COMPLEX NETWORKS, CSYS303

HW05 WRITE-UP

David W. Landay
University of Vermont
Graduate Student, Comp. Systems and Data Sci.

Problem 1

For each of our main six networks, compute and present distributions of the shortest path length between all pairs of nodes. Notation: d_{ij} is the shortest distance between nodes i and j. Also compute the average shortest path length, $\langle d \rangle$.

With limited time and computational resources, we will only perform these tasks on the four smallest networks; these are the `C. Elegans`, `Dolphins`, `Karate Club`, and `Political Books` networks. The following output describes the average shortest path length over each network:

Network Name	Avg. Shortest Path Length
C.Elegans	3.9744
Dolphins	3.3028
Karate	2.3374
Pol.Books	3.0494

Figure 1: Average shortest path lengths of the four smallest networks.

The distributions of shortest path lengths over the same four networks are shown in the following figure:

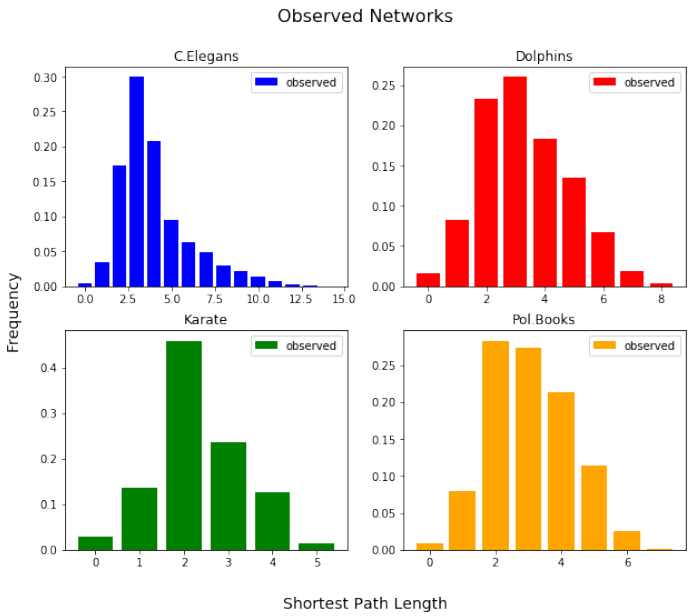

Figure 2: Distribution of shortest path lengths in each of the four smallest networks.

Problem 2

Generate ensembles of random networks of the same 'size' as the six networks.Process 1 random network and then scale up as computing power/time/sanity permits. 1000 random networks would be good.Size here means having the same number of nodes and the same number of edges. As for the real networks, compute the shortest path lengths for these random networks and present frequency distributions.

For each of the four networks, we create 1000 random networks that are of the same size as the original network. This means that every network in the random ensemble has the same number of nodes and edges. Using the `networkX` package, there are several convenient ways to generate random graphs, but it is just as easy to create 1000 random networks by assigning k random edges to n generated nodes, where k is the same quantity as $|E|$ and n is the same quantity as N ($|E|$ is the number of edges, and N is the number of nodes in the original network). The distribution of shortest path lengths for each ensemble of random networks is displayed in the following figure:

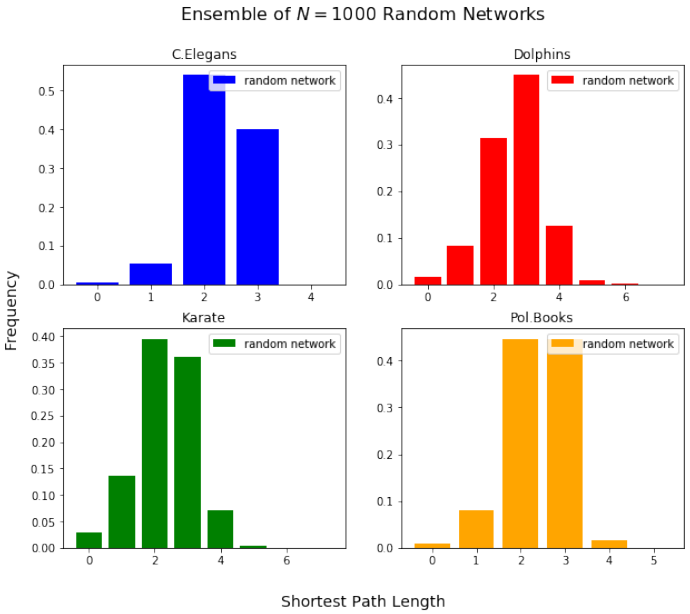

Figure 3: Distributions of shortest path lengths averaged over an ensemble of 1000 random networks. Each random network in the ensemble is the same size as the four original network.

The average shortest path length over the ensembles of networks is given in the following output:

Network Name	Avg. Shortest Path Length
C.Elegans	2.3426
Dolphins	2.6146
Karate	2.3241
Pol.Books	2.3813

Figure 4: Average shortest path length taken over each of the four ensembles of random networks.

