18.419 Random Walks and Polynomial-Time Algorithms February 14, 2002

Lecture 4:Discrete Random Walks

Lecturer: Santosh Vempala *Scribe: Grant Wang*

Introduction

In this lecture, we will cover some basic parameters of discrete random walks on graphs, as well as some basic results that will help us gain intuition. The idea behind a random walk on a graph $G = (V, E)$ is simple: we are at some vertex v, and we choose a vertex u uniformly at randomly from the set of neighbors of v. If we repeat this process, we have a random walk on the graph G.

Wuch a random walk is a Markov chain with random variables $X_1, X_2, \ldots X_{n+1}$, where $\Pr[X_{t+1} = x | X_1 \ldots X_t] = \Pr[X_{t+1} = x | X_t]$. That is, our random walk is such that the probability we are at a node is only dependent on the last node we were at, rather than on all the previous nodes. The properties of the graph G determine the type of Markov chain that models the random walk. For example, if G is undirected, we have a time-reversible Markov chain. If G is undirected and regular, we have a symmetric Markov chain.

Natural questions

There are many natural questions to ask about random walks on graphs. One interesting theorem is the following, as proved by Polya.

Theorem 1 (Polya). *Consider an infinite random walk on a d-dimensional grid, starting at the origin. If $d \leq 2$, the walk comes back to the origin infinitely often. But if $d = 3$, the walk comes back to the origin a finite number of times.*

Other natural questions include:

- How many steps does it take to visit every node in the graph?
- How many steps does it take to arrive at node j, starting at node i?

- At each step, there is a probability distribution P_t, where $P_t(i)$ is the probability we are at node i on step t. As we continue walking randomly, is there a value t for which $P_t = P_{t+1}$?

Notation and basic parameters

At this point, we are ready to formalize these questions and develop some general notation for random walks on general, finite graphs. Let $G = (V, E)$, and let $v_0, v_1, \ldots v_t$ be the random walk, i.e. v_i is the state of the random walk. Denote by $P_t(i)$ the probability that at time t we are at vertex i, i.e. $P_t(i) = \Pr[v_t = i]$. Recall that a random walk on a graph consists of choosing a neighbor uniformly at random, and walking to that node. Therefore, we can define the matrix M, where $M_{ij} = p_{ij} = \frac{1}{d(i)}$ for $(i, j) \in E$, where $d(i)$ is the degree of node i. That is, the ijth entry of the matrix encodes the probability we step from node i to j. Note that if we let A be the adjacency matrix of G, then we just have

$$M = \begin{pmatrix} \frac{1}{d(1)} & 0 & \cdots \\ 0 & \frac{1}{d(2)} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} A$$

We also have the following equalities:

$$\begin{aligned} P_{t+1}(i) &= \sum_j P_t(j) p_{ji} \\ P_t &= M^T P_{t-1} = (M^T)^2 P_{t-2} = (M^T)^t P_0 \end{aligned}$$

Definition 2 (Stationary distribution). *The probability distribution P_t is **stationary** if $P_{t+1} = P_t$.*

Stationary distribution

Several natural questions arise about stationary distributions: do all graphs have stationary distributions? Is this stationary distribution ever reached if we do a random walk from any node? Is the stationary distribution unique? We will first show that all undirected graphs do have a stationary distribution.

Let $\pi(i) = \frac{d(i)}{2m}$, where $m = |E|$. We show that this distribution is stationary.

Lemma 3. *π is a stationary distribution for any undirected graph $G = (V, E)$.*

Proof. Consider the probability we are at node i on the next step from the distribution π. From above, we have that:

$$P(i) = \sum_j \pi_j p_{ji} = \sum_j \frac{d(j)}{2m} p_{ji} = \sum_{(i,j)\in E} \frac{d(j)}{2m} \frac{1}{d(j)} = \frac{d(i)}{2m} = \pi(i)$$

□

We can also answer the question whether the stationary distribution will be reached for general, undirected graphs. The answer is no; consider a bipartite graph (V_1, V_2, E). If the random walk starts on some $v \in V_i$, on all odd k we have $P_k(u) = 0$ for all $u \notin V_i$, and vice versa for even k.

Is the stationary distribution unique for undirected graphs? We can also answer this in the negative – consider a graph with two disconnected components: G_1, G_2. Then there are two stationary distributions: one in which the probability we are in G_1 is 0, and the other where the probability we are in G_2 is 0. However, if we ensure that the graph is both undirected and connected, we can show that such a stationary distribution is unique.

