Problem 1 (Adding a Link to a Network)**.** Fix an undirected network. For each of the following network statistics, when a link is added to the network, does the statistic always increase, always decrease, or sometimes increase and sometime decrease (depending on the network and the location of the new link)? For each answer, either give a proof that the statistic always increases or always decreases, or an example that shows it can go either way.

(a) average degree

(b) diameter

(c) average path length

(d) overall clustering coefficient

(e) average clustering coefficient

(f) decay centrality, for an arbitrary node i

(g) betweenness centrality, for an arbitrary node i

Solution.

Supporting Claim 1. Adding an edge doesn't increase the shortest path length $\ell(u, v)$ for any pair of vertices (u, v).

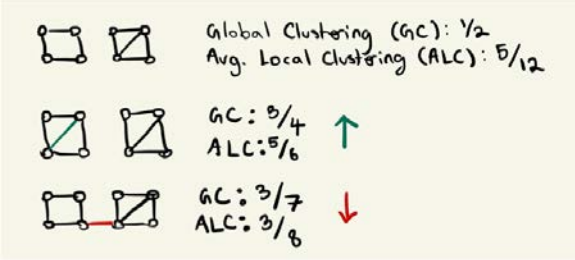
Proof of claim. All paths in the old graph remain in the new graph, so the new shortest u v path can be no longer than the old one. □

a. Each edge is incident to two vertices, so if E is the total number of edges then $\sum_v d_v = 2E$. Therefore the average degree is $2E/n$ and *increases* when we add an edge.

b. The diameter is
$$\max_{x,y \in G} \ell(x, y).$$
Since all of the $\ell(x, y)$ are non-increasing, so is the diameter.

c. Similarly, since $\ell(x, y)$ is non-increasing, so is the average shortest path length
$$\binom{n}{2}^{-1} \sum_{(x,y) \in G} \ell(x, y).$$

d, e. Both of these can go in either direction.

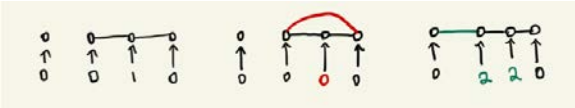


f. Since $\delta \leq 1$ and $\ell(x, y)$ is non-increasing, $\delta^{\ell(x,y)}$ is non-decreasing. So the decay centrality

$$\sum_{(x,y)\in G} \delta^{\ell(x,y)}$$

is non-decreasing.

g. Can go in either direction.



Problem 2 (The Adjacency Matrix)**.** Let $\mathbf{g}$ be the adjacency matrix of an undirected network, and let $\mathbf{1}$ be the column vector whose elements are all 1. In terms of these quantities write expressions for:

(a) The vector $\mathbf{d}$ whose elements are the degrees d_i of the nodes.

(b) The number m of edges in the network.

(c) The matrix $\mathbf{N}$ whose element N_{ij} is the number of common neighbors of nodes i and j.

(d) The total number of triangles in the network, where a triangle means three nodes, each connected by edges to both of the others.

Solution.

(a) Note that the ith entry of $\mathbf{d}$ is given by

$$\sum_{j=1}^{n} g_{ij} = (\mathbf{g1})_i.$$

We conclude that $\mathbf{d} = \mathbf{g1}$.

(b) Each edge (i, j) corresponds to two 1s appearing among the entries of $\mathbf{g}$, namely the entries g_{ij} and g_{ji}. It follows that

$$2m = \sum_{i,j=1}^{m} g_{ij} = \mathbf{1}^\intercal \mathbf{g1}.$$

(c) For each common neighbor k of nodes i and j, there is a unique length-2 walk $i \quad k \quad j$, and vice-versa. Thus, the number of such common neighbors N_{ij} is equal to the number of $i \quad j$ walks of length 2, namely $\mathbf{g}^2$ (as proved in lecture).

(d) Similarly, each triangle corresponds to a unique length-3 walk beginning and ending at a give node. This is exactly the sum of the diagonal entries of $\mathbf{g}^3$ or trace($\mathbf{g}^3$).