Problem 3

Determine how well/poorly random networks produce the shortest path distributions of real world networks. Using whatever tests you like, show how well both the average shortest path length and the full distributions compare between the real network and their random counterparts.

The below figure shows the distributions from **Fig.2** and **Fig.3** superimposed on one another:

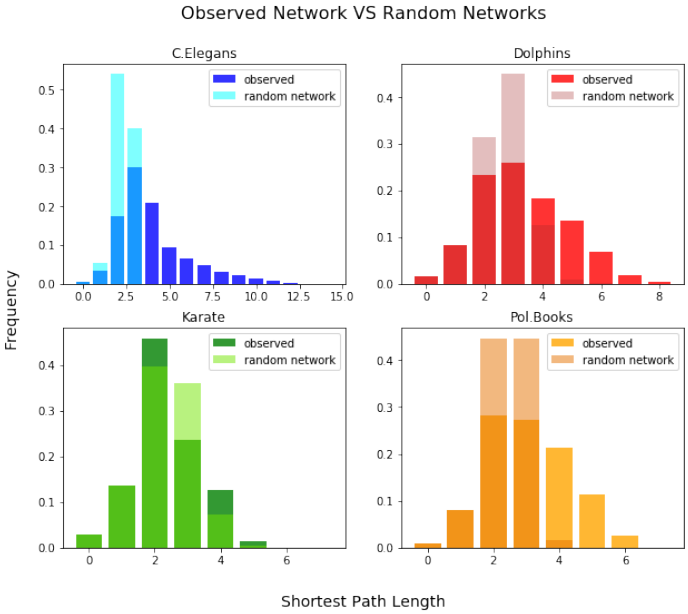

Figure 5: The shortest path length distribution for each of the four networks and that of the corresponding random ensemble superimposed.

Especially for the `C. Elegans` network, which is directed, we can tell that the distribution of shortest path lengths for the four networks, may not be well represented by the distributions generated from the ensemble of random networks. We would like to test whether or not this is true.

We cannot assume that our shortest path length distributions are normal (this is clearly true in the case of our directed network), so we must use an appropriate non-parametric test to understand whether the observed and random distributions were drawn from the same sample. We turn to the two-sample Kolmogorov-Smirnov (KS) test to challenge the null-hypothesis that, indeed, the distributions were drawn from the same sample. In the KS-test, the test statistic D is the maximum absolute difference between the two cumulative distribution functions. We look to calculate a p-value - the probability of observing a D-statistic at least as large as what we found - to test whether our null hypothesis is true. If we set our α-level to be 0.05, we get the following results:

Network Name	D (KS-Statistic)	D_α	p-Value
C.Elegans	0.4882	0.0052	0.0000
Dolphins	0.2720	0.0219	0.0000
Karate	0.0631	0.0400	0.0002
Pol.Books	0.3369	0.0130	0.0000

Figure 6: Results of the KS-test, to see whether the distributions were drawn from the same sample. Given the low p-values, we may reject the null-hypothesis that these distributions (corresponding to the same networks) were drawn from the same sample. Another way of saying this is that our observed D-statistic was greater than the D_α corresponding to the calculated p-value.

Problem 4

(a) Determine all possible three node motifs for undirected networks and sketch them. (Do not consider motifs for which one or more nodes are disconnected.)

Network motifs are subgraphs that may repeat themselves in a network or set of networks. They reveal structural elements of complex networks, and may reflect functional properties of the network (like flow). In undirected networks, there are only two possible

three node motifs. They are as follows:

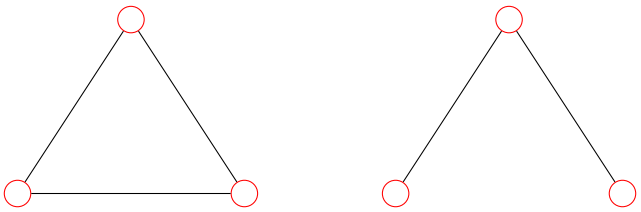

Figure 7: Possible three node motifs for an undirected network.

(b) Do the same as above for directed networks, allowing that a pair of nodes may have two edges traveling opposite ways behind them.

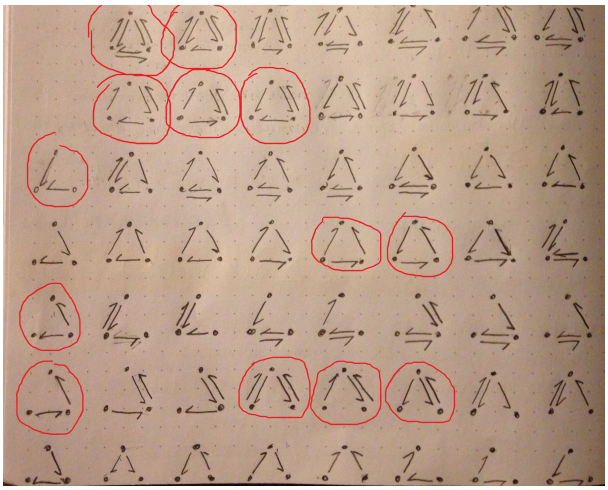

Figure 8: Motifs that are circled are the 13 total three node motifs. The remaining subgraphs are the total unique three node motifs if we care about the ordering of nodes. There are 54 in total.

