Basic models and questions in statistical network analysis*

Miklós Z. Rácz† and Sébastien Bubeck†

Microsoft Research, Redmond, WA, USA
e-mail: miracz@microsoft.com*;* sebubeck@microsoft.com

Abstract: Extracting information from large graphs has become an important statistical problem since network data is now common in various fields. In this minicourse we will investigate the most natural statistical questions for three canonical probabilistic models of networks: (i) community detection in the stochastic block model, (ii) finding the embedding of a random geometric graph, and (iii) finding the original vertex in a preferential attachment tree. Along the way we will cover many interesting topics in probability theory such as Pólya urns, large deviation theory, concentration of measure in high dimension, entropic central limit theorems, and more.

MSC 2010 subject classifications: Primary 62-02, 05C80, 60C05.
Keywords and phrases: Networks, statistical inference, random graphs, random trees, community detection, stochastic block model, random geometric graphs, evolving random graphs, preferential attachment.

Contents

1. Lecture 1: A primer on exact recovery in the general stochastic block model

Community detection is a fundamental problem in many sciences, such as sociology (e.g., finding tight-knit groups in social networks), biology (e.g., detecting protein complexes), and beyond. Given its importance, there have been a plethora of algorithms developed in the past few decades to detect communities. But how can we test whether an algorithm performs well? What are the fundamental limits to *any* community detection algorithm? Often in real data the ground truth is not known (or there is not even a well-defined ground truth), so judging the performance of algorithms can be difficult. Probabilistic generative models can be used to model real networks, and even if they do not fit the data perfectly, they can still be useful: they can act as benchmarks for comparing different clustering algorithms, since the ground truth is known.

Perhaps the most widely studied generative model that exhibits community structure is the stochastic block model (SBM). The SBM was first introduced in sociology [37] and was then studied in several different scientific communities, including mathematics, computer science, physics, and statistics [23, 24, 35, 10, 43, 52].[1] It gives a distribution on graphs with n vertices with a hidden partition of the nodes into k communities. The relative sizes of the communities, and the edge densities connecting communities are parameters of the general SBM. The statistical inference problem is then to recover as much of the community structure as possible given a realization of the graph, but without knowing any of the community labels.

1.1. The stochastic block model and notions of recovery

The general stochastic block model is a distribution on graphs with latent community structure, and it has three parameters: n, the number of vertices; a probability distribution $p = (p_1, \ldots, p_k)$ that describes the relative sizes of the communities; and $Q \in [0,1]^{k \times k}$, a symmetric $k \times k$ matrix that describes the probabilities with which two given vertices are connected, depending on which communities they belong to. The number of communities, k, is implicit in this notation; in these notes we assume that k is a fixed constant. A random graph from $\text{SBM}(n, p, Q)$ is defined as follows:

- The vertex set of the graph is $V = \{1, \ldots, n\} \equiv [n]$.
- Every vertex $v \in V$ is independently assigned a (hidden) label $\sigma_v \in [k]$ from the probability distribution p on $[k]$. That is, $\mathbb{P}(\sigma_v = i) = p_i$ for every $i \in [k]$.
- Given the labels of the vertices, each (unordered) pair of vertices $(u, v) \in V \times V$ is connected independently with probability Q_{σ_u, σ_v}.

---

[1]Disclaimer: the literature on community detection is vast and rapidly growing. It is not our intent here to survey this literature; we refer the interested reader to the papers we cite, as well as the recent survey by Abbe [1], for further references.
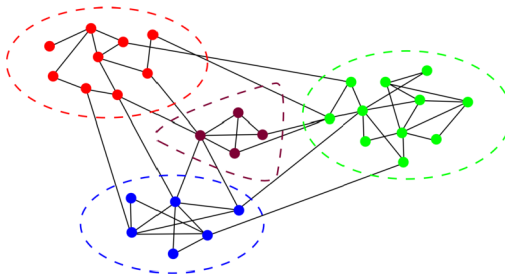
Fɪɢ 1. *A schematic of the general stochastic block model.*

Example 1.1 (Symmetric communities)**.** *A simple example to keep in mind is that of symmetric communities, with more edges within communities than between communities. This is modeled by the SBM with $p_i = 1/k$ for all $i \in [k]$ and $Q_{i,j} = a$ if $i = j$ and $Q_{i,j} = b$ otherwise, with $a > b > 0$.*

We write $G \sim \mathrm{SBM}(n, p, Q)$ for a graph generated according to the SBM without the hidden vertex labels revealed. The goal of a statistical inference algorithm is to recover as many labels as possible using only the underlying graph as an observation. There are various notions of success that are worth studying.

- **Weak recovery** (also known as detection). An algorithm is said to *weakly recover* or *detect* the communities if it outputs a partition of the nodes which is positively correlated with the true partition, with high probability (whp)[2].
- **Partial recovery.** How much can be recovered about the communities? An algorithm is said to *recover communities with an accuracy of $\alpha \in [0, 1]$* if it outputs a labelling of the nodes which agrees with the true labelling on a fraction α of the nodes whp. An important special case is when only $o(n)$ vertices are allowed to be misclassified whp, known as *weak consistency* or *almost exact recovery.*
- **Exact recovery** (also known as recovery or strong consistency). The strongest notion of reconstruction is to recover the labels of all vertices exactly whp. When this is not possible, it can still be of interest to understand which communities can be exactly recovered, if not all; this is sometimes known as *"partial-exact-recovery".*

In all the notions above, the agreement of a partition with the true partition is maximized over all relabellings of the communities, since we are not interested in the specific original labelling per se, but rather the partition (community structure) it induces.

The different notions of recovery naturally lead to studying different regimes of the parameters. It is well-known that in the Erdős-Rényi random graph—

---

[2]In these notes "with high probability" stands for with probability tending to 1 as the number of nodes in the graph, n, tends to infinity.

which is the SBM with a single block, or, equivalently, the SBM where all the connection probabilities are equal—the threshold connection probability for the graph to have a giant component is $1/n$ (see [32]) and the threshold connection probability for it to be connected is $\ln(n)/n$ (see [31]). These results suggest what the relevant regimes are for the different notions of recovery in the SBM. For weak recovery to be possible, many vertices in all but one community should be non-isolated (in the symmetric case this means that there should be a giant component), requiring the edge probabilities to be $\Omega(1/n)$. For exact recovery, all vertices in all but one community should be non-isolated (in the symmetric case this means that the graph should be connected), requiring the edge probabilities to be $\Omega(\ln(n)/n)$. In these regimes it is natural to scale the edge probability matrices accordingly, i.e., to consider SBM $(n, p, Q/n)$ or SBM $(n, p, \ln(n) Q/n)$, where $Q \in \mathbb{R}_+^{k \times k}$.

There has been lots of work in the past few years understanding the fundamental limits to recovery under the various notions discussed above. For weak recovery there is a sharp phase transition, the threshold of which was first conjectured in [26]. This was proven first for two symmetric communities [47, 48] and then for multiple communities [4]. Partial recovery is less well understood, and finding the fraction of nodes that can be correctly recovered for a given set of parameters is an open problem; see [49] for results in this direction for two symmetric communities.

In this lecture we are interested in exact recovery, for which Abbe and Sandon gave the value of the threshold for the general SBM, and showed that a quasilinear time algorithm works all the way to the threshold [3] (building on previous work that determined the threshold for two symmetric communities [2, 50]). The remainder of this lecture is an exposition of their main results and a few of the key ideas that go into proving and understanding it.

1.2. From exact recovery to testing multivariate Poisson distributions

Recall that we are interested in the logarithmic degree regime for exact recovery, i.e., we consider $G \sim \mathrm{SBM}(n, p, \ln(n)Q/n)$, where $Q \in \mathbb{R}_+^{k \times k}$ is independent of n. We also assume that the communities have linear size, i.e., that p is independent of n, and $p_i \in (0, 1)$ for all i. Our goal is to recover the labels of all the vertices whp.

As a thought experiment, imagine that not only is the graph G given, but also all vertex labels are revealed, except for that of a given vertex $v \in V$. Is it possible to determine the label of v?

Understanding this question is key for understanding exact recovery, since if the error probability of this is too high, then exact recovery will not be possible. On the other hand, it turns out that in this regime it is possible to recover all but $o(n)$ labels using an initial partial recovery algorithm. The setup of the thought experiment then becomes relevant, and if we can determine the label of v given the labels of all the other nodes with low error probability, then we can
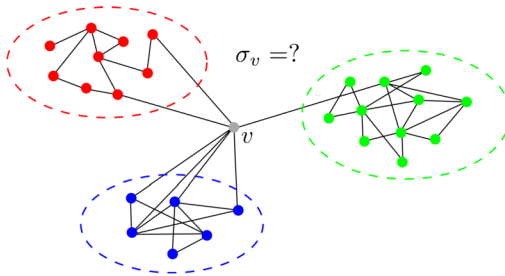
FIG 2. *Suppose all community labels are known except that of vertex v. Can the label of v be determined based on its neighbors' labels?*

correct all errors made in the initial partial recovery algorithm, leading to exact recovery. We will come back to the connection between the thought experiment and exact recovery; for now we focus on understanding this thought experiment.

Given the labels of all vertices except v, the information we have about v is the number of nodes in each community it is connected to. In other words, we know the *degree profile* $d(v)$ of v, where, for a given labelling of the graph's vertices, the i-th component $d_i(v)$ is the number of edges between v and the vertices in community i.

The distribution of the degree profile $d(v)$ depends on the community that v belongs to. Recall that the community sizes are given by a multinomial distribution with parameters n and p, and hence the relative size of community $i \in [k]$ concentrates on p_i. Thus if $\sigma_v = j$, the degree profile $d(v) = (d_1(v), \ldots, d_k(v))$ can be approximated by independent binomials, with $d_i(v)$ approximately distributed as $\mathrm{Bin}(np_i, \ln(n)Q_{i,j}/n)$, where $\mathrm{Bin}(m, q)$ denotes the binomial distribution with m trials and success probability q. In this regime, the binomial distribution is well-approximated by a Poisson distribution of the same mean. In particular, Le Cam's inequality gives that

$$\mathrm{TV}\left(\mathrm{Bin}\left(na, \ln(n)b/n\right), \mathrm{Poi}\left(ab\ln(n)\right)\right) \leq \frac{2ab^2\left(\ln(n)\right)^2}{n},$$

where $\mathrm{Poi}(\lambda)$ denotes the Poisson distribution with mean λ, and TV denotes the total variation distance[3]. Using the additivity of the Poisson distribution and the triangle inequality, we get that

$$\mathrm{TV}\left(\mathcal{L}\left(d(v)\right), \mathrm{Poi}\left(\ln(n)\sum_{i\in[k]} p_i Q_{i,j} e_i\right)\right) = O\left(\frac{\left(\ln(n)\right)^2}{n}\right),$$

where $\mathcal{L}(d(v))$ denotes the law of $d(v)$ conditionally on $\sigma_v = j$ and e_i is the i-th unit vector.

---

[3]Recall that the total variation distance between two random variables X and Y taking values in a finite space $\mathcal{X}$ with laws μ and ν is defined as $\mathrm{TV}(\mu, \nu) \equiv \mathrm{TV}(X, Y) = \frac{1}{2}\sum_{x\in\mathcal{X}}|\mu(x) - \nu(x)| = \sup_A |\mu(A) - \nu(A)|$.

Thus the degree profile of a vertex in community j is approximately Poisson distributed with mean $\ln(n) \sum_{i \in [k]} p_i Q_{i,j} e_i$. Defining $P = \mathrm{diag}(p)$, this can be abbreviated as $\ln(n) (PQ)_j$, where $(PQ)_j$ denotes the j-th column of the matrix PQ. We call the quantity $(PQ)_j$ the *community profile* of community j; this is the quantity that determines the distribution of the degree profile of vertices from a given community.

Our thought experiment has thus been reduced to a Bayesian hypothesis testing problem between k multivariate Poisson distributions. The prior on the label of v is given by p, and we get to observe the degree profile $d(v)$, which comes from one of k multivariate Poisson distributions, which have mean $\ln(n)$ times the community profiles $(PQ)_j$, $j \in [k]$.

1.3. Testing multivariate Poisson distributions

We now turn to understanding the testing problem described above; the setup is as follows. We consider a Bayesian hypothesis testing problem with k hypotheses. The random variable H takes values in $[k]$ with prior given by p, i.e., $\mathbb{P}(H = j) = p_j$. We do not observe H, but instead we observe a draw from a multivariate Poisson distribution whose mean depends on the realization of H: given $H = j$, the mean is $\lambda(j) \in \mathbb{R}_+^k$. In short:

$$ D \,|\, H = j \sim \mathrm{Poi}(\lambda(j)), \qquad j \in [k]. $$

In more detail:

$$ \mathbb{P}(D = d \,|\, H = j) = \mathcal{P}_{\lambda(j)}(d), \qquad d \in \mathbb{Z}_+^k, $$

where

$$ \mathcal{P}_{\lambda(j)}(d) = \prod_{i \in [k]} \mathcal{P}_{\lambda_i(j)}(d_i) $$

and

$$ \mathcal{P}_{\lambda_i(j)}(d_i) = \frac{\lambda_i(j)^{d_i}}{d_i!} e^{-\lambda_i(j)}. $$

Our goal is to infer the value of H from a realization of D. The error probability is minimized by the maximum a posteriori (MAP) rule, which, upon observing $D = d$, selects

$$ \arg\max_{j \in [k]} \mathbb{P}(D = d \,|\, H = j)\, p_j $$

as an estimate for the value of H, with ties broken arbitrarily. Let P_e denote the error of the MAP estimator. One can think of the MAP estimator as a tournament of $k-1$ pairwise comparisons of the hypotheses: if $\mathbb{P}(D = d \,|\, H = i)\, p_i > \mathbb{P}(D = d \,|\, H = j)\, p_j$ then the MAP estimate is not j. The probability that one makes an error during such a comparison is exactly

$$ P_e(i,j) := \sum_{x \in \mathbb{Z}_+^k} \min \{ \mathbb{P}(D = x \,|\, H = i)\, p_i, \mathbb{P}(D = x \,|\, H = j)\, p_j \}. \qquad (1.1) $$

For finite k, the error of the MAP estimator is on the same order as the largest pairwise comparison error, i.e., $\max_{i,j} P_e(i,j)$. In particular, we have that

$$\frac{1}{k-1} \sum_{i=1}^{k} \sum_{j=i+1}^{k} P_e(i,j) \leq P_e \leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} P_e(i,j). \qquad (1.2)$$

Exercise 1.1. *Show* (1.2).

Thus we desire to understand the magnitude of the error probability $P_e(i,j)$ in (1.1) in the particular case when the conditional distribution of D given H is a multivariate Poisson distribution with mean vector on the order of $\ln(n)$. The following result determines this error up to first order in the exponent.

