Searching for (sharp) thresholds in random structures: where are we now?

Will Perkins[*]

August 7, 2024

Abstract

We survey the current state of affairs in the study of thresholds and sharp thresholds in random structures on the occasion of the recent proof of the Kahn–Kalai Conjecture by Park and Pham and the fairly recent proof of the satisfiability conjecture for large k by Ding, Sly, and Sun. Random discrete structures appear as fundamental objects of study in many scientific and mathematical fields including statistical physics, combinatorics, algorithms and complexity, social choice theory, coding theory, and statistics. While the models and properties of interest in these fields vary widely, much progress has been made through the development of general tools applicable to large families of models and properties all at once. Historically these tools originated to solve or make progress on specific, difficult conjectures in the areas mentioned above. We will survey recent progress on some of these hard problems and describe some challenges for the future.

This survey was prepared in conjunction with a talk for the Current Events Bulletin at the 2024 Joint Mathematics Meetings in San Francisco.

1   Introduction

Randomness is a powerful tool for algorithm design, scientific discovery, modeling the world, engineering, and mathematical proof. To use randomness effectively and confidently, we would like to accurately understand the properties of a typical random object or structure. Perhaps most familiar is the use of the Central Limit Theorem to understand statistical significance and margin of error in scientific studies and political polling.

In modern applications in algorithms, physics, social science, and other fields the random objects of interest can be very large and complex (think large computer networks, social networks, neural networks, or models of many interacting particles). We would like to be able to understand with as much accuracy as possible, what properties random structures typically possess, and how these typical properties change as underlying parameters change.

Even for the simplest class of random structures (those in which elements of a structured set are included independently with the same probability) this question can be extremely challenging, depending on the complexity of the property of interest. Over the past 40 or more years, specific problems have driven different fields (including probabilistic combinatorics, algorithms, and statistical physics) to develop powerful tools for investigating this type of problem. Some of these tools are designed for specific settings while others are very general.

Recently major progress on both general and specific problems has been made: Park and Pham proved the 'Kahn–Kalai Conjecture' [?] and Ding, Sly, and Sun proved the 'Satisfiability Conjecture'

---

[*]School of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA

for large k [?]. The aim of this survey will be to put these exciting developments in a shared context and give an idea of some of the remaining and pressing challenges in the general area of thresholds in random structures. A key theme will be the way in which probability, combinatorics, and algorithms interact in these questions.

1.1 Random structures, monotone properties, and thresholds

The main setting for this survey will be p-biased product probability measures on the discrete cube $\{0,1\}^N$; that is, associating to a vector $x \in \{0,1\}^N$ the set $S_x = \{i \in [N] : x_i = 1\}$, probability measures of the form

$$\mu_p(S) = p^{|S|}(1-p)^{N-|S|}.$$

Typically we will think of N being large, and p small, possibly depending on N. This is simply a random subset with elements chosen independently with the same probability. Often the set of coordinates $[N] = \{1, \ldots, N\}$ will have additional structure; the coordinates might represent vertices or edges of some graph, or elements of an ordered set, or many other possibilities. Then we can think of the random set $S \sim \mu_p$ as being a *random structure*.

We will see numerous examples shortly, but perhaps the first example to have in mind is flipping a sequence of N independent coins each with probability p of landing heads. The structure here is the ordering of coordinates first to last.

A *property* $\mathcal{F}$ is simply an event in this probability space, $\mathcal{F} \subseteq \{0,1\}^N$. A property $\mathcal{F}$ is non-trivial if $\mathcal{F} \neq \emptyset$, $\mathcal{F} \neq \{0,1\}^N$. In our example, non-trivial properties include 'Flipping all heads'; 'Flipping at least $N/3$ heads'; 'Flipping an odd number of heads', 'Flipping a tail on the 7th and 11th flips', and so on. A property is *monotone* if it is closed under changing 0's to 1's; that is, if $x \in \mathcal{F}$ and $y \geq x$ coordinate-wise, then y must also be in $\mathcal{F}$. In the examples above about coin flips, all properties are monotone (or have a monotone complement) except the property of flipping an odd number of heads.

For a non-trivial monotone property, the probability that $\mathcal{F}$ holds under μ_p increases from 0 to 1 as p increases from 0 to 1, and we will be interested in how rapidly it does so. Take the property of flipping at least $N/3$ heads. By using a normal approximation to a binomial, we see that when $p = \frac{1}{3} + \frac{c\sqrt{2}}{3\sqrt{N}}$, $\mu_p(\mathcal{F})$ is approximately equal to the probability that a standard normal is at least $-c$; and so as p passes $1/3$ in an interval of length on the order $N^{-1/2}$, $\mu_p(\mathcal{F})$ jumps rapidly from near 0 to near 1. The rest of this survey is devoted to understanding similar phenomena for much more complex properties of random structures.

1.2 Examples

Here we give some examples from combinatorics, probability, statistical physics, computer science, and social choice theory. These examples have driven considerable interest in thresholds in random structures from different fields and their study from these different perspectives has also given the field shared language and intuition.

1.2.1 Probabilistic combinatorics

A central example comes from combinatorics. Here we take $n \in \mathbb{N}$ and let $N = \binom{n}{2}$, with coordinates representing edges of the complete graph K_n on n vertices. Then the elements of $\{0,1\}^N$ are in correspondence with (labeled) graphs on n vertices. A sample from μ_p is a random graph on n vertices in which each possible edge is included independently with probability p; the model is known as the *Erdős-Rényi random graph* [?, ?, ?, ?]. See two samples from $G(n,p)$ in Figures ?? and ??. Despite the simplicity of its definition, the Erdős-Rényi random graph exhibits a wide range of fascinating
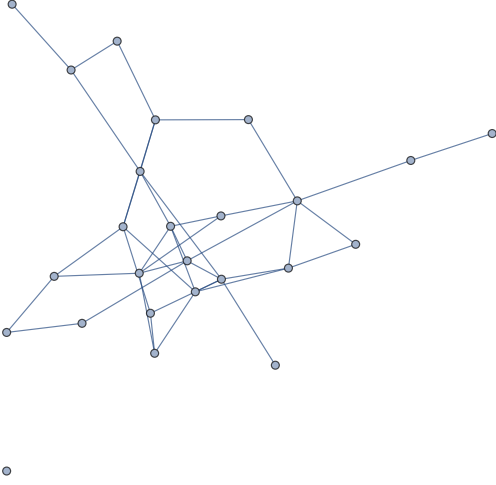
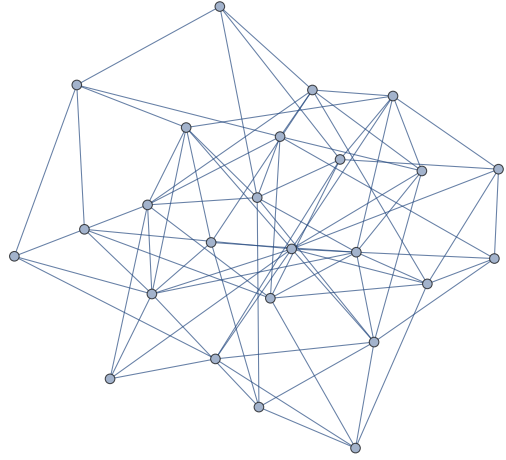Figure 1: A realization of the random graph $G(25, 1/8)$

Figure 2: A realization of the random graph $G(25, 1/4)$

behaviors, and has been studied from many angles: as an interesting random structure in its own right, as a source of examples and counterexamples in graph theory, as a source of conjectured hard instances of combinatorial optimization problems, and as a host graph for random processes in probability and statistical physics.

1.2.2 Computer science

In computer science, a computational decision problem can be represented by a property of the discrete cube; or equivalently, *boolean function* $f : \{0, 1\}^N \to \{0, 1\}$, where the input x is encoded as a binary string and $f(x) = 1$ if x is a YES instance and 0 otherwise. A graph-theoretic decision problem corresponds directly to a property of graphs, as above. The focus in computer science is on the computational resources (time, space, etc.) required to compute a given boolean function f, either in the worst case over inputs, or, as we discuss more below, for typical or random instances x, such as those drawn according to μ_p.

1.2.3 Statistical physics

Much of the language and intuition around thresholds and sharp thresholds comes from physical systems that exhibit sudden changes in qualitative behavior as some parameter is changed in a small way. This is the phenomenon of a *phase transition*. The phase transitions of water from liquid to solid as temperature drops below $32°$ F or from liquid to gas as temperature rises past $212°$ F are familiar to everyone.

Mathematically, phase transitions are non-analytic points (or discontinuities in functions or their derivatives) of observables of infinite systems as some parameter varies. This is not quite the setting described above of sequences of finite random structures with parameters that vanish as the system size grows, but there is much to be gained by pushing this analogy in both directions; in Section **??** we describe the impact of the phase transition perspective on computer science, while [**?**] takes the analogy in the other direction.

The statistical physics model perhaps most related to μ_p is that of *percolation* [**?**, **?**, **?**]. Here the setting is typically an infinite graph, say a lattice like $\mathbb{Z}^d$ or the hexagonal lattice. In site (resp.
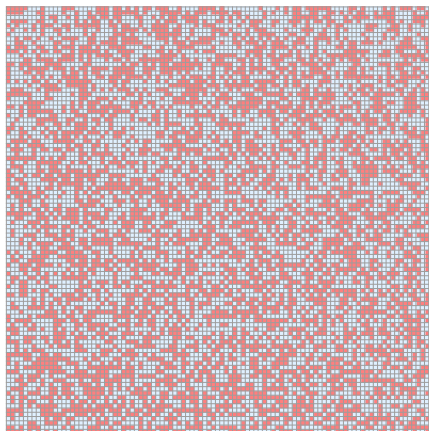
3

Figure 3: Site percolation on a 100×100 grid with $p = 1/2$

bond) percolation each vertex (resp. edge) is declared 'open' independently with probability p; the main question is about the existence or non-existence of an infinite, connected 'open' component. See Figure ?? for a depiction of site percolation on a finite portion of $\mathbb{Z}^2$.

1.2.4 Social choice theory

Social choice theory is the study of collective decision making, including the study of voting mechanisms. We can view an election between two candidates with N voters as a boolean function $f : \{0,1\}^N \to \{0,1\}$, where the candidates are labeled 0 and 1, and a vector $x \in \{0,1\}^N$ is the list of the votes of the N voters. The function f is a rule for determining the winner of the election given the votes. A monotone function f can be thought of as a rule with the sensible property that if a voter changes their vote from 0 to 1 the outcome of the election cannot change from 1 to 0. Social choice theory lends a lot of evocative terminology to the study of boolean functions and thresholds. A 'dictator' boolean function is one that depends only on one coordinate; a 'junta' is one that depends only on a small number of coordinates. The majority function (i.e. equal representation democracy) has some extremal properties with applications in approximation algorithms ('Majority is stablest' [**?**, **?**]), while the 'tribes' function, a hierarchical majority function from [**?**], inspired conjectures on general properties of boolean functions that have proved essential to the study of thresholds.

1.3 Outline

In Section **??** we define critical probabilities, thresholds, sharp and coarse thresholds, and scaling windows, then illustrate these notions with three examples from the study of random graphs. In the next three sections, we present some problems and conjectures that have driven progress on thresholds in different fields. In Section **??** we present the first of these main motivating questions, the question of thresholds for spanning structures in random graphs and hypergraphs. We discuss perfect matchings in hypergraphs, Latin squares, and the Kahn–Kalai Conjecture. In Section **??** we discuss the random k-SAT model, the satisfiability conjecture, and statistical physics predictions for thresholds in random computational problems. In Section **??** we describe the stochastic block model and the predicted information-theoretic and algorithmic thresholds in the model.

We conclude in Section **??** with some questions and open problems.

