Internet Analytics (COM-308)

Problem Set 2 - Solutions

Problem 1

(a) Assume $G(V, E)$ is a connected, undirected, non-bipartite graph. We run a random walk $\{X_t\}$ on this graph, as seen in class. For some arbitrary edge $e \in E$, compute the probability that the random walk traverses edge e, asymptotically for $t \to \infty$.

Under the assumptions on G, the random walk is ergodic. Therefore, it has a unique stationary distribution π, which we saw in class satisfies $\pi_i = d_i/2m$, with d_i the degree of vertex i, and $m = |E|$. The event { the random walk traverses edge $e = (u, v)$ } can be written as $\{X_t = u, X_{t+1} = v \cup X_t = v, X_{t+1} = u\}$. Given that under π, $\mathbb{P}(X_t = u) = d_u/2m$, and $\mathbb{P}(X_{t+1} = v|X_t = u) = 1/d_u$, $\mathbb{P}(X_t = u, X_{t+1} = v) = 1/2m$. The probability of the reverse traversal $v \to u$ is also $1/2m$, and the two events are disjoint. Hence, the total probability is $1/m$ for every edge $e \in E$. Note that when we considered the Friendship Paradox, we considered uniform edge sampling.

(b) To convince ourselves that the second formula in class for $\widehat{F}$ is correct, compute both the expectation under π of the numerator and of the denominator of the expression for the "self-normalized" estimator.

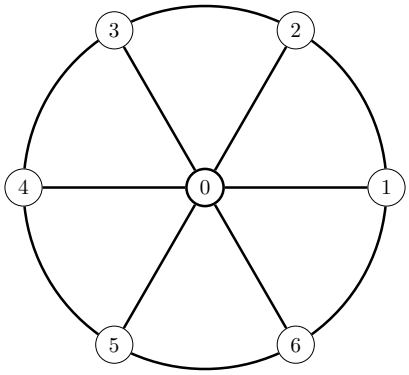
Numerator:

$$\mathbb{E}_\pi \sum_{t=1}^{T} f(X_t)/d(X_t) = \sum_{t=1}^{T} \mathbb{E}_\pi f(X)/d_X = \sum_{t=1}^{T} \sum_{v \in V} \frac{d_v}{2m} \frac{f(v)}{d_v} = \frac{nT}{2m} F, \tag{1}$$

while the denominator equals

$$\mathbb{E}_\pi \sum_{t=1}^{T} 1/d(X_t) = \sum_{t=1}^{T} \mathbb{E}_\pi 1/d(X) = \sum_{t=1}^{T} \sum_{v \in V} \frac{d_v}{2m} \frac{1}{d_v} = \frac{nT}{2m}. \tag{2}$$

(c) Consider the undirected graph G_1 (Figure 1). Find the stationary distribution of the random walk on this graph. Which node has the highest visiting probability?



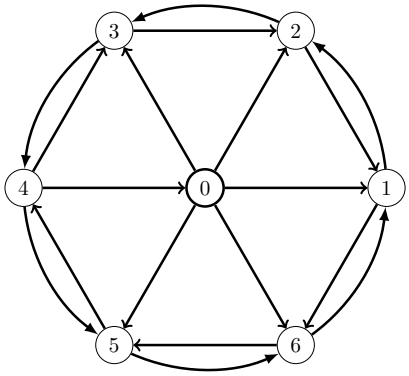
Undirected graph G_1.

We have seen that a random walk on an undirected graph G has stationary distribution $\pi(.) \propto d(.)$. Assume π_i is the visiting probability of node i in the stationary regime. This implies that

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6) = \left(\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right),$$

where the node 0 has the highest visiting probability.

(d) Consider the directed graph G_2 (Figure 2). This graph is a directed version of graph G_1. Does the random walk have a stationary distribution on this graph? If there is, compute it, and identify the node with the highest visiting probability; otherwise, explain why there is no stationary distribution.



Directed graph G_2.

The random walk on this graph is both irreducible and aperiodic (non-bipartite). Therefore, there is a stationary probability distribution. To find the stationary distribution $\pi = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6)$, we solve the following set of equations:

$$\begin{aligned}
\pi_0 &= \pi_4/3 \\
\pi_1 &= \pi_0/5 + \pi_2/2 + \pi_6/2 \\
\pi_2 &= \pi_6 = \pi_0/5 + \pi_1/2 + \pi_3/2 \\
\pi_3 &= \pi_5 = \pi_0/5 + \pi_2/2 + \pi_4/3 \\
\pi_4 &= \pi_3/2 + \pi_5/2 = \pi_3 = \pi_5 \\
\pi_0 &+ \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 = 1.
\end{aligned}$$

This yields

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6) = \left(\frac{1}{21}, \frac{19}{105}, \frac{6}{35}, \frac{3}{21}, \frac{3}{21}, \frac{3}{21}, \frac{6}{35}\right),$$

where node 1 has the highest visiting probability.

(e) Check whether the stationary distribution π for the directed graph G_2 is proportional to either the node in–degrees and/or their out–degrees?

No, it is neither. The nodes $1, 2, 3, 5$ and 6 have matching in– and out–degrees, but they do not have the same visiting probabilities.

Note that in the graph G_2, the high degree node 0 has many outgoing edges, but it has just one incoming edge. Intuitively, once the RW leaves node 0, it takes a long time to visit it again.

Problem 2

The conductance Φ of a graph measures how well different node subsets are connected to their complements. We saw in class that this has connections to the mixing time of a random walk on the graph.

We want to gain some intuition about this measure through examples. For this, assume n is even, and compute the conductance of the following three graphs:

(a) the cycle C_n.

(b) the complete graph K_n.

(c) two copies of $K_{n/2}$ connected by a single edge.

As seen in class, the conductance of a graph $G(V, E)$, with n nodes and m edges, is defined as

$$\Phi = \min_{S \subset V} \frac{|\delta S|}{2m\pi(S)\pi(S')}, \tag{3}$$

where S is a non empty set of nodes, $S' = V \setminus S$, and δS is the set of edges between S and S'. For each of these graphs, you first have to guess the bottleneck, i.e., the sets $\{S, S'\}$ that achieve the minimum.

(a) $\Phi = \dfrac{2}{2m\frac{1}{2}\frac{1}{2}} = \dfrac{4}{m} = \dfrac{4}{n}$. The bottleneck cut separates the cycle into two line graphs, with δS of size 2. Note that the denominator is maximized when the two line graphs are of identical size, so the bottleneck is between two equal halves. (Note: the number of nodes n would need to be odd for the RW to be ergodic.)

(b) $\Phi = \min_{S \subset V} \dfrac{|S|(n-|S|)}{2m\frac{|S|}{n}\frac{n-|S|}{n}} = \dfrac{n}{n-1}$. For this graph, the conductance is equal for all possible cuts, so there is no dominating bottleneck.

(c) $\Phi = \dfrac{2}{m} = \dfrac{2}{\frac{n}{2}(\frac{n}{2}-1)+1} = \dfrac{8}{n^2-2n+4}$.

As expected, the graph with the lowest conductance is the graph with two copies of $K_{n/2}$ connected by a single edge. A random walk on this graph converges slowly because it is very hard to get from one $K_{n/2}$ to the other $K_{n/2}$.