Average-case complexity and statistical inference

We give an example of the sort of question we will look at. The idea is to 'hide' some structure within a high-dimensional random object. We may then ask if we can *detect* this, i.e. whether we can distinguish the vanilla random object from the random object plus planted structure, also if we may *recover*, i.e. 'find' the planted structure from within the random object.

A favourite combinatorial random object is the Erdős-Rènyi random graph, $G_{n,1/2}$, take n vertices and between each pair of vertices independently place an edge with probability $1/2$. We are interested in the typical behaviour for large n. The structure we plant is a clique, i.e. between a subset S^* of the vertices of the graph, we place all the possible edges. Now, the random graph $G_{n,1/2}$ without the planted structure will naturally have some cliques by chance, and indeed the largest of these will have size approximately $2\log_2 n$ with probability tending to 1 as $n \to \infty$; which suggests it might not be possible to detect or recover a planted structure of size smaller than $2\log_2 n$. This turns out to be true, as we shall see. Interestingly, there is another phase transition. Fast algorithms finding the clique, e.g. picking the vertices of highest degrees, are only known when the planted clique has size about $n^{1/2}$ or higher; which is considerably larger than $2\log_2 n$. There is some 'evidence' that this $n^{1/2}$ threshold is fundamental: by evidence we mean rigorous statements we can prove which *suggest* that there are no polynomial time algorithms.

These ideas will be made precise in the course as we investigate what forms this evidence can take. We illustrate the techniques on three running examples, planted clique, as described above, as well a generalisation of it planted dense subgraph, and a Gaussian planted structure problem biclustering - see Appendix A for a list and the phase transitions in each model. This area is very active and many of the techniques and results presented here have been developed within the last decade and some within the last year.

1 Detection

1.1 Definitions

Problem Setup We specify a dimension n, and parameters (e.g. k size of planted structure, λ strength of 'signal', p, q probabilities of 'community' edges and 'non-community' edges respectively). For each fixed set of parameters we are interested in the behaviour for large n or as $n \to \infty$.

For a detection problem, under H_0 the *null hypothesis*, we sample from the probability space $\mathcal{Q}_n$ and under H_1 the *alternate hypothesis* we sample from the probability space $\mathcal{P}_n$. We write $\mathbb{P}_0(G = g)$ to denote the probability that a random sample G from probability distribution $\mathcal{Q}_n$ is the deterministic g (and similarly $\mathbb{P}_1(G = g)$ to denote the same for $\mathcal{P}_n$). We will try to stick to the convention of denoting random variables, random graphs or random matrices by capital letters and deterministic values, graphs and matrices by lower case letters.

A *test* is a function ϕ_n on the union of the supports of $\mathcal{Q}_n$ and $\mathcal{P}_n$, with $\phi_n(g) \in \{0,1\}$[1]. We need a notion of how 'good' a test is at distinguishing $\mathcal{Q}_n$ and $\mathcal{P}_n$ and will use risk. The *risk* of a test ϕ, denoted $r(\phi)$ is

$$r(\phi) = \sum_{g:\,\phi(g)=1} \mathbb{P}_0(\phi(G) = g) + \sum_{g:\,\phi(g)=0} \mathbb{P}_1(\phi(G) = g) = \mathbb{P}_0(\phi(G) = 1) + \mathbb{P}_1(\phi(G) = 0)$$

[1]Suppose for now that ϕ_n is deterministic, later we may have random tests.

Observe that it is easy to design a function which achieves risk 1. We can take $\phi_{\text{guess null}}(g) = 0$ for all g in the support of $\mathcal{P}_n$ and $\mathcal{Q}_n$, or we could take the random test $\phi_p(g)$ which takes value 1 with probability p and 0 otherwise. Both of these have risk 1 for each input g.

We say a test ϕ_n achieves *strong detection* between H_0 and H_1 if $r(\phi_n) \to 0$ as $n \to \infty$. Similarly, we say a test ϕ_n achieves *weak detection* between H_0 and H_1 if there exists $\varepsilon > 0, n_0$ such that $r(\phi_n) < 1 - \varepsilon$ for all $n > n_0$.

