Signal in Noise – Graph versions

A clique in a graph is a subset of vertices, all of which are adjacent to each other. A planted clique is a clique created from another graph by adding edges between all pairs of a selected subset of vertices.

The planted clique problem can be formalized as a decision problem over a random distribution on graphs, parameterized by two numbers, n (the number of vertices), and k (the size of the clique). These parameters may be used to generate a graph, by the following random process:[1]

1. Create an Erdős–Rényi random graph on n vertices by choosing independently for each pair of vertices whether to include an edge connecting that pair, with probability 1/2 for each pair.
2. Decide whether or not to add a clique to the graph, with probability 1/2; if not, return the graph formed in step 1.
3. Choose randomly a subset of k of the n vertices and add an edge (if one is not already present) between each pair of the selected vertices.

The problem is then to determine algorithmically whether one of the graphs resulting from this process contains a clique of at least k vertices.

Above is from wikipedia….put simply (but less elegantly) – given a graph that you suspect might contain a clique of size k, in a background of noise. or might just be all "noise" (an Erdos-Renyi graph) – can you perform a test to give a high probability correct answer to "is this all noise or is there a k-clique in it"?

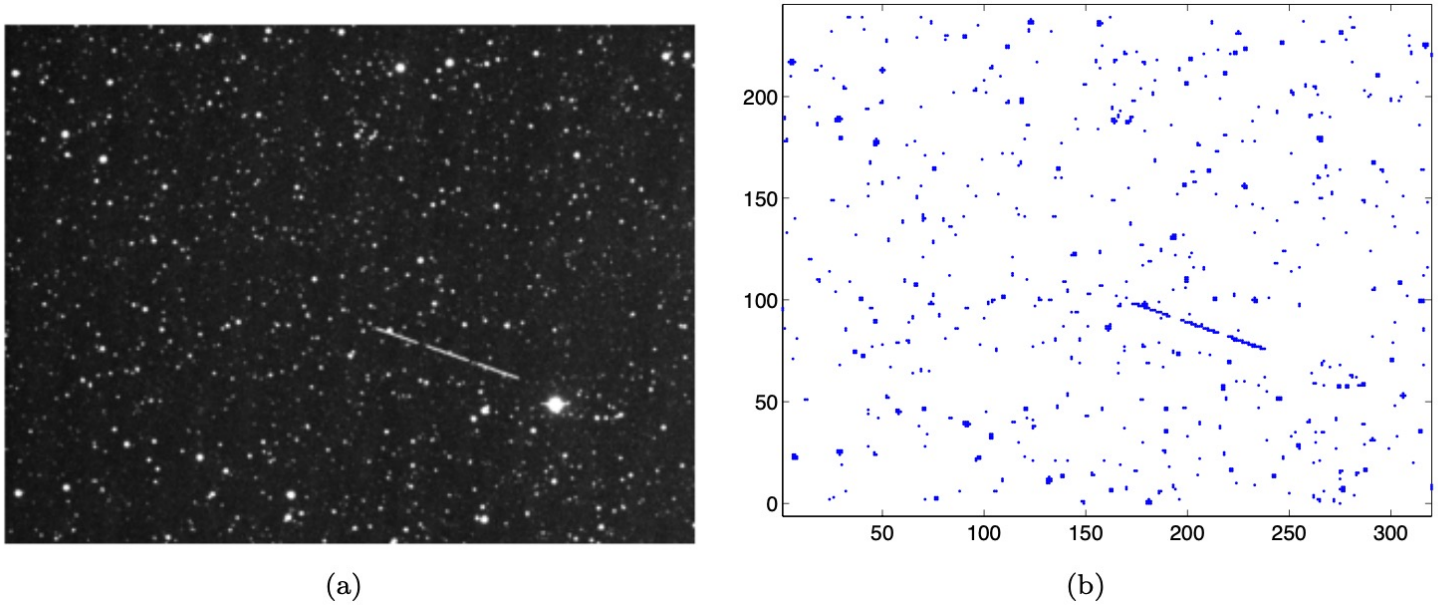
Of course variations

- Of the detection problem
  - Is there a clique of size at least k etc.
- Of the search problem
  - Find the clique, or
  - Approximately find the clique (for whatever sensible notion of approximately you think relevant).