Lemma 4. *Let $G = (V, E)$ be an undirected, connected graph. Then G has a unique stationary distribution π.*

Proof. Suppose not, i.e. G has two stationary distributions π_1, π_2, i.e. $M^T\pi_1 = \pi_1, M^T\pi_2 = \pi_2$. Let $\pi_3 = \pi_2 - \pi_1$; without loss of generality, at least one element of π_3 is negative. Consider $\pi_1 + \alpha\pi_3$, where α is chosen so that $\pi_1 + \alpha\pi_3 \geq 0$, and there is some index j for which it takes 0. Then $\pi_1+\alpha\pi_3$ is also stationary, since $M^T(\pi_1+\alpha(\pi_2-\pi_1)) = \pi_1+\alpha(\pi_2-\pi_1)$. But then the set of vertices with zero probability must be disconnected from the rest of the graph, which contradicts the fact that G is connected. □

So, each undirected, connected graph has a unique stationary distribution. But we know for the bipartite graph, that a random walk can never achieve such a distribution. Is this particular for the bipartite graph, or are there other examples of undirected, connected graphs for which a random walk will never achieve the unique stationary distribution. In the next lecture, we answer this question in the negative:

Theorem 5. *Let $G = (V, E)$ be an undirected, connected graph that is not bipartite. Then a random walk on G converges to a unique stationary distribution.*

More parameters

One special case of a random walk is when our probability matrix M, is symmetric, i.e. $p_{ij} = p_{ji}$. In this case, the stationary distribution is defined by $\pi(i) = \frac{1}{n}$. We verify that this is indeed the stationary distribution:

$$P(i) = \sum_j \pi_j p_{ji} = \frac{1}{n} \sum_j p_{ji} = \frac{1}{n} \sum_j p_{ij} = \frac{1}{n} = \pi(i)$$

We now define a last few parameters for random walks:

Definition 6 (Time-reversible). *We call a random walk* ***time-reversible*** *if* $\pi_i p_{ij} = \pi_j p_{ji}$.

Definition 7 (Hitting time). *The* ***hitting time****,* $H(i, j)$ *from* i *to* j *in a random walk is the expected number of steps to begin at vertex* i *and walk to vertex* j.

Definition 8 (Cover time). *The* ***cover time*** *of a random walk is the maximum over all starting vertices of the expected number of steps needed to touch every vertex of the graph.*

Next lecture, we will define the **mixing time** or **mixing rate** of a graph in more detail. For now, we can think of it as how fast the random walk approaches the stationary distribution.

Examples

To gain some intuition about these parameters, and random walks in general, we go over a few examples:

Line graph

Consider a line graph on $n+1$ vertices, and n edges. We would like to determine the hitting time between any two vertices, i, k, i.e. $H(i, k)$. We first start off with a particular, simple case: $H(k-1, k)$. In this case, consider the line graph on $k+1$ vertices and k edges. We know that the stationary distribution is $\pi(k) = \frac{d(i)}{2k} = \frac{1}{2k}$. Therefore, in a stationary walk, if we are at the last node k, the expected number of times before we return to node k is $2k$. It follows that the hitting time from $k-1$ to k is just $2k-1$, i.e. we do not need to take the first step to the left.

From this, we can derive a recursive formula for $H(i,k)$. The idea is that to walk from i to k, we will first walk from i to $k-1$, and then from $k-1$ to k:

$$\begin{aligned} H(i,k) &= H(i,k-1) + H(k-1,k) \\ &= H(i,k-1) + 2k - 1 \\ &= (2i+1) + (2i+3) + \ldots (2k-1) \\ &= k^2 - i^2 \end{aligned}$$

It follows that the cover time for a line graph is $H(0,n) = (n-1)^2$.

Complete graph

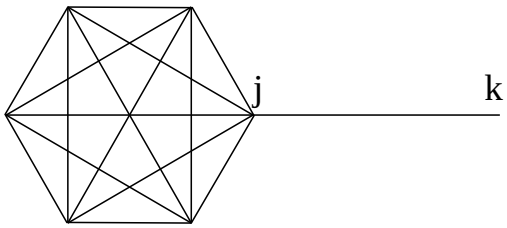
Here, we consider a complete graph on n vertices, and we would like to know the expected time to visit every vertex. This is simply the coupon collector problem – at each step, we visit some vertex chosen at random. How long does it take to visit each vertex in the graph? Let τ_i be a random variable that is the number of steps to visit i distinct vertices for the first time. We are interested in $E[\tau_n]$. Note that $\tau_{i+1} - \tau_i$ is the number of steps to visit a new vertex, after we have visited i vertices. The probability that we visit a new vertex is just $\frac{n-i}{n-1}$, so the expectation is simply $\frac{n-1}{n-i}$ steps. Therefore, we have:

$$\mathrm{E}[\tau_n] = \sum_i^{n-1} \mathrm{E}[\tau_{i+1} - \tau_i] = \sum_i^{n-1} \frac{n-1}{n-i} = n \sum_i^{n-1} \frac{1}{i} = O(n \log n)$$