1.4 Further reading

This survey is not meant to be exhaustive at all, but rather aims to describe some recent results and open problems in different areas, all around the topic of thresholds and sharp thresholds in random structures. This area is very fortunate to have several excellent books and surveys devoted to the topic from different perspectives. Some of these sources are listed below with a few remarks.

Friedgut's 2005 survey 'Hunting for sharp thresholds' [?], following the developments in [?, ?, ?], is devoted to the question of which properties have sharp thresholds (defined below), criteria for proving sharpness of a threshold, and some intuition and meta conjectures on the topic.

Kalai and Safra's 2006 survey [?] takes a broad look at threshold phenomena in computer science, mathematics, and social choice theory and explains how notions arising in the latter (of 'influence' or 'pivotality' of coordinates) along with Fourier analysis of boolean functions can be used to study the sharpness or coarseness of thresholds. O'Donnell's textbook on discrete Fourier analysis of boolean functions [?] is a great reference to learn about these tools.

From a different perspective, Duminil-Copin's 2019 survey [?] describes how general tools from the study of sharp thresholds of boolean functions can be used to study phase transitions in classical statistical physics models like percolation on infinite graphs.

Rao's recent survey [?] describes the 2021 breakthrough of Alweiss, Lovett, Wu, and Zhang on sunflowers [?], the method of which led to the proof of the fractional Kahn–Kalai conjecture by Frankston, Kahn, Narayanan, and Park [?]. Even more recently Park has written an expository article [?] explaining the intuition behind the Kahn–Kalai conjecture and the consequences of its proof by Park and Pham [?].

Finally, in Sections **??** and **??** on random computational problems and statistical inference we discuss areas that have been shaped to a great extent by the field of statistical physics through questions, conjectures, and methods. The textbook of Mezard and Montanari [?] is a great resource for understanding these methods, while the surveys of Zdeborová and Krzakala [?], Moore [?], and Abbe [?] give an account of developments in this area in the last decade.

2 General questions, notions, and phenomena

In what follows, the perspective will be asymptotic as some underlying parameter, usually N or n, tends to infinity. We use standard asymptotic notation: $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$, $\Theta(\cdot)$ to compare growth rates of functions. We use a subscript, e.g. $O_\varepsilon(\cdot)$, to indicate the implied constant may depend on ε. When discussing probabilities we say an event A holds 'with high probability' or 'whp' as $N \to \infty$ if $\lim_{N\to\infty} \Pr(A) = 1$; that is, $\Pr(A) = 1 - o(1)$. Here we implicitly consider a sequence of probability spaces.

For a non-trivial monotone property $\mathcal{F}$, the probability $\mu_p(\mathcal{F})$, as a function of p, is a strictly increasing function that starts with $\mu_0(\mathcal{F}) = 0$ and $\mu_1(\mathcal{F}) = 1$ (in fact it is a polynomial in p). The basic question about a monotone property of a random structure is how $\mu_p(\mathcal{F})$ increases from 0 to 1 as p increases. More generally we will think of a sequence of random structures indexed by N along with a sequence of monotone properties, and ask how $\mu_p(\mathcal{F}_N)$ increases from 0 to 1 as a function of N. From here on, we will write $\mathcal{F}$ and μ_p even though both implicitly depend on N.

Because $\mu_p(\mathcal{F})$ is strictly increasing, we can define the *critical probability* of $\mathcal{F}$, $p_c(\mathcal{F})$, as

$$p_c(\mathcal{F}) = \{p : \mu_p(\mathcal{F}) = 1/2\}.$$

The first task in studying a particular property $\mathcal{F}$ is to identify, at least approximately, the critical probability.

Task 1. *For a given monotone property $\mathcal{F}$, determine $p_c(\mathcal{F})$ asymptotically as $N \to \infty$, or at least determine the asymptotic order of $p_c(\mathcal{F})$.*

Obtaining the asymptotics of p_c means finding some $f(n)$ so that $p_c(\mathcal{F}) = (1 + o(1))f(n)$; finding the asymptotic order means finding $f(n)$ so that $p_c(\mathcal{F}) = \Theta(f(n))$.

In the context of random graphs, this task was put forward in the original paper of Erdős and Rényi [?] and has been a central topic in probabilistic combinatorics since. In Section **??** we discuss one of the most challenging classes of problems for this task: that of determining the asymptotic order of p_c for the existence of different kinds of spanning subgraphs (including perfect matchings for random graphs and hypergraphs, H-factors of random graphs, and other combinatorial designs (including Latin squares) in random structures). We will discuss the recent breakthrough work on the Kahn–Kalai conjecture in [**?, ?**] that has led to a new powerful tool for determining the asymptotic order of p_c.

After Task **??**, the most fundamental question about monotone properties is how quickly $\mu_p(\mathcal{F})$ increases from near 0 to near 1. To quantify this, we define the *width of the scaling window* of a monotone property. Let

$$T_\varepsilon(\mathcal{F}) = \left\{ p_{1-\varepsilon} - p_\varepsilon : \mu_{p_{1-\varepsilon}}(\mathcal{F}) = 1 - \varepsilon \text{ and } \mu_{p_\varepsilon}(\mathcal{F}) = \varepsilon \right\}.$$

In words T_ε is the amount p has to increase for the probability of $\mathcal{F}$ to increase from ε to $1 - \varepsilon$.

Erdős and Rényi [?] defined a *threshold function* for a monotone property as follows: $p_t(n)$ is a threshold function for a monotone property $\mathcal{F}$ if the following hold:

1. If $p = \omega(p_t(n))$, then $\mu_p(\mathcal{F}) = 1 - o(1)$.

2. If $p = o(p_t(n))$, then $\mu_p(\mathcal{F}) = o(1)$.

Equivalently, $\mathcal{F}$ has a threshold function (and we may take the function to be $p_c(\mathcal{F})$) if for every small $\varepsilon > 0$, $T_\varepsilon(\mathcal{F}) = O_\varepsilon(p_c(\mathcal{F}))$.

The following result of Bollobás and Thomason justifies the abstract study of thresholds in monotone properties and the specific definition of $p_c(\mathcal{F})$ above.

Theorem 1 (Bollobás–Thomason [?]). *Every non-trivial monotone property has a threshold function; moreover one can take $p_c(\mathcal{F})$ to be this threshold function.*

The original proof of this theorem uses the Kruskal–Katona theorem, but a simple modern proof is as follows. Let K be a large positive integer. If $\mu_p(\mathcal{F}) = \varepsilon$, then $\mu_{Kp}(\mathcal{F}) \geq 1 - (1 - \varepsilon)^K$ by superimposing K independent copies of the random structure and applying monotonicity; taking $K \approx \log(1/\varepsilon)/\varepsilon$ proves the theorem.

Erdős and Rényi further classified thresholds as *sharp* or *coarse* according to how $T_\varepsilon(\mathcal{F})$ compares to $p_c(\mathcal{F})$. We say $\mathcal{F}$ has a *sharp threshold* if for every $\varepsilon > 0$, $T_\varepsilon(\mathcal{F}) = o_\varepsilon(p_c(\mathcal{F}))$; otherwise $\mathcal{F}$ has a coarse threshold. Equivalently, $\mathcal{F}$ has a sharp threshold if for every $\varepsilon > 0$,

1. If $p \geq (1 + \varepsilon)p_c(\mathcal{F})$, then $\mu_p(\mathcal{F}) = 1 - o(1)$.

2. If $p \leq (1 - \varepsilon)p_c(\mathcal{F})$, then $\mu_p(\mathcal{F}) = o(1)$.

Task 2. *Determine if a monotone property $\mathcal{F}$ has a sharp or coarse threshold.*

In fact, in very general settings Task **??** has been solved, with the machinery of discrete Fourier analysis. Roughly, as Friedgut describes in [?] and Kalai and Safra write in [?], we should expect a sharp threshold unless there is reason to expect a coarse threshold; and the reason to expect a coarse

threshold is if $\mathcal{F}$ is essentially determined by a small number of coordinates or by the presence or absence of a small substructure. This is made rigorous in increasing generality in Friedgut's theorem on sharp thresholds [?], Bourgain's theorem in the appendix of [?], and Hatami's result in [?]. We describe in Section ?? the random k-SAT model, a main motivation for Friedgut's important result.

A final task in understanding a particular property $\mathcal{F}$ with a sharp threshold is getting more precise bounds on the scaling window.

Task 3. *Determine the asymptotic order of the width of the scaling window of a monotone property* $\mathcal{F}$.

Unlike with Tasks ?? and ??, as of yet there is no general principle in determining the magnitude of $T_\varepsilon(\mathcal{F})$ beyond distinguishing sharp from coarse: the width of the scaling window really seems to depend on the particular details of the property $\mathcal{F}$ in question, and being able to determine the asymptotic order of the width indicates a near-complete understanding of the property.

2.1 Examples

We give three examples to illustrate these different tasks. The examples are the properties in the random graph $G(n, p)$ of *containing a triangle*, *being connected*, and *being 3-colorable*; call these properties $\mathcal{F}_{K_3}$, $\mathcal{F}_{\text{connected}}$, $\mathcal{F}_{3-\text{col}}$ respectively.

To understand the probability that $G(n, p)$ contains a triangle, we can use the first- and second-moment methods (see e.g. [?] for an exposition). Let X be the number of triangles. Then $\mathbb{E}X = \binom{n}{3}p^3$ and $\mathbb{V}\mathrm{ar}(X) = (1 + o(1))\binom{n}{3}p^3$. Using this we can determine that $p_c(\mathcal{F}_{K_3}) = \Theta(1/n)$. Markov's Inequality yields:

$$\Pr(\mathcal{F}_{K_3}) = \Pr(X \geq 1) \leq \mathbb{E}X = (1 + o(1))\frac{n^3 p^3}{6} = o(1)$$

when $p = o(1/n)$. Chebyshev's Inequality yields

$$\Pr(\overline{\mathcal{F}_{K_3}}) = \Pr(X = 0) \leq \Pr(|X - \mathbb{E}X| \geq \mathbb{E}X) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{(\mathbb{E}X)^2} = (1 + o(1))\frac{6}{n^3 p^3} = o(1)$$

when $p = \omega(1/n)$. This pair of facts tells us that $p = c/n$ (for any constant $c > 0$) is a threshold function for $\mathcal{F}_{K_3}$. In fact much more is known: when $p = c/n$, the distribution of X converges to a Poisson($c^3/6$) random variable. Thus the scaling window is of length $\Theta(1/n)$ and the threshold is coarse; see Figure ??.

Next we turn to the property $\mathcal{F}_{\text{connected}}$. Being connected is a more 'global' property than that of containing a triangle, but there is an easy lower bound on $p_c(\mathcal{F}_{\text{connected}})$: if G has an isolated vertex (and $n > 1$) then G must be disconnected. Or, taking complements, $\mathcal{F}_{\text{connected}} \subseteq \mathcal{F}_{\text{no isolated vertex}}$. As with triangles it is straightforward to understand the threshold for containing isolated vertices using the method of moments. Let Y be the number of isolated vertices; then $\mathbb{E}Y = n(1 - p)^{n-1}$. As with triangles, when $\mathbb{E}Y$ tends to a positive constant, the distribution of Y converges to a Poisson random variable; when $\mathbb{E}Y \to 0$ whp $Y = 0$; and when $\mathbb{E}Y \to \infty$, whp $Y \geq 1$. This tells us the exact scaling window for $\mathcal{F}_{\text{no isolated vertex}}$: if $p = \frac{\log n + c}{n}$ with $c \in \mathbb{R}$ constant, then $\Pr(\mathcal{F}_{\text{no isolated vertex}}) = (1 + o(1))e^{-e^{-c}}$, see Figure ??. In this case the width of the scaling window is $\Theta(1/n) = o(p_c)$ and so the threshold is sharp.

