1. Networks and degree distributions. Consider a network with N total vertices and E total undirected, unweighted edges.

   (a) What is the probability f that an edge connects two vertices connected at random?

   $$f = \frac{E}{\binom{N}{2}} = \frac{2E}{N(N-1)}$$

   (b) What is the average number of neighbors J per protein?

   $$J = (N-1)f = \frac{2E}{N}$$

   (c) The probability p_n that a protein has exactly n neighbors follows a binomial distribution. Provide this distribution.

   $$p_n = \binom{N-1}{n} f^n (1-f)^{N-1-n}$$

   (d) Show that in the limit that $f \to 0$, $N \to \infty$, and J a finite constant, that the binomial probability distribution for p_n approaches a Poisson distribution. Provide the Poisson distribution.

   $$\begin{aligned} p_n = & \binom{N-1}{n} f^n (1-f)^{N-1-n} & (1) \\ \approx & \binom{N-1}{n} f^n (e^{-f})^{N-1-n} & (2) \\ \approx & \frac{(N-1)(N-2)...(N-1-n)}{n!} \left(\frac{J}{N-1}\right)^n e^{-fN} e^{f(1+n)} & (3) \\ \approx & \frac{e^{-J} J^n}{n!} & (4) \end{aligned}$$

   So, p_n approximately follows a poisson distribution with parameter J.

   (e) Based on the Poisson distribution, for what value of J do we expect on average only one vertex in the entire network to have no neighbors?

   $$p_0 = \frac{e^{-J} * J^n}{n!}|_{n=0} = e^{-J}$$

   Thus, the average number of nodes with no neighbors is equal to:

   $$n_0 = Np_0 = Ne^{-J}$$

   For n_0 to be equal to one, we should have $J = ln(N)$.

2. Network motifs. Again consider a network with N total vertices and E total edges. The edge probability f and the average degree J can be calculated from N and E and can also be used in your answers. You can assume the thermodynamic limit that $f \to 0$, $N \to \infty$, and $fN \to J$.

   (a) What is the probability that three vertices, selected at random, have all edges (a 3-clique)? How many 3-cliques exist in the network? The probability that three vertices have all edges is $p = f^3$.
   There are $\binom{N}{3}$ total possible 3-cliques, so the expected number is $\binom{N}{3} f^3$.

   (b) How many edges are possible for a group of k vertices? Given a set of k vertices, what is the probability that all these edges exist, giving a k-clique? How many k-cliques are expected in a random network?
   The number of possible edges is

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

   The probability of all edges existing is $f^{\binom{k}{2}}$. The expected number of k-cliques is $\binom{N}{k} f^{\binom{k}{2}}$

   (c) Now suppose that the N vertices are not randomly organized. Instead, there are M modules, each with n vertices within the module. Edges occur only within modules, with no edges between modules. For example, in a social network, modules might be families, workplaces, or university classes. In a biological network, modules might be protein complexes. In this case, f for edges between modules is 0. What is f within a module? What is the expected number of 3-cliques as a function of M? What is the clustering coefficient as a function of M, defined as the number of 3-cliques for the given M relative to the number for a non-modular network with $M = 1$? There are $\frac{E}{M}$ edges per module. The edge probability within a module is

$$\begin{aligned} f_M &= \frac{\text{edges per module}}{\text{possible edges per module}} \\ &= \frac{\frac{E}{M}}{\binom{n}{2}} \\ &= \frac{2E}{Mn(n-1)} \end{aligned}$$

   The expected number of 3-cliques is $M\binom{n}{3} f_M^3$. The clustering coefficient is

$$\frac{\frac{E(n-2)}{3} f_M^2}{\binom{N}{3} f^3}$$

3. Disease spread, the giant component, and network diameter. Suppose a network has N total vertices, representing people. Each infected person transmits the disease to an additional J people.

