CS 4850 - Mathematical Foundations March 3 and 5, 2010

Lecture 16 and 17: Random Walks

Professor: John Hopcroft *Scribe: Matt Paff and Matt Weinberg*

1 Harmonic Functions on a Graph

Let $G = (V, E)$ be a connected graph. Fix a partition of the vertices into two sets, the "internal" nodes and the "external" nodes. Mark each external node v with a number n_v. For each internal node u, specify a weighted sum in terms of the numbers on that node's neighbors. Denote this weighted sum as $f_u : \mathbb{R}^{d_u} \to \mathbb{R}$, where d_u is the degree of u. Then a *harmonic function* on G is a function $\varphi : V \to \mathbb{R}$ such that:

- For all external nodes v, $\varphi(v) = n_v$. So the numbers marked on the external nodes can be viewed like boundary conditions in differential equations.
- For all internal nodes u, $\varphi(u) = f_u(\varphi(x_1), \ldots, \varphi(x_{d_u}))$ where $x_1, \ldots, x_{d_u}$ are the neighbors of u.

So the external nodes are fixed, but an internal node is only fixed once all of the values of its neighbors are chosen. We will now present three lemmas which will be useful later.

Lemma 1: Let u be any internal node, let N_u be the neighbors of u, and let φ be a harmonic function. Then:

$$\min_{x \in N_u} \varphi(x) \leq \varphi(u) \leq \max_{x \in N_u} \varphi(x)$$

In words, the value of u is not less than the smallest value of its neighbors, and the value of u is not greater than the biggest value of its neighbors.

Proof: The follows direction from the fact that the value of u is a weighted sum of the values of its neighbors. Weighted sums cannot possibly be either smaller than or greater than all of the terms.

Lemma 2: A harmonic function takes on its max and min on the external nodes. [Note, an internal node may also have the same max/min value - this is simply saying that there cannot be an internal node with a smaller (or larger) value than every external node.]

Proof: This follows from the following fact: Since weighted sums require that each weight is positive, the value of an internal node is equal to the minimum (or maximum) of the values of its neighbors if and only if all of its neighbors have the same value. Using this fact, a simple induction on the shortest path from u to any external node gives us that there must exist an external node with the same value.

Lemma 3: Harmonic functions are unique.

Proof: Suppose f, g are two harmonic functions. Let $h = f - g$. h is a harmonic function on the same graph, just with a different set of boundary conditions. But f and g agree on the boundary, so h is 0 everywhere on the boundary. By lemma 2, this implies that h is 0 on all internal nodes as well, so h is identically 0. Therefore, f and g must agree on every node, so $f = g$.

2 Random Walks and Electrical Networks

We can view a graph as an electrical network by viewing each edge as a resistor. For each edge (x, y), let R_{xy} be the resistance on that edge, I_{xy} the current, V_{xy} the voltage, and C_{xy} the conductance (so by definition $C_{xy} = 1/R_{xy}$). For any node x, let $C_x = \sum_{y \in N_x} C_{xy}$, where as above N_x is the set of neighbors of x. Then let the probability of taking edge (x, y) when you are at vertex x to be $P_{xy} = C_{xy}/C_x$. Note that P_{xy} is not necessarily equal to P_{yx} since C_x may not be equal to C_y.

Fix two nodes a and b, and attach a power source to them so that $V_a = 1$ and $V_b = 0$ (where V_x is the voltage at node x). So current will start running through the network.

Claim 1: V_x = the probability of reaching vertex a before b when starting a random walk at x.

Proof: To prove this claim, we will show that both sides of that equation satisfy the same harmonic function on G, and hence must be equal by lemma 3.

First consider the voltages. The boundary conditions on the voltages are $V_a = 1$ and $V_b = 0$. From electrical engineering, we know that for any edge (x, y):

$$I_{xy} = (V_x - V_y)C_{xy}$$

And for any node x other than a, b:

$$\sum_{y \in N_x} I_{xy} = 0$$

Therefore, for any node $x \neq a, b$, we have:

$$\sum_{y \in N_x} (V_x - V_y)C_{xy} = 0$$

From this, we get the following chain of equalities:

$$V_x \sum_{y \in N_x} C_{xy} = \sum_{y \in N_x} V_y C_{xy}$$
$$V_x C_x = \sum_{y \in N_x} V_y C_{xy}$$

And then solving for V_x, noting that $P_{xy} = C_{xy}/C_x$ by definition, we have:

$$V_x = \sum_{y \in N_x} V_y P_{xy}$$

Since $\sum_{y \in N_x} P_{xy} = 1$, this is a weighted sum in terms of the voltages of x's neighbors, so the voltages do in fact satisfy a harmonic function on G.

Now consider the probabilities. Let P_x be the probability that a random walk starting at x reaches a before b. Obviously, $P_a = 1$ and $P_b = 0$, so these probabilities satisfy the same boundary conditions as the voltages. Also, $P_x = \sum_{y \in N_x} P_{xy} P_y$ is obvious from the definition of P_x. If you start the random walk at x, your next location will be some $y \in N_x$, chosen randomly with probabilities P_{xy}. So that equality trivially follows. Hence, the P_x's satisfy the same harmonic function as the voltages, and thus by lemma 3, $P_x = V_x$ for all x, as claimed.

Now we will consider a probabilistic interpretation of current.

Claim 2: I_{xy} = the expected net frequency a random walk starting at a goes along the (x, y) edge from $x \to y$ before reaching b ("net frequency" means the number of times we move from x to y along the (x, y) edge minus the number of times we move from y to x along the (x, y) edge).

