing the function $\mathfrak{F}$ requires tight control of the "collision probability" of two nonequal assignments τ_1 and τ_2. This is the probability that the random histograms $\boldsymbol{h}(\tau_1) = (|{\tau_1}^{-1}(1) \cap S|, \dots, |{\tau_1}^{-1}(d) \cap S|)$ and $\boldsymbol{h}(\tau_2) = (|{\tau_2}^{-1}(1) \cap S|, \dots, |{\tau_2}^{-1}(d) \cap S|)$ generated from a random draw of a pool S coincide. The collision probability roughly measures the correlation strength between the two assignments. Specifically, we will be interested in the collision probabilities of the pairs (τ^*, τ) where τ^* is the planted assignment and τ is any candidate assignment. Its decay reveals how long an assignment τ "survives" as a satisfying assignment to the HQP as $n \to \infty$. The study of these collision probabilities requires the evaluation of certain Gaussian integrals over the space of *Eulerian flows* of a weighted graph on d vertices that is defined based on the overlap structure of τ and τ^*. We prove a family of identities that relate these integrals to some combinatorial polynomials in the weights of the graph: the spanning tree and spanning forest polynomials. We believe that these identities are of independent interest beyond the problem studied in this paper. Once these collision probabilities are controlled, the computation of $\mathfrak{F}(\gamma)$ per se requires the analysis of a certain sequence of optimization problems. We show that the sequence of maximum values converges to a finite limit that yields the value of the annealed free energy.

On the other hand, the proof of the first part of Theorem 1.1 is straightforward—it is an extension of a standard counting argument used in [45, 42]. The argument goes as follows: if m is too small, then the number of possible histograms one could potentially observe is exponentially smaller than the number of assignments of n variables that agree with $\boldsymbol{\pi}$. Therefore, when the planted assignment τ^* is drawn at random, there will exist at least one $\tau \neq \tau^*$ that satisfies the constraints of the CSP with overwhelming probability. We begin with this argument in the next section and then turn to the more challenging computation of the upper bound.

2. Proof of Theorem 1.1.

Notation. We denote vectors in $\mathbb{R}^d$ in bold lower case letters, e.g., $\boldsymbol{x}$, and matrices in $\mathbb{R}^{d\times d}$ will be written in bold lower case underlined letters, e.g., $\underline{\boldsymbol{x}}$. We denote the coordinates of such vectors and matrices as x_r and x_{rs}, respectively. Matrices that act either as linear operators on the space $\mathbb{R}^{d\times d}$ or that are functions of elements in this space are written in bold upper case letters, e.g., $\boldsymbol{M}\underline{\boldsymbol{x}}$ and $\boldsymbol{L}(\underline{\boldsymbol{x}})$, for $\underline{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$. These choices will be clear from the context. We may write $\underline{\boldsymbol{x}}/\underline{\boldsymbol{y}}$ to indicate coordinatewise division. Additionally, for two $d \times d$ matrices $\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}} \in \mathbb{R}^{d\times d}$, $\underline{\boldsymbol{a}} \odot \underline{\boldsymbol{b}} \in \mathbb{R}^{d\times d}$ is their Hadamard, or entrywise product. We let $\mathbf{1} \in \mathbb{R}^d$ be the all-ones vector.

2.1. The first part of Theorem 1.1: The lower bound.

Let $m = \gamma\frac{n}{\log n}$, with $\gamma > 0$. The number of potential histograms one could possibly observe in a single query with pool size $|S| = k$ is $f(k,d) := \binom{d+k-1}{d-1} \le (k+1)^{d-1}$ (see, e.g., Lemma 2.2 in [19]). Since the queries are independent, the number of collections of histograms $\{\boldsymbol{h}_a\}_{1\le a\le m}$ one could potentially observe in m queries is $\prod_{a=1}^m f(|S_a|,d)$. On the other hand, the number of possible assignments $\tau : \{1,\dots,n\} \mapsto \{1,\dots,d\}$ satisfying the constraint $\boldsymbol{\pi} = \frac{1}{n}\left(\left|\tau^{*-1}(1)\right|,\dots,\left|\tau^{*-1}(d)\right|\right)$ is $\binom{n}{n\boldsymbol{\pi}} = \binom{n}{n\pi_1,\dots,n\pi_d} \ge C(\boldsymbol{\pi})n^{-(d-1)/2}\exp(H(\boldsymbol{\pi})n)$ for some constant $C(\boldsymbol{\pi}) > 0$ depending on $\boldsymbol{\pi}$. (This lower bound follows from Stirling's formula.)

Now the probability that τ^* is the unique satisfying assignment of the CSP with constraints given by the random histograms $\{\boldsymbol{h}_a(\tau^*)\}_{1\le a\le m}$, averaged over the random choice of $\tau^* \sim \boldsymbol{\pi}$, is

$$
\begin{aligned}
&\mathbb{E}_{\tau^*\sim\boldsymbol{\pi}}\,\mathbb{E}_{\{S_a\}}\Big[\mathbb{1}\{\forall\tau \in \{1,\dots,d\}^n : \boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1,\dots,m\} \implies \tau = \tau^*\}\Big]\\
&\qquad\le \binom{n}{n\boldsymbol{\pi}}^{-1} \cdot \mathbb{E}_S\left[f(|S|,d)\right]^m\\
&\qquad\le \binom{n}{n\boldsymbol{\pi}}^{-1} \cdot \mathbb{E}_S\left[(|S|+1)^{d-1}\right]^m\\
&\qquad\le C(\boldsymbol{\pi})\; n^{(d-1)/2} \cdot \exp\Big(-H(\boldsymbol{\pi})n\Big) \cdot (n+1)^{m(d-1)}\\
&\qquad\le C(\boldsymbol{\pi})\; n^{(d-1)/2} \cdot \exp\Big((\gamma(d-1) - H(\boldsymbol{\pi}))n\Big).
\end{aligned}
$$

If $\gamma < \gamma_{\mathrm{low}}$, the last quantity tends to 0 as $n \to \infty$. This concludes the proof of the first assertion of the theorem.

2.2. The second part of Theorem 1.1: The upper bound.

We use a first moment method to show that when γ is greater than γ_{up}, the only assignment satisfying the HQP is τ^* with high probability. Let $\mathcal{Z}$ be the number of satisfying assignments to the HQP:

$$
\mathcal{Z} := \Big|\{\tau \in \{1,\dots,d\}^n \;:\; \boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1,\dots,m\}\}\Big|. \tag{2.1}
$$

The planted assignment τ^* is obviously a solution, so we always have $\mathcal{Z} \ge 1$. Recall the definition of the annealed free energy

$$
\mathfrak{F}(\gamma) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\mathcal{Z} - 1\right]. \tag{2.2}
$$

Also, recall that for $1 \le k \le d-1$, $\boldsymbol{\pi}^{(k)} \in \Delta^{k-1}$ is defined as $\pi_1^{(k)} = \sum_{r=1}^{d-k+1} \pi_{[r]}$ and $\pi_l^{(k)} = \pi_{[d-k+l]}$ for all $2 \le l \le k$ (if $k \ge 2$).

Theorem 2.1. *Let $m = \gamma \frac{n}{\log n}$, with $\gamma > 0$. The limit (2.2) exists for all $\gamma > 0$, and its value is*

$$\mathfrak{F}(\gamma) = \max_{1 \le k \le d-1} \left\{ H(\boldsymbol{\pi}) - H(\boldsymbol{\pi}^{(k)}) - \frac{\gamma}{2}(d-k) \right\}. \tag{2.3}$$

We can deduce from Theorem 2.1 the smallest value of γ past which $\mathfrak{F}(\gamma)$ becomes negative. In particular, we see that $\mathfrak{F}$ is a decreasing function of γ that crosses the horizontal axis at

$$\gamma_{\text{up}} = 2 \max_{1 \le k \le d-1} \frac{H(\boldsymbol{\pi}) - H(\boldsymbol{\pi}^{(k)})}{d-k}.$$

From this result, it is easy to prove the second assertion of Theorem 1.1. By averaging over τ^* and applying Markov's inequality, we have

$$\begin{aligned} &\mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \Pr\left(\exists \tau \in \{1, \dots, d\}^n : \tau \ne \tau^*, \boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right) \\ &\qquad = \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \Pr\left(\mathcal{Z} \ge 2\right) \le \mathbb{E}[\mathcal{Z} - 1]. \end{aligned}$$

For $\gamma > \gamma_{\text{up}}$, it is clear that $\mathfrak{F}(\gamma) < 0$. Therefore,

$$\begin{aligned} &\lim \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \Pr\left(\exists \tau \in \{1, \dots, d\}^n \ :\ \tau \ne \tau^*, \boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right) \\ &\qquad = \lim \exp(n\ \mathfrak{F}(\gamma)) = 0. \end{aligned}$$

Now it remains to prove Theorem 2.1, and this represents the main technical thrust of our paper.

2.3. Collisions, overlaps, and the first moment.

Preliminaries. We begin by presenting the main quantities to be analyzed in our application of the first moment method. We have

$$\begin{aligned} \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \mathbb{E}_{\{S_a\}}[\mathcal{Z} - 1] &= \mathbb{E}_{\tau^* \sim \boldsymbol{\pi}} \left[\sum_{\substack{\tau \in \{1, \dots, d\}^n \\ \tau \ne \tau^*}} \Pr\left(\boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right) \right] \\ &= (d^n - 1) \Pr_{\tau, \tau^*, \{S_a\}} \left(\boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right), \end{aligned}$$

where $\tau^* \sim \boldsymbol{\pi}$, $\tau \sim \text{Unif}(\{1, \dots, d\}^n \backslash \{\tau^*\})$. By conditional independence,

$$\begin{aligned} &\Pr_{\tau, \tau^*, \{S_a\}} \left(\boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right) \\ &= \mathbb{E}_{\tau, \tau^*} \left[\Pr_{\{S_a\}} \left(\boldsymbol{h}_a(\tau) = \boldsymbol{h}_a(\tau^*)\ \forall a \in \{1, \dots, m\}\right) \right] = \mathbb{E}_{\tau, \tau^*} \left[\Pr_{S} \left(\boldsymbol{h}(\tau) = \boldsymbol{h}(\tau^*)\right)^m \right]. \end{aligned}$$

Next, we write the *collision probability*, $\Pr_S(\boldsymbol{h}(\tau) = \boldsymbol{h}(\tau^*))$, for fixed τ and τ^* in a convenient form. Let us first define the *overlap matrix*, $\underline{\boldsymbol{\mu}}(\tau, \tau^*) = (\mu_{rs})_{1 \le r,s \le d} \in \mathbb{Z}_+^{d \times d}$, of τ and τ^*, by

$$\mu_{rs} = \left|\tau^{-1}(r) \cap \tau^{*-1}(s)\right| \quad \text{for all } r, s = 1, \dots, d. \tag{2.4}$$

Remark that $\boldsymbol{h}(\tau) = \boldsymbol{h}(\tau^*)$ if and only if $\left|S \cap \tau^{-1}(r)\right| = \left|S \cap \tau^{*-1}(r)\right|$ for all $r \in \{1, \dots, d\}$. Since the collection of sets $\{\tau^{-1}(r)\}_{1 \le r \le d}$ forms a partition of $\{1, \dots, n\}$, and similarly with τ^*, the event $\{\boldsymbol{h}(\tau) = \boldsymbol{h}(\tau^*)\}$ is the same as

$$\left\{\sum_{s=1}^{d} \left|S \cap \tau^{-1}(r) \cap \tau^{*-1}(s)\right| = \sum_{s=1}^{d} \left|S \cap \tau^{-1}(s) \cap \tau^{*-1}(r)\right| \ \ \forall r \in \{1, \dots, d\}\right\}.$$

Recall that each i is included in S independently with probability α. Therefore, the probability that two assignments τ and τ^* collide on a random pool S—meaning that their histograms formed on the pool S coincide—is

$$\Pr_S(\boldsymbol{h}(\tau) = \boldsymbol{h}(\tau^*)) = \sum_{\underline{\boldsymbol{\nu}}} \left(\prod_{r,s=1}^{d} \binom{\mu_{rs}}{\nu_{rs}} \alpha^{\nu_{rs}} (1-\alpha)^{\mu_{rs} - \nu_{rs}}\right) \mathbb{1}\left\{\underline{\boldsymbol{\nu}}\mathbf{1} = \underline{\boldsymbol{\nu}}^\intercal \mathbf{1}\right\}, \tag{2.5}$$

where the outer sum is over all arrays of integer numbers $\underline{\boldsymbol{\nu}} = (\nu_{rs})_{1 \le r,s \le d}$ such that $0 \le \nu_{rs} \le \mu_{rs}$ for all r, s. We see from the above expression that the collision probability of τ and τ^* only depends on the overlap matrix $\underline{\boldsymbol{\mu}}(\tau, \tau^*)$. We henceforth denote the probability in (2.5) by $q(\underline{\boldsymbol{\mu}})$, where we dropped the dependency on τ and τ^*. Remark that $\tau = \tau^*$ if and only if their overlap matrix $\underline{\boldsymbol{\mu}}$ is diagonal. Thus, we can rewrite the expected number of solutions as

$$\mathbb{E}[\mathcal{Z} - 1] = \binom{n}{n\boldsymbol{\pi}}^{-1} \cdot \sum_{\underline{\boldsymbol{\mu}}} \binom{n}{\underline{\boldsymbol{\mu}}} q(\underline{\boldsymbol{\mu}})^m \ \mathbb{1}\left\{\underline{\boldsymbol{\mu}}^\intercal \mathbf{1} = n\boldsymbol{\pi}\right\}, \tag{2.6}$$

where the sum is over all nondiagonal arrays $\underline{\boldsymbol{\mu}} = (\mu_{rs})_{1 \le r,s \le d}$ with nonnegative integer entries that sum to n, and $\binom{n}{\underline{\boldsymbol{\mu}}} = \frac{n!}{\prod_{r,s} \mu_{rs}!}$.

