Learning on graphs : phase transitions.

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.

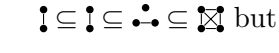
A favourite combinatorial random object is the Erdős–Rényi random graph, $G_{n,1/2}$, take n vertices and between each pair of vertices independently place an edge with probability $1/2$. We are interested in the typical behaviour for large n. The structure we plant is a clique, i.e. between a subset 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 $n \to \infty$; 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.

1 Phase transitions in random graphs[1]

1.1 Graphs and random graphs

Define a (simple) graph[2] $g = (V, E)$ to be a set of labelled vertices $[n] = \{1, 2, \ldots, n\}$ and set of pairs of vertices E which we call edges, with no loops or multiple edges. Write $e(g)$ for the number of edges $|E|$. Technically the edge between vertices i and j should be denoted $\{i, j\}$ but we will use the standard shorthand ij or ji interchangeably.

For graphs $g = (V(g), E(g))$ and $h = (V(h), E(h))$ we say that h and g are *isomorphic*, denoted $h \approx g$ if there is a bijective function $\phi : V(h) \to V(g)$ such that $ij \in E(h)$ if and only if $\phi(i)\phi(j) \in E(g)$. For example  and  . Similarly, we say that graph h is a subgraph of g, denoted $h \subset g$ if there is an injective map $\phi : V(h) \to V(g)$ with $\phi(i)\phi(j) \in E(g)$ for each edge $ij \in E(h)$. For example  but and , since we asked our map to be injective.

Given an integer n and a real number $0 \leqslant p \leqslant 1$, sample random graph $G \sim G(n, p)$ by taking the graph with vertex set $[n] = \{1, 2, ..., n\}$ in which each possible edge ij, $1 \leqslant i < j \leqslant n$, is present with probability p, independently of the others. For any given graph g on $[n]$, the probability of g depends only on the number of edges in g,

$$\mathbb{P}(G = g) = p^{e(g)}(1-p)^{\binom{n}{2}-e(g)}.$$

[1] Section 1 can be skipped by those familiar with the first and second moment method.

[2] We use the convention that random graphs are denoted by G, deterministic graphs by g.

For $G(3,p)$ where the set of possible graphs is $\{$  $\}$. (Note that because the graphs are labelled  $\neq$  .) If we sample a graph $H \sim G(3,p)$ then H is with probability p^3, for each of  ,  , the probability is $p^2(1-p)$, etc. In the special case that $p = 1/2$, then all $\binom{n}{2}$ graphs on vertex set $[n]$ are equally likely.

1.2 Thresholds

Consider how the probability that $G \sim G(n,p)$ contains a triangle changes for different values of p. Clearly, for any n, for $G \sim G(n,0)$ then G has no triangles and for $G \sim G(n,1)$ then G has triangles for $n \geqslant 3$. One can also see that the probability of containing a triangle in increasing in p. We will find that for $p(n)$ and $p'(n)$ 'not too far apart' that for $G \sim G(n,p)$ and $G' \sim G(n,p')$ we have $\mathbb{P}($$\subset G) \to 0$ while $\mathbb{P}($$\subset G) \to 1$ a sort of a 'phase transition' in the behaviour.

Let $\mathcal{A}$ denotes a set of graphs. We say the function $p^* = p^*(n)$ is a (coarse) *threshold* for $\mathcal{A}$ if $\mathbb{P}(G \in \mathcal{A}) \to 0$ for $G \sim G(n,p)$ with $p/p^* \to 0$ and $\mathbb{P}(G \in \mathcal{A}) \to 1$ for $G \sim G(n,p)$ with $p/p^* \to \infty$. Similarly, we say the function $p^* = p^*(n)$ is a *sharp threshold* if the above holds for $p < (1-\varepsilon)p^*$ and $p > (1+\varepsilon)p^*$ respectively.

graph property	(coarse) threshold $p^*(n)$
contains	$n^{-3/2}$
contains cycle	n^{-1}
contains $\triangle$	n^{-1}
non-planar	n^{-1}
connected	$n^{-1}\log n$
contains	$n^{-2/3}$

1.3 First and second moment methods for proving thresholds

A standard technique, the *first moment method*, to show that the probably a random graph contains a subgraph is to show that the expected number of them goes to zero. Similarly, to show that a random graph contains a subgraph it is enough to show that the variance of the number of subgraphs is of smaller order than the expectation squared, and this is termed the *second moment method*.