It turns out that a first-moment argument on connected components of size ≥ 2 from Erdős and Rényi [?] shows that in fact $\mathcal{F}_{\text{no isolated vertex}}$ approximates $\mathcal{F}_{\text{connected}}$ very well indeed around the threshold. An effective way to state the approximation is as a *hitting time result*, which we now describe.
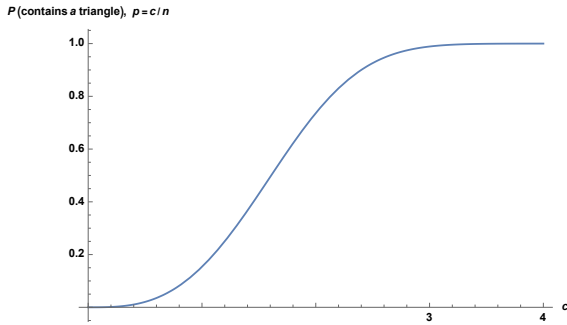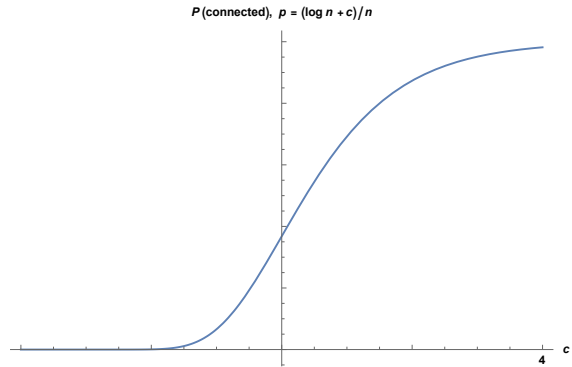
Figure 4: Scaling window of $\mathcal{F}_{K_3}$

Figure 5: Scaling window of $\mathcal{F}_{\text{connected}}$

We can couple the random graphs $G(n,p)$ for all $p \in [0,1]$ simultaneously by drawing iid Uniform$[0,1]$ random variables U_{ij}, $1 \le i < j \le n$. Then we form $G(n,p)$ by including $(ij) \in E$ if and only if $U_{ij} \le p$. Under this coupling $G(n,p')$ is a subgraph of $G(n,p)$ if $p' \le p$. We can then define the 'random graph process' as the discrete time process $G_0, G_1, \ldots, G_{\binom{n}{2}}$ where at each step a uniformly random edge (not already present) is added to the graph. This process is what results from raising p from 0 to 1 and marking the appearance of new edges in the coupling.

Theorem 2 (Hitting time for connectivity [?]). *Whp over the random graph process we have*

$$\min_t \{G_t \text{ is connected}\} = \min_t \{G_t \text{ has no isolated vertices}\}.$$

That is, the edge that touches the last isolated vertex in the process also connects the graph, with high probability over the process. This implies that the threshold and scaling widow are the same for $\mathcal{F}_{\text{connected}}$ and $\mathcal{F}_{\text{no isolated vertex}}$.

We will see below in Section **??** further examples of hitting time results along the same lines for more complicated properties than connectivity.

Finally, consider a property for which we do not yet have a full understanding. Let $\mathcal{F}_{3-\text{col}}$ be the property that G has a proper 3-coloring of its vertices (no monochromatic edges). The complement of this property is a non-trivial monotone property.

One can show that if $p = c/n$, $c < 1$, then whp $G(n,p)$ is 3-colorable by showing that whp all connected components are trees or unicyclic (and thus 3-colorable). For an upper bound on p_c one can use the first moment method on Z, the number of 3-colorings of G. As a first attempt we can bound $\mathbb{E}Z \le 3^n(1-p)^{\frac{n^2}{6}}$ using the fact that a balanced partition of n vertices has the fewest potential monochromatic edges. This tells us that $G(n,c/n)$ is not 3-colorable whp when $c > 6\log 3 \approx 6.5917$. However, we can do something a little bit more clever: whp $G(n,c/n)$ has $\frac{cn}{2} + o(n)$ edges and so we can condition on the number of edges m, and compute $\mathbb{E}Z$ in the random graph G_m from the random graph process. This gives the bound

$$\mathbb{E}_{G_m} Z \le 3^n \left(1 - \frac{1}{3}\right)^m$$

which tends to 0 when $m \sim cn/2$, $c > \frac{2\log 3}{\log 3 - \log 2} \approx 5.419$. We will see more of this idea of combining the moment method with conditioning on typical events in Section **??**.

Together these facts tell us that $p_c(\mathcal{F}_{3-\text{col}}) = \Theta(1/n)$ and the width of the scaling window is $O(1/n)$, uniformly over ε. What are the asymptotics of p_c? Is there some d_3 so that $G(n,d/n)$ is 3-colorable whp when $d < d_3$ and whp not 3-colorable when $d > d_3$? In fact this is the only remaining

8

open problem from the original Erdős and Rényi paper on random graphs [**?**]. In Section **??** we describe predictions for this problem made by statistical physicists.

3 Perfect matchings, Latin squares, Kahn–Kalai, and spread distributions

From the beginning of the study of random graphs, a key question has been 'For a given subgraph H, when in the evolution of the random graph does a copy of H first appear?' That is, what is the threshold for the property $\mathcal{F}_H$ of containing a copy of H.

For fixed size H, the answer is now known completely [**?**], though more delicate in general than the case of $H = K_3$ above. A threshold function for $\mathcal{F}_H$ is the *expectation threshold* for the densest subgraph H' of H (maximizing the ratio of edges to vertices over all induced subgraphs of H); the value $p_E(H')$ so that the expected number of copies of H' in $G(n, p_E(H'))$ is 1. For $p = o(p_E(H'))$, the expected number of copies of H' tends to 0, and so whp $G(n, p)$ has no copies of H' and thus no copies of H. When the expected number of copies of H' for each $H' \subseteq H$ tends to infinity, it has been showed that whp a copy of H exists.

It can be much more difficult to determine the threshold for the appearance of subgraphs whose size grows with n; in particular, for *spanning subgraphs* that include each of the n vertices. Two classical examples are those of H being a perfect matching (with n even) and H being a Hamilton cycle. As with connectivity, there are natural lower bounds for $p_c(\mathcal{F}_H)$ in these cases, coming from local obstructions. If G has an isolated vertex, it cannot have a perfect matching; if G has a vertex of degree at most 1 it cannot have a Hamilton cycle. In both cases these obstructions determine the thresholds, even in the strong form of hitting-time results [**?, ?, ?**] that give a complete understanding of the scaling windows.

Other spanning subgraph problems remained open for many years. One famous such problem is Shamir's Problem [**?**] (mentioned first in print by Erdős in [**?**]): the problem of determining the threshold for a random k-uniform hypergraph to contain a perfect matching.

A k-uniform hypergraph $G = (V, E)$ is a set of vertices V along with a collection E of k-sets of V (the case $k = 2$ is a usual graph). The random k-uniform hypergraph $G^{(k)}(n, p)$ is a k-uniform hypergraph on n vertices in which each possible k-set is an edge independently with probability p. A perfect matching in a k-uniform hypergraph on n vertices is a set of n/k vertex-disjoint hyperedges that (necessarily) cover all n vertices; n must be divisible by k for a perfect matching to be possible. For $k = 2$ (a graph) this coincides with the usual definition of a perfect matching.

As with perfect matchings in graphs, the natural obstruction to a perfect matching in a hypergraph is a vertex not appearing in any hyperedges; this immediately yields a lower bound of $\Omega(\log n \cdot n^{1-k})$ for the threshold for perfect matchings.

After many years of effort, Johansson, Kahn, and Vu proved that the lower bound is tight up to constants.

Theorem 3 (Johansson–Kahn–Vu [**?**])**.** *Let $\mathcal{F}$ be the property of $G^{(k)}(n, p)$ containing a perfect matching when k divides n. Then*

$$p_c(\mathcal{F}) = \Theta_k(\log n \cdot n^{1-k}).$$

Much more recently Kahn proved much finer results: first-order asymptotics of $p_c(\mathcal{F})$ in [**?**] and the definitive hitting-time result [**?**] (see also the related [**?, ?, ?**]).

3.1 The Kahn–Kalai conjectures

Motivated by challenges like Shamir's Problem, Kahn and Kalai [?] made two bold conjectures about thresholds. These conjectures arise from the fact that it is often straightforward to give a lower bound for $p_c(\mathcal{F})$; that is, to show that when p is small enough $\mu_p(\mathcal{F}) = o(1)$. A matching or near matching upper bound is often much more difficult. This is especially apparent in the examples above (connectivity, perfect matchings, etc.) in which there is a trivial local obstruction to some property. Kahn and Kalai's two conjectures both have a similar flavor: 'The best possible easy lower bound on p_c is not too far from the truth.' We present the two conjectures in the opposite order they appear in [?] (but in the order they arrived at them).

Their second conjecture is about the property $\mathcal{F}_H$ of containing an isomorphic copy of some subgraph H in a random graph (or hypergraph). We have seen several examples of this for different H, including a triangle, a perfect matching, a Hamilton cycle, and a hypergraph perfect matching. As we saw for fixed-size subgraphs, it can be crucial to consider the rarest (in terms of expected number of copies) subgraphs of H. This motivates the following definition of the *subgraph expectation threshold* of H:

$$p_E(H) = \min\{p : \mathbb{E}_{G(n,p)} X_{H'} \geq 1 \,\forall \text{ subgraphs } H' \subseteq H \,\}.$$

Clearly $p_E(H)$ bounds below the threshold of $\mathcal{F}_H$; if $p = o(p_E(H))$ then there is some subgraph $H' \subseteq H$ that whp does not appear in $G(n,p)$. Kahn and Kalai conjecture this cannot be too far from the truth.

Conjecture 4 (Kahn and Kalai [?]). *There exists an absolute constant K so that for all graphs H,*

$$p_c(\mathcal{F}_H) \leq K \cdot p_E(H) \cdot \log |V(H)|\,.$$

For H of fixed size, Conjecture ?? asserts that $p_e(H)$ is a threshold for $\mathcal{F}_H$ and this is known to be true as described above. The real content of the conjecture is for H that grow with n. In particular Conjecture ?? is tight (up to the constant K) for perfect matchings in graphs and hypergraphs.

Their first conjecture was in the more general setting of monotone properties of $\{0,1\}^N$. To state it we must define a more abstract notion of an expectation threshold (following [?, ?, ?]). A property $\mathcal{F} \subseteq \{0,1\}^N$ is *p-small* if there exists a 'cover' $\mathcal{G} \subseteq \{0,1\}^N$ so that:

1. $\forall T \in \mathcal{F} \,\exists\, S \in \mathcal{G}, S \subseteq T$.

2. $\sum_{S \in \mathcal{G}} p^{|S|} \leq \frac{1}{2}$.

To give some intuition for this definition, if H is a graph and $H' \subseteq H$, then $\mathcal{F}_{H'}$ is a cover of $\mathcal{F}_H$, satisfying the first condition above.

The *expectation threshold* of $\mathcal{F}$, $q(\mathcal{F})$, is the largest p for which $\mathcal{F}$ is p-small. Finally let $L(\mathcal{F})$ be the maximum size of a minimal element of $\mathcal{F}$.

Notice that $q(\mathcal{F})$ is a lower bound for $p_c(\mathcal{F})$. Let $q = q(\mathcal{F})$. Then

$$\mu_q(\mathcal{F}) \leq \sum_{S \in \mathcal{G}} \sum_{T \in \mathcal{F}, T \supseteq S} \mu_q(T)$$
$$\leq \sum_{S \in \mathcal{G}} \sum_{T \supseteq S} \mu_q(T)$$
$$= \sum_{S \in \mathcal{G}} q^{|S|} \leq \frac{1}{2}\,.$$

Kahn and Kalai conjectured that in the abstract setting this cannot be far from the truth for p_c. Park and Pham then proved this.