The detect signal in noise is an important problem!

- Practically – most algorithms assume there is a signal. When given “white noise” – they still return an answer.
- Theoretically – since detection ought to be typically easier than finding…then if you can’t detect with reliability/speed then you almost certainly can’t find (in a reasonable way)
- There is a subtlety here – the Erdos-Renyi graph process will almost certainly have a k-clique above a certain edge density/probability of adding an edge. So finding a k-clique doesn’t mean it is a “meaningful k-clique”. Notion of statistically impossible to decide (if the clique ius planted….meaningful).
- This is not just a theoretical issue…arises in my own area of research - robust fitting…is the “structure” you found just noise (outliers) or a real structure.

When is your structure actual structure or “accident”. Signal in noise. Verses ”just noise”



Imagine “even worse” – more background noise (stars) – smaller part of ”streak” or “detections on streak more widely spaced”

Figure 1.1: Estimating satellite trajectory by finding linear streaks. (a) Input image containing an orbiting satellite obtained via telescope (courtesy of Defence Science Technology Group, Australia’s primary defence science research organisation). (b) A set of points obtained by intensity thresholding, removal of large blobs and centroiding. Observe that there exist a significant number of outliers, *i.e.*, points not lying on the target line.

Cliques and Independent sets – two sides of a coin (homogeneous sets)

- Informal – loose – thinking/description.
- A clique is “structure” - ”everybody holding a property in common – with all other members of the clique”
- An independent set is the “opposite” – “no-one having that property in common with anybody else in that set”
- Actually, since a clique in the graph is an independent set in the complement graph, they are “two sides of the same coin”. The independent set vertices do share a common property – the negation of the property defining the clique!

Cliques and Independent sets – two sides of a coin (homogeneous sets)

- So in a sense they BOTH describe “order”/ “structure”.
- So in any given graph “how much order/disorder” can there be?
  - What is the maximum size of any clique or independent set in that given graph?
  - Over any choice of graph – what is the minimum amount of “order” you will see? (In some sense “order”, as a clique, must show up in either the graph or its complement).
  - It is known that any graph must have a homogeneous set (clique or independent set) of size at least O(log(N)). i.e., $\max(\alpha(G),w(G))>O(\log(N))$ for any graph. So there is a fixed floor for the amount of “disorganization”/ “lack of structure” in any graph – even one that “ought to be, with high probability, maximum disordered”.

Cliques and Independent sets – two sides of a coin (homogeneous sets)

- Note: a randomly chosen graph has no reason to "push" the order into the graph or the complement of the graph and so will likely have a clique of that size (and also likely to have an independent set of that size).
- But if you bias the "random" choice towards higher numbers of edges then of course you bias the max(α(G),w(G)) to be given by the max clique size, rather than the max independent set size.
- More precisely (and choosing what "random" means) for the Erdos-Renyi "random noise" graph (p=0.5) and a clique planted in that:

> There exists a function $f(n) \sim 2\log_2 n$ such that asymptotically almost surely, the size of the largest clique in an n-vertex random graph is either $f(n)$ or $f(n)+1$,[2] and there exists some constant c such that the expected number of cliques of size $\geq f(n) - c$ converges to infinity. Consequently, one should expect that the planting a clique of size $\sim 2\log_2 n$ cannot be detected with high probability.