Lemma 1.1. *Show that the function $p^*(n) = \frac{1}{n^{2/3}}$ is a threshold for $G(n,p)$ containing* *as a subgraph.*

Proof. Let p be such that $p/p^* \to 0$, i.e. $n^{2/3}p \to 0$ and let $G_n \sim G(n,p)$. Let $Y_n = Y_n(G_n)$ count the number of in G_n. For each set S of 4 vertices from $[n]$, let A_S be the event that G_n restricted to the vertices S is a . We can write Y_n in terms of these indicator random variables.

$$Y_n = \sum_{S \in \binom{[n]}{4}} 1_{A_S}$$

As expectation is linear, the expected number of in G_n is

$$\mathbb{E}(Y_n) = \sum_{S \in \binom{[n]}{4}} \mathbb{E}(1_{A_S}) \leqslant n^4 p^6 = (n^{2/3}p)^6.$$

Hence $\mathbb{E}(Y_n) \to 0$ for $n^{2/3}p \to 0$. Observe $\mathbb{P}(G_n \text{ contains a } \boxtimes) = \mathbb{P}(Y_n > 0) \leqslant \sum_{k \geqslant 1} k\mathbb{P}(Y_n = k) \leqslant \mathbb{E}(Y_n)$ and so for $n^{2/3}p \to 0$ whp G_n does not contain $\boxtimes$ as a subgraph.

Now it remains to show that for $p/p^* \to \infty$, i.e. for $n^{2/3}p \to \infty$ that whp $G_n \sim G(n,p)$ contains a $\boxtimes$. For this part of the proof we calculate the variance of Y_n by writing $Y_n = \sum_S 1_{A_S}$ and expanding. Write $\sum_S$ for $\sum_{S \in \binom{[n]}{4}}$.

$$\mathrm{Var}(Y_n) = \mathbb{E}(Y_n^2) - \mathbb{E}(Y_n)^2 \quad = \quad \mathbb{E}\Big(\Big(\sum_S 1_{A_S}\Big)^2\Big) - \Big(\sum_S \mathbb{E}(1_{A_S})\Big)^2.$$

We can rearrange a little to get an expression for the variance in terms of the probabilities of the events A_S and A_T

$$\begin{aligned}
\mathrm{Var}(Y_n) &= \mathbb{E}\Big(\sum_S 1_{A_S} \sum_{T \in [n]^3} 1_{A_T}\Big) - \sum_S \mathbb{E}(1_{A_S}) \sum_T \mathbb{E}(1_{A_T}) \\
&= \sum_{S,T} \Big(\mathbb{E}(1_{A_S} 1_{A_T}) - \mathbb{E}(1_{A_S})\mathbb{E}(1_{A_T})\Big) \\
&= \sum_{S,T} \Big(\mathbb{P}(A_S \,\&\, A_T) - \mathbb{P}(A_S)\mathbb{P}(A_T)\Big). \qquad (1.1)
\end{aligned}$$

If $S \cap T = \varnothing$, i.e. the vertex subsets S and T are disjoint then the events A_S and A_T are independent. Notice this is also true if S and T intersect in one vertex because they still share no edges in common. Hence if $|S \cap T| \leqslant 1$ then $\mathbb{P}(A_S \,\&\, A_T) = \mathbb{P}(A_S)\mathbb{P}(A_T)$ and these terms cancel in the expression for the variance (1.1) above.

