Page Ranks and Random Walks

Yufei Tao

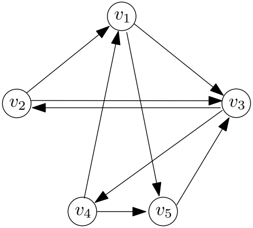
Department of Computer Science and Engineering
Chinese University of Hong Kong

We will discuss **page ranks** on a directed graph, which reflect vertices' "importance". We will also take the opportunity to discuss the theory of **random walks** (a.k.a. **Markov chains**), which generalize the stochastic process underlying page ranks.

Internet as a Graph

Represent WWW as a directed graph $G = (V, E)$:

- Each webpage is a node in V.
- E has an edge from v_1 to v_2 if page v_1 has a link to page v_2.



If a page v has no links, add a link to itself.

Random Surfing

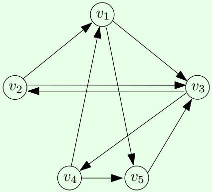
1. u = the page we are currently at (initially, u = an arbitrary page).
2. Toss a coin with a "heads" probability α.
3. If the coin comes up heads, follow a random link in u and set u to the new page
4. Otherwise (tails), set u to a random page in G – call this a **reset**.
5. Repeat from Step 1.

Page Rank

A page's **page rank** is the probability of being the t-th page visited when $t = \infty$.

The probability is not affected by the choice of the first page (this will become clear later).

Example:



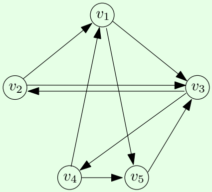
Assume that $\alpha = 4/5$ and the 1st page chosen is v_1. What is the probability of the event "2nd page $= v_3$"? This happens if one of the following takes place:

- The coin comes up heads and we follow the link from v_1 to v_3; probability $= \frac{4}{5} \cdot \frac{1}{2} = \frac{2}{5}$.
- Tails and the reset picks v_3; probability $= \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{25}$.

Hence, the probability is $\frac{1}{25} + \frac{2}{5} = \frac{11}{25}$.

Access Probability

Example (cont.):



What is the probability of "3rd page $= v_4$"? This happens if one of the following takes place:

- 2nd page $= v_3$, the coin comes up heads, and we follow the link from v_3 to v_4; probability $= \frac{11}{25} \cdot \frac{4}{5} \cdot \frac{1}{2} = \frac{22}{125}$.
- Tails and the reset picks v_4; probability $= \frac{1}{25}$.

Hence, the probability is $\frac{22}{125} + \frac{1}{25} = \frac{27}{125}$.

Access Probability

Given a vertex $v \in V$ and an integer $t \geq 1$, define $p(v, t)$ to be the probability of "$v =$ the t-th page". Then:

$$p(v, t+1) \quad = \quad \frac{1-\alpha}{|V|} + \alpha \cdot \sum_{u \in in(v)} \frac{p(u, t)}{outdeg(u)}$$

where

- $in(v)$ is the set of **in-neighbors** of v;
- $outdeg(v)$ is the **out-degree** of v.

Access Probability $\Rightarrow$ Page Rank

When $t \to \infty$, we **always** have:

$$p(v, t+1) = p(v, t)$$

for all $v \in V$. The value of $p(v, t)$ at that moment is the **page rank** of v.

Next, we will discuss how page ranks are related to the well-established theory of random walks. We will see that page ranks form an eigenvector of a matrix that depends only on G and α.

An $n \times n$ matrix M is called a **stochastic matrix** if:

- every value in M is non-negative;
- the values of each column sum up to 1.

Random Walk

Every stochastic matrix M defines a **random walk**:

- Define a directed graph G_{markov} with nodes $v_1, ..., v_n$. For every non-zero entry $M[j, i]$ of M $(1 \leq i, j \leq n)$, G_{markov} has an edge from v_i to v_j.
- Initially, pick an arbitrary vertex as the **first stop**.
- Inductively, assuming that v_i is the t-th stop $(t \geq 1)$, move to an out-neighbor v_j with probability $M[j, i]$. That neighbor is the **$(t + 1)$-th stop**.

The above stochastic process is also called a **Markov chain**.