Lollipop graph

Here, we show a graph where the hitting time is $O(n^3)$. Consider a graph on n nodes where $n/2$ are a complete subgraph, and there exists a line graph of $n/2$ nodes connected to a node j in the complete subgraph. Let k be the last node on the line graph, and let i be any node in the complete subgraph. Then $H(i,k) = \Omega(n^3)$. Note that to get to j, we need in expectation $\frac{n}{2}$ steps. It follows that to make one step onto the line path, from any node other than j on the subgraph, we need in expectation $\frac{n^2}{2}$ many steps. Lastly, since on a walk from one end to the other on a path of length $n/2$, we expect to return to the first node on the path $n/2$ times, it follows that $H(i,k) = \Omega(n^3)$.

For directed graphs, the hitting time can be exponential. Consider a line graph of n nodes, with edges oriented from left to right. Construct an edge $(i,0)$ for $i \in V$. Then to go from 0 to n takes at least 2^n steps in expectation.



We show that the lollipop graph is actually the worst case for undirected graphs:

Theorem 9. *The cover time for any connected, undirected graph G is $O(mn) = O(n^3)$.*

Proof. First, we show that for $(i,j) \in E$, $H(i,j) + H(j,i) \leq 2m$. We can bound this by above by considering the expected number of steps to walk from i to j and then use the edge (j,i). Recall that in a stationary walk, if we have just traversed an edge, the expected number of steps before we traverse the same edge in the same direction is $2m$. Therefore, if we are at node i with adjacent node j, the expected number of steps before the edge (j,i) is traversed is $2m$. This is the desired upper bound on $H(i,j) + H(j,i)$.

Now, consider a spanning tree T of G constructed by depth-first search. We show that the cover time is bounded above by $2m(n-1)$. We traverse the tree in a depth-first fashion $(v_1, v_2, \ldots v_{2n-1} = v_1)$, so that each edge is traversed twice. Let T be the cover time – then we have:

$$T \leq \sum_{i=1}^{2n-1} H(v_i, v_{i+1}) = \sum_{(i,j)\in T} H(i,j) + H(j,i) \leq 2m(n-1)$$

It follows that T is $O(n^3)$. □

Universal traversal sequences

Let G be graph on n nodes, where the degree of any node is d. Suppose at each node, the edges are labeled $1 \ldots d$. We would like to define a sequence $(h_1, h_2 \ldots h_t) \subseteq \{1, \ldots, d\}^t$ such that if we start from any node in the graph, taking the h_ith edge out of the node at step i ensures that at time t, we have visited all the nodes in the graph. A universal traverse sequence (UTS) for n, d is a sequence that, for all d-regular graphs on n nodes, and for every labelling of it, starting from any start vertex, and walking according to the

sequence and the labelling, we will have touched each node in the graph. We prove that such sequences exist.

Theorem 10. *There exists a UTS of size $O(d^2n^3 \log n)$.*

Proof. The proof relies on the probabilistic method. Consider a random sequence of length $2dn^2(dn+2\log n)$. In particular, let $s_1, \ldots, s_n$ be $t = dn+2\log n$ subsequences of length $2dn^2$ each. Next, consider any d-regular graph, and any start vertex v. Note that the expected cover time for a d-regular graph is at most dn^2, using Theorem 9. So the probability that the cover time is at most twice its expectation is at least $\frac{1}{2}$, and hence the probability that a particular s_i (the length of s_i is twice the expected cover time) covers G for a given v is at least $\frac{1}{2}$. Thus, the probability that one of the $s_i's$ covers G is at least $1 - \frac{1}{2^t}$.

Define the random variable $X_{G,i}$ to be 0 if G is covered by sequence s_i, and 1 otherwise. Then we have:

$$\begin{aligned} \Pr[X_{G,i}] &\leq \frac{1}{2^t} \\ \mathrm{E}[\sum_{G,i} X_{G,i}] &\leq \#\text{ of graphs (n,d)} \cdot \#\text{ of starting points} \cdot \frac{1}{2^t} \\ &\leq n^{nd} \cdot n \cdot \frac{1}{n^{nd+2}} = \frac{1}{n} < 1 \end{aligned}$$

Since the expectation is less than 1 and this is a positive integer-valued random variable, there exists a sequence for which the value is 0. □