MS&E 233 Lectures 6 and 7: Reputation Systems and PageRank

Ashish Goel, scribed by RIley Matthews

April 18

1 Lecture 6

- **Why do we (now) need reputation systems?**
  - Earlier: centralized reputation systems, e.g., NYT editor deciding what is fit to print
  - Now: decentralized creation demands decentralized curation
  - Need to connect and interact with complete strangers
  - Examples of reputation systems: amazon, yelp, quora, google search
    - Consider google search: Huge number of pages returned for a given string query – the search engine then orders them using both relevance and "reputation". The information behind this reputation is infact largely "crowdsourced" from the network of content creators and web-surfers, who implicitly reveal important pages by linking to or clicking on them.
- **How to measure/define reputation?**
  - Working Measure: *You are highly reputed if other highly reputed individuals think highly of you.*
    - Problem? This definition is cyclic/ self-referential. But so are real reputations!
  - Let's apply this measure to web pages:
    - Let the graph in figure 1 represent a realy tiny web graph. An edge from one node to another indicates that the first web-page links to the second. One can use the same basic idea to model social networks, where an edge from a node to another indicates that the first user is following the second. Let us use $\pi(X)$ to denote the reputation of node X. Assume $\pi(X) \geq 0$ for all X. As our first attempt to apply the informal recursive definition, consider the following simple set of equations:

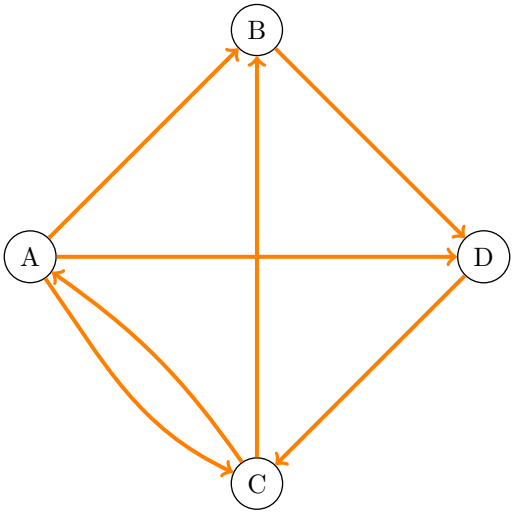

Figure 1: Simple Network

$$\pi(D) = \pi(A) + \pi(B)$$
$$\pi(B) = \pi(A) + \pi(C)$$
$$\pi(C) = \pi(A) + \pi(D)$$
$$\pi(A) = \pi(C)$$

Summing all rows yields $2\pi(A) + \pi(C) = 0$ which is absurd; all values are zero or all are ∞. There is no good solution.

* What is conceptually wrong?
  - Which of A and B is expressing stronger preference for D? B! Shouldn't treat A and B's vote the same.
* Now try:

$$\pi(D) = \frac{1}{3}\pi(A) + \pi(B)$$
$$\pi(B) = \frac{1}{3}\pi(A) + \frac{1}{2}\pi(C)$$
$$\pi(C) = \frac{1}{3}\pi(A) + \pi(D)$$
$$\pi(A) = \frac{1}{2}\pi(C)$$

The rationale is that by linking to 3 other nodes, the node A is casting only (1/3)-rd of a vote for each node it links to; the same

Figure 2:

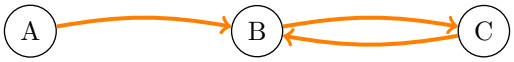

Figure 3:

logic dictates the rest of the terms. Summing theses equations gives $\pi(A)+\pi(B)+\pi(C)+\pi(D) = \pi(A)+\pi(B)+\pi(C)+\pi(D)$, i.e., no contradiction. HOWEVER, there are infinite solutions to these equations.

- · We fix that issue by normalizing the reputations i.e., adding the constraint that $1 = \pi(A)+\pi(B)+\pi(C)+\pi(D)$.
- · Notice that the reputations can then be thought of a probabilities.