The rest of the proof. From here, the proof of Theorem 2.1 roughly breaks into three parts:

(i) One needs to have tight asymptotic control on the collision probability $q(\underline{\boldsymbol{\mu}})$ when any subset of the entries of $\underline{\boldsymbol{\mu}}$ become large. This will be achieved via the Laplace method (see, e.g., [21]). The outcome of this analysis is an asymptotic estimate that exhibits two different speeds of decay, polynomial or exponential, depending on the "balancedness" of $\underline{\boldsymbol{\mu}}$ as its entries become large. This notion of balancedness, namely that $\underline{\boldsymbol{\mu}}$ must have equal row- and column-sums,[1] is specific to the histogram setting and departs from the usual "double stochasticity" that arises in other more classical problems, such as Graph Coloring and Community Detection under the stochastic block model [3, 2, 14, 5, 4]. As we will explain in the next section, configurations (τ, τ^*) with an unbalanced overlap matrix have an exponentially

[1] These are exactly the constraints on $\underline{\boldsymbol{\nu}}$ showing up in (2.5).

decaying collision probability; i.e., they exhibit weak correlation and disappear very early on as $n \to \infty$ under the scaling $m = \gamma \frac{n}{\log n}$. On the other hand, those configurations with balanced overlap exhibit a slow decay of correlation: their collision probability decays only polynomially, and these are the last surviving configurations in expression (2.6) as $n \to \infty$.

(ii) Understanding the above-mentioned polynomial decay of $q(\underline{\boldsymbol{\mu}})$ requires the evaluation of a multivariate Gaussian integral (which is a product of the above analysis) over the space of constraints of the array $\underline{\boldsymbol{\nu}}$ in (2.5), the latter being the space of *Eulerian flows* on the graph on d vertices whose edges are weighted by the (large) entries of $\underline{\boldsymbol{\mu}}$. We show that this integral, properly normalized, evaluates to *the inverse square root of the spanning tree (or forest) polynomial* of this graph. This identity seems to be new, to the best of our knowledge, and may be of independent interest. We therefore provide two very different proofs of it, each highlighting different combinatorial aspects.

(iii) Finally, armed with these estimates, we show the existence of, and compute the exact value of, the annealed free energy of the model in the thermodynamic limit, thereby completing the proof of Theorem 2.1. This last part requires the analysis of a certain optimization problem involving an entropy term and an "energy" term accounting for the correlations discussed above. Here we can exactly characterize the maximizing configurations for large n, and this allows the computation of the value of $\mathfrak{F}(\gamma)$. We note once more that this situation contrasts with the more traditional case of Graph Coloring, where we lack a rigorous understanding of the maximizing configurations of the second moment, except when certain additional constraints are imposed on their overlap matrix.

3. Bounding the collision probabilities.

Here we provide tight asymptotic bounds on the collision probabilities $q(\underline{\boldsymbol{\mu}})$ defined in (2.5). Consider the following subspace of $\mathbb{R}^{d\times d}$, which will play a key role in the analysis:

$$\mathcal{F} := \left\{ \underline{\boldsymbol{x}} \in \mathbb{R}^{d\times d} \ : \ \sum_{s=1}^{d} x_{rs} = \sum_{s=1}^{d} x_{sr} \ \forall r \in \{1,\dots,d\} \right\}. \tag{3.1}$$

This is a linear subspace of dimension $(d-1)^2 + d$ in $\mathbb{R}^{d\times d}$. For $p, q \in (0,1)$, let $D(p \parallel q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ be the Kullback–Leibler divergence of the Bernoulli laws with parameters p and q. Let $G = (V, E)$ be an undirected graph on d vertices where we allow up to two parallel edges between each pair of vertices, i.e., $V = \{1, \dots, d\}$, and $E \subseteq \{(r,s) \ : \ r, s \in V, \ r \neq s\}$. For $\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}} \in \mathbb{R}_+^{d\times d}$, $\underline{\boldsymbol{x}} \in [0,1]^{d\times d}$, let

$$\varphi_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{x}}) := \sum_{(r,s)\in E} \mu_{rs} D(x_{rs} \parallel \alpha), \tag{3.2}$$

and recalling that $\odot$ represents the Hadamard (entrywise) product, we let

$$\vartheta(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}}) := \min_{\substack{\underline{\boldsymbol{x}}\in[0,1]^{d\times d} \\ \boldsymbol{M}_G(\underline{\boldsymbol{x}}\odot\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\nu}})\in\mathcal{F}}} \varphi_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{x}}), \tag{3.3}$$

where for two $d \times d$ matrices $\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}$, $\boldsymbol{M}_G(\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}})$ is the $d \times d$ matrix with entries a_{rs} if $(r,s) \in E$ and b_{rs} otherwise. It is a simple consequence of Lagrange duality of convex programming and

Slater's conditions (see, e.g., [8, 36]), that the function (3.3) can be written in an equivalent form as

$$\begin{aligned}\vartheta(\underline{\boldsymbol{\nu}},\underline{\boldsymbol{\mu}}) &= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^d}\left\{\sum_{(r,s)\notin E}\nu_{rs}(\lambda_r-\lambda_s)+\sum_{(r,s)\in E}\mu_{rs}\log\left(\frac{e^{\lambda_r-\lambda_s}}{\alpha+(1-\alpha)e^{\lambda_r-\lambda_s}}\right)\right\}\\ &= \phi^*_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{\nu}}\mathbf{1}-\underline{\boldsymbol{\nu}}^\intercal\mathbf{1}),\end{aligned}$$

where $\phi^*_{\underline{\boldsymbol{\mu}}}$ is the Legendre–Fenchel transform of the (convex) function

$$\phi_{\underline{\boldsymbol{\mu}}}(\boldsymbol{\lambda}) := -\sum_{(r,s)\in E}\mu_{rs}\log\left(\frac{e^{\lambda_r-\lambda_s}}{\alpha+(1-\alpha)e^{\lambda_r-\lambda_s}}\right).$$

We may note that since $\phi^*_{\underline{\boldsymbol{\mu}}}$ is convex on $\mathbb{R}^d$, ϑ is a continuous function of its first argument. Before we state our bounds on the collision probability, we recall the following concept from algebraic graph theory. Define *the spanning tree polynomial* of G as

$$T_G(\underline{\boldsymbol{z}}) := \frac{1}{\mathsf{nst}(G)}\sum_T\prod_{(r,s)\in T} z_{rs} \tag{3.4}$$

for $\underline{\boldsymbol{z}}\in\mathbb{R}_+^{d\times d}$, where the sum is over all spanning trees of G, and $\mathsf{nst}(G)$ is the number of spanning trees of G. In cases where G is not connected, we define the following polynomial:

$$P_G := \prod_{l=1}^{\mathsf{ncc}(G)} T_{G_l}, \tag{3.5}$$

where G_l is the lth connected component of G, and we denote by $\mathsf{ncc}(G)$ the number of connected components of G. This polynomial may be interpreted as the generating polynomial of *spanning forests* of G having exactly $\mathsf{ncc}(G)$ trees. The polynomials T_G and P_G are multi-affine, homogeneous of degree $d-1$ for T_G (when G is connected) and $d-\mathsf{ncc}(G)$ for P_G, and do not depend on the diagonal entries $\{z_{rr}: 1\le r\le d\}$. Furthermore, letting $z_{rs}=1$ for all $r\neq s$, we have $P_G(\underline{\boldsymbol{z}})=T_G(\underline{\boldsymbol{z}})=1$. We now provide tight asymptotic bounds on the collision probability $q(\underline{\boldsymbol{\mu}})$ when a subset E of the entries of $\underline{\boldsymbol{\mu}}$ become large.

Theorem 3.1. *Let $G=(V,E)$, with $V=\{1,\dots,d\}$, $E=\{(r,s)\in V^2 \;:\; r\neq s\}$, and $\epsilon\in(0,1)$. There exist two constants $0<c_u<c_l$ depending on ϵ, d, and α such that for all n sufficiently large, and all $\underline{\boldsymbol{\mu}}\in\{0,\dots,n\}^{d\times d}$ with $\mu_{rs}\ge\epsilon n$ if and only if $(r,s)\in E$, we have*

$$c_l\frac{e^{-\vartheta_l(\underline{\boldsymbol{\mu}})}}{P_G(\underline{\boldsymbol{\mu}})^{1/2}}\le q(\underline{\boldsymbol{\mu}})\le c_u\frac{e^{-\vartheta_u(\underline{\boldsymbol{\mu}})}}{P_G(\underline{\boldsymbol{\mu}})^{1/2}},$$

with

$$\vartheta_u(\underline{\boldsymbol{\mu}})=\inf_{\underline{\boldsymbol{\nu}}}\{\vartheta(\underline{\boldsymbol{\nu}},\underline{\boldsymbol{\mu}}):0\le\nu_{rs}\le\mu_{rs}\;\forall(r,s)\notin E\}$$

and

$$\vartheta_l(\underline{\boldsymbol{\mu}})=\sup_{\underline{\boldsymbol{\nu}}}\{\vartheta(\underline{\boldsymbol{\nu}},\underline{\boldsymbol{\mu}}):0\le\nu_{rs}\le\mu_{rs}\;\forall(r,s)\notin E\}.$$

Let us now expand on the above result and derive some special cases and corollaries. First, we see that the collision probabilities can decay at two different speeds—polynomial or exponential—in the entries of the overlap matrix $\underline{\boldsymbol{\mu}}$, depending on whether $\vartheta_u(\underline{\boldsymbol{\mu}})$ (and/or $\vartheta_l(\underline{\boldsymbol{\mu}})$) is zero or strictly negative. Second, the apparent gap in the exponential decay of $q(\underline{\boldsymbol{\mu}})$ in the above characterization is artificial; one can make ϑ_u and ϑ_l equal by taking $\mu_{rs} = 0$ for all $(r,s) \notin E$. Alternatively, they could be made arbitrarily close to each other under an appropriate limit: Assume for simplicity that $\mu_{rs} = nw_{rs} > 0$ for all $(r,s) \in E$ for some $\underline{\boldsymbol{w}} \in [0,1]^{d\times d}$. We have

$$\vartheta(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}}) = n\vartheta(\underline{\boldsymbol{\nu}}/n, \underline{\boldsymbol{w}}).$$

For $(r,s) \notin E$, we have $\mu_{rs} < \epsilon n$, and therefore

$$\vartheta_u(\underline{\boldsymbol{\mu}})/n \le \inf_{\underline{\boldsymbol{x}}} \{\vartheta(\underline{\boldsymbol{x}}, \underline{\boldsymbol{w}}) : 0 \le x_{rs} \le \epsilon \ \forall (r,s) \notin E\} \xrightarrow[\epsilon \to 0]{} \vartheta(\mathbf{0}, \underline{\boldsymbol{w}}).$$

The last step is justified by the continuity of $\vartheta(\cdot, \underline{\boldsymbol{w}})$. The same argument holds for $v_l(\underline{\boldsymbol{\mu}})$. Denoting the limiting function under this operation as $\vartheta(\underline{\boldsymbol{w}})$, we obtain

$$\vartheta(\underline{\boldsymbol{w}}) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^d} \sum_{(r,s)\in E} w_{rs} \log\left(\frac{e^{\lambda_r - \lambda_s}}{\alpha + (1-\alpha)e^{\lambda_r - \lambda_s}}\right) = \min_{\substack{\underline{\boldsymbol{x}} \in [0,1]^{d\times d} \\ \underline{\boldsymbol{w}} \odot \underline{\boldsymbol{x}} \in \mathcal{F}}} \varphi_{\underline{\boldsymbol{w}}}(\underline{\boldsymbol{x}}).$$

The function ϑ can be seen as the exponential rate of decay of $q(\underline{\boldsymbol{\mu}})$. The reason ϑ_u and ϑ_l cannot (in general) be replaced by ϑ in Theorem 3.1 is that all control on the constants c_u and c_l is lost when $\epsilon \to 0$. Next, we identify the cases where this exponential decay is nonvacuous.