We may now define what we mean by EASY and POSSIBLE detection. Say for $H_0 : \mathcal{Q}_n(\alpha, \beta)$ vs $H_1 : \mathcal{P}_n(\alpha, \beta)$ that *strong detection for H_0 vs H_1 is* EASY *for parameters* α, β if there exists a test ϕ_n implementable as a polynomial time algorithm such that $r(\phi_n) \to 0$ as $n \to \infty$. The definition for weak detection being EASY is similar, just replace the condition on $r(\phi_n)$ as required.

Say for $H_0 : \mathcal{Q}_n(\alpha, \beta)$ vs $H_1 : \mathcal{P}_n(\alpha, \beta)$ that *strong detection for H_0 vs H_1 is* POSSIBLE *for parameters* α, β if there exists a test ϕ_n such that $r(\phi_n) \to 0$ as $n \to \infty$. In particular ϕ_n may be a brute-force algorithm. The definition for weak detection being POSSIBLE is similar.

Later we will be able to talk about detection problems being HARD if they are POSSIBLE and we have 'evidence of hardness'. We will usually specify which evidence of hardness.

1.2 The spiked matrix model

We define a probability space BC_n over $n \times n$ matrices with real entries. The detection problem will be to distinguish between a matrix sampled from the 'null' where all entries normally distributed with mean zero and variance one, from the 'alternate' with a planted submatrix of expected size k which has entries with mean λ and variance one. See Fig 3.

Given the number of vertices n, total community size k and signal strength $\lambda > 0$, define the additive (independent) Gaussian model $\mathrm{BC}_n = \mathrm{BC}_n(k, \lambda)$ as follows. Under BC_n, independently for each $i \in [n] := \{1, 2, \ldots, n\}$, the community label σ_i is sampled such that $\sigma_i = 1$ with probability k/n and $\sigma_i = 0$ with probability $1 - k/n$, then set $X_{ij} = \sigma_i \sigma_j$ For each pair of vertices $i, j \in [n]$ the matrix entry Y_{ij} is sampled from

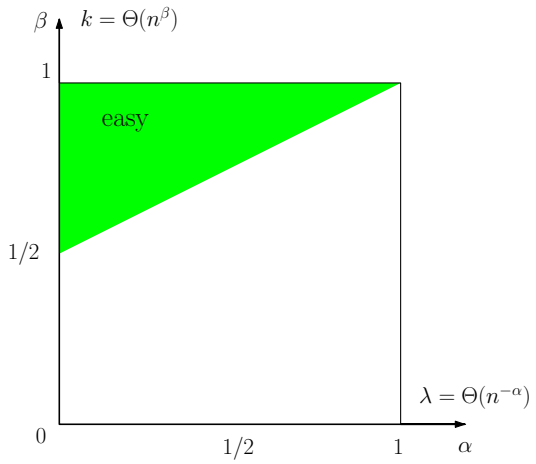
$$Y_{ij}|X_{ij} \sim \begin{cases} \mathcal{N}(\lambda, 1), & X_{ij} = 1 \quad (\text{i.e. } \sigma_i = \sigma_j = 1) \\ \mathcal{N}(0, 1), & X_{ij} = 0. \end{cases}$$

Note that if $ij \neq kl$ then $Y_{ij}|X_{ij}$ is independent from $Y_{kl}|X_{kl}$. We define the (fixed number) Gaussian model $\mathrm{BC}_n = \mathrm{BC'}_n(k, \lambda)$ as above except that we choose a set S uniformly from all subsets of $[n]$ of size k, and set $\sigma_i = 1$ if $i \in S$ and $\sigma_i = 0$ if $i \notin S$. Thus in this model we have exactly k 'community' indices.

