Massive Graph Management & Analytics

Random Walk on Graphs

Nacéra Seghouani

Computer Science Department, CentraleSupélec
Laboratoire Interdisciplinaire des Sciences du Numérique, LISN
nacera.seghouani@centralesupelec.fr

2024-2025

First Perron-Frobenius Theory

☞ Perron-Frobenius vector for non-negative matrices leads to the characterization of non-negative primary eigenvectors, such as stationary distributions of Markov chains and Google's PageRank.

$\rightarrow$ Note that positive or non negative matrix iff $\mathbf{A}_{ij} > 0$ resp. $\mathbf{A}_{ij} \geq 0$ is different from positive definite or semi-positive definite matrix

☞ Perron theorem for positive matrix $\mathbf{A}_{ij} > 0$

- ✓ $\exists \lambda^* > 0, \mathbf{v}^* > \mathbf{0}, ||\mathbf{v}^*||_2 = 1$ s.t $\mathbf{A}\mathbf{v}^* = \lambda^* \mathbf{v}^*$ right column eigenvector
  $\exists \lambda^* > 0, \mathbf{w} > \mathbf{0}, ||\mathbf{w}||_2 = 1$ s.t $\mathbf{w}\mathbf{A} = \lambda^* \mathbf{w}$ left row eigenvector
- ✓ $\forall \lambda$ other eigenvalue of $\mathbf{A}, |\lambda| < \lambda^*$, dominant eigenvalue (the largest absolute value)
- ✓ λ^* is simple (multiplicity 1) and $\mathbf{v}^*$ is unique (up to rescaling).
  Such eigenvectors will be called Perron vectors.

☞ Perron Theorem for non negative Matrix: $\mathbf{v}^*$ are non negative rather than positive and the uniqueness of $\mathbf{v}^*$ is not guaranteed

First Perron-Frobenius Theory

- **Definition: Irreducible**
  Connected graph $\mathcal{G}(V,E) \Leftrightarrow$ for any $1 \leq i,j \leq |V|, \exists k \in \mathbb{N}^*, \mathbf{A}_{ij}^k > 0$
- **Definition: Primitive**
  strengths this condition to k-connected, *i.e.* every pair of nodes are connected by a path of length $k \Rightarrow \lambda^*$ is unique and $\lambda^* = max|\lambda|$
  - ✓ $\mathbf{A}$ is connected (irreducible) and $\mathbf{A}_{ii} > 0$ for some i is sufficient for primitivity but not necessary
- **Definition: Primitive + Non-negative**
  - ✓ $\exists\lambda^* > 0, \mathbf{v}^* > \mathbf{0}, ||\mathbf{v}^*||_2 = 1$ s.t $\mathbf{A}\mathbf{v}^* = \lambda^*\mathbf{v}^*$ right column eigenvector
    $\exists\mathbf{w} > \mathbf{0}, ||\mathbf{w}||_2 = 1$ s.t $\mathbf{w}\mathbf{A} = \lambda^*\mathbf{w}$ left row eigenvector
  - ✓ $\forall\lambda$ other eigenvalue of $\mathbf{A}, |\lambda| < \lambda^*$, dominant eigenvalue (the largest absolute value)
  - ✓ $\mathbf{v}^*$ is unique (up to rescaling).

Random walks on graphs

- A random walk on a graph $\mathcal{G}(E, V)$ is a random process that starts from some vertex v_i, and repeatedly moves to a neighbor v_j chosen uniformly at random. ξ_t is a random variable describing the position of a random walk after t steps.