Lemma 1.2 (Abbe and Sandon [3]). *For any $c_1, c_2 \in (0, \infty)^k$ with $c_1 \neq c_2$ and $p_1, p_2 > 0$, we have*

$$\sum_{x \in \mathbb{Z}_+^k} \min\left\{ \mathcal{P}_{\ln(n)c_1}(x) p_1, \mathcal{P}_{\ln(n)c_2}(x) p_2 \right\} = O\left(n^{-D_+(c_1,c_2) - \frac{\ln\ln(n)}{2\ln(n)}} \right), \qquad (1.3)$$

$$\sum_{x \in \mathbb{Z}_+^k} \min\left\{ \mathcal{P}_{\ln(n)c_1}(x) p_1, \mathcal{P}_{\ln(n)c_2}(x) p_2 \right\} = \Omega\left(n^{-D_+(c_1,c_2) - \frac{k\ln\ln(n)}{2\ln(n)}} \right), \qquad (1.4)$$

where

$$D_+(c_1, c_2) = \max_{t \in [0,1]} \sum_{i \in [k]} \left(t c_1(i) + (1-t) c_2(i) - c_1(i)^t c_2(i)^{1-t} \right). \qquad (1.5)$$

We do not go over the proof of this statement—which we leave to the reader as a challenging exercise—but we provide some intuition in the univariate case. Figure 3 illustrates the probability mass function of two Poisson distributions, with
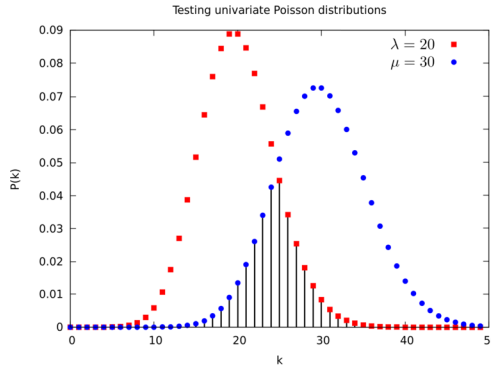

FIG 3. *Testing univariate Poisson distributions. The figure plots the probability mass function of two Poisson distributions, with means $\lambda = 20$ and $\mu = 30$, respectively.*

means $\lambda = 20$ and $\mu = 30$, respectively. Observe that $\min\{\mathcal{P}_\lambda(x), \mathcal{P}_\mu(x)\}$ decays rapidly away from $x_{\max} := \arg\max_{x \in \mathbb{Z}_+} \min\{\mathcal{P}_\lambda(x), \mathcal{P}_\mu(x)\}$, so we can obtain a good estimate of the sum $\sum_{x \in \mathbb{Z}_+} \min\{\mathcal{P}_\lambda(x), \mathcal{P}_\mu(x)\}$ by simply estimating the term $\min\{\mathcal{P}_\lambda(x_{\max}), \mathcal{P}_\mu(x_{\max})\}$. Now observe that $x_{\max}$ must satisfy $\mathcal{P}_\lambda(x_{\max}) \approx \mathcal{P}_\mu(x_{\max})$; after some algebra this is equivalent to $x_{\max} \approx \frac{\lambda-\mu}{\log(\lambda/\mu)}$. Let t^* denote the maximizer in the expression of $D_+(\lambda, \mu)$ in (1.5). By differentiating in t, we obtain that t^* satisfies $\lambda - \mu - \log(\lambda/\mu) \cdot \lambda^{t^*} \mu^{1-t^*} = 0$, and so $\lambda^{t^*} \mu^{1-t^*} = \frac{\lambda-\mu}{\log(\lambda/\mu)}$. Thus we see that $x_{\max} \approx \lambda^{t^*} \mu^{1-t^*}$, from which, after some algebra, we get that $\mathcal{P}_\lambda(x_{\max}) \approx \mathcal{P}_\mu(x_{\max}) \approx \exp(-D_+(\lambda, \mu))$.

The proof of (1.4) in the multivariate case follows along the same lines: the single term corresponding to

$$x_{\max} := \arg\max_{x \in \mathbb{Z}_+^k} \min\left\{\mathcal{P}_{\ln(n)c_1}(x), \mathcal{P}_{\ln(n)c_2}(x)\right\}$$

gives the lower bound. For the upper bound of (1.3) one has to show that the other terms do not contribute much more.

Exercise 1.2. *Prove Lemma 1.2.*

Our conclusion is thus that the error exponent in testing multivariate Poisson distributions is given by the explicit quantity D_+ in (1.5). The discussion in Section 1.2 then implies that D_+ plays an important role in the threshold for exact recovery. In particular, it intuitively follows from Lemma 1.2 that a necessary condition for exact recovery should be that

$$\min_{i,j \in [k], i \neq j} D_+\left((PQ)_i, (PQ)_j\right) \geq 1.$$

Suppose on the contrary that $D_+\left((PQ)_i, (PQ)_j\right) < 1$ for some i and j. This implies that the error probability in the testing problem is $\Omega(n^{\varepsilon-1})$ for some $\varepsilon > 0$ for all vertices in communities i and j. Since the number of vertices in these communities is linear in n, and most of the hypothesis testing problems are approximately independent, one expects there to be no error in the testing problems with probability at most $\left(1 - \Omega(n^{\varepsilon-1})\right)^{\Omega(n)} = \exp(-\Omega(n^\varepsilon)) = o(1)$.

1.4. Chernoff-Hellinger divergence

Before moving on to the threshold for exact recovery in the general SBM, we discuss connections of D_+ to other, well-known measures of divergence. Writing

$$D_t(\mu, \nu) := \sum_{x \in [k]} \left(t\mu(x) + (1-t)\nu(x) - \mu(x)^t \nu(x)^{1-t}\right)$$

we have that

$$D_+(\mu, \nu) = \max_{t \in [0,1]} D_t(\mu, \nu).$$

For any fixed t, D_t can be written as

$$D_t\left(\mu,\nu\right) = \sum_{x\in[k]} \nu\left(x\right) f_t\left(\frac{\mu\left(x\right)}{\nu\left(x\right)}\right),$$

where $f_t\left(x\right) = 1 - t + tx - x^t$, which is a convex function. Thus D_t is an f-divergence, part of a family of divergences that generalize the Kullback-Leibler (KL) divergence (also known as relative entropy), which is obtained for $f(x) = x\ln(x)$. The family of f-divergences with convex f share many useful properties, and hence have been widely studied in information theory and statistics. The special case of $D_{1/2}\left(\mu,\nu\right) = \frac{1}{2}\left\|\sqrt{\mu} - \sqrt{\nu}\right\|_2^2$ is known as the Hellinger divergence. The Chernoff divergence is defined as

$$C_*\left(\mu,\nu\right) = \max_{t\in(0,1)} -\log\sum_x \mu(x)^t \nu(x)^{1-t},$$

and so if μ and ν are probability vectors, then $D_+\left(\mu,\nu\right) = 1 - e^{-C_*(\mu,\nu)}$. Because of these connections, Abbe and Sandon termed D_+ the *Chernoff-Hellinger divergence*.

While the quantity D_+ still might seem mysterious, even in light of these connections, a useful point of view is that Lemma 1.2 gives D_+ an *operational meaning*.

1.5. Characterizing exact recoverability using CH-divergence

Going back to the exact recovery problem in the general SBM, let us jump right in and state the recoverability threshold of Abbe and Sandon: exact recovery in $\mathrm{SBM}(n, p, \ln(n)Q/n)$ is possible if and only if the CH-divergence between all pairs of community profiles is at least 1.

Theorem 1.3 (Abbe and Sandon [3]). *Let $k \in \mathbb{Z}_+$ denote the number of communities, let $p \in (0,1)^k$ with $\|p\|_1 = 1$ denote the community prior, let $P = \mathrm{diag}(p)$, and let $Q \in (0,\infty)^{k\times k}$ be a symmetric $k\times k$ matrix with no two rows equal. Exact recovery is solvable in $\mathrm{SBM}\left(n, p, \ln(n)Q/n\right)$ if and only if*

$$\min_{i,j\in[k],i\neq j} D_+\left((PQ)_i, (PQ)_j\right) \geq 1. \tag{1.6}$$

This theorem thus provides an operational meaning to the CH-divergence for the community recovery problem.

Example 1.4 (Symmetric communities). *Consider again k symmetric communities, that is, $p_i = 1/k$ for all $i \in [k]$, $Q_{i,j} = a$ if $i = j$, and $Q_{i,j} = b$ otherwise, with $a, b > 0$. Then exact recovery is solvable in $\mathrm{SBM}\left(n, p, \ln(n)Q/n\right)$ if and only if*

$$\left|\sqrt{a} - \sqrt{b}\right| \geq \sqrt{k}. \tag{1.7}$$

We note that in this case D_+ is the same as the Hellinger divergence.

Exercise 1.3. *Deduce from Theorem 1.3 that (1.7) gives the threshold in the example above.*

1.5.1. *Achievability*

Let us now see how Theorem 1.3 follows from the hypothesis testing results, starting with the achievability. When the condition (1.6) holds, then Lemma 1.2 tells us that in the hypothesis testing problem between Poisson distributions the error of the MAP estimate is $o(1/n)$. Thus if the setting of the thought experiment described in Section 1.2 applies to every vertex, then by looking at the degree profiles of the vertices we can correctly reclassify all vertices, and the probability that we make an error is $o(1)$ by a union bound. However, the setting of the thought experiment does not quite apply. Nonetheless, in this logarithmic degree regime it is possible to partially reconstruct the labels of the vertices, with only $o(n)$ vertices being misclassified. The details of this partial reconstruction procedure would require a separate lecture—in brief, it determines whether two vertices are in the same community or not by looking at how their $\log(n)$ size neighborhoods interact—so now we will take this for granted; we refer the interested reader to [3] for more.

It is possible to show that there exists a constant δ such that if one estimates the label of a vertex v based on classifications of its neighbors that are wrong with probability x, then the probability of misclassifying v is at most $n^{\delta x}$ times the probability of error if all the neighbors of v were classified correctly. The issue is that the standard partial recovery algorithm has a constant error rate for the classifications, thus the error rate of the degree profiling step could be n^c times as large as the error in the hypothesis testing problem, for some $c > 0$. This is an issue when $\min_{i \neq j} D_+ \left((PQ)_i, (PQ)_j \right) < 1 + c$.

To get around this, one can do multiple rounds of more accurate classifications. First, one obtains a partial reconstruction of the labels with an error rate that is a sufficiently low constant. After applying the degree-profiling step to each vertex, the classification error at each vertex is now $O(n^{-c'})$ for some $c' > 0$. Hence after applying another degree-profiling step to each vertex, the classification error at each vertex will now be at most $n^{\delta \times O(n^{-c'})} \times o(1/n) = o(1/n)$. Thus applying a union bound at this stage we can conclude that all vertices are correctly labelled whp.

1.5.2. *Impossibility*

The necessity of condition (1.6) was already described at a high level at the end of Section 1.3. Here we give some details on how to deal with the dependencies that arise.

Assume that (1.6) does not hold, and let i and j be two communities that violate the condition, i.e., for which $D_+ \left((PQ)_i, (PQ)_j \right) < 1$. We want to argue that vertices in communities i and j cannot all be distinguished, and so any classification algorithm has to make at least one error whp. An important fact that we use is that the lower bound (1.4) arises from a particular choice of degree profile that is both likely for the two communities. Namely, define the degree

profile x by

$$x_\ell = \left\lfloor (PQ)_{\ell,i}^{t^*} (PQ)_{\ell,j}^{1-t^*} \ln(n) \right\rfloor$$

for every $\ell \in [k]$, where $t^* \in [0,1]$ is the maximizer in $D_+ \left((PQ)_i, (PQ)_j \right)$, i.e., the value for which $D_+ \left((PQ)_i, (PQ)_j \right) = D_{t^*} \left((PQ)_i, (PQ)_j \right)$. Then Lemma 1.2 tells us that for any vertex in community i or j, the probability that it has degree profile x is at least

$$\Omega \left(n^{-D_+\left((PQ)_i, (PQ)_j \right)} / (\ln(n))^{k/2} \right),$$

which is at least $\Omega \left(n^{\varepsilon-1} \right)$ for some $\varepsilon > 0$ by assumption.

To show that this holds for many vertices in communities i and j at once, we first select a random set S of $n/(\ln(n))^3$ vertices. Whp the intersection of S with any community ℓ is within $\sqrt{n}$ of the expected value $p_\ell n/(\ln(n))^3$, and furthermore a randomly selected vertex in S is not connected to any other vertex in S. Thus the distribution of a vertex's degree profile excluding connections to vertices in S is essentially a multivariate Poisson distribution as before. We call a vertex in S *ambiguous* if for each $\ell \in [k]$ it has exactly x_ℓ neighbors in community ℓ that are not in S. By Lemma 1.2 we have that a vertex in S that is in community i or j is ambiguous with probability $\Omega \left(n^{\varepsilon-1} \right)$. By definition, for a fixed community assignment and choice of S, there is no dependence on whether two vertices are ambiguous. Furthermore, due to the choice of the size of S, whp there are at least $\ln(n)$ ambiguous vertices in community i and at least $\ln(n)$ ambiguous vertices in community j that are not adjacent to any other vertices in S. These $2\ln(n)$ are indistinguishable, so no algorithm classifies all of them correctly with probability greater than $1/\binom{2\ln(n)}{\ln(n)}$, which tends to 0 as $n \to \infty$.

1.5.3. The finest exact partition recoverable

We conclude by mentioning that this threshold generalizes to finer questions. If exact recovery is not possible, what is the finest partition that can be recovered? We say that exact recovery is solvable for a community partition $[k] = \sqcup_{s=1}^t A_s$, where A_s is a subset of $[k]$, if there exists an algorithm that whp assigns to every vertex an element of $\{A_1, \ldots, A_t\}$ that contains its true community. The finest partition that is exactly recoverable can also be expressed using CH-divergence in a similar fashion. It is the largest collection of disjoint subsets such that the CH-divergence between these subsets is at least 1, where the CH-divergence between two subsets is defined as the minimum of the CH-divergences between any two community profiles in these subsets.

Theorem 1.5 (Abbe and Sandon [3]). *Under the same settings as in Theorem 1.3, exact recovery is solvable in* $\mathrm{SBM}(n, p, \ln(n)Q/n)$ *for a partition* $[k] = \sqcup_{s=1}^t A_s$ *if and only if*

$$D_+ \left((PQ)_i, (PQ)_j \right) \geq 1$$

for every i and j in different subsets of the partition.

1.6. Outlook: open problems and extensions

There has been a huge recent interest in the SBM, initiated by the work of Decelle et al. [26], which has led to enormous advancements in this area. Nonetheless, plenty of exciting avenues for further research remain. We refer the reader to the recent survey by Abbe [1] for a thorough treatment of the SBM and recent developments, including a more detailed discussion of open problems; here we mention just a few of them.

In this lecture we discussed exact recovery in the case of linear-size communities. A natural open problem is to understand what happens in the case of sublinear-size communities: where do the phase transitions occur?

In the case of weak recovery, Abbe and Sandon [4] recently showed that in the case of $k \geq 4$ communities, it is possible to information-theoretically recover the communities below the so-called KS threshold. Is it possible to locate the precise information-theoretic threshold in this setting? Is it possible to show that the computational threshold is the KS threshold, thereby proving that there exists an information-computation gap?

In many applications, communities are not disjoint but overlapping. How well can we detect overlapping communities? What are the fundamental limits? To understand the utility of the SBM we must also study how robust the obtained results are to changes in the model. What can we say about variants of the SBM, such as including degree corrections or allowing adversaries to modify the graph?