1.3 When detection in matrix model is EASY

Lemma 1.1. *Fix α, β such that $0 < \alpha, \beta < 1$ and $\beta > \alpha/2 + 1/2$ (i.e. the green region in Fig 1). Let $\mathcal{Q}_n = \mathcal{Q}_n(\alpha, \beta)$ be $BC(n, 0, 0)$ and $\mathcal{P}_n = \mathcal{P}_n(\alpha, \beta)$ be $BC(n, k = n^\beta, \lambda = n^{-\alpha})$. Then the strong detection problem is* EASY *for α, β.*

end L1



(a) detection easy

Figure 1: Region considered in Lemma 1.1, see also Figure 3

Proof. (sketch)
We show the follow test distinguishes with high probability. Define $\phi_{\text{sum}} : \mathbb{R}^{n^2} \to \{0,1\}$ by

$$\phi_{\text{sum}}(x) = \begin{cases} 1, & \text{if } \sum_{i,j} x_{ij} > \frac{1}{2}\lambda k^2, \\ 0, & \text{otherwise.} \end{cases}$$

Recall that if $X_1 \sim N(\mu_1, s_1^2)$ and $X_2 \sim N(\mu_2, s_2^2)$ and X_1 and X_2 are independent then the sum is distributed $X_1 + X_2 \sim N(\mu_1 + \mu_2, s_1^2 + s_2^2)$. Note under H_0 we have n^2 variables distributed as $N(0,1)$, and under H_0 we have k^2 variables distributed as $N(\lambda, 1)$ and $n^2 - k^2$ distributed as $N(0,1)$, and thus

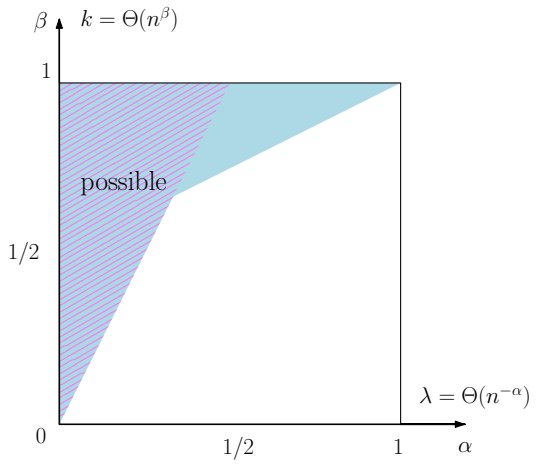
$$\sum_{i,j} X_{ij} = \begin{cases} N(0, n^2), & \text{under } H_0, \\ N(\lambda k^2, n^2), & \text{under } H_1. \end{cases}$$

The proof may now be completed using Lemma B.2[2].

□

Lemma 1.2. *Fix $\alpha, \beta \in (0,1)$ such that either of the following holds $\beta > \alpha/2 + 1/2$ or $\beta > 2\alpha$ (i.e. the shaded dashed and non-dashed regions in Fig 2). Let $\mathcal{Q}_n = \mathcal{Q}_n(\alpha, \beta)$ be $BC'(n, 0, 0)$ and $\mathcal{P}_n = \mathcal{P}_n(\alpha, \beta)$ be $BC'(n, k, \lambda)$. Then the strong detection problem is* POSSIBLE *for α, β.*

[2]Intuition: we succeed when the difference in means $\gg$ square-root of the variance. In this case the difference in means of the test statistic is $\lambda k^2 = n^{-\alpha+2\beta}$, and the variance is n^2 under both H_0 and H_1. So this intuition would say we succeed when $n^{-\alpha+2\beta} \gg n$, i.e. when $-\alpha + 2\beta > 1$, which is what we are aiming for.



(a) detection possible

Figure 2: Region considered in Lemma 1.2, since we already have a test for the green region in Figure **??**, we need only prove there is a (brute-force) test for the dashed region. See also Figure 3

Proof. Note that by Lemma 1.1 we have a test which distinguishes whp for $\alpha, \beta \in (0,1)$ with $\beta > \alpha/2 + 1/2$ (i.e. the green region in Fig 1). Thus it suffices to construct a test which distinguishes whp for fixed $\alpha, \beta \in (0,1)$ with $\beta > 2\alpha$ (i.e. the purple dashed area in Fig 2.