Lemma 3.2. *Let $\alpha \in (0,1)$, and let $\underline{\boldsymbol{\mu}} \in \mathbb{R}_+^{d\times d}$. We have the following:*

(i) *$\vartheta(\underline{\boldsymbol{\mu}}) = 0$ if and only if $\underline{\boldsymbol{\mu}} \in \mathcal{F}$.*

(ii) *$\vartheta_u(\underline{\boldsymbol{\mu}}) = 0$ if and only if $\boldsymbol{M}_G(\alpha\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\nu}}) \in \mathcal{F}$ for* some *$\underline{\boldsymbol{\nu}} \in \mathbb{R}_+^{d\times d}$ such that $0 \le \nu_{rs} \le \mu_{rs}$ for all $(r,s) \notin E$.*

Now we specialize Theorem 3.1 to the case where the entries of the overlap matrix are either zero or grow proportionally to n. From Theorem 3.1 and Lemma 3.2, we deduce a key corollary on the convergence of the properly rescaled logarithm of the collision probabilities.

Corollary 3.3. *Given a graph $G = (V,E)$, let $\underline{\boldsymbol{w}} \in [0,1]^{d\times d}$ be such that $w_{rs} > 0$ if and only if $(r,s) \in E$. If $\underline{\boldsymbol{w}} \in \mathcal{F}$, then*

$$\lim_{n\to\infty} \frac{\log q(n\underline{\boldsymbol{w}})}{\log n} = -\frac{d - \mathsf{ncc}(G)}{2}.$$

Otherwise, if $\underline{\boldsymbol{w}} \notin \mathcal{F}$, then

$$\lim_{n\to\infty} \frac{\log q(n\underline{\boldsymbol{w}})}{n} = -\vartheta(\underline{\boldsymbol{w}}).$$

We see that the assignments τ such that $\underline{\boldsymbol{\mu}}(\tau, \tau^*) \in \mathcal{F}$ exhibit a much stronger correlation to τ^* than those for which this overlap matrix does not belong to $\mathcal{F}$ and will hence survive much longer as $n \to \infty$. This has an intuitive explanation: pools S that do not sample a fraction α of $\tau^{-1}(r) \cap \tau^{*-1}(s)$ for all r, s make an exponentially small contribution to the sum (2.5) defining $q(\mu)$, since they do not maximize the binomial terms in (2.5). If $\mu(\tau, \tau^*) \notin \mathcal{F}$, then this "fair" sampling is not allowed by the constraints defining the sum. As a result, the whole sum is exponentially small.

Proof of Lemma 3.2. Let $\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\nu}} \in \mathbb{R}_+^{d\times d}$, with $\underline{\boldsymbol{\mu}} \neq \underline{\mathbf{0}}$. Let $\alpha \in (0,1)$, and let $G = (V,E)$ denote a graph on d vertices. The function $\varphi_{\underline{\boldsymbol{\mu}}}$ defined in (3.2) is strictly convex on the support of $\underline{\boldsymbol{\mu}}$, i.e., on the subspace induced by the nonzero coordinates of $\underline{\boldsymbol{\mu}}$, so it admits a unique minimizer on the closed convex set $\{\underline{\boldsymbol{x}} \in [0,1]^{d\times d} \;:\; \boldsymbol{M}_G(\underline{\boldsymbol{x}}^* \odot \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\nu}}) \in \mathcal{F}\}$ intersected with that subspace. Let $\underline{\boldsymbol{x}}^*$ be this minimizer. By differentiating the associated Lagrangian, the entries of $\underline{\boldsymbol{x}}^*$ admit the expressions

$$x^*_{rs} = \frac{\alpha}{\alpha + (1-\alpha)e^{\lambda_r - \lambda_s}}$$

for all $(r,s) \in E$ (recall that $\mu_{rs} > 0$ for all such (r,s)) and where the vector $\boldsymbol{\lambda} \in \mathbb{R}^d$ is the unique solution up to global shifts of the system of equations: for all $r \in \{1,\dots,d\}$,

$$\sum_{s:(r,s)\in E} \frac{\alpha\mu_{rs}}{\alpha + (1-\alpha)e^{\lambda_r-\lambda_s}} + \sum_{s:(r,s)\notin E} \nu_{rs} = \sum_{s:(r,s)\in E} \frac{\alpha\mu_{sr}}{\alpha + (1-\alpha)e^{\lambda_s-\lambda_r}} + \sum_{s:(r,s)\notin E} \nu_{sr}. \tag{3.6}$$

The claims of the lemma follow directly from the system of equations (3.6) and the fact that the nonnegative function $\varphi_{\underline{\boldsymbol{\mu}}}$ vanishes if and only if $x^*_{rs} = \alpha$ for all $(r,s) \in E$: to show (i), we take $\underline{\boldsymbol{\nu}} = \underline{\mathbf{0}}$. It is clear from the equations that $\underline{\boldsymbol{\mu}} \in \mathcal{F}$ if and only if $\boldsymbol{\lambda} = c\mathbf{1}$, $c \in \mathbb{R}$, is a solution to the above equations; and this is equivalent to $x^*_{rs} = \alpha$ whenever $\mu_{rs} > 0$. This is in turn equivalent to $\vartheta(\underline{\boldsymbol{\mu}}) = \varphi_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{x}}^*) = 0$. The same strategy is employed to show (ii), in conjunction with the continuity of the function $\underline{\boldsymbol{\nu}} \mapsto \vartheta(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}})$ over a compact domain (the infimum defining ϑ_u is attained). ■

Proof of Corollary 3.3. Fix $G = (V,E)$, let $\underline{\boldsymbol{w}} \in (0,1)^{d\times d}$ with $w_{rs} > 0$ if and only if $(r,s) \in E$, and let n be an integer. For simplicity, assume that for $n\underline{\boldsymbol{w}}$ is an array of integer entries. The noninteger part introduces easily manageable error terms. Applying Theorem 3.1 with $\epsilon = \min_{(r,s)\in E} w_{rs}$, we have for n large

$$c_l P_G(n\underline{\boldsymbol{w}})^{-1/2} \exp -\vartheta_l(n\underline{\boldsymbol{w}}) \le q(n\underline{\boldsymbol{w}}) \le c_u P_G(n\underline{\boldsymbol{w}})^{-1/2} \exp -\vartheta_u(n\underline{\boldsymbol{w}}).$$

Moreover, since $w_{rs} = 0$ for $(r,s) \notin E$, we have

$$\vartheta_u(n\underline{\boldsymbol{w}}) = \vartheta_l(n\underline{\boldsymbol{w}}) = n\vartheta(\underline{\boldsymbol{w}}).$$

On the other hand, by homogeneity of the polynomial P_G (3.5), $P_G(n\underline{\boldsymbol{w}}) = n^{d-\mathsf{ncc}(G)} P_G(\underline{\boldsymbol{w}})$. Applying Lemma 3.2 yields the desired result: If $\underline{\boldsymbol{w}} \in \mathcal{F}$, then

$$\lim_{n\to\infty} \frac{\log q(n\underline{\boldsymbol{w}})}{\log n} = -\frac{d - \mathsf{ncc}(G)}{2}.$$

Otherwise,

$$\lim_{n\to\infty} \frac{\log q(n\underline{\boldsymbol{w}})}{n} = -\vartheta(\underline{\boldsymbol{w}}).$$

■

3.1. A Gaussian integral. One important step in proving Theorem 3.1 (specifically for obtaining the polynomial decay part of $q(\underline{\boldsymbol{\mu}})$) is the following identity relating the Gaussian integral on a linear space $\mathcal{F}(G)$ defined based on a graph G to the spanning tree/forest polynomial of G. We denote by K_d the complete graph on d vertices where every pair of distinct vertices is connected by *two parallel edges*.

Proposition 3.4. *Let $G = (V, E)$ be a graph on d vertices, where self-loops and up to two parallel edges are allowed: $V = \{1, \dots, d\}$, $E \subseteq V \times V$. Further, let*

$$\mathcal{F}(G) = \Big\{ \underline{\boldsymbol{x}} \in \mathcal{F} \ : \ x_{rs} = 0 \textit{ for } (r, s) \notin E \Big\}.$$

For any array of positive real numbers $(w_{rs})_{(r,s)\in E}$, we have

$$\int_{\mathcal{F}(G)} e^{-\sum_{rs} x_{rs}^2/2w_{rs}} \, \mathrm{d}\underline{\boldsymbol{x}} = \left((2\pi)^{\dim(\mathcal{F}(G))} \, \frac{\prod_{r,s} w_{rs}}{P_G(\underline{\boldsymbol{w}})} \right)^{1/2}.$$

In the case where G is the complete graph K_d, $\mathcal{F}(G) = \mathcal{F}$, $\dim(\mathcal{F}) = (d-1)^2 + d$, and $P_G = T_G = (2^{d-1} d^{d-2})^{-1} \sum_T \prod_{(r,s)\in T} w_{rs}$, where the sum is over all spanning trees of K_d. The prefactor in the last expression comes from Cayley's formula for the number of spanning trees of the complete graph [12]. We will show that it suffices to prove Proposition 3.4 in the case where $G = K_d$ in order to establish it for any graph G. We were not able to locate this identity in the literature. Our proof proceeds by relating the above Gaussian integral to the characteristic polynomial of the Laplacian matrix of G then invoking the Principal Minors Matrix-Tree theorem (see, e.g., [12]).

4. Computing the annealed free energy.

In this section, we establish the existence of $\mathfrak{F}(\gamma)$ and compute its value for all $\gamma > 0$. For $1 \le k \le d$, let $\mathcal{D}_k$ denote the set of binary matrices $\boldsymbol{X} \in \{0, 1\}^{k\times d}$ such that each column of $\boldsymbol{X}$ contains *exactly* one nonzero entry and each row contains *at least* one nonzero entry. The elements of $\mathcal{D}_k$ represent partitions of the set $\{1, \dots, d\}$ into k nonempty subsets.

Proposition 4.1. *Let $m = \gamma \frac{n}{\log n}$, with $\gamma > 0$ fixed for all $n \ge 2$. We have*

$$\mathfrak{F}(\gamma) = \max_{1\le k\le d-1} \left\{ H(\boldsymbol{\pi}) - \min_{\boldsymbol{X}\in\mathcal{D}_k} H(\boldsymbol{X}\boldsymbol{\pi}) - \frac{\gamma}{2}(d-k) \right\}.$$

Moreover, the inner minimization problem in the above expression can be solved explicitly.