Proof: The proof of this claim will be similar to the first claim. Let μ_x be the expected number of visits to x on a walk from a to b before reaching b. For $x \neq a, b$, it is clear that:

$$\mu_x = \sum_{y \in N_x} \mu_y P_{yx}$$

This holds since every time you visit x, you must have just visited some $y \in N_x$ and gone along the (x, y) edge to x. This looks like the same weighted sum as before, except with P_{yx} instead of P_{xy}. Recall that those are not necessarily equal, however we do have:

$$P_{yx} = \frac{C_{yx}}{C_y} = \frac{C_{xy}}{C_y} = P_{xy}\frac{C_x}{C_y}$$

So then we get that:

$$\mu_x = \sum_{y \in N_x} \mu_y P_{xy} \frac{C_x}{C_y}$$

Which implies:

$$\frac{\mu_x}{C_x} = \sum_{y \in N_x} \frac{\mu_y}{C_y} P_{xy}$$

So μ_x/C_x is harmonic. $\mu_b/C_b = 0$ since $\mu_b = 0$, but μ_a/C_a is not necessarily 1. So this satisfies different boundary conditions than the voltages above. However, we can simply change the voltage on a (since we are determining that anyway using the power source), so set $V_a = \mu_a/C_a$ instead of 1. Then $V_x = \mu_x/C_x$ for all x since they satisfy the same harmonic function. Then using the equation $I_{xy} = (V_x - V_y)C_{xy}$ from electrical engineering, we have:

$$\begin{aligned}
I_{xy} &= (V_x - V_y)C_{xy} \\
&= \left(\frac{\mu_x}{C_x} - \frac{\mu_y}{C_y}\right) C_{xy} \\
&= \mu_x P_{xy} - \mu_y P_{yx} \\
&= E[\#\text{ of traversals from } x \to y] - E[\#\text{ of traversals from } y \to x] \\
&= E[\#\text{ of net traversals from } x \to y]
\end{aligned}$$

This is exactly the equality we claimed, completing the proof.

Now we will express the "escape probability" in terms of the electrical network. The escape probability is defined as the probability that a random walk starting at a reaches b before returning to a, and is denoted P_{esc}.

We can view the whole network as one big resistor between a and b. So define V_{ab} as the difference in voltage between a and b, and I_{ab} to be the total amount of current running through the whole network. Then define "R-effective" to be $R_{\text{eff}} = \frac{V_{ab}}{I_{ab}}$. Further, define $C_{\text{eff}} = 1/R_{\text{eff}}$.

From electrical engineering, since a is the only source of current in the graph and all current must exit the graph at node b, we know that:

$$I_{ab} = \sum_{y \in N_a} I_{ay} = \sum_{y \in N_a} (V_a - V_y)C_{ay}$$

Set $V_a = 1$ and $V_b = 0$ (so $V_{ab} = 1$), and then we get:

$$\begin{aligned}
I_{ab} &= \sum_{y \in N_a} (1 - V_y)C_{ay} \\
&= \sum_{y \in N_a} C_{ay} - C_a \sum_{y \in N_a} V_y \frac{C_{ay}}{C_a} \\
&= C_a - C_a \sum_{y \in N_a} V_y P_{ay} \\
&= C_a \left(1 - \sum_{y \in N_a} P_{ay} \Pr\left[\text{random walk starting at } y \text{ reaches } a \text{ before } b\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= C_a \left(1 - \Pr\left[\text{random walk starting at } a \text{ returns to } a \text{ before reaching } b\right]\right) \\
&= C_a \Pr\left[\text{random walk starting at } a \text{ reaches } b \text{ before returning to } a\right] \\
&= C_a P_{\text{esc}}
\end{aligned}
$$

Combining this with $R_{\text{eff}} I_{ab} = V_{ab} = 1$, we get $1/R_{\text{eff}} = I_{ab} = C_a P_{\text{esc}}$, which implies:

$$P_{\text{esc}} = \frac{1}{R_{\text{eff}} C_a} = \frac{C_{\text{eff}}}{C_a}$$

Here's some intuition concerning this formula. Conductance is a measure of how easily electricity flows through a circuit. So C_a is a measure of how easily electricity in the rest of the graph will flow back to a. And C_{eff} is a measure of how easily electricity will flow from a to b. So it makes sense that the higher C_{eff} is, the higher the escape probability, and the higher C_a is, the lower the escape probability.

3 Page Rank

The PageRank system created a billion dollar industry to improve a webpage's PageRank. We will consider some of the ways to do so, and how to counter them.

Recall a webpage's PageRank is dependent on the number of incoming links, and incoming links from "important" websites (ie, ones with high PageRanks) count more. Google uses random walks around the internet to compute the PageRanks.

First, suppose a webpage's PageRank were simply determined by how many times the random walk goes to that website. Here are two simple ways to improve the PageRank of one's page:

1. "Capture" the random walk: If a webpage has no outgoing edges, then Google detects that and restarts the random walk at another vertex. So simply deleting all of the outgoing links doesn't work. And Google also completely ignores self loops. However, you can create a dummy site and have your website only link to the dummy site, and have the dummy site only link to your site. Then once the random walk enters your site, it will keep switching back and forth between your site and the dummy site, thereby increasing both sites' PageRanks.

2. "Spam Farm": Purchase old urls that had decent PageRanks, and add links on them to your website.

One way to counter the first is the following: Instead of counting the number of times a random walk is in a site, count the expected "hitting time" (which is definied as the amount of time it takes to reach the site when starting at a website picked uniformaly at random). The "capture" technique obviously has no effect anymore. We will continue to discuss this technique next lecture.