10/8: Random Walks (cont.)

Outline

- We've been covering basic random walks on graphs, following Doyle and Snell's "Random walks and electric networks": `http://front.math.ucdavis.edu/math.PR/0001057`
- Today we'll look at Google's PageRank algorithm and related techniques.

Recap

Recall that solving the boundary hitting probability problem (Dirichlet problem) led to

$$\begin{aligned} Pf &= f \\ \begin{pmatrix} I & 0 \\ R & Q \end{pmatrix} \begin{pmatrix} f_B \\ f_D \end{pmatrix} &= \begin{pmatrix} f_B \\ f_D \end{pmatrix} \end{aligned}$$

with the matrix solution

$$f_D = (I - Q)^{-1} R f_B$$

Stationary Distribution

The stationary distribution of an ergodic Markov chain satisfies

$$\pi P = \pi$$

Or, π is principal eigenvector of $P^{\top}$.

Product Method

Since

$$\begin{aligned} \pi P &= \pi P^2 \\ &= \pi P^n \\ &\cdots \\ &\longrightarrow \pi \end{aligned}$$

Product Method

Suggests an iterative algorithm

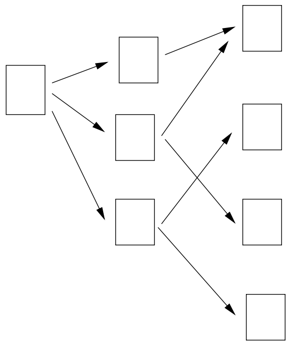
$$\begin{aligned} \pi^{(0)} &= \text{arbitrary} \\ \pi^{(k+1)} &= \pi^{(k)} P \end{aligned}$$

Called the *product method*. If the chain is ergodic, will converge to π (geometrically)

PageRank

(Brin and Page, 1998)

This is the basis for the famous *PageRank* algorithm.



Why not just count incoming links?

PageRank

Linkage matrix:

$$A_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

Note A is not symmetric. Normalize A to get M, stochastic.

Random Surfing

- Choose random page from set of pages current page points to
- With probability $(1-\epsilon)$ move to that page
- With probability ϵ, move to a random page

$$P = (1-\epsilon)M + \epsilon U$$

where $U_{ij} = \frac{1}{n}$.

Product Method

Want π so that

$$((1-\epsilon)M + \epsilon U)^\top \pi = \pi$$

Note: chain is not reversible, so can't read off π. But is ergodic.

$$\pi^{(n)} = (1-\epsilon)M^\top \pi^{(n-1)} + \frac{\epsilon}{n}(1,1,\ldots,1)^\top$$

Product Method

Sparse matrix multiply,

$$\begin{aligned}(M^\top \pi)_k &= \sum_j M^\top_{kj}\pi_j \\ &= \sum_j M_{jk}\pi_j\end{aligned}$$

Google and PageRank

- The PageRank score for a web page i is π_i
- The method for combining π_i with content score for query is a closely guarded secret! (Why?)
- (Show PageRank scores)

HITS Algorithm

(Kleinberg, 1998)

Good *authority* is page linked to by pages with high hub scores

Good *hub* is page that links to pages with high authority scores

HITS Algorithm

From matrix

$$A_{ij} = \begin{cases} 1 & \text{if page } i \text{ links to page } j \\ 0 & \text{otherwise} \end{cases}$$

we define

$$\begin{aligned} a_i^{(t+1)} &= \sum_{j:j\to i} h_j^{(t)} \\ h_i^{(t+1)} &= \sum_{j:i\to j} a_j^{(t+1)} \end{aligned}$$

HITS Algorithm – Matrix form

$$\begin{aligned} a^{(t+1)} &= A^\top h^{(t)} = (A^\top A)a^{(t)} \\ h^{(t+1)} &= Aa^{(t+1)} = (AA^\top)h^{(t)} \end{aligned}$$

So:

$a = \lim_{t\to\infty} a^{(t)}$ is principal eigenvector of $A^\top A$

$h = \lim_{t\to\infty} h^{(t)}$ is principal eigenvector of $AA^\top$.

HITS vs. PageRank

Usually HITS is viewed as being run on a small collection, and PageRank on a large collection, but this is not essential.

Back to PageRank

Without random jumping (prob ϵ), *transient* vertices would have $\pi_i = 0$, or there might not be unique solution.

However, PageRank satisfies $\pi_i \geq \frac{\epsilon}{n}$.