Let $\phi_{\text{search}} : \mathbb{R}^{n^2} \to \{0,1\}$ be defined by

$$\phi_{\text{search}}(x) = \begin{cases} 1, & \text{if } \max_{S\subset[n],\ |S|=k} \sum_{i,j\in S} x_{ij} > \frac{1}{2}\lambda k^2, \\ 0, & \text{otherwise.} \end{cases}$$

Define the random variable $T_{\text{search}} = \max_{S\subset[n],\ |S|=k} \sum_{i,j\in S} X_{ij}$, i.e. the quantity that is thresholded in ϕ_{search}.

Under H_0, T_{search} is the max of $\binom{n}{k}$ variables, $T_1, \dots, T_{\binom{n}{k}}$ where each $T_i \sim N(0, n^2)$.

Recall that if $X_1 \sim N(\mu_1, s_1^2)$ then for constant a, we have $aX_1 \sim N(a\mu_1, a^2 s_1^2)$. Thus $n^{-1}T_i \sim N(0,1)$, and so we can apply Lemma C.1. (Note the T_i's are not independent, they are sums of overlapping parts of the matrix, but we don't need independence to apply the lemma.) By Lemma C.1 whp

$$T_{\text{search}} = \max_{i=1,\dots,m} \frac{1}{n} T_i \leqslant \sqrt{(2+\varepsilon)\log\left(\binom{n}{k}\right)} \leqslant \sqrt{(2+\varepsilon)k\log n}$$

where we used the fact that $\binom{n}{k} \leqslant n^k$.

Under H_1, again T_{search} is the max of $\binom{n}{k}$ variables, including the sum over the planted submatrix, and since it is a max it is at least as big as the sum over the planted set S^*. Let $T_{\text{planted}} = \sum_{i,j\in S^*} X_{i,j}$. Since T_{planted} is the sum of k^2 $N(\lambda, 1)$ random variables, $T_{\text{planted}} \sim N(\lambda k^2, k^2)$ and $\mathbb{E}[T_{\text{planted}}] = \lambda k^2$.

We want to show $T_{\text{planted}} > \lambda k^2/2$ whp. Since T_{planted} has expected value λk^2 it is enough to show that whp T_{planted} is at most $\lambda k^2/3$ away from its expectation - this shows whp $T_{\text{planted}} > 2\lambda k^2/3$ - see also the chat in Section B. By Lemma B.2, with $t = \lambda k^2/3$

$$\mathbb{P}(|T_{\text{planted}} - \lambda k^2| \geqslant \lambda k^2/3) \leqslant 2\exp(-(\lambda k^2/3)^2/(2k^2)) = 2\exp(-\lambda^2 k^2/6).$$

Now, note that $2\exp(-\lambda^2 k^2/6) = o(1)$ if $\lambda k \to \infty$, i.e. $n^{\beta-\alpha} \to \infty$ which is true whenever $\beta > \alpha$. Hence for $\beta > \alpha$ $T_{\text{planted}} > \lambda k^2/2$ whp. Since $T_{\text{search}} \geqslant T_{\text{planted}}$ always. Thus under H_1 whp $\phi_{search} = 1$ as required.

2 Planted Clique

For further details see [1] which we follow in this section.