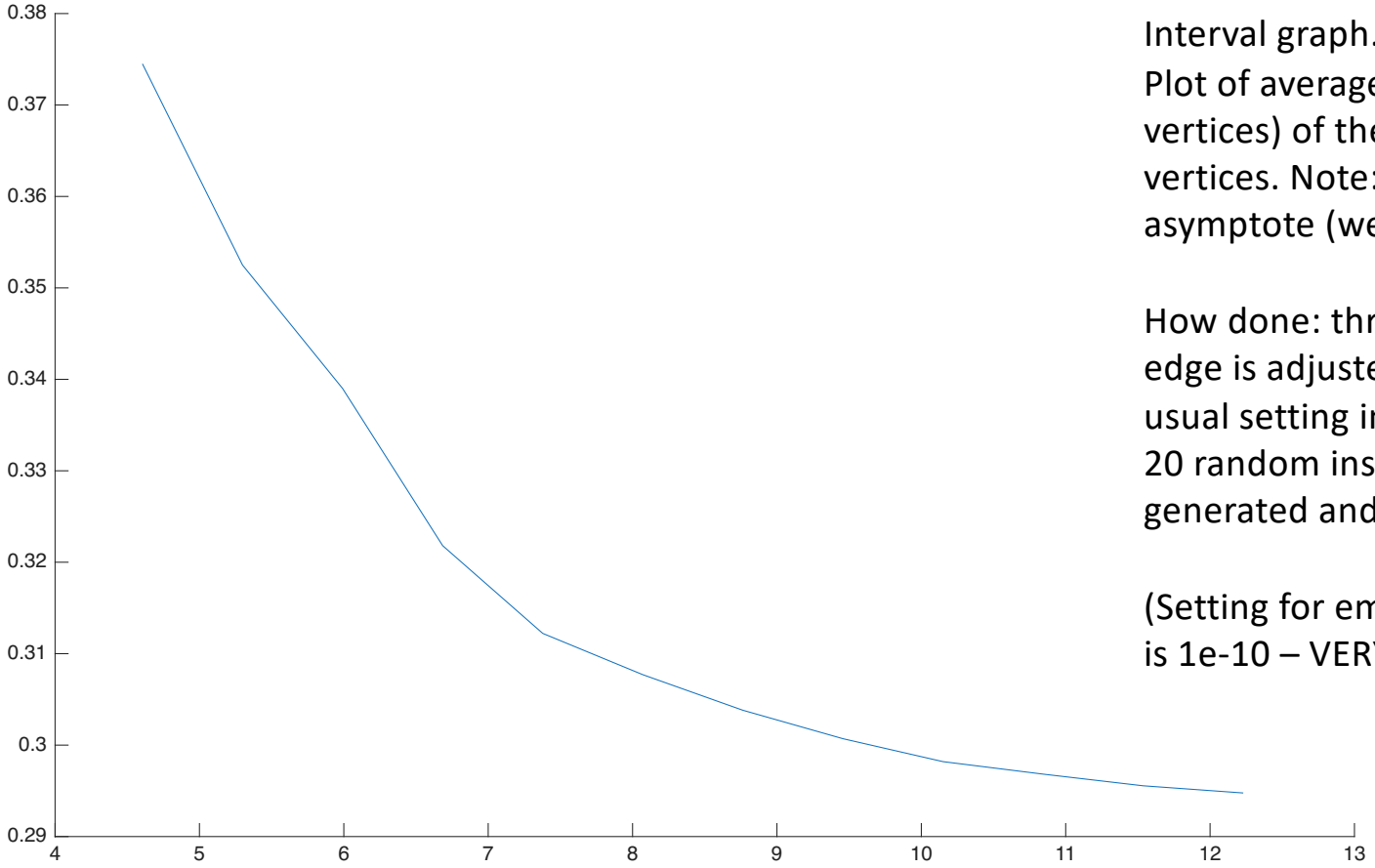
Cliques and Independent sets – two sides of a coin (homogeneous sets)

- This has significant implications for “finding structure in data”
- For example – any meaningful structure (i.e., not naturally occurring in *any* given graph/data set) must be larger than O(log(N)) as for even the “complete noise” graph will still have a homogeneous set (structure) of that size.
- Put another way, if the “signal” (structure) you are trying to recover (identify/detect etc) is of size O(log(N)) – in a graph on N vertices – then it is *informationally* impossible. Your SNR has to be higher!
- Above this information threshold, it *ought* to be possible to recover the signal (and indeed brute force/exhaustive search will) but it turns out that in most situations examines we have good reasons ro believe that probably no “efficient” algorithm can unless the structure reaches some higher threshold in size. This is a very interesting an active research area (planted partitions, planted clique, planted low rank submatrix, etc. etc.)