Lemma 4.2. *Let $\boldsymbol{\pi}_{[\cdot]}$ be a permutation of the vector $\boldsymbol{\pi}$ such that $\pi_{[1]} \ge \pi_{[2]} \ge \cdots \ge \pi_{[d]}$. And for $1 \le k \le d-1$, let $\boldsymbol{\pi}^{(k)} \in \Delta^{k-1}$ be defined as $\pi_1^{(k)} = \sum_{r=1}^{d-k+1} \pi_{[r]}$ and $\pi_l^{(k)} = \pi_{[d-k+l]}$ for all $2 \le l \le k$. Then*

$$\min_{\boldsymbol{X}\in\mathcal{D}_k} H(\boldsymbol{X}\boldsymbol{\pi}) = H(\boldsymbol{\pi}^{(k)}).$$

Theorem 2.1 follows from Proposition 4.1 and Lemma 4.2. We begin with the proof of the latter and devote the next subsection to the lengthier proof of the former.

Proof of Lemma 4.2. We start with an arbitrary partition of $\boldsymbol{\pi}$ into k groups and define a sequence of operations on the set of k-partitions of $\boldsymbol{\pi}$ that strictly decreases $H(\underline{\boldsymbol{X}}\boldsymbol{\pi})$ at each step and, irrespective of the starting point, always converges to $\boldsymbol{\pi}^{(k)}$. Starting with an arbitrary k-partition, write down the groups from left to right in decreasing order of the total weight of each group. Initially, every group is marked *incomplete*. Then we perform the following operations:

1. Start with the rightmost incomplete group.
2. If it has more than one element, transfer the largest element to the leftmost group. This strictly decreases the entropy, since the heaviest group gets heavier and the lightest group gets lighter. Repeat this step until the rightmost group has exactly one element, and then move to the next step.
3. Consider this (now singleton) group. If there is no element to its left that is lighter than it, mark the group as complete. Otherwise, swap this element with the lightest element to its left, and then mark it complete. Then go back to step 1. ∎

4.1. Proof of Proposition 4.1.

Let $m = \gamma\frac{n}{\log n}$. Recall from (2.6) that

$$\mathbb{E}[\mathcal{Z}-1] = \binom{n}{n\boldsymbol{\pi}}^{-1} \cdot \sum_{\underline{\boldsymbol{\mu}}} \binom{n}{\underline{\boldsymbol{\mu}}} q(\underline{\boldsymbol{\mu}})^m \; \mathbb{1}\left\{\underline{\boldsymbol{\mu}}^\intercal \mathbf{1} = n\boldsymbol{\pi}\right\},$$

where the sum is over all arrays $\underline{\boldsymbol{\mu}} \in \mathbb{Z}_+^{d\times d}$ such that $\mathbf{1}^\intercal\underline{\boldsymbol{\mu}}\mathbf{1} = n$, $1 \le \sum_{r\neq s}\mu_{rs}$ (recall that this last constraint excludes τ^* from the sum). Since the sum defining $\mathbb{E}[\mathcal{Z}-1]$ is larger than its maximum term and smaller than the maximum term times $(n+1)^{d^2}$, we only need to understand the convergence of the sequence

$$\begin{aligned}\mathfrak{F}_n &:= \frac{1}{n}\log\left(\max_{\substack{\underline{\boldsymbol{\mu}}\in\{0,\ldots,n\}^{d\times d}\\ \text{nondiagonal}}} \binom{n}{\underline{\boldsymbol{\mu}}} q(\underline{\boldsymbol{\mu}})^m \; \mathbb{1}\left\{\underline{\boldsymbol{\mu}}^\intercal\mathbf{1} = n\boldsymbol{\pi}\right\}\right)\\ &= \max\left\{\frac{1}{n}\log\binom{n}{\underline{\boldsymbol{\mu}}} + \gamma\frac{\log q(\underline{\boldsymbol{\mu}})}{\log n} \; : \; \underline{\boldsymbol{\mu}} \in \{0,\ldots,n\}^{d\times d}, \sum_{r\neq s}\mu_{rs} \geq 1, \underline{\boldsymbol{\mu}}^\intercal\mathbf{1} = n\boldsymbol{\pi}\right\}.\end{aligned}$$

If this sequence converges, we would have

$$\mathfrak{F}(\gamma) = -H(\boldsymbol{\pi}) + \lim_{n\to\infty}\mathfrak{F}_n, \tag{4.1}$$

since $\frac{1}{n}\log\binom{n}{n\boldsymbol{\pi}} \to H(\boldsymbol{\pi})$ by Stirling's formula. Next, we show that the above limit indeed exists. Let

$$\psi_n(\underline{\boldsymbol{w}}) := \frac{1}{n}\log\binom{n}{n\underline{\boldsymbol{w}}} + \gamma\frac{\log q(n\underline{\boldsymbol{w}})}{\log n}. \tag{4.2}$$

By Corollary 3.3, the function

$$\psi(\underline{\boldsymbol{w}}) := \begin{cases} H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - \mathsf{ncc}(\underline{\boldsymbol{w}})) & \text{if } \underline{\boldsymbol{w}} \in \mathcal{F},\\ -\infty & \text{otherwise}\end{cases} \tag{4.3}$$

is the pointwise limit of the sequence of functions $\{\psi_n\}_{n\geq 2}$ on $\Delta^{d\times d-1}$. Next, we use the following lemma, which states that any nondiagonal sequence of maximizers $\{\underline{\boldsymbol{\mu}}^{(n)}\}_n$ of ψ_n is such that $\sum_{r\neq s}\mu_{rs}^{(n)}$ grows proportionally to n.

Lemma 4.3. *For all $n \geq 2$, let*

$$\underline{\boldsymbol{\mu}}^{(n)} \in \arg\max\left\{\psi_n(\underline{\boldsymbol{\mu}}/n) \;:\; \underline{\boldsymbol{\mu}} \in \{0,\dots,n\}^{d\times d},\; 1 \leq \sum_{r\neq s}\mu_{rs} \leq n,\; \underline{\boldsymbol{\mu}}^\intercal\mathbf{1} = n\boldsymbol{\pi}\right\}.$$

It holds that

$$\liminf_{n\to\infty}\frac{\sum_{r\neq s}\mu_{rs}^{(n)}}{n} > 0.$$

By Lemma 4.3, which we prove at the end of the current argument, we can safely restrict the set of candidate maximizers to those $\underline{\boldsymbol{\mu}}$ such that $\sum_{r\neq s}\mu_{rs} \geq c_0 n$ for some fixed but small $c_0 > 0$. From here, and by a change of variables $\underline{\boldsymbol{\mu}} = n\underline{\boldsymbol{w}}$, mere pointwise convergence suffices to interchange lim inf and sup:

$$\begin{aligned}\liminf_{n\to\infty}\mathfrak{F}_n &\geq \liminf_{n\to\infty}\;\sup\left\{\psi_n(\underline{\boldsymbol{w}}) \;:\; \begin{array}{c}\underline{\boldsymbol{w}} \in \{i/n : 0\leq i\leq n\}^{d\times d},\; \underline{\boldsymbol{w}}^\intercal\mathbf{1} = \boldsymbol{\pi},\\ c_0 \leq \sum_{r\neq s} w_{rs} \leq 1\end{array}\right\}\\ &\geq \sup\left\{\psi(\underline{\boldsymbol{w}}) \;:\; \underline{\boldsymbol{w}} \in [0,1]^{d\times d}\cap\mathcal{F},\; c_0 \leq \sum_{r\neq s} w_{rs} \leq 1,\; \underline{\boldsymbol{w}}^\intercal\mathbf{1} = \boldsymbol{\pi}\right\}.\end{aligned} \tag{4.4}$$

Now we present a matching upper bound for $\limsup\mathfrak{F}_n$. For $\epsilon > 0$, let $G_n = (\{1,\dots,d\}, E_n)$ be defined such that $(r,s) \in E_n$ if and only if $w_{rs}^{(n)} \geq \epsilon$. Let $(G_l)_{l=1}^k$ denote the connected components of the graph G_n, $k = \mathsf{ncc}(G_n)$. Also, for $\underline{\boldsymbol{w}}$ an array for positive entries, let $\mathsf{ncc}^\epsilon(\underline{\boldsymbol{w}})$ denote the number of connected components of the graph $G(\underline{\boldsymbol{w}},\epsilon) = (V, E(\underline{\boldsymbol{w}},\epsilon))$, $V = \{1,\dots,d\}$, $E(\underline{\boldsymbol{w}},\epsilon) = \{(r,s) \;:\; r\neq s,\; w_{rs} > \epsilon\}$, and let

$$\vartheta^\epsilon(\underline{\boldsymbol{w}}) := \inf_{\underline{\boldsymbol{x}}}\{\vartheta(\underline{\boldsymbol{x}},\underline{\boldsymbol{w}}) : 0 \leq x_{rs} \leq \epsilon \;\forall (r,s) \notin E(\underline{\boldsymbol{w}},\epsilon)\}.$$

We will also write $\mathsf{ncc}(\underline{\boldsymbol{w}})$ for $\mathsf{ncc}^0(\underline{\boldsymbol{w}})$. Let $\underline{\boldsymbol{w}}^{(n)} = \underline{\boldsymbol{\mu}}^{(n)}/n$ for all $n \geq 2$, where $\underline{\boldsymbol{\mu}}^{(n)}$ is defined as in Lemma 4.3. By Theorem 3.1, we have for n sufficiently large

$$q(n\underline{\boldsymbol{w}}^{(n)}) \leq c_u(\epsilon,d,\alpha)P_{G_n}(n\underline{\boldsymbol{w}}^{(n)})^{-1/2}\exp -\vartheta^\epsilon(n\underline{\boldsymbol{w}}^{(n)}).$$

Since $w_{rs}^{(n)} \geq \epsilon$ of all the edges (r,s) of G_n^ϵ, $\prod_l T_{G_l}(\underline{\boldsymbol{w}}^{(n)})$ is bounded below by ϵ^d independently of n. Therefore, for n sufficiently large,

$$\begin{aligned}\psi_n(\underline{\boldsymbol{w}}^{(n)}) &= \frac{1}{n}\log\binom{n}{n\underline{\boldsymbol{w}}^{(n)}} + \gamma\frac{\log q(n\underline{\boldsymbol{w}}^{(n)})}{\log n}\\ &\leq \frac{1}{n}\log\binom{n}{n\underline{\boldsymbol{w}}^{(n)}} - \frac{\gamma}{2}(d - \mathsf{ncc}^{\epsilon}(\underline{\boldsymbol{w}}^{(n)})) - \frac{\gamma n}{\log n}\vartheta^{\epsilon}(\underline{\boldsymbol{w}}^{(n)})\\ &\quad + \mathcal{O}\left(\frac{\log c_u(\epsilon, d, \alpha) + d\log(1/\epsilon)}{\log n}\right)\\ &\leq \sup\Bigg\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - \mathsf{ncc}^{\epsilon}(\underline{\boldsymbol{w}})) - \frac{\gamma n}{\log n}\vartheta^{\epsilon}(\underline{\boldsymbol{w}}) \ :\\ &\qquad\qquad \underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ c_0 \leq \sum_{r\neq s} w_{rs} \leq 1,\ \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1} = \boldsymbol{\pi}\Bigg\}\\ &\quad + \mathcal{O}\left(\frac{\log c_u(\epsilon, d, \alpha) + d\log(1/\epsilon)}{\log n}\right),\end{aligned}$$

where the last inequality is obtained by Stirling's formula and taking a supremum over all $\underline{\boldsymbol{w}}$. By Lemma 3.2, $\vartheta^{\epsilon}(\underline{\boldsymbol{w}}) = 0$ if and only if $\boldsymbol{M}_G(\alpha\underline{\boldsymbol{w}}, \underline{\boldsymbol{x}}) \in \mathcal{F}$ for some $\underline{\boldsymbol{x}} \in [0,1]^{d\times d}$ such that $0 \leq x_{rs} \leq \epsilon$ for all $(r,s) \notin E$, $G = (V,E)$ being the graph whose edges are $(r,s) : w_{rs} \geq \epsilon$. This constrains the supremum to be achieved in the space of such $\underline{\boldsymbol{w}}$ for n sufficiently large. Moreover, this condition implies in particular that

$$\|\underline{\boldsymbol{w}}\mathbf{1} - \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1}\|_{\ell_\infty} \leq 2d\alpha^{-1}\epsilon,$$

where $\|\cdot\|_{\ell_\infty}$ is the ℓ_∞ norm of a vector in $\mathbb{R}^d$. Consequently, this yields the following upper bound as $n \to \infty$:

$$\limsup_{n\to\infty} \mathfrak{F}_n \leq \sup\left\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - \mathsf{ncc}^{\epsilon}(\underline{\boldsymbol{w}})) \ : \ \begin{array}{c}\underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ \|\underline{\boldsymbol{w}}\mathbf{1} - \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1}\|_{\ell_\infty} \leq 2d\alpha^{-1}\epsilon,\\ c_0 \leq \sum_{r\neq s} w_{rs} \leq 1,\ \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1} = \boldsymbol{\pi}\end{array}\right\} \tag{4.5}$$

for all $\epsilon > 0$. Next, we argue that as $\epsilon \to 0$, the right-hand side of the above inequality converges to

$$\sup\left\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - \mathsf{ncc}(\underline{\boldsymbol{w}})) \ : \ \underline{\boldsymbol{w}} \in [0,1]^{d\times d} \cap \mathcal{F},\ c_0 \leq \sum_{r\neq s} w_{rs} \leq 1,\ \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1} = \boldsymbol{\pi}\right\},$$

thereby establishing the existence of the limit $\lim \mathfrak{F}_n$ along with its precise value. Since the function $\epsilon \to \mathsf{ncc}^{\epsilon}(\underline{\boldsymbol{w}})$ is nondecreasing for any fixed $\underline{\boldsymbol{w}}$, the limit of the right-hand side of (4.5) as $\epsilon \to 0$ exists by monotone convergence. The limit can be decomposed as

$$\begin{aligned}&\lim_{\epsilon\to 0}\sup\left\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - \mathsf{ncc}^{\epsilon}(\underline{\boldsymbol{w}})) : \begin{array}{c}\underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ \|\underline{\boldsymbol{w}}\mathbf{1} - \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1}\|_{\ell_\infty} \leq 2d\alpha^{-1}\epsilon,\\ c_0 \leq \sum_{r\neq s} w_{rs} \leq 1,\ \underline{\boldsymbol{w}}^{\mathsf{T}}\mathbf{1} = \boldsymbol{\pi}\end{array}\right\}\\ &= \max_{1\leq k\leq d}\max_{\{V_l\}_{l=1}^k}\lim_{\epsilon\to 0}\ \sup\left\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d-k) \ : \ \text{such that (4.6) holds}\right\},\end{aligned}$$

$$\left\{\begin{array}{c} \underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ \|\underline{\boldsymbol{w}}\mathbf{1} - \underline{\boldsymbol{w}}^\intercal\mathbf{1}\|_{\ell_\infty} \le 2d\alpha^{-1}\epsilon, \\ w_{rs} \le \epsilon\ \forall (r,s) \in V_l \times V_{l'},\ l \ne l', \\ G_l(\underline{\boldsymbol{w}}) \text{ is connected } \forall l,\ c_0 \le \sum_{r\ne s} w_{rs} \le 1,\ \underline{\boldsymbol{w}}^\intercal\mathbf{1} = \boldsymbol{\pi}, \end{array}\right. \tag{4.6}$$

where $\{V_l\}_{l=1}^k$ ranges over partitions of the set $\{1,\dots,d\}$ with k nonempty subsets, and $G_l(\underline{\boldsymbol{w}}) = (V_l, \{(r,s) \in V_l \times V_l\ :\ w_{rs} > \epsilon\})$ for all $1 \le l \le k$. Letting $\epsilon < c_0$, the range of the outermost maximum becomes $1 \le k \le d-1$. By concavity of the entropy, the constraint that the graphs $G_l(\underline{\boldsymbol{w}})$ must be connected can be safely removed from the maximization problem without changing its maximum value, since it will be automatically satisfied. Thus, the innermost optimization problem is that of a continuous function on a closed and bounded domain that shrinks with ϵ. Its value is therefore a continuous function of ϵ. Hence, by sending ϵ to 0, in conjunction with the lower bound (4.4), we conclude that

$$\lim_{n\to\infty} \mathfrak{F}_n = \sup\left\{\psi(\underline{\boldsymbol{w}})\ :\ \underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ c_0 \le \sum_{r\ne s} w_{rs} \le 1,\ \underline{\boldsymbol{w}}\mathbf{1} = \underline{\boldsymbol{w}}^\intercal\mathbf{1} = \boldsymbol{\pi}\right\}. \tag{4.7}$$

As a final step, we make the above expression a bit more explicit. As argued previously, the supremum in (4.7) can be decomposed such that one first takes the maximum of $\psi(\underline{\boldsymbol{w}})$ over all $\underline{\boldsymbol{w}}$ such that $w_{rs} = 0$ for all $(r,s) \in V_l \times V_{l'},\ l \ne l'$, where $\{V_l\}_{1\le l\le k}$ is a fixed partition of $\{1,\dots,d\}$ into nonempty subsets, then maximizes over all such partitions, and then maximizes over all $1 \le k \le d-1$. The first optimization problem has a value

$$\sup\left\{H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d-k)\ :\ \begin{array}{c}\underline{\boldsymbol{w}} \in [0,1]^{d\times d},\ \underline{\boldsymbol{w}}\mathbf{1} = \underline{\boldsymbol{w}}^\intercal\mathbf{1} = \boldsymbol{\pi}, \\ w_{rs} = 0\ \forall (r,s) \in V_l \times V_{l'},\ l \ne l'\end{array}\right\},$$

where the constraint $c_0 \le \sum_{r\ne s} w_{rs} \le 1$ is not active for c_0 small enough and hence can be removed. Let $\underline{\boldsymbol{w}}$ be in the above constraint set. Then $H(\underline{\boldsymbol{w}}) = -\sum_{l=1}^k \sum_{(r,s)\in V_l\times V_l} w_{rs}\log w_{rs}$, and this is maximized at

$$w^*_{rs} = \begin{cases} (\pi_r\pi_s)/\sum_{r'\in V_l}\pi_{r'} & \text{if } (r,s) \in V_l \times V_l,\ l \in \{1,\dots,k\}, \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

with maximum value

$$\begin{aligned} H(\underline{\boldsymbol{w}}^*) &= 2H(\boldsymbol{\pi}) + \sum_{l=1}^k \left(\sum_{r\in V_l}\pi_r\right)\log\left(\sum_{r\in V_l}\pi_r\right) \\ &= 2H(\boldsymbol{\pi}) - H(\boldsymbol{X}\boldsymbol{\pi}), \end{aligned} \tag{4.9}$$

where $\boldsymbol{X} \in \{0,1\}^{k\times d}$, $X_{l,r} = 1$ if and only if $r \in V_l$. Note that $\mathcal{D}_k$ is the set of all such matrices (each one corresponding to a partition $\{V_l\}$ of $\{1,\dots,d\}$). Finally, by maximizing over all possible partitions and using (4.1), we get

$$\mathfrak{F}(\gamma) = \max_{1\le k\le d-1}\left\{H(\boldsymbol{\pi}) - \min_{\boldsymbol{X}\in\mathcal{D}_k} H(\boldsymbol{X}\boldsymbol{\pi}) - \frac{\gamma}{2}(d-k)\right\}.$$

This completes the proof of Proposition 4.1, except for the proof of Lemma 4.3, which we provide below.

Proof of Lemma 4.3. Let

$$\underline{\boldsymbol{\mu}}^{(n)} \in \arg\max \left\{ \psi_n(\underline{\boldsymbol{\mu}}/n) \ : \ \underline{\boldsymbol{\mu}} \in \{0,\dots,n\}^{d\times d},\ 1 \le \sum_{r\neq s} \mu_{rs},\ \underline{\boldsymbol{\mu}}^\intercal \mathbf{1} = n\boldsymbol{\pi} \right\}.$$

We show that

$$\liminf_{n\to\infty}\ n^{-1}\sum_{r\neq s}\mu_{rs}^{(n)} > 0.$$

Let us first show that

$$\frac{(\log n)^3}{n}\sum_{r\neq s}\mu_{rs}^{(n)} \longrightarrow \infty,$$

and then remove the logarithmic factor. We proceed by contradiction, by showing that if the above statement is not true, then the expected number of nonplanted solutions $\mathbb{E}[\mathcal{Z}-1]$ vanishes as $n\to\infty$ for any $\gamma > 0$, which contradicts our lower bound of Theorem 1.1. We have

$$\mathbb{E}\left[\mathcal{Z}-1\right] \le \binom{n}{n\boldsymbol{\pi}}^{-1}\cdot (n+1)^{d^2}\cdot \binom{n}{\underline{\boldsymbol{\mu}}^{(n)}}\cdot q_{\max}^{\gamma n/\log n},$$

with $q_{\max} = \sup\left\{q(\underline{\boldsymbol{\mu}}) \ : \ 1 \le \sum_{r\neq s}\mu_{rs},\ n\in\mathbb{N}\right\}$. It is not difficult to see that $q_{\max} < 1$: since $q(\underline{\boldsymbol{\mu}})$ decays to zero when any subset of nondiagonal entries of μ become large (Theorem 3.1), it is enough to check that $q(\underline{\boldsymbol{\mu}}) < 1$ for finitely many $\underline{\boldsymbol{\mu}}$'s (such that $1 \le \sum_{r\neq s}\mu_{rs}$) to certify that $q_{\max} < 1$, since the supremum must be achieved at a finite n. Additionally, notice that $q(\underline{\boldsymbol{\mu}}) = 1$ if and only if $\underline{\boldsymbol{\mu}}$ is a diagonal matrix. With this we conclude that $q_{\max} < 1$. Moreover,

$$\binom{n}{\underline{\boldsymbol{\mu}}^{(n)}} = \binom{n}{n\boldsymbol{\pi}}\prod_{r=1}^{d}\frac{(n\pi_r)!}{\prod_{s\neq r}\mu_{sr}!(n\pi_r - \sum_{s\neq r}\mu_{sr})!} \le \binom{n}{n\boldsymbol{\pi}}\prod_{r=1}^{d}(n\pi_r)^{\sum_{s\neq r}\mu_{sr}}.$$

If $\sum_{r\neq s}\mu_{rs}^{(n)} \le Cn/(\log n)^3$ for some constant $C > 0$, then

$$\mathbb{E}\left[\mathcal{Z}-1\right] \le (n+1)^{d^2}\cdot n^{Cn/(\log n)^3}\cdot q_{\max}^{\gamma n/\log n} \xrightarrow[n\to\infty]{} 0$$

for all $\gamma > 0$, and this contradicts the fact that below γ_{low} there are exponentially many distinct satisfying assignments.

Now let us assume that $\frac{(\log n)^3}{n}\sum_{r\neq s}\mu_{rs}^{(n)} \to \infty$ but $\liminf n^{-1}\sum_{r\neq s}\mu_{rs}^{(n)} = 0$. We proceed by contradiction once more and construct a sequence of points that have a higher objective value than $\underline{\boldsymbol{\mu}}^{(n)}$. Instead of working with convergent subsequences, we may as well assume that $\{\underline{\boldsymbol{\mu}}^{(n)}\}$ is convergent. Let

$$E_n = \left\{(r,s) \ : \ r\neq s,\ \mu_{rs}^{(n)} > \epsilon\sum_{r\neq s}\mu_{rs}^{(n)}\right\}$$

and

$$E_\infty = \left\{(r,s) \ : \ r\neq s,\ \liminf_{n\to\infty}\frac{\mu_{rs}^{(n)}}{\sum_{r\neq s}\mu_{rs}^{(n)}} > 0\right\}$$

for all n and some $\epsilon > 0$ sufficiently small. Let $k_n = \mathsf{ncc}(G_n)$ be the number of connected components of the graph $G_n = (\{1, \dots, d\}, E_n)$, and, similarly, let $k_\infty = \mathsf{ncc}(G_\infty)$, with $G_\infty = (\{1, \dots, d\}, E_\infty)$. Observe that E_∞ and E_n are both nonempty sets, and hence $k_\infty, k_n \le d-1$ for all n.