Theorem 5 (Kahn–Kalai Conjecture, now Park–Pham Theorem [**?**])**.** *There exists an absolute constant K so that for every monotone $\mathcal{F}$,*

$$p_c(\mathcal{F}) \leq K \cdot q(\mathcal{F}) \cdot \log L(\mathcal{F}).$$

Theorem **??** gives a powerful method for bounding p_c from above: find a good (or the best) $\mathcal{G}$ with the properties above, and this will determine p_c up to a $\log N$ factor or better. Looking ahead, we can characterize $q(\mathcal{F})$ via the following integer optimization problem over the variables $g(x)$, $x \in \{0,1\}^N$ which we interpret as the indicator vector of a potential cover $\mathcal{G}$ witnessing $\mathcal{F}$ being p-small.

$$V(\mathcal{F}, p) = \min \sum_{x \in \{0,1\}^N} p^{|x|} g(x)$$

subject to

$$g(x) \in \{0,1\} \text{ for all } x \in \{0,1\}^N$$

$$\sum_{S \subseteq T} g(x_S) \geq 1 \text{ for all } T \in \mathcal{F}$$

In particular, if $V(\mathcal{F}, p) \geq 1/2$, then $q(\mathcal{F}) \leq p$, and by Theorem **??**, $p_c(\mathcal{F}) \leq K \cdot p \cdot \log L(\mathcal{F})$. However, solving an integer program over 2^N variables (for all large N) seems somewhat formidable.

3.2 Duality and spread distributions

Before Park and Pham proved Theorem **??**, Frankston, Kahn, Narayanan, and Park [**?**] (following the breakthrough of Alweiss, Lovett, Wu, and Zhang on the sunflower conjecture [**?**]) proved a related but weaker conjecture of Talagrand from [**?**], dubbed the 'Fractional Kahn–Kalai Conjecture'. To state this, we first relax the integrality constraint above and define a linear program (in 2^N variables)

$$V_f(\mathcal{F}, p) = \min \sum_{x \in \{0,1\}^N} p^{|x|} g(x)$$

subject to

$$g(x) \in [0,1] \text{ for all } x \in \{0,1\}^N$$

$$\sum_{S \subseteq T} g(x_S) \geq 1 \text{ for all } T \in \mathcal{F}.$$

Then define $q_f(\mathcal{F})$, the *fractional expectation threshold* of $\mathcal{F}$, to be the largest p for which $V_f(\mathcal{F}, p) \leq 1/2$. Again it is easy to see that $p_c(\mathcal{F}) \geq q_f(\mathcal{F})$, with the same proof as above.

Moreover, we have the relations

$$q(\mathcal{F}) \leq q_f(\mathcal{F}) \leq p_c(\mathcal{F}) \leq K \cdot q(\mathcal{F}) \cdot \log L(\mathcal{F}) \leq K \cdot q_f(\mathcal{F}) \cdot \log L(\mathcal{F}),$$

where the second-to-last inequality is Theorem **??**; the weaker inequality $p_c(\mathcal{F}) \leq K \cdot q_f(\mathcal{F}) \cdot \log L(\mathcal{F})$ is the main result of [**?**].

Theorem 6 (Frankston, Kahn, Narayanan, and Park [**?**])**.** *There exists an absolute constant K so that for every monotone $\mathcal{F}$,*

$$p_c(\mathcal{F}) \leq K \cdot q_f(\mathcal{F}) \cdot \log L(\mathcal{F}).$$

To bound p_c with Theorem **??** it suffices to lower bound $V_f(\mathcal{F}, p)$: if $V_f(\mathcal{F}, p) \geq 1/2$ then $p_c(\mathcal{F}) \leq K \cdot p \cdot \log L(\mathcal{F})$. The nice thing about the linear programing formulation is that we can use duality to

give a lower bound. We can write the dual as a linear program with variables $\nu(T)$ for each $T \in \mathcal{F}$ and constraints for each $S \subseteq \{0,1\}^N$.

$$V_f(\mathcal{F}, p) = \max \sum_{T \in \mathcal{F}} \nu(T)$$

$$\text{subject to}$$

$$\nu(T) \geq 0 \text{ for all } T \in \mathcal{F}$$

$$\sum_{T \supseteq S} \nu(T) \leq p^{|S|} \text{ for all } S \subseteq \{0,1\}^N.$$

We are interested in showing $V_f(\mathcal{F}, p) \geq 1/2$ to bound p_c and via duality this can be accomplished by exhibiting a good $\nu(\cdot)$. A slightly more elegant formulation is due to Talagrand [**?**], who made the following definition of a *spread probability distribution*.

Definition 7. *Let $\mathcal{F} \subseteq \{0,1\}^N$. A probability measure ν supported on $\mathcal{F}$ is p-spread if for all $S \subseteq \{0,1\}^N$,*

$$\sum_{T \supseteq S} \nu(T) \leq 2p^{|S|}.$$

Putting all of the above together we obtain a very useful theorem.

Theorem 8. *There is an absolute constant K so that the following is true. Let $\mathcal{F}$ be a monotone property that supports a p-spread probability measure ν. Then*

$$p_c(\mathcal{F}) \leq K \cdot p \cdot \log L(\mathcal{F}).$$

Theorem **??** follows from Theorem **??**, duality, and the extra factor 2 allowing us to take ν to be a probability measure.

3.3 Applications

Theorem **??** is a brand-new tool for proving upper bounds on $p_c(\mathcal{F})$ and we will recount here some spectacular applications. But first as a warm-up we will see how it can be used to establish $p_c(\mathcal{F}_{\text{perfect matching}}) = \Theta\left(\frac{\log n}{n}\right)$.

Let n be even, and let ν be the uniform distribution on perfect matchings of the complete graph K_n. Let A be some set of edges of K_n and let $M \sim \nu$. We want to bound the probability that $A \subseteq M$. First note that if A is not a matching then this probability is 0. Now suppose A is a matching of size k. Then, letting $\text{pm}(G)$ denote the number of perfect matchings of G and $(a)_b = \frac{a!}{(a-b)!}$,

$$\Pr(A \subseteq M) = \frac{\text{pm}(K_{n-2k})}{\text{pm}(K_n)} = \frac{(n-2k)!}{2^{n/2-k}(n/2-k)!} \frac{2^{n/2}(n/2)!}{n!} = \frac{2^k(n/2)_k}{(n)_{2k}} \leq \left(\frac{e}{n}\right)^k,$$

and so ν is $\frac{e}{n}$-spread. Applying Theorem **??** gives the result, though note that this is weaker than the sharp threshold obtained in [**?**]. Here spreadness is proved by counting, and it helps a lot that we have an explicit formula for $\text{pm}(K_n)$ to use.

From Theorem **??** and similar counting arguments (see e.g. [**?**]), one can also rather easily obtain the asymptotic order of the threshold for perfect matchings in hypergraphs and K_r-factors as well as the threshold for containing any bounded-degree spanning tree, previously known through the long proofs of Johansson, Kahn, and Vu [**?**] and Montgomery [**?**] respectively.

Perhaps even more exciting are applications which before the theorem were completely out of reach but which now can be approached through finding spread distributions. We describe one application here but see also [?, ?, ?] for other applications of either Theorem ?? or the notion of spread distributions.

A *Latin square* of order n is an $n \times n$ matrix with entries in $\{1, \ldots, n\}$ in which no row or column contains repeated entries. A Latin square of order 8 is given in Table ??. One can ask about the

1	3	4	8	2	6	7	5
6	1	2	5	7	4	8	3
5	2	3	1	8	7	4	6
4	6	5	3	1	8	2	7
2	4	6	7	3	5	1	8
7	8	1	2	6	3	5	4
8	5	7	6	4	1	3	2
3	7	8	4	5	2	6	1

Table 1: A Latin square of order 8

existence of Latin squares of order n, the number of Latin squares of order n, or the existence of Latin squares with additional properties. Johansson [?] asked a very natural threshold question about Latin squares: if each position (ij) in an $n \times n$ matrix is assigned a list L_{ij} of allowed symbols in $\{1, \ldots, n\}$ by including each element in L_{ij} independently with probability p, what is the threshold in p for the existence of a Latin square in which each entry appears in the corresponding list L_{ij}? The lower bound comes from a simple local obstruction: for $p \leq (1 - \varepsilon)\frac{\log n}{n}$ there will be empty lists whp. A natural conjecture given the results and conjectures on perfect matchings and factors in random graphs is that this local obstruction determines the threshold: its asymptotic order, first-order asymptotics, and even in the form of a hitting time. This conjecture is stated explicitly in [?], and see also the closely related conjectures in [?, ?, ?, ?].

Conjecture 9. *The property $\mathcal{F}_{Latin}$ has a sharp threshold at $p = \frac{\log n}{n}$*

Before Theorem ??, there was really no effective way to to prove upper bounds on $p_c(\mathcal{F}_{\text{Latin}})$: the best known result was that $p_c \leq 1 - \delta$ for some fixed $\delta > 0$ [?].

Using Theorem ?? and finding a sufficiently spread distribution on Latin squares, Sah, Sawhney and Simkin proved Conjecture ?? up to subpolynomial factors: $p_c = n^{-1+o(1)}$ [?]. Intriguingly, this bound already established that the threshold is sharp without determining the threshold, by applying Friedgut's theorem discussed in the next section. Following this, Kang, Kelly, Kühn, Methuku, and Osthus [?] proved $p_c = O(\log^2 n/n)$, within a factor $\log n$ of the conjecture. Then very recently Jain and Pham, and independently Keevash, established the correct order of the threshold.

Theorem 10 (Jain–Pham [?]; Keevash [?])**.**

$$p_c(\mathcal{F}_{Latin}) = \Theta\left(\frac{\log n}{n}\right) .$$

Using Theorem ??, it suffices to construct an $O(1/n)$-spread distribution on Latin squares. It is natural to expect that (as in the case of perfect matchings) the *uniform distribution* is $O(1/n)$ spread, but unlike perfect matchings it is very challenging to enumerate Latin squares. So instead, the above authors have constructed non-uniform spread distributions using sophisticated tools from probabilistic and extremal combinatorics, namely iterative absorption (in [?, ?]), analyzing the Lovász Local Lemma probability distribution (in [?]), and analyzing a random greedy stochastic process (in [?]).

4 Random k-SAT and the satisfiability conjecture

The most important open problem in theoretical computer science (and one of the most important in all of mathematics) is the P vs NP question: can computational decision problems with polynomial-time checkable certificates be solved in polynomial time?

Most computer scientists and mathematicians believe that P$\neq$NP and that many hard computational problems exist; these include classic problems like MAX-CUT, Max Independent Set, traveling salesman, boolean satisfiability, graph coloring, among many others. It is believed that these problem are computationally intractable in the worst-case over instances. On the other hand, large instances of these problems are solved every day (see for example a survey on the success of SAT solvers in [**?**] or exact solutions to very large instances of the traveling salesman problem in e.g. [**?**, **?**]). What can explain this discrepancy? One possibility is that 'typical' instances of certain NP-hard problems are tractable while hard instances are exceptional.

This is one motivation for the study of *average-case complexity*: the computational tractability or intractability of random instances of computational problems. Average-case complexity is a huge topic with many fascinating facets and many mysteries (see e.g. [**?**, **?**]). Here we will deal with a small slice of the topic.

Recall that a boolean CNF formula is the AND of OR's of literals (boolean variables and their negations). A k-CNF formula is the AND of clauses of k literals each. For example, the following is a small 3-SAT formula:

$$(x_7 \vee \overline{x}_2 \vee \overline{x}_3) \wedge (\overline{x}_1 \vee x_8 \vee \overline{x}_4) \wedge (x_2 \vee x_4 \vee x_1).$$

The k-SAT decision problem is to determine if there exists an assignment of True and False to the boolean variables so that the given k-CNF formula evaluates to true; if such an assignment exists the formula is satisfiable and unsatisfiable otherwise.

