Threshold Phenomena for Random Discrete Structures

Jinyoung Park

1. Erdős–Rényi Model

To begin, we briefly introduce a model of random graphs. Recall that a graph is a mathematical structure that consists of vertices (nodes) and edges.

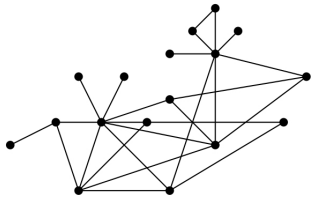


Figure 1. A graph.

Roughly speaking, a random graph in this article means that, given a vertex set, the existence of each potential edge is decided at random. We will specifically focus on the

Jinyoung Park is an assistant professor of mathematics at Courant Institute of Mathematical Sciences, NYU. Her email address is jinyoungpark@nyu.edu.
Her research is supported by NSF grant DMS-2153844.

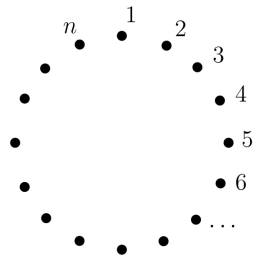
Communicated by Notices *Associate Editor Emilie Purvine.*

For permission to reprint this article, please contact: reprint-permission@ams.org.

DOI: https://doi.org/10.1090/noti2802

Erdős–Rényi random graph (denoted by $G_{n,p}$), which is defined as follows.

Consider n vertices that are labelled from 1 to n.



Observe that on those n vertices, there are potentially $\binom{n}{2}$ edges, that is, the edges labelled $\{1,2\},\{1,3\},\dots,\{n-1,n\}$. Given a probability $p \in [0,1]$, include each of the $\binom{n}{2}$ potential edges with probability p, where the choice of each edge is made independently from the choices of the other edges.

Example 1.1. As a toy example of the Erdős–Rényi random graph, let's think about what $G_{n,p}$ looks like when $n = 3$ and the value of p varies. First, if $p = 1/2$, then $G_{n,p}$ has the probability distribution as in Figure 2, defined on the collection of eight graphs. Observe that each graph is equally likely (since each potential edge is present with probability 1/2 independently).

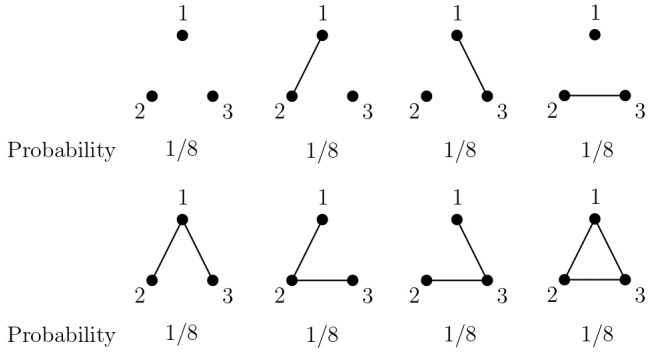


Figure 2. $G_{3,1/2}$.

Of course, we will have a different probability distribution if we change the value of p. For example, if p is closer to 0, say 0.01, then $G_{n,p}$ has the distribution as in Figure 3, where sparser graphs are more likely (as expected). On the other hand, if p is closer to 1, then denser graphs will be more likely.

In reality, when we consider $G_{n,p}$, n is a large (yet finite) number that tends to infinity, and $p = p(n)$ is usually a function of n that tends to zero as $n \to \infty$. For example, $p = 1/n$, $p = \log n/n$, etc.

As we saw in Example 1.1, a random graph is a random variable with a certain probability distribution (as opposed to a fixed graph) that depends on the values of n and p. Assuming n is given, the structure of $G_{n,p}$ changes as the value of p changes, and in order to understand $G_{n,p}$, we ask questions about the structure of $G_{n,p}$ such as

What's the probability that $G_{n,p}$ is connected?

or

What's the probability that $G_{n,p}$ is planar?

Basically, for *any* property $\mathcal{F}(= \mathcal{F}_n)$ of interest, we can ask

What's the probability that $G_{n,p}$ satisfies property $\mathcal{F}$?

In those questions, usually we are interested in understanding the *typical* structure/behavior of $G_{n,p}$. Observe that, unless $p = 0$ or 1, there is always a positive probability that all of the edges in $G_{n,p}$ are absent, or all of them are present (see Examples 1.2, 1.3). But in this article, we would rather ignore such extreme events that happen with a tiny probability, and focus on properties that $G_{n,p}$ possesses with a probability close to 1.

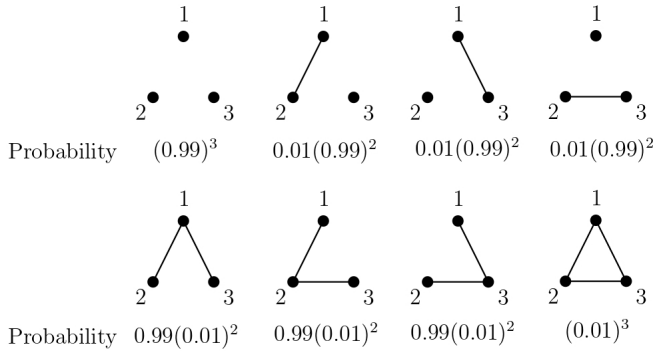
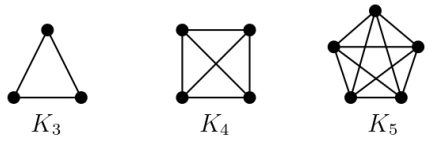


Figure 3. $G_{3,0.01}$.

We often use languages and tools from probability theory to describe/understand behaviors of $G_{n,p}$. Below we discuss some very basic examples.

We will write $f(n) \ll g(n)$ if $\frac{f(n)}{g(n)} \to 0$ as $n \to \infty$.

Example 1.2. One important object in graph theory is the *complete graph*, a graph with all the potential edges present. The complete graph on n vertices is denoted by K_n.



We can easily imagine that, unless p is very close to 1, it is extremely unlikely that $G_{n,p}$ is complete. Indeed,

$$\mathbb{P}(G_{n,p} = K_n) = p^{\binom{n}{2}}$$

(since we want all the edges present), which tends to 0 unless $1 - p$ is of order at most n^{-2}.

