PageRank Notes

August 8, 2019

1 View 1: Random Walks

Imagine a walk that goes to a random neighbor with probability $(1-\alpha)$, and returns to node u with probability α. The stationary distribution is:

$$\mathbf{p}_u = \alpha\chi_u + (1-\alpha)\mathbf{W}\mathbf{p}_u \iff \mathbf{p}_u = (\mathbf{I} - (1-\alpha)\mathbf{W})^{-1}\alpha\chi_u \tag{1}$$

The inverse exists because $\mathbf{W}$ has eigenvalues between -1 and 1.

2 View 2: Spilling Paint

Imagine we start off with 1 unit of paint at node u. In each step, α fraction of the paint dries, while $(1-\alpha)$ fraction takes a *lazy* random walk: stay with probability 1/2 and go to a random neighbor with probability 1/2. Let $\widehat{\mathbf{W}}$ be the lazy random walk, $\mathbf{s}^t$ denote the dried paint at time t ($\mathbf{s}^0 = \chi_u$), and $\mathbf{r}^t$ denote the wet paint at time t. We have

$$\begin{aligned}\mathbf{s}^{t+1} &= \mathbf{s}^t + \alpha\mathbf{r}^t \\ \mathbf{r}^{t+1} &= (1-\alpha)\widehat{\mathbf{W}}\mathbf{r}^t\end{aligned}$$

We can solve for $\mathbf{s}^\infty$:

$$\mathbf{s}^\infty = \sum_{t=0}^{\infty} \alpha(1-\alpha)^t \widehat{\mathbf{W}}^t \chi_u = \alpha\left(\mathbf{I} - (1-\alpha)\widehat{\mathbf{W}}\right)^{-1}\chi_u$$

So s^∞ is the same as p_u up to the difference between $\mathbf{W}$ and $\widehat{\mathbf{W}}$. **If we view the paint particles as executing random walks in parallel, then the connection between the two views makes sense.** With this view, we can even ensure $\mathbf{W} = \widehat{\mathbf{W}}$ by changing the value of α in View 2 to some β. Essentially, we set β so that a particle of paint has probability α of drying before moving to a neighbor (that is, chose lazy random walk and actually moved instead of staying in place). So from now on, we always use $\mathbf{W}$ instead of $\widehat{\mathbf{W}}$.

3 Local Updates

How do we compute $p_u = \mathbf{s}^\infty$? It turns out that View 2 is more helpful, and allows us to **compute it completely asynchronously**. In particular, initialize $\mathbf{s} \leftarrow \mathbf{0}$, $\mathbf{r} \leftarrow \chi_u$ and keep on performing the following three updates simultaneously for different values of u:

$$\begin{aligned}\mathbf{s}(u) &\leftarrow \mathbf{s}(u) + \alpha\mathbf{r}(u) \\ \mathbf{r}(u) &\leftarrow \mathbf{0} \\ \mathbf{r}(v) &\leftarrow \mathbf{r}(v) + \frac{1-\alpha}{d(u)}\mathbf{r}(u)\end{aligned}$$

The claim is that we always maintain

$$\mathbf{p}_u = \mathbf{s} + \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \mathbf{W}^t \mathbf{r}$$

This can be proven with the paint particles simulating parallel random walks analogy.

4 PageRank Analysis

For this entire section, fix the starting vertex u, and set $\mathbf{p} := \mathbf{p}_u$. Define

$$\mathbf{q}(v) := \frac{\mathbf{p}(v)}{d(v)}$$

It makes sense to normalize by $d(v)$, since higher-degree nodes will naturally have more stationary probability. Let us order the vertices $1, 2, \ldots, n$ so that

$$\mathbf{q}(1) \geq \mathbf{q}(2) \geq \cdots \geq \mathbf{q}(n)$$

We will regularly use the set $[k] = \{1, 2, \ldots, k\}$, the k vertices with highest $\mathbf{q}$ values.