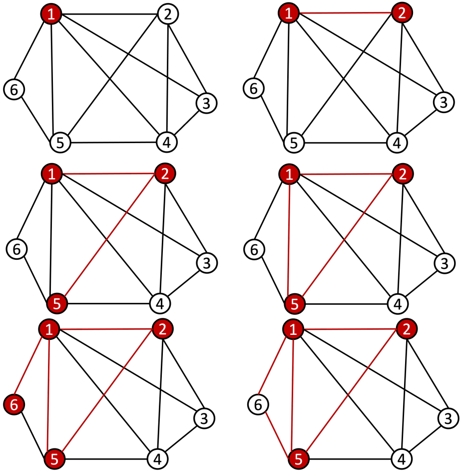
$$\mathbf{P}_{ij} = P(\xi_{t+1} = j | \xi_t = i)$$

- The sequence can be regarded as a category of Markov chain (discrete time stochastic process) where the position ξ_0 is the initial state according to the distribution P^0 and next state depends only on the current state. The t steps transition probability is:

$$\mathbf{P}^t_{ij} = P(\xi_t = j | \xi_0 = i)$$

- Some examples: path traced by a molecule in a liquid or a gas (Brownian motion) ,
the price of a fluctuating stock, the financial status of a gambler, ... The term random walk was first introduced by Karl Pearson in 1905.

Random walks on graphs: Example



Random walks on graphs

- Consider a random walk with $\mathbf{P}_{ij} = P(\xi_{t+1} = j | \xi_t = i) \geq 0$, thus $\mathbf{P}$ is a row-stochastic or row-Markov matrix:

$$\sum_j \mathbf{P}_{ij} = 1, \mathbf{P} \times \mathbf{1} = \mathbf{1} \in \mathbb{R}^n \text{ right Perron eigenvector} > 0$$

- From Perron theorem for non-negative matrices, we know:
  - ✓ $\mathbf{v}^* = \mathbf{1}$ is a right Perron eigenvector of $\mathbf{P}$
  - ✓ $|\lambda| \leq \lambda^* = 1$, is a Perron eigenvalue
  - ✓ $\exists$ left Perron row eigenvector $\pi\mathbf{P} = \pi$
  - ✓ $\mathbf{P}$ is primitive $\Rightarrow \pi$ is unique

Random walks on graphs: Stationary distribution

- Let be π^t the row vector giving the probability distribution of ξ_t. π_i^t be the probability that a walk is on v_i at t.

$$\pi^{t+1} = \pi^t \mathbf{P}$$
$$\pi^{t+1} = \pi^0 \mathbf{P}^t$$

$$lim_{t\to\infty} \pi^{t+1} = lim_{t\to\infty} \pi^t \mathbf{P}$$

$$\lambda\pi = \pi\mathbf{P} \text{ with } \lambda = 1$$

- This means if we take powers of $\mathbf{P}$, *i. e.*, $\mathbf{P}^k$, all rows will converge to the stationary distribution π if primitivity holds.
- What about: $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Random walks on graphs

- What does "random" mean? If the walk is on i at t, the single step transition probability refers to the uniform probability that the random walk moves to j at $t+1$.

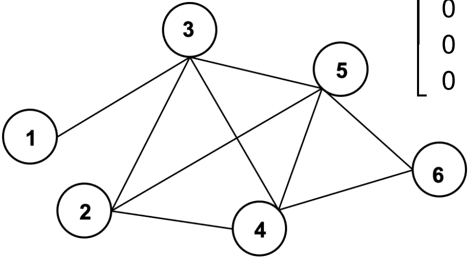
$$\mathbf{P}_{ij} = P(\xi_{t+1} = j | \xi_t = i) = \begin{cases} \frac{1}{d_i} & \forall v_i, v_j \in V \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mathbf{P}_{ij} = \frac{\mathbf{A}_{ij}}{\sum_{v_j \in V} \mathbf{A}_{ij}} = \frac{\mathbf{A}_{ij}}{d_i} = \mathbf{D}_{ii}^{-1} \mathbf{A}_{ij}$$

- The random sequence of vertices $\xi_0, \xi_1, \ldots \xi_t, \xi_{t+1}, \ldots$ visited on $\mathcal{G}(E, V)$ is a Markov chain with state space V and matrix transition probabilities

$$\mathbf{P} = \mathbf{D}^{-1} \mathbf{A}$$

Random walks on graphs



$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 & 0 \\ 0 & 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