Example 1.3. Similarly, we can compute the probability that $G_{n,p}$ is "empty" (let's denote this by $\emptyset$) meaning that

no edges are present.[1] The probability for this event is

$$\mathbb{P}(G_{n,p} = \emptyset) = (1-p)^{\binom{n}{2}}.$$

When p is small, $1-p$ is approximately e^{-p}, so the above computation tells us that

$$\mathbb{P}(G_{n,p} = \emptyset) \to \begin{cases} 0 & \text{if } p \gg 1/n^2; \\ 1 & \text{if } p \ll 1/n^2. \end{cases}$$

Example 1.4. How many edges does $G_{n,p}$ typically have? The natural first step to answer this question is computing the *expected* number of edges in $G_{n,p}$. Using *linearity of expectation*,

$$\begin{aligned} &\mathbb{E}[\text{number of edges in } G_{n,p}] \\ &= \sum_{i<j} \mathbb{P}(\text{edge } \{i,j\} \text{ is present in } G_{n,p}) \\ &= (\text{number of edges in } K_n) \times \mathbb{P}(\text{each edge is present}) \\ &= \binom{n}{2} p. \end{aligned}$$

Remark 1.5. For example, if $p = 1/n$, then the expected number of edges in $G_{n,p}$ is $\frac{n-1}{2}$. But does this really imply that $G_{n,1/n}$ *typically* has about $\frac{n-1}{2}$ edges? The answer to this question is related to the fascinating topic of "concentration of a probability measure." We will very briefly discuss this topic in Example 2.5.

Example 1.6. Similarly, we can compute the expected number of *triangles* (the complete graph K_3) in $G_{n,p}$.

$$\begin{aligned} &\mathbb{E}[\text{number of triangles in } G_{n,p}] \\ &= \sum_{i<j<k} \mathbb{P}(\text{triangle } \{i,j,k\} \text{ is present in } G_{n,p}) \\ &= (\text{number of triangles in } K_n) \times \mathbb{P}(\text{each triangle is present}) \\ &= \binom{n}{3} p^3. \end{aligned}$$

The above computation tells us that

$$\mathbb{E}[\text{number of triangles in } G_{n,p}] \to \begin{cases} 0 & \text{if } p \ll 1/n; \\ \infty & \text{if } p \gg 1/n, \end{cases}$$

from which we can conclude that $G_{n,p}$ is typically triangle-free if $p \ll 1/n$. (If the expectation tends to 0, then there is little chance that $G_{n,p}$ contains a triangle.)

Remark 1.7. On the contrary, we *cannot* conclude that $G_{n,p}$ *typically* contains many triangles for $p \gg 1/n$ from the above expectation computation. Just think about a lottery of the prize money 10^{1000} dollars with the chance of winning 10^{-100}, to see that a large expectation does not necessarily imply a high chance of the occurence of an event. In general, showing that a desired structure typically *exists* in $G_{n,p}$ is a very challenging task, and this became a motivation for the *Kahn–Kalai conjecture* that we will discuss in the latter sections.

[1] *By the definition, $G_{n,p}$ has n vertices as a default.*

2. Threshold Phenomena

One striking thing about $G_{n,p}$ is that appearance and disappearance of certain properties are abrupt. Probably one of the most well-known examples that exhibit threshold phenomena of $G_{n,p}$ is the appearance of the *giant component*. A *component* of a graph is a maximal connected subgraph. For example, the graph in Figure 4 consists of four components, and the size (the number of vertices) of each component is 1, 2, 6, and 8.

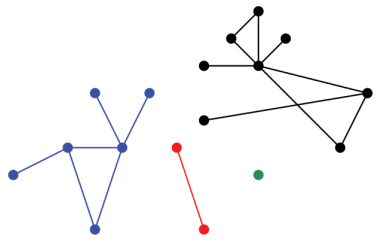


Figure 4. A graph that consists of four components.

For $G_{n,p}$, observe that, when $p = 0$, the size of a largest component of $G_{n,p}$ is 1; in this case all of the edges are absent with probability 1, so each of the components is an isolated vertex. On the other hand, when $p = 1$, $G_{n,p}$ is the complete graph with probability 1, so the size of its largest component is n.

Figure 5. $G_{n,0}$ and $G_{n,1}$.

Then what if p is strictly between 0 and 1?

Question 2.1. What's the (typical) size of a largest component in $G_{n,p}$?

Of course, one would naturally guess that as p increases from 0 to 1, the typical size of a largest component in $G_{n,p}$ would also increase from 1 to n. But what is really interesting here is that there is a "sudden jump" in this increment.

In the following statement and everywhere else, *with high probability* means that the probability that the event under consideration occurs tends to 1 as $n \to \infty$.

Theorem 2.2 (Erdős–Rényi [6]). *For any $\varepsilon > 0$, the size of a largest component of $G_{n,p}$ is*

$$\begin{cases} \le C_1(\varepsilon) \log n & \text{if } np < 1-\varepsilon \\ \ge C_2(\varepsilon) n & \text{if } np > 1+\varepsilon \end{cases}$$

with high probability, where $C_1(\varepsilon), C_2(\varepsilon)$ depend only on ε.

The above theorem says that if p is "slightly smaller" than $\frac{1}{n}$, then typically all of the components of $G_{n,p}$ are very small (note that $\log n$ is much smaller than the number of vertices, n).

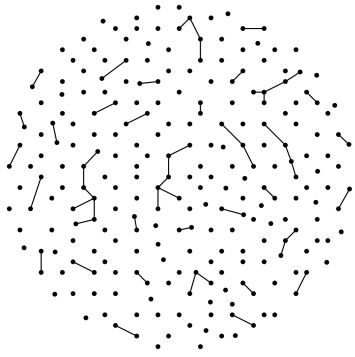


Figure 6. $G_{n,p}$ with all components small ($np < 1-\varepsilon$).

On the other hand, if p is "slightly larger" than $\frac{1}{n}$, then the size of a largest component of $G_{n,p}$ is as large as linear in n. It is also well-known that all other components are very small (at most of order $\log n$), and this unique largest component is called the *giant component*.

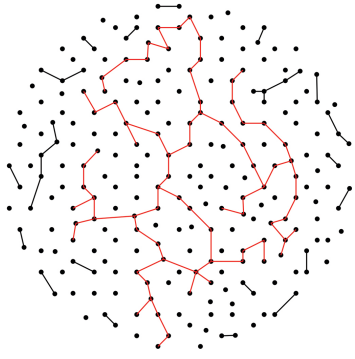


Figure 7. $G_{n,p}$ with the giant component ($np > 1-\varepsilon$).

So around the value $p = \frac{1}{n}$, the giant component "suddenly" appears, and therefore the structure of $G_{n,p}$ also drasitically changes. This is one example of the *threshold phenomena* that $G_{n,p}$ exhibits, and the value $p = \frac{1}{n}$ is a *threshold function* for $G_{n,p}$ of having the giant component. (The formal definition of a threshold function is given in Definition 2.3. See also the definition of *the threshold* in Section 5.)

