MAE 298, Network Theory and Applications
Winter 2008
Problem Set # 2, Due Feb 11

Analysis of a real-world network

For this problem you must find a data set of a real-world network. It could be the call-graph in a software program, a recommendation network of books constructed via amazon.com, a flight network for an airline, a collaboration network of scientists or movie actors, a protein-interaction/gene-interaction network, a piece of the Amtrak rail network, etc. When choosing this data set, keep in mind the class project (ideally, try and make these overlap). The network should have at least 200 nodes.

a) Describe your data set and where/how you obtained it. Is this a directed or undirected graph? Are there several components, or is it all one connected component?

b) How many nodes and edges are present? What is the average degree? (If it is a directed graph give values for both average in- and out-degree.)

c) Plot the degree distribution (again, if directed, plot both in- and out-degree distributions). Describe qualitatively the shape of the distribution (Is this distribution "heavy-tailed"? Does it decay exponentially? etc.)

d) Visualize the network (i.e., draw a picture of this network using any means you like – graph drawing software, pencil and paper, etc). Try to use color or size to display interesting attributes of your data (degree, age, high-clustering, etc). You may want to label the nodes with their identities (or not).

(f) For this part, assume all links are bi-drectional (i.e., if you have a directed graph, assume all edges go both ways). Determine the clustering coefficient of each vertex in the network and present the results as a plot (*i.e.*, plot the clustering coefficient distribution). What is the average clustering coefficient for the network?

The clustering coefficient C_i for a vertex i is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them:

C_i = (# number of links between neighbors of i, excluding i) / (total number of links that could exist between neighbors).

In more formal terms: $C_i = 2|e_{jk}|/k_i(k_i - 1)$,

where $|e_{jk}|$ is the total number of links between all nodes j and k that are connected to node i (NOT including i), and k_i is the degree of node i. Note the factor of 2 comes from the fact we are considering undirected edges, so the total number of edges that could exist between neighbors of i is $k_i(k_i - 1)/2$. If a node is disconnected (*i.e.*, $k_i = 0$), then assume $C_i = 0$ for that node.

For more information on the clustering coefficient, see for instance,
`http://en.wikipedia.org/wiki/Clustering_coefficient`