Interpretation from Bayes Rule

Can interpret PageRank as a query-independent prior on pages being of interest to the user

$$
\begin{aligned}
\mathrm{Pr}(\mathrm{doc} = d \,|\, \mathrm{query} = q) &= \frac{\mathrm{Pr}(\mathrm{query} = q \,|\, \mathrm{doc} = d)\,\mathrm{Pr}(\mathrm{doc} = d)}{\mathrm{Pr}(\mathrm{query} = q)} \\
&\propto \mathrm{Pr}(\mathrm{query} = q \,|\, \mathrm{doc} = d)\,\mathrm{Pr}(\mathrm{doc} = d)
\end{aligned}
$$

Simplest case: $\mathrm{Pr}(\mathrm{doc} = d) = \pi_d$ and

$$
\mathrm{Pr}(\mathrm{query} = q \,|\, d) = \frac{1_{\text{doc matches query}}}{\sum_{q'} 1_{\text{doc matches query } q'}}
$$

Generalized PageRank

Suppose that

$$
P(x, y) = \epsilon\mu(y) + (1 - \epsilon)Q(x, y)
$$

Convergence is geometric:

$$
\begin{aligned}
\|P^n(x, \cdot) - \pi(\cdot)\|_1 &= \sum_y |P^n(x, y) - \pi(y)| \\
&\leq C\epsilon^n
\end{aligned}
$$

for arbitrary x.

(In fact, much faster convergence for expander graphs.)

Generalized PageRank

Expression for stationary distribution:

$$
\begin{aligned}
\pi(y) &= \epsilon \sum_{k=0}^{\infty} (1 - \epsilon)^k \sum_{x \in G} \mu(x) Q^k(x, y) \\
\pi &= \epsilon\mu(I - (1 - \epsilon)Q)^{-1}
\end{aligned}
$$

PageRank Data Bias

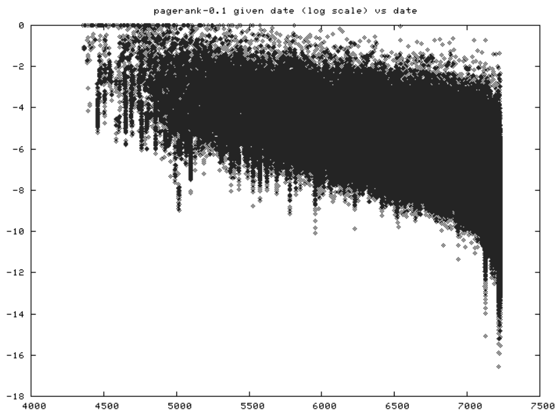
PageRank tends to favor older pages.

Why?

How to remove this bias?

PageRank Date Bias

(Breyer, 2002)



PageRank Date Bias

Let τ_x be the time stamp of page x. The smaller the stamp, the older the page is.

Define a new transition function:

$$Q_D(x, y) = \frac{1}{C_\lambda(x)} \exp\left(-\lambda \left(\frac{\tau_x - \tau_y}{\tau_{\text{max}} - \tau_{\text{min}}}\right)_+\right)$$

where $(a)_+ = a$ if $a \geq 0$ and $= 0$ otherwise.

Stability Issues

If the graph changes slightly, we'd like for the rankings not to change dramatically.

[Sample rankings]

Stability Issues

For PageRank, $\epsilon U + (1-\epsilon)M$, as long as ϵ is not too small, and modified pagesdo not have high overall PageRank scores, the perturbed PageRank scores $\tilde{\pi}$ will not be too far from the original scores π.

Stability

(Ng, Zheng and Jordan, 2001)

Suppose pages with index

$$i_1, i_2, \ldots, i_k$$

are changed. Let $\tilde{M}$ be the new transition matrix. Then

$$\|\tilde{\pi} - \pi\|_1 \leq \frac{2\sum_{j=1}^{k} \pi_{i_j}}{\epsilon}$$

Stability (cont.)

Idea: construct two chains X_t and Y_t so that

- Start out in the same state
- Resets occur in lock step
- Transitions are "correlated"
- Different only when one was at a "changed" document on previous time step
- Asymptotic distributions are as desired, π and $\tilde{\pi}$.

Then we'll show that as t gets large, the probability that X_t and Y_t differ is not very large.

Stability (cont.)