The abrupt appearance of the giant component of $G_{n,p}$ is just one instance of vast threshold phenomena for *random discrete structures*. In this article, we will mostly deal with $G_{n,p}$ for the sake of concreteness, but there will be a brief discussion about a more general setting in Section 5.

Now we introduce the formal definition of a threshold function due to Erdős and Rényi. Recall that, in $G_{n,p}$, all the vertices are labelled $1, \dots, n$. A *graph property* is a property that is invariant under graph isomorphisms (i.e., relabelling the vertices), such as {connected}, {planar}, {triangle-free}, etc. We use $\mathcal{F}(= \mathcal{F}_n)$ for a graph property, and $G_{n,p} \in \mathcal{F}$ denotes that $G_{n,p}$ has property $\mathcal{F}$.

Definition 2.3. Given a graph property $\mathcal{F}(= \mathcal{F}_n)$, we say that $p_0 = p_0(n)$ is a *threshold function*[2] (or simply a *threshold*) for $\mathcal{F}$ if

$$\mathbb{P}(G_{n,p} \in \mathcal{F}) \to \begin{cases} 0 & \text{if } p \ll p_0 \\ 1 & \text{if } p \gg p_0. \end{cases}$$

For example, $p_0 = \frac{1}{n}$ is a threshold function for the existence of the giant component.

Note that it is not obvious at all whether a given graph property would admit a threshold function. Erdős and Rényi proved that many graph properties have a threshold function, and about 20 years later, Bollobás and Thomason proved that, in fact, there is a wide class of properties that admit a threshold function. In what follows, an *increasing (graph) property* is a property that is preserved under addition of edges. For example, connectivity is an increasing property, because if a graph is connected then it remains connected no matter what edges are additionally added.

Theorem 2.4 (Bollobas–Thomason [5]). *Every increasing property has a threshold function.*

Now it immediately follows from the above theorem that all the properties that we have mentioned so far—connectivity, planarity,[3] having the giant component, etc.—have a threshold function (thus exhibit a threshold phenomenon). How fascinating it is!

On the other hand, knowing that a property $\mathcal{F}$ has a threshold function $p_0 = p_0(\mathcal{F})$ does not tell us anything about the value of p_0. So it naturally became a central interest in the study of random graphs to find a threshold function for various increasing properties. One of the most studied classes of increasing properties is *subgraph containment*, i.e., the question of for what $p = p(n)$, $G_{n,p}$ is likely/unlikely to contain a copy of the given graph. Figure 8 shows some of the well-known threshold functions for various subgraph containments (and that for connectivity).

[2] *By the definition, a threshold function is determined up to a constant factor thus not unique, but conventionally people also call this the threshold function. In this article, we will separately define the threshold in Section 5, which is distinguished from a threshold function.*

[3] *We can apply the theorem for nonplanarity, which is an increasing property.*

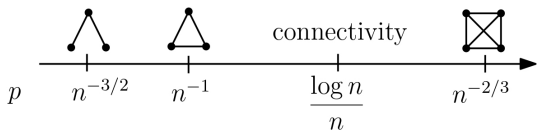


Figure 8. Some well-known thresholds.

Example 2.5. Figure 8 says that $p = \frac{1}{n}$ is a threshold function for the property $\mathcal{F} = \{\text{contains a triangle}\}$. Recall from the definition of a threshold that this means

(i) if $p \ll \frac{1}{n}$ then $\mathbb{P}(G_{n,p} \text{ contains a triangle}) \to 0$; and

(ii) if $p \gg \frac{1}{n}$ then $\mathbb{P}(G_{n,p} \text{ contains a triangle}) \to 1$.

We have already justified (i) in Example 1.2 by showing that

$$\mathbb{E}[\text{number of triangles in } G_{n,p}] \to 0 \text{ if } p \ll \frac{1}{n}.$$

However, showing (ii) is an entirely different story. As discussed in Remark 1.7, the fact that

$$\mathbb{E}[\text{number of triangles in } G_{n,p}] \to \infty$$

does not necessarily imply that $G_{n,p}$ typically contains many triangles. Here we briefly describe one technique, which is called the *second moment method*, that we can use to show (ii): let X be the number of triangles in $G_{n,p}$, noting that then X is a random variable. By showing that the variance of X is very small, which implies that X is "concentrated around" $\mathbb{E}X$, we can derive (from the fact that $\mathbb{E}X$ is huge) that typically the number of triangles in $G_{n,p}$ is huge. We remark that the second moment method is only a tiny part of the much broader topic of *concentration of a probability measure*.

We stress that, in general, finding a threshold function for a given increasing property is a very hard task. To illustrate this point, let's consider one of the most basic objects in graph theory, a *spanning tree*—a tree that contains all of the vertices.

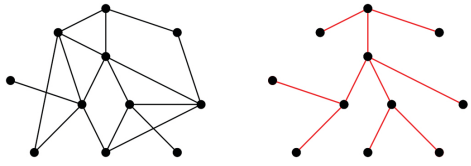


Figure 9. A connected graph and a spanning tree in it.

The question of finding a threshold function for $G_{n,p}$ of containing a spanning tree[4] was one of the first questions studied by Erdős and Rényi. Already in their seminal paper [6], Erdős and Rényi showed that a threshold function for containing a spanning tree is $p_0 = \frac{\log n}{n}$. However, the difficulty of this problem immensely changes if we require $G_{n,p}$ to contain a *specific* (up to isomorphisms) spanning tree (or more broadly, a spanning graph.[5]) For example, one of the biggest open questions in this area back in 1960s was finding a threshold function for a *Hamiltonian cycle* (a cycle that contains all of the vertices).

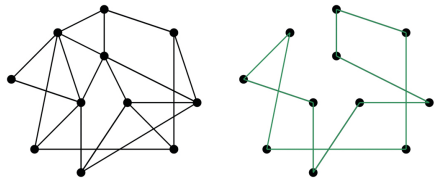


Figure 10. A graph and a Hamiltonian cycle in it.

This problem was famously solved by Pósa in 1976.

Theorem 2.6 (Pósa [16]). *A threshold function for $G_{n,p}$ to contain a Hamiltonian cycle is*

$$p_0(n) = \frac{\log n}{n}.$$

Note that both threshold functions for {contain any spanning tree} and {contain a Hamiltonian cycle} are of order $\frac{\log n}{n}$, even though the latter is a stronger requirement. Later we will see (in the discussion that follows Example 4.6) that $\frac{\log n}{n}$ is actually an easy lower bound on both threshold functions. It has long been conjectured that for *any* spanning tree[6] with a constant maximum degree, its threshold function is of order $\frac{\log n}{n}$. This conjecture was only very recently proved by Montgomery [14].