Cliques and Independent sets – two sides of a coin (homogeneous sets)

- But what if we know our data must be restricted to some “special sets of graphs”. Is O(log(N)) now the information theoretic limit for exact recovery?
- There is a famous conjecture (Erdos-Hajnal) that if you forbid any given induced subgraph then the “situation” changes to now $O(N^c)$ where c is a constant depending on the subgraph. It is not yet determined if it is true (in general), but it is true for some specific cases. In other words, “random” (likely to have maximum disorder for that class) chosen graph from the given class (excluding the subgraph) must still have order as large as $O(N^c)$. Of course this is only meaningful for $c<1$
- Restricting you class of graphs INCREASES the unavoidable “background”/inherent structure! Of course, if you also bias the number of edges then you also incrwase/decrease the likelihood the maximum “false” structure is buried in the graph or the complement of the graph.

Expected clique sizes in random geometric graphs

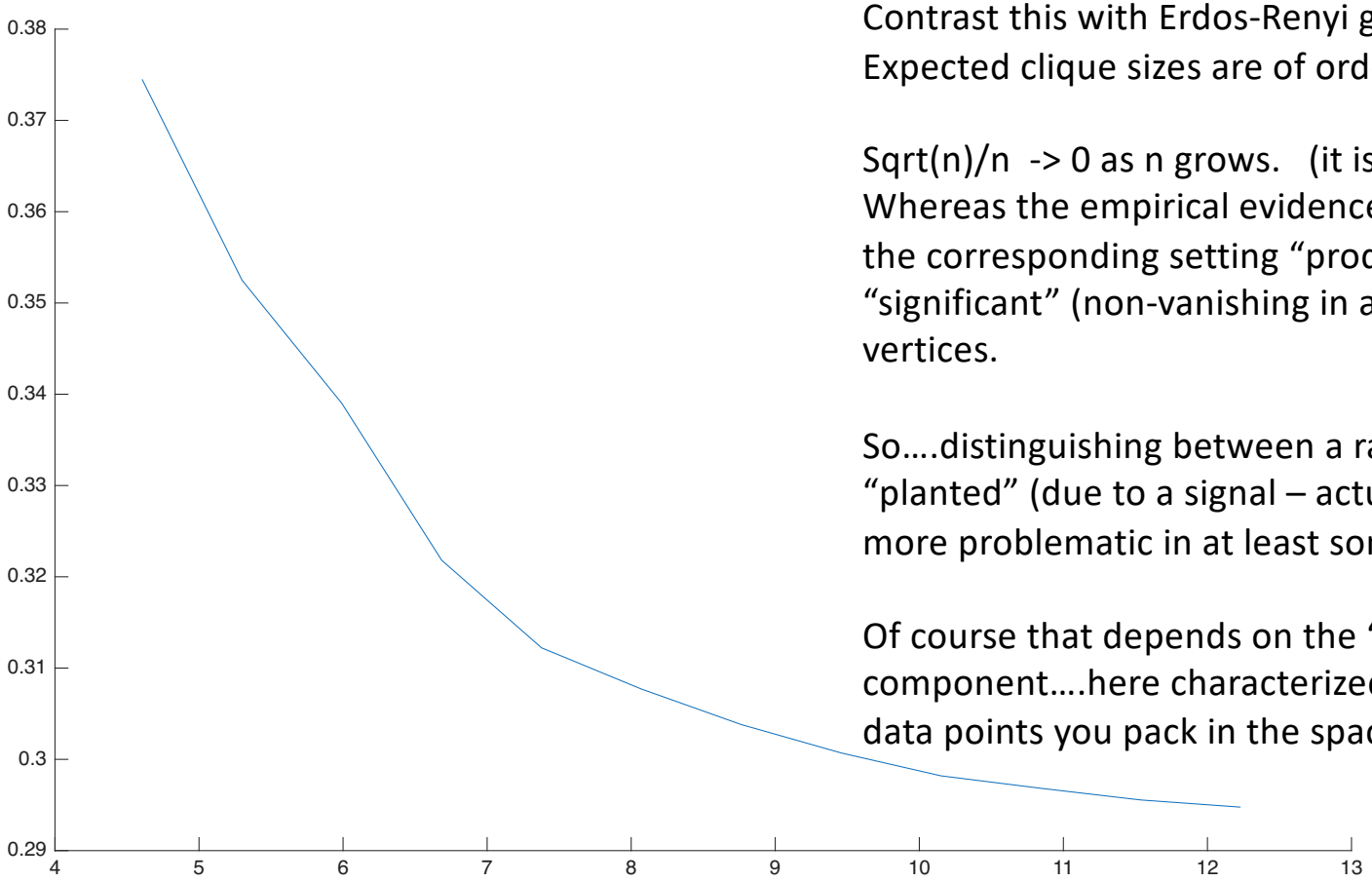


Interval graph.
Plot of average size (as a ratio with the number of vertices) of the maximum clique vs log number of vertices. Note: as n->infinity the graph may asymptote (weak evidence).

How done: threshold for “close” enough to declare edge is adjusted until p(edge)=0.5 (matching the usual setting in the Erdos-Renyi random graph).
20 random instances at each choice of n are generated and maximum clique found.

(Setting for empirical edge density deviation from 0.5 is 1e-10 – VERY close!)

Expected clique sizes in random geometric graphs



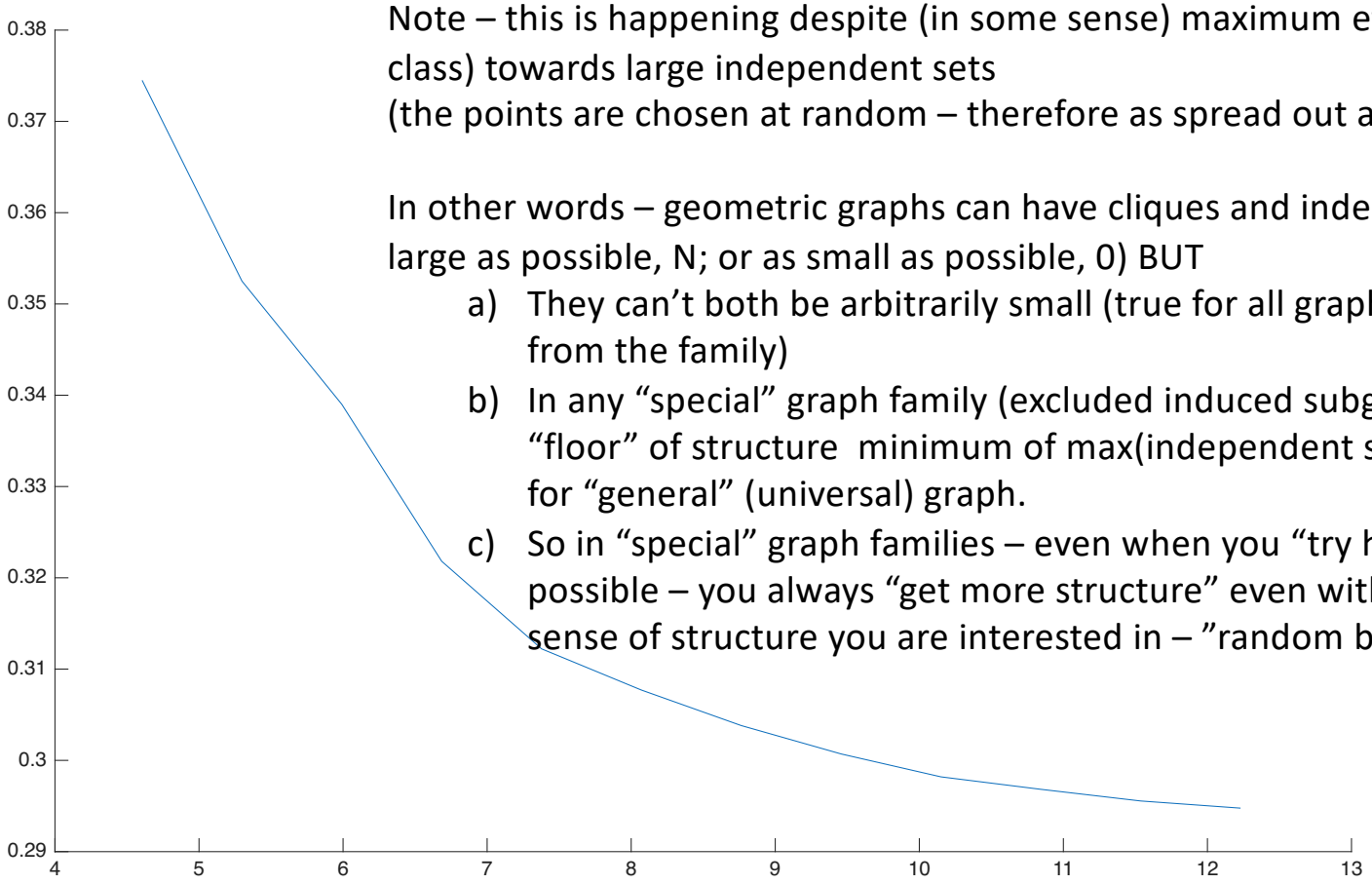
Contrast this with Erdos-Renyi graph p=0.5.
Expected clique sizes are of order sqrt(n).