So by this observation and (1.1),

$$\mathrm{Var}(Y_n) \quad \leqslant \quad \sum_{|S \cap T| = \{2,3,4\}} \mathbb{P}(A_S \,\&\, A_T). \qquad (1.2)$$

We now consider the three options: $|S \cap T| = 2, 3, 4$. For each of these, for $S, T \in \binom{[n]}{4}$ with the given intersection we want to calculate $\mathbb{P}(A_S \,\&\, A_T)$. For $|S \cap T| = 2$, one edge is shared. There are 10 other edges that need to be present in order to have $\boxtimes$ on both S and on T. Hence $\mathbb{P}(A_S \,\&\, A_T) = p^{11}$ for $|S \cap T| = 2$. Similarly, for $|S \cap T| = 3$, we get $\mathbb{P}(A_S \,\&\, A_T) = p^9$ and for $|S \cap T| = 4$, we get $\mathbb{P}(A_S \,\&\, A_T) = \mathbb{P}(A_S) = p^6$.

The aim is to find an upper bound for the right hand side of (1.2). Hence we want to know how many $S, T \in \binom{[n]}{4}$ for each of the possible overlaps. When S and T overlap on 2 vertices, the number of ways to pick them is to first pick the set of vertices in S then pick the two vertices in S that will ovelap with T, and lastly pick the last two vertices in T (the ones that don't overlap with S). This makes $\binom{n}{4}\binom{4}{2}\binom{n}{2}$. Actually all we need is that the number of $S, T \in \binom{[n]}{4}$ which overlap on two vertices is at most n^6. Similarly the number that overlap on three vertices is at most n^5 and the number overlapping on all four vertices is at most n^4. Thus, from (1.2),

$$\mathrm{Var}(Y_n) \quad \leqslant \quad n^6 p^{11} + n^5 p^9 + n^4 p^6. \qquad (1.3)$$

Now we have a good upper bound on the variance. What we actually want to show is that whp G_n contains a $\boxtimes$. In other words we want to show whp $Y_n > 0$.

We use the following non-obvious idea. I have some b for which I know $b > 0$ and I want to use this to show that $a > 0$. Notice it is enough to show that $|b - a| < b$. (Or, equivalently that it is unlikely that $|b - a| \geqslant b$.) We show Y_n is likely non-zero by showing it is sufficiently close to $\mathbb{E}[Y_n]$ which we know to be positive. By some re-arranging and Chebyshev,

$$\mathbb{P}(Y_n = 0) \leqslant \mathbb{P}\Big(|Y_n - \mathbb{E}(Y_n)| \geqslant \mathbb{E}(Y_n)\Big) \leqslant \mathbb{P}\Big(|Y_n - \mathbb{E}(Y_n)| \geqslant \mathbb{E}(Y_n)\Big) \leqslant \frac{\mathrm{Var}(Y_n)}{\mathbb{E}(Y_n)^2}.$$

The problem is now reduced to terms we have already calculated. By (1.3),

$$\mathbb{P}(Y_n = 0) \leqslant \frac{n^6 p^{11} + n^5 p^9 + n^4 p^6}{(\binom{n}{4} p^6)^2}. \tag{1.4}$$

For $n^{2/3}p \to \infty$ the fraction in (1.4) goes to zero, and thus whp G contains a ⊠ as a subgraph. □

2 Detection

2.1 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 $\phi_n(g) \in \{0, 1\}$[3]. 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

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

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 (regardless of the input graph g). Both of these have risk 1.

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

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

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 not known to be EASY) and we have 'evidence of hardness'. We will specify which evidence of hardness.

3 Planted Clique

Our approach for the possible and impossible regions of planted clique follows closely that of [5].

3.1 When planted clique is POSSIBLE

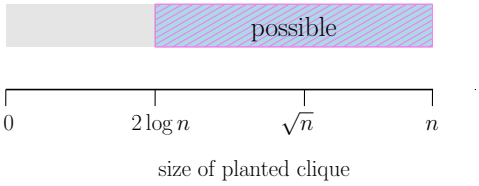


Figure 1: We show the detecting planted clique is possible in the dashed region in Lemma 3.1. We prove there is a (brute-force) test that distinguishes H_0: $G(n, 1/2)$ and H_1: $G'(n, k, 1/2)$ with high probability.

