

Due October 8 2009 in class

CS322: Network Analysis

Problem Set 1 - Fall 2009

If you have any questions regarding the problems set, send an email to the course assistant: simlac@stanford.edu. Please write the name of your collaborators on your problem set. You can use existing software or code to compute the answers, you don't have to submit the source code.

The goal of this problem set is to practice implementing some basic network analysis techniques on a moderate-sized network dataset. You will be working on a co-authorship network, from the e-print arXiv that covers scientific collaborations between authors papers submitted to High Energy Physics – Theory category.

This dataset has entries with the following comma-separated format for each paper:

`<year>,<volume>,<journal>, <author1>& ... &<authorN>, <title>`

You can find the data at: <http://snap.stanford.edu/na09/hep-th.txt>

So, a sample line from the data will be as,

```
1999, 40, J.Math.Phys., Clark & Alamos & Brown, PT-Symmetric Mechanics
2000, 41, J.Math.Phys., Alamos & Davis, QU-Symmetric Quantum Mechanics
```

From the bibliography, you should construct the co-authorship network as follows.

- There must be one node for each author.
- There should be an undirected edge between two nodes if and only if they are coauthors of at least one publication. (Even if a pair of authors has coauthored on multiple papers, there should still be a single edge joining them.)

Our example above would create a graph with nodes C, A, B, D corresponding to Clark, Alamos, Brown and Davis, and edges A–B, A–C, B–C and A–D.

A broad community of researchers is interested in scientific co-authorship networks precisely because they form detailed, pre-digested snapshots of a rich form of social interaction that unfolds over a long period of time. By using on-line bibliographic records, one can often track the patterns of collaboration within a field across a century or more, and thereby attempt to extrapolate how the social structure of collaboration may work across a range of harder to measure settings as well.

The Problems

Problem 1.1

Recall that degree of a node is the number of edges the node has to the other nodes. We will first look at how the degrees of the nodes in the network are distributed. Let k_j be the number of the nodes with degree exactly j .

Draw a scatterplot of the ordered pairs (j, k_j) for j such that $j > 0$ and $k_j > 0$. Also draw a scatterplot of the ordered pairs $(\log j, \log k_j)$.

What do you notice?

Problem 1.2

(a) Find the number of nodes in the largest connected component. How many nodes are there in the network overall? Can we think of this network as having a *giant* component?

(b) Let n_i be the number of connected components of size i . Draw a scatterplot of the ordered pairs $(\log i, \log n_i)$ for those i such that $i > 0$ and $n_i > 0$.

(c) We start by fixing a node n from the network. Let n be the author called *Ambjorn*. Now we will perform Breadth First Search from this node and count how many new nodes we encounter as we expand away from n .

Let r_j be the number of the nodes that can reach n by a path of length j , i.e., r_1 is equal to the number of collaborators (degree) of n , r_2 is the number of collaborators of collaborators of n , and so on.

Draw a scatterplot of the ordered pairs (j, r_j) . What can you say about the expansion α around n as a function of j ?

Problem 1.3

(a) Recall that the local clustering coefficient for undirected graph $G(V, E)$ is defined as $C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)}$: $v_j, v_k \in N_i$, $e_{jk} \in E$ and k_i is the degree of node i , N_i is a set of network neighbors of node i .

Find the average clustering coefficient, $C = \frac{1}{|V|} \sum_{i \in V} C_i$, for the co-authorship network.

(b) Next we will create a random network with the same degree distribution as the original network. Let's call this graph a rewired network. To generate a rewired network with the same degree distribution as the original co-authorship network we use the *configuration model*: For each node n_i in the random network, create d_i spokes (half-edges), where d_i is the degree of the node i in the original network. Now randomly connect endpoints of spokes. One possible way to do this is create a random ordering of the spokes and then simply connect spokes that are adjacent in the ordering – connect spokes at positions $2i + 1$ and

$2i$, for $i = 1 \dots$. Notice that this procedure will likely create a multi-graph – a graph with multiple edges (pairs of connected spokes) between a pair of nodes, and self-loops on nodes.

Now, simplify the resulting multi-graph by collapsing multiple edges between a pair of nodes into a single edge and remove self loops.

On the same figure superimpose the degree distributions of the real and random networks on log-log scales (as in second part of Problem 1).

Find the average clustering coefficient for the random network. How does it compare to the real network?

Problem 1.4

(a) For each edge e_{ij} in the coauthorship network, let w_{ij} be the number of shared neighbors between i and j . Order the edges in decreasing order of w_{ij} and delete them from the network in this particular order. Plot the size of the largest connected versus percentage of the deleted edges. (since it will be computationally challenging to compute connected components after deletion of each single edge, delete edges in batches of X (say $X = |E|/100$, where $|E|$ is the total number of edges in the network). First delete X edges, compute components, delete additional X edges, compute components and so on.

How does the size of the largest component change?

(b) For each edge e_{ij} in the co-authorship network, let z_{ij} be the number of papers where i and j were co-authors (appeared together on the list of authors). Delete the edges from the network in decreasing order of z_{ij} . Plot the size of the largest connected versus percentage of the deleted edges.

How does the size of the largest component change?

Problem 1.5

When we think about a single aggregate measure to summarize the distances between the nodes in a given graph, there are two natural quantities that come to mind. One is the diameter, which we define to be the maximum distance between any pair of nodes in the graph. Another is the average distance, which — as the term suggests — is the average distance over all pairs of nodes in the graph. In many graphs, these two quantities are close to each other in value. But there are graphs where they can be very different.

(a) Describe an example of a graph where the diameter is more than three times as large as the average distance.

(b) Describe how you could extend your construction to produce graphs in which the diameter exceeds the average distance by as large a factor as you'd like. (That is, for every number c , can you produce a graph in which the diameter is more than c times as large as the average distance?)