3. The Kahn–Kalai Conjecture: A Preview

Henceforth, $\mathcal{F}$ always denotes an increasing property.

In 2006, Jeff Kahn and Gil Kalai [12] posed an extremely bold conjecture that captures the location of threshold functions for *any* increasing properties. Its formal statement will be given in Conjecture 4.11 (graph version) and Theorem 5.7 (abstract version), and in this section we will give an informal description of this conjecture first. All of the terms not defined here will be discussed in the forthcoming sections.

Given an $\mathcal{F}$, we are interested in locating its threshold function, $p_0(\mathcal{F})$.[7] But again, this is in general a very hard task.

Kahn and Kalai introduced another quantity which they named the *expectation threshold* and denoted by $p_{\mathbb{E}}(\mathcal{F})$, which is associated with some sort of expectation

[4] *This is equivalent to $G_{n,p}$ is connected.*

[5] *A spanning graph means a graph that contains all of the vertices*

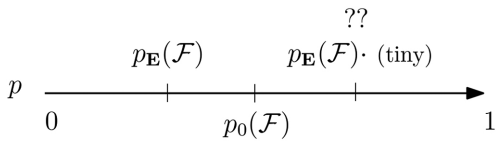
[6] *More precisely, for any sequence of spanning trees $\{T_n\}$*

[7] *We switch the notation from $p_0(n)$ to $p_0(\mathcal{F})$ to emphasize its dependence on $\mathcal{F}$.*

calculations as its name indicates. By its definition (Definition 4.5),

$$p_{\mathbb{E}}(\mathcal{F}) \le p_0(\mathcal{F}) \text{ for any } \mathcal{F},$$

and, in particular, $p_{\mathbb{E}}(\mathcal{F})$ is easy to compute for many interesting increasing properties $\mathcal{F}$. So $p_{\mathbb{E}}(\mathcal{F})$ provides an "easy" lower bound on the hard parameter $p_0(\mathcal{F})$. A really fascinating part is that then Kahn and Kalai conjectured that $p_0(\mathcal{F})$ is, in fact, bounded *above* by $p_{\mathbb{E}}(\mathcal{F})$ multiplied by some tiny quantity!



So this conjecture asserts that, for any $\mathcal{F}$, $p_0(\mathcal{F})$ is actually well-predicted by (much) easier $p_{\mathbb{E}}(\mathcal{F})$!

The graph version of this conjecture (Conjecture 4.11) is still open, but the abstract version (Theorem 5.7) is recently proved in [15].

4. Motivating Examples

The conjecture of Kahn and Kalai is very strong, and even the authors of the conjecture wrote in their paper [12] that "it would probably be more sensible to conjecture that it is *not* true." The fundamental question that motivated this conjecture was:

Question 4.1. What drives thresholds?

All of the examples in this section are carefully chosen to show the motivation behind the conjecture.

Recall that the definition of a threshold (Definition 2.3) doesn't distinguish constant factors. So in this section, we will use the convenient notation $\gtrsim, \lesssim$, and $\asymp$ to mean (respectively) $\ge, \le$, and $=$ up to constant factors. Finally, write $p_0(H)$ for a threshold function for $G_{n,p}$ of containing a copy of H, for notational simplicity.

Example 4.2. Let H be the graph in Figure 11. Let's find $p_0(H)$.

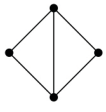


Figure 11. Graph H.

In Example 2.5, we observed that there is a connection between a threshold function and computing expectations. As we did in Examples 1.4 and 1.6,

$$\begin{aligned}&\mathbb{E}[\text{number of } H\text{'s in } G_{n,p}]\\ &= (\text{number of (labelled) } H\text{'s in } K_n)\times\\ &\quad \mathbb{P}(\text{each (labelled) copy of } H \text{ is present in } G_{n,p})\\ &\overset{(\dagger)}{\asymp} n^4 p^5,\end{aligned}$$

where (†) is because the number of H's in K_n is of order n^4 (since H has four vertices), and $\mathbb{P}$(each copy of H is present) is precisely p^5 (since H has five edges). So we have

$$\mathbb{E}[\text{number of } H\text{'s in } G_{n,p}] \to \begin{cases} 0 & \text{if } p \ll n^{-4/5};\\ \infty & \text{if } p \gg n^{-4/5}, \end{cases} \tag{1}$$

and let's (informally) call the value $p = n^{-4/5}$

"the threshold for the expectation of H."

This name makes sense since $p = n^{-4/5}$ is where the expected number of H's drastically changes. Note that (1) tells us that

$$\mathbb{P}(G_{n,p} \supseteq H) \to 0 \quad \text{if } p \ll n^{-4/5},$$

so, by the definition of a threshold, we have

$$n^{-4/5} \lesssim p_0(H).$$

This way, we can always easily find a lower bound on $p_0(F)$ for any graph F.

What is interesting here is that, for H in Figure 11, we can actually show that

$$\mathbb{P}(G_{n,p} \supseteq H) \to 1 \quad \text{if } p \gg n^{-4/5}$$

using the second moment method (discussed in Example 2.5). This tells us a rather surprising fact that $p_0(H)$ is actually *equal* to the threshold for the expectation of H.

> **Dream.** Maybe $p_0(F)$ is always equal to the threshold for the expectation of F for *any* graph F?

The next example shows that the above dream is too dreamy to be true.

Example 4.3. Consider $\tilde{H}$ in Figure 12 this time. Notice that $\tilde{H}$ is H in Figure 11 with a "tail."

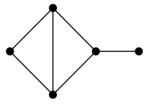


Figure 12. Graph $\tilde{H}$.

By repeating a similar computation as before, we have

$$\mathbb{E}[\text{number of } \tilde{H}\text{'s in } G_{n,p}] \to \begin{cases} 0 & \text{if } p \ll n^{-5/6};\\ \infty & \text{if } p \gg n^{-5/6}, \end{cases}$$

so the threshold for the expectation of $\tilde{H}$ is $n^{-5/6}$. Again, this gives that

$$\mathbb{P}(G_{n,p} \supseteq \tilde{H}) \to 0 \quad \text{if } p \ll n^{-5/6},$$

so we have $n^{-5/6} \lesssim p_0(\tilde{H})$. However, the actual threshold $p_0(\tilde{H})$ is $n^{-4/5}$, which is much larger than the lower bound.

