

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-Rényi 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 $n \rightarrow \infty$; 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 $n \rightarrow \infty$.

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

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

¹Suppose 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 $r(\phi_n) \rightarrow 0$ as $n \rightarrow \infty$. 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 $r(\phi_n) \rightarrow 0$ as $n \rightarrow \infty$. The definition for weak detection being EASY is similar, just replace the condition on $r(\phi_n)$ as required.

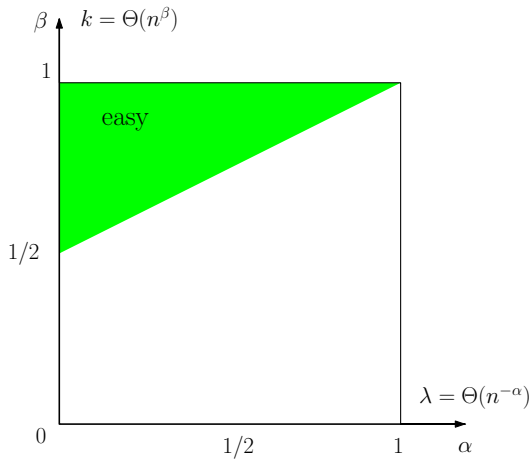
Say for $H_0 : \mathcal{Q}_n(\alpha, \beta)$ vs $H_1 : \mathcal{P}_n(\alpha, \beta)$ that *strong detection for H_0 vs H_1 is POSSIBLE for parameters α, β* if there exists a test ϕ_n such that $r(\phi_n) \rightarrow 0$ as $n \rightarrow \infty$. In particular ϕ_n may be a brute-force algorithm. The definition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

Proof. Let $k = n^\beta$ and $\lambda = n^{-\alpha}$ and to be continued, see also the notes □



(a) detection

Figure 1: see also Figure 2

A List of Planted problems

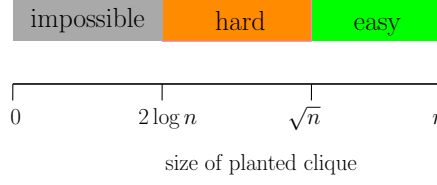
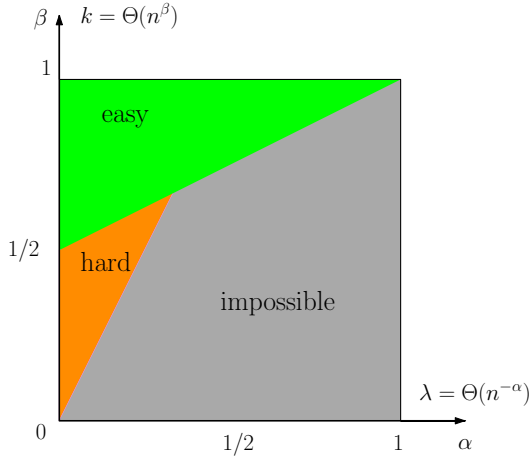


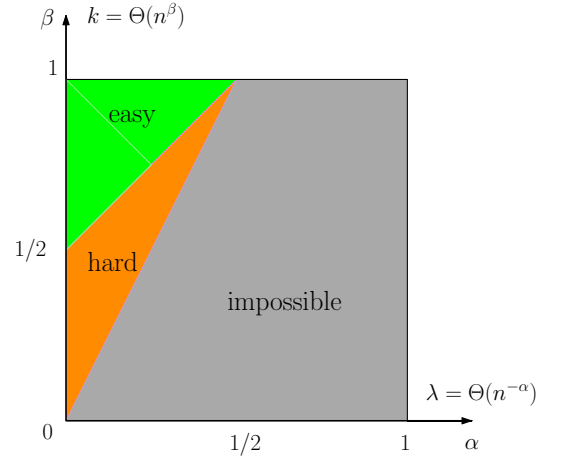
Figure 1: **Planted clique.**

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 2: **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.

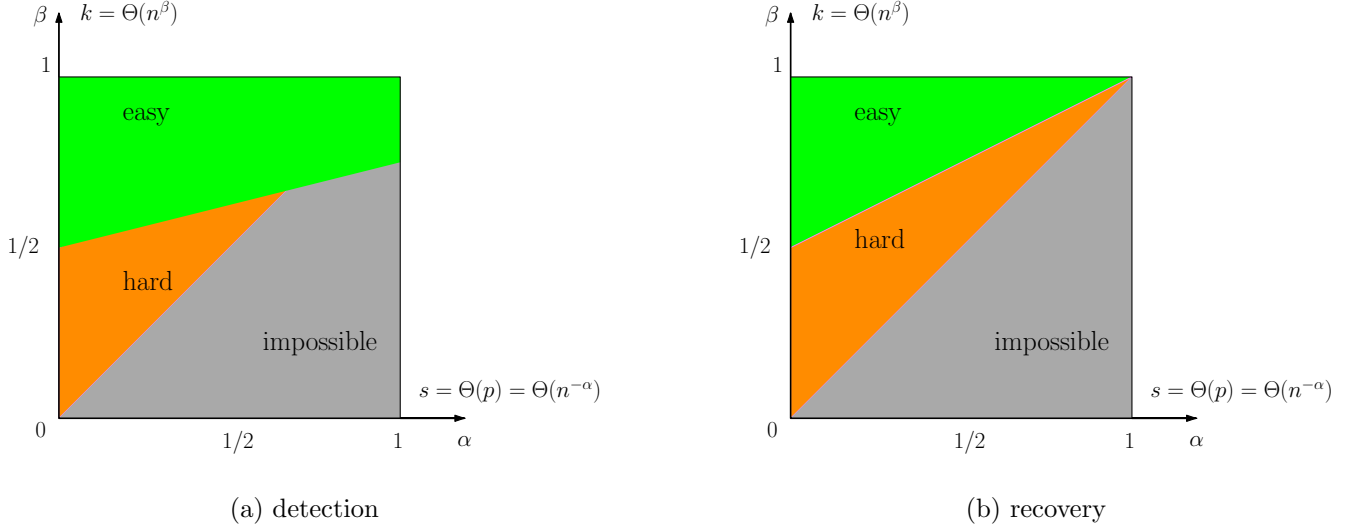


Figure 3: **Planted dense subgraph.**

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 $\mathbb{P}(|X_n - c_n| > (a_n - b_n)/2) \rightarrow 0$ as $n \rightarrow \infty$. Note these two steps together prove that X_n lies in $[a_n, b_n]$ whp, i.e. that $\mathbb{P}(a_n \leq X_n \leq b_n) \rightarrow 1$ as $n \rightarrow \infty$.

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 $S = X_1 + \dots + X_n$ where X_1, \dots, X_n are independent and $a \leq X_i \leq b$ for all i . Then*

$$\mathbb{P}(|S - \mathbb{E}[S]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{n(a-b)^2}\right).$$

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

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Index

$r(\phi)$, 1

detection

easy, 2

possible, 2

strong, 2

weak, 2

risk, 1

strong detection, 2

test, 1

risk, 1

weak detection, 2