Lemma 3.1. *Let $k = k(n) > 2\log_2 n + 3$. Then for $H_0 : G(n, 1/2)$ vs $H_1 : G'(n, k, 1/2)$ strong detection is* POSSIBLE.

For graph g, define $\omega(g)$ to be the size of the largest clique in g, i.e. the size of the largest set of vertices S such that each pair of vertices in S is connected by an edge in graph g.

Proof. Our test will work by thresholding on the size of the largest clique in the graph. Let

$$\phi_n(g) = \begin{cases} 1, & \text{if } \omega(g) > 2\log_2 n + 3, \\ 0, & \text{otherwise.} \end{cases}$$

Then the risk of this test is

$$r(\phi_n) = \mathbb{P}_0(\phi_n = 1) + \mathbb{P}_1(\phi_n = 0) = \mathbb{P}_0\big(\omega(G) > 2\log_2 n + 3\big) + \mathbb{P}_1\big(\omega(G) \leqslant \log_2 n + 3\big).$$

Note that the size of the largest clique in the planted model is at least the size of the planted clique (it might be bigger if there is another vertex which happens to be connected to each vertex in the planted clique). Since, in P_n, we have planted a clique of size $2\log_2 n$, we have

$$\mathbb{P}_1(\phi_n(G) = 0) = \mathbb{P}_1(\omega(G) \leqslant \log_2 n + 3) = 0.$$

Thus the risk simplifies to consider only the size of the largest clique in $G(n, 1/2)$

$$r(\phi_n) = \mathbb{P}_0\big(\omega(G) > 2\log_2 n + 3\big).$$

Let N_m be the number of cliques of size m. Since $\mathbb{P}_0(N_m \geqslant 1) \leqslant \mathbb{E}_0[N_m]$ it suffices to bound the expected number of cliques. (This is an example of the 'first moment method'.)

Then we may calculate

$$\begin{aligned}\mathbb{E}_0[N_m] &= \sum_{S:\,|S|=m} \mathbb{P}_0\big(S \text{ is a clique in} G\big) \\ &= \binom{n}{m} 2^{-\binom{m}{2}} \\ &\leqslant n^m 2^{-m(m-1)/2} \\ &= (n2^{-(m-1)/2})^m.\end{aligned}$$

One may then check that for $m \leqslant 2\log_2 n + 3$ that $n2^{-(m-1)/2} \leqslant 1/2$. And thus for $m \leqslant 2\log_2 n + 3$

$$\mathbb{E}_0[N_m] \leqslant 2^{-m}.$$

and hence for $m \leqslant 2\log_2 n + 3$, $E_0[N_m] \to 0$ as $n \to \infty$.

Hence, since $\mathbb{P}_0(N_m \geqslant 1) \leqslant \mathbb{E}_0[N_m]$, we have $\mathbb{P}_0(\omega(G) \geqslant 2\log_2 n + 3) \to 0$ as $n \to \infty$ and thus we have that the risk of our test goes to zero as n goes to ∞ as required. □

4 Likelihood ratio and risk

We define the likelihood ratio between discrete probability spaces $H_0 : Q$ and $H_1 : P$ by

$$L(g) = \frac{\mathbb{P}_1(G = g)}{\mathbb{P}_0(G = g)}. \tag{4.1}$$

Define $\phi^* = \phi^*(P, Q)$, the *likelihood ratio test* to be the following test

$$\phi^*(g) = \begin{cases} 1, & \text{if } L(g) > 1, \\ 0, & \text{if } L(g) \leqslant 1. \end{cases}$$

Lemma 4.1. *Suppose P and Q are discrete probability spaces. The test ϕ^* achieves minimal risk over tests to distinguish $H_0 : Q$ and $H_1 : P$.*