Lemma 1. *For every k,*

$$\sum_{i \leq k < j} \mathbf{q}(i) - \mathbf{q}(j) \leq \alpha \tag{2}$$

Proof. Intuition: when $\alpha = 0$, the stationary distribution of v is proportional to $d(v)$, so all $\mathbf{q}(v)$ are equal, so expression (2) is 0. (2) represents the *asymmetric imbalance* of the random walk, which is "at most" α because that's the probability of traveling the directed edge back to u.

$$\begin{aligned} \chi_{[k]}^T \mathbf{W}\mathbf{p} &= (\text{volume initially in } [k]) - (\text{change in volume going out of } [k]) \\ &= \chi_{[k]}^T \mathbf{p} - \sum_{(i,j) \in E, i \in [k], j \notin [k]} (\mathbf{q}(i) - \mathbf{q}(j)) \end{aligned}$$

Also,

$$\begin{aligned} \mathbf{p} \overset{(1)}{=} \alpha \chi_u + (1-\alpha)\mathbf{W}\mathbf{p} \leq \alpha \chi_u + \mathbf{W}\mathbf{p} &\iff \mathbf{W}\mathbf{p} \geq \mathbf{p} - \alpha \chi_u \\ \implies \chi_{[k]}^T \mathbf{W}\mathbf{p} \geq \chi_{[k]}^T \mathbf{p} - \alpha \chi_{[k]}^T \chi_u &\geq \chi_{[k]}^T \mathbf{p} - \alpha \end{aligned}$$

Combining the two proves the lemma. □

Lemma 2. *If $\phi([j]) \geq 2\theta$ for some θ, then there exists $k > j$ with*

$$d([k]) \geq (1+\theta) d([j]) \qquad \text{and} \qquad \mathbf{q}(k) \geq \mathbf{q}(j) - \frac{\alpha}{\theta d([j])}. \tag{3}$$

Proof. Let $k > j$ be the smallest with $d([k]) \geq (1+\theta)d([j])$. There are $< \theta d([j])$ many edges between a vertex in $[j]$ and one in $[j+1, k-1]$. Since $\phi([j]) \geq 2\theta$, there are $\geq 2\theta d([j])$ many edges out of $[j]$. So there are $\geq 2\theta d([j]) - \theta d([j]) = \theta d([j])$ many edges from $[j]$ to $[k, n]$. Since $\mathbf{q}$ is decreasing,

$$\sum_{\substack{(a,b) \in E \\ a \leq j,\, b > k}} (\mathbf{q}(a) - \mathbf{q}(b)) \geq \sum_{\substack{(a,b) \in E \\ a \leq j,\, b > k}} (\mathbf{q}(j) - \mathbf{q}(k)) \geq \theta d([j])(\mathbf{q}(j) - \mathbf{q}(k))$$

Intuition: $\sum_{i\le k<j} \mathbf{q}(i) - \mathbf{q}(j)$ is always bounded by α, but we have lots of edges, so the difference of α is spread out over many edges. We also have

$$\sum_{\substack{(a,b)\in E\\ a\le j,\, b>k}} (\mathbf{q}(a) - \mathbf{q}(b)) \le \sum_{\substack{(a,b)\in E\\ a\le j,\, b>j}} (\mathbf{q}(a) - \mathbf{q}(b)) \overset{(2)}{\le} \alpha$$

Therefore,

$$\theta d([j])(\mathbf{q}(j) - \mathbf{q}(k)) \le \alpha \iff \mathbf{q}(k) \ge \mathbf{q}(j) - \frac{\alpha}{\theta d([j])}.$$

□

Lemma 3. *Suppose $\phi(G) \ge \Omega(\theta)$. Let h be smallest s.t. $d([h]) \ge 2m/3$. For every $i \le h$:*

$$\mathbf{q}(h) \ge \mathbf{q}(i) - \frac{2\alpha}{\theta^2 d([i])}.$$

This means that we can start with a large conductance set and blow it up to a set $[h]$ of volume $\Theta(m)$, while making sure $\mathbf{q}(h) \approx \mathbf{q}(i)$.