Sqrt(n)/n -> 0 as n grows. (it is said to be o(1))
Whereas the empirical evidence is that for the random interval graph the corresponding setting "produces large" clique sizes. O(N) some "significant" (non-vanishing in any sense) fraction of the number of vertices.

So….distinguishing between a randomly occurring clique and one "planted" (due to a signal – actual structure in the data) could be even more problematic in at least some senses.

Of course that depends on the "strength" of the background random component….here characterized by p. (Roughly – how much random data points you pack in the space).

Expected clique sizes in random geometric graphs

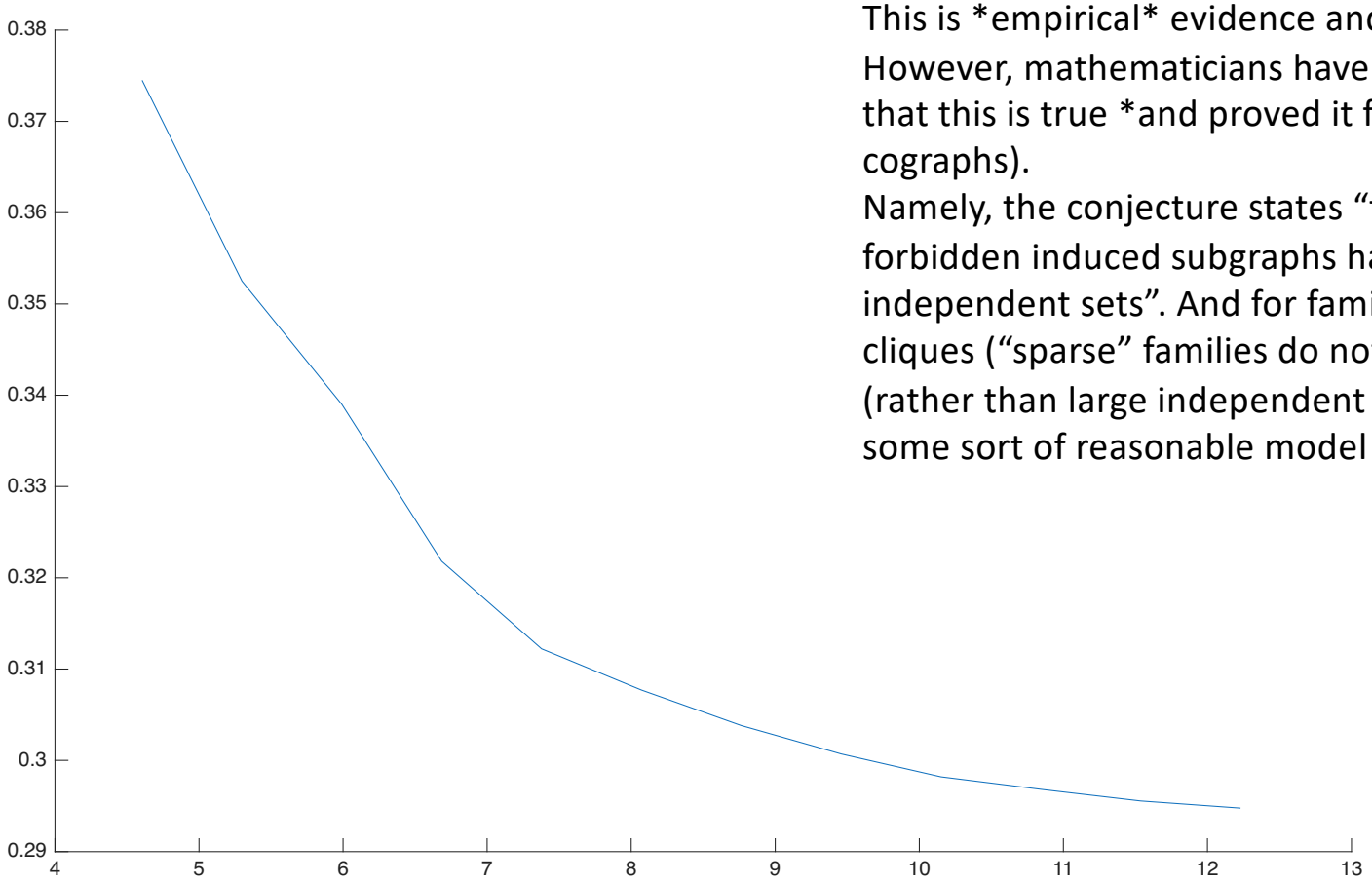


Note – this is happening despite (in some sense) maximum effort to bias (within the chosen graph class) towards large independent sets
(the points are chosen at random – therefore as spread out as possible – in some sense).

In other words – geometric graphs can have cliques and independent sets of "any sizes" (even as large as possible, N; or as small as possible, 0) BUT

a) They can't both be arbitrarily small (true for all graphs) in the same instance (chosen graph from the family)
b) In any "special" graph family (excluded induced subgraphs – special property) – the minimal "floor" of structure minimum of max(independent set,clique) – is biased to larger than it is for "general" (universal) graph.
c) So in "special" graph families – even when you "try hardest" to have as minimal structure as possible – you always "get more structure" even within situations biased to minimize the sense of structure you are interested in – "random background" to signal.

Expected clique sizes in random geometric graphs



This is *empirical* evidence and in some sense “weak”.
However, mathematicians have conjectured (Erdos-Hajnal conjecture) that this is true *and proved it for many special cases*, including cographs).
Namely, the conjecture states “that families of graphs defined by forbidden induced subgraphs have either large cliques or large independent sets”. And for families that “favour” in some sense – large cliques (“sparse” families do not) the conjecture says that large cliques (rather than large independent sets) are expected “on average” – under some sort of reasonable model of randomness for that class.

Planted dense subgraph

- A k-clique is the densest (most fully connected) graph on k-vertices.
- Maybe our problem only requires finding (or checking whether exists) a dense subgraph (for suitable definition of just how dense)?
- Reasonably obvious generalization/weakening of planted clique.
  - Model
    - 1. generate $G_{n,p}$
    - Select randomly k-vertices
    - For pairs of vertices in this k-subset, if not already connected (by the noise - $G_{n,p}$) then add an edge with suitable probability (to get the required density of edges)
- Unsurprisingly similar algorithms can be use to find/detect, and similar things can be said about the thresholds at which is it impossible (statistically), or hard, or easy…..
- See for example https://arxiv.org/abs/2004.13978 or do your own web search! ☺

