CS485 Spring 2007
Homework 1

Due Date: Feb 2 2007

Pick a real world dataset containing a large network, with at least 100 nodes and 1000 edges (e.g. look at `http://www.cosin.org/extra/data/data.html`). Treat it as an undirected graph. Please include the reference.

1. Plot a node degree distribution for the dataset.

2. Find sizes of its components.

3. Compute the clustering coefficient defined as:

$$C = \frac{\text{number of triangles in the network}}{\text{number of connected triples of vertices}}$$

where a "triangle" is a fully connected set of 3 vertices, and a "connected triplet" is a connected set of 3 vertices (i.e. one edge may or may not be missing). The clustering coefficient tells how well transitivity is preserved in the network.

4. Compare to $G(n, p)$, analytically or empirically. Let n be equal to the number of nodes in your chosen dataset, and choose p such that the expected number of edges also matches. Then:

   (a) Plot the degree distribution for such $G(n, p)$ graph.

   (b) Find sizes of components of the $G(n, p)$ graph.

   (c) Compute the clustering coefficient for the $G(n, p)$ graph.

   (d) Briefly describe which quantities differ between the original dataset and the $G(n, p)$ graph, and what it might mean.