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.

A List of Planted problems

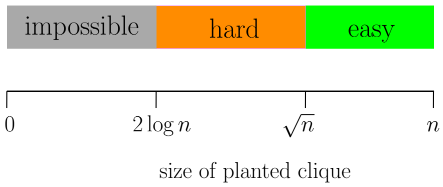
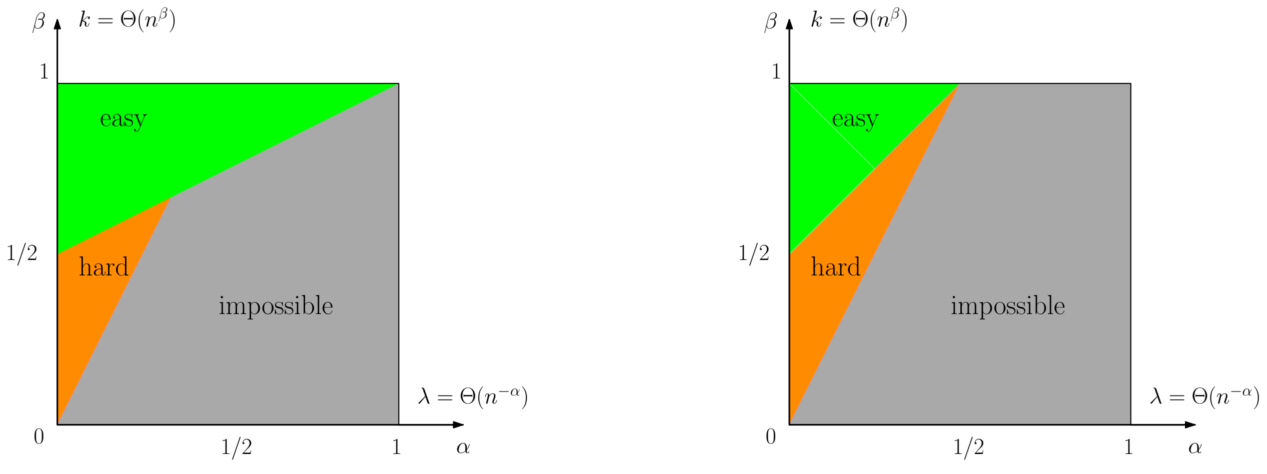


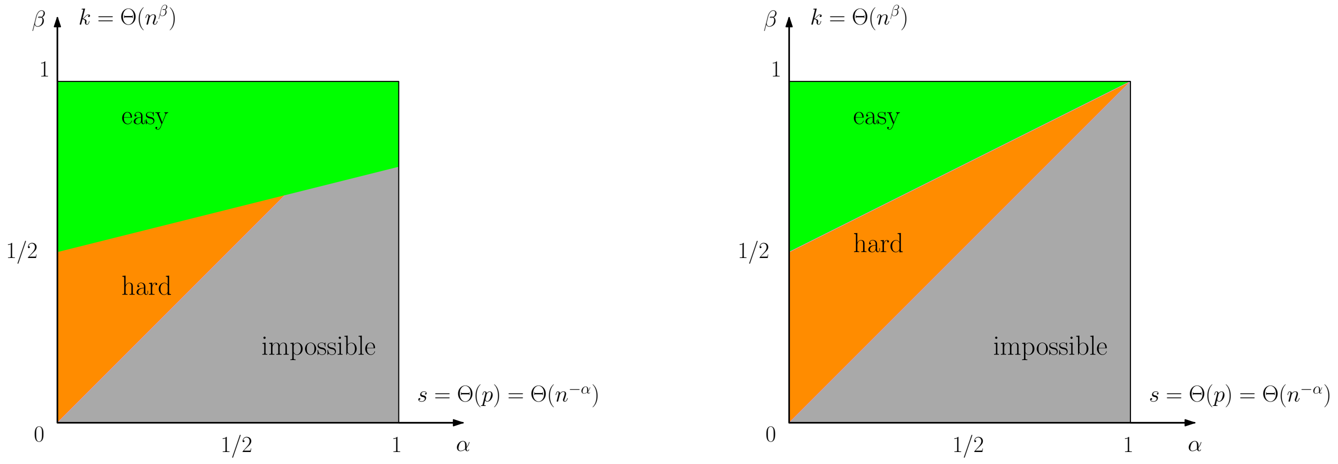
Figure 0: **Planted clique.**
H_0: $G(n, \frac{1}{2})$ random graph on n vertices where each edge is present independently with probability 1/2.
H_1: $G(n, k, \frac{1}{2})$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability 1 if $i, j \in S$ or with probability 1/2 otherwise.



(a) detection

(b) recovery

Figure 1: **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 2: **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.