A random walk is **irreducible** if the nodes of G_{markov} are mutually reachable.

A random walk is **aperiodic** if the following is true: every vertex in G_{markov} has a non-zero probability of being visited at every $t \geq t_0$ for some sufficiently large t_0.

An $n \times 1$ vector P is a **probability vector** if:

- each component in P is a value between 0 and 1;
- all components of P sum up to 1.

Theorem: Let M be a stochastic matrix describing an irreducible and aperiodic random walk. Then, there is a unique probability vector P satisfying $P = MP$.

The proof is non-trivial and omitted.

P is the **stationary probability vector** of the random walk. Note that it is an eigenvector of M corresponding to the eigenvalue 1.

Random Surfing = Random Walk

The random surfing process we saw earlier is a random walk. Given v_i as the current stop, we jump to v_j with probability

- $\frac{1-\alpha}{n}$ if v_i has no link to v_j;
- $\frac{1-\alpha}{n} + \frac{\alpha}{outdeg(v_i)}$ otherwise.

Define M as an $n \times n$ matrix with $M[j, i]$ set to the above probability.

Think: Why is the random walk irreducible and aperiodic?

Random Surfing = Random Walk

As before, let $p(v_i, t)$ $(1 \leq i \leq n)$ be the probability of "v_i = the t-th stop". Define

$$P(t) = \begin{bmatrix} p(v_1, t) \\ p(v_2, t) \\ ... \\ p(v_n, t) \end{bmatrix}$$

From Slide 8, we know:

$$P(t+1) = M \cdot P(t).$$

Random Surfing = Random Walk

When $P(t+1) = P(t)$, $P(t)$ is the solution of P in

$$P = MP.$$

By the theorem in Slide 14, P uniquely exists, which proves the uniqueness of page ranks.

Power Method

We can calculate P with the following algorithm (known as the **power method**):

1. $P(1) \leftarrow (1, 0, ..., 0)^T$ and $t \leftarrow 1$
2. **for** $t = 2, 3, ...$ **do**
3. $\quad P(t+1) = M \cdot P(t)$

In practice, terminate the algorithm at some reasonably large t (e.g., 100). Next, we will show that the algorithm converges quickly.

Define r_i $(1 \le i \le n)$ as the page rank of v_i. We will consider the following error metric:

$$Err(t) = \sum_{i=1}^{n} |p(v_i, t) - r_i|. \quad (1)$$

We will prove:

Lemma: $Err(t) \le \alpha \cdot Err(t-1)$.

This implies $Err(t) \le \alpha^t \cdot Err(0)$ and, hence, $Err(t) \le \epsilon$ after $t = O(\log \frac{1}{\epsilon})$ rounds.

Proof

By definition of stationary vector, we know that for each $i \in [1, n]$,

$$r_i = \frac{1-\alpha}{n} + \alpha \cdot \sum_{\text{in-neighbor } v_j \text{ of } v_i} \frac{r_j}{outdeg(v_j)}.$$

By how the power method runs, we know:

$$p(v_i, t) = \frac{1-\alpha}{n} + \alpha \cdot \sum_{\text{in-neighbor } v_j \text{ of } v_i} \frac{p(v_j, t-1)}{outdeg(v_j)}.$$

The above equations yield

$$|p(v_i, t) - r_i| \leq \alpha \cdot \sum_{\text{in-neighbor } v_j \text{ of } v_i} \frac{|p(v_j, t-1) - r_j|}{outdeg(v_j)}. \quad (2)$$

Proof

By combining (1) and (2), we have:

$$Err(t) \quad \leq \quad \alpha \cdot \sum_{v_i} \sum_{\text{in-neighbor } v_j \text{ of } v_i} \frac{|p(v_j, t-1) - r_j|}{outdeg(v_j)}.$$

Observe that $\frac{|p(v_j, t-1) - r_j|}{outdeg(v_j)}$ is added exactly $outdeg(v_j)$ times on the right hand side. Therefore:

$$Err(t) \quad \leq \quad \alpha \cdot \sum_{v_i} |p(v_i, t-1) - r_i| = \alpha \cdot Err(t-1)$$

which completes the proof. □