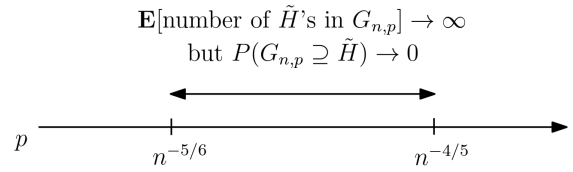


Figure 13. Gap between $p_0(\tilde{H})$ and the expectational lower bound.

This is interesting, because Figure 13 tells us that when $n^{-5/6} \ll p \ll n^{-4/5}$, $G_{n,p}$ contains a huge number of $\tilde{H}$ "on average," but still it is very unlikely that $G_{n,p}$ actually contains $\tilde{H}$. What happens in this inverval?

Here is an explanation. Recall from Example 4.2 that if $p \ll n^{-4/5}$, then $G_{n,p}$ is unlikely to contain H. But

$$\text{the absence of } H \text{ implies the absence of } \tilde{H},$$

because H is a subgraph of $\tilde{H}$!

So when $n^{-5/6} \ll p \ll n^{-4/5}$, it is highly unlikely that $G_{n,p}$ contains $\tilde{H}$ because it is already unlikely that $G_{n,p}$ contains H. However, if $G_{n,p}$ happens to contain H, then that copy of H typically has lots of "tails" as in Figure 14. This produces a huge number of copies of $\tilde{H}$'s in $G_{n,p}$.

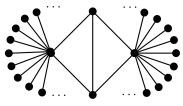


Figure 14. H with many "tails."

Maybe you have noticed the similarity between this example and the example of a lottery in Remark 1.7.

In Example 4.3, $p_0(\tilde{H})$ is not predicted by the expected number of $\tilde{H}$, thus the **Dream** is broken. However, it still shows that $p_0(\tilde{H})$ is predicted by the expected number of *some* subgraph of $\tilde{H}$, and, intriguingly, this holds true in general. To provide its formal statement, define the *density* of a graph F by

$$\text{density}(F) = \frac{(\text{the number of edges of } F)}{(\text{the number of vertices of } F)}.$$

The next theorem tells us the exciting fact that we can find $p_0(F)$ by just looking at its densest subgraph, as long as F is fixed.[8]

[8] *For example, a Hamiltonian cycle is not a fixed graph, since it grows as n grows.*

Theorem 4.4 (Bollobás [4]). *For any* fixed *graph F, $p_0(F)$ is equal to the threshold for the expectation of the* densest *subgraph of F.*

For example, in Example 4.2, the densest subgraph of H is H itself, so $p_0(H)$ is determined by the expectation of H. This also determines $p_0(\tilde{H})$ in Example 4.3, since the densest subgraph of $\tilde{H}$ is again H.

Motivated by the preceding examples and Theorem 4.4, we give a formal definition of the *expectation threshold*.

Definition 4.5 (Expectation threshold). For any graph F, the expectation threshold for F is

$$p_{\mathbb{E}}(F) = \min\{p : \mathbb{E}[\text{number of } F' \text{ in } G_{n,p}] \geq 1 \quad \forall F' \subseteq F\}.$$

Observe that

$$p_{\mathbb{E}}(F) \lesssim p_0(F) \text{ for any } F, \tag{2}$$

and in particular, Theorem 4.4 gives that

$$p_{\mathbb{E}}(F) \asymp p_0(F) \text{ for any fixed } F.$$

Note that this gives a beautiful answer to Question 4.1 whenever $\mathcal{F}$ is a property of containing a fixed graph.

Example 4.6. Theorem 4.4 characterizes threshold functions for any fixed graphs. To extend our exploration, in this example we consider a graph that *grows* as n grows. We say a graph M is a *matching* if M is a disjoint union of edges. M is a *perfect matching* if M is a matching that contains all the vertices. Write PM for perfect matching.

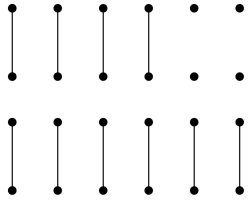


Figure 15. A matching (above) and a perfect matching (below).

Keeping Question 4.1 in mind, let's first check the validity of Theorem 4.4 to a perfect matching,[9] which is not a fixed graph. By repeating a similar computation as before, we obtain that

$$\mathbb{E}[\text{number of PM's in } G_{n,p}] \asymp (n/e)^{n/2} p^{n/2},$$

which tends to 0 if $p \ll 1/n$. In fact, it is easy to compute (by considering all subgraphs of a perfect matching) that $p_{\mathbb{E}}(\text{PM}) \asymp 1/n$, so by (2),

$$p_0(\text{PM}) \gtrsim 1/n.$$

However, unlike threshold functions for fixed graphs, $p_0(\text{PM})$ is *not* equal to $p_{\mathbb{E}}(\text{PM})$; it was proved by Erdős and Rényi that

$$p_0(\text{PM}) \asymp \frac{\log n}{n} (\gg p_{\mathbb{E}}(\text{PM})). \tag{3}$$

[9] *We assume $2|n$ to avoid a trivial obstruction from having a perfect matching.*

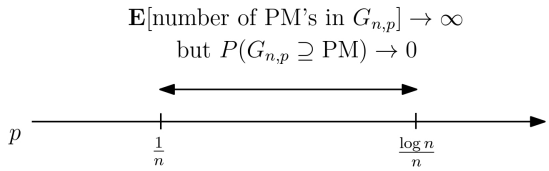


Figure 16. Gap between p_0(PM) and $p_{\mathbb{E}}$(PM).

Notice that, in Figure 16, what happens in the gap is *fundamentally* different from that in Figure 13. When $\frac{1}{n} \ll p \ll \frac{\log n}{n}$, $G_{n,p}$ contains huge numbers of PMs and *all its subgraphs* "on average." This means the absence of a subgraph of a PM is not the obstruction for $G_{n,p}$ from having a PM when $p \gg 1/n$. Then what happens here, and what's the real obstruction?

It turned out, we have

$$p_0(\text{PM}) \gtrsim \frac{\log n}{n}$$

for a very simple reason: the existence of an isolated vertex[10] in $G_{n,p}$. It is well-known that if $p \ll \frac{\log n}{n}$, then $G_{n,p}$ contains an isolated vertex with high probability (this phenomenon is elaborated in Example 4.7). Of course, if there is an isolated vertex in a graph, then this graph cannot contain a perfect matching.