Examples of what is known

http://proceedings.mlr.press/v125/brennan20a/brennan20a.pdf

B.6. Semirandom Planted Dense Subgraph and the Recovery Conjecture

In the planted dense subgraph model of single community recovery, the observation is a sample from $\mathcal{G}(n, k, P_1, P_0)$ which is formed by planting a random subgraph on k vertices from $\mathcal{G}(k, P_1)$ inside a copy of $\mathcal{G}(n, P_0)$, where $P_1 > P_0$ are allowed to vary with n and satisfy that $P_1 = O(P_0)$. Detection and recovery of the hidden community in this model have been studied extensively (Arias-Castro and Verzelen, 2014; Butucea and Ingster, 2013; Verzelen and Arias-Castro, 2015; Hajek et al., 2015; Chen and Xu, 2016; Hajek et al., 2016c; Montanari, 2015; Candogan and Chandrasekaran, 2018) and this model has emerged as a canonical example of a problem with a detection-recovery computational gap. While it is possible to efficiently detect the presence of a hidden subgraph of size $k = \tilde{\Omega}(\sqrt{n})$ if $(P_1 - P_0)^2/P_0(1 - P_0) = \tilde{\Omega}(n^2/k^4)$, the best known polynomial time algorithms to *recover* the subgraph require a higher signal at the Kesten-Stigum threshold of $(P_1 - P_0)^2/P_0(1 - P_0) = \tilde{\Omega}(n/k^2)$.

Examples of what is known

http://proceedings.mlr.press/v125/brennan20a/brennan20a.pdf

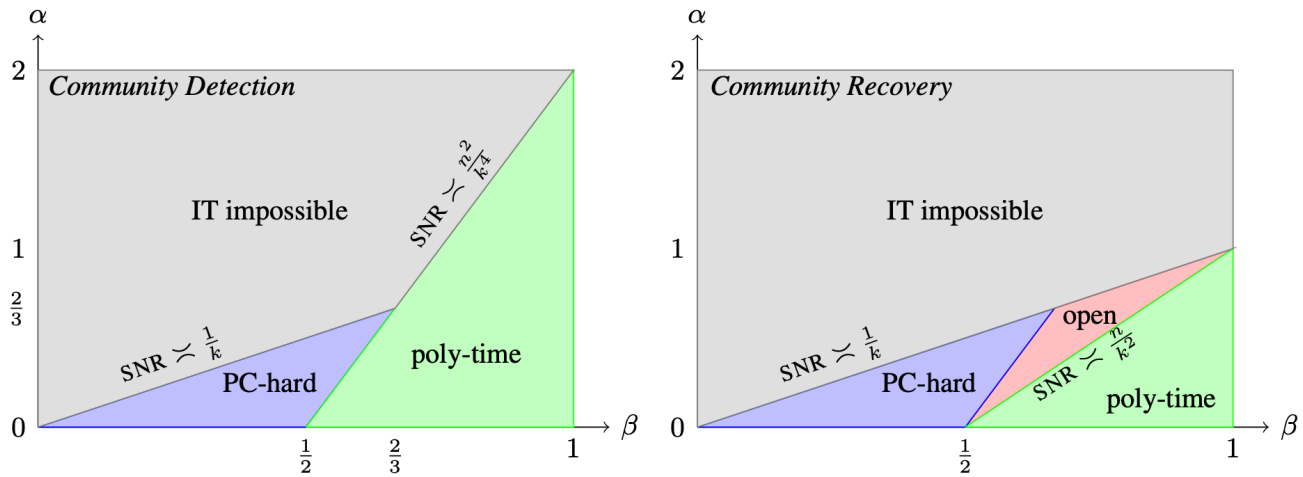


Figure 2: Prior computational and statistical barriers in the detection and recovery of a single hidden community from the PC conjecture (Hajek et al., 2015; Brennan et al., 2018, 2019a). The axes are parameterized by α and β where $\text{SNR} = \frac{(P_1-P_0)^2}{P_0(1-P_0)} = \tilde{\Theta}(n^{-\alpha})$ and $k = \tilde{\Theta}(n^{\beta})$. The red region is conjectured to be hard but no PC reductions showing this are known.

Whole “Zoo” of planted models…

- Get model of backround noise for that situation (Graph, Hypergraph, Simplicial Complex, etc.)
- Define way of “planting” a signal in that noise
- Look at ways to theoretically determine what is possible (for detection, find solution)
- Look at ways to efficiently get a good solution in practice
- Look at situations that this model (and associated algorithms and theory) might be useful (explain, give good ways to solve etc.).