Proof. Fix $\phi \neq \phi^*$. We prove the lemma by showing that $r(\phi) \geqslant r(\phi^*)$. We calculate

$$\begin{aligned} r(\phi) - r(\phi^*) &= \sum_{g:\phi(g)=1} \mathbb{P}_0(G = g) + \sum_{g:\phi(g)=0} \mathbb{P}_1(G = g) - \sum_{g\,:\,\phi^*(g)=1} \mathbb{P}_0(G = g) - \sum_{g\,:\,\phi^*(g)=0} \mathbb{P}_1(G = g) \\ &= \sum_{g:\phi^*(g)=1,\phi(g)=0} \underbrace{\mathbb{P}_1(G = g) - \mathbb{P}_0(G = g)}_{>0} + \sum_{g:\phi^*(g)=0,\phi(g)=1} \underbrace{\mathbb{P}_0(G = g) - \mathbb{P}_1(G = g)}_{\geqslant 0} \\ &\geqslant 0. \end{aligned}$$

□

Having established that ϕ^* achieves minimal risk over all tests, we may now calculate the minimal risk possible. Recall the total variation distance may be defined,

$$d_{\mathrm{TV}}(P, Q) = \sum_g |\mathbb{P}_P(G = g) - \mathbb{P}_Q(G = g)|.$$

Lemma 4.2. *Suppose P and Q are finite discrete probability spaces. For $H_0 : G \sim Q$ and $H_1 : G \sim P$, the likelihood ratio test ϕ^* satisfies*

$$r(\phi^*) = 1 - \tfrac{1}{2}\mathbb{E}_0[\,\big|L(G) - 1\big|\,] = 1 - \tfrac{1}{2}d_{\mathrm{TV}}(P, Q)$$

Proof. We first note that

$$\mathbb{E}_0[\,\big|L(G) - 1\big|\,] = \sum_g \mathbb{P}_0(G = g)\,\mathbb{E}_0[\,\big|L(g) - 1\big|\,] = \sum_g \mathbb{P}_0(G = g)\,\mathbb{E}_0[\,\big|\frac{\mathbb{P}_1(G = g)}{\mathbb{P}_0(G = g)} - 1\big|\,] = d_{\mathrm{TV}}(P, Q)$$

Now by definition,

$$r(\phi^*) = \sum_g \mathbb{P}_0(G = g)\,\mathbf{1}[L(g) \geqslant 1] + \sum_g \mathbb{P}_1(G = g)\,\mathbf{1}[L(g) < 1]$$

Then noting we may make the substitution $\mathbf{1}[L(g) < 1] = \frac{1}{2}\mathbf{1}[L(g) < 1] + \frac{1}{2} - \frac{1}{2}\mathbf{1}[L(g) \geqslant 1]$ and symmetrically for the other indicator we obtain;

$$\begin{aligned} r(\phi^*) &= 1 + \tfrac{1}{2}\sum_g \mathbf{1}[L(g) \geqslant 1]\big(\mathbb{P}_0(G = g) - \mathbb{P}_1(G = g)\big) + \tfrac{1}{2}\sum_g \mathbf{1}[L(g) < 1](\mathbb{P}_1(G = g) - \mathbb{P}_0(G = g)) \\ &= 1 + \tfrac{1}{2}\sum_{g\,:\,L(g)\geqslant 1} \underbrace{\big(1 - \frac{\mathbb{P}_1(G = g)}{\mathbb{P}_1(G = g)}\big)}_{(*)}\,\mathbb{P}_0(G = g) + \tfrac{1}{2}\sum_{g\,:\,L(g)<1} \underbrace{(\frac{\mathbb{P}_1(G = g)}{\mathbb{P}_0(G = g)} - 1)}_{(*)}\,\mathbb{P}_0(G = g) \end{aligned}$$

Since both expressions denoted by $(*)$ equate to $|L(g)-1|$, we obtain that $r(\phi) = 1-\mathbb{E}_0[\,\big|L(G)-1\big|\,]$. □

Now, recall that for a random variable X, $\mathbb{E}[|X|] \leqslant \sqrt{\mathbb{E}[X^2]}$ (to see this note $\mathrm{Var}[|X|] \geqslant 0$). Thus, by the result above,

$$r(\phi^*) = 1 - \mathbb{E}_0[\,\big|L(G) - 1\big|\,] \geqslant 1 - (\mathbb{E}_0[\,\big(L(G) - 1\big)^2\,])^{1/2} = 1 - (\mathbb{E}_0[L(G)^2] - 1)^{1/2}$$

where the last inequality follows since $\mathbb{E}_0[L(G)] = \sum_g \mathbb{P}_0[G = g]\,L(g) = 1$. This gives the following corollary.