An empirical observation about algorithms brought the study of random computational problems to the attention of computer scientists, statistical physicists, and probabilists. In 1992, Mitchell, Selman, and Levesque [**?**] (working on problems in Artificial Intelligence and computational deductive reasoning) generated uniformly random 3-SAT instances on n variables with m constraints, for fairly large n and different values of m. They ran standard heuristic SAT solving and SAT refutation algorithms and observed the following: the running time required to find a solution (or find a proof that none existed) showed a sharp peak (as a function of m) right around the point at which an estimate of the probability of such a random instance being satisfiable made a sharp decrease from near 1 to near 0.

These two empirical observations – that random k-SAT exhibits a sharp threshold and that instances near the threshold are computationally hard – set off an explosion of work on random k-SAT and related models (random graph coloring, k-NAE-SAT, k-XOR-SAT, etc.) in many different directions.

We can model random k-SAT in the setting of this paper. For a given n and p, let $F_k(n,p)$ be a random k-SAT formula generated by including each of the $\binom{2n}{k}$ possible k-clauses independently with probability p. The property of being satisfiable is a non-trivial property with a monotone complement (and thus has a threshold function).

One long-standing conjecture is that the random k-SAT model exhibits a sharp threshold.

Conjecture 11 (Satisfiability Conjecture). *For each $k \geq 2$, there exists $c_k > 0$ so that for every $\varepsilon > 0$, the following hold:*

- *If $p \leq (1-\varepsilon)c_k n^{-1/(k-1)}$, then whp $F_k(n,p)$ is satisfiable.*

- *If $p \geq (1+\varepsilon)c_k n^{-1/(k-1)}$, then whp $F_k(n,p)$ is unsatisfiable.*

Conjecture **??** was proved early on for the special case of random 2-SAT [**?**]; and in fact the precise scaling window was determined [**?**]. The reason for this is that determining satisfiability of a 2-SAT formula can be reduced to determining the strongly connected components of the implication graph on literals; this gives a linear-time algorithm for 2-SAT in the worst case and gives a strong analogy between the random 2-SAT threshold and the emergence of a giant component in a random graph. The case $k \geq 3$ is fundamentally different.

Random k-SAT is an example of a *random constraint satisfaction problem* (random CSP); there is an initial set of possible solutions (here all possible assignments to n boolean variables) and a set of random constraints is selected, each of which rules out some solutions (here each clause rules out a 2^{1-k} fraction of all possible solutions). Random graph coloring is another random CSP: the q^n colorings of n vertices are the possible solutions; each edge rules out $\frac{1}{q}$-fraction of the solutions.

How can one locate the satisfiability threshold in a random CSP? As with many of the examples above, a starting point is to try the first- and second-moment methods on the random variable Z that counts the number of solutions. If p is large enough that $\mathbb{E}Z \to 0$, we know $p_c \leq p$; conversely if p is such that $\mathbb{E}Z \to \infty$ and, say, $\mathbb{E}(Z^2) \leq C(\mathbb{E}Z)^2$, then (via the Paley–Zygmund Inequality and Friedgut's Theorem, see Sections **??** and **??**) we know $p_c \geq p$.

This approach, combined with smart ideas of conditioning, replacing Z with a related random variable, and solving difficult optimization problems to bound $\mathbb{E}(Z^2)$, has been able to pin down the satisfiability threshold in many random CSP's to within a small constant factor (e.g. [**?**, **?**, **?**, **?**]). These probabilistic methods face a barrier, however, and cannot address Conjecture **??** or determine p_c. Instead, new tools and ideas from mathematics and statistical physics have been brought to bear on the problem.

4.1 Friedgut's Theorem

Major progress towards Conjecture **??** was made by Friedgut in [**?**] who proved an n-dependent sharp threshold.

Theorem 12 (Friedgut). *For each $k \geq 3$ there is a function $c_k(n)$ bounded above and below by constants so that for every $\varepsilon > 0$ the following hold:*

- *If $p \leq (1-\varepsilon)c_k(n) \cdot n^{-1/(k-1)}$, then whp $F_k(n,p)$ is satisfiable.*

- *If $p \geq (1+\varepsilon)c_k(n) \cdot n^{-1/(k-1)}$, then whp $F_k(n,p)$ is unsatisfiable.*

This theorem shows that the scaling window of random k-SAT is $o(p_c)$ but it leaves open the possibility that the critical density c_k fluctuates as $n \to \infty$.

Though motivated by the Satisfiability Conjecture, Friedgut proved much more in [**?**]: he gave a very general characterization of which monotone properties of random graphs can have a coarse threshold: those that are well approximated by the property of containing some bounded-size subgraph from a bounded-size list of subgraphs. All other properties have a sharp threshold (in the non-uniform sense of Theorem **??**).

Returning to the example from Section **??**, Achlioptas and Friedgut [**?**] use [**?**] to show that for $q \geq 3$ the property of $G(n,p)$ being q-colorable exhibits a sharp threshold, in the sense of Theorem **??**. It remains open to show that there exists constants d_q so that a sharp threshold for q-colorability occurs at d_q/n.

4.2 The cavity method and the structure of solution spaces

Statistical physicists soon turned their attention to random k-SAT and related models using tools and intuition from the study of spin glasses [**?**, **?**, **?**, **?**, **?**, **?**]. The replica and cavity methods they developed
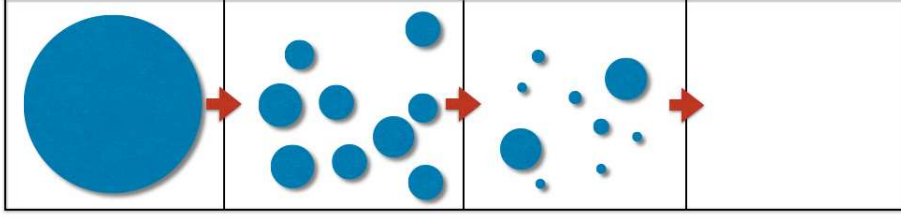
Figure 6: Cartoon of the evolution of the solution space for random k-SAT or random graph coloring as the edge density increases.

are powerful analytic tools for analyzing disordered (random) systems, based on assumptions that to a large extent have not been proved mathematically. Both are methods for computing $\mathbb{E} \log Z$, the expectation of a log partition function, the normalizing constant of a Gibbs measure[1]. The replica method is based on the identity $\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$, and rather magically using computations for large integer n to deduce the value of the limit. See [?, Chapter 1.13] for a mathematical introduction. The cavity method is based on a coupling of random graphs or hypergraphs on n vertices and on $n+1$ vertices, and understanding the expected change in the partition function, $\log Z_{n+1} - \log Z_n$, as a function of the Gibbs measure and its correlation properties. See [?, Chapter 19] for an introduction to the method. Under strong assumptions, these two methods are capable of making very detailed predictions for thresholds and phase transitions in a broad class of random structures, *Gibbs measures on random graphs*.

Applying these methods to random CSP's involves shifting perspective and instead of asking whether or not there is a solution, asking 'what does the uniform distribution on solutions (if they exist) look like?' Then the main properties of interest are *correlations* and *overlaps*. Correlations are measured, for instance, by the covariance in the indicator random variables that two variables take the value True (or two vertices take the color Red). Overlap is the random variable measuring the fraction of coordinates on which two independent samples from the uniform distribution on solutions agree.

Figure **??** depicts a cartoon of how the space of solutions in, say, random k-SAT or random graph q-coloring, changes (or is predicted to change) as the density of constraints increases. Three distinct thresholds are pictured. At low densities, solutions lie in one large component, connected under single variable changes. Then after the *shattering threshold* solutions break apart into exponentially many clusters of roughly equal exponential size, each separated by linear Hamming distance ($\Theta(n)$ variables must be changed to move between clusters). Next, at the *condensation threshold* a constant number of large clusters contain almost all solutions, while there are exponentially many smaller clusters. Finally, after the *satisfiability threshold*, no solutions remain.

The condensation threshold is critical for correlations and overlap. In the random graph coloring model, below the condensation threshold average correlations between vertices vanish as $n \to \infty$, and the overlap concentrates on $1/q$ (as it would if the graph were empty and we were sampling uniformly from $[q]^n$). Above the threshold, average correlations are bounded away from 0 and overlap concentrates on two points, $1/q$ and some $\eta > 1/q$, with the second corresponding to the case in which both samples come from the same dominant cluster. In the language of the cavity method, the model is *replica symmetric* below the condensation threshold and exhibits *replica symmetry breaking* above the threshold.

In terms of identifying the satisfiability threshold, condensation is important because it presents a fundamental obstacle to applying the second-moment method to the random variable Z counting

---

[1]Here the setting is always at positive temperature so $Z > 0$ and taking a logarithm is justified; models with hard constraints like random k-SAT or q-coloring are analyzed by taking a limit as the temperature tends to 0.

16

the number of solutions. We briefly describe the connection at a high level. Let d denote the average vertex or variable degree of a random CSP. The exponential order of the first moment, $f_1(d) := \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}Z$, is a decreasing function of d and the first-moment threshold d_{first} is the point at which it hits 0. For $d > d_{\text{first}}$ whp there are no solutions by Markov's Inequality; for $d < d_{\text{first}}$ the expected number of solutions is exponentially large. To show that there is at least one solution with probability bounded away from 0, we would like to apply the second-moment method via the Paley–Zygmund Inequality: $\Pr(Z > 0) \geq \frac{(\mathbb{E}Z)^2}{\mathbb{E}(Z^2)}$. Writing $\mathbb{E}(Z^2) = \sum_{\sigma, \tau} \Pr[\sigma, \tau \text{ both solutions}]$ one sees that on an exponential scale, the second moment is at most the square of the first moment if and only if the dominant contribution to $\mathbb{E}(Z^2)$ comes from pairs σ, τ of uncorrelated vectors. Past the condensation threshold, pairs of solutions with non-trivial correlation are significant and so the second moment should be expected to be exponentially larger than the square of the first moment. See the discussion in [?, ?, ?, ?, ?] for much more detail. To prove a random CSP is whp satisfiable beyond the condensation threshold requires a significantly different approach.

4.3   Proof of the satisfiability conjecture for large k

How can one locate the satisfiability threshold in light of condensation blocking the second-moment method? One very successful solution is to use the cavity method predictions to select a different random variable on which to perform the second-moment method. In particular, the predicted '1-RSB' behavior of both random k-SAT and random graph coloring suggests that while the uniform distribution over solutions exhibits replica symmetry breaking above the condensation threshold, the uniform distribution on *clusters* does not.

The innovation of Coja-Oghlan and Panagiotou in [?] was to design (for the closely related random k-NAE SAT model[2]) a random variable Z_β counting the number of solutions with βn 'blocked' variables that is a stand-in for counting the number of clusters of a given (exponential) size. The same authors apply this idea to random k-SAT in [?], additionally conditioning on typical vertex degrees. This kind of random variable was also used in [?, ?] to determine thresholds and limiting values of optimization problems in random regular models (in which each vertex or variable has the same edge or constraint degree).

In [?], Ding, Sly, and Sun used this cavity-method inspired second-moment argument along with conditioning on the typical empirical distribution of arbitrary depth neighborhoods of variables to resolve Conjecture ?? for k large enough.

Theorem 13 (Ding, Sly, Sun ??). *There exists $k_0 > 0$ so that for each $k \geq k_0$ there exists c_k so that the following hold.*

- *If $m \leq (1 - \varepsilon)c_k n$, then whp $F_k(n, m)$ is satisfiable.*

- *If $m \geq (1 + \varepsilon)c_k n$, then whp $F_k(n, m)$ is unsatisfiable.*

The proof is a mathematical tour-de-force, employing several sophisticated tools (for both the upper and lower bounds) and performing very challenging probabilistic and combinatorial calculations. Notably, to prove Conjecture ?? in this case they establish the exact threshold as predicted by the 1-RSB cavity method.