2. Lecture 2: Estimating the dimension of a random geometric graph on a high-dimensional sphere

Many real-world networks have strong structural features and our goal is often to recover these hidden structures. In the previous lecture we studied the fundamental limits of inferring communities in the stochastic block model, a natural generative model for graphs with community structure. Another possibility is *geometric structure*. Many networks coming from physical considerations naturally have an underlying geometry, such as the network of major roads in a country. In other networks this stems from a latent feature space of the nodes. For instance, in social networks a person might be represented by a feature vector of their interests, and two people are connected if their interests are close enough; this latent metric space is referred to as the *social space* [36].

In such networks the natural questions probe the underlying geometry. Can one detect the presence of geometry? If so, can one estimate various aspects of the geometry, e.g., an appropriately defined dimension? In this lecture we study these questions in a particularly natural and simple generative model of a random geometric graph: n points are picked uniformly at random on the d-dimensional sphere, and two points are connected by an edge if and only if they are sufficently close.[4]

---

[4]This lecture is based on [19].

We are particularly interested in the *high-dimensional* regime, motivated by recent advances in all areas of applied mathematics, and in particular statistics and learning theory, where high-dimensional feature spaces are becoming the new norm. While the low-dimensional regime has been studied for a long time in probability theory [51], the high-dimensional regime brings about a host of new and interesting questions.

2.1. A simple random geometric graph model and basic questions

Let us now define more precisely the random geometric graph model we consider and the questions we study. In general, a geometric graph is such that each vertex is labeled with a point in some metric space, and an edge is present between two vertices if the distance between the corresponding labels is smaller than some prespecified threshold. We focus on the case where the underlying metric space is the Euclidean sphere $\mathbb{S}^{d-1} = \left\{ x \in \mathbb{R}^d : \|x\|_2 = 1 \right\}$, and the latent labels are i.i.d. uniform random vectors in $\mathbb{S}^{d-1}$. We denote this model by $G(n, p, d)$, where n is the number of vertices and p is the probability of an edge between two vertices (p determines the threshold distance for connection). This model is closely related to latent space approaches to social network analysis [36], and it has applications to remote sensing and finance as well [27]. Locally sparsified versions of related random geometric graphs also serve as models for wireless networks, see, e.g., [15, 16, 17] and the references therein.

Slightly more formally, $G(n, p, d)$ is defined as follows. Let $X_1, \ldots, X_n$ be independent random vectors, uniformly distributed on $\mathbb{S}^{d-1}$. In $G(n, p, d)$, distinct vertices $i \in [n]$ and $j \in [n]$ are connected by an edge if and only if $\langle X_i, X_j \rangle \geq t_{p,d}$, where the threshold value $t_{p,d} \in [-1, 1]$ is such that $\mathbb{P}\left(\langle X_1, X_2 \rangle \geq t_{p,d} \right) = p$. For example, when $p = 1/2$ we have $t_{p,d} = 0$.

The most natural random graph model without any structure is the standard Erdős-Rényi random graph $G(n, p)$, where any two of the n vertices are independently connected with probability p.

We can thus formalize the question of detecting underlying geometry as a simple hypothesis testing question. The null hypothesis is that the graph is drawn from the Erdős-Rényi model, while the alternative is that it is drawn from $G(n, p, d)$. In brief:

$$H_0 : G \sim G(n, p), \qquad H_1 : G \sim G(n, p, d). \tag{2.1}$$

To understand this question, the basic quantity we need to study is the total variation distance between the two distributions on graphs, $G(n, p)$ and $G(n, p, d)$, denoted by $\mathrm{TV}\left(G(n, p), G(n, p, d) \right)$; recall that the total variation distance between two probability measures P and Q is defined as $\mathrm{TV}\left(P, Q \right) = \frac{1}{2} \|P - Q\|_1 = \sup_A |P(A) - Q(A)|$. We are interested in particular in the case when the dimension d is *large*, growing with n.

It is intuitively clear that if the geometry is too high-dimensional, then it is impossible to detect it, while a low-dimensional geometry will have a strong

effect on the generated graph and will be detectable. How fast can the dimension grow with n while still being able to detect it? Most of this lecture will focus on this question.

If we can detect geometry, then it is natural to ask for more information. Perhaps the ultimate goal would be to find an embedding of the vertices into an appropriate dimensional sphere that is a *true representation*, in the sense that the geometric graph formed from the embedded points is indeed the original graph. More modestly, can the dimension be estimated? We touch on this question at the end of the lecture.

2.2. The dimension threshold for detecting underlying geometry

The high-dimensional setting of the random geometric graph $G(n, p, d)$ was first studied by Devroye, György, Lugosi, and Udina [27], who showed that if n is fixed and $d \to \infty$, then

$$\mathrm{TV}\left(G(n, p), G(n, p, d)\right) \to 0,$$

that is, geometry is indeed lost in high dimensions. More precisely, they show that this convergence happens when $d \gg n^7 2^{n^2/2}$.[5] This follows by observing that for fixed n, the multivariate central limit theorem implies that as $d \to \infty$, the inner products of the latent vectors converge in distribution to a standard Gaussian:

$$\left(\frac{1}{\sqrt{d}} \langle X_i, X_j \rangle\right)_{\{i,j\} \in \binom{[n]}{2}} \stackrel{d \to \infty}{\Longrightarrow} \mathcal{N}\left(0, I_{\binom{n}{2}}\right).$$

The Berry-Esseen theorem gives a convergence rate, which then allows to show that for any graph G on n vertices, $|\mathbb{P}\left(G(n, p) = G\right) - \mathbb{P}\left(G(n, p, d) = G\right)| = O\left(\sqrt{n^7/d}\right)$; the factor of $2^{n^2/2}$ comes from applying this bound to every term in the L_1 distance.

However, the result above is not tight, and we seek to understand the fundamental limits to detecting underlying geometry. The dimension threshold for dense graphs was recently found in [19], and it turns out that it is $d \approx n^3$, in the following sense.

Theorem 2.1 (Bubeck, Ding, Eldan, Rácz [19]). *Let $p \in (0, 1)$ be fixed. Then*

$$\mathrm{TV}\left(G(n, p), G(n, p, d)\right) \to \begin{cases} 0, & \text{if } d \gg n^3, \qquad (2.2) \\ 1, & \text{if } d \ll n^3. \qquad (2.3) \end{cases}$$

Moreover, in the latter case there exists a computationally efficient test to detect underlying geometry (with running time $O\left(n^3\right)$).

Most of the lecture will be devoted to understanding this theorem. At the end we will consider this same question for *sparse graphs* (where $p = c/n$), where determining the dimension threshold is an intriguing open problem.

---

[5]Throughout these notes we use standard asymptotic notation; for instance, $f(t) \ll g(t)$ as $t \to \infty$ if $\lim_{t \to \infty} f(t)/g(t) = 0$.

2.3. The triangle test

A natural test to uncover geometric structure is to count the number of triangles in G. Indeed, in a purely random scenario, vertex u being connected to both v and w says nothing about whether v and w are connected. On the other hand, in a geometric setting this implies that v and w are close to each other due to the triangle inequality, thus increasing the probability of a connection between them. This, in turn, implies that the expected number of triangles is larger in the geometric setting, given the same edge density. Let us now compute what this statistic gives us.
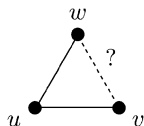
FIG 4. *Given that u is connected to both v and w, v and w are more likely to be connected under $G(n, p, d)$ than under $G(n, p)$.*

For a graph G, let A denote its adjacency matrix, i.e., $A_{i,j} = 1$ if vertices i and j are connected, and 0 otherwise. Then $T_G(i, j, k) := A_{i,j} A_{i,k} A_{j,k}$ is the indicator variable that three vertices i, j, and k form a triangle, and so the number of triangles in G is

$$T(G) := \sum_{\{i,j,k\} \in \binom{[n]}{3}} T_G(i, j, k).$$

By linearity of expectation, for both models the expected number of triangles is $\binom{n}{3}$ times the probability of a triangle between three specific vertices. For the Erdős-Rényi random graph the edges are independent, so the probability of a triangle is p^3, and thus we have

$$\mathbb{E}\left[T\left(G(n, p)\right)\right] = \binom{n}{3} p^3.$$

For $G(n, p, d)$ it turns out that for any fixed $p \in (0, 1)$ we have

$$\mathbb{P}\left(T_{G(n,p,d)}(1, 2, 3) = 1\right) \approx p^3 \left(1 + \frac{C_p}{\sqrt{d}}\right) \tag{2.4}$$

for some constant $C_p > 0$, which gives that

$$\mathbb{E}\left[T\left(G(n, p, d)\right)\right] \geq \binom{n}{3} p^3 \left(1 + \frac{C_p}{\sqrt{d}}\right).$$

Showing (2.4) is somewhat involved, but in essence it follows from the *concentration of measure* phenomenon on the sphere, namely that most of the mass on the high-dimensional sphere is located in a band of $O\left(1/\sqrt{d}\right)$ around the
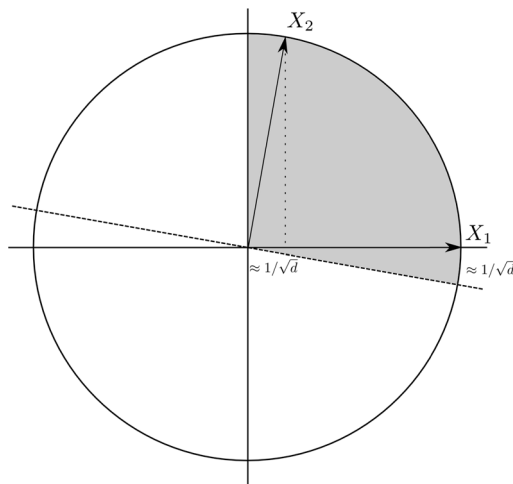
FIG 5. *If X_1 and X_2 are two independent uniform points on the d-dimensional sphere $\mathbb{S}^{d-1}$, then their inner product $\langle X_1, X_2 \rangle$ is on the order of $1/\sqrt{d}$ due to the concentration of measure phenomenon on the sphere. This then implies that the probability of a triangle in $G(n, 1/2, d)$ is $(1/2)^3 + c/\sqrt{d}$ for some constant $c > 0$.*

equator. We sketch here the main intuition for $p = 1/2$, which is illustrated in Figure 5.

Let X_1, X_2, and X_3 be independent uniformly distributed points in $\mathbb{S}^{d-1}$. Then

$$
\begin{aligned}
\mathbb{P}\left(T_{G(n,1/2,d)}(1,2,3) = 1\right) & \\
&= \mathbb{P}\left(\langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0, \langle X_2, X_3 \rangle \geq 0\right) \\
&= \mathbb{P}\left(\langle X_2, X_3 \rangle \geq 0 \,|\, \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0\right) \\
&\quad \times \mathbb{P}\left(\langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0\right) \\
&= \frac{1}{4} \times \mathbb{P}\left(\langle X_2, X_3 \rangle \geq 0 \,|\, \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0\right),
\end{aligned}
$$

where the last equality follows by independence. So what remains is to show that this latter conditional probability is approximately $1/2 + c/\sqrt{d}$. To compute this conditional probability what we really need to know is the typical angle is between X_1 and X_2. By rotational invariance we may assume that $X_1 = (1, 0, 0, \ldots, 0)$, and hence $\langle X_1, X_2 \rangle = X_2(1)$, the first coordinate of X_2. One way to generate X_2 is to sample a d-dimensional standard Gaussian and then normalize it by its length. Since the norm of a d-dimensional standard Gaussian is very well concentrated around $\sqrt{d}$, it follows that $X_2(1)$ is on the order of $1/\sqrt{d}$. Conditioned on $X_2(1) \geq 0$, this typical angle gives the boost in the conditional probability that we see. See Figure 5 for an illustration.

Thus we see that the boost in the number of triangles in the geometric setting is $\Theta\left(n^3/\sqrt{d}\right)$ in expectation:

$$\mathbb{E}\left[T\left(G(n,p,d)\right)\right] - \mathbb{E}\left[T\left(G(n,p)\right)\right] \geq \binom{n}{3}\frac{C_p}{\sqrt{d}}.$$

To be able to tell apart the two graph distributions based on the number of triangles, the boost in expectation needs to be much greater than the standard deviation.

Exercise 2.1. *Show that*

$$\mathrm{Var}\left(T\left(G\left(n,p\right)\right)\right) = \binom{n}{3}\left(p^3 - p^6\right) + \binom{n}{4}\binom{4}{2}\left(p^5 - p^6\right)$$

and that $\mathrm{Var}\left(T\left(G\left(n,p,d\right)\right)\right) \leq n^4$.

Exercise 2.2. *Show that if*

$$|\mathbb{E}\left[T\left(G(n,p,d)\right)\right] - \mathbb{E}\left[T\left(G(n,p)\right)\right]|$$
$$\gg \max\left\{\sqrt{\mathrm{Var}\left(T\left(G\left(n,p\right)\right)\right)}, \sqrt{\mathrm{Var}\left(T\left(G\left(n,p,d\right)\right)\right)}\right\},$$

then

$$\mathrm{TV}\left(G(n,p), G(n,p,d)\right) \to 1.$$

Putting together Exercises 2.1 and 2.2 we see that $\mathrm{TV}\left(G(n,p), G(n,p,d)\right) \to 1$ if $n^3/\sqrt{d} \gg \sqrt{n^4}$, which is equivalent to $d \ll n^2$.

2.4. *Signed triangles are more powerful*

While triangles detect geometry up until $d \ll n^2$, are there even more powerful statistics that detect geometry for larger dimensions? One can check that longer cycles also only work when $d \ll n^2$, as do several other natural statistics. Yet it turns out that the underlying geometry can be detected even when $d \ll n^3$.

The simple idea that leads to this improvement is to consider *signed triangles*. We have already noticed that triangles are more likely in the geometric setting than in the purely random setting. This also means that induced wedges (i.e., when there are exactly two edges among the three possible ones) are less likely in the geometric setting. Similarly, induced single edges are more likely, and induced independent sets on three vertices are less likely in the geometric setting. Figure 6 summarizes these observations.
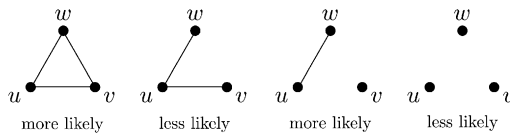

FIG 6. *This figure summarizes which patterns are more or less likely in the geometric set-ting than in the purely random setting. The signed triangles statistic reweights the different patterns with positive and negative weights.*

The signed triangles statistic incorporates these observations by giving the different patterns positive or negative weights. More precisely, we define

$$\tau\left(G\right) := \sum_{\{i,j,k\}\in\binom{[n]}{3}} \left(A_{i,j}-p\right)\left(A_{i,k}-p\right)\left(A_{j,k}-p\right).$$

The key insight motivating this definition is that the variance of signed triangles is *much smaller* than the variance of triangles, due to the cancellations introduced by the centering of the adjacency matrix: the $\Theta\left(n^4\right)$ term vanishes, leaving only the $\Theta\left(n^3\right)$ term.

Exercise 2.3. *Show that*
$$\mathbb{E}\left[\tau\left(G(n,p)\right)\right] = 0$$

and
$$\mathrm{Var}\left(\tau\left(G(n,p)\right)\right) = \binom{n}{3}p^3\left(1-p\right)^3.$$

On the other hand it can be shown that

$$\mathbb{E}\left[\tau\left(G(n,p,d)\right)\right] \geq c_p n^3/\sqrt{d}, \tag{2.5}$$

so the gap between the expectations remains. Furthermore, it can also be shown that the variance also decreases for $G(n,p,d)$ and we have

$$\mathrm{Var}\left(\tau\left(G(n,p,d)\right)\right) \leq n^3 + \frac{3n^4}{d}. \tag{2.6}$$

Putting everything together and using Exercise 2.2 for the signed triangles statistic τ, we get that $\mathrm{TV}\left(G(n,p),G(n,p,d)\right) \to 1$ if $n^3/\sqrt{d} \gg \sqrt{n^3+n^4/d}$, which is equivalent to $d \ll n^3$. This concludes the proof of (2.3) from Theorem 2.1.

