PROBLEM SET 6 – Week 8

We will be using the ACL (Association for Computational Linguistics) anthology, composed by Mark Joseph & Drago Radev:
http://tangra.si.umich.edu/clair/anthology/

We will just use two networks derived from this painstakingly put together data (currently maintained by Bryan Gibson, who was kind enough to share it with us). They are attached as a tarball (for those of you who love data), but I have also provided Pajek files you can use directly instead. The files are the weighted co-authorship network CoAuthorshipNetwork.net (how many papers two people co-authored) and the weighted citation network (how many papers of co-author A cite papers of co-author B) AuthorCitationNetwork.net. In each of these networks, an author is only included if they have at least 10 papers in the ACL dataset.

Your tasks are the following:

1. Load the two networks. They should both have the same number of vertices: 1559. Compute the density of both: Info>Network>General. Which one has the higher density? Why could this be the case?

2. Compute the clustering coefficient of both. Net > Vector > Clustering Coefficients > CC1. And then Info > Vector. Interpret the difference. (do this on an undirected version of the citation network Net>Transform>Arcs->Edges), but for the rest of the assignment use the directed version).

3. In the co-authorship network, compute the degree, closeness, and betweenness of each author (refer to previous assignments if you don't remember how to do this). BTW, here finally, is a (rather complicated) way to sort the vertices by their centralities:
    a. Apply the centrality measure so that you have a vector of values for each vertex
    b. With that vector selected in the drop-down menu, select Vector>Make Permuation
    c. With that permuation selected in the permutation drop-down menu, select Operations>Reoder>Network. This will create a new network
    d. Re-calculate the centrality for the ordered network. Click on the 'edit' button next to the new centrality vector. Now the vertices are ordered from least to most central, so scroll to the bottom to get the top 5.
    2. Just have a look at how these. You don't need to include anything in the assignment at this point, you'll be correlating the co-authorship measures with citation measures.

4. In the citation network, compute the indegree and proximity prestige of each author.

- for proximity prestige, you are getting the input domain of the vertex (everyone who cites that person directly or indirectly), and dividing by the average distance to those vertices. You will use Net>Partitions>Domain>Input
- this will produce two things: a partition with the size of the input domain of each vertex, and a vector of average distances to vertices in the input domain
- create a vector from the input domain size partition Partition > Make Vector
- then select the second drop down menu for the vector to be the average distance
- select Vectors > Divide First by Second. This will be the input prestige of each vertex

Look for the highest correlation in a centrality measure for the co-authorship network and prestige (indegree or proximity prestige) for the citation network. Please give all pairwise correlations. Which two measures are the most correlated? Interpret. Caution! Make sure that you are using the centrality/proximity measures with the original vertex ordering, and then find the correlations.

- Select a centrality measure as the first vector in the vector drop down menu
- There is a second drop-down menu right below it, select a prestige measure
- Select Vectors>Info. This will give you the Pearson correlation coefficient
- Make sure that the measures were applied to the original ordering of the vertices, so that you are correlating values for the same vertex

5. Finally, load the file CitationNetWoCoauthors.net. This is the citation network with citations between co-authors removed (the reason being that an author may be citing their own paper and in the process citing their co-authors). We're trying to get a more "unbiased" prestige measure where we don't take direct citations by co-authors into account. It's not perfect. Compare the density of this network with the complete author citation network. What percentage of the citation edges was from co-authors?