One could hope to apply the same strategy to random graph coloring, but this seems hopelessly complex: having $q > 2$ possibilities for each variable makes everything more complicated, and the proof in [?] is already over 300 pages.

---

[2]'not-all-equal' SAT: an assignment is satisfying if every clause has at least one true and at least one false literal.

4.4 Algorithmic thresholds

What about the other question asked by Mitchell, Selman, and Levesque – is it true that random k-SAT instances near the satisfiability threshold are computationally intractable?

There was some hope among statistical physicists that *survey propagation*, a message-passing algorithm based on the 1-RSB predictions for random k-SAT and other models, might provide an efficient search algorithm for instances near the satisfiability threshold [?]. However (at least for large k) this was disproved [?, ?].

Instead, evidence of computational hardness based on the structure of the solution space has emerged. Achlioptas and Coja-Oghlan proved the shattering of the solution space in [?] and observed that the threshold for shattering approximately coincides with the density above which no efficient search algorithms are known (though see below for a caveat in linking the two). While it did not establish a direct link between solution space structure and algorithms, this paper was innovative both in its techniques (using the *planted model*, described more below) and in making a conceptual link between the two.

A more direct link with algorithms is the Overlap Gap Property (OGP) pioneered by Gamarnik and Sudan [?, ?]. The basic OGP states that there is an interval (a, b) so that whp over an instance of a random computational problem, no pair of solutions have their normalized overlap in (a, b). This can be proved using a first-moment method. Amazingly, this simple property then implies that entire classes of algorithms (local algorithms, low-degree algorithms) cannot find a solution whp. For random k-SAT the OGP perspective has been applied to determine approximately the low-degree algorithmic threshold of the problem [?].

4.5 Scaling windows

With Friedgut's theorem establishing a sharp threshold in the random k-SAT model, one could hope to say something quantitatively stronger about scaling windows in this and related models. Friedgut's theorem (or Bourgain's more abstract theorem) gives bounds on the k-SAT scaling window of the form $T_\varepsilon = O(p_c / \log \log n)$, just barely enough for a sharp threshold. Surprisingly, Abbe and Montanari [?] show that even a very mild improvement to this bound would resolve Conjecture **??**

Proposition 14 ([?]). *If for any fixed $\delta > 0$, the scaling window for random k-SAT satisfies $T_\varepsilon = O(p_c / \log^{1+\delta} n)$, then Conjecture* **??** *holds.*

The statistical physics cavity method does not give much guidance on what to expect for scaling windows. In 2002, Wilson [?] used basic probabilistic arguments to prove generic polynomial lower bounds on the scaling window for random k-SAT and other models, thus disproving a number of conjectures from statistical physics. Good upper bounds for random k-SAT or random graph q-coloring are completely lacking currently.

In a different random CSP, however, scaling windows have recently been pinned down precisely. The *binary perceptron* [?] arose in the 1960's as a toy model of a neural network then attracted attention in statistical physics via the work of Gardner and others [?, ?, ?]. A symmetric variant was recently introduced in [?] and has some remarkably nice properties from a mathematical point of view. For one, the plain first- and second-moment methods (along with some concentration arguments) suffice to pin down precisely the satisfiability threshold in the model [?, ?, ?]. Closely related is the fact that the solution space looks very different than the cartoon in Figure **??**: for all constraint densities below the satisfiability threshold, whp over the instance, almost all solutions are isolated, at linear Hamming distance from the nearest other solution. The corresponding cartoon would be a sprinkling of points.

The nature of the solution space in this model raises a lot of questions: if almost all solutions are

isolated should the search problem be hard at all densities? In fact there are efficient search algorithms at low densities [**?**, **?**] even when almost all solutions are isolated; these algorithms in fact find solutions lying in very rare clusters (with maximum possible diameter) [**?**], as predicted by physicists working on machine learning problems [**?**]. These algorithms are 'multiscale majority algorithms' in which entries of a potential solution vector X are set round by round to optimize a greedy objective. The symmetric binary perceptron is a random model of the combinatorial discrepancy problem, and so constructive discrepancy minimization algorithms also provide provably efficient search algorithms at low constraint densities [**?**, **?**, **?**, **?**]. These algorithms can also be adapted for the original (non-symmetric) perceptron model [**?**]. The success of these algorithms indicates that the cartoon in Figure **??** is not really relevant to algorithms: the cartoon depicts properties of typical solutions, while efficient algorithms may indeed find rare solutions. The structural properties like the OGP that apply to *all* solutions (or tuples of solutions) are more algorithmically relevant.

Finally, in a recent breakthrough Altschuler proved something remarkable about the scaling window: the scaling window, measured in number of constraints, is of width $O(\log n)$ (while the critical number of constraints is $O(n)$) [**?**]. Contrast this to what is known via abstract results like Friedgut's theorem which would give a bound of $O(n/\log\log n)$; and to what is known for models like random k-SAT and random graph coloring which is only what the abstract results give. Even more recently Sah and Sawhney determined the scaling window completely, showing it is of width $O(1)$ and giving the limiting probability of satisfiability inside the window [**?**].

5   Statistical inference and the stochastic block model

A fundamental statistical question can be phrased as 'Under what circumstances can a signal be recovered from a noisy observation of that signal?' Or - 'Can we distinguish a signal from noise?'

As the study of statistics has evolved in the age of fast computers and massive data sets, these same questions remain fundamental but the data sets of interest are now very high dimensional and the question of efficient computation becomes paramount. Just as we want to know whether 'typical' instances of 3-SAT are algorithmically tractable we also want to know if 'typical' statistical inference problems are algorithmically tractable.

A very useful framework for the rigorous study of these questions is the *teacher–student* framework (see e.g., [**?**]). In this framework, a teacher describes a generative probabilistic model of data to a student. The model takes as an input a 'ground truth' and adds to this noise of some form. The teacher chooses a ground truth from some known prior distribution, generates the data from the model, and presents the data to the student. The student's task is to recover the ground truth from the data, using knowledge of the generative model and prior distribution.

Perhaps the most studied teacher–student model in the *stochastic block model*, a toy model for the statistical and machine learning task of clustering: partitioning a set of data points into subsets with similar characteristics. The stochastic block model deals with a simple specialization of clustering: community detection in which the the data is a graph and the task is to partition the vertex set into subsets of vertices with similar connectivity structure; this could mean finding a partition in which most edges lie within parts, or finding a partition in which most edges cross the parts.

Formally, a symmetric version of the stochastic block model is defined as follows. Fix an integer $q \geq 2$, and $p_{\text{in}}, p_{\text{out}} \in [0,1]$, $p \neq q$, and let n denote the number of vertices of the graph.

- Choose a partition $\sigma \in [q]^n$ of $[n]$ into q parts uniformly at random.

- For each pair $i, j \in [n]$, $i \neq j$, include the edge (i, j) with probability p_{in} if $\sigma(i) = \sigma(j)$ and with probability p_{out} if $\sigma(i) \neq \sigma(j)$, all edges independently of the others.

- Call the resulting graph $\hat{G}$.

If $p_{\text{in}} > p_{\text{out}}$, then, on average, more edges will be drawn between vertices with the same label (we say the model is *assortative*); if $p_{\text{in}} > p_{\text{out}}$, the opposite is true (the model is *disassortative*). If $p_{\text{in}} = p_{\text{out}}$ the model is simply $G(n, p)$.

The inference task is to *recover* the partition σ given the graph $\hat{G}$. Recovery can be defined in different ways: exact recovery (up to a permutation of the q labels); recovery of almost all the labels up to permutation; or recovery of a partition σ' that (after a permutation) agrees with σ on $\frac{1}{q} + \varepsilon$ fraction of vertices, for some small $\varepsilon > 0$; that is, just a tiny bit better than random guessing. We focus here on the last notion, called *weak recovery*.

Though we have left the setting of monotone properties of the hypercube under μ_p, we have retained some of the essential features: this model has independent edges and an important monotonicity property. If we fix the ratio $\frac{p_{\text{in}}}{p_{\text{out}}} =: \alpha$ and set $p_{\text{in}} = p\alpha$ and $p_{\text{out}} = p$ then the weak recovery problem only becomes *easier* as p increases: we have more data (observed edges) from which to deduce the signal (the underlying partition). That is, if $p > p'$ we can reduce the model with parameter p to that with parameter p' by independently deleting each edge with probability p'/p.

5.1 Sharp thresholds for inference

Though the stochastic block model was defined independently in different fields in the 1980's [?, ?, ?], mathematical interest in the model exploded in 2011 when Decelle, Krzakala, Moore, and Zdeborová [?] used the cavity method to make a series of beautiful predictions about sharp thresholds in the model.

We focus on their conjectures for thresholds for weak recovery when $p_{\text{in}}, p_{\text{out}} = O(1/n)$ and so the random graph has constant average degree.

Fix a number of parts $q \geq 2$ and a ratio $\alpha > 0$ of $p_{\text{in}}/p_{\text{out}}$. There is an *information theoretic threshold* at d_{inf} if

1. when $d < d_{\text{inf}}$, $p_{\text{in}} = \alpha\frac{d}{n}$, $p_{\text{out}} = \frac{d}{n}$, there is no algorithm (efficient or not) that whp finds a partition σ' that agrees with σ on $\frac{1}{q} + \varepsilon$ fraction of vertices (after a permutation of the parts) for any fixed $\varepsilon > 0$.

2. when $d > d_{\text{inf}}$, $p_{\text{in}} = \alpha\frac{d}{n}$, $p_{\text{out}} = \frac{d}{n}$, there is $\varepsilon > 0$ and an algorithm (perhaps inefficient) that whp finds a partition σ' that agrees with σ on $\frac{1}{q} + \varepsilon$ fraction of vertices (after a permutation of the parts).

There is an *algorithmic threshold* at d_{alg} if

1. when $d < d_{\text{alg}}$, $p_{\text{in}} = \alpha\frac{d}{n}$, $p_{\text{out}} = \frac{d}{n}$, there is no polynomial-time algorithm that whp finds a partition σ' that agrees with σ on $\frac{1}{q} + \varepsilon$ fraction of vertices (after a permutation of the parts) for any fixed $\varepsilon > 0$.

2. when $d > d_{\text{alg}}$, $p_{\text{in}} = \alpha\frac{d}{n}$, $p_{\text{out}} = \frac{d}{n}$, there is $\varepsilon > 0$ and a polynomial-time algorithm that whp finds a partition σ' that agrees with σ on $\frac{1}{q} + \varepsilon$ fraction of vertices (after a permutation of the parts).

For the case $q = 2, 3$ they conjectured that for any α, $d_{\text{inf}} = d_{\text{alg}}$, but for $q \geq 5$ they conjectured a gap: $d_{\text{inf}} < d_{\text{alg}}$ (now known as a statistical–computational gap, and the subject of great recent interest in statistics and computer science). In general the conjectured value d_{alg} is explicit (it is the 'Kestum–Stigum threshold' [?]), while the conjectured value of d_{inf} is given implicitly as the solution of a complicated optimization problem over probability measures. The $q = 2$ conjectures were proved in celebrated works of Mossel–Neeman–Sly [?, ?] and Massoulié [?].

For $q \geq 3$, partial progress has been made. Abbe and Sandon [?] proved that for any q and any α, weak recovery is possible above the Kestum–Stigum threshold, i.e. the positive part of the conjecture on the algorithmic threshold. For $q \geq 3$ and $\alpha < 1$ (the disassortative case), the conjecture on the value of the information–theoretic threshold (as the solution of the complicated optimization problem) was proved in [?] by implementing a rigorous version of the cavity method. Recently, it was proved in [?] that for $q = 3$ (and for $q = 4$ and α sufficiently close to 1), $d_{\mathrm{inf}} = d_{\mathrm{alg}}$.