Corollary 4.1. *Suppose P and Q are discrete probability spaces. The likelihood ratio test ϕ^* has risk $r(\phi^*) \geqslant 1 - \frac{1}{2}(\mathbb{E}_0[L(G)^2] - 1)^{1/2}$.*

(actual) end L1

4.1 When planted clique is IMPOSSIBLE

It turns out that we have very good control on the likelihood ratio for $H_0 : G(n, 1/2)$ vs $H_1 : G'(n, k, 1/2)$ and this will allow us to show that $r(\phi^*) \to 1$ for k sufficiently below $2\log_2 n$.

Lemma 4.3. *Let $k = k(n) < 2\log_2 n - 5\log_2\log_2 n$. Then for $H_0 : G(n, 1/2)$ vs $H_1 : G'(n, k, 1/2)$ strong detection is* IMPOSSIBLE.

Proof. First observe that for $G \sim G(n, 1/2)$ all (labelled) graphs on n vertices are equally likely, and hence $\mathbb{P}_0(G = g) = (1/2)^{\binom{n}{2}}$. For the planted clique model let S^* denote the set of planted vertices, in $G'(n, k, 1/2)$, S^* is distributed uniformally over all k-element subsets of $[n]$, i.e. $S^* \in^u \binom{[n]}{k}$. Then we may calculate,

$$\begin{aligned} \mathbb{P}_1(G = g) &= \sum_{|S|=k} \mathbb{P}_1(G = g \,|\, S^* = S)\,\mathbb{P}(S^* = S) \\ &= \frac{(1/2)^{\binom{n}{2}-\binom{k}{2}}}{\binom{n}{k}} \sum_{|S|=k} \mathbb{P}(\mathbf{1}[S \text{ is a clique in } g]) \end{aligned}$$

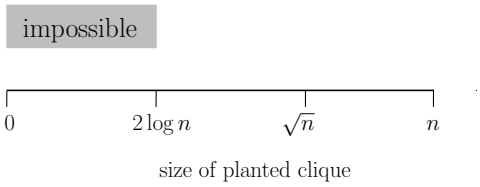


Figure 2: We show the detecting planted clique is impossible in the grey region. See Lemma 4.3.

Hence we have an exact expression for the likelihood ratio

$$L(g) = 2^{\binom{k}{2}} \binom{n}{k}^{-1} \sum_{|S|=k} \mathbf{1}[S \text{ is a clique in } g].$$

and can thus calculate $\mathbb{E}_0[L(G)^2]$,

$$\begin{aligned}
\mathbb{E}_0[L(G)^2] &= \sum_g \mathbb{P}_0(G=g) L(g)^2 \\
&= \frac{2^{k(k-1)}}{\binom{n}{k}^2} \sum_g \mathbb{P}_0(G=g) \Big(\sum_{|S|=k} \mathbf{1}[S \text{ is a clique in } g] \Big)^2 \\
&= \frac{2^{k(k-1)}}{\binom{n}{k}^2} \sum_{|S|,|T|=k} \mathbb{P}_0\big(S, T \text{ both cliques in } G\big) \\
&= \frac{2^{k(k-1)}}{\binom{n}{k}^2} \sum_i \sum_{|S|,|T|=k, |S\cap T|=i} (\frac{1}{2})^{2\binom{k}{2} - \binom{i}{2}} \\
&= \frac{1}{\binom{n}{k}^2} \sum_i \sum_{|S|,|T|=k, |S\cap T|=i} 2^{\binom{i}{2}}
\end{aligned}$$

