ECS 253 / MAE 253, Network Theory and Applications

Spring 2018

Advanced Problem Set # 2, Due May 3

Problem 1: Clustering coefficient of the Erdos-Renyi random graph

Consider the random graph $G(n, p)$ with average degree denoted by m.

a) Show that in the limit of large n the expected number of triangles in the network is $\frac{1}{6}m^3$, where m is the mean degree. In other words, the number of triangles depends on the density of edges, not the size of the network. (Hint 1: Recall that edges are all added independently and for two independent events X and Y, the joint probability $p(X, Y) = p(X)p(Y)$. Hint 2: Think of how many distinct triangles can exist among n nodes.)

b) Using arguments similar to part (a) show that the expected number of connected triples is $\frac{1}{2}nm^2$. Note for a particular choice of 3 vertices (for instance, labeled "A", "B" and "C"), there are three ways to form a connected triple: (i) A "V" with "A" as the central vertex; (ii) A "V" with "B" as the central vertex; (iii) A "V" with "C" as the central vertex.

c) The average clustering coefficient for a network is the expected number of triangles divided by the expected number of connected triples. Write an expression for the average clustering coefficient, $\mathcal{C}$, for $G(n, p)$ as a function of n and m.

d) Random graphs are typically analyzed in the $n \to \infty$ limit. What is the value of clustering coefficient $\mathcal{C}$ in the limit $n \to \infty$ while mean degree m is held constant?

Problem 2 Fitting Power Law Distributions

In this exercise, you will understand more about fitting power laws to data. First you will generate a synthetic data set, with 100,000 elements sampled from a power law distribution with exponent $\gamma = 2.5$. Then you will analyze the data.

a) The "transformation method" is a way to generate a set of random variables

sampled from a desired distribution. First generate 100,000 random numbers r on the unit interval $0 \leq r < 1$. Now apply the following transformation, $x = (1-r)^{-1/(\gamma-1)}$. x will be a random power law distributed real number in the range $1 \leq x < \infty$. Plot a histogram of size of x versus frequency, first on a linear-linear scale, then on a log-log scale. (The size will be the horizontal axes; frequency of occurrence will be the vertical.)

b) The slope of the plot in part (a) is the power law exponent. But it is hard to estimate from the data as you should find that the right "tail" is noisy. Do you think that noise in the tail leads to an overestimate or underestimate of the exponent of the power law? (Explain.)

c) One way to eliminate noise in the tail is to bin the horizontal axes into exponentially wider bins. The width of the bins is hence 1, 2, 4, 8, 16, etc. This means bin one encompasses the range $1 \leq x < 2$; bin 2 encompasses the range $2 \leq x < 4$; bin 3 encompasses the range $4 \leq x < 8$; bin 4 encompasses the range $8 \leq x < 16$; etc. Bin the data you generated in part (a) accordingly and plot the result on a log-log scale.

d) Estimate the slope of the plot in part (c) using a linear least-squares fit.

e) A different way to reduce the noise in the tail, rather than exponential binning, is to consider the cumulative distribution. Given a Probability Density Function $P(x)$, the corresponding Complimentary Cumulative Density Function $C(x)$ is the probability that a random variable drawn from the distribution has a value greater than x:

$$C(x) = \int_x^\infty P(x)dx.$$

Show that if $P(x)$ is a power law (properly normalized of course), then $C(x)$ is also a power law and derive the mathematical expression for $C(x)$.

f) Plot $C(x)$ versus x on a log-log scale and now estimate the exponent γ using linear least squares.

g) Which method is superior: exponential binning in part (c) or the cumulative distribution in part (f)?

Problem 3: Resolution limit of modularity function

Modularity aims to quantify the quality of a community partition of a network by comparing the number of links found inside communities compared to what is expected by chance, and it can be written in the form

$$Q = \sum_{c=1}^{C} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right], \tag{1}$$

where C is the number of communities, l_c is the number of links in the cth community, d_c is the total degree of nodes in the cth community and L is the total number of links in the network.

Consider a network consisting of a ring of n cliques of size m, such that each neighboring clique is connected by a single link and n is an even number. A clique is a fully connected subgraph; therefore a clique of m nodes has $m(m-1)/2$ links. This network has an intuitive community structure: each clique corresponds to a community.

3.a) Determine the modularity Q_{single} of the intuitive partition and the modularity Q_{pairs} of the partition in which pairs of neighboring cliques are merged into a single community. Express Q in terms of the number of cliques n and the total number of links $L = nm(m-1)/2 + n$.

3.b) Show that only for $n < \sqrt{2L}$ will the modularity maximum corresponds to the intuitively correct community division.

3.c) Discuss the consequences of the above inequality for community detection.