Chapter IV: Link Analysis

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Wintersemester 2013/14

Friendship Networks, Citation Networks, …

- **Link analysis** studies the **relationships** (e.g., friendship, citation) between **objects** (e.g., people, publications) to find out about their **characteristics** (e.g., popularity, impact)

- **Social Network Analysis** (e.g., on a friendship network)
  - **Closeness centrality** of a person v is the **fraction of shortest paths** between any two persons (u, w) that pass through v

- **Bibliometrics** (e.g., on a citation network)
  - **Co-citation** measures how many papers cite both u and v
  - **Co-reference** measures how many common papers both u and v refer to

…, and the Web?

- **World Wide Web** can be seen as **directed graph** $G(V, E)$
  - **web pages** correspond to **vertices** (or, nodes) V
  - **hyperlinks** between them correspond to **edges** E

- Link analysis on the **Web graph** can give us clues about
  - which web pages are **important** and should thus be ranked higher
  - which pairs of web pages are **similar to each other**
  - which web pages are probably **spam** and should be ignored
  - …

Chapter IV: Link Analysis

IV.1 The World Wide Web as a Graph

1. **How Big is the Web?**
2. **Degree Distributions**
3. **Random-Graph Models**
4. **Bow-Tie Structure**

Based on **MRS Chapter 21**

1. How Big is the Web?

- How big is the entire World Wide Web?
  - **quasi-infinite** when you consider all (dynamic) URLs (e.g., of calendars)

- **Indexed Web** is a more reasonable notion to look at
  - [Gulli and Signori ’05] estimated it as 11.5 billions (10^9) in 2005
  - Google claimed to know about more than 1 trillion (10^{12}) URLs in 2008
  - WorldWideWebSize.com provides daily estimates obtained by extrapolating from the number of results returned by Google and Bing on the basis of Zipf’s law (currently: 3.6 billion – 38 billion)

2. Degree Distributions

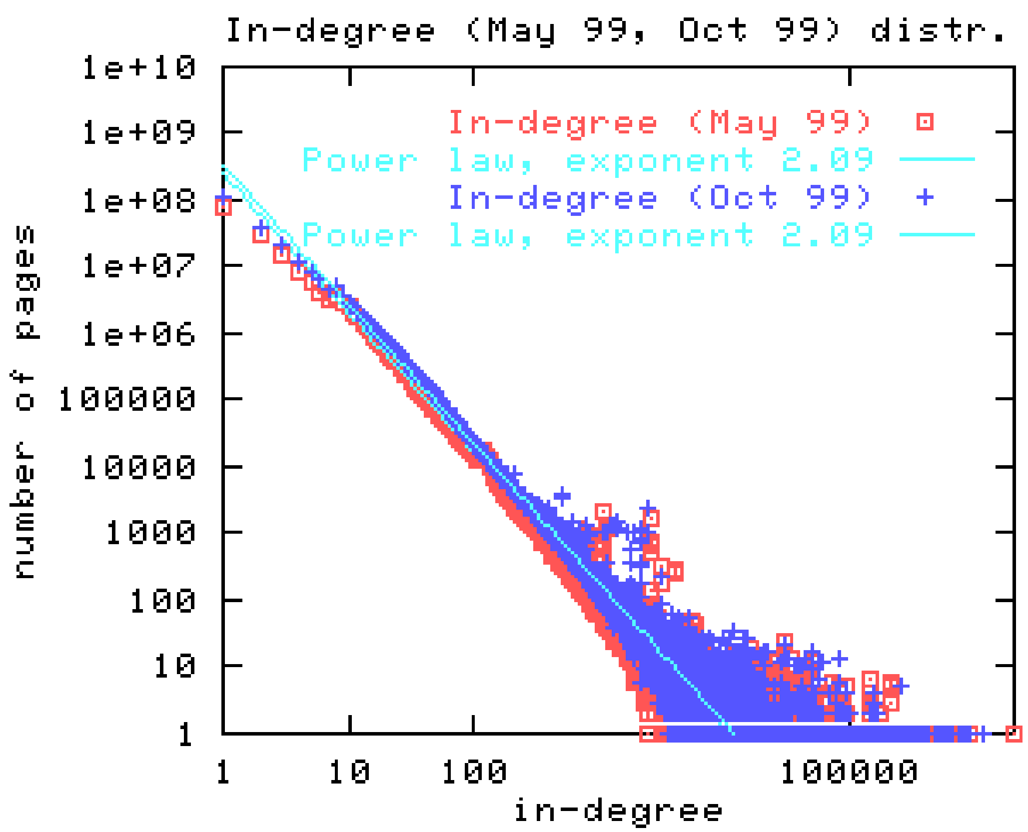
- What is the distribution of in-/out-degrees on the Web graph?
  - *in-degree*(v) of vertex v is the number of incoming edges (u, v)
  - *out-degree*(v) of vertex v is the number of outgoing edges (v, w)
- **Zipfian distribution** has probability mass function