We may then note that the number of ways to choose k-element sets S, T with overlap i is $\binom{n}{k}\binom{k}{i}\binom{n-k}{k-i}$ – first choose S then the i vertices to overlap then choose the remainder of T. Hence (with the convention $\binom{0}{2} = \binom{1}{2} = 0$),

$$\mathbb{E}_0[L(G)^2] = \frac{1}{\binom{n}{k}} \sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{k-i} 2^{\binom{i}{2}}.$$

Now, since $\binom{n}{k} = \sum_i \binom{n-k}{k-j}\binom{k}{j}$, by subtracting 1 we lose the first two terms of the sum,

$$\mathbb{E}_0[L(G)^2] - 1 = \frac{1}{\binom{n}{k}} \sum_{i=0}^{k} \binom{k}{i} \binom{n-k}{k-i} (2^{\binom{i}{2}} - 1) \leqslant \frac{1}{\binom{n}{k}} \sum_{i=2}^{k} \binom{k}{i} \binom{n-k}{k-i} 2^{\binom{i}{2}}. \tag{4.2}$$

Hence to show asymptotically trivial risk for the likelihood ratio test i.e. that $r(\phi^*) = 1 - o(1)$, by Corollary 4.1 it is enough to show that the final expression in (4.2) is $o(1)$ for $k < 2\log_2 n - 5\log_2\log_2 n$: see [2, Thm 7.3], and we are done. □

(plan) end L1

5 Acknowledgements and bibliographic notes for each section.

Grateful to many for discussions which influenced these note including Misha Isaev, Gabor Lugosi, Svante Janson, Cindy Rush, Anda Skeja and Alex Wein. Also, many parts of these notes are inspired by expositions of others, and I recommend these references for further reading: particularly lecture notes by Lugosi [5] and by Wu and Xu [6], for an introduction to low-degree the introduction and Chapter 2 of [4], and for further theory on reductions in total variation distance the paper of Brennan, Bresler and Huleihel. For general theory of random graphs, Frieze and Karoński's textbook is excellent and available free online [2].

5.1 Sections 3 and 4

Sections 3 and 4 concern the IT-threshold, for POSSIBLE vs IMPOSSIBLE for testing between binomial random graphs and the planted clique model. See [1] and references therein for results on this testing problem (as well as other combinatorial testing problems). Our treatment follows notes of Lugosi [5].

We note much sharper results are known for POSSIBLE and IMPOSSIBLE regions for planted clique. For the test function considered, where we threshold on the size of the largest clique, the important behaviour to understand was the size of the largest clique in $G(n, 1/2)$. For constant $0 < p < 1$, take $k_0 = \lceil 2\log_{1/p} n - 2\log_{1/p}\log_{1/p} n + 2\log_{1/p} e/2 + 1 + o(1)\rceil$. Then it is known that for $G \sim G(n,p)$, whp $k_0 - 1 \leqslant w(G) \leqslant k_0$, i.e. two-point concentration! This was proven independently by Bollobas and Matula and is important for estimating the chromatic number of $G(n, 1/2)$. For discussion see also equation (2) and Figure 1 of [3].