2.5. Barrier to detecting geometry: when Wishart becomes GOE

We now turn to proving (2.2), which, together with (2.3), shows that the threshold dimension for detecting geometry is n^3. This also shows that the signed triangle statistic is near-optimal, since it can detect geometry whenever $d \ll n^3$.

There are essentially three main ways to bound the total variation of two distributions from above: (i) if the distributions have nice formulas associated with them, then exact computation is possible; (ii) through *coupling* the distributions; or (iii) by using inequalities between probability metrics to switch the problem to bounding a different notion of distance between the distributions. Here, while the distribution of $G(n,p,d)$ does not have a nice formula associated with it, the main idea is to view this random geometric graph as a function of an $n\times n$ Wishart matrix with d degrees of freedom—i.e., a matrix of inner products of n d-dimensional Gaussian vectors—denoted by $W(n,d)$. It turns out that one can view $G(n,p)$ as (essentially) the same function of an $n\times n$ GOE random matrix—i.e., a symmetric matrix with i.i.d. Gaussian entries on and above the

diagonal—denoted by $M(n)$. The upside of this is that both of these random matrix ensembles have explicit densities that allow for explicit computation. We explain this connection here in the special case of $p = 1/2$ for simplicity; see [19] for the case of general p.

Recall that if Y_1 is a standard normal random variable in $\mathbb{R}^d$, then $Y_1/\|Y_1\|$ is uniformly distributed on the sphere $\mathbb{S}^{d-1}$. Consequently we can view $G(n, 1/2, d)$ as a function of an appropriate Wishart matrix, as follows. Let Y be an $n \times d$ matrix where the entries are i.i.d. standard normal random variables, and let $W \equiv W(n, d) = YY^T$ be the corresponding $n \times n$ Wishart matrix. Note that $W_{ii} = \langle Y_i, Y_i \rangle = \|Y_i\|^2$ and so $\langle Y_i/\|Y_i\|, Y_j/\|Y_j\| \rangle = W_{ij}/\sqrt{W_{ii}W_{jj}}$. Thus the $n \times n$ matrix A defined as

$$A_{i,j} = \begin{cases} 1 & \text{if } W_{ij} \geq 0 \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

has the same law as the adjacency matrix of $G(n, 1/2, d)$. Denote the map that takes W to A by H, i.e., $A = H(W)$.

In a similar way we can view $G(n, 1/2)$ as a function of an $n \times n$ matrix drawn from the Gaussian Orthogonal Ensemble (GOE). Let $M(n)$ be a symmetric $n \times n$ random matrix where the diagonal entries are i.i.d. normal random variables with mean zero and variance 2, and the entries above the diagonal are i.i.d. standard normal random variables, with the entries on and above the diagonal all independent. Then $B = H(M(n))$ has the same law as the adjacency matrix of $G(n, p)$. Note that B only depends on the sign of the off-diagonal elements of $M(n)$, so in the definition of B we can replace $M(n)$ with $M(n, d) := \sqrt{d}M(n) + dI_n$, where I_n is the $n \times n$ identity matrix.

We can thus conclude that

$$\mathrm{TV}(G(n, 1/2, d), G(n, 1/2)) = \mathrm{TV}(H(W(n, d)), H(M(n, d)))$$
$$\leq \mathrm{TV}(W(n, d), M(n, d)).$$

The densities of these two random matrix ensembles are well known. Let $\mathcal{P} \subset \mathbb{R}^{n^2}$ denote the cone of positive semidefinite matrices. When $d \geq n$, $W(n, d)$ has the following density with respect to the Lebesgue measure on $\mathcal{P}$:

$$f_{n,d}(A) := \frac{(\det(A))^{\frac{1}{2}(d-n-1)} \exp\left(-\frac{1}{2}\mathrm{Tr}(A)\right)}{2^{\frac{1}{2}dn}\pi^{\frac{1}{4}n(n-1)}\prod_{i=1}^{n}\Gamma\left(\frac{1}{2}(d+1-i)\right)},$$

where $\mathrm{Tr}(A)$ denotes the trace of the matrix A. It is also known that the density of a GOE random matrix with respect to the Lebesgue measure on $\mathbb{R}^{n^2}$ is $A \mapsto (2\pi)^{-\frac{1}{4}n(n+1)} 2^{-\frac{n}{2}} \exp\left(-\frac{1}{4}\mathrm{Tr}(A^2)\right)$ and so the density of $M(n, d)$ with respect to the Lebesgue measure on $\mathbb{R}^{n^2}$ is

$$g_{n,d}(A) := \frac{\exp\left(-\frac{1}{4d}\mathrm{Tr}\left((A - dI_n)^2\right)\right)}{(2\pi d)^{\frac{1}{4}n(n+1)} 2^{\frac{n}{2}}}.$$

These explicit formulas allow for explicit calculations. In particular, one can show that the log-ratio of the densities is $o(1)$ with probability $1-o(1)$ according to the measure induced by $M(n, d)$. This follows from writing out the Taylor expansion of the log-ratio of the densities and using known results about the empirical spectral distribution of Wigner matrices (in particular that it converges to a semi-circle law). The outcome of the calculation is the following result, proven independently and simultaneously by Bubeck et al. and Jiang and Li.

Theorem 2.2 (Bubeck, Ding, Eldan, Rácz [19]; Jiang, Li [39]). *Define the random matrix ensembles $W(n, d)$ and $M(n, d)$ as above. If $d/n^3 \to \infty$, then*

$$\mathrm{TV}\left(W\left(n, d\right), M\left(n, d\right)\right) \to 0.$$

We conclude that it is impossible to detect underlying geometry whenever $d \gg n^3$.

2.6. Estimating the dimension

Until now we discussed *detecting* geometry. However, the insights gained above allow us to also touch upon the more subtle problem of *estimating* the underlying dimension d.

Dimension estimation can also be done by counting the "number" of signed triangles as in Section 2.4. However, here it is necessary to have a bound on the difference of the expected number of signed triangles between consecutive dimensions; the lower bound of (2.5) is not enough. Still, we believe that the right hand side of (2.5) should give the true value of the expected value for an appropriate constant c_p, and hence we expect to have that

$$\mathbb{E}\left[\tau\left(G(n, p, d)\right)\right] - \mathbb{E}\left[\tau\left(G(n, p, d+1)\right)\right] = \Theta\left(\frac{n^3}{d^{3/2}}\right). \tag{2.7}$$

Thus, using the variance bound in (2.6), we get that dimension estimation should be possible using signed triangles whenever $n^3/d^{3/2} \gg \sqrt{n^3 + n^4/d}$, which is equivalent to $d \ll n$.

Showing (2.7) for general p seems involved; Bubeck et al. showed that it holds for $p = 1/2$, which can be considered as a proof of concept. We thus have the following.

Theorem 2.3 (Bubeck, Ding, Eldan, Rácz [19]). *There exists a universal constant $C > 0$ such that for all integers n and $d_1 < d_2$, one has*

$$\mathrm{TV}\left(G(n, 1/2, d_1), G(n, 1/2, d_2)\right) \geq 1 - C\left(\frac{d_1}{n}\right)^2.$$

This result is tight, as demonstrated by a result of Eldan [29], which states that when $d \gg n$, the Wishart matrices $W(n, d)$ and $W(n, d+1)$ are indistinguishable. By the discussion in Section 2.5, this directly implies that $G(n, 1/2, d)$ and $G(n, 1/2, d+1)$ are indistinguishable.

Theorem 2.4 (Eldan [29]). *There exists a universal constant $C > 0$ such that for all integers $n < d$,*

$$\mathrm{TV}\left(G(n, 1/2, d), G(n, 1/2, d+1)\right) \leq \mathrm{TV}\left(W(n, d), W(n, d+1)\right)$$
$$\leq C\sqrt{\left(\frac{d+1}{d-n}\right)^2 - 1}.$$

2.7. The mysterious sparse regime

The discussion so far has focused on dense graphs, i.e., assuming $p \in (0, 1)$ is constant, where Theorem 2.1 tightly characterizes when the underlying geometry can be detected. The same questions are interesting for *sparse graphs* as well, where the average degree is constant or slowly growing with n. However, since there are so few edges, this regime is much more challenging.

It is again natural to consider the number of triangles as a way to distinguish between $G(n, c/n)$ and $G(n, c/n, d)$. A calculation shows that this statistic works whenever $d \ll \log^3(n)$.

Theorem 2.5 (Bubeck, Ding, Eldan, Rácz [19]). *Let $c > 0$ be fixed and assume $d/\log^3(n) \to 0$. Then*

$$\mathrm{TV}\left(G\left(n, \frac{c}{n}\right), G\left(n, \frac{c}{n}, d\right)\right) \to 1.$$

In contrast with the dense regime, in the sparse regime the signed triangle statistic τ does not give significantly more power than the triangle statistic T. This is because in the sparse regime, with high probability, the graph does not contain any 4-vertex subgraph with at least 5 edges, which is where the improvement comes from in the dense regime.

The authors also conjecture that $\log^3(n)$ is the correct order where the transition happens.

Conjecture 2.6 (Bubeck, Ding, Eldan, Rácz [19]). *Let $c > 0$ be fixed and assume $d/\log^3(n) \to \infty$. Then*

$$\mathrm{TV}\left(G\left(n, \frac{c}{n}\right), G\left(n, \frac{c}{n}, d\right)\right) \to 0.$$

The main reason for this conjecture is that, when $d \gg \log^3(n)$, $G(n, c/n)$ and $G(n, c/n, d)$ seem to be locally equivalent; in particular, they both have the same Poisson number of triangles asymptotically. Thus the only way to distinguish between them would be to find an emergent global property which is significantly different under the two models, but this seems unlikely to exist. Proving or disproving this conjecture remains a challenging open problem. The best known bound is n^3 from (2.2) (which holds uniformly over p).

2.8. Outlook: open problems and extensions

High-dimensional random geometric graphs promise to be a rich source of exciting open problems. The main open problem that this lecture leaves open is understanding the critical dimension in the sparse setting; see Conjecture 2.6 above. More generally, it is interesting to understand how the critical dimension depends on p.

Understanding the robustness of these results is important—this is also the topic of the next lecture, in the setting of the Wishart to GOE transition for random matrices. Recently Eldan and Mikulincer studied the effect of anisotropy on the power of detecting geometry in random geometric graphs [30]. The authors introduce new notions of dimensionality and prove a theorem similar to Theorem 2.1 with appropriate upper and lower bounds on the "effective critical dimension".

Perhaps the ultimate goal is to find good representations of network data, and hence to faithfully embed the graph of interest into an appropriate metric space. While the worst-case version of this problem—recognizing if a graph can be realized as a geometric graph—is known to be NP-hard [14], the probabilistic setting should provide lots of opportunities for exciting research.

3. Lecture 3: Introduction to entropic central limit theorems and a proof of the fundamental limits of dimension estimation in random geometric graphs

Recall from the previous lecture that the dimension threshold for detecting geometry in $G(n, p, d)$ for constant $p \in (0, 1)$ is $d = \Theta\left(n^3\right)$. What if the random geometric graph model is not $G(n, p, d)$? How robust are the results presented in the previous lecture? We have seen that the detection threshold is intimately connected to the threshold of when a Wishart matrix becomes GOE. Understanding the robustness of this result on random matrices is interesting in its own right, and this is what we will pursue in this lecture.[6] Doing so also gives us the opportunity to learn about the fascinating world of entropic central limit theorems.

3.1. Setup and main result: the universality of the threshold dimension

Let $\mathbb{X}$ be an $n \times d$ random matrix with i.i.d. entries from a distribution μ that has mean zero and variance 1. The $n \times n$ matrix $\mathbb{X}\mathbb{X}^T$ is known as the Wishart matrix with d degrees of freedom. As we have seen in the previous lecture, this arises naturally in geometry, where $\mathbb{X}\mathbb{X}^T$ is known as the Gram matrix of inner products of n points in $\mathbb{R}^d$. The Wishart matrix also appears naturally in statistics as the sample covariance matrix, where d is the number of samples

---

[6]This lecture is based on [21].

and n is the number of parameters.[7] We refer to [21] for further applications in quantum physics, wireless communications, and optimization.

We consider the Wishart matrix with the diagonal removed, and scaled appropriately:

$$\mathcal{W}_{n,d} = \frac{1}{\sqrt{d}} \left(\mathbb{X}\mathbb{X}^T - \mathrm{diag}\left(\mathbb{X}\mathbb{X}^T \right) \right).$$

In many applications—such as to random graphs, as we have seen in the previous lecture—the diagonal of the matrix is not relevant, so removing it does not lose information. Our goal is to understand how large does the dimension d have to be so that $\mathcal{W}_{n,d}$ is approximately like $\mathcal{G}_n$, which is defined as the $n \times n$ Wigner matrix with zeros on the diagonal and i.i.d. standard Gaussians above the diagonal. In other words, $\mathcal{G}_n$ is drawn from the Gaussian Orthogonal Ensemble (GOE) with the diagonal replaced with zeros.

A simple application of the multivariate central limit theorem gives that if n is fixed and $d \to \infty$, then $\mathcal{W}_{n,d}$ converges to $\mathcal{G}_n$ in distribution. The main result of Bubeck and Ganguly [21] establishes that this holds as long as $d \widetilde{\gg} n^3$ under rather general conditions on the distribution μ.

Theorem 3.1 (Bubeck and Ganguly [21]). *If the distribution μ is log-concave[8] and $\frac{d}{n^3 \log^2(d)} \to \infty$, then*

$$\mathrm{TV}\left(\mathcal{W}_{n,d}, \mathcal{G}_n \right) \to 0. \tag{3.1}$$

On the other hand, if μ has a finite fourth moment and $\frac{d}{n^3} \to 0$, then

$$\mathrm{TV}\left(\mathcal{W}_{n,d}, \mathcal{G}_n \right) \to 1. \tag{3.2}$$

This result extends Theorems 2.1 and 2.2, and establishes n^3 as the universal critical dimension (up to logarithmic factors) for sufficiently smooth measures μ: $\mathcal{W}_{n,d}$ is approximately Gaussian if and only if d is much larger than n^3. For random graphs, as seen in Lecture 2, this is the dimension barrier to extracting geometric information from a network: if the dimension is much greater than the cube of the number of vertices, then all geometry is lost. In the setting of statistics this means that the Gaussian approximation of a Wishart matrix is valid as long as the sample size is much greater than the cube of the number of parameters. Note that for some statistics of a Wishart matrix the Gaussian approximation is valid for much smaller sample sizes (e.g., the largest eigenvalue behaves as in the limit even when the number of parameters is on the same order as the sample size [42]).