* What we've just described is **Naive Pagerank**.
* What is wrong with it? Consider a graph that at first is as stated in figure 2. The PageRanks are given by $\pi(A) = \pi(B) = 1/2$. Then, node B makes a fake webpage C to increase its reputation, as described in figure 3. The PageRanks now become $\pi(A) = 0; \pi(B) = \pi(C) = 1/2$.
  - · Thus, the naive PageRank system can be easily gamed. Both in practice and theory, web-designers gaming the system to increas PageRank can be a big problem[1]
  - · Another problem: consider the graph in figure 4 which is partitioned into two disconnected components.
    Infinite number of solutions again. For example, $\pi(A) = \pi(B) = \pi(C) = 1/3; \pi(D) = \pi(E) = 0$ and $\pi(A) = \pi(B) = \pi(C) = 0; \pi(D) = \pi(E) = 1/2$ are both valid solutions, as is any of their weighted averages. We will come back and address this issue a little later.

– Random Walk on a Graph

* Imagine that we have a random web surfer, or a "PageRank monkey", clicking a link uniformly at random on whichever page it happens to be visiting right not (thus going to a different page).
* New way to think of reputation: $f(X)$ = fraction of time the monkey spends at X in his random surf.
* Over a long period of time we'd expect (referencing figure 1 network):

[1] See "Making eigenvector-based reputation systems robust to collusion" for a detailed discussion. `http://www.stanford.edu/~ashishg/papers/WAW_adapt.ps` .

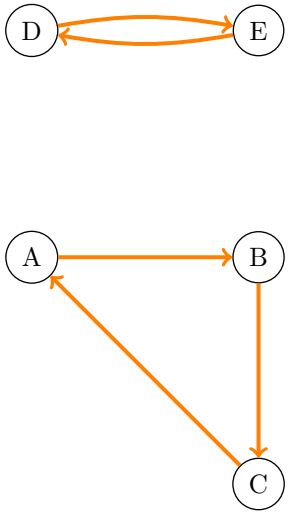

Figure 4:

.

$$f(A) + f(B) + f(C) + f(D) = 1$$

.

$$f(D) = \frac{1}{3}f(A) + f(B)$$
$$f(B) = \frac{1}{3}f(A) + \frac{1}{2}f(C)$$
$$f(C) = \frac{1}{3}f(A) + f(D)$$
$$f(A) = \frac{1}{2}f(C)$$

- These are the same equations as PageRank. Hence, frequencies f defined above are the same as π, and we can use the two interpretations interchangeably.
- As promised earlier, we will now try to address the problems with naive PageRank. At every step, with a small probability ϵ, the monkey jumps uniformly at random to a node in the graph (regardless if there is a link from his current web page).
- This random walk with the occasional random jumps is the final version of the final PageRank definition, specified as the algorihm followed by the PageRank monkey:

– **Final Page Rank Definition**

Currently at page X:

With probability ϵ, jump to a random web page (this is known as "teleportation")
With probability $1-\epsilon$, click on a random hyperlink within page X

- Reputation of web page X: fraction of the time this random surfer monkey spends at X.

2 Lecture 7

- Review
  - Recall from last lecture our abstract measure of reputation: *You are highly reputed if other highly reputed individuals think highly of you*
  - Also recall our discussion of naive PageRank, which we now summarize mathemtically:
    - Naive PageRank:

$$\pi(i) = \sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|}$$

      . Here, out(j) denotes the set of links going out of j.
    - PageRank random (monkey) walk algorithm:
      - Currently at page i:
        With probability ϵ, jump to a random web page ("teleportation")
        With probability $1-\epsilon$, click on a random hyperlink within page i
      - Simulate the above over a large number of iterations and then set $\pi(i)$ = frequency of visits to page i

- New: How should we calculate the $\pi(i)$ values?
  - Note that if we let the random monkey run for a long time we'd expect that for each i

$$\pi(i) = \frac{\epsilon}{N} + (1-\epsilon) \sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|}$$

    where N is the number of nodes in the network.
  - How can we compute these $\pi(i)$'s?
    - Solve the system of equations numerically
    - One can simulate the monkey for a long time (in fact, around n/ϵ suffices, where n is the number of nodes) and see what fraction the monkey spends at each node.
    - Or...Power Iteration

– **Power Iteration**

* Initialize:

$$\pi_0(i) = \frac{1}{N}$$

* Iteration:

$$\pi_{r+1}(i) = \frac{\epsilon}{N} + (1-\epsilon) \sum_{j:j \text{ links to i}} \frac{\pi_r(j)}{|\text{out}(j)|}$$

– What's the difference? We use the r^{th} iteration's π's to compute the $(r+1)^{\text{th}}$ iteration's.

