PROBLEM SET 5 – Week 7

1. The Girvan-Newman betweenness clustering algorithm (50pts)

Select a network of up to 200 nodes, preferably one where you suspect there might be interesting structure. For 90% credit (45pts), you can use the poliblogrecip.gdf file, which I had demoed in lecture, but for max credit, you need to construct a network of your own (based on real data, or an artificial topology of your choosing). Your network should be undirected. Load the data file into Guess and run the script betweennessclustering.py. The GUESS tool has a button, which will remove the edge with the highest betweenness (marked red) at each click, and another button to keep removing edges until the next community is broken off (or all edges are removed). You need to decide when to stop removing edges (fyi, this is just an educational tool implementing a slow algorithm and it cannot cluster big data sets). Please answer the following:

- Does the algorithm allow you to identify underlying communities in the network?

- Is the removal of a leaf node a good stopping criterion for the algorithm? That is, if you stop removing edges before a node with degree one is removed, does this reveal informative community structure, or would you have wanted to stop earlier (or keep going after removing a leaf node).

- Which nodes are not very embedded in their communities? How does the algorithm reveal this fact?

- Are any nodes 'misclassified' by being placed in a cluster where you think they may not belong according to a node attribute? How can such a misclassification be informative?

- Turn in one image that shows community structure in this network *I*.

2. Hierarchical clustering (50 pts)

Use Pajek to hierarchically cluster the very same network (again 45 pts at maximum if you use the political blog network, 50pts if you use your own network).

- Visualize the permuted adjacency matrix (*I*)

- Visualize the dendrogram. (*I*)

- Compare the hierarchical clustering with the betweenness-based clustering at a point where both methods produce the same number of communities (e.g. both have 2 communities). Are there any differences in the groupings? Which nodes were classified in different communities? Use your knowledge of what the network represents to discuss why some nodes were placed in different communities by the two different techniques.