To distinguish the random matrix ensembles, we have seen in Lecture 2 that signed triangles work up until the threshold dimension in the case when μ is standard normal. It turns out that the same statistic works in this more general setting; when the entries of the matrices are centered, this statistic can be written as $A \mapsto \mathrm{Tr}\left(A^3 \right)$. Similarly to the calculations in Section 2.4, one can

---

[7]In statistics the number of samples is usually denoted by n, and the number of parameters is usually denoted by p; here our notation is taken with the geometric perspective in mind.

[8]A measure μ with density f is said to be log-concave if $f(\cdot) = e^{-\varphi(\cdot)}$ for some convex function φ.

show that under the two measures $\mathcal{W}_{n,d}$ and $\mathcal{G}_n$, the mean of $\mathrm{Tr}\left(A^3\right)$ is 0 and $\Theta\left(n^3/\sqrt{d}\right)$, respectively, whereas the variances are $\Theta\left(n^3\right)$ and $\Theta\left(n^3 + n^5/d^2\right)$, respectively. Then (3.2) follows by an application of Chebyshev's inequality. We leave the details as an exercise for the reader.

We note that for (3.1) to hold it is necessary to have some smoothness assumption on the distribution μ. For instance, if μ is purely atomic, then so is the distribution of $\mathcal{W}_{n,d}$, and thus its total variation distance to $\mathcal{G}_n$ is 1. The log-concave assumption gives this necessary smoothness, and it is an interesting open problem to understand how far this can be relaxed.

3.2. Pinsker's inequality: from total variation to relative entropy

Our goal is now to bound the total variation distance $\mathrm{TV}\left(\mathcal{W}_{n,d}, \mathcal{G}_n\right)$ from above. In the general setting considered here there is no nice formula for the density of the Wishart ensemble, so $\mathrm{TV}\left(\mathcal{W}_{n,d}, \mathcal{G}_n\right)$ cannot be computed directly. Coupling these two random matrices also seems challenging.

In light of these observations, it is natural to switch to a different metric on probability distributions that is easier to handle in this case. We refer the reader to the excellent paper [34] which gathers ten different probability metrics and many relations between then. Here we use Pinsker's inequality to switch to relative entropy:

$$\mathrm{TV}\left(\mathcal{W}_{n,d}, \mathcal{G}_n\right)^2 \leq \frac{1}{2}\mathrm{Ent}\left(\mathcal{W}_{n,d}\,\|\,\mathcal{G}_n\right), \tag{3.3}$$

where $\mathrm{Ent}\left(\mathcal{W}_{n,d}\,\|\,\mathcal{G}_n\right)$ denotes the relative entropy of $\mathcal{W}_{n,d}$ with respect to $\mathcal{G}_n$. In the following subsection we provide a brief introduction to entropy; the reader familiar with the basics can safely skip this. We then turn to entropic central limit theorems and techniques involved in their proof, before finally coming back to bounding the right hand side in (3.3).

3.3. A brief introduction to entropy

The *entropy* of a discrete random variable X taking values in $\mathcal{X}$ is defined as

$$H(X) \equiv H(p) = -\sum_{x\in\mathcal{X}} p\left(x\right)\log\left(p\left(x\right)\right),$$

where p denotes the probability mass function of X. The log is commonly taken to have base 2, in which case entropy is measured in bits; if one considers the natural logarithm ln then it is measured in nats. Note that entropy is always nonnegative, since $p(x) \leq 1$ for every $x \in \mathcal{X}$. This is a measure of uncertainty of a random variable. It measures how much information is required on average to describe the random variable. Many properties of entropy agree with the intuition of what a measure of information should be. A useful way of thinking about entropy is the following: if we have an i.i.d. sequence of random variables and we know that the source distribution is p, then we can construct a code with average description length $H(p)$.

Example 3.2. *If X is uniform on a finite space $\mathcal{X}$, then $H(X) = \log |\mathcal{X}|$.*

The *conditional entropy* of Y given X is defined as

$$H\left(Y \mid X\right) = \sum_{x \in \mathcal{X}} p\left(x\right) H\left(Y \mid X = x\right),$$

where p denotes the probability mass function of X and $H\left(Y \mid X = x\right)$ is the entropy of the random variable Y conditioned on the event $\{X = x\}$. It measures how much information is required on average to describe the random variable Y, given that the value of the random variable X is known.

For continuous random variables the *differential entropy* is defined as

$$h\left(X\right) \equiv h\left(f\right) = -\int f\left(x\right) \log f\left(x\right) dx,$$

where f is the density of the random variable X.

Example 3.3. *If X is uniform on the interval $[0, a]$, then $h(X) = \log\left(a\right)$. If X is Gaussian with mean zero and variance σ^2, then $h(X) = \frac{1}{2}\log\left(2\pi e \sigma^2\right)$.*

Note that these examples show that differential entropy can be negative. One way to think of differential entropy is to think of $2^{h(X)}$ as "the volume of the support".

The *relative entropy* (also known as *Kullback-Leibler divergence*) of two distributions P and Q on a discrete space $\mathcal{X}$ is defined as

$$D\left(P \,\|\, Q\right) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

For two distributions with densities f and g the relative entropy is defined as

$$D\left(f \,\|\, g\right) = \int_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}.$$

Relative entropy is always nonnegative; this follows from Jensen's inequality. Relative entropy can be interpreted as a measure of distance between two distributions, although it is not a metric: it is not symmetric and it does not obey the triangle inequality. It can be thought of as a measure of inefficiency of assuming that the source distribution is q when it is really p. If we use a code for distribution q but the source is really from p, then we need $H(p) + D(p \,\|\, q)$ bits on average to describe the random variable.

In the following we use Ent to denote all notions of entropy and relative entropy. We also slightly abuse notation and interchangeably use a random variable or its law in the argument of entropy and relative entropy.

Entropy and relative entropy satisfy useful chain rules; we leave the proof of the following identities as an exercise for the reader. For entropy we have:

$$\text{Ent}\left(X_1, X_2\right) = \text{Ent}\left(X_1\right) + \text{Ent}\left(X_2 \mid X_1\right).$$

For relative entropy we have:

$$\mathrm{Ent}\left((Y_1, Y_2) \,\|\, (Z_1, Z_2)\right) = \mathrm{Ent}\left(Y_1 \,\|\, Z_1\right) + \mathbb{E}_{y \sim \lambda_1} \mathrm{Ent}\left(Y_2 \,|\, Y_1 = y \,\|\, Z_2 \,|\, Z_1 = y\right),$$
$$(3.4)$$

where λ_1 is the marginal distribution of Y_1 and $Y_2 \,|\, Y_1 = y$ denotes the distribution of Y_2 conditionally on the event $\{Y_1 = y\}$.

Let ϕ denote the density of γ_n, the n-dimensional standard Gaussian distribution, and let f be an isotropic density with mean zero, i.e., a density for which the covariance matrix is the identity I_n. Then

$$0 \leq \mathrm{Ent}\left(f \,\|\, \phi\right) = \int f \log f - \int f \log \phi$$
$$= \int f \log f - \int \phi \log \phi = \mathrm{Ent}\left(\phi\right) - \mathrm{Ent}\left(f\right),$$

where the second equality follows from the fact that $\log \phi\left(x\right)$ is quadratic in x, and the first two moments of f and ϕ are the same by assumption. We thus see that the standard Gaussian maximizes entropy among isotropic densities.

3.4. An introduction to entropic CLTs

At this point we are ready to state the entropic central limit theorem. The central limit theorem states that if $Z_1, Z_2, \ldots$ are i.i.d. real-valued random variables with zero mean and unit variance, then $S_m := \left(Z_1 + \cdots + Z_m\right)/\sqrt{m}$ converges in distribution to a standard Gaussian random variable as $m \to \infty$. There are many other senses in which S_m converges to a standard Gaussian, the entropic CLT being one of them.

Theorem 3.4 (Entropic CLT). *Let $Z_1, Z_2, \ldots$ be i.i.d. real-valued random variables with zero mean and unit variance, and let $S_m := \left(Z_1 + \cdots + Z_m\right)/\sqrt{m}$. If $\mathrm{Ent}\left(Z_1 \,\|\, \phi\right) < \infty$, then*

$$\mathrm{Ent}\left(S_m\right) \nearrow \mathrm{Ent}\left(\phi\right)$$

as $m \to \infty$. Moreover, the entropy of S_m increases monotonically, that is, $\mathrm{Ent}\left(S_m\right) \leq \mathrm{Ent}\left(S_{m+1}\right)$ for every $m \geq 1$.

The condition $\mathrm{Ent}\left(Z_1 \,\|\, \phi\right) < \infty$ is necessary for an entropic CLT to hold; for instance, if the Z_i are discrete, then $h\left(S_m\right) = -\infty$ for all m.

The entropic CLT originates with Shannon in the 1940s and was first proven by Linnik [45] in 1959 (without the monotonicity part of the statement). The first proofs that gave explicit convergence rates were given independently and at roughly the same time by Artstein, Ball, Barthe, and Naor [7, 5, 6], and Johnson and Barron [41] in the early 2000s, using two different techniques.

The fact that $\mathrm{Ent}\left(S_1\right) \leq \mathrm{Ent}\left(S_2\right)$ follows from the entropy power inequality, which goes back to Shannon [56] in 1948. This implies that $\mathrm{Ent}\left(S_m\right) \leq \mathrm{Ent}\left(S_{2m}\right)$ for all $m \geq 0$, and so it was naturally conjectured that $\mathrm{Ent}\left(S_m\right)$ increases monotonically. However, proving this turned out to be challenging. Even the inequality $\mathrm{Ent}\left(S_2\right) \leq \mathrm{Ent}\left(S_3\right)$ was unknown for over fifty years, until Artstein, Ball, Barthe, and Naor [5] proved in general that $\mathrm{Ent}\left(S_m\right) \leq \mathrm{Ent}\left(S_{m+1}\right)$ for all $m \geq 1$.

In the following we sketch some of the main ideas that go into the proof of these results, in particular following the techniques of Artstein, Ball, Barthe, and Naor [7, 5, 6].

3.5. *From relative entropy to Fisher information*

Our goal is to show that some random variable Z, which is a convolution of many i.i.d. random variables, is close to a Gaussian G. One way to approach this is to *interpolate* between the two. There are several ways of doing this; for our purposes interpolation along the Ornstein-Uhlenbeck semigroup is most useful. Define

$$P_t Z := e^{-t} Z + \sqrt{1 - e^{-2t}} G$$

for $t \in [0, \infty)$, and let f_t denote the density of $P_t Z$. We have $P_0 Z = Z$ and $P_\infty Z = G$. This semigroup has several desirable properties. For instance, if the density of Z is isotropic, then so is f_t. Before we can state the next desirable property that we will use, we need to introduce a few more useful quantities.

For a density function $f : \mathbb{R}^n \to \mathbb{R}_+$, let

$$\mathcal{I}(f) := \int \frac{\nabla f (\nabla f)^T}{f} = \mathbb{E}\left[(\nabla \log f)(\nabla \log f)^T\right]$$

be the *Fisher information matrix*. Let μ denote the mean of the density f. The Cramér-Rao bound states that for any unbiased estimator $\widehat{\mu}$ of μ, the covariance matrix $\text{Cov}(\widehat{\mu}) = \mathbb{E}\left[(\widehat{\mu} - \mu)(\widehat{\mu} - \mu)^T\right]$ satisfies

$$\text{Cov}(\widehat{\mu}) \succeq \mathcal{I}(f)^{-1},$$

that is, $\text{Cov}(\widehat{\mu}) - \mathcal{I}(f)^{-1}$ is positive semidefinite. The *Fisher information* is defined as

$$I(f) := \text{Tr}(\mathcal{I}(f)).$$

It is sometimes more convenient to work with the Fisher information distance, defined as $J(f) := I(f) - I(\phi) = I(f) - n$. Similarly to the discussion above, one can show that the standard Gaussian minimizes the Fisher information among isotropic densities, and hence the Fisher information distance is always nonnegative.

Now we are ready to state the De Bruijn identity [57], which characterizes the change of entropy along the Ornstein-Uhlenbeck semigroup via the Fisher information distance:

$$\partial_t \text{Ent}(f_t) = J(f_t).$$

This implies that the relative entropy between f and ϕ—which is our quantity of interest—can be expressed as follows:

$$\text{Ent}(f \,\|\, \phi) = \text{Ent}(\phi) - \text{Ent}(f) = \int_0^\infty J(f_t)\, dt. \tag{3.5}$$

Thus our goal is to bound the Fisher information distance $J(f_t)$.

3.6. Bounding the Fisher information distance

We first recall a classical result by Blachman [11] and Stam [57] that shows that Fisher information decreases under convolution.

Theorem 3.5 (Blachman [11]; Stam [57])**.** *Let $Y_1, \ldots, Y_d$ be independent random variables taking values in $\mathbb{R}$, and let $a \in \mathbb{R}^d$ be such that $\|a\|_2 = 1$. Then*

$$I\left(\sum_{i=1}^d a_i Y_i\right) \leq \sum_{i=1}^d a_i^2 I(Y_i).$$

In the i.i.d. case, this bound becomes $\|a\|_2^2 I(Y_1)$.

Artstein, Ball, Barthe, and Naor [7, 5] gave the following variational characterization of the Fisher information, which gives a particularly simple proof of Theorem 3.5.

Theorem 3.6 (Variational characterization of Fisher information [7, 5])**.** *Let $w : \mathbb{R}^d \to (0, \infty)$ be a sufficiently smooth[9] density on $\mathbb{R}^d$, let $a \in \mathbb{R}^d$ be a unit vector, and let h be the marginal of w in direction a. Then we have*

$$I(h) \leq \int_{\mathbb{R}^d} \left(\frac{\mathrm{div}(pw)}{w}\right)^2 w \tag{3.6}$$

for any continuously differentiable vector field $p : \mathbb{R}^d \to \mathbb{R}^d$ with the property that for every x, $\langle p(x), a \rangle = 1$. Moreover, if w satisfies $\int \|x\|^2 w(x) < \infty$, then there is equality for some suitable vector field p.

The Blachman-Stam theorem follows from this characterization by taking the constant vector field $p \equiv a$. Then we have $\mathrm{div}(pw) = \langle \nabla w, a \rangle$, and so the right hand side of (3.6) becomes $a^T \mathcal{I}(w) a$, where recall that $\mathcal{I}$ is the Fisher information matrix. In the setting of Theorem 3.5 the density w of $(Y_1, \ldots, Y_d)$ is a product density: $w(x_1, \ldots, x_d) = f_1(x_1) \times \cdots \times f_d(x_d)$, where f_i is the density of Y_i. Consequently the Fisher information matrix is a diagonal matrix, $\mathcal{I}(w) = \mathrm{diag}(I(f_1), \ldots, I(f_d))$, and thus $a^T \mathcal{I}(w) a = \sum_{i=1}^d a_i^2 I(f_i)$, concluding the proof of Theorem 3.5 using Theorem 3.6.

Given the characterization of Theorem 3.6, one need not take the vector field to be constant; one can obtain more by optimizing over the vector field. Doing this leads to the following theorem, which gives a rate of decrease of the Fisher information distance under convolutions.

Theorem 3.7 (Artstein, Ball, Barthe, and Naor [7, 5, 6])**.** *Let $Y_1, \ldots, Y_d$ be i.i.d. random variables with a density having a positive spectral gap c.[10] Then*

---

[9]It is enough that w is continuously twice differentiable and satisfies $\int \|\nabla w\|^2 / w < \infty$ and $\int \|\mathrm{Hess}(w)\| < \infty$.