– For example,

$$\pi_1(i) = \frac{\epsilon}{N} + (1-\epsilon) \sum_{j:j \text{ links to i}} \frac{\pi_0(j)}{|\text{out}(j)|}$$

– Problem: as $r \to \infty$, will the $\pi_r(i)$ necessarily converge to $\pi(i)$?

– It turns out the answer is yes, but we need to prove this.

- **Proof of Convergence for Power Iteration**:

Define

$$\Delta_r = \sum_i |\pi_r(i) - \pi(i)|$$

as measure of distance between the iterative and true π values. We want to show that $\lim_{r \to +\infty} \Delta_r = 0$. From the above discussion we know:

$$\pi_{r+1}(i) = \frac{\epsilon}{N} + (1-\epsilon) \sum_{j:j \text{ links to i}} \frac{\pi_r(j)}{|\text{out}(j)|} \tag{1}$$

$$\pi(i) = \frac{\epsilon}{N} + (1-\epsilon) \sum_{j:j \text{ links to i}} \frac{\pi(j)}{|\text{out}(j)|} \tag{2}$$

Take the difference, (1)-(2)

$$\pi_{r+1}(i) - \pi(i) = (1-\epsilon) \sum_{j:j \text{ links to } i} \frac{\pi_r(j) - \pi(j)}{|\text{out}(j)|} \tag{3}$$

Applying the triangle inequality – $|a_1 + a_2 + \cdots + a_k| \leq |a_1| + |a_2| + \cdots + |a_k|$ – to (3) we have

$$|\pi_{r+1}(i) - \pi(i)| \leq (1-\epsilon) \sum_{j:j \text{ links to } i} \frac{|\pi_r(j) - \pi(j)|}{|\text{out}(j)|}$$

Summing over all i yields

$$\begin{aligned}\Big(\Delta_{r+1} =\Big) \ \sum_i |\pi_{r+1}(i) - \pi(i)| &\le (1-\epsilon)\sum_i \sum_{j:j \text{ links to } i} \frac{|\pi_r(j) - \pi(j)|}{|\text{out}(j)|} \\ &= (1-\epsilon)\sum_j \sum_{i:j \text{ links to } i} \frac{|\pi_r(j) - \pi(j)|}{|\text{out}(j)|} \\ &= (1-\epsilon)\sum_j |\pi_r(j) - \pi(j) = (1-\epsilon)\Delta_r\end{aligned}$$

Thus $\Delta_{r+1} \le (1-\epsilon)\Delta_r$ which implies that $\Delta_r \le (1-\epsilon)^r \Delta_0$ and therefore $\Delta_r \to 0$ as $r \to \infty$. I.e., for every i the $\pi_r(i)$ converge to $\pi(i)$. □

- **Personalized PageRank**
  - Personalized to a certain node x in the graph/network
  - Relevant to advertising, personalized search, etc
  - PR monkey's algorithm personalized to page x:
    - Currently at page i:
      With probability ϵ, **jump to x**
      With probability $1-\epsilon$, click on a random hyperlink within page i
    - With this modified PR random walk, the long run frequencies would satisfy:

$$\pi(i) = \begin{cases} (1-\epsilon) \displaystyle\sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|} & \text{if } i \neq x \\ \epsilon + (1-\epsilon) \displaystyle\sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|} & \text{if } i = x \end{cases}$$

  - More generally, if we wanted to personalize PageRank to a certain *subset* of the nodes in the network (e.g., personalize to all Californians, men, women, 18-27-year-olds, etc.) the PageRank algorithm personalized to a subset A would be:
    - Currently at page i:
      With probability ϵ, **jump to a random node in** A
      With probability $1-\epsilon$, click on a random hyperlink within page i
    - With the frequencies satisfying

$$\pi(i) = \begin{cases} (1-\epsilon) \displaystyle\sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|} & \text{if } i \notin A \\ \frac{\epsilon}{|A|} + (1-\epsilon) \displaystyle\sum_{j:j \text{ links to } i} \frac{\pi(j)}{|\text{out}(j)|} & \text{if } i \in A \end{cases}$$