So (3) says the very compelling fact that once p is large enough that $G_{n,p}$ avoids isolated vertices, $G_{n,p}$ contains a perfect matching!

The existence of an isolated vertex in $G_{n,p}$ is essentially equivalent to the *coupon collector's problem*:

Example 4.7 (Coupon collector's problem). Each box of cereal contains a random coupon, and there are n different types of coupons. If all coupons are equally likely, then how many boxes of cereal do we (typically) need to buy to collect all n coupons?

The well-known answer to this question is that we need to buy $\gtrsim n \log n$ boxes of cereal. This phenomenon can be translated to $G_{n,p}$ in the following way: in $G_{n,p}$, the n vertices are regarded as coupons. If a vertex is contained in a (random) edge in $G_{n,p}$, then that is regarded as being "collected." Note that if $p \ll \frac{\log n}{n}$, then typically the number of edges in $G_{n,p}$ is $\binom{n}{2}p \ll n \log n$, and then the coupon collector's problem says that there is typically an "uncollected coupon," which is an isolated vertex.

Observe that, in Example 4.6, the "coupon-collector behavior" of $G_{n,p}$ provides another lower bound on p_0(PM), pushing up the first lower bound, $p_{\mathbb{E}}$(PM), by $\log n$. And it turned out that this second (better) lower bound is equal to the threshold.

[10] *a vertex not contained in any edges*

Lower bounds		Threshold
Expectation threshold	$p_0 \gtrsim p_{\mathbb{E}}$	$p_0 \asymp p_{\mathbb{E}} \log n$
Coupon collector	$p_0 \gtrsim p_{\mathbb{E}} \log n$	

Figure 17. Bounds on p_0(PM).

Hitting time. Again, the existence of an isolated vertex in a graph trivially blocks this graph from containing any spanning subgraphs. In Example 4.6, we observed the compelling phenomenon that if p is large enough that $G_{n,p}$ typically avoids isolated vertices, then for those p, $G_{n,p}$ contains a perfect matching with high probability. Would this mean that, for $G_{n,p}$, isolated vertices are the *only* barriers to the existence of spanning subgraphs?

To investigate this question, we consider a random process defined as below. Consider a sequence of graphs on n vertices

$$G_0 = \emptyset, G_1, G_2, \dots, G_{\binom{n}{2}} = K_n,$$

where G_{m+1} is obtained from G_m by adding a random edge.

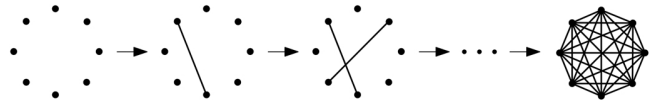


Figure 18. Random process.

Then G_m, the m-th graph in this sequence, is the random variable that is uniformly distributed among all the graphs on n vertices with m edges. The next theorem tells us that, indeed, isolated vertices are *the* obstructions for a random graph to having a perfect matching.

Theorem 4.8 (Erdős–Rényi [7]). *Let m_0 denote the first time that G_m contains no isolated vertices. Then, with high probability, G_{m_0} contains a perfect matching.*

We remark that Theorem 4.8 gives much more precise information about p_0(PM) (back in $G_{n,p}$ setting). For example, Theorem 4.8 implies:

Theorem 4.9. *Let $p = \frac{\log n + c_n}{n}$. Then*

$$\lim_{n\to\infty} \mathbb{P}(G_{n,p} \supseteq PM) = \begin{cases} 0 & \text{if } c_n \to -\infty \\ e^{-e^{-c}} & \text{if } c_n \to c \\ 1 & \text{if } c_n \to \infty \end{cases}.$$

We observe a similar phenomenon for Hamiltonian cycles. Notice that in order for a graph to contain a Hamiltonian cycle, a minimum requirement is that every vertex is contained in at least *two* edges. The next theorem tells us that, again, this naive requirement is essentially the only barrier.

Theorem 4.10 (Ajtai-Komlós-Szemerédi [1], Bollobás [3]). *Let m_1 denote the first time that every vertex in G_m is contained in at least two edges. Then, with high probability, G_{m_1} contains a Hamiltonian cycle.*

Returning to Question 4.1, so far we have established that there are two factors that affect threshold functions. We first observed that $p_{\mathbb{E}}$ always gives a lower bound on p_0. We then observed that, when it applies, the coupon-collector behavior of $G_{n,p}$ pushes up this expectational lower bound by $\log n$. Conjecture 4.11 below dauntingly proposes that there are *no other factors* that affect thresholds.

Conjecture 4.11 (Kahn–Kalai [12]). *For any graph F,*

$$p_0(F) \lesssim p_{\mathbb{E}}(F) \log n.$$

Conjecture 4.11 is still wide open even after the "abstract version" of this conjecture is proved. We close this section with a very interesting example in which p_0 lies strictly in between $p_{\mathbb{E}}$ and $p_{\mathbb{E}} \log n$. A *triangle factor* is a (vertex-) disjoint union of triangles that contains *all* the vertices.

Figure 19. A triangle factor.

The question of a threshold function for a triangle-factor[11] was famously solved by Johansson, Kahn, and Vu [10]. Observe that an obvious obstruction for a graph from having a triangle factor is the existence of a vertex that is not contained in any triangles. The result below is the hitting time version of [10], which is obtained by combining [11] and [9].

Theorem 4.12. *Let m_2 denote the first time that every vertex in G_m is contained in at least one triangle. Then, with high probability, G_{m_2} contains a triangle factor.*

The above theorem implies that

$$p_0(\text{triangle factor}) \asymp p_{\mathbb{E}}(\text{triangle factor}) \cdot (\log n)^{1/3}.$$

5. The Expectation Threshold Theorem

The abstract version of the Kahn–Kalai conjecture, which is the main content of this section, is recently proved in [15]. We remark that the discussion in this section is not restricted by the languages in graph theory anymore.

We introduce some more definitions for this general setting. Given a finite set X, the *p-biased product probability measure*, μ_p, on 2^X is defined by

$$\mu_p(A) = p^{|A|}(1-p)^{|X\setminus A|} \quad \forall A \subseteq X.$$

We use X_p for the random variable whose law is

$$\mathbb{P}(X_p = A) = \mu_p(A) \quad \forall A \subseteq X.$$

[11] *or, more generally, a K_r-factor for any fixed r*

In other words, X_p is a "p-random subset" of X, which means X_p contains each element of X with probability p independently.

Example 5.1. If $X = \binom{[n]}{2}$, then

$$X_p = G_{n,p}.$$

So $G_{n,p}$ is a special case of the random model X_p.