□

A List of Planted problems

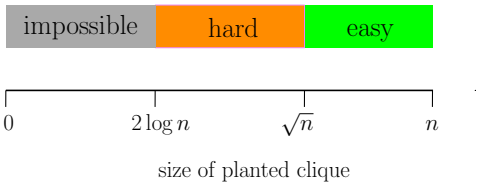
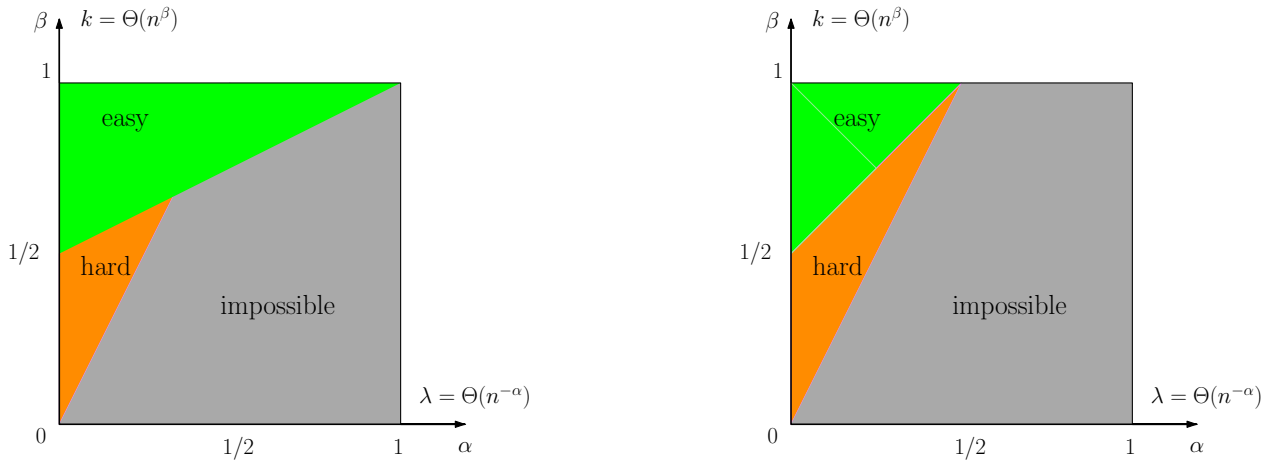
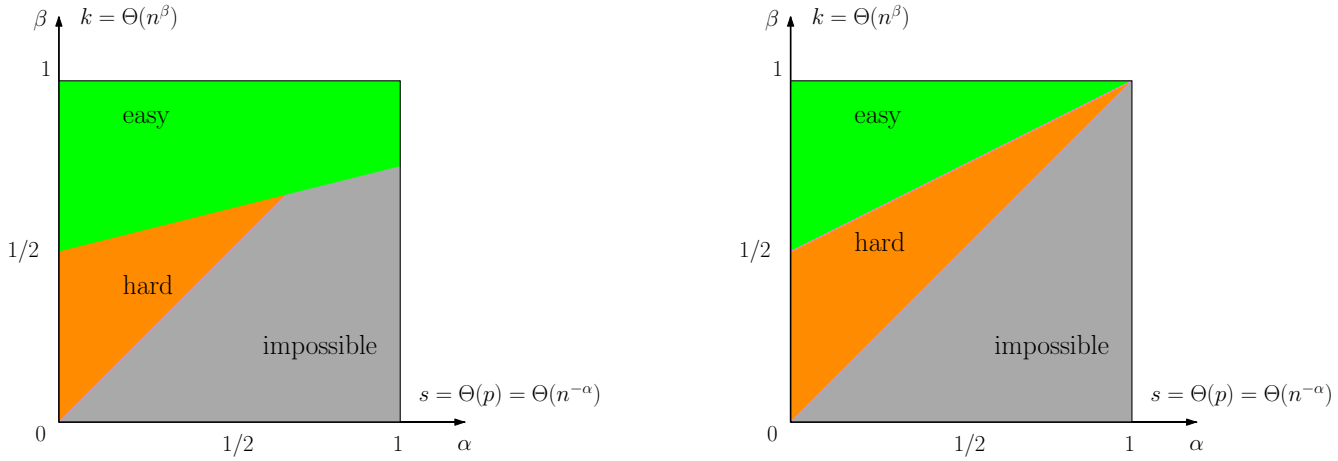


Figure 2: **Planted clique.**
H_0: $G(n, \frac{1}{2})$ random graph on n vertices where each edge is present independently with probability 1/2.
H_1: $G(n, k, \frac{1}{2})$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability 1 if $i, j \in S$ or with probability 1/2 otherwise.