   (a) Define $C_n(J)$ as the number of people infected through a path of length n, with $C_0(J) = 1$ and $C_1(J) = J$. Provide an expression for the general case, $C_n(J)$.

   $$C_n(J) = J^n$$

   (b) Define the total number of infected people as $T(J) = \sum_{n=0}^{\infty} C_n(J)$. Provide $T(J)$ for $J < 1$ and for $J > 1$. Remember that $T(J)$ must be $\leq N$.

   $$T_J = \sum_{n=0}^{\infty} C_n(J) = \begin{cases} \frac{1}{1-J} & J < \frac{N-1}{N} \\ N & J \geq \frac{N-1}{N} \end{cases}$$

   (c) For $J > 1$, the network radius ρ is roughly defined as $C_\rho(J) = N$, the number of steps to infect the entire network. Provide ρ in terms of N and J.

   $$\rho = \frac{\log N}{\log J}$$

4. Graph diffusion. A network has 4 vertices, numbered $1, 2, 3, 4$. The network has 4 undirected edges: $1 \sim 2$, $1 \sim 3$, $2 \sim 3$, $3 \sim 4$.

   (a) Provide the adjacency matrix $\mathbf{A}$ and the degree matrix $\mathbf{D}$.

   $$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

   (b) Define the matrix $\mathbf{A}(n)$ as having matrix elements $a(n)_{ij}$ that give the number of paths of length n starting at j and ending at i. Provide $\mathbf{A}(n)$ for $n = 0, 1, 2$. Describe how to calculate $\mathbf{A}(n)$ for other values of n.

   $$\mathbf{A}(n) = \mathbf{A}^n$$
   $$\mathbf{A}(0) = \mathbf{I}$$
   $$\mathbf{A}(1) = \mathbf{A}$$
   $$\mathbf{A}(2) = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

(c) Suppose that a random walk leaving a vertex chooses each of the outgoing edges with equal probability. Let the matrix $\mathbf{P}(n)$ have matrix elements $p(n)_{ij}$ giving the probability that an n-step random walk that started at j ends at i. Provide the matrix $\mathbf{P}(n)$ for $n = 0, 1, 2$. Describe how to calculate $\mathbf{P}(n)$ for other values of n.

$$\begin{aligned}
\mathbf{P}(n) &= \left(\mathbf{A}\mathbf{D}^{-1}\right)^n \\
\mathbf{P}(0) &= \mathbf{I} \\
\mathbf{P}(1) &= \mathbf{A}\mathbf{D}^{-1} \\
&= \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 \end{bmatrix} \\
\mathbf{P}(2) &= \mathbf{P}(1)\mathbf{A}\mathbf{D}^{-1} \\
&= \begin{bmatrix} \frac{5}{12} & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{6} & \frac{5}{12} & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{2}{3} & 0 \\ \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} \end{bmatrix}
\end{aligned}$$

5. Graph diffusion. Now consider a general graph with adjacency matrix $\mathbf{A}$ and degree matrix $\mathbf{D}$.

   (a) The graph Laplacian $\mathbf{L}$ is defined as $\mathbf{D} - \mathbf{A}$. Prove the the column vector $\mathbf{1}$, a column vector with each element 1, is an eigenvector of $\mathbf{L}$ with eigenvalue 0.
   If d_{ij} and a_{ij} are the i, j-th element of $\mathbf{D}$ and $\mathbf{A}$, respectively, then:

$$\begin{aligned}
\mathbf{L}\mathbf{1} &= 0 \\
(\mathbf{D} - \mathbf{A})\mathbf{1} &= 0 \\
\sum_{i,j}(d_{ij} - a_{ij}) &= 0 \\
\sum_{i,j} d_{ij} - \sum_{i,j} a_{ij} &= 0 \\
2E - 2E &= 0
\end{aligned}$$

The last step holds because the first sum is over the degree of all nodes, which equals $2E$, and the second sum is over the adjacency matrix which has two elements of value 1 for each edge and therefore also equals $2E$.