We redefine increasing property in our new setup. A *property* is a subset of 2^X, and $\mathcal{F} \subseteq 2^X$ is an *increasing property* if

$$B \supseteq A \in \mathcal{F} \Rightarrow B \in \mathcal{F}.$$

Informally, a property is increasing if we cannot "destroy" this property by adding elements. Note that in this new definition, $\mathcal{F}$ is not required to possess strong symmetry as in increasing *graph* properties; for example, there is no longer a requirement "invariant under graph isomorphisms."

Observe that $\mu_p(\mathcal{F})(:= \sum_{A\in\mathcal{F}} \mu_p(A) = \mathbb{P}(X_p \in \mathcal{F}))$ is a polynomial in p, thus continuous. Furthermore, it is a well-known fact that $\mu_p(\mathcal{F})$ is strictly increasing in p unless $\mathcal{F} = \emptyset, 2^X$ (see Figure 20). For the rest of this section, we always assume $\mathcal{F} \neq \emptyset, 2^X$.

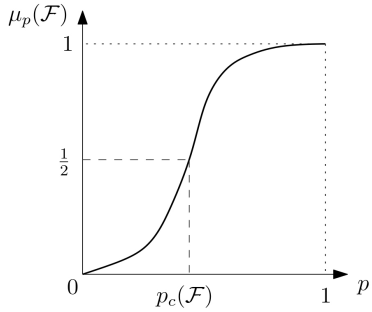


Figure 20. $\mu_p(\mathcal{F})$ for $p \in [0,1]$, and $p_c(\mathcal{F})$.

Because $\mu_p(\mathcal{F})$ is continuous and strictly increasing in p, there exists a unique $p_c(\mathcal{F})$ for which $\mu_{p_c}(\mathcal{F}) = 1/2$. This $p_c(\mathcal{F})$ is called *the threshold* for $\mathcal{F}$.

Remark 5.2. The definition of $p_c(\mathcal{F})$ does not require sequences. Incidentally, by Theorem 2.4, for any increasing *graph* property $\mathcal{F}(= \mathcal{F}_n)$, $p_c(\mathcal{F})$ is an (Erdős–Rényi) threshold function for $\mathcal{F}$.

For a general increasing property $\mathcal{F} \subseteq 2^X$, the definition of $p_{\mathbb{E}}$ is not applicable anymore. Kahn and Kalai introduced the following generalized notion of the expectation threshold, which is also introduced by Talagrand [17].

Definition 5.3. Given a finite set X and an increasing property $\mathcal{F} \subseteq 2^X$, $q(\mathcal{F})$ is the maximum of $q \in [0,1]$ for which there exists $\mathcal{G} \subseteq 2^X$ satisfying the following two properties.

(a) Each $A \in \mathcal{F}$ contains some member of $\mathcal{G}$.
(b) $\sum_{S\in\mathcal{G}} q^{|S|} \leq 1/2$.

A family $\mathcal{G} \subseteq 2^X$ that satisfies (a) is called a *cover* of $\mathcal{F}$.

Remark 5.4. The definition of $q(\mathcal{F})$ eliminates the "symmetry" requirement—which seems very natural (and seemingly easier to deal with) in the context of thresholds for subgraph containments—from the definition of $p_{\mathbb{E}}$. It is worth noting that this flexibility is crucially used in the proof of Theorem 5.7 in [15].

The next proposition says that $q(\mathcal{F})$ still provides a lower bound on the threshold.

Proposition 5.5. *For any finite set X and increasing property $\mathcal{F} \subseteq 2^X$,*

$$q(\mathcal{F}) \le p_c(\mathcal{F}).$$

Justification. Write $q = q(\mathcal{F})$. By the definition of $p_c(\mathcal{F})$, it suffices to show that $\mu_q(\mathcal{F}) \le 1/2$. We have

$$\mu_q(\mathcal{F}) \le \sum_{S\in\mathcal{G}} \sum_{S\subseteq A\in\mathcal{F}} \mu_q(A) \le \sum_{S\in\mathcal{G}} \sum_{B\supseteq S} \mu_q(B) = \sum_{S\in\mathcal{G}} q^{|S|} \le 1/2,$$

where the first inequality uses the fact that $\mathcal{G}$ covers $\mathcal{F}$. □

For a graph F, write $\mathcal{F}_F$ for the increasing graph property of containing a copy of F. The example below illustrates the relationship between $p_{\mathbb{E}}(F)$ and $q(\mathcal{F}_F)$.

Example 5.6 (Example 4.3 revisited). For $X = \binom{[n]}{2}$ (so $X_p = G_{n,p}$) and the increasing property $\mathcal{F} = \{\text{contain } \tilde{H}\} (\subseteq 2^X)$,

$$\mathcal{G}_1 := \{\text{all the (labelled) copies of } \tilde{H} \text{ in } K_n\}$$

is a cover of $\mathcal{F}$. The left-hand side of Definition 5.3 (b) is

$$\sum_{S\in\mathcal{G}_1} q^{|S|} = (\text{number of } \tilde{H}\text{'s in } K_n) \times \mathbb{P}(\text{each copy of } \tilde{H} \text{ is present in } G_{n,p}),$$

which is precisely the expected number of $\tilde{H}$'s in $G_{n,p}$. Combined with Proposition 5.5, the above computation gives that $n^{-5/6} \lesssim p_c(\mathcal{F})$.

On the other hand, we have (implicitly) discussed in Example 4.3 that there is another cover that gives a lower bound better than that of $\mathcal{G}_1$; if we take

$$\mathcal{G}_2 := \{\text{all the (labelled) copies of } H \text{ in } K_n\},$$

then the computation in Definition 5.3 (b) gives that $n^{-4/5} \lesssim p_c(\mathcal{F})$.

The above discussion shows that, for any (not necessarily fixed) graph F,

$$p_{\mathbb{E}}(F) \lesssim q(\mathcal{F}_F)$$

(whether $p_{\mathbb{E}}(F) \asymp q(\mathcal{F}_F)$ is unknown). The abstract version of the Kahn–Kalai conjecture is similar to its graph version, with $p_{\mathbb{E}}$ replaced by $q(\mathcal{F})$. This is what's proved in [15].

Theorem 5.7 (Park–Pham [15], conjectured in [12, 17]). *There exists a constant K such that for any finite set X and increasing property $\mathcal{F} \subseteq 2^X$,*

$$p_c(\mathcal{F}) \le Kq(\mathcal{F}) \log \ell(\mathcal{F})$$

where $\ell(\mathcal{F})$ is the size of a largest minimal element of $\mathcal{F}$.