5.2   The planted model

To wrap things up, we discuss a concept with connections to all of the developments recounted in Sections ??, ??, and ??.

The stochastic block model is an example of a *planted model*. A solution (in this case a partition into q parts) is chosen, then a random instance is drawn consistent with this *planted solution*. The inference task above is to recover information about the planted solution given the random instance. An even easier computation task is to distinguish the planted model from the purely random model (in this case the Erdős-Rényi random graph $G(n, p)$).

It is straightforward to devise planted models for random k-SAT and random graph coloring: pick a solution σ uniformly at random and choose independent constraints or edges among those that are satisfied by σ.

Achlioptas and Coja-Oghlan [?] show that if the number of solutions in the random model is sufficiently concentrated, then high probability results about the planted solution in the planted model can be transferred to high probability statements about uniformly random solutions in the random model. This is the key to establishing results on the structure of the solution space (and is used in the symmetric perceptron results [?, ?] as well).

Returning again to random graph coloring, the planted model is exactly the extreme case ($\alpha = 0$) of the disassortative stochastic block model. In [?, ?, ?] the condensation threshold for random graph q-coloring is determined precisely, in terms of a solution to a variational problem arising from the cavity method. This bound is the best known lower bound on the q-colorability threshold for $q \geq 4$ (the best bound for $q = 3$ is in [?]). In [?], a precise connection between condensation thresholds and information theoretic thresholds is made: the two thresholds coincide for a large class of models exhibiting some symmetry and convexity; these include random graph coloring, the anti-ferromagnetic Potts model, k-NAE-SAT. They do not include random k-SAT (asymmetry) or the assortative stochastic block model (lacking the needed convexity). Overcoming these technical obstacles and being able to identify condensation thresholds in asymmetric models is a major challenge and could eventually lead to a generic approach to Conjecture ??. See related discussion in [?, ?].

To conclude, somewhat surprisingly the planted model is closely related to the developments recounted in Section ?? as well. Mossel, Niles-Weed, Sun and Zadik [?] recently showed that a well-chosen planted model and a second-moment argument can be used to prove the 'spread lemma', the key technical ingredient in the proof of improved sunflower bounds [?] and the fractional Kahn–Kalai conjecture [?].

6   Conclusions and questions

As we saw in Section ??, the fractional Kahn–Kalai theorem has sufficed for all known applications, but one can ask if there are applications that require Theorem ??.

Question 1. *Are there threshold applications that need the full power of Theorem ?? rather than the fractional version of Theorem ??? Or is there a constant K so that $q_f(\mathcal{F}) \leq Kq(\mathcal{F})$?*

The second is conjectured in the affirmative by Talagrand in [**?**].

As we have seen, one effective way of determining sharp thresholds and scaling windows is to prove a hitting time result, relating a complex property to a simple property.

Question 2. *Is there a general way to find first-order asymptotics of p_c for the property of $G(n,p)$ containing a subgraph H when there is not a natural hitting-time conjecture?*

A nice conjecture for one such property is by Kahn, Narayanan, and Park on the threshold for the existence of the square of a Hamilton cycles in $G(n,p)$; that is, the existence of a cyclic ordering of the n vertices so that there is an edge between every pair of neighbors and second neighbors.

Conjecture 15 (Kahn, Narayanan, and Park [**?**]). *For the property $\mathcal{F}$ of $G(n,p)$ containing the square of a Hamilton cycle,*

$$p_c(\mathcal{F}) = (1 + o(1))\sqrt{\frac{e}{n}}.$$

In [**?**] they prove $p_c = \Theta(n^{-1/2})$ using a finer understanding of the spreadness of the uniform distribution on squares of Hamilton cycles; a direct application of Theorem **??** would have lost a $\log n$ factor. See also [**?**, **?**] for generalizations.

Question 3. *Are there any general-purpose tools for bounding the width of the scaling window of a monotone property with a sharp threshold, beyond the bounds given by Friedgut's theorem? Can Conjecture **??** be proved in a generic way, without the need to precisely identify the threshold?*

Random geometric graphs

We conclude the survey by stating one of the author's favorite open problems on thresholds in random structures.

In this survey we have discussed the Erdős-Rényi random graph in depth and another model of a random graph in the stochastic block model; both models have the important property that edges are independent. There are many other important random graph models that do not have this property. Here we discuss one, the *random geometric graph* [**?**, **?**], in which edges between randomly placed points are determined geometrically.

In particular, let $\mathbb{S}^{d-1}$ be the unit sphere in d-dimensions[3]. The random graph $G_d(n,p) = (V, E)$ is formed as follows:

- Let $V = [n]$

- Select n points $x_1, \ldots, x_n$ independently and uniformly from $\mathbb{S}^{d-1}$.

- Let $E = \{(i,j) : x_i \cdot x_j \geq \tau_p\}$ where τ_p is chosen so that $\Pr[x_i \cdot x_j \geq \tau_p] = p$.

In particular, like the Erdős-Rényi random graph $G(n,p)$, $G_d(n,p)$ is a random graph on n vertices with edge probability p; but now the edges are not independent.

Thresholds in a random geometric graphs have been studied extensively; see the textbook of Penrose [**?**] for results on connectivity, subgraph containment, existence of a giant component. More recently, following the influential paper of Devroye, György, Lugosi, and Udina [**?**], thresholds in $G_d(n,p)$ when $d = d(n) \to \infty$ have been studied, along with a different kind of threshold: for a given $p(n)$ for which values of $d(n)$ are the two random graphs $G(n,p)$ and $G_d(n,p)$ statistically distinguishable? [**?**, **?**].

---

[3]We alternatively could choose the underlying space to be the d-dimensional unit torus, $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, the results and questions below would still apply.

Just as in $G(n, p)$, a general study of monotone properties is possible in $G_d(n, p)$. In particular, let $\mathcal{F}$ be a non-trivial monotone property (where, as always, monotone pertains to adding edges). Then again the probability that $\mathcal{F}$ holds in $G_d(n, p)$ is a strictly increasing function of p, and so $p_c(\mathcal{F})$ can be uniquely defined, and all of the same questions from Section **??** can be asked in this setting too.

Perhaps surprisingly then, the analogue of Theorem **??**, the Bollobás-Thomason theorem, that lays the foundation for the general study of thresholds of monotone properties in random graphs, is not known in general for random geometric graphs. In the special case of $d = 1$ (points on a circle), McColm proved that every monotone property has a threshold [**?**]. Moreover, Goel, Rai, and Krishnamachari [**?**] proved that the scaling window of every monotone property is bounded by a function (depending on d) that vanishes as $n \to \infty$, the analogue of the result of Friedgut and Kalai [**?**] for $G(n, p)$. The general statement, however, remains open.

Conjecture 16. *Every non-trivial monotone property has a threshold in $G_d(n, p)$.*

Acknowledgements

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] E. Abbe, S. Li, and A. Sly. Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 860–873, 2022.

[3] E. Abbe, S. Li, and A. Sly. Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 327–338. IEEE, 2022.

[4] E. Abbe and A. Montanari. On the concentration of the number of solutions of random satisfiability formulas. *Random Structures & Algorithms*, 45(3):362–382, 2014.

[5] E. Abbe and C. Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 71(7):1334–1406, 2018.

[6] D. Achlioptas and A. Coja-Oghlan. Algorithmic barriers from phase transitions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 793–802. IEEE, 2008.

[7] D. Achlioptas and E. Friedgut. A sharp threshold for k-colorability. *Random Structures & Algorithms*, 14(1):63–70, 1999.

[8] D. Achlioptas and C. Moore. Almost all graphs with average degree 4 are 3-colorable. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 199–208, 2002.

[9] D. Achlioptas and C. Moore. The asymptotic order of the random k-SAT threshold. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 779–788. IEEE, 2002.

[10] D. Achlioptas and C. Moore. Random k-SAT: Two moments suffice to cross a sharp threshold. *SIAM Journal on Computing*, 36(3):740–762, 2006.

[11] D. Achlioptas, A. Naor, and Y. Peres. Rigorous location of phase transitions in hard optimization problems. *Nature*, 435(7043):759–764, 2005.

[12] D. Achlioptas and Y. Peres. The threshold for random k-SAT is $2^k \ln 2 - O(k)$. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 223–231, 2003.

[13] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[14] D. J. Altschuler. Critical window of the symmetric perceptron. *Electronic Journal of Probability*, 28:1–28, 2023.

[15] R. Alweiss, S. Lovett, K. Wu, and J. Zhang. Improved bounds for the sunflower lemma. *Annals of Mathematics*, 194(3):795–815, 2021.

[16] M. Anastos and D. Chakraborti. Robust Hamiltonicity in families of Dirac graphs. *arXiv preprint arXiv:2309.12607*, 2023.

[17] L. J. Andrén, C. J. Casselgren, and L.-D. Öhman. Avoiding arrays of odd order by Latin squares. *Combinatorics, Probability and Computing*, 22(2):184–212, 2013.

[18] D. Applegate, R. Bixby, V. Chvátal, and W. Cook. On the solution of traveling salesman problems. *Documenta Mathematica*, pages 645–656, 1998.

[19] B. Aubin, W. Perkins, and L. Zdeborová. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 2019.

[20] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Physical Review Letters*, 115(12):128101, 2015.

[21] N. Bansal and J. H. Spencer. On–line balancing of random inputs. *Random Structures & Algorithms*, 57(4):879–891, 2020.

[22] V. Bapst, A. Coja-Oghlan, S. Hetterich, F. Raßmann, and D. Vilenchik. The condensation phase transition in random graph coloring. *Communications in Mathematical Physics*, 341:543–606, 2016.

[23] M. Ben-Or and N. Linial. Collective coin flipping, robust voting schemes and minima of Banzhaf values. In *26th Annual Symposium on Foundations of Computer Science (sfcs 1985)*, pages 408–416. IEEE, 1985.

[24] A. Bogdanov, L. Trevisan, et al. Average-case complexity. *Foundations and Trends® in Theoretical Computer Science*, 2(1):1–106, 2006.

[25] B. Bollobás. *Random graphs*. Springer, 1998.

[26] B. Bollobás, C. Borgs, J. T. Chayes, J. H. Kim, and D. B. Wilson. The scaling window of the 2-SAT transition. *Random Structures & Algorithms*, 18(3):201–256, 2001.

[27] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press, 2006.

[28] B. Bollobás and A. Thomason. Random graphs of small order. In *North-Holland Mathematics Studies*, volume 118, pages 47–97. Elsevier, 1985.

[29] B. Bollobás and A. G. Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987.

[30] A. Braunstein, M. Mézard, and R. Zecchina. Survey propagation: An algorithm for satisfiability. *Random Structures & Algorithms*, 27(2):201–226, 2005.

[31] G. Bresler and B. Huang. The algorithmic phase transition of random k-SAT for low degree polynomials. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 298–309. IEEE, 2022.

[32] S. Bubeck, J. Ding, R. Eldan, and M. Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

[33] T. N. Bui, S. Chaudhuri, F. T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7:171–191, 1987.

[34] C. J. Casselgren and R. Häggkvist. Coloring complete and complete bipartite graphs from random lists. *Graphs and Combinatorics*, 32:533–542, 2016.

[35] V. Chvátal, W. Cook, G. B. Dantzig, D. R. Fulkerson, and S. M. Johnson. Solution of a large-scale traveling-salesman problem. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, pages 7–28, 2010.

[36] V. Chvátal and B. Reed. Mick gets some (the odds are on his side) (satisfiability). In *Proceedings., 33rd Annual Symposium on Foundations of Computer Science*, pages 620–627. IEEE Computer Society, 1992.

[37] A. Coja-Oghlan, C. Efthymiou, N. Jaafari, M. Kang, and T. Kapetanopoulos. Charting the replica symmetric phase. *Communications in Mathematical Physics*, 359(2):603–698, 2018.