[10]We say that a random variable has spectral gap c if for every sufficiently smooth g, we have $\mathrm{Var}(g) \leq \frac{1}{c} \mathbb{E} g'^2$. In particular, log-concave random variables have a positive spectral gap, see [12].

for any $a \in \mathbb{R}^d$ with $\|a\|_2 = 1$ we have that

$$J\left(\sum_{i=1}^{d} a_i Y_i\right) \leq \frac{2\|a\|_4^4}{c + (2-c)\|a\|_4^4} J(Y_1).$$

When $a = \frac{1}{\sqrt{d}}\mathbf{1}$, then $\frac{2\|a\|_4^4}{c+(2-c)\|a\|_4^4} = O(1/d)$, and thus using (3.5) we obtain a rate of convergence of $O(1/d)$ in the entropic CLT.

A result similar to Theorem 3.7 was proven independently and roughly at the same time by Johnson and Barron [41] using a different approach involving score functions.

3.7. A high-dimensional entropic CLT

The techniques of Artstein, Ball, Barthe, and Naor [7, 5, 6] generalize to higher dimensions, as was recently shown by Bubeck and Ganguly [21]. A result similar to Theorem 3.7 can be proven, from which a high-dimensional entropic CLT follows, together with a rate of convergence, by using (3.5) again.

Theorem 3.8 (Bubeck and Ganguly [21]). *Let $Y \in \mathbb{R}^d$ be a random vector with i.i.d. entries from a distribution ν with zero mean, unit variance, and spectral gap $c \in (0,1]$. Let $A \in \mathbb{R}^{n \times d}$ be a matrix such that $AA^T = I_n$, the $n \times n$ identity matrix. Let $\varepsilon = \max_{i \in [d]} (A^T A)_{i,i}$ and $\zeta = \max_{i,j \in [d], i \neq j} \left|(A^T A)_{i,j}\right|$. Then we have that*

$$\text{Ent}(AY \,\|\, \gamma_n) \leq n \min\left\{2\left(\varepsilon + \zeta^2 d\right)/c, 1\right\} \text{Ent}(\nu \,\|\, \gamma_1),$$

where γ_n denotes the standard Gaussian measure in R^n.

To interpret this result, consider the case where the matrix A is built by picking rows one after the other uniformly at random on the Euclidean sphere in $\mathbb{R}^d$, conditionally on being orthogonal to previous rows (to satisfy the isotropicity condition $AA^T = I_n$). We then expect to have $\varepsilon \simeq n/d$ and $\zeta \simeq \sqrt{n}/d$ (we leave the details as an exercise for the reader), and so Theorem 3.8 tells us that $\text{Ent}(AY \,\|\, \gamma_n) \lesssim n^2/d$.

3.8. Back to Wishart and GOE

We now turn our attention back to bounding the relative entropy $\text{Ent}(\mathcal{W}_{n,d} \,\|\, \mathcal{G}_n)$ between the $n \times n$ Wishart matrix with d degrees of freedom (with the diagonal removed), $\mathcal{W}_{n,d}$, and the $n \times n$ GOE matrix (with the diagonal removed), $\mathcal{G}_n$; recall (3.3). Since the Wishart matrix contains the (scaled) inner products of n vectors in $\mathbb{R}^d$, it is natural to relate $\mathcal{W}_{n+1,d}$ and $\mathcal{W}_{n,d}$, since the former comes from the latter by adding an additional d-dimensional vector to the n vectors already present. Specifically, we have the following:

$$\mathcal{W}_{n+1,d} = \begin{pmatrix} \mathcal{W}_{n,d} & \frac{1}{\sqrt{d}}\mathbb{X}X \\ \frac{1}{\sqrt{d}}(\mathbb{X}X)^T & 0 \end{pmatrix},$$

where X is a d-dimensional random vector with i.i.d. entries from μ, which are also independent from $\mathbb{X}$. Similarly we can write the matrix $\mathcal{G}_{n+1}$ using $\mathcal{G}_n$:

$$\mathcal{G}_{n+1} = \begin{pmatrix} \mathcal{G}_n & \gamma_n \\ \gamma_n^T & 0 \end{pmatrix}.$$

This naturally suggests to use the chain rule for relative entropy and bound $\mathrm{Ent}\left(\mathcal{W}_{n,d} \,\|\, \mathcal{G}_n\right)$ by induction on n. By (3.4) we get that

$$\mathrm{Ent}\left(\mathcal{W}_{n+1,d} \,\|\, \mathcal{G}_{n+1}\right) = \mathrm{Ent}\left(\mathcal{W}_{n,d} \,\|\, \mathcal{G}_n\right) + \mathbb{E}_{W_{n,d}}\left[\mathrm{Ent}\left(\tfrac{1}{\sqrt{d}}\mathbb{X}X \,|\, \mathcal{W}_{n,d} \,\|\, \gamma_n\right)\right].$$

By convexity of the relative entropy we also have that

$$\mathbb{E}_{W_{n,d}}\left[\mathrm{Ent}\left(\tfrac{1}{\sqrt{d}}\mathbb{X}X \,|\, \mathcal{W}_{n,d} \,\|\, \gamma_n\right)\right] \le \mathbb{E}_{\mathbb{X}}\left[\mathrm{Ent}\left(\tfrac{1}{\sqrt{d}}\mathbb{X}X \,|\, \mathbb{X} \,\|\, \gamma_n\right)\right].$$

Thus our goal is to understand and bound $\mathrm{Ent}\left(AX \,\|\, \gamma_n\right)$ for $A \in \mathbb{R}^{n \times d}$, and then apply the bound to $A = \frac{1}{\sqrt{d}}\mathbb{X}$ (followed by taking expectation over $\mathbb{X}$). This is precisely what was done in Theorem 3.8, the high-dimensional entropic CLT, for A satisfying $AA^T = I_n$. Since $A = \frac{1}{\sqrt{d}}\mathbb{X}$ does not necessarily satisfy $AA^T = I_n$, we have to correct for the lack of isotropicity. This is the content of the following lemma, the proof of which we leave as an exercise for the reader.

Lemma 3.9 ([21]). *Let $A \in \mathbb{R}^{n \times d}$ and $Q \in \mathbb{R}^{n \times n}$ be such that $QA\left(QA\right)^T = I_n$. Then for any isotropic random variable X taking values in $\mathbb{R}^d$ we have that*

$$\mathrm{Ent}\left(AX \,\|\, \gamma_n\right) = \mathrm{Ent}\left(QAX \,\|\, \gamma_n\right) + \frac{1}{2}\mathrm{Tr}\left(AA^T\right) - \frac{n}{2} + \frac{1}{2}\log\left|\det\left(Q\right)\right|. \quad (3.7)$$

We then apply this lemma with $A = \frac{1}{\sqrt{d}}\mathbb{X}$ and $Q = \left(\frac{1}{d}\mathbb{X}\mathbb{X}^T\right)^{-1/2}$. Observe that $\mathbb{E}\mathrm{Tr}\left(AA^T\right) = \frac{1}{d}\mathbb{E}\mathrm{Tr}\left(\mathbb{X}\mathbb{X}^T\right) = \frac{1}{d} \times n \times d = n$, and hence in expectation the middle two terms of the right hand side of (3.7) cancel each other out.

The last term in (3.7), $-\frac{1}{4}\log\det\left(\frac{1}{d}\mathbb{X}\mathbb{X}^T\right)$, should be understood as the relative entropy between a centered Gaussian with covariance given by $\frac{1}{d}\mathbb{X}\mathbb{X}^T$ and a standard Gaussian in $\mathbb{R}^n$. Controlling the expectation of this term requires studying the probability that $\mathbb{X}\mathbb{X}^T$ is close to being non-invertible, which requires bounds on the left tail of the smallest singular of $\mathbb{X}$. Understanding the extreme singular values of random matrices is a fascinating topic, but it is outside of the scope of these notes, and so we refer the reader to [21] for more details on this point.

Finally, the high-dimensional entropic CLT can now be applied to see that $\mathrm{Ent}\left(QAX \,\|\, \gamma_n\right) \lesssim n^2/d$. From the induction on n we get another factor of n, arriving at $\mathrm{Ent}\left(\mathcal{W}_{n,d} \,\|\, \mathcal{G}_n\right) \lesssim n^3/d$. We conclude that the dimension threshold is $d \approx n^3$, and the information-theoretic proof that we have outlined sheds light on why this threshold is n^3.

3.9. Outlook: open problems and extensions

The work of Bubeck and Ganguly [21] leaves open many questions; we mention here some of them and refer to [21] for more. Perhaps the main open problem is to find the optimal conditions under which the phase transition described in Theorem 3.1 holds. For instance, does it hold if the rows (or columns) of $\mathbb{X}$ are i.i.d. from a log-concave distribution in $\mathbb{R}^d$ (or $\mathbb{R}^n$)? Several estimates still work under this assumption, but it seems that the induction step of the proof requires the independence of the entries of $\mathbb{X}$.

Another interesting question is whether Eldan's result on Wishart matrices, Theorem 2.4, can be extended to Wishart matrices with more general entries. One difficulty is that the tools described in this lecture are based on measuring relative entropy with respect to a standard Gaussian (since this maximizes entropy), while the problem concerns only Wishart matrices.

A different direction of inquiry is to study higher order interactions. Denoting the ith column of $\mathbb{X}$ by $\mathbb{X}_i$, we can write $\mathbb{X}\mathbb{X}^T = \sum_{i=1}^d \mathbb{X}_i \otimes \mathbb{X}_i = \sum_{i=1}^d \mathbb{X}_i^{\otimes 2}$. Now for any $k \in \mathbb{N}$ one may consider the distribution $\mathcal{W}_{n,d}^{(k)}$ of $\frac{1}{\sqrt{d}} \sum_{i=1}^d \mathbb{X}_i^{\otimes k}$. How large does d need to be as a function of n and k so that $\mathcal{W}_{n,d}^{(k)}$ is approximately Gaussian?

4. Lecture 4: Confidence sets for the root in uniform and preferential attachment trees

In the previous lectures we studied random graph models with community structure and also models with an underlying geometry. While these models are important and lead to fascinating problems, they are also static in time. Many real-world networks are constantly evolving, and their understanding requires models that reflect this. This point of view brings about a host of new interesting and challenging statistical inference questions that concern the temporal dynamics of these networks.

In the last lecture we will study such questions: given the current state of a network, can one infer the state at some previous time? Does the initial *seed* graph have an influence on how the network looks at large times? If so, is it possible to find the origin of a large growing network? We will focus in particular on this latter question. More precisely, given a model of a randomly growing graph starting from a single node, called the *root*, we are interested in the following question. Given a large graph generated from the model, is it possible to find a small set of vertices for which we can guarantee that the root is in this set with high probability? Such *root-finding algorithms* can have applications to finding the origin of an epidemic or a rumor.

4.1. Models of growing graphs

A natural general model of randomly growing graphs can be defined as follows. For $n \geq k \geq 1$ and a graph S on k vertices, define the random graph $G(n, S)$

by induction. First, set $G(k, S) = S$; we call S the *seed* of the graph evolution process. Then, given $G(n, S)$, $G(n + 1, S)$ is formed from $G(n, S)$ by adding a new vertex and some new edges according to some adaptive rule. If S is a single vertex, we write simply $G(n)$ instead of $G(n, S)$. There are several rules one can consider; here we study perhaps the two most natural rules: uniform attachment and preferential attachment. Moreover, for simplicity we focus on the case of growing *trees*, where at every time step a single edge is added.

Uniform attachment trees are perhaps the simplest model of randomly growing graphs and are defined as follows. For $n \geq k \geq 1$ and a tree S on k vertices, the random tree $UA(n, S)$ is defined as follows. First, let $UA(k, S) = S$. Then, given $UA(n, S)$, $UA(n + 1, S)$ is formed from $UA(n, S)$ by adding a new vertex u and adding a new edge uv where the vertex v is chosen uniformly at random among vertices of $UA(n, S)$, independently of all past choices.
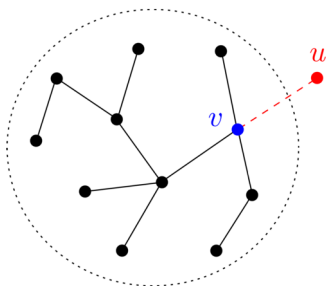

FIG 7. *Growing trees: add a new vertex u and attach it to an existing vertex v according some adaptive probabilistic rule.*

In preferential attachment the vertex is chosen with probability proportional to its degree [46, 8, 13]. For a tree T denote by $d_T(u)$ the degree of vertex u in T. For $n \geq k \geq 2$ and a tree S on k vertices we define the random tree $PA(n, S)$ by induction. First, let $PA(k, S) = S$. Then, given $PA(n, S)$, $PA(n+1, S)$ is formed from $PA(n, S)$ by adding a new vertex u and a new edge uv where v is selected at random among vertices in $PA(n, S)$ according to the following probability distribution:

$$\mathbb{P}\left(v = i \,|\, PA(n, S)\right) = \frac{d_{PA(n,S)}(i)}{2\left(n - 1\right)}.$$

4.2. Questions: detection and estimation

The most basic questions to consider are those of *detection* and *estimation*. Can one detect the influence of the initial seed graph? If so, is it possible to estimate the seed? Can one find the root if the process was started from a single node? We introduce these questions in the general model of randomly growing graphs described above, even though we study them in the special cases of uniform and preferential attachment trees later.

The detection question can be rephrased in the terminology of hypothesis testing. Given two potential seed graphs S and T, and an observation R which is a graph on n vertices, one wishes to test whether $R \sim G(n, S)$ or $R \sim G(n, T)$. The question then boils down to whether one can design a test with asymptotically (in n) nonnegligible power. This is equivalent to studying the total variation distance between $G(n, S)$ and $G(n, T)$, so we naturally define

$$\delta(S, T) := \lim_{n \to \infty} \mathrm{TV}(G(n, S), G(n, T)),$$

where $G(n, S)$ and $G(n, T)$ are random elements in the finite space of unlabeled graphs with n vertices. This limit is well-defined because $\mathrm{TV}(G(n, S), G(n, T))$ is nonincreasing in n (since if $G(n, S) = G(n, T)$, then the evolution of the random graphs can be coupled such that $G(n', S) = G(n', T)$ for all $n' \geq n$) and always nonnegative.

If the seed has an influence, it is natural to ask whether one can estimate S from $G(n, S)$ for large n. If so, can the subgraph corresponding to the seed be located in $G(n, S)$? We study this latter question in the simple case when the process starts from a single vertex called the *root*.[11] A *root-finding algorithm* is defined as follows. Given $G(n)$ and a target accuracy $\varepsilon \in (0, 1)$, a root-finding algorithm outputs a set $H(G(n), \varepsilon)$ of $K(\varepsilon)$ vertices such that the root is in $H(G(n), \varepsilon)$ with probability at least $1 - \varepsilon$ (with respect to the random generation of $G(n)$).

An important aspect of this definition is that the size of the output set is allowed to depend on ε, but not on the size n of the input graph. Therefore it is not clear that root-finding algorithms exist at all. Indeed, there are examples when they do not exist: consider a path that grows by picking one of its two ends at random and extending it by a single edge. However, it turns out that in many interesting cases root-finding algorithms do exist. In such cases it is natural to ask for the best possible value of $K(\varepsilon)$.

