Detecting Structure in Graphs

Gautam Rayaprolu

McGill University

gautam.rayaprolu@mail.mcgill.ca

December 13, 2017

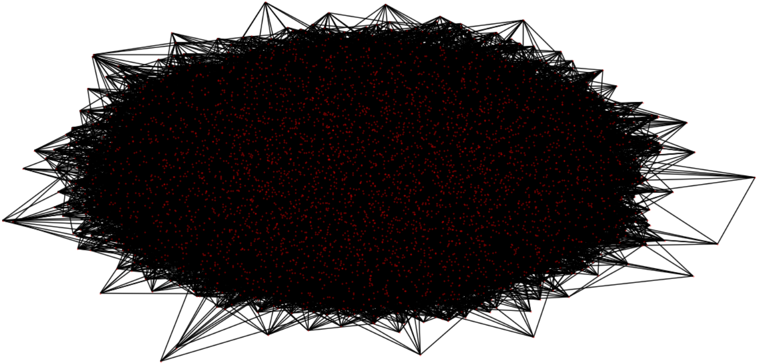
Introduction

- An Erdos Renyi Random Graph $G(n, p)$ is a graph on n vertices where each of the $\binom{n}{2}$ edges exists independently with probability p
- The largest clique on this graph is $O(log(n))$ with high probability
- As part of this project I showed that the largest hypercube on this graph is $O(log(log(n)))$
- You can use spectral methods to detect a planted clique of size atleast $O(\sqrt{n})$ on this graph

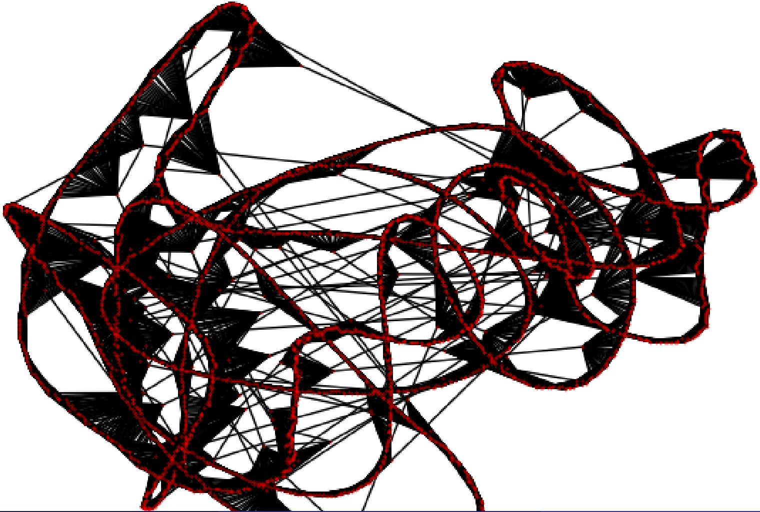
Random Graphs as a model of networks

- ER graphs have a low clustering coefficient and their degrees are distributed as a Poisson Distribution.
- Many real world graphs have been shown to a power law degree distribution. i.e. $P(k) \approx k^{-\gamma}$
- The clustering coefficient shows how close neighbours are to being a clique
- Many models have been suggested, most notably the watts-strogatz small world graph

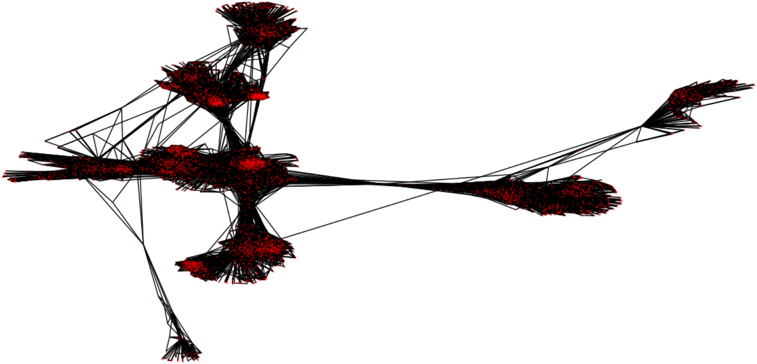
Erdos Renyi Graph



Watts Strogatz Small World Graph



Facebook Network Graph



Kronecker Graphs

- The Kronecker product of 2 matrices $A \otimes B$ is the process where each element a_{ij} of A is replaced by the matrix $a_{ij}B$
- The adjacency matrix of a Kronecker graph is constructed by taking an 'initiator' adjacency matrix P and repeatedly applying the Kronecker Product with itself
- We could alternative take values between 0 and 1 and interpret them as probabilities of the edge occurring