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

[39] A. Coja-Oghlan and K. Panagiotou. The asymptotic k-sat threshold. *Advances in Mathematics*, 100(288):985–1068, 2016.

[40] A. Coja-Oghlan and L. Zdeborová. The condensation transition in random hypergraph 2-coloring. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 241–250. SIAM, 2012.

[41] A. Coja-Oglan and K. Panagiotou. Catching the k-NAESAT threshold. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 899–908, 2012.

[42] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, (3):326–334, 1965.

[43] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical review E*, 84(6):066106, 2011.

[44] L. Devroye, A. György, G. Lugosi, and F. Udina. High-dimensional random geometric graphs and their clique number. *Electronic Journal of Probability*, 16:2481–2508, 2011.

[45] J. Ding, A. Sly, and N. Sun. Maximum independent sets on random regular graphs. *Acta Math*, 217:263–340, 2016.

[46] J. Ding, A. Sly, and N. Sun. Satisfiability threshold for random regular NAE-SAT. *Communications in Mathematical Physics*, 341(2), 2016.

[47] J. Ding, A. Sly, and N. Sun. Proof of the satisfiability conjecture for large k. *Annals of Mathematics*, 196(1):1–388, 2022.

[48] H. Duminil-Copin. Sharp threshold phenomena in statistical physics. *Japanese Journal of Mathematics*, 14:1–25, 2019.

[49] M. E. Dyer and A. M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

[50] R. Eldan and M. Singh. Efficient algorithms for discrepancy minimization in convex sets. *Random Structures & Algorithms*, 53(2):289–307, 2018.

[51] P. Erdős. On the combinatorial problems which I would most like to see solved. *Combinatorica*, 1(1):25–42, 1981.

[52] P. Erdős and A. Rényi. On random graphs I. *Publ. math. debrecen*, 6(290-297):18, 1959.

[53] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

[54] P. Erdős and A. Rényi. On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hungar*, 17:359–368, 1966.

[55] A. Espuny Diaz and Y. Person. Spanning f-cycles in random graphs. *Combinatorics, Probability and Computing*, 32(5):833–850, 2023.

[56] K. Frankston, J. Kahn, B. Narayanan, and J. Park. Thresholds versus fractional expectation-thresholds. *Annals of Mathematics*, 194(2):475–495, 2021.

[57] E. Friedgut. Sharp thresholds of graph properties, and the k-SAT problem. *Journal of the American mathematical Society*, 12(4):1017–1054. Appendix by J. Bourgain, 1999.

[58] E. Friedgut. Hunting for sharp thresholds. *Random Structures & Algorithms*, 26(1-2):37–51, 2005.

[59] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American mathematical Society*, 124(10):2993–3002, 1996.

[60] E. Friedgut and M. Krivelevich. Sharp thresholds for certain Ramsey properties of random graphs. *Random Structures & Algorithms*, 17(1):1–19, 2000.

[61] A. Frieze and M. Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[62] D. Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021.

[63] D. Gamarnik and M. Sudan. Limits of local algorithms over sparse random graphs. *The Annals of Probability*, pages 2353–2376, 2017.

[64] V. Ganesh and M. Y. Vardi. On the unreasonable effectiveness of SAT solvers. In T. Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 547–566. Cambridge University Press, 2020.

[65] E. Gardner. Maximum storage capacity in neural networks. *Europhysics Letters*, 4(4):481, 1987.

[66] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[67] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[68] E. N. Gilbert. Random plane networks. *Journal of the society for industrial and applied mathematics*, 9(4):533–543, 1961.

[69] A. Goel, S. Rai, and B. Krishnamachari. Monotone properties of random geometric graphs have sharp thresholds. *Annals of Applied Probability*, pages 2535–2552, 2005.

[70] G. Grimmett and G. Grimmett. *What is percolation?* Springer, 1999.

[71] H. Hatami. A structure theorem for Boolean functions with small total influences. *Annals of Mathematics*, pages 509–533, 2012.

[72] A. Heckel. Random triangles in random graphs. *Random Structures & Algorithms*, 59(4):616–621, 2021.

[73] A. Heckel, M. Kaufmann, N. Müller, and M. Pasch. The hitting time of clique factors. *arXiv preprint arXiv:2302.08340*, 2023.

[74] S. Hetterich. Analysing survey propagation guided decimation on random formulas. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[75] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[76] V. Jain and H. T. Pham. Optimal thresholds for Latin squares, Steiner triple systems, and edge colorings. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1425–1436. SIAM, 2024.

[77] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*. John Wiley & Sons, 2011.

[78] A. Johansson. Triangle factors in random graphs, 2006.

[79] A. Johansson, J. Kahn, and V. Vu. Factors in random graphs. *Random Structures & Algorithms*, 33(1):1–28, 2008.

[80] J. Kahn. Hitting times for Shamir's problem. *Transactions of the American Mathematical Society*, 375(1):627–668, 2022.

[81] J. Kahn. Asymptotics for Shamir's problem. *Advances in Mathematics*, 422:109019, 2023.

[82] J. Kahn and G. Kalai. Thresholds and expectation thresholds. *Combinatorics, Probability and Computing*, 16(3):495–502, 2007.

[83] J. Kahn, B. Narayanan, and J. Park. The threshold for the square of a Hamilton cycle. *Proceedings of the American Mathematical Society*, 149(8):3201–3208, 2021.

[84] G. Kalai and S. Safra. Threshold phenomena and influence with some perspectives from mathematics, computer science, and economics. *Computational complexity and statistical physics*, page 25, 2006.

[85] D. Y. Kang, T. Kelly, D. Kühn, A. Methuku, and D. Osthus. Thresholds for Latin squares and Steiner triple systems: Bounds within a logarithmic factor. *Transactions of the American Mathematical Society*, 2023.

[86] P. Keevash. The optimal edge-colouring threshold. *arXiv preprint arXiv:2212.04397*, 2022.

[87] T. Kelly, A. Müyesser, and A. Pokrovskiy. Optimal spread for spanning subgraphs of Dirac hypergraphs. *arXiv preprint arXiv:2308.08535*, 2023.

[88] H. Kesten. *Percolation theory for mathematicians*, volume 2. Springer, 1982.

[89] H. Kesten and B. P. Stigum. A limit theorem for multidimensional Galton-Watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.

[90] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

[91] J. H. Kim and J. R. Roche. Covering cubes by random half cubes, with applications to binary neural networks. *Journal of Computer and System Sciences*, 56(2):223–252, 1998.

[92] A. Kireeva and J.-C. Mourrat. Breakdown of a concavity property of mutual information for non-Gaussian channels. *Information and Inference: A Journal of the IMA*, 13(2):iaae008, 2024.

[93] L. M. Kirousis, E. Kranakis, D. Krizanc, and Y. C. Stamatiou. Approximating the unsatisfiability threshold of random formulas. *Random Structures & Algorithms*, 12(3):253–269, 1998.

[94] J. Komlós and E. Szemerédi. Limit distribution for the existence of Hamiltonian cycles in a random graph. *Discrete mathematics*, 43(1):55–63, 1983.

[95] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066, 1989.

[96] F. Krzakała, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

[97] S. Li, T. Schramm, and K. Zhou. Discrepancy algorithms for the binary perceptron. *arXiv preprint arXiv:2408.00796*, 2024.

[98] S. Liu, S. Mohanty, T. Schramm, and E. Yang. Testing thresholds for high-dimensional sparse random geometric graphs. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 672–677, 2022.

[99] S. Lovett and R. Meka. Constructive discrepancy minimization by walking on the edges. *SIAM Journal on Computing*, 44(5):1573–1582, 2015.

[100] Z. Luria and M. Simkin. On the threshold problem for Latin boxes. *Random Structures & Algorithms*, 55(4):926–949, 2019.

[101] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.

[102] G. L. McColm. Threshold functions for random graphs on a line segment. *Combinatorics, Probability and Computing*, 13(3):373–387, 2004.

[103] S. Mertens, M. Mézard, and R. Zecchina. Threshold values of random k-SAT from the cavity method. *Random Structures & Algorithms*, 28(3):340–373, 2006.

[104] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[105] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[106] M. Mézard, G. Parisi, and R. Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[107] D. Mitchell, B. Selman, H. Levesque, et al. Hard and easy distributions of SAT problems. In *Aaai*, volume 92, pages 459–465, 1992.

[108] R. Monasson and R. Zecchina. Statistical mechanics of the random k-satisfiability model. *Physical Review E*, 56(2):1357, 1997.

[109] A. Montanari, F. Ricci-Tersenghi, and G. Semerjian. Clusters of solutions and replica symmetry breaking in random k-satisfiability. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(04):P04004, 2008.

[110] R. Montgomery. Spanning trees in random graphs. *Advances in Mathematics*, 356:106793, 2019.

[111] C. Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *Bulletin of EATCS*, 1(121), 2017.

[112] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015.

[113] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[114] E. Mossel, J. Niles-Weed, N. Sun, and I. Zadik. A second moment proof of the spread lemma. *arXiv preprint arXiv:2209.11347*, 2022.

[115] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 21–30. IEEE, 2005.

[116] E. Mossel, A. Sly, and Y. Sohn. Exact phase transitions for stochastic block models and reconstruction on trees. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 96–102, 2023.

[117] R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[118] J. Park. Threshold phenomena for random discrete structures. *Notices of the American Mathematical Society*, 70(10):1615–1625, 2023.

[119] J. Park and H. Pham. A proof of the Kahn–Kalai conjecture. *Journal of the American Mathematical Society*, 37(1):235–243, 2024.

[120] M. Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

[121] W. Perkins and C. Xu. Frozen 1-RSB structure of the symmetric Ising perceptron. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1579–1588, 2021.

[122] C. Pohoata, L. Sauermann, and D. Zakharov. Sharp bounds for rainbow matchings in hypergraphs. *arXiv preprint arXiv:2212.07580*, 2022.

[123] A. Rao. Sunflowers: from soil to oil. *Bulletin of the American Mathematical Society*, 60(1):29–38, 2023.

[124] O. Riordan. Random cliques in random graphs and sharp thresholds for F-factors. *Random Structures & Algorithms*, 61(4):619–637, 2022.

[125] T. Rothvoss. Constructive discrepancy minimization for convex sets. *SIAM Journal on Computing*, 46(1):224–234, 2017.

[126] A. Sah and M. Sawhney. Distribution of the threshold for the symmetric perceptron. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2369–2382. IEEE, 2023.

[127] A. Sah, M. Sawhney, and M. Simkin. Threshold for Steiner triple systems. *Geometric and Functional Analysis*, pages 1–32, 2023.

[128] J. Schmidt and E. Shamir. A threshold for perfect matchings in random d-pure hypergraphs. *Discrete mathematics*, 45(2-3):287–295, 1983.

[129] M. Simkin. $(n, k, k-1)$-Steiner systems in random hypergraphs. *arXiv preprint arXiv:1711.01975*, 2017.

[130] S. Spiro. A smoother notion of spread hypergraphs. *Combinatorics, Probability and Computing*, 32(5):809–818, 2023.

[131] M. Talagrand. Are all sets of positive measure essentially convex? In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1992–94*, pages 295–310. Springer, 1995.

[132] M. Talagrand. Selector processes on classes of sets. *Probability theory and related fields*, 135:471–486, 2006.

[133] M. Talagrand. Are many small sets explicitly small? In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 13–36, 2010.

[134] M. Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[135] A. Wigderson. P, NP and mathematics–a computational complexity perspective. In *Proceedings of the ICM*, volume 6, pages 665–712, 2006.

[136] D. B. Wilson. On the critical exponents of random k-SAT. *Random Structures & Algorithms*, 21(2):182–195, 2002.

[137] Q. Yu and Y. Polyanskiy. Ising model on locally tree-like graphs: Uniqueness of solutions to cavity equations. *IEEE Transactions on Information Theory*, 2023.

[138] L. Zdeborová and F. Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.