Now we consider an arbitrary partition of the set of vertices $\{1, \dots, d\}$ into k_∞ subsets $\{V_l\}_{l=1}^{k_\infty}$ and let G be the graph on d vertices with edge set $\cup_{l=1}^{k_\infty} V_l \times V_l$; i.e., G is the union of k_∞ *complete* connected components. Finally, let $\underline{\boldsymbol{v}}^{(n)} := n\underline{\boldsymbol{w}}$ for all n, with

$$w_{rs} = \begin{cases} (\pi_r \pi_s)/\sum_{r' \in V_l} \pi_{r'} & \text{if } (r,s) \in V_l \times V_l,\ l \in \{1, \dots, k_\infty\}, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that this construction provides one of the candidate maximizers of the annealed free energy (see (4.8)). Observe that $\underline{\boldsymbol{v}}^{(n)}$ satisfies all the constraints satisfied by $\underline{\boldsymbol{\mu}}^{(n)}$ and, additionally, $\underline{\boldsymbol{v}}^{(n)} \in \mathcal{F}$. Therefore, by Corollary 3.3, we have

$$\psi_n(\underline{\boldsymbol{v}}^{(n)}/n) = H(\underline{\boldsymbol{w}}) - \frac{\gamma}{2}(d - k_\infty) + o_n(1).$$

Recall that the function ψ_n is defined in (4.2). On the other hand, to study the asymptotics of $\psi_n(\underline{\boldsymbol{\mu}}^{(n)}/n)$, we apply Theorem 3.1 with n replaced by $\sum_{r \neq s} \mu_{rs}^{(n)}$ (which grows to infinity), and we get

$$\psi_n(\underline{\boldsymbol{\mu}}^{(n)}/n) \le H(\boldsymbol{\pi}) - \frac{\gamma}{2}(d - k_n)\left(1 - 3\frac{\log\log n}{\log n}\right) - \frac{\vartheta_u(\underline{\boldsymbol{\mu}}^{(n)})}{\log n} + o_n(1).$$

The term in the right-hand side follows from Stirling's formula and the fact that $\mu_{rs}^{(n)}/n \to 0$ for all $r \neq s$. The second term follows from the fact that

$$P_{G_n}(\underline{\boldsymbol{\mu}}^{(n)}) \ge \left(\epsilon \sum_{r \neq s} \mu_{rs}^{(n)}\right)^{d-k_n} \gg \left(\frac{n}{(\log n)^3}\right)^{d-k_n}.$$

Next, we argue based on these estimates that $\psi_n(\underline{\boldsymbol{v}}^{(n)}/n) > \psi_n(\underline{\boldsymbol{\mu}}^{(n)}/n)$ for all n large enough. First, the term involving ϑ_u in the upper bound on $\psi_n(\underline{\boldsymbol{\mu}}^{(n)}/n)$ can be dropped, since it is always nonnegative. By direct computation (we already showed this in (4.9)), we have

$$H(\underline{\boldsymbol{w}}) - H(\boldsymbol{\pi}) = H(\boldsymbol{\pi}) - H(\boldsymbol{p}),$$

with $\boldsymbol{p} \in \Delta^{k_\infty - 1}$ with $p_l = \sum_{r \in V_l} \pi_r$ for all $1 \le l \le k_\infty$. We show that the right-hand side of this equality is strictly positive:

$$\begin{aligned}
H(\boldsymbol{\pi}) - H(\boldsymbol{p}) &= -\sum_{r=1}^{d} \pi_r \log \pi_r + \sum_{l=1}^{k_\infty} \left(\sum_{r\in V_l} \pi_r\right) \log \left(\sum_{r\in V_l} \pi_r\right) \\
&= -\sum_{l=1}^{k_\infty} \sum_{r\in V_l} \pi_r \log\left(\frac{\pi_r}{p_l}\right) \\
&= -\sum_{l=1}^{k_\infty} p_l \sum_{r\in V_l} \frac{\pi_r}{p_l} \log\left(\frac{\pi_r}{p_l}\right) \\
&\geq -\sum_{l=1}^{k_\infty} p_l \log\left(\frac{\sum_{r\in V_l}\pi_r^2}{p_l^2}\right) \\
&\geq 0.
\end{aligned}$$

We used Jensen's inequality on the concave function $x \mapsto \log x$ and the fact that $\sum_{r\in V_l} \pi_r^2 \leq p_l \sum_{r\in V_l} \pi_r = p_l^2$ for all l. Moreover, since all coordinates of $\boldsymbol{\pi}$ are strictly positive, equality holds if and only if $\pi_r = p_l$ for all l and $r \in V_l$, which implies that the partition must be trivial; i.e., $k_\infty = d$. Recall that this does not happen, since E_∞ is nonempty.

On the other hand, by setting ϵ sufficiently small (smaller than all the limits in the definition of E_∞), any edge in E_∞ will eventually (and permanently from then on) be in E_n. Therefore, the number of connected components of G_n does not exceed that of G_∞: $k_n \leq k_\infty$ for n sufficiently large. We conclude that $\psi_n(\underline{\boldsymbol{v}}^{(n)}/n) > \psi_n(\underline{\boldsymbol{\mu}}^{(n)}/n)$ for all n large enough. Therefore, $\underline{\boldsymbol{\mu}}^{(n)}$ is not always a maximizer of ψ_n, and this leads to a contradiction. ■

5. Proof of Theorem 3.1.

Our proof is based on the method of Laplace from asymptotic analysis: when the entries of $\underline{\boldsymbol{\mu}}$ are large, the sum defining $q(\underline{\boldsymbol{\mu}})$ is dominated by its largest term corrected by a subexponential term which is represented by a Gaussian integral (see, e.g., [21] for the univariate case). Since we are in a multivariate situation, the asymptotics of q depend on which subset of the entries of $\underline{\boldsymbol{\mu}}$ is large. Our approach is inspired by [3]. We recall that for $\underline{\boldsymbol{\mu}} \in \mathbb{Z}_+^{d\times d}$,

$$q(\underline{\boldsymbol{\mu}}) = \sum_{\substack{\underline{\boldsymbol{\nu}}\in\mathbb{Z}_+^{d\times d}\cap\mathcal{F} \\ 0\leq\nu_{rs}\leq\mu_{rs}}} \left(\prod_{r,s=1}^{d} \binom{\mu_{rs}}{\nu_{rs}} \alpha^{\nu_{rs}}(1-\alpha)^{\mu_{rs}-\nu_{rs}}\right).$$

Let $G = (V, E)$, with $V = \{1, \ldots, d\}$ and $E \subseteq \{(r, s) \in V^2 \ : \ r \neq s\}$. The graph G will be used to store information about which entries of $\underline{\boldsymbol{\mu}}$ are going to infinity linearly in n and which entries are not. We can split the sum defining q into a double sum, one involving the large terms (A in subsequent notation), and the rest

$$q(\underline{\boldsymbol{\mu}}) = \sum_{\substack{0\leq\nu'_{rs}\leq\mu_{rs} \\ (r,s)\notin E}} \prod_{(r,s)\notin E} \binom{\mu_{rs}}{\nu'_{rs}} \alpha^{\nu'_{rs}}(1-\alpha)^{\mu_{rs}-\nu'_{rs}} A(\underline{\boldsymbol{\nu}}', \underline{\boldsymbol{\mu}}),$$

with

$$A(\underline{\boldsymbol{\nu}}', \underline{\boldsymbol{\mu}}) = \sum_{\substack{0 \le \nu_{rs} \le \mu_{rs} \\ (r,s) \in E}} \prod_{(r,s)\in E} \binom{\mu_{rs}}{\nu_{rs}} \alpha^{\nu_{rs}} (1-\alpha)^{\mu_{rs}-\nu_{rs}} \mathbb{1}\left\{ \boldsymbol{M}_G(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\nu}}') \in \mathcal{F} \right\},$$

where for two $d \times d$ matrices $\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}$, $\boldsymbol{M}_G(\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}})$ is the $d \times d$ matrix with entries a_{rs} if $(r, s) \in E$ and b_{rs} otherwise. The quantity A will be approximated using the Laplace method. Recall from the expressions (3.2) and (3.3) that

$$\varphi_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{x}}) = \sum_{(r,s)\in E} \mu_{rs} D(x_{rs} \parallel \alpha)$$

and

$$\vartheta(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}}) = \min_{\substack{\underline{\boldsymbol{x}} \in [0,1]^{d\times d} \\ \boldsymbol{M}_G(\underline{\boldsymbol{x}} \odot \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\nu}}) \in \mathcal{F}}} \varphi_{\underline{\boldsymbol{\mu}}}(\underline{\boldsymbol{x}}).$$

Let $\underline{\boldsymbol{x}}^*(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}})$ be the optimal solution of the above optimization problem.

Before stating our asymptotic approximation result for A, we state an important lemma on the boundedness of the entries of $\underline{\boldsymbol{x}}^*(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}})$, where the bounds depend only on ϵ and α.

Lemma 5.1 (proved in section SM2 of the supplementary materials, linked from the main article webpage). *Let G be fixed as above, and let $\alpha \in (0, 1)$ and $\epsilon \in (0, 1)$. There exist two constants $0 < c_l \le c_u < 1$ depending only on d, α, and ϵ such that the following is true: For all integers $n \ge 1$, and $\underline{\boldsymbol{\mu}} \in \{0, \dots, n\}^{d\times d}$ such that $\mu_{rs} \ge \epsilon n$ if and only if $(r, s) \in E$, and all $\underline{\boldsymbol{\nu}}' \in \{0, \dots, n\}^{\bar{E}}$ such that $0 \le \nu'_{rs} \le \mu_{rs}$ for all $(r, s) \notin E$, we have*

$$c_l \le \min_{(r,s)\in E} x^*_{rs} \le \max_{(r,s)\in E} x^*_{rs} \le c_u.$$

Therefore, the entries of $\underline{\boldsymbol{x}}^*$ can effectively be treated as constants throughout the rest of the proof. Now we state our asymptotic estimate for A.

Proposition 5.2 (proved in section SM1 of the supplementary materials). *Let G be fixed as above, and let $\epsilon > 0$. For all n sufficiently large, all $\underline{\boldsymbol{\mu}} \in \{0, \dots, n\}^{d\times d}$ with $\mu_{rs} \ge \epsilon n$ if and only if $(r, s) \in E$, and all $\underline{\boldsymbol{\nu}}' \in \{0, \dots, n\}^{\bar{E}}$ such that $0 \le \nu'_{rs} \le \mu_{rs}$ for all $(r, s) \notin E$, we have*

$$A(\underline{\boldsymbol{\nu}}', \underline{\boldsymbol{\mu}}) \quad \asymp_{G,d,\epsilon,\alpha} \frac{e^{-\vartheta(\underline{\boldsymbol{\nu}}', \underline{\boldsymbol{\mu}})}}{P_G(\underline{\boldsymbol{\mu}})^{1/2}}.$$

Here the symbol "$\asymp_{G,d,\epsilon,\alpha}$" means that the ratio is upper- and lower-bounded by constants depending only on G, d, ϵ, and α.

By the above proposition, we have

$$q(\underline{\boldsymbol{\mu}}) \quad \asymp_{G,d,\epsilon,\alpha} \sum_{\substack{\underline{\boldsymbol{\nu}} \in \mathbb{Z}_+^{\bar{E}} \\ 0 \le \nu_{rs} \le \mu_{rs}}} \left(\prod_{(r,s)\notin E} \binom{\mu_{rs}}{\nu_{rs}} \alpha^{\nu_{rs}} (1-\alpha)^{\mu_{rs}-\nu_{rs}} \right) \frac{e^{-\vartheta(\underline{\boldsymbol{\nu}}, \underline{\boldsymbol{\mu}})}}{P_G(\underline{\boldsymbol{\mu}})^{1/2}}.$$

The estimate above (ignoring the term $P_G(\underline{\boldsymbol{\mu}})$) can be interpreted as the expected value of the function $e^{-\vartheta(\underline{\boldsymbol{\nu}},\underline{\boldsymbol{\mu}})}$ under the law of the random variable $\underline{\boldsymbol{\nu}}$ where each entry ν_{rs} for $(r,s) \notin E$ is independently binomial with parameters α and μ_{rs}. From here, the bounds claimed in Theorem 3.1 follow immediately.