Problem 5

For a uniformly distributed population, to minimize the average distance between individuals and their nearest facility, we've made a claim that facilities would be placed at the centres of the tiles on a hexagonal lattice (or the vertices of a triangular lattice). Why is this?

If we construct a triangular lattice, notice that if we place a facility on any vertex of

the lattice, then that facility is equidistant to the six nearest/adjacent vertices.Furthermore, as demonstrated in this diagram, any vertex is the center of a hexagon and thus constructs the same argument for hexagonal lattices

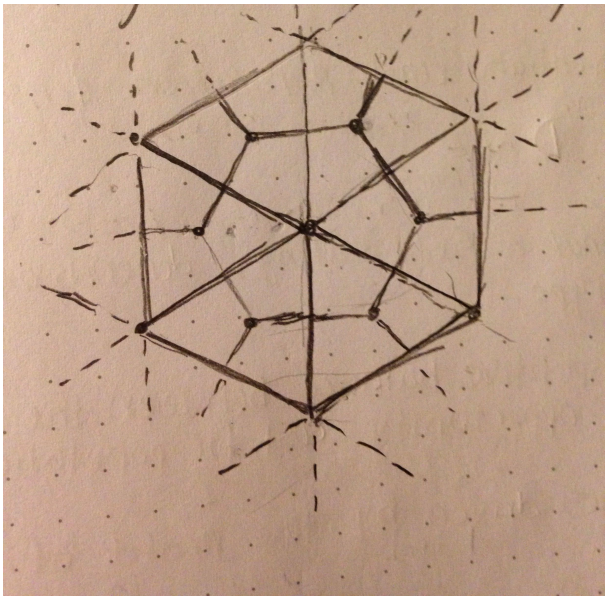

Figure 9: Construction of triangular and hexagonal lattices; proving that these network configurations minimize the distance between individuals and their nearest facility

Also notice that placing the facilities in the center of the triangular tiles still gives us that the facilities are equidistant to any three sides of the triangle they are centered within. If we draw lines to connect the facilities, notice that now we are constructing a hexagonal lattice, on which each facility has three equally distant neighbors. Hence, the configurations that minimize the distance between individuals and their nearest facilities are the triangular lattice and hexagonal lattice.

Problem 7

Following Um et al.'s approach, obtain a more general scaling for mixed public-private facilities in two dimensions. Use the cost function:

$$c_i = n_i \langle r \rangle^{\beta}, \text{ for } 0 \leq \beta \leq 1$$

where, respectively, n_i and $\langle r \rangle$ are the population and the average 'source to sink'distance for the population of the i_{th} Voronoi cell (which surrounds the i_{th} facility). Note that $\beta = 0$ corresponds to purely commercial facilities, and $\beta = 1$ corresponds to strongly social ones.

The following relationships are outlined in the Um et. al paper:

$$\rho(r) = \frac{n_i}{s_i} \tag{1}$$

$$D(r) = \frac{1}{s_i} \tag{2}$$

$$\langle r \rangle \approx g\sqrt{s_i} \tag{3}$$

Here, **(1)** denotes the population density within a Voronoi cell i, **(2)** denotes the facility density within that same cell, and **(3)** denotes the average distane to the i_{th} facility governed by the product of g, some geometrical constant $\sim \mathcal{O}(1)$, and the root of the area i. From the outlined relationships **(1)**, **(2)**, and **(3)**, we get the following:

$$c_i = n_i g^{\beta} s_i^{\frac{\beta}{2}}$$

$$= \rho(r) s_i g^{\beta} s_i^{\frac{\beta}{2}}$$

$$= \rho(r) g^{\beta} s_i^{\frac{\beta}{2}+1} \tag{4}$$

We may re-write **(4)** in the following way:

$$s_i^{\frac{\beta}{2}+1} = s_i^{\frac{2+\beta}{2}}$$

$$\implies s_i^{\frac{2+\beta}{2}} = c_i g^{-\beta} \rho(r)^{-1} \tag{5}$$

Finally, **(5)** and **(2)** give us the solution:

$$(5)\&(2) \implies \rho(r)^{-1} \propto D(r)^{-\frac{2+\beta}{2}}$$

$$\implies \rho(r)^{\frac{2}{2+\beta}} \propto D(r) \therefore \tag{6}$$