Random walks on graphs: Balance condition

✓ If we find a probability distribution π which satisfies balance condition, then:

$$\pi_i \mathbf{P}_{ij} = \pi_j \mathbf{P}_{ji} \; \forall v_i, v_j \in V$$

$$\pi_i \frac{\mathbf{A}_{ij}}{d_i} = \pi_j \frac{\mathbf{A}_{ji}}{d_j} \;\Rightarrow\; \frac{\pi_i}{d_i} = \frac{\pi_j}{d_j} = const$$

$$\Rightarrow \sum_j \pi_j = \sum_j \frac{\pi_i}{d_i} d_j = const \sum_j d_j = 1$$

$$\pi_i = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2|E|}$$

The stationary probabilities are proportional to the degrees of the vertices.

Random walks on graphs: Balance condition

- ✓ In particular, if $\mathcal{G}$ is d-regular (nodes with equal degrees d),

$$\pi_{\mathbf{i}} = \frac{d}{2m} = \frac{1}{n} \; \forall v_i \in V$$

  is the uniform distribution: a random walk moves along every edge with the same frequency.
- ✓ The balance condition implies time-reversibility. The reversed walk is also a Markov chain. Suppose that the random walk has the stationary distribution and consider the reversed walk $\rho_t = \xi_{r-t}$ with $r = 0, \ldots, t$

Random walks on graphs: Hitting time

- The expected hitting probability: $\mathbf{h}_{ij}$ is the probability of hitting j starting from i

$$\mathbf{h}_{ij} = \begin{cases} \sum_k \mathbf{P}_{ik}\mathbf{h}_{kj} & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

- The expected hitting time: $\mathbf{H}_{ij}$ is the expected number of walks before hitting j in a random walk starting from i (conditioning by the first walk)

$$\mathbf{H}_{ij} = \begin{cases} 1 + \sum_k \mathbf{P}_{ik}\mathbf{H}_{kj} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

  In general $\mathbf{H}_{ij} \neq \mathbf{H}_{ji}$. $\mathbf{H}$ matrix is not symmetric; it follows the triangle inequality.

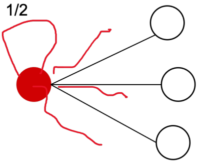
- The commute time $\mathbf{C}_{ij} = \mathbf{H}_{ij} + \mathbf{H}_{ji}$ means the excepted number of steps in a random walk starting at i, before accessing the node j and then reaching i again.

- Example:

$$\begin{bmatrix} 3/4 & 1/4 & 0 \\ 3/4 & 0 & 1/4 \\ 0 & 1/4 & 3/4 \end{bmatrix}$$

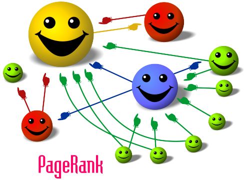
Random walks on graphs: Lazy random walk

- Lazy random walk: either stay on the current node with the probability $\frac{1}{2}$ or walk on a neighbor



- Give the matrix form $\mathbf{P}^t$ and $\mathbf{P}^{t+1}$. What do you observe comparing to simple random walk.

PageRank



- Formalize the problem in the case of directed graph by taking into account only out going edges
- The web is very heterogeneous by its nature, and certainly huge, we do not expect its graph to be connected. The solution of Page and Brin: fix a positive constant p between 0 and 1, called the damping factor (a typical value for p is 0.15). Define the Page Rank matrix

$$\mathbf{P}_g = (1-p)\mathbf{P} + p\mathbf{B} \text{ where } \mathbf{B} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

  Most of the time, a surfer will follow the outgoing links and move on to one of the neighbors. A smaller percentage of the time, the surfer will dump the current page and choose arbitrarily a different page from the web. The damping factor p reflects the probability that the surfer quits the current page and jump to a new one .

- Prove that $\mathbf{P}_g$ remains stochastic.

First mini-Project (notebook)

- ☞ Give some examples of specific/known transition matrices to show convergence, hitting probabilities and hitting times expectations (formalisation + program)
- ☞ Give some examples to compare random walk and lazy random walk (formalisation + program)