Proof. Apply Lemma 2 repeatedly starting from i until we reach h. This ensures that $d([i])$ grows geometrically, so that we don't lose too much in the bound $\mathbf{q}(k) \ge \mathbf{q}(j) - \frac{\alpha}{\theta d([j])}$ from (3). We have

$$\begin{aligned}
\mathbf{q}(h) &\ge \mathbf{q}(i) - \frac{\alpha}{\theta d([i])} - \frac{\alpha}{\theta(1+\theta)d([i])} - \frac{\alpha}{\theta(1+\theta)^2 d([i])} - \cdots \\
&= \mathbf{q}(i) - \frac{\alpha}{\theta d([i])} \cdot \left(1 + \frac{1}{1+\theta} + \frac{1}{(1+\theta)^2} + \cdots\right) \\
&= \mathbf{q}(i) - \frac{\alpha}{\theta d([i])} \cdot \frac{1+\theta}{\theta} \\
&\ge \mathbf{q}(i) - \frac{2\alpha}{\theta^2 d([i])}.
\end{aligned}$$

Note that this is where we pick up another factor θ, resulting in the square-root in PageRank approximation.

□

Suppose we run PageRank so that there is a *small* set S, $d(S) \le O(m/\log m)$, with large probability mass, say, $\mathbf{p}(S) \ge 2/3$. We show that we can find a small conductance cut.

Lemma 4. *For any set S with $\mathbf{p}(S) \ge 2/3$, let j be smallest with $d([j]) \ge d(S)$. There is an $i \in [j]$ s.t.*

$$\mathbf{q}(i) \ge \frac{2/3}{d([i])H(2m)}.$$

Proof. This is a typical harmonic-series-type bound. Suppose not: $\mathbf{q}(i) = \frac{\mathbf{p}(i)}{d(i)} < \frac{2/3}{d([i])H(2m)}$ for all i. Sum $d(i)$ copies of $\mathbf{q}(i)$ for each i with $d([i]) \le d(S)$:

$$\mathbf{p}(S) \le \mathbf{p}([j]) = \sum_{i\le j} \mathbf{p}(i) = \sum_{i\le j} d(i)\mathbf{q}(i) < \sum_{i\le j} d(i)\frac{2/3}{d([i])H(2m)} \le \sum_{\ell=1}^{2m} \frac{2/3}{\ell \cdot H(2m)} = 2/3,$$

contradicting the assumption that $\mathbf{p}(S) \ge 2/3$. □

Lemma 5. *Suppose there is S with $d(S) \le O(m/\log m)$ and $\mathbf{p}(S) \ge 2/3$. Then, we can find a low-conductance cut. In particular, there is a set $[j]$, $j \in [n]$, with $\phi([j]) \le \sqrt{6\alpha H(2m)}$.*

Proof. Suppose not. Define $\theta := \sqrt{6\alpha H(2m)}$ and suppose that $\phi([j]) > \theta$ for all $j \in [n]$. Pick i as in Lemma 4, so that

$$\mathbf{q}(i) \geq \frac{2/3}{d([i])H(2m)}.$$

Define h as in Lemma 3, and apply Lemma 3 to i, so that $d([h]) \geq 2m/3$ and $\mathbf{q}(h) \geq \mathbf{q}(i) - \frac{2\alpha}{\theta^2 d([i])}$. We have

$$\mathbf{q}(h) \geq \mathbf{q}(i) - \frac{2\alpha}{\theta^2 d([i])} = \frac{2/3}{d([i])H(2m)} - \frac{2\alpha}{6\alpha H(2m)d([i])} = \frac{1}{3H(2m)d([i])}.$$

Since $\mathbf{q}$ is decreasing,

$$1 \geq \mathbf{p}([h]) = \sum_{i \in [h]} \mathbf{p}(i) = \sum_{i \in [h]} d(i)\mathbf{q}(i) \overset{\mathbf{q}\text{ decr.}}{\geq} \sum_{i \in [h]} d(i)\mathbf{q}(h) = d([h])\mathbf{q}(h) \geq \frac{2m}{3} \cdot \frac{1}{3H(2m)d([i])}$$
$$\implies d([i]) \geq \Theta(m/\log m)$$

By choice of i in Lemma 4, we know that $d([i-1]) \leq d(S) \leq O(m \log m)$. Let's also assume that degrees are not extremely large: $\deg(i) \leq o(m/\log m)$. This means that

$$d([i-1]) \approx d([i]) \geq \Theta(m/\log m),$$

contradiction. □

Q: What about when $d(S) = \Theta(m)$, like in our case?