Graph Mining
SD212
3. PageRank

Thomas Bonald

2019 – 2020

TELECOM Paris
IP PARIS

Motivation

How to identify the most "important" nodes in a graph, either globally or relatively to some other nodes?

Useful for:

- information retrieval
- content recommendation
- local clustering

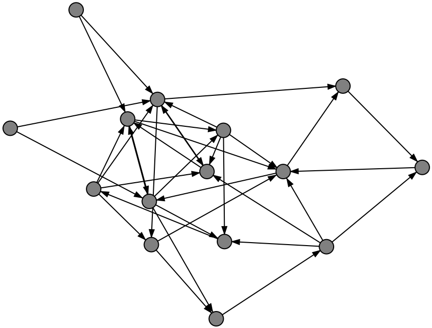
We focus on PageRank, originally proposed by Google's founders in 1999 to rank Web pages: popular pages are typically visited more frequently by a random Web surfer.

Outline

Setting

Consider a directed graph $G = (V, E)$:

- n nodes, m edges
- A, adjacency matrix
- $d^+ = A1, d^- = A^T1$, vectors of out-degrees and in-degrees



Random walk

In the **absence** of sinks ($d^+ > 0$):

- A Markov chain $X_0, X_1, X_2, \ldots$ of transition matrix $P = D^{-1}A$ with $D = \mathrm{diag}(d^+)$
- Probability distribution π_t at time t (row vector)
- Dynamics $\pi_{t+1} = \pi_t P$

Stationary distribution

If the graph is **strongly connected** and **aperiodic**,

$$\lim_{t \to +\infty} \pi_t = \pi \quad \text{with} \quad \pi = \pi P$$

Computation

Stationary distribution

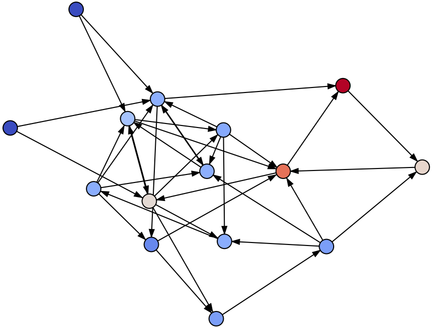
Input:
P, transition matrix
K, number of iterations

Do:
$\pi \leftarrow \frac{1}{n}(1, \dots, 1)$
For $t = 1, \dots, K$, $\pi \leftarrow \pi P$

Output:
π, (approximate) stationary distribution

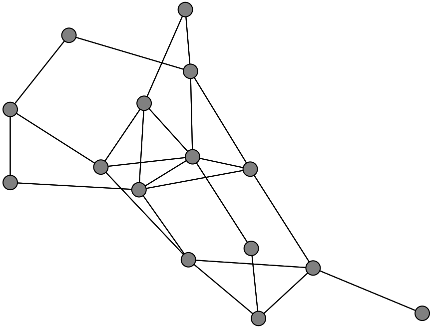
Complexity: $O(Km)$ in time, $O(n)$ in memory

Example



The case of undirected graphs

We have $d = d^+ = d^-$

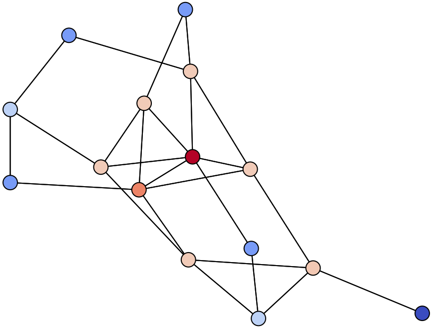


Stationary distribution

If the graph is **connected**, the stationary distribution is proportional to the degrees:

$$\pi \propto d$$

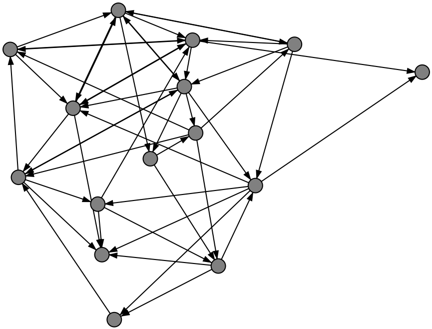
Example



Outline

1. Random walk
2. **PageRank**
3. Personalized PageRank
4. BiPageRank

Accounting for sinks



Random walk with forced restarts

$$P_{ij} = \begin{cases} \frac{A_{ij}}{d_i^+} & \text{if } d_i^+ > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

PageRank

Random walk with **restarts**:

- Fix $\alpha \in (0, 1)$
- Walk with probability α, restart (e.g., to a random node) with probability $1 - \alpha$
- An irreducible Markov chain with transition matrix:

$$P^{(\alpha)} = \alpha P + (1 - \alpha) \frac{11^T}{n}$$

PageRank

Unique solution to the equations:

$$\pi^{(\alpha)} = \alpha \pi^{(\alpha)} P + (1 - \alpha) \frac{1^T}{n}$$

Computation

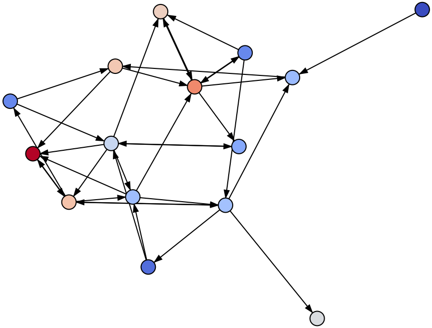
PageRank

Input:
P, transition matrix (with forced restarts)
α, damping factor
K, number of iterations

Do:
$\pi \leftarrow \frac{1}{n}(1, \ldots, 1)$
For $t = 1, \ldots, K$, $\pi \leftarrow \alpha \pi P + (1 - \alpha)\frac{1}{n}(1, \ldots, 1)$

Output:
π, (approximate) PageRank vector

Example ($\alpha = 0.85$)



Setting the damping factor

- The path length before restart (in the absence of sinks) has a **geometric distribution** with parameter $1 - \alpha$
- Average path length:
$$\frac{\alpha}{1-\alpha}$$
- For $\alpha = 0.85$, we get about 5.7, a typical distance between two nodes in real graphs (cf. the **six degrees of separation**).

Expression of the PageRank vector

Proposition

$$\pi^{(\alpha)} = (1-\alpha)\sum_{t=0}^{+\infty}\alpha^t \pi_t$$

Limiting cases

- **No restarts** $(\alpha \to 1)$

$$\pi^{(\alpha)} \to \pi = \lim_{t\to+\infty} \pi_t$$

- **Frequent restarts** $(\alpha \to 0)$

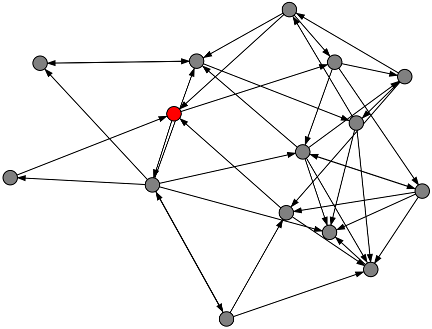
$$\pi^{(\alpha)} = (1-\alpha)\pi_0 + \alpha\pi_1 + o(\alpha)$$

Ranking equivalent to neighbor sampling

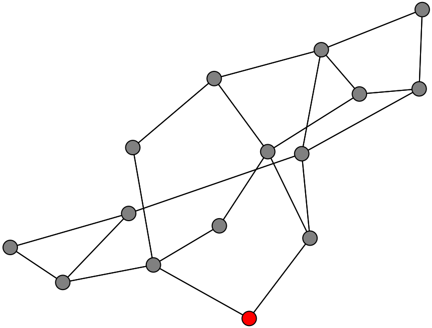
Outline

1. Random walk
2. PageRank
3. **Personalized PageRank**
4. BiPageRank

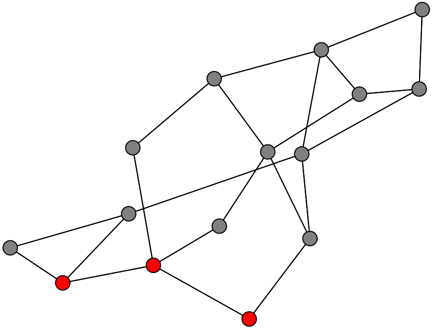
Personalization



Personalization



Personalization



Personalized PageRank

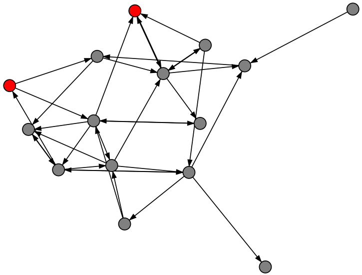
Let μ be some distribution on $S \subset V$ (e.g., uniform)

- Forced restarts:

$$P_{ij} = \begin{cases} \frac{A_{ij}}{d_i^+} & \text{if } d_i^+ > 0 \\ \mu_j & \text{otherwise} \end{cases}$$