Theorem 5.7 is extremely powerful; for instance, its immediate consequences include historically difficult results such as the resolutions of *Shamir's problem* [10] and the "*tree conjecture*" [14]. Here we mention one smaller consequence:

Example 5.8. If F is a fixed graph, then $\ell(\mathcal{F}_F)$ is the number of edges in F, thus a constant. So in this case Theorem 5.7 says $p_c(\mathcal{F}) \asymp q(\mathcal{F})$, which recovers Theorem 4.4.

The sunflower conjecture, and "fractional" Kahn–Kalai. The proof of Theorem 5.7 is strikingly easy given its powerful consequences. The approach is inspired by remarkable work of Alweiss, Lovett, Wu, and Zhang [2] on the *Erdős–Rado sunflower conjecture*, which seemingly has no connection to threshold phenomena. This totally unexpected connection was first exploited by Frankston, Kahn, Nayaranan, and the author in [8], where a "fractional" version of the Kahn–Kalai conjecture (conjectured by Talagrand [17]) was proved, illustrating how two seemingly unrelated fields of mathematics can be nicely connected!

Note that $q(\mathcal{F})$ is in theory hard to compute. For instance, in Example 4.3, we can estimate $p_{\mathbb{E}}(\tilde{H})$ by finding $F \subseteq \tilde{H}$ with the maximum $e(F)/v(F)$. On the other hand, to compute $q(\mathcal{F}_{\tilde{H}})$, we should in principle consider all possible covers of $\mathcal{F}_{\tilde{H}}$, which is typically not feasible. The good news is that there is a convenient way to find an upper bound on $q(\mathcal{F})$, which is often of the correct order. Namely, Talagrand [17] introduced a notion of *fractional expectation threshold*, $q_f(\mathcal{F})$, satisfying

$$q(\mathcal{F}) \le q_f(\mathcal{F}) \le p_c(\mathcal{F})$$

for any increasing property $\mathcal{F}$. He conjectured (and it was proved in [8]) that the (abstract) Kahn–Kalai conjecture (now Theorem 5.7) holds with $q_f(\mathcal{F})$ in place of $q(\mathcal{F})$. This puts us in linear programming territory: by LP duality, a bound $q_f(\mathcal{F}) \le \alpha$ ($\alpha \in [0,1]$) is essentially equivalent to existence of an "α-spread" probability measure on $\mathcal{F}$. In all applications of Theorem 5.7 to date, what is actually used to upper bound $q(\mathcal{F})$ is an appropriately spread measure.[12] So all these applications actually follow from the weaker Talagrand version.

We close this article with a very interesting conjecture of Talagrand [17] that would imply the equivalence of Theorem 5.7 and its fractional version:

[12] *The problem of constructing well-spread measure is getting growing attention now; see, e.g.,* [13] *for a start.*

Conjecture 5.9. *There exists a constant K such that for any finite set X and increasing property $\mathcal{F} \subseteq 2^X$,*

$$q_f(\mathcal{F}) \leq Kq(\mathcal{F}).$$

ACKNOWLEDGMENT. The author is grateful to Jeff Kahn for his helpful comments.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi, *First occurrence of Hamilton cycles in random graphs*, Cycles in graphs (Burnaby, B.C., 1982), North-Holland Math. Stud., vol. 115, North-Holland, Amsterdam, 1985, pp. 173–178, DOI 10.1016/S0304-0208(08)73007-X. MR821516

[2] R. Alweiss, S. Lovett, K. Wu, and J. Zhang, *Improved bounds for the sunflower lemma*, Ann. of Math. (2) **194** (2021), no. 3, 795–815, DOI 10.4007/annals.2021.194.3.5. MR4334977

[3] B. Bollobás, *The evolution of sparse graphs*, Graph theory and combinatorics (Cambridge, 1983), Academic Press, London, 1984, pp. 35–57. MR777163

[4] B. Bollobás, *Random graphs*, 2nd ed., Cambridge Studies in Advanced Mathematics, vol. 73, Cambridge University Press, Cambridge, 2001, DOI 10.1017/CBO9780511814068. MR1864966

[5] B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica **7** (1987), no. 1, 35–38, DOI 10.1007/BF02579198. MR905149

[6] P. Erdős and A. Rényi, *On the evolution of random graphs* (English, with Russian summary), Magyar Tud. Akad. Mat. Kutató Int. Közl. **5** (1960), 17–61. MR125031

[7] 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** (1966), 359–368, DOI 10.1007/BF01894879. MR200186

[8] K. Frankston, J. Kahn, B. Narayanan, and J. Park, *Thresholds versus fractional expectation-thresholds*, Ann. of Math. (2) **194** (2021), no. 2, 475–495, DOI 10.4007/annals.2021.194.2.2. MR4298747

[9] A. Heckel, M. Kaufmann, N. Müller, and M. Pasch, *The hitting time of clique factors,* Preprint, arXiv 2302.08340

[10] A. Johansson, J. Kahn, and V. Vu, *Factors in random graphs*, Random Structures Algorithms **33** (2008), no. 1, 1–28, DOI 10.1002/rsa.20224. MR2428975

[11] J. Kahn, *Hitting times for Shamir's problem*, Trans. Amer. Math. Soc. **375** (2022), no. 1, 627–668, DOI 10.1090/tran/8508. MR4358678

[12] J. Kahn and G. Kalai, *Thresholds and expectation thresholds*, Combin. Probab. Comput. **16** (2007), no. 3, 495–502, DOI 10.1017/S0963548307008474. MR2312440

[13] D. 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, to appear.

[14] R. Montgomery, *Spanning trees in random graphs*, Adv. Math. **356** (2019), 106793, 92, DOI 10.1016/j.aim.2019.106793. MR3998769

[15] J. Park and H. T. Pham, *A proof of the Kahn–Kalai conjecture*, J. Amer. Math. Soc., electronically published on August 7, 2023, DOI: `https://doi.org/10.1090/jams/1028` (to appear in print).

[16] L. Pósa, *Hamiltonian circuits in random graphs*, Discrete Math. **14** (1976), no. 4, 359–364, DOI 10.1016/0012-365X(76)90068-6. MR389666

[17] M. Talagrand, *Are many small sets explicitly small?*, STOC'10—Proceedings of the 2010 ACM International Symposium on Theory of Computing, ACM, New York, 2010, pp. 13–35. MR2743011

Jinyoung Park

Credits

Opening graphic is courtesy of enjoynz via Getty.

Figures 1–20 and photo of Jinyoung Park are courtesy of Jinyoung Park.