15-396A Science of the Web
Assignment 1
Due September 15, 2011 @ 3:00pm

Signed Graphs (10 points each)

1. Prove that if a signed complete graph has a clustering, then this clustering is unique.
2. Assume **G** is a signed graph that has at least one clustering. State an algorithm to find a clustering in **G**. Your algorithm must run in time polynomial in the number of nodes of G.
3. Using part 2, prove that a signed graph has a clustering if and only if it contains no cycles with exactly one negative edge.

Normalized Betweenness Centrality (10 points each)

1. We saw in class that the star network contains a vertex with normalized betweenness centrality (NBC) of 1. Are there any other graphs that contain a vertex with NBC of 1? Make sure to justify your answer--- either give a graph and show the calculations of NBC, or prove that no such graph exists.
2. Let $\mathbf{T_n}$ be the complete binary tree of height **n**. What is the NBC of the root node? What about for nodes at depth **k** in the tree (**k** = 0 is the root node, nodes with **k** = **n** are the leaf nodes)?
3. Let **T** be an arbitrary tree. Given a vertex **v**, what is a simple expression for the **betweenness centrality** of **v**? Hint: consider rooting **T** at **v** and examine the resulting subtrees $\mathbf{T_1}$, $\mathbf{T_2}$, …, $\mathbf{T_k}$.

Clustering Coefficient and NBC (10 points)

You saw how the clustering coefficient of a vertex measures how densely connected its neighborhood is. It turns out that having a highly clustered neighborhood prevents a vertex from having a high normalized betweenness centrality. Show that for any vertex **v** in a connected graph **G** with **n** nodes,

$$C_{NB}(v) \le 1 - [CC(v)\ deg(v)\ (deg(v)-1) / ((N-1)(N-2))].$$

Furthermore, show that this bound is tight by exhibiting a graph in which the inequality above is an equality for some vertex in the graph.

Graph Exploration with Gephi (30 Points)

In this question you are going to explore graphs using Gephi, an open source graph exploration and analysis tool.

Getting Gephi and some data:

1. Download Gephi from http://gephi.org. It runs on Windows, Mac, and Linux, so this shouldn't be a problem for anyone in the class. Let us know if you have trouble installing the program.
2. Go to Mark Newman's website (http://www-personal.umich.edu/~mejn/netdata/) and download the Zachary Karate Club graph and "books about US politics" graph. The graph files are the *.gml files after you unzip them. They can be easily opened in Gephi.

Task 1: Calculating network properties (20 points)

We're going to examine some properties of these two graphs. Be careful when opening the graph files, as you can choose between interpreting them as either a directed or undirected graph. Make sure to load them as **undirected** graphs (as well as when doing subsequence analyses).

For the karate club graph and graph of political books, answer the following questions:

1. Which node (or nodes) has the highest degree, and what is this degree? Give the node ID and label, if there is one. What's the average degree in the graph? What is the most common vertex degree in the graph?
2. How many triangles are in the graph? What is the average clustering coefficient in the graph? After you calculate the number of triangles, plot the number of triangles in which a node participates versus the degree of the node. What do you notice about the trend? How does this compare to a model in which nodes are connected at random to each other?
3. Calculate the betweenness and closeness centralities (making sure to normalize). Which four nodes in each network have the greatest closeness centrality, and what is the NBC of these nodes? It's sufficient to provide two tables with node IDs and NBC values.

Task 2: Create a visual artifact (10 points)

This is your chance to get creative! Using any data set that you can find or generate, create a visual artifact of the network pick. Play around with different layouts (you can even manually manipulate the layout), node size, node color, edge color, labels, etc. Submit a PDF (instructions for how to do so will be announced later) of your artifact along with a description of the network you are visualizing and the way you picked different node properties. We'll show off some of these visualizations in class.

Stay away from me! (20 BONUS Points)

Nadnerbs are very antisocial creatures and dislike being near other Nadnerbs in their social graph. In fact, it's very important that each Nadnerb gets as far away as possible from other Nadnerbs. However, Nadnerbs are not completely anitsocial and want to form a connected social network.

Given a connected social graph $G = (V,E)$, we say that a Nadnerb v is ε-sad if $|\{u \text{ in } V \text{ such that } d(u,v) < \text{diam}(G)\}| > \varepsilon|G|$. That is to say, v is ε-sad if it's not as far away as possible from more than an ε fraction of nodes in the graph.

You get to determine the way in which Nadnerbs become friends (the friend recommendation system of the Nadnerb world). You are given an arbitrary $\varepsilon > 0$. As the population grows, you want the fraction of Nadnerbs who are ε-sad to go to zero. Furthermore, it's important that as the population grows, the distance between Nadnerbs increases (otherwise things would get rather cramped). Can you come up with a sequence $\{G_n = (V_n, E_n)\}$ of social graphs that satisfies these conditions?

Formally, your sequence should satisfy the following properties:

1. Diameter(G_n) → ∞ as n → ∞
2. |{v in V_n such that v is ε-sad in G_n}| / |V_n| → 0 as n → ∞

Make sure to prove that your construction is correct!