The KronFit Algorithm

- This algorithm was described by [3] in 2010
- Given a graph G with $N = N_1^k$ nodes we wish the compute a MLE estimator for the stochastic initiator matrix P of size N_1
- $log(l(\Theta)) = log(P(G|\Theta) = log \sum_{\sigma} P(G|\Theta, \sigma)P(\sigma)$ where sigma represents a permutation of labellings
- Use Metropolis sampling to get an estimate of the sum , which has $N!$ terms.
- Note that there are several global maxima corresponding to different permutations of the Initiator Matrix

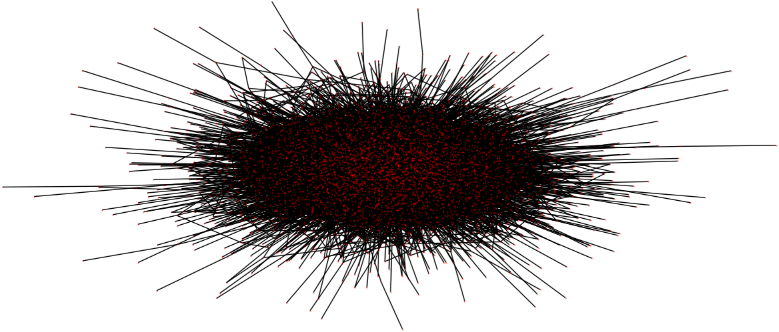
Optimizations

- Naively Calculating the likelihood takes $O(N^2)$ time
- For the metropolis sampling algorithm we uniformly pick 2 indices and swap the nodes at those indices with probability $\frac{P(\sigma^i|G,\Theta)}{P(\sigma^{i-1}|G,\Theta)}$
- This difference only takes $O(N)$ to calculate
- The adjacency matrix of the training graph is sparse, so we can also estimate the full likelihood in $O(|E|)$ time which is roughly linear for real world graphs

Experimental Results

- I implemented the Kronfit algorithm using scipy's TNC minimizer and used it to fit 1024 nodes on the facebook graph.
- Due to performance constraints I had to restrict the number of permutation samples to 100
- The average clustering coeffecient of the facebook graph was around 0.7 but that of the kronecker graph was around 0.03 which is very similar to a $G(n, p)$
- The log likehood's on the training data were of the order of 10^{-4} and did not vary much with different training sets

Learned Kronecker Graph



Learned Kronecker Graph

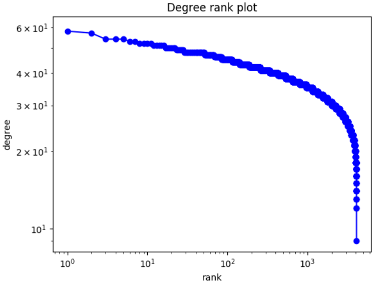


Figure: Kronecker Degree Distribution

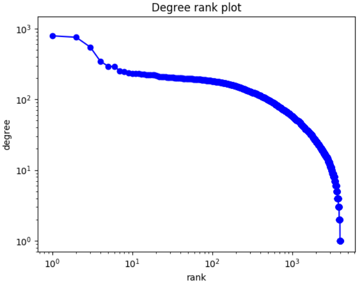


Figure: Facebook Degree Distribution

Issues and Improvements

- [3] says that there is a narrow band of parameters where kronecker graphs have the interesting properties of real world graphs
- The largest graph simulated was only of size 2^{12} which may not be enough to see the clustering behaviour
- Might need a larger number of samples from the permutation space to get a better estimate of the likelihood
- Consider a bayesian approach instead of using maximum likelihood, however this might be prohibitively expensive

Conclusions

- Spectral methods can be used to detect planted cliques of size $O(\sqrt{n})$ on a $G(n, p)$
- Real world networks are more clustered than random graphs
- Kronecker Matrices can be used to model real world graph structures
- Training Kronecker Matrices can be quite expensive and may not work on small training sets.

References

- Berthet Quentin, Rigollet Philippe, Computational Lower Bounds for Sparse PCA,(2013).
- Cook Alexis B, Miller Benjamin A, Planted clique detection below the noise floor using low-rank sparse PCA,(2015).
- Jon Kleinberg, Jure Leskovec, Deepayan Chakrabarti, Christos Faloutsos, Zoubin Ghahramani Kronecker Graphs: An Approach to Modelling Networks,(2010).
- Jure Leskovec, Andrej Krevl SNAP Datasets: Stanford Large Network Dataset Collection

The End