4.3. The influence of the seed

Consider distinguishing between a preferential attachment tree started from a star with 10 vertices, S_{10}, and a preferential attachment tree started from a path with 10 vertices, P_{10}. Since the preferential attachment mechanism incorporates the rich-get-richer phenomenon, one expects the degree of the center of the star in $\mathrm{PA}(n, S_{10})$ to be significantly larger than the degree of any of the initial vertices in the path in $\mathrm{PA}(n, P_{10})$. This intuition guided Bubeck, Mossel, and Rácz [22] when they initiated the theoretical study of the influence of the seed in preferential attachment trees. They showed that this intuition is correct: the limiting distribution of the maximum degree of the preferential attachment tree indeed depends on the seed. Using this they were able to show that for any two

---

[11]In the case of preferential attachment, starting from a single vertex is not well-defined; in this case we start the process from a single edge and the goal is to find one of its endpoints.

seeds S and T with at least 3 vertices[12] and different degree profiles we have $\delta_{\mathrm{PA}}(S,T) > 0$.

However, statistics based solely on degrees cannot distinguish all pairs of nonisomorphic seeds. This is because if S and T have the same degree profiles, then it is possible to couple $\mathrm{PA}(n,S)$ and $\mathrm{PA}(n,T)$ such that they have the same degree profiles for every n. In order to distinguish between such seeds, it is necessary to incorporate information about the graph structure into the statistics that are studied. This was done successfully by Curien, Duquesne, Kortchemski, and Manolescu [25], who analyzed statistics that measure the *geometry* of large degree nodes. These results can be summarized in the following theorem.

Theorem 4.1. *The seed has an influence in preferential attachment trees in the following sense. For any trees S and T that are nonisomorphic and have at least 3 vertices, we have $\delta_{\mathrm{PA}}(S,T) > 0$.*

In the case of uniform attachment, degrees do not play a special role, so initially one might even think that the seed has no influence in the limit. However, it turns out that the right perspective is not to look at degrees but rather the sizes of appropriate subtrees (we shall discuss such statistics later). By extending the approach of Curien et al. [25] to deal with such statistics, Bubeck, Eldan, Mossel, and Rácz [20] showed that the seed has an influence in uniform attachment trees as well.

Theorem 4.2. *The seed has an influence in uniform attachment trees in the following sense. For any trees S and T that are nonisomorphic and have at least 3 vertices, we have $\delta_{\mathrm{UA}}(S,T) > 0$.*

These results, together with a lack of examples showing opposite behavior, suggest that for most models of randomly growing graphs the seed has influence.

Question 4.3. *How common is the phenomenon observed in Theorems 4.1 and 4.2? Is there a natural large class of randomly growing graphs for which the seed has an influence? That is, models where for any two seeds S and T (perhaps satisfying an extra condition), we have $\delta(S,T) > 0$. Is there a natural model where the seed has no influence?*

The extra condition mentioned in the question could be model-dependent, but should not be too restrictive. It would be fascinating to find a natural model where the seed has no influence in a strong sense. Even for models where the seed does have an influence, proving the statement in full generality is challenging and interesting.

---

[12]This condition is necessary for a simple reason: the unique tree on 2 vertices, S_2, is always followed by the unique tree on 3 vertices, S_3, and hence $\delta(S_2, S_3) = 0$ for any model of randomly growing trees.

4.4. Finding Adam

These theorems about the influence of the seed open up the problem of *finding* the seed. Here we present the results of Bubeck, Devroye, and Lugosi [18] who first studied root-finding algorithms in the case of uniform attachment and preferential attachment trees.

They showed that root-finding algorithms indeed exist for preferential attachment trees and that the size of the best confidence set is polynomial in $1/\varepsilon$.

Theorem 4.4. *There exists a polynomial time root-finding algorithm for preferential attachment trees with $K(\varepsilon) \leq c\frac{\log^2(1/\varepsilon)}{\varepsilon^4}$ for some finite constant c. Furthermore, there exists a positive constant c' such that any root-finding algorithm for preferential attachment trees must satisfy $K(\varepsilon) \geq \frac{c'}{\varepsilon}$.*

They also showed the existence of root-finding algorithms for uniform attachment trees. In this model, however, there are confidence sets whose size is *subpolynomial* in $1/\varepsilon$. Moreover, the size of any confidence set has to be at least *superpolylogarithmic* in $1/\varepsilon$.

Theorem 4.5. *There exists a polynomial time root-finding algorithm for uniform attachment trees with $K(\varepsilon) \leq \exp\left(c\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\right)$ for some finite constant c. Furthermore, there exists a positive constant c' such that any root-finding algorithm for uniform attachment trees must satisfy $K(\varepsilon) \geq \exp\left(c'\sqrt{\log(1/\varepsilon)}\right)$.*

These theorems show an interesting quantitative difference between the two models: finding the root is exponentially more difficult in preferential attachment than in uniform attachment. While this might seem counter-intuitive at first, the reason behind this can be traced back to the rich-get-richer phenomenon: the effect of a rare event where not many vertices attach to the root gets amplified by preferential attachment, making it harder to find the root.

In the remaining part of this lecture we explain the basic ideas that go into proving Theorems 4.4 and 4.5 and prove some simpler special cases. Before we do so, we give a primer on Pólya urns, whose variants appear throughout the proofs. If the reader is familiar with Pólya urns, then the following subsection can be safely skipped.

4.5. Pólya urns: the building blocks of growing graph models

While uniform attachment and preferential attachment are arguably the most basic models of randomly growing graphs, the evolution of various simple statistics, such as degrees or subtree sizes, can be described using even simpler building blocks: Pólya urns. This subsection aims to give a brief introduction into the well-studied world of Pólya urns, while simultaneously showing examples of how these urn models show up in uniform attachment and preferential attachment.

4.5.1. The classical Pólya urn

The classical Pólya urn [28] starts with an urn filled with b blue balls and r red balls. Then at every time step you put your hand in the urn, without looking at its contents, and take out a ball sampled uniformly at random. You observe the color of the ball, put it back into the urn, together with another ball of the same color. This process is illustrated in Figure 8.
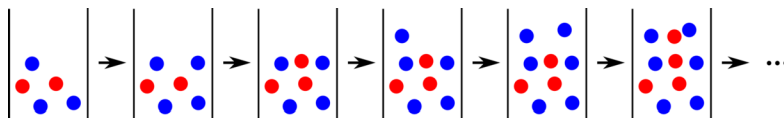

FIG 8. *A realization of Pólya's urn with $b = 3$ blue balls and $r = 2$ red balls initially.*

We are interested in the fraction of blue and red balls in the urn at large times. Let X_n denote the number of blue balls in the urn when there are n balls in the urn in total; initially we have $X_{b+r} = b$. Furthermore, let $x_n = X_n/n$ denote the fraction of blue balls when there are n balls in total.

Let us start by computing the expected increase in the number of blue balls at each time step:

$$\mathbb{E}\left[X_{n+1} \mid X_n\right] = (X_n + 1) \times \frac{X_n}{n} + X_n \times \left(1 - \frac{X_n}{n}\right) = \left(1 + \frac{1}{n}\right) X_n.$$

Dividing this by $(n + 1)$ we obtain that

$$\mathbb{E}\left[x_{n+1} \mid \mathcal{F}_n\right] = x_n,$$

where $\mathcal{F}_n$ denotes the filtration of the process up until time n (when there are n balls in the urn); since X_n is a Markov process, this is equivalent to conditioning on X_n. Thus the fraction of blue balls does not change in expectation; in other words, x_n is a martingale. Since x_n is also bounded ($x_n \in [0, 1]$), it follows that x_n converges almost surely to a limiting random variable. Readers not familiar with martingales should not be discouraged, as it is simple to see heuristically that x_n converges: when there are n balls in the urn, the change in x_n is on the order of $1/n$, which converges to zero fast enough that one expects x_n to converge.[13]

Our next goal is to understand the limiting distribution of x_n. First, let us compute the probability of observing the first five draws as in Figure 8, starting with a blue ball, then a red, then two blue ones, and lastly another red: this probability is $\frac{3}{5} \times \frac{2}{6} \times \frac{4}{7} \times \frac{5}{8} \times \frac{3}{9}$. Notice that the probability of obtaining 3 blue balls and 2 red ones in the first 5 draws is the same regardless of the order in which we draw the balls. This property of the sequence X_n is known as

---

[13]The reader can convince themselves that $\mathbb{E}\left[(x_{n+1} - x_n)^2 \mid \mathcal{F}_n\right] = \frac{x_n(1-x_n)}{(n+1)^2}$, and so the sum of the variances from time N onwards is bounded by $\sum_{n \geq N} (n+1)^{-2} \leq 1/N$.

exchangeability and has several useful consequences (most of which we will not explore here). It follows that the probability of seeing k blue balls in the first n draws takes on the following form:

$$\mathbb{P}\left(X_{n+b+r} = b+k\right)$$
$$= \binom{n}{k} \frac{b\left(b+1\right)\ldots\left(b+k-1\right) \times r\left(r+1\right)\ldots\left(r+n-k-1\right)}{\left(b+r\right)\left(b+r+1\right)\ldots\left(b+r+n-1\right)}$$

From this formula one can read off that $X_{n+b+r} - b$ is distributed according to the *beta-binomial distribution* with parameters (n, b, r). An alternative way of sampling from the beta-binomial distribution is to first sample a probability p from the beta distribution with parameters b and r (having density $x \mapsto \frac{\Gamma(b+r)}{\Gamma(b)\Gamma(r)} x^{b-1} \left(1-x\right)^{r-1} \mathbf{1}_{\{x \in [0,1]\}}$), and then conditionally on p, sample from the binomial distribution with n trials and success probability p. Conditionally on p, the strong law of large numbers applied to the binomial distribution thus tells us that $\left(X_{n+b+r} - b\right)/n$ converges almost surely to p. Since $x_n = \left(X_{n+b+r} - b\right)/n + o(1)$, it follows that $x_n \to p$ almost surely. We have thus derived the following theorem.

Theorem 4.6. *Let x_n denote the fraction of blue balls at time n (when there are n balls in total) in a classical Pólya urn which starts with b blue balls and r red balls. Then*

$$\lim_{n\to\infty} x_n = x$$

almost surely, where $x \sim \text{Beta}\left(b, r\right)$.

Example 4.7. *The classical Pólya urn shows up in uniform attachment trees as it describes the evolution of subtree sizes as follows. Pick an edge of a tree,*
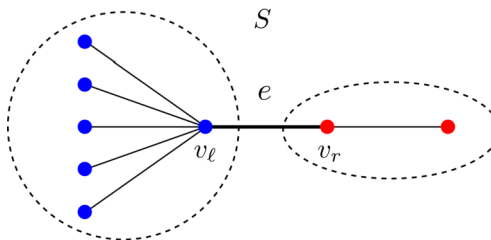

FIG 9. *The subtree sizes in uniform attachment trees evolve according to the classical Pólya urn.*

such as edge e in tree S in Figure 9, with endpoints v_ℓ and v_r. This edge partitions the tree into two parts on either side of the edge: a subtree under v_ℓ and a subtree under v_r. The sizes of these subtrees (i.e., the number of vertices they contain) evolve exactly like the classical Pólya urn described above (in the example depicted in Figure 9 we have $b = 6$ and $r = 2$ initially).

4.5.2. Multiple colors

A natural generalization is to consider multiple colors instead of just two. Let m be the number of colors, let $\underline{X}_n = (X_{n,1}, \ldots, X_{n,m})$ denote the number of balls of each color when there are n balls in the urn in total, and let $\underline{x}_n = \underline{X}_n/n$. Assume that initially there are r_i balls of color i.

In this case the fraction of balls of each color converges to the natural multivariate generalization of the beta distribution: the Dirichlet distribution. The Dirichlet distribution with parameters $(r_1, \ldots, r_m)$, denoted $\mathrm{Dir}\,(r_1, \ldots, r_m)$, has density

$$\underline{x} = (x_1, \ldots, x_m) \mapsto \frac{\Gamma\left(\sum_{i=1}^m r_i\right)}{\prod_{i=1}^m \Gamma(r_i)} x_1^{r_1-1} \ldots x_m^{r_m-1} \mathbf{1}_{\left\{\forall i \,:\, x_i \in [0,1], \sum_{i=1}^m x_i = 1\right\}}.$$

It has several natural properties that one might expect, for instance the aggregation property, that if one groups coordinates i and j together, then the resulting distribution is still Dirichlet, with parameters r_i and r_j replaced by $r_i + r_j$. This also implies that the univariate marginals are beta distributions.

The convergence result for multiple colors follows similarly to the one for two colors, so we simply state the result.

Theorem 4.8. *Let $\underline{x}_n$ denote the fraction of balls of each color at time n (when there are n balls in total) in a classical Pólya urn of m colors which starts with r_i balls of color i. Then*

$$\lim_{n \to \infty} \underline{x}_n = \underline{x}$$

almost surely, where $\underline{x} \sim \mathrm{Dir}\,(r_1, \ldots, r_m)$.

Example 4.9. *A Pólya urn with multiple colors shows up in uniform attachment trees when we partition the tree into multiple subtrees. Picking a subtree*
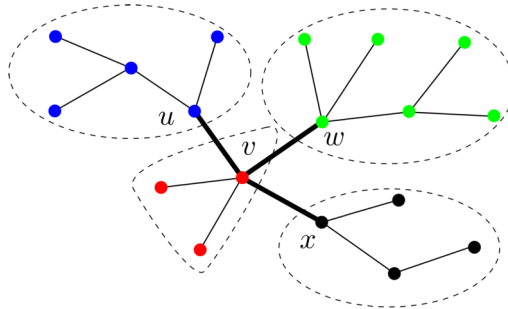

FIG 10. *The sizes of multiple subtrees in uniform attachment trees evolve according to a Pólya urn with multiple colors.*

of m vertices as highlighted in bold in Figure 10, the tree is partitioned into m subtrees. The sizes of these subtrees (i.e., the number of vertices they contain) evolve exactly like the classical Pólya urn with m colors described above.

4.5.3. Adding multiple balls at a time

It is also natural to consider adding more than one extra ball at each time step. The effect of this is to change the parameter of the limiting Dirichlet distribution.

Theorem 4.10. *Let $\underline{x}_n$ denote the fraction of balls of each color at time n (when there are n balls in total) in a Pólya urn of m colors which starts with r_i balls of color i and where k balls of the same color are added at each time step. Then*

$$\lim_{n \to \infty} \underline{x}_n = \underline{x}$$

almost surely, where $\underline{x} \sim \mathrm{Dir}\,(r_1/k, \ldots, r_m/k)$.

Example 4.11. *Pólya urns where two balls of the same color are added at each time step appear in preferential attachment trees as follows. Consider partitioning the tree into m subtrees as in Figure 10, but now define the size of a subtree to be the sum of the degrees of the vertices in it. Consider which subtree the new incoming vertex attaches to. In the preferential attachment process each subtree is picked with probability proportional to its size and whichever subtree is picked, the sum of the degrees (i.e., the size) increases by 2 due to the new edge. Thus the subtree sizes evolve exactly according to a Pólya urn described above with $k = 2$.*

4.5.4. More general urn models

More generally, one can add some number of balls of each color at each time step. The replacement rule is often described by a *replacement matrix* of size $m \times m$, where the ith row of the matrix describes how many balls of each color to add to the urn if a ball of color i is drawn. The urn models studied above correspond to replacement matrices that are a constant multiple of the identity. The literature on general replacement matrices is vast and we do not intend to discuss it here; our goal is just to describe the simple case when the replacement matrix is $\left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. We refer to [38] for detailed results on triangular replacement matrices, and to the references therein for more general replacement rules.