6. Proof of Proposition 3.4.

We first reduce the proof to the case where $G = K_d$ by a limiting argument. Let $G = (V, E)$ be a graph on d vertices. If G is not connected, then the constraints defining the space $\mathcal{F}(G)$ decouple across the connected components of G and so does the integrand $\exp -\frac{1}{2}\sum_{(r,s)\in E} x_{rs}^2/w_{rs}$; therefore, the Gaussian integral factors across the connected components of G. Hence, we may assume that G is connected. Now, if

$$\int_{\mathcal{F}} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/w_{rs}} \,\mathrm{d}\underline{\boldsymbol{x}} = (2\pi)^{((d-1)^2+d)/2} \left(\frac{\prod_{r,s} w_{rs}}{T(\underline{\boldsymbol{w}})}\right)^{1/2}$$

for all $\underline{\boldsymbol{w}} \in \mathbb{R}_+^{d\times d}$, where $T = T_{K_d}$ (3.4), then taking a limit $w_{rs} \to 0$ for all $(r,s) \notin E$, we get

$$\frac{1}{\left(\prod_{(r,s)\notin E} w_{rs}\right)^{1/2}} \int_{\mathcal{F}} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/w_{rs}} \,\mathrm{d}\underline{\boldsymbol{x}} \longrightarrow c(G) \int_{\mathcal{F}(G)} e^{-\frac{1}{2}\sum_{(r,s)\in E} x_{rs}^2/w_{rs}} \,\mathrm{d}\underline{\boldsymbol{x}},$$

where $c(G) > 0$ is a constant that only depends on G. On the other hand,

$$T(\underline{\boldsymbol{w}}) \longrightarrow \frac{\mathsf{nst}(G)}{2^{d-1}d^{d-2}}\, T_G(\underline{\boldsymbol{w}}),$$

where the denominator is a formula for $\mathsf{nst}(T_{K_d})$ (Cayley's formula). Therefore,

$$c(G) \int_{\mathcal{F}(G)} e^{-\frac{1}{2}\sum_{(r,s)\in E} x_{rs}^2/w_{rs}} \,\mathrm{d}\underline{\boldsymbol{x}} = (2\pi)^{((d-1)^2+d)/2} \left(\frac{2^{d-1}d^{d-2}}{\mathsf{nst}(G)} \frac{\prod_{(r,s)\in E} w_{rs}}{T_G(\underline{\boldsymbol{w}})}\right)^{1/2}.$$

Now we set $w_{rs} = 1$ for all $(r,s) \in E$ to clear out the constants. Since $\int_{\mathcal{F}(G)} e^{-\frac{1}{2}\sum_{(r,s)\in E} x_{rs}^2} \,\mathrm{d}\underline{\boldsymbol{x}} = (2\pi)^{\dim(\mathcal{F}(G))/2}$, we get

$$\int_{\mathcal{F}(G)} e^{-\frac{1}{2}\sum_{(r,s)\in E} x_{rs}^2/w_{rs}} \,\mathrm{d}\underline{\boldsymbol{x}} = (2\pi)^{\dim(\mathcal{F}(G))/2} \left(\frac{\prod_{(r,s)\in E} w_{rs}}{T_G(\underline{\boldsymbol{w}})}\right)^{1/2}.$$

Now it remains to prove the proposition for the complete graph. The approach, which was suggested to us by Andrea Sportiello, relies on an interpolation argument that involves expressing the Gaussian integral over $\mathcal{F}$ as the *limit* of another parameterized Gaussian integral, when the parameter tends to zero. This latter integral can, on the other hand, be written in closed form, by relating it to the characteristic polynomial of a Laplacian matrix. Then the Principal Minors Matrix-Tree theorem is invoked to finish the argument. For $\delta > 0$, let

$$I(\delta) = \frac{1}{(2\pi\delta^2)^{(d-1)/2}} \int_{\mathbb{R}^{d\times d}} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/w_{rs}}\, e^{-\frac{1}{2\delta^2}\|(\underline{\boldsymbol{x}}-\underline{\boldsymbol{x}}^\intercal)\mathbf{1}\|_{\ell_2}^2} \,\mathrm{d}\underline{\boldsymbol{x}}.$$

The additional Gaussian term in $I(\delta)$ gradually concentrates the mass of the integral on $\mathcal{F}$ as δ becomes small, and we have the following limiting statement.

Lemma 6.1. *We have*

$$\lim_{\delta\to 0} I(\delta) = c_d \int_{\mathcal{F}} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/2w_{rs}}\ \mathrm{d}\underline{\boldsymbol{x}},$$

with

$$c_d = \frac{1}{(2\pi)^{(d-1)/2}} \int_{\mathcal{F}^\perp} e^{-2\|\underline{\boldsymbol{z}}\mathbf{1}\|_{\ell_2}^2}\ \mathrm{d}\underline{\boldsymbol{z}} = (2d)^{-(d-1)/2}.$$

On the other hand, a straightforward computation allows us to write $I(\delta)$ in closed form.

Lemma 6.2. *Let $G = (V, E)$ be the complete graph ($V = \{1, \dots, d\}$, $E = \{(r, s) \in V \times V, r \neq s\}$) where the edges are weighted by the array $\underline{\boldsymbol{w}} \in \mathbb{R}_+^{d\times d}$. Let $\boldsymbol{L}(\underline{\boldsymbol{w}}) \in \mathbb{R}^{d\times d}$ be the Laplacian matrix of G. For all $\delta > 0$, it holds that*

$$I(\delta) = (2\pi)^{((d-1)^2+d)/2} \left(\prod_{r,s} w_{rs}\right)^{1/2} \frac{\delta}{\mathrm{Det}\left(\delta^2\boldsymbol{I} + \boldsymbol{L}(\underline{\boldsymbol{w}})\right)^{1/2}}.$$

Now, by the Principal Minors Matrix-Tree theorem (see, e.g., [12]), the characteristic polynomial of the Laplacian matrix of a graph admits the following expansion:

$$\mathrm{Det}\left(x\boldsymbol{I} + \boldsymbol{L}(\underline{\boldsymbol{w}})\right) = \sum_F x^{|\mathrm{roots}(F)|} \prod_{(r,s)\in F} w_{rs},$$

where the sum is over all rooted spanning forests F of the graph. We finish the argument by taking a limit in δ:

$$\delta^{2(d-1)}\mathrm{Det}\left(\boldsymbol{I} + \delta^{-2}\boldsymbol{L}(\underline{\boldsymbol{w}})\right) = \delta^{-2}\mathrm{Det}\left(\delta^2\boldsymbol{I} + \boldsymbol{L}(\underline{\boldsymbol{w}})\right) \xrightarrow[\delta\to 0]{} d\sum_T \prod_{(r,s)\in T} w_{rs}.$$

The latter is $(2d)^{d-1}T(\underline{\boldsymbol{w}})$, since the above limit singles out the rooted spanning forests with exactly one root—i.e., rooted spanning trees—from the characteristic polynomial, and there are d ways of choosing the root of a spanning tree. This exactly leads to the desired identity

$$\int_{\mathcal{F}} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/2w_{rs}}\ \mathrm{d}\underline{\boldsymbol{x}} = (2\pi)^{((d-1)^2+d)/2} \left(\frac{\prod_{r,s} w_{rs}}{T(\underline{\boldsymbol{w}})}\right)^{1/2}.$$

Proof of Lemma 6.1. We decompose $\mathbb{R}^{d\times d}$ into the direct sum $\mathcal{F} \oplus \mathcal{F}^\perp$. It is easy to see that $\mathcal{F}^\perp = \{\underline{\boldsymbol{z}} = \boldsymbol{\lambda}\mathbf{1}^\intercal - \mathbf{1}\boldsymbol{\lambda}^\intercal,\ \boldsymbol{\lambda} \in \mathbb{R}^d\}$, which is a $(d-1)$-dimensional space. For $\underline{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$, let $\underline{\boldsymbol{y}} \in \mathbb{R}^{d\times d}$ be its orthogonal projection on $\mathcal{F}$, and let $\underline{\boldsymbol{z}} = \underline{\boldsymbol{x}} - \underline{\boldsymbol{y}}$. Therefore, $(\underline{\boldsymbol{x}} - \underline{\boldsymbol{x}}^\intercal)\mathbf{1} = (\underline{\boldsymbol{z}} - \underline{\boldsymbol{z}}^\intercal)\mathbf{1} = 2\underline{\boldsymbol{z}}\mathbf{1} = 2(d\boldsymbol{\lambda} - (\mathbf{1}^\intercal\boldsymbol{\lambda})\mathbf{1})$. For $\delta > 0$, we have

$$I(\delta) = \frac{1}{(2\pi\delta^2)^{(d-1)/2}} \int_{F\times F^\perp} e^{-\frac{1}{2}\sum_{r,s}(y_{rs}+z_{rs})^2/w_{rs}}\ e^{-\frac{2}{\delta^2}\|\underline{\boldsymbol{z}}\mathbf{1}\|_{\ell_2}^2}\ \mathrm{d}\underline{\boldsymbol{y}}\mathrm{d}\underline{\boldsymbol{z}}.$$

We make the change of variables $\underline{\boldsymbol{z}}' = \underline{\boldsymbol{z}}/\delta$:

$$I(\delta) = \frac{1}{(2\pi)^{(d-1)/2}} \int_{\mathcal{F}\times\mathcal{F}^\perp} e^{-\frac{1}{2}\sum_{r,s}(y_{rs}+\delta z'_{rs})^2/w_{rs}}\ e^{-2\|\underline{\boldsymbol{z}}'\mathbf{1}\|_{\ell_2}^2}\ \mathrm{d}\underline{\boldsymbol{y}}\mathrm{d}\underline{\boldsymbol{z}}'.$$

By dominated convergence,

$$\begin{aligned}\lim_{\delta\to 0} I(\delta) &= \frac{1}{(2\pi)^{(d-1)/2}} \int_{\mathcal{F}\times\mathcal{F}^\perp} e^{-\frac{1}{2}\sum_{r,s} y_{rs}^2/w_{rs}}\; e^{-2\|\underline{\boldsymbol{z}}\mathbf{1}\|_{\ell_2}^2}\, \mathrm{d}\underline{\boldsymbol{y}}\mathrm{d}\underline{\boldsymbol{z}}\\ &= \frac{1}{(2\pi)^{(d-1)/2}} \int_{\mathcal{F}} e^{-\frac{1}{2}\sum_{r,s} y_{rs}^2/w_{rs}} \mathrm{d}\underline{\boldsymbol{y}} \int_{\mathcal{F}^\perp} e^{-2\|\underline{\boldsymbol{z}}\mathbf{1}\|_{\ell_2}^2} \mathrm{d}\underline{\boldsymbol{z}}.\end{aligned}$$

Moreover,

$$\int_{\mathcal{F}^\perp} e^{-2\|\underline{\boldsymbol{z}}\mathbf{1}\|_{\ell_2}^2} \mathrm{d}\underline{\boldsymbol{z}} = (2d)^{(d-1)/2} \int_{\{\boldsymbol{\lambda}\in\mathbb{R}^d, \mathbf{1}^\intercal\boldsymbol{\lambda}=0\}} e^{-2d^2\|\boldsymbol{\lambda}\|_{\ell_2}^2}\, \mathrm{d}\boldsymbol{\lambda} = (2\pi)^{(d-1)/2}(2d)^{-(d-1)/2},$$

where the prefactor in the first equality comes from the fact that $\|\underline{\boldsymbol{z}}\|_F^2 = 2d\,\|\boldsymbol{\lambda}\|_{\ell_2}^2$ for $\underline{\boldsymbol{z}} = \boldsymbol{\lambda}\mathbf{1}^\intercal - \mathbf{1}\boldsymbol{\lambda}^\intercal,\ \boldsymbol{\lambda}\in\mathbb{R}^d,\ \mathbf{1}^\intercal\boldsymbol{\lambda} = 0$. ■

Proof of Lemma 6.2. Let $\delta > 0$. We linearize the quadratic term $\|(\underline{\boldsymbol{x}} - \underline{\boldsymbol{x}}^\intercal)\mathbf{1}\|_{\ell_2}^2$ in $I(\delta)$ by writing the corresponding Gaussian as the Fourier transform of another Gaussian: for all $\underline{\boldsymbol{x}} \in \mathbb{R}^{d\times d}$,

$$e^{-\frac{1}{2\delta^2}\|(\underline{\boldsymbol{x}}-\underline{\boldsymbol{x}}^\intercal)\mathbf{1}\|_{\ell_2}^2} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\mathrm{i}\delta^{-1}\boldsymbol{y}^\intercal(\underline{\boldsymbol{x}}-\underline{\boldsymbol{x}}^\intercal)\mathbf{1}-\frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2}\, \mathrm{d}\boldsymbol{y},$$

where $\mathrm{i}^2 = -1$. Then

$$I(\delta) = \frac{1}{(2\pi\delta^2)^{(d-1)/2}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^{d\times d}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}\sum_{rs} x_{rs}^2/w_{rs}}\; e^{-\mathrm{i}\delta^{-1}\boldsymbol{y}^\intercal(\underline{\boldsymbol{x}}-\underline{\boldsymbol{x}}^\intercal)\mathbf{1}-\frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2}\, \mathrm{d}\underline{\boldsymbol{x}}\mathrm{d}\boldsymbol{y}.$$