Problem 3 (Betweenness Centrality in Trees)**.** Consider an undirected (connected) tree of n vertices. Suppose that a particular vertex k in the tree has degree d, so that its removal would divide the tree into d disjoint regions, and suppose that the sizes of those regions are $n_1, \ldots, n_d$.

(a) Show that the betweenness centrality of the vertex is

$$B_k = 1 - \sum_{m=1}^{d} \frac{n_m(n_m - 1)}{(n-1)(n-2)}.$$

(b) Using this result, calculate the betweenness of the ith vertex from the end of a "line graph" of n vertices, i.e., n vertices in a row.

Solution. (a) Let R_m denote region m. If i and j are two vertices in the same region, none of the shortest paths between them passes through vertex k, while if they are in different regions, all of the shortest paths between them pass through vertex k.

Therefore the betweenness centrality of k is equal to the number of pairs of vertices *not* belonging to the same region divided by the total number of pairs of vertices excluding k. The total number of pairs of vertices excluding k is

$$\binom{n-1}{2} = \frac{(n-1)(n-2)}{2}.$$

The number of vertex pairs not belonging to the same region the total number of vertex pairs minus those that *do* belong to the same region, which is

$$\binom{n-1}{2} - \sum_{m=1}^{d} \binom{n_m}{2}.$$

Dividing the two quantities gives the formula we wanted.

(b) The "line graph" has the form described above, with regions of size $l-1$ and $n-l$. Plugging this into the above, we get

$$B_l = 1 - \frac{(l-1)(l-2)}{(n-1)(n-2)} - \frac{(n-l)(n-l-1)}{(n-1)(n-2)}.$$

Problem 4 (Expected Degree)**.** First, some definitions. Fix an undirected graph $G = (N, E)$ with finitely many nodes, none of which have degree zero. We will use $N(i)$ to denote the *neighborhood* of a node i—that is, the set of all nodes j that share some edge with i. Define d_i to be the degree of node i, which is also the size of i's neighborhood: $d_i = |N(i)|$.[1] Let $P(d)$ be the fraction of the nodes in the graph with degree d.

(a) Suppose we pick an edge uniformly at random[2] and then pick either node of that edge with equal probability. Call that node i. Let D be the degree of i; it is a random variable because the node was random. What is the expectation of D? Write your answer in terms of P.

(b) Prove that the expectation of D is at least as large as the mean of P.

(c) We make a definition to keep track of how popular i's neighbors are, on average.

Definition. Define δ_i to be the arithmetic mean of the degree of i's neighbors. That is,

$$\delta_i = \frac{\sum_{j \in N(i)} d_j}{d_i}.$$

Theorem. For graph $G = (N, E)$ satisfying the conditions given in this problem,

$$\frac{1}{|N|} \sum_{i \in N} d_i \leq \frac{1}{|N|} \sum_{i \in N} \delta_i.$$

Prove this statement.

Solution.

a. To understand this, let us start with a simple case: a network on four nodes with three links: $\{12, 23, 34\}$. So, one half of the nodes have degree 1 and one half have degree 2. That is, $P(1) = \frac{1}{2} = P(2)$.

It is easy to see that if we randomly pick a link and then randomly pick an end of it, there is a $\frac{2}{3}$ chance that we find a node of degree 2 and a $\frac{1}{3}$ chance that we find a node of degree 1. This just reflects the fact that higher degree nodes are involved in a proportionately higher percentage of the links.

In fact, their degree determines relatively how many more links they are involved with. In particular, if we randomly pick a link and a node at the end

[1] Recall we use the notation $|S|$ to denote the number of elements in the set S.

[2] That is, each edge of the graph is chosen with equal probability.

of it, and we consider two nodes of degrees d_j and d_k, then node k is relatively d_k/d_j times more likely to be the one we find than node j.

Let $\widetilde{P}$ be the distribution of D. From the above we can gyess that that probability of a degree-d node being sampled is proportional to degree, making it natural to conjecture that

$$\widetilde{P}(d) = \frac{P(d)d}{\mu_P},$$

where $\mu_P = \sum_d P(d)d$ is the expected degree under the distribution P.