The urn model with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ can also be described as the classical Pólya urn with two colors as described in Section 4.5.1, but in addition a blue ball is always added at each time step. It is thus natural to expect that there will be many more blue balls than red balls in the urn at large times. It turns out that the number of red balls at time n scales as $\sqrt{n}$ instead of linearly in n. The following result is a special case of what is proved in [38].

Theorem 4.12. *Let (X_n, Y_n) denote the number of blue and red balls, respectively, at time n (when there are n balls in total) in a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. Assume that initially there are some red balls in the urn. Then $Y_n/\sqrt{n}$ converges in distribution to a nondegenerate random variable.*

Example 4.13. *The evolution of the degree of any given vertex in a preferential attachment tree can be understood through such a Pólya urn. More precisely, fix a vertex v in the tree, let Y_n denote the degree of v when there are n vertices in total, and let X_n denote the sum of the degrees of all other vertices. Then (X_n, Y_n) evolves exactly according to a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. This implies that the degree of any fixed vertex scales as $\sqrt{n}$ in the preferential attachment tree.*

4.6. Proofs using Pólya urns

With the background on Pólya urns covered, we are now ready to understand some of the proofs of the results concerning root-finding algorithms from [18].

4.6.1. A root-finding algorithm based on the centroid

We start by presenting a simple root-finding algorithm for uniform attachment trees. This algorithm is not optimal, but its analysis is simple and highlights the basic ideas.

For a tree T, if we remove a vertex $v \in V(T)$, then the tree becomes a forest consisting of disjoint subtrees of the original tree. Let $\psi_T(v)$ denote the size (i.e., the number of vertices) of the largest component of this forest. For example, in Figure 9 if we remove v_r from S, then the tree breaks into a singleton and a star consisting of 6 vertices; thus $\psi_S(v_r) = 6$. A vertex v that minimizes $\psi_T(v)$ is known as a *centroid* of T; one can show that there can be at most two centroids. We define the confidence set H_ψ by taking the set of K vertices with smallest ψ values.

Theorem 4.14. *[18] The centroid-based H_ψ defined above is a root-finding algorithm for the uniform attachment tree. More precisely, if $K \geq \frac{5}{2} \frac{\log(1/\varepsilon)}{\varepsilon}$, then*

$$\liminf_{n \to \infty} \mathbb{P}\left(1 \in H_\psi\left(\mathrm{UA}\,(n)^\circ\right)\right) \geq 1 - \frac{4\varepsilon}{1 - \varepsilon},$$

where 1 denotes the root, and $\mathrm{UA}\,(n)^\circ$ denotes the unlabeled version of $\mathrm{UA}\,(n)$.

Proof. We label the vertices of the uniform attachment tree in chronological order. We start by introducing some notation that is useful throughout the proof. For $0 \leq i \leq k$, denote by $T_{i,k}$ the tree containing vertex i in the forest obtained by removing in $\mathrm{UA}\,(n)$ all edges between vertices $\{1, \ldots, k\}$. Also, let $|T|$ denote the size of a tree T, i.e., the number of vertices it contains. Note that the vector $(|T_{1,k}|, \ldots, |T_{k,k}|)$ evolves according to the classical Pólya urn with k colors as described in Section 4.5.2, with initial state $(1, \ldots, 1)$. Therefore, by Theorem 4.8, the normalized vector $(|T_{1,k}|, \ldots, |T_{k,k}|)/n$ converges in distribution to a Dirichlet distribution with parameters $(1, \ldots, 1)$.

Now observe that

$$\mathbb{P}\left(1 \notin H_\psi\right) \leq \mathbb{P}\left(\exists i > K : \psi(i) \leq \psi(1)\right)$$
$$\leq \mathbb{P}\left(\psi(1) \geq (1 - \varepsilon)\,n\right) + \mathbb{P}\left(\exists i > K : \psi(i) \leq (1 - \varepsilon)\,n\right).$$

We bound the two terms appearing above separately, starting with the first one. Note that $\psi(1) \leq \max\{|T_{1,2}|, |T_{2,2}|\}$, and both $|T_{1,2}|/n$ and $|T_{2,2}|/n$ converge in distribution to a uniform random variable in $[0, 1]$. Hence a union bound gives us that

$$\limsup_{n \to \infty} \mathbb{P}\left(\psi(1) \geq (1 - \varepsilon) n\right) \leq 2 \lim_{n \to \infty} \mathbb{P}\left(|T_{1,2}| \geq (1 - \varepsilon) n\right) = 2\varepsilon.$$

For the other term, first observe that for any $i > K$ we have

$$\psi(i) \geq \min_{1 \leq k \leq K} \sum_{j=1, j \neq k}^{K} |T_{j,K}|.$$

Now using the results on Pólya urns from Section 4.5 we have that for every k such that $1 \leq k \leq K$, the random variable $\frac{1}{n} \sum_{j=1, j \neq k}^{K} |T_{j,K}|$ converges in distribution to the $\mathrm{Beta}(K - 1, 1)$ distribution. Hence by a union bound we have that

$$\limsup_{n \to \infty} \mathbb{P}\left(\exists i > K : \psi(i) \leq (1 - \varepsilon) n\right)$$

$$\leq \lim_{n \to \infty} \mathbb{P}\left(\exists 1 \leq k \leq K : \sum_{j=1, j \neq k}^{K} |T_{j,K}| \leq (1 - \varepsilon) n\right)$$

$$\leq K(1 - \varepsilon)^{K-1}.$$

Putting together the two bounds gives that

$$\limsup_{n \to \infty} \mathbb{P}\left(1 \notin H_\psi\right) \leq 2\varepsilon + K(1 - \varepsilon)^{K-1},$$

which concludes the proof due to the assumption on K. $\qquad\square$

The same estimator H_ψ works for the preferential attachment tree as well, if one takes $K \geq C \frac{\log^2(1/\varepsilon)}{\varepsilon^4}$ for some positive constant C. The proof mirrors the one above, but involves a few additional steps; we refer to [18] for details.

For uniform attachment the bound on K given by Theorem 4.14 is not optimal. It turns out that it is possible to write down the maximum likelihood estimator (MLE) for the root in the uniform attachment model; we do not do so here, see [18]. One can view the estimator H_ψ based on the centroid as a certain "relaxation" of the MLE. By constructing a certain "tighter" relaxation of the MLE, one can obtain a confidence set with size subpolynomial in $1/\varepsilon$ as described in Theorem 4.5. The analysis of this is the most technical part of [18] and we refer to [18] for more details.

4.6.2. Lower bounds

As mentioned above, the MLE for the root can be written down explicitly. This aids in showing a lower bound on the size of a confidence set. In particular,

Bubeck et al. [18] define a set of trees whose probability of occurrence under the uniform attachment model is not too small, yet the MLE provably fails, giving the lower bound described in Theorem 4.5. We refer to [18] for details.

On the other hand, for the preferential attachment model it is not necessary to use the structure of the MLE to obtain a lower bound. A simple symmetry argument suffices to show the lower bound in Theorem 4.4, which we now sketch.

First observe that the probability of error for the optimal procedure is non-decreasing with n, since otherwise one could simulate the process to obtain a better estimate. Thus it suffices to show that the optimal procedure must have a probability of error of at least ε for some finite n. We show that there is some finite n such that with probability at least 2ε, the root is isomorphic to at least $2c/\varepsilon$ vertices in PA(n). Thus if a procedure outputs at most c/ε vertices, then it must make an error at least half the time (so with probability at least ε).

Observe that the probability that the root is a leaf in PA(n) is $\frac{1}{2} \times \frac{3}{4} \times \cdots \times \left(1 - \frac{1}{2n}\right) = \Theta\left(1/\sqrt{n}\right)$. By choosing $n = \Theta\left(1/\varepsilon^2\right)$, this happens with probability $\Theta\left(\varepsilon\right)$. Furthermore, conditioned on the root being a leaf, with constant probability vertex 2 is connected to $\Theta\left(\sqrt{n}\right) = \Theta\left(1/\varepsilon\right)$ leaves (here we use Theorem 4.12), which are then isomorphic to the root.

4.7. Outlook: open problems and extensions

There are many open problems and further directions that one can pursue; the four main papers we have discussed [22, 25, 20, 18] contain 20 open problems and conjectures alone. For instance, can the bounds on the size of the optimal confidence set be improved and ultimately tightened? What about other tree growth models? What happens when we lose the tree structure and consider general graphs, e.g., by adding multiple edges at each time step?

When the tree growth model is not as combinatorial as uniform attachment or preferential attachment, then other techniques might be useful. In particular, many tree growth models can be embedded into continuous time branching processes and then the full machinery of general branching processes can be brought to the fore and applied; see [53, 9] and the references therein for such results. This approach can also be used to obtain finite confidence sets for the root, as demonstrated recently in [40] for sublinear preferential attachment trees.

A closely related problem to those discussed in this lecture is that of detecting the source of a diffusion spreading on an underlying network. The results are very similar to those above: the rumor source can be efficiently detected in many settings, see, e.g., [54, 55, 44]. A different twist on this question is motivated by anonymous messaging services: can one *design protocols* for spreading information that preserve anonymity by minimizing the probability of source detection? Fanti et al. [33] introduced a process, termed adaptive diffusion, that indeed achieves this goal. Understanding the tradeoffs between privacy and other desiderata is timely and should lead to lots of interesting research.

References

[1] E. Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, to appear, 2017.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016. MR3447993

[3] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2015. MR3473334

[4] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. Preprint available at `http://arxiv.org/abs/1512.09080`, 2015.

[5] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004. MR2083473

[6] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. On the rate of convergence in the entropic central limit theorem. *Probability Theory and Related Fields*, 129(3):381–390, 2004. MR2128238

[7] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41–63, 2003. MR1991646

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. MR2091634

[9] S. Bhamidi. Universal techniques to analyze preferential attachment trees: Global and Local analysis. Available online at `http://www.unc.edu/~bhamidi/preferent.pdf`, 2007.

[10] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[11] N. Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information Theory*, 11(2):267–271, 1965. MR0188004

[12] S. G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999. MR1742893

[13] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The Degree Sequence of a Scale-Free Random Graph Process. *Random Structures & Algorithms*, 18(3):279–290, 2001. MR1824277

[14] H. Breu and D. G. Kirkpatrick. Unit disk graph recognition is NP-hard. *Computational Geometry*, 9(1):3–24, 1998. MR1492799

[15] N. Broutin, L. Devroye, N. Fraiman, and G. Lugosi. Connectivity threshold of Bluetooth graphs. *Random Structures & Algorithms*, 44(1):45–66, 2014. MR3143590

[16] N. Broutin, L. Devroye, and G. Lugosi. Connectivity of sparse Bluetooth

networks. *Electronic Communications in Probability*, 20(48):1–10, 2015. MR3358970

[17] N. Broutin, L. Devroye, and G. Lugosi. Almost optimal sparsification of random geometric graphs. *The Annals of Applied Probability*, 26(5):3078–3109, 2016. MR3563202

[18] S. Bubeck, L. Devroye, and G. Lugosi. Finding Adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017. MR3607120

[19] 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. MR3545825

[20] S. Bubeck, R. Eldan, E. Mossel, and M. Z. Rácz. From trees to seeds: on the inference of the seed from large trees in the uniform attachment model. *Bernoulli*, 23(4A):2887–2916, 2017. MR3648049

[21] S. Bubeck and S. Ganguly. Entropic CLT and Phase Transition in High-dimensional Wishart Matrices. *International Mathematics Research Notices*, to appear, 2016.

[22] S. Bubeck, E. Mossel, and M. Z. Rácz. On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, 2(1):30–39, 2015. MR3361606

[23] T. N. Bui, S. Chaudhuri, F. T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, 1987. MR0905164

[24] A. Condon and R. M. Karp. Algorithms for Graph Partitioning on the Planted Partition Model. *Random Structures and Algorithms*, 18(2):116–140, 2001. MR1809718

[25] N. Curien, T. Duquesne, I. Kortchemski, and I. Manolescu. Scaling limits and influence of the seed graph in preferential attachment trees. *Journal de l'École polytechnique — Mathématiques*, 2:1–34, 2015. MR3326003

[26] 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.

[27] 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. MR2861682

[28] F. Eggenberger and G. Pólya. Über die Statistik verketteter Vorgänge. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 3(4):279–289, 1923.

[29] R. Eldan. An efficiency upper bound for inverse covariance estimation. *Israel Journal of Mathematics*, 207(1):1–9, 2015. MR3358038

[30] R. Eldan and D. Mikulincer. Information and dimensionality of anisotropic random geometric graphs. Preprint available at https://arxiv.org/abs/1609.02490, 2016.

[31] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959. MR0120167

[32] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, V.A(1–2):17–60, 1960. MR0125031

[33] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath. Spy vs. Spy: Rumor Source Obfuscation. In *ACM SIGMETRICS*, volume 43, pages 271–284. ACM, 2015.

[34] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[35] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. MR1908073

[36] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. MR1951262

[37] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. MR0718088

[38] S. Janson. Limit theorems for triangular urn schemes. *Probability Theory and Related Fields*, 134(3):417–452, 2006. MR2226887

[39] T. Jiang and D. Li. Approximation of Rectangular Beta-Laguerre Ensembles and Large Deviations. *Journal of Theoretical Probability*, 28:804–847, 2015. MR3413957

[40] V. Jog and P.-L. Loh. Analysis of Centrality in Sublinear Preferential Attachment Trees via the Crump-Mode-Jagers Branching Process. *IEEE Transactions on Network Science and Engineering*, 4(1):1–12, 2017. MR3625951

[41] O. Johnson and A. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129(3):391–409, 2004. MR2128239

[42] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001. MR1863961

[43] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011. MR2788206

[44] J. Khim and P.-L. Loh. Confidence Sets for the Source of a Diffusion in Regular Trees. *IEEE Transactions on Network Science and Engineering*, 4(1):27–40, 2017. MR3625953

[45] Y. V. Linnik. An Information-Theoretic Proof of the Central Limit Theorem with Lindeberg Conditions. *Theory of Probability & Its Applications*, 4(3):288–299, 1959. MR0124081

[46] H. M. Mahmoud. Distances in random plane-oriented recursive trees. *Journal of Computational and Applied Mathematics*, 41(1-2):237–245, 1992. MR1181723

[47] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 694–703. ACM, 2014. MR3238997

[48] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. Preprint available at http://arxiv.org/abs/1311.4115, 2013.

[49] E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction, and optimal recovery of block models. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, 2014. MR3543895

[50] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing (STOC)*, pages 69–75. ACM, 2015. MR3388184

[51] M. Penrose. *Random Geometric Graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, 2003. MR1986198

[52] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011. MR2893856

[53] A. Rudas, B. Tóth, and B. Valkó. Random trees and general branching processes. *Random Structures & Algorithms*, 31(2):186–202, 2007. MR2343718

[54] D. Shah and T. Zaman. Rumors in a Network: Who's the Culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011. MR2849111

[55] D. Shah and T. Zaman. Finding Rumor Sources on Random Trees. *Operations Research*, 64(3):736–755, 2016. MR3515208

[56] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948. MR0026286

[57] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959. MR0109101