(b) Suppose that random walks are sampled over many lengths, with the probability of n steps following a Poisson distribution with mean λ. Define the matrix $\mathbf{P}(\lambda)$ as having matrix elements $p(\lambda)_{ij}$ giving the probability that a random walk following this process and starting at j ends at i. Provide a closed-form expression for $\mathbf{P}(\lambda)$.

$$\begin{aligned}\mathbf{P}(\lambda) &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \left(\mathbf{A}\mathbf{D}^{-1}\right)^k \\ &= \sum_{k=0}^{\infty} \frac{\left(\lambda \mathbf{A}\mathbf{D}^{-1}\right)^k}{k!} e^{-\lambda} e^{-\lambda \mathbf{A}\mathbf{D}^{-1}} e^{\lambda \mathbf{A}\mathbf{D}^{-1}} \\ &= e^{-\lambda} e^{\lambda \mathbf{A}\mathbf{D}^{-1}} \sum_{k=0}^{\infty} \frac{\left(\lambda \mathbf{A}\mathbf{D}^{-1}\right)^k}{k!} e^{-\lambda \mathbf{A}\mathbf{D}^{-1}} \\ &= e^{-\lambda} e^{\lambda \mathbf{A}\mathbf{D}^{-1}} \\ &= e^{-\lambda\left(\mathbf{I}-\mathbf{A}\mathbf{D}^{-1}\right)} \\ &= e^{-\lambda\left(\mathbf{I}-\mathbf{A}\mathbf{D}^{-1}\right)\mathbf{D}\mathbf{D}^{-1}} \\ &= e^{-\lambda \mathbf{L}\mathbf{D}^{-1}}\end{aligned}$$

(c) Your answer to the previous question should have had the form $e^{-\lambda \mathbf{M}}$, where $\mathbf{M}$ is a matrix. It is possible to prove that $\mathbf{M}$ is positive semidefinite: it has at least one eigenvector with eigenvalue 0, and all other eigenvectors have eigenvalues ≥ 0. Suppose that random walks are started at time 0 according to column vector β, with β_i equal to the probability that the random walk starts at vertex i. Prove that in the limit $\lambda \to 0$, the steady-state distribution is proportional to the eigenvector with eigenvalue 0.
TODO

(d) Provide a closed-form expression for an eigenvector of $\mathbf{M}$ with eigenvalue 0.
If e_i is the ith element of eigenvector $\mathbf{E}$,

$$\begin{aligned}\mathbf{L}\mathbf{D}^{-1}\mathbf{E} &= 0 \\ \mathbf{D}^{-1}\mathbf{E} &= \mathbf{1} \\ e_i &= d_{i,i}\end{aligned}$$

6. Flux-balance analysis of metabolic networks. Suppose that an organism has X muscle cells and Y neural cells. The organism can generate 1000 units of energy per minute. Muscle

cells require 1 unit/min and neural cells require 2 units/min. The organism needs at least 200 muscle cells and 100 neural cells to survive.

(a) What are the coordinates defining the vertices of the feasible region in (X,Y) space?

$$(200,100),(200,400),(800,100)$$

(b) Suppose that the fitness $\phi(X,Y)$ of the organism is a linear function,

$$\phi(X,Y) = aX + bY,$$

with a and b positive. What condition on (a,b) permits multiple optimal solutions? What condition gives $X > Y$ at the optimum? What condition gives $X < Y$ at the optimum?

There are multiple optimal solutions when the slope of the fitness equation is the same as the slope of the outside of the feasible region. The slope of the fitness is $\frac{dy}{dx} = -\frac{a}{b}$. The slope of the feasible region is -2. So we have multiple optimal solutions when $\frac{a}{b} = 2$. We can check this by seeing that if $a = 2b$, two vertices of the feasible region have the same cost. If $\frac{a}{b} > 2$, then the only optimum is $(800,100)$. If $\frac{a}{b} < 2$, then the only optimum is $(200,400)$.