- Random restarts:

$$P^{(\alpha)} = \alpha P + (1 - \alpha) 1 \mu$$



Computation

Personalized PageRank

Input:
P, transition matrix (with forced restarts)
μ, personalization row vector
α, damping factor
K, number of iterations

Do:
$\pi \leftarrow \mu$
For $t = 1, \ldots, K$, $\pi \leftarrow \alpha \pi P + (1 - \alpha)\mu$

Output:
π, (approximate) PageRank vector

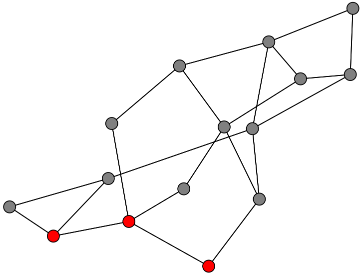
Expression of the Personalized PageRank vector

Proposition

In the absence of sinks,

$$\pi^{(\alpha)} = \sum_{s \in S} \mu_s \pi_s^{(\alpha)}$$

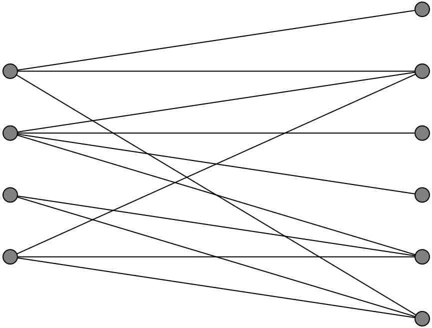
where $\pi_s^{(\alpha)}$ is the Personalized PageRank vector associated with s



Outline

1. Random walk
2. PageRank
3. Personalized PageRank
4. **BiPageRank**

Case of bipartite graphs



$$A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \quad D = \mathrm{diag}(d) = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

BiPageRank

Given a bipartite graph $G = (V_1, V_2, E)$, consider a random walk with restarts in V_1 (e.g., uniform)

Transition matrices:

- $P_1 = D_1^{-1}B$ from V_1 to V_2
- $P_2 = D_2^{-1}B^T$ from V_2 to V_1

BiPageRank

Unique solution to the equations:

$$\pi_1 = \alpha \pi_2 P_2 + (1-\alpha)\mu_1$$
$$\pi_2 = \alpha \pi_1 P_1$$

with μ_1 uniform on V_1

Computation

BiPageRank

Input:
P_1, transition matrix from V_1 to V_2
P_2, transition matrix from V_2 to V_1
α, damping factor
K, number of iterations

Do:
$\pi_1 \leftarrow \frac{1}{n_1}(1, \ldots, 1)$, $\pi_2 \leftarrow 0$
For $t = 1, \ldots, K$,
$\quad \pi_1 \leftarrow \alpha \pi_2 P_2 + (1 - \alpha) \frac{1}{n_1}(1, \ldots, 1)$
$\quad \pi_2 \leftarrow \alpha \pi_1 P_1$

Output:
π_1, π_2, BiPageRank vectors

Expression of the BiPageRank vector

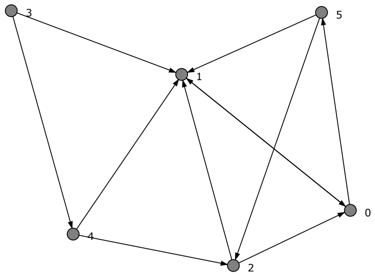
Let $\pi(t)$ be the distribution after t time steps of the pure random walk starting from V_1.

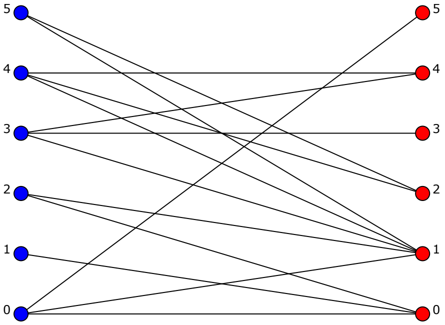
Proposition

$$\pi_1^{(\alpha)} = (1-\alpha) \sum_{t \in 2\mathbb{N}} \alpha^t \pi_1(t)$$

$$\pi_2^{(\alpha)} = (1-\alpha) \sum_{t \in 2\mathbb{N}+1} \alpha^t \pi_2(t)$$

Directed graphs as bipartite graphs





PageRank vs BiPageRank

Consider some target node $s \in V$, and let $\alpha \to 0$

- Successors of s are best ranked with PageRank
- Nodes having many common successors with s are best ranked with BiPageRank