A List of Planted problems

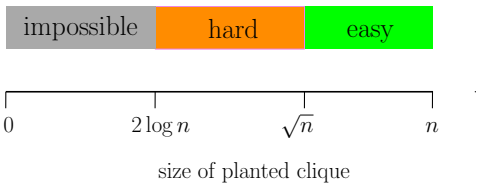
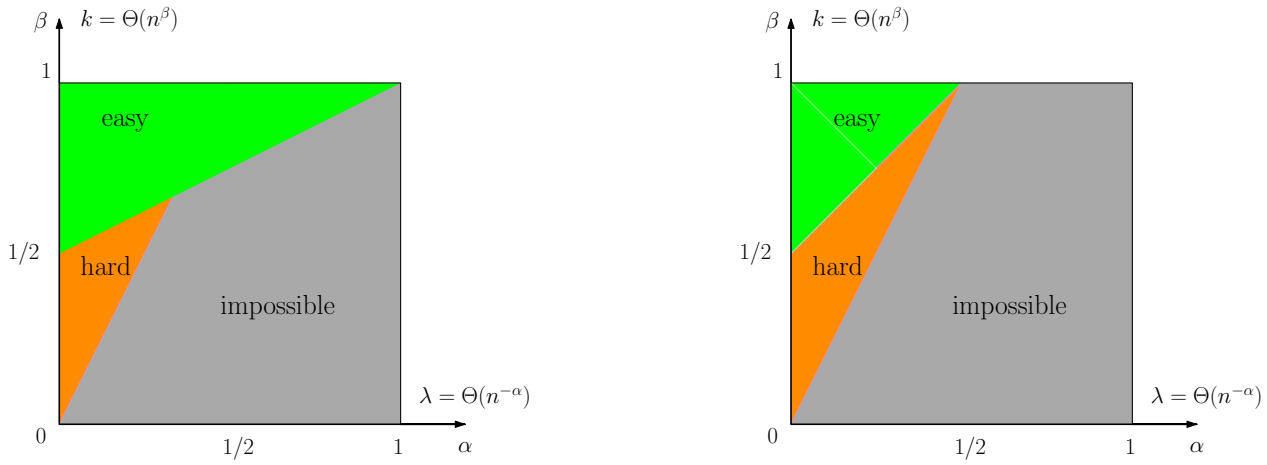


Figure 2: **Planted clique.**
H_0: $G(n, \frac{1}{2})$ random graph on n vertices where each edge is present independently with probability 1/2.
H_1: $G(n, k, \frac{1}{2})$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present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



(a) detection (b) recovery

Figure 3: **Spiked Matrix Model** (planted submatrix with elevated mean).
H_0: a random $n \times n$ matrix with each entry independent with distribution $N(0,1)$.
H_1: $\mathrm{BC}(n,k,\lambda)$, an $n \times n$ 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.

B Probability Background

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

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

B.1 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*

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

$$P(\mathrm{Bin}(n,p) > c_0/p_0) \leqslant P(|\mathrm{Bin}(n,p) - np| > np - c_0/p_0)$$

So in the notation above $t = np - c_0/p_0$, want $n(p - c_0/(p_0 n))^2$ large.

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

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

B.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 B.3. *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 $m \to \infty$.

It will also be useful to approximate the binomial distribution with the normal distribution. Below we state the Berry-Esseen theorem, in the special case of comparing the cumulative distributions functions (CDFs) of the binomial and normal distributions [6, Lemma 2.4].

Theorem B.1 (Berry-Esseen)**.** *There exists an absolute constant C such that*

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(\mathrm{Binom}(n,p) \leqslant x) - \mathbb{P}(N(np, np(1-p)) \leqslant x)| \leqslant \frac{C}{\sqrt{np}}.$$

We will denote the CDF of the standard normal by $\Phi(x) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\,dx$, and the complementary CDF by $\overline{\Phi}(x) = 1 - \Phi(x)$.

C 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

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

The following form of Stirlings approximation for binomials can also be useful.

$$\sqrt{2\pi n}\big(\frac{n}{e}\big)^n < n! < e\sqrt{n}\big(\frac{n}{e}\big)^n.$$

We will use 'Big 'o" notation $O(.)$, $o(.)$, $\omega(.)$ and $\Omega(.)$, see definitions in [2].

Index

References

[1] Louigi Addario-Berry et al. "On combinatorial testing problems". In: *The Annals of Statistics* 38.5 (2010), pp. 3063–3092.

[2] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.

[3] Annika Heckel and Oliver Riordan. "How does the chromatic number of a random graph vary?" In: *Journal of the London Mathematical Society* 108.5 (2023), pp. 1769–1815.

[4] Samuel Hopkins. *Statistical inference and the sum of squares method*. Cornell University, 2018.

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

[6] Yihong Wu and Jiaming Xu. "Statistical Inference on Graphs: Selected Topics". In: *Lecture notes.* (2022). URL: `https://people.duke.edu/~jx77/stats-graphs.pdf`.