Graph Mining: Page Ranks and Random Walks

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

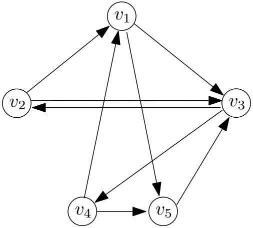
This lecture will discuss

- **page ranks** for measuring vertex importance in directed graphs, and
- the underlying theory on **random walks** (a.k.a. **Markov chains**).

Internet as a Graph

To start our discussion, let us represent WWW as a directed graph $G = (V, E)$:

- Each webpage is a node in V.
- E has an edge (v_1, v_2) if page v_1 has a hyper-link to page v_2.
- If a page v has no outgoing links, add a self-loop (v, v) to E.



Random Surfing

1. u = the page we are visiting (initially, set u to an arbitrary page).
2. Toss a coin with heads probability α.
3. If the coin comes up heads, follow a random out-edge (u, v) of u; set u to v.
4. Otherwise (tails), set u to a random page in G; call this a **reset**.
5. Repeat from Step 1.

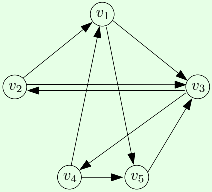
Page Rank

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

The lecture will answer the FAQs below:

- Would the probability converge for every vertex for $t = \infty$?
- How fast is the convergence?
- Do page ranks depend on the choice of the first page?
- How to compute the page ranks?

Example: Assume that $\alpha = 4/5$ and the 1st page chosen is v_1.

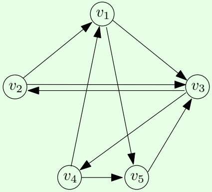


What is the probability of "2nd page=v_3"? The event happens if

- The coin comes up heads and we follow the link $(v_1, v_3) \Rightarrow$ probability $= \frac{4}{5} \cdot \frac{1}{2} = \frac{2}{5}$;
- tails and the reset picks $v_3 \Rightarrow$ probability $= \frac{1}{5} \cdot \frac{1}{5} = \frac{1}{25}$.

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

Example (cont.):



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

- 2nd page $= v_3$, the coin comes up heads, and we follow the link $(v_3, v_4) \Rightarrow$ 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

$$p(v, t) = \textbf{\textit{Pr}}[v \text{ is the } t\text{-th page visited}].$$

Then:

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

where

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

Access Probability ⇒ Page Rank

When $t \to \infty$,

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

definitely holds for all $v \in V$.

The converged value of $p(v, t)$ is the **page rank** of v.

Before delving into the theory of page ranks, we need to first understand some basic results from the theory of random walks.

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.

An $n \times n$ matrix $\boldsymbol{M}$ is called a **stochastic matrix** if every column is a probability vector.

Random Walk

Every stochastic matrix $\boldsymbol{M}$ defines a **random walk** as follows.

- Build a directed graph G_{markov} with vertices $v_1, ..., v_n$. For every non-zero entry $\boldsymbol{M}[j, i]$ of $\boldsymbol{M}$, add an edge (v_i, v_j) to G_{markov}.
- Pick an arbitrary vertex as the **first stop**.
- Inductively, assuming that the t-th stop $(t \geq 1)$ is at v_i, move to an out-neighbor v_j with probability $\boldsymbol{M}[j, i]$ as 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.

Theorem 1: Let $\boldsymbol{M}$ be a stochastic matrix describing an irreducible and aperiodic random walk. Then, all the following are true.

- There is a unique probability vector P satisfying $P = \boldsymbol{M}P$.
- When $t \to \infty$, $\boldsymbol{Pr}[v_i$ is the t-th node visited] equals $P[i]$ for each $i \in [1, n]$.

The proof is non-trivial and omitted.

P is the **stationary probability vector** of the random walk.

- P an eigenvector of $\boldsymbol{M}$ corresponding to the eigenvalue 1.

Random Surfing = Random Walk

The random surfing process 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.

Think: What is $\boldsymbol{M}$? Why is the random walk irreducible and aperiodic?

Random Surfing = Random Walk

Recall: $p(v_i, t) = \boldsymbol{Pr}[v_i \text{ is the } t\text{-th visited}]$, for each $i \in [1, n]$.

Define

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

From Slide 8, we know:

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

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

$$P = \boldsymbol{M} P.$$

Theorem 1 implies that $P(t) \to P$ when $t \to \infty$.

Finally, we will analyze how fast $P(t)$ will converge to P. Our analysis will also serve as another proof for the convergence of $P(t)$.

Power Method

Consider the following algorithm for computing $P(t)$ iteratively:

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

Next, we will show that the algorithm converges quickly.

Let r_i = the page rank of v_i (for each $i \in [1, n]$).

Define:

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

We will prove:

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

This implies $Err(t) \le \alpha^t \cdot Err(0)$.
In turn, this shows that $Err(t) \le \epsilon$ 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 have:

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

The above equations yield

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

Proof

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

$$Err(t) \quad \leq \quad \alpha \cdot \sum_{i=0}^{n} \sum_{\text{in-neighbor } v_j \text{ of } v_i} \frac{|p(v_j, t-1) - r_j|}{outdeg(v_j)}.$$

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. □