Construct "coupled" Markov chain (X_t, Y_t) with

$$X_0 = Y_0 \sim \pi$$

On step t, with probability ϵ reset *both* chains to the *same* page.

With probability $1-\epsilon$:

If $X_{t-1} = Y_{t-1}$ is not one of the changed documents, then $X_t = Y_t$ according to M.

Otherwise, X_t and Y_t chosen independently, according to the proper distribution (M or $\tilde{M}$).

Stability (cont.)

Then

$$
\begin{aligned}
X_t &\sim \epsilon U + (1-\epsilon) M \\
Y_t &\sim \epsilon U + (1-\epsilon) \tilde{M}
\end{aligned}
$$

with stationary distributions π and $\tilde{\pi}$. (Note: X_t already starts in the stationary distribution, so stays there.)

Stability (cont.)

Let $d_t = \Pr(X_t \neq Y_t)$. Then

$$
\begin{aligned}
d_{t+1} &= \Pr(X_{t+1} \neq Y_{t+1}) \\
&= \epsilon \Pr(X_{t+1} \neq Y_{t+1} \,|\, \text{reset at } t+1) + \\
&\quad (1-\epsilon)\Pr(X_{t+1} \neq Y_{t+1} \,|\, \text{don't reset at } t+1) \\
&= (1-\epsilon)\Pr(X_{t+1} \neq Y_{t+1} \,|\, \text{don't reset at } t+1) \\
&= (1-\epsilon)\left[\Pr(X_{t+1} \neq Y_{t+1}, X_t \neq Y_t \,|\, \text{don't reset at } t+1) + \right. \\
&\quad \left. + \Pr(X_{t+1} \neq Y_{t+1}, X_t = Y_t \,|\, \text{don't reset at } t+1)\right]
\end{aligned}
$$

Stability (cont.)

Let $d_t = \Pr(X_t \neq Y_t)$. Then

$$
\begin{aligned}
d_{t+1} &= \Pr(X_{t+1} \neq Y_{t+1}) \\
&= (1-\epsilon)\left[\Pr(X_{t+1} \neq Y_{t+1}, X_t \neq Y_t \,|\, \text{don't reset at } t+1) \right. \\
&\quad \left. + \Pr(X_{t+1} \neq Y_{t+1}, X_t = Y_t \,|\, \text{don't reset at } t+1)\right] \\
&\leq (1-\epsilon)\left[\Pr(X_t \neq Y_t \,|\, \text{don't reset at } t+1) + \right. \\
&\quad \left. + \Pr(X_{t+1} \neq Y_{t+1}, X_t = Y_t, X_t \in \mathcal{P} \,|\, \text{don't reset at } t+1)\right] \\
&\leq (1-\epsilon)\left(\Pr(X_t \neq Y_t) + \Pr(X_t \in \mathcal{P})\right) \\
&= (1-\epsilon)(d_t + \sum_{i \in \mathcal{P}} \pi_i)
\end{aligned}
$$

where $\mathcal{P} = \{i_1, i_2, \ldots, i_k\}$.

Stability (cont.)

Thus, if (X_∞, Y_∞) is drawn from the stationary distribution of the corresponding chains, then

$$
\Pr(X_\infty \neq Y_\infty) \leq \frac{\sum_{i \in \mathcal{P}} \pi_i}{\epsilon}
$$

Coupling Lemma

Suppose the $X \sim p$ and $Y \sim q$. Then

$$\frac{1}{2}\|p-q\|_1 \;\leq\; \Pr(X \neq Y)$$

Proof:

$$\begin{aligned}
\|p-q\|_1 &= \sum_x |p(x)-q(x)| \\
&= \sum_x |p(x)(1-q(x)) - q(x)(1-p(x))| \\
&\leq \sum_x |p(x)(1-q(x))| + \sum_x |q(x)(1-p(x))| \\
&= 2\Pr(X \neq Y)
\end{aligned}$$

Stability (cont.)

So, from the coupling lemma and the previous calculation, have that

$$\begin{aligned}
\|\pi - \tilde{\pi}\| &\leq 2\Pr(X_\infty \neq Y_\infty) \\
&\leq \frac{2\sum_{i\in\mathcal{P}} \pi_i}{\epsilon}
\end{aligned}$$

Gives an probabilistic explanation of the stability of the PageRank algorithm.

The HITS algorithm suffers from "instability." However, can easily stabilize it by introducing a "reset" mechanism similar to PageRank's.