To be a bit more formal or precise, let us fix a graph (N, E), where $N = \{1, \ldots, n\}$ is the set of notes and $E \subseteq N \times N$ is the set of edges. Since $P(d)$ is the fraction of the nodes in the graph with degree d, there are $nP(d)$ nodes with degree d.

Each edge is picked uniformly at random, and then either node of that edge is picked with equal probability. So we have $2|E|$ events corresponding to the results of the node selection process, each occurs with probability $\frac{1}{2|E|}$. Among those events, each node with degree d is chosen d times, and there are exactly $nP(d)$ nodes with degree d. Thus, a node with degree d is chosen with probability $\frac{nP(d)d}{2|E|}$. It is familiar that the sum of degrees across all nodes is equal to $2|E|$ (see Problem 1), so the expected degree $\mu_P = \frac{2|E|}{n}$, giving rise to the formula of $\widetilde{P}$ above.

The random variable D is distributed according to $\widetilde{P}$. Therefore, the expectation of D is

$$\mu_D = \sum_d \widetilde{P}(d)d = \sum_d \frac{P(d)d^2}{\mu_P}.$$

b. Start from the fact that the variance of P is nonnegative:

$$0 \leq \sigma_P^2 = \sum_d P(d)\left(d - \mu_P\right)^2 = \sum_d P(d)d^2 - \mu_P^2.$$

We obtain

$$\sum_d P(d)d^2 \geq \mu_P^2 \quad \Leftrightarrow \quad \mu_D = \sum_d \frac{P(d)d^2}{\mu_P} \geq \mu_P.$$

For an alternative version of this proof, define a random variable X with distribution P. A standard fact (which you can work out from basic algebra with

quadratics) is that

$$\mu_D = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} = \frac{\mathbb{E}[X]^2 + \mathrm{Var}[X]}{\mathbb{E}[X]}.$$

So we conclude that

$$\mu_D = \mathbb{E}[X] + \frac{\mathrm{Var}[X]}{\mathbb{E}[X]} = \mu_P + \text{a positive number.}$$

The positive number is $\frac{\sigma_P^2}{\mu_P}$, which tells us that $\mu_D - \mu_P$ is larger the more variable is P (in the sense captured by variance over mean).

c. It is equivalent to show

$$\sum_{i \in V} d_i \leq \sum_{i \in V} \delta_i. \tag{1}$$

Since $d_i = |N(i)|$, we can write the left-hand side of equation (1) as

$$\sum_{i \in V} d_i = \sum_{i \in V} \sum_{j \in N(i)} 1 = \sum_{j<i,\,(ij) \in E} 2.$$

The last equality is due to the fact that each edge is counted twice within the summation $i \in V, j \in N(i)$.

Similarly, by definition, the right-hand side of equation (1) can be written as

$$\sum_{i \in V} \delta_i = \sum_{i \in V} \frac{\sum_{j \in N(i)} d_j}{d_i} = \sum_{i \in V} \sum_{j \in N(i)} \frac{d_j}{d_i} = \sum_{j<i,\,(ij) \in E} \left(\frac{d_j}{d_i} + \frac{d_i}{d_j} \right).$$

Therefore, equation (1) is equivalent to

$$\sum_{j<i,\,(ij) \in E} 2 \leq \sum_{j<i,\,(ij) \in E} \left(\frac{d_j}{d_i} + \frac{d_i}{d_j} \right).$$

Because

$$\frac{d_j}{d_i} + \frac{d_i}{d_j} \geq 2$$

for any d_i and d_j, the final inequality is satisfied. Thus by the series of equivalences the initial inequality is satisfied.

Note that the fact $\frac{d_j}{d_i} + \frac{d_i}{d_j} \geq 2$ is a special case of the fact that $x + \frac{1}{x} \geq 2$ for any real number $x > 0$, which can be easily verified using high school calculus.

14.15 / 6.207 Networks

Spring 2022