We complete the square involving x_{rs} in the exponentiated expression

$$-\frac{1}{2}\sum_{r,s}\frac{x_{rs}^2}{w_{rs}} - \mathrm{i}\delta^{-1}\boldsymbol{y}^\intercal(\underline{\boldsymbol{x}} - \underline{\boldsymbol{x}}^\intercal)\mathbf{1} = -\frac{1}{2}\sum_{r,s}\frac{\left(\left(x_{rs} + \mathrm{i}\frac{w_{rs}}{\delta}(y_r - y_s)\right)^2 + \frac{w_{rs}^2}{\delta^2}(y_r - y_s)^2\right)}{w_{rs}}.$$

Then, by Fubini's theorem,

$$\begin{aligned} I(\delta) = \frac{1}{(2\pi\delta^2)^{(d-1)/2}} \frac{1}{(2\pi)^{d/2}} &\int_{\mathbb{R}^d} e^{-\frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2 - \frac{1}{2}\sum_{rs}\frac{w_{rs}}{\delta^2}(y_r-y_s)^2}\\ &\times\left(\int_{\mathbb{R}^{d\times d}} e^{-\frac{1}{2}\sum_{rs}\frac{1}{w_{rs}}\left(x_{rs}+\mathrm{i}\frac{w_{rs}}{\delta}(y_r-y_s)\right)^2} \mathrm{d}\underline{\boldsymbol{x}}\right)\mathrm{d}\boldsymbol{y}.\end{aligned}$$

The inner integral evaluates to $\left(\prod_{r,s} 2\pi w_{rs}\right)^{1/2}$. Hence

$$\begin{aligned} I(\delta) &= \frac{(2\pi)^{(d-1)^2/2}}{\delta^{d-1}}\left(\prod_{r,s} w_{rs}\right)^{1/2} \int_{\mathbb{R}^d} e^{-\frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2 - \frac{1}{2}\sum_{r,s}\frac{w_{rs}}{\delta^2}(y_r-y_s)^2}\, \mathrm{d}\boldsymbol{y}\\ &= \frac{(2\pi)^{((d-1)^2+d)/2}}{\delta^{d-1}}\left(\prod_{r,s} w_{rs}\right)^{1/2} \mathrm{Det}\left(\boldsymbol{I} + \delta^{-2}\boldsymbol{L}(\underline{\boldsymbol{w}})\right)^{-1/2},\end{aligned}$$

where $\boldsymbol{L}(\underline{\boldsymbol{w}}) \in \mathbb{R}^{d\times d}$ is the Laplacian matrix of the weighted graph G. ■

7. Discussion. Our main result, Theorem 1.1, leaves a gap of essentially a factor of two between γ_{low} and γ_{up}. This is a limitation of the methods employed. In particular, it may seem plausible that the upper bound is loose due to a possible lack of concentration of the random variable $\mathcal{Z}$ about its mean, and this translates to the possibility of existence of a nontrivial interval inside $[\gamma_{\text{low}}, \gamma_{\text{up}}]$ where $\mathcal{Z}$ is typically close to 1 while its expectation is exponentially large. This is a standard issue in the use of the first (or second) moment method encountered in many random CSPs. Surprisingly enough—and as mentioned below Theorem 1.1—this is not the case in the HQP, as it was recently shown [37] that γ_{up} is the sharp threshold. Therefore, the first moment method does identify the phase transition in this problem.

Beyond our setting, the "sparse" regime where the sets S_a are of constant size k (exactly or on average) could also be of interest. Here the relevant scaling is one where m is proportional to n. The lower bound argument could be easily extended and yields a bound of $\frac{H(\boldsymbol{\pi})}{(d-1)\log k}$. As for the upper bound, one could in principle follow the same first moment strategy, but our analysis breaks in a quite serious fashion, in that none of our asymptotic estimates hold true in this regime.

Acknowledgments. We benefited from insightful conversations with many people. We thank Cris Moore for bringing the work of Achlioptas and Naor [3] to our attention. We thank Ross Boczar for his input using Mathematica in the trial-and-error process that led to the discovery of Proposition 3.4. We thank Andrea Sportiello for suggesting the interpolation method that is used in our proof of Proposition 3.4. We thank Nikhil Srivastava for bringing the survey [7] to our attention. Part of this work was performed during the spring of 2016 when FK and LZ were visiting the Simons Institute for the Theory of Computing at UC Berkeley.

REFERENCES

[1] D. Achlioptas and A. Coja-Oghlan, *Algorithmic barriers from phase transitions*, in Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, 2008, pp. 793–802.
[2] D. Achlioptas and C. Moore, *The chromatic number of random regular graphs*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, Berlin, Heidelberg, 2004, pp. 219–228.
[3] D. Achlioptas and A. Naor, *The two possible values of the chromatic number of a random graph*, Ann. of Math. (2), 162 (2005), pp. 1335–1351.
[4] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, *Information-theoretic thresholds for community detection in sparse networks*, in Proceedings of the 49th Annual Conference on Learning Theory, 2016, pp. 383–416.
[5] V. Bapst, A. Coja-Oghlan, S. Hetterich, F. Raßmann, and D. Vilenchik, *The condensation phase transition in random graph coloring*, Comm. Math. Phys., 341 (2016), pp. 543–606.
[6] M. Bayati, M. Lelarge, and A. Montanari, *Universality in polytope phase transitions and message passing algorithms*, Ann. Appl. Probab., 25 (2015), pp. 753–822.
[7] N. Biggs, *Algebraic potential theory on graphs*, Bull. London Math. Soc., 29 (1997), pp. 641–682.
[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[9] E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.
[10] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[11] E. J. Candés and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.

[12] S. Chaiken, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Algebraic Discrete Methods, 3 (1982), pp. 319–329, https://doi.org/10.1137/0603033.

[13] A. Coja-Oghlan, *Random Constraint Satisfaction Problems*, preprint, https://arxiv.org/abs/0911.2322, 2009.

[14] A. Coja-Oghlan, C. Efthymiou, and S. Hetterich, *On the chromatic number of random regular graphs*, J. Combin. Theory Ser. B, 116 (2016), pp. 367–439.

[15] A. Coja-Oghlan and A. Frieze, *Analyzing* `Walksat` *on random formulas*, SIAM J. Comput., 43 (2014), pp. 1456–1485, https://doi.org/10.1137/12090191X.

[16] A. Coja-Oghlan, A. Haqshenas, and S. Hetterich, `Walksat` *Stalls Well Below the Satisfiability Threshold*, preprint, https://arxiv.org/abs/1608.00346, 2016.

[17] A. Coja-Oghlan, E. Mossel, and D. Vilenchik, *A spectral approach to analysing belief propagation for* 3*-colouring*, Combin. Probab. Comput., 18 (2009), pp. 881–912.

[18] A. Coja-Oghlan and W. Perkins, *Belief propagation on replica symmetric random factor graph models*, Ann. Inst. Henri Poincaré D, 5 (2018), pp. 211–249.

[19] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, Cambridge, UK, 2011.

[20] V. Dani, C. Moore, and A. Olson, *Tight bounds on the threshold for permuted k-colorability*, in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, Berlin, Heidelberg, 2012, pp. 505–516.

[21] N. G. De Bruijn, *Asymptotic Methods in Analysis*, Dover, New York, 1970.

[22] J. Ding, A. Sly, and N. Sun, *Proof of the satisfiability conjecture for large k*, in Proceedings of the 47th Annual ACM Symposium on Theory of Computing, 2015, pp. 59–68.

[23] J. Ding, A. Sly, and N. Sun, *Satisfiability threshold for random regular* NAE-SAT, Comm. Math. Phys., 341 (2016), pp. 435–489.

[24] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[25] D. L. Donoho, *For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution*, Comm. Pure Appl. Math., 59 (2006), pp. 797–829.

[26] D. L. Donoho, A. Javanmard, and A. Montanari, *Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing*, IEEE Trans. Inform. Theory, 59 (2013), pp. 7434–7464.

[27] D. Du and F. Hwang, *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*, Ser. Appl. Math. 18, World Scientific, Hackensack, NJ, 2006.

[28] A. El Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. I. Jordan, *Decoding from pooled data: Phase transitions of message passing*, in Proceedings of the 2017 IEEE International Symposium on Information Theory, 2017, pp. 2780–2784.

[29] M. Fazel, *Matrix Rank Minimization with Applications*, Ph.D. thesis, Stanford University, Stanford, CA, 2002; available online from http://faculty.washington.edu/mfazel/thesis-final.pdf.

[30] M. Heo, R. L. Leibel, B. B. Boyer, W. K. Chung, M. Koulu, M. K. Karvonen, U. Pesonen, A. Rissanen, M. Laakso, M. I. J. Uusitupa, Y. Chagnon, C. Bouchard, P. A. Donohoue, T. L. Burns, A. R. Shuldiner, K. Silver, R. E. Andersen, O. Pedersen, S. Echwald, T. I. A. Sørensen, P. Behn, M. A. Permutt, K. B. Jacobs, R. C. Elston, D. J. Hoffman, and D. B. Allison, *Pooling analysis of genetic data: The association of leptin receptor (LEPR) polymorphisms with variables related to human adiposity*, Genetics, 159 (2001), pp. 1163–1178.

[31] F. Krzakala, M. Mézard, and L. Zdeborová, *Reweighted belief propagation and quiet planting for random K-SAT*, J. Satisf. Boolean Model. Comput., 8 (2012/14) pp. 149–171.

[32] F. Krzakała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, *Gibbs states and the set of solutions of random constraint satisfaction problems*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 10318–10323.

[33] F. Krzakala and L. Zdeborová, *Hiding quiet solutions in random constraint satisfaction problems*, Phys. Rev. Lett., 102 (2009), 238701.

[34] M. Mézard and C. Toninelli, *Group testing with random pools: Optimal two-stage algorithms*, IEEE Trans. Inform. Theory, 57 (2011), pp. 1736–1745.

[35] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501, https://doi.org/10.1137/070697835.

[36] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[37] J. Scarlett and V. Cevher, *Phase transitions in the pooled data problem*, in Proceedings of the Thirty-First Annual Conference on Neural Information Processing Systems, 2017.

[38] A. Sebő, *On two random search problems*, J. Statist. Plann. Inference, 11 (1985), pp. 23–31.

[39] P. Sham, J. S. Bader, I. Craig, M. O'Donovan, and M. Owen, *DNA pooling: A tool for large-scale association studies*, Nat. Rev. Genet., 3 (2002), pp. 862–871.

[40] A. Sly, N. Sun, and Y. Zhang, *The number of solutions for random regular NAE-SAT*, in Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science, 2016, pp. 724–731.

[41] T. Tanaka, *A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors*, IEEE Trans. Inform. Theory, 48 (2002), pp. 2888–2910.

[42] I.-H. Wang, S.-L. Huang, K.-Y. Lee, and K.-C. Chen, *Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms*, in Proceedings of the IEEE International Symposium on Information Theory, 2016, pp. 1386–1390.

[43] Y. Wu and S. Verdú, *Fundamental limits of almost lossless analog compression*, in Proceedings of the IEEE International Symposium on Information Theory, 2009, pp. 359–363.

[44] L. Zdeborová and F. Krzakala, *Statistical Physics of Inference: Thresholds and Algorithms*, preprint, https://arxiv.org/abs/1511.02476, 2015.

[45] P. Zhang, F. Krzakala, M. Mézard, and L. Zdeborová, *Non-adaptive pooling strategies for detection of rare faulty items*, in Proceedings of the IEEE International Conference on Communications Workshop, 2013, pp. 1409–1414.

[46] K. S. Zigangirov, *Theory of Code Division Multiple Access Communication*, John Wiley & Sons, Hoboken, NJ, 2004.