Inference, Models and Simulation for Complex Systems
CSCI 7000-003, Fall 2010
Prof. Aaron Clauset
Problem Set 4, due 10/27

1. **Structural balance**

   In social network theory, the notion of *structural balance* says that certain sets of social relationships are stable while others are not. In its conventional form, structural balance only applies to fully connected or complete simple networks (no self-loops or multi-edges) whose edges have weights $w_{ij} \in \{-,+\}$ and to the stability of triangles. The idea is that relationships between three people—triangles—with either all positive ties $\{+,+,+\}$ or with one positive tie $\{+,-,-\}$ are stable, while those with either all negative ties $\{-,-,-\}$ or with one negative tie $\{-,+,+\}$ are unstable and will convert into one of the two stable forms.

   Consider the following signed graph:

$$\begin{aligned} G = \{&(1,2,+),(1,3,-),(1,4,+),(1,5,-),(2,3,-),\\ &(2,4,+),(2,5,-),(3,4,-),(3,5,+),(4,5,+)\} \ . \end{aligned}$$

   (a) Is G structurally balanced? If so, why? If not, identify a minimal set of edges that need to be flipped in sign to make it balanced.

   (b) Structural balance is not defined for networks whose weights are real valued $w_{ij} \in (-\infty,\infty)$, which makes it somewhat unrealistic. After all, some social relationships are stronger than others. Briefly describe what changes would need to be made in order to generalize structural balance theory to this case.

2. **Modular graphs**

   (a) Consider a "line graph" consisting of n vertices where each vertex connects to exactly two others, in a line, except for the two end points, which have degree one. That is, a network of n vertices and $m = n-1$ edges whose diameter is m. Show that if we divide this network into any two contiguous groups, such that one group has r connected vertices and the other has $n-r$, the modularity takes the value

$$Q = \frac{3 - 4n + 4rn - 4r^2}{2(n-1)^2} \ .$$

(b) Again considering the line graph, show that when n is even, the optimal division, in terms of modularity Q, is the division that splits the network exactly down the middle, into two parts of equal size.

(c) Now consider a "ring graph" made of k cliques, each containing c vertices, arranged in circle, where each clique is connected by one edge to each its two neighbors. Let each edge have unit weight; let k be an even number; let P_1 be a partition with k groups where each group contains exactly one of the k cliques; and let P_2 be a partition with $k/2$ groups where each group contains one pair of adjacent cliques.

Derive an expression for the difference in modularity scores $\Delta Q = Q_2 - Q_1$ and show that this difference is positive whenever $k > 2\left[\binom{c}{2} + 1\right]$. This is the so-called *resolution limit* of the modularity function, which says that at some size of the network, merging smaller module-like structures—here, the cliques—becomes more favorable under the modularity function than keeping them separate. Thus, finding the partition that maximizes Q will miss these small structures.
(Hint: for each partition, begin by writing expressions for e_i and d_i for a group.)

(d) Again consider the ring graph, but now connect each clique to every other clique with edges of weight $2/(k-1)$. Derive an expression for the difference in modularity scores $\Delta Q = Q_2 - Q_1$. (To generalize Q to a weighted graph, let e_i be the total edge weight within group i, let m be the total edge weight in the graph, and let d_i be the total weight of edges with an endpoint in group i.) For what value of k is this expression positive? Briefly discuss what this result means for the resolution limit.
(Hint: for each of the two partitions, again start with expressions for e_i and d_i.)

3. **Data analysis**

Download the "Political blogs", "Coauthorships in network science" and "Internet" data sets from Mark Newman's website:
`http://www-personal.umich.edu/~mejn/netdata/`
These networks are in the GML format, which is a kind of markup language for graph structures. (If necessary, convert each network to be undirected, and throw out any self-loops or multiedges.) For each data set, provide the following:

- A good visualization of the network. The Java program yEd (linked from the class webpage) can understand the GML format and has a version of the Fruchterman-Reingold spring embedder under the submenu Layout → Organic.

- A table giving the names of top ten vertices, in order of their closeness centrality (as defined in Lecture 8), each with their closeness score and their degree.
- A figure showing the vertex closeness centrality as a function of vertex degree (this may look better on log-log axes). Overlay on this data a line showing the same for a random graph with the same degree sequence as the empirical network (à la the configuration model, in Lecture 9), but averaged over several instances.
- A brief discussion of your results.

4. **(optional) Spatial networks**

Spatial networks are those whose vertices are embedded in some metric space, e.g., $\mathbf{z}_i \in \mathbb{R}^N$ where typically $N = 2$. In many spatial networks, the points are fixed in their location and our task is to build a network that both connects them and minimizes some kind of cost function over the properties of the network. For instance, in an airline network, the locations of airports are fixed but we can choose which airports to connect by flights. Cost functions are typically some tradeoff of the total length of all edges in the network (which we seek to minimize) against the efficiency of the network (which we seek to maximize).

Consider the following spatial network growth mechanism. Place $n-1$ points uniformly at random on the unit square (i.e., $0 \leq \mathbf{z}_x, \mathbf{z}_y \leq 1$) and one point in the exact center. This point is vertex 0. Now, add $n-1$ edges, one at a time, so that each edges connects one of the still disconnected nodes to the growing network. At each time step, add the edge (i, j) that has minimum weight under this function

$$w_{ij} = d_{ij} + \alpha \frac{d_{ij} + \ell_{j0}}{d_{i0}} \ ,$$

where d_{ij} is the Euclidean distance between vertices i and j, ℓ_{ij} is the distance along the shortest path in the network between i and j, and α is a free parameter. The first term is the length of the prospective edge, while the second term represents the routing time to the center of the network. Study the relationship between the *route factor*, defined as

$$q = \frac{1}{n} \sum_{i=1}^{n} \frac{\ell_{i0}}{d_{i0}} \ ,$$

and the free parameter α in the weight function. Present your results. Include example visualizations of the networks grown for a few values of α.