(a) detection (b) recovery

Figure 3: **Spiked Matrix Model** (planted submatrix with elevated mean).
H_0: random $n \times n$ matrix with each entry independent with distribution $N(0,1)$.
H_1: $n \times n$ matrix with each index in set S independently with probability k/n. Each entry independent with distribution $N(\lambda, 1)$ if $i, j \in S$ and with distribution $N(0,1)$ otherwise.



(a) detection (b) recovery

Figure 4: **Planted dense subgraph**.
H_0: $G(n,q)$ random graph on n vertices where each edge is present independently with probability q.
H_1: $G(n,k,q,s)$ with $s > 0$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability $q+s$ if $i, j \in S$ or with probability q otherwise.

B Concentration Inequalities

Sometimes we are interested in a random variable X_n which is very likely to fall within some interval $[a_n, b_n]$ (and this can be very useful for us!). Often we can prove this statement in two steps. First we calculate the expected value. Let $c_n = \mathbb{E}[X_n]$ and suppose for simplicity that $c_n = (a_n + b_n)/2$. The second step is to show it is unlikely that X_n is far from its expected value c_n; i.e. to show $\mathbb{P}(|X_n - c_n| > (a_n - b_n)/2) \to 0$ as $n \to \infty$. Note these two steps together prove that X_n lies in $[a_n, b_n]$

whp, i.e. that $\mathbb{P}(a_n \leqslant X_n \leqslant b_n) \to 1$ as $n \to \infty$.

We say in this case (i.e. when the second step works), that a random variable is *concentrated about its mean* and refer to the bounds below as *concentration inequalities*. We will use these often so collect them in this section of the appendix for easy reference.

Lemma B.1 (Hoeffding's inequality)**.** *Let $S = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are independent and $a \leqslant X_i \leqslant b$ for all i. Then*

$$\mathbb{P}\big(\,|\,S - \mathbb{E}[S]\,|\,\big) \geqslant t \leqslant 2\exp\bigg(-\frac{2t^2}{n(a-b)^2}\bigg).$$

Lemma B.2. *Let $X \sim N(\mu, \sigma^2)$. Then*

$$\mathbb{P}\big(\,|\,X - \mathbb{E}[X]\,|\,\big) \geqslant t \leqslant 2\exp\bigg(-\frac{t^2}{2\sigma^2}\bigg).$$

C Probability Background

We will use many properties of the distributions, we collate these here for reference while reading the proofs or doing exercises.

C.1 General Probability

We say a sequence of events E_n holds whp 'with high probability' if $\mathbb{P}(E_n) \to 1$ as $n \to \infty$.

C.2 Normal Distribution

The following lemma shows the max of m $N(0,1)$ variables is not too big. Note the variables $X_1, \ldots, X_m$ need not be independent.

Lemma C.1. *Let $\varepsilon > 0$. Suppose $X_1, \ldots, X_m \sim N(0,1)$. Then*

$$X_{\max} = \max_{i \in 1,\ldots,m} X_i \leqslant \sqrt{(2+\varepsilon)\log m}$$

with probability tending to 1 as $m \to \infty$.

D Helpful Combinatorial notation and inequalities

The notation $\binom{n}{k}$ read 'n choose k' is the number of ways to pick a set of k items from a set of n items,

$$\binom{n}{k} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots 1} = \frac{n!}{(n-k)!k!}$$

and

$$\frac{(n-k+1)^k}{k^k} \leqslant \binom{n}{k} \leqslant n^k.$$

D.1 Big 'o' notation

We will use notation $O(.)$, $o(.)$, $\omega(.)$ and $\Omega(.)$.

Index

References

[1] Gábor Lugosi. "Lectures on combinatorial statistics". In: *47th Probability Summer School, Saint-Flour* (2017), pp. 1–91.