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

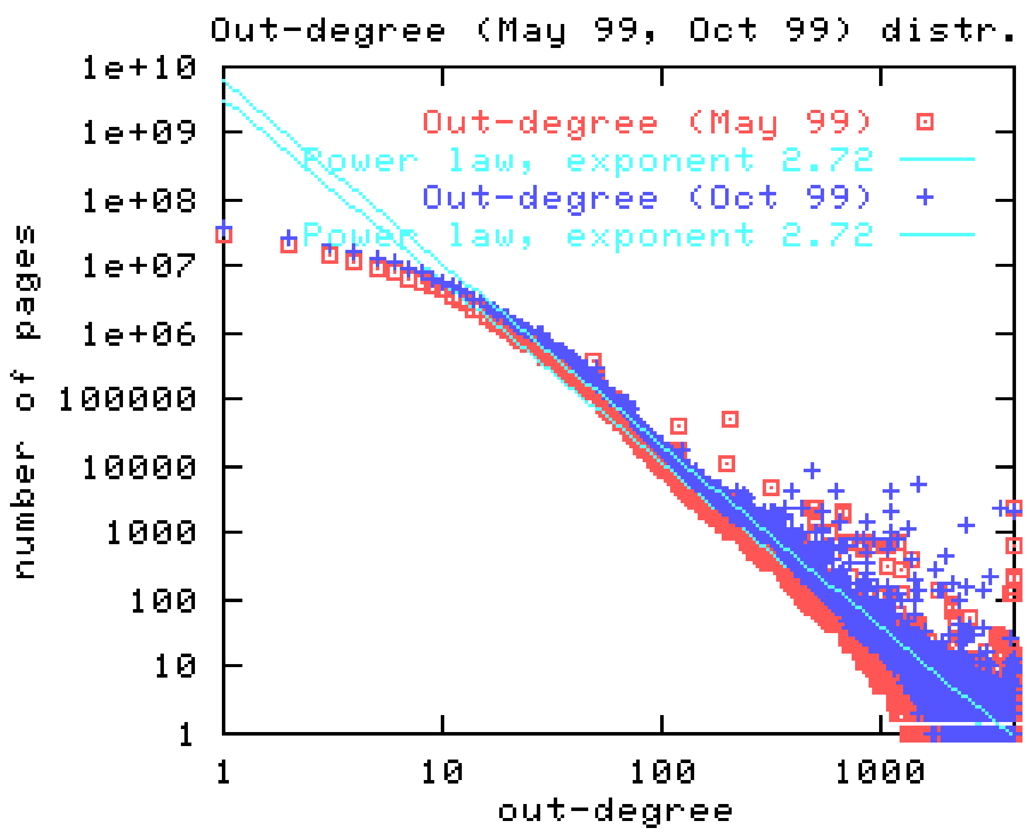
with rank k, parameter s, and total number of objects N

  - provides good model of **many real-world phenomena**, e.g., word frequencies, city populations, corporation sizes, income rankings
  - appear as **straight line** with slope -s in **log-log-plot**

Degree Distributions



$$s = 2.10$$



$$s = 2.72$$

- Full details: [Broder et al. ‘00]

3. Random-Graph Models

- **Generative models** of undirected or undirected graphs

- **Erdös-Renyi Model** $G(n, p)$ generates a graph consisting of n vertices; each possible edge (u, w) exists with probability p

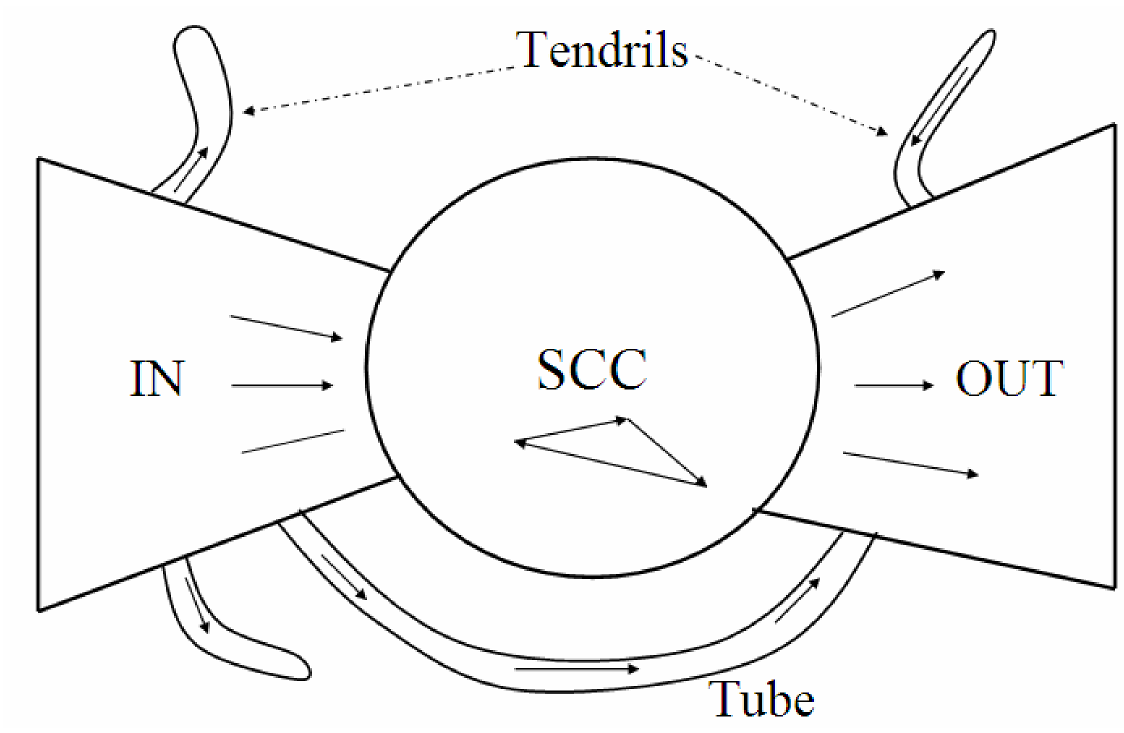
- **Barabási-Albert Model** generates a graph by successively adding vertices u with m edges; the edge (u, v) attaches to vertex v with probability proportional to $deg(v)$

- **Preferential attachment** ("*the rich get richer*") in the Barabási-Albert Model yields graphs with properties similar to Web graph

- Full details: [Barabási and Albert '99]

4. Bow-Tie Structure

- The Web graph looks a lot like a **bow tie** [Broder et al. '00]



- **Strongly Connected Component** (SCC) of web pages that are reachable from each other by following a few hyperlinks
- **IN** consisting of web pages from which SCC is reachable
- **OUT** consisting of web pages reachable from SCC

Additional Literature for IV.1

- **A.-L. Barabási and R. Albert:** *Emergence of Scaling in Random Networks*, Science 1999

- **A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener**: *Graph Structure in the Web*, Computer Networks 33:309-320, 2000

- **A. Gulli and A. Signori**: *The Indexable Web is More than 11.5 Billion Pages*, WWW 2005

- **R. Meusel, O. Lehmberg, C. Bizer**: *Topology of the WDC Hyperlink Graph* http://webdatacommons.org/hyperlinkgraph/topology.html, 2013

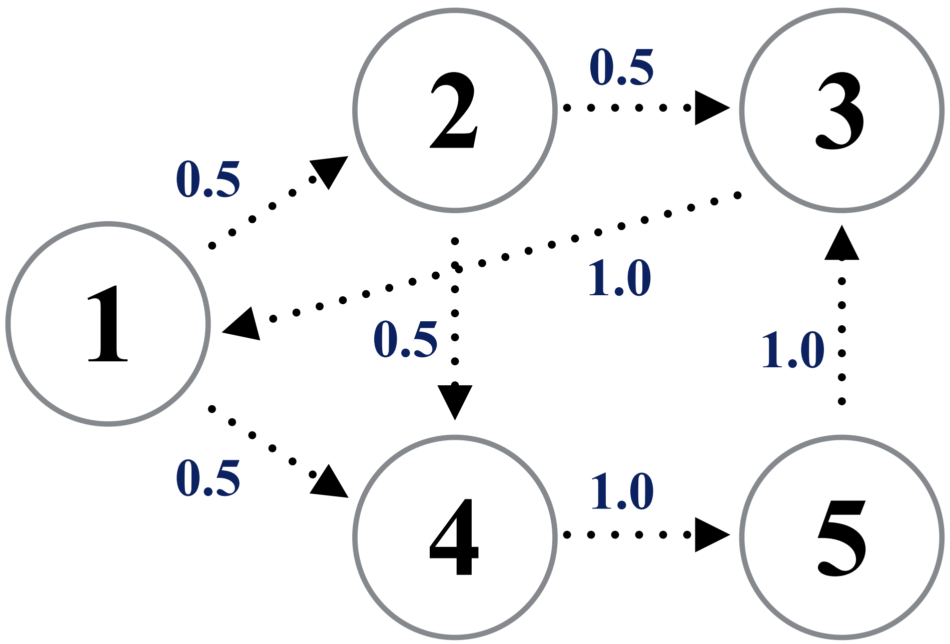
IV.2 PageRank

- **Hyperlinks** distinguish the Web from other document collections and can be interpreted as **endorsements** for the target web page
- **In-degree** as a measure of the **importance/authority/popularity** of a web page v is **easy to manipulate** and does not consider the **importance of the source web pages**
- **PageRank** considers a web page v **important if many important** web pages link to it

Larry Page & Sergey Brin

- **Random surfer model**
  - follows a uniform random outgoing link with probability $(1-\varepsilon)$
  - jumps to a uniform random web page with probability ε
- <u>Intuition</u>: Important web pages are the ones that are visited often

Markov Chains



$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$S = \{1, \ldots, 5\}$$

Stochastic Processes & Markov Chains

- **Discrete stochastic process** is a family of random variables

$$\{X_t \mid t \in T\}$$

with $T = \{0, 1, 2 \ldots\}$ as **discrete time domain**

- Stochastic process is a **Markov chain** if

$$\begin{aligned} & P[X_t = x \mid X_{t-1} = w, \ldots, X_0 = a] \\ = \quad & P[X_t = x \mid X_{t-1} = w] \end{aligned}$$

holds, i.e., it is **memoryless**

- Markov chain is **time-homogeneous** if for all times *t*

$$P[X_{t+1} = x \mid X_t = w] = P[X_t = x \mid X_{t-1} = w]$$

holds, i.e., **transition probabilities** do not depend on time

State Space & Transition Probability Matrix

- **State space** of a Markov chain $\{ X_t \mid t \in T \}$ is the countable set S of all values that X_t can assume
  - $X_t : \Omega \rightarrow S$
  - Markov chain **is in state *s* at time *t*** if $X_t = s$
  - Markov chain $\{ X_t \mid t \in T \}$ is **finite** if it has a finite state space
- If a Markov chain $\{ X_t \mid t \in T \}$ is **finite** and **time-homogeneous**, its **transition probabilities** can be described as a matrix $\mathbf{P} = (p_{ij})$

$$p_{ij} = P[X_t = j \mid X_{t-1} = i]$$

- For $|S| = n$ the transition probability matrix $\mathbf{P}$ is a ***n*-by-*n* right-stochastic matrix** (i.e., its rows sum up to 1)

$$\forall\, i \;:\; \sum_j p_{ij} = 1$$

Properties of Markov Chains

- State i is **reachable** from state j if there exists a n $\geq$ 0 such that $(\mathbf{P}^n)_{ij} > 0$ (with $\mathbf{P}^n = \mathbf{P} \times \ldots \times \mathbf{P}$ as n-th exponent of $\mathbf{P}$)
- States i and j **communicate** if i is reachable from j and vice versa
- Markov chain is **irreducible** if all states $i, j \in S$ communicate
- Markov chain is **positive recurrent** if the **recurrence probability** is 1 and the **mean recurrence time** is finite for every state i

$$\sum_{k=1}^{\infty} P[X_k = i \wedge \forall 1 \leq j < k \;:\; X_j \neq i \,|\, X_0 = i] = 1$$

$$\sum_{k=1}^{\infty} k\, P[X_k = i \wedge \forall 1 \leq j < k \;:\; X_j \neq i \,|\, X_0 = i] < \infty$$

Properties of Markov Chains

- Markov chain is **aperiodic** if every state *i* has period 1 defined as

$$gcd\;\{\, k \;:\; P[X_k = i \wedge \forall 1 \leq j < k \;:\; X_j \neq i \,|\, X_0 = i] > 0 \,\}$$

- Markov chain is **ergodic** if it is time-homogeneous, irreducible, positive recurrent, and aperiodic

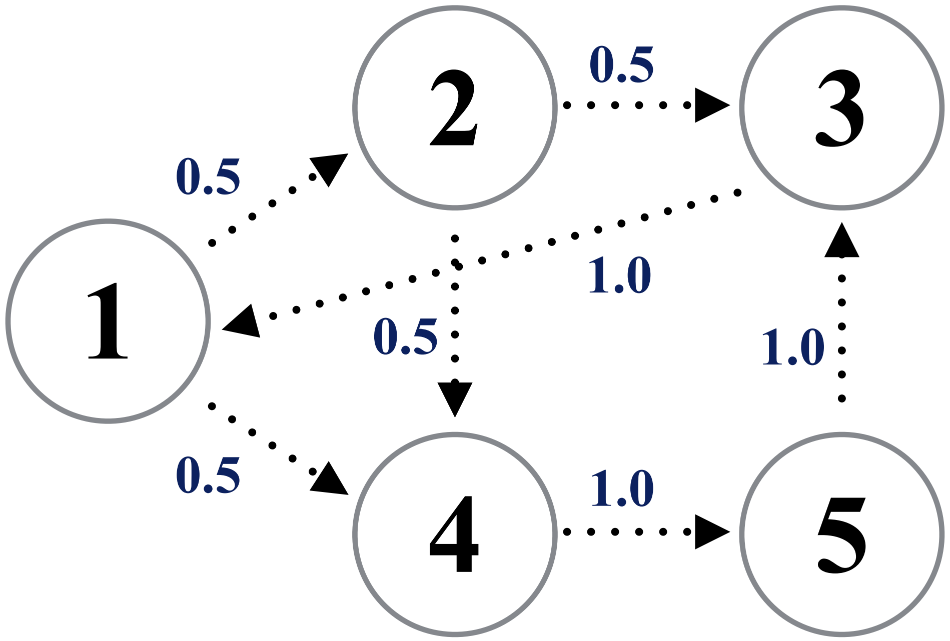
- The 1-by-*n* vector $\boldsymbol{\pi}$ is the **stationary state distribution** of the Markov chain described by $\mathbf{P}$ if $\pi_i \geq 0$, $\Sigma\, \pi_i = 1$, and

$$\boldsymbol{\pi}\,\mathbf{P} = \boldsymbol{\pi}$$

  - π_i is the limit probability that Markov chain is in state *i*
  - $1/\pi_i$ reflects the average time until the Markov chain returns to state *i*

- <u>Theorem</u>: If a Markov chain is **finite** and **ergodic**, then there exists a **unique stationary state distribution** $\boldsymbol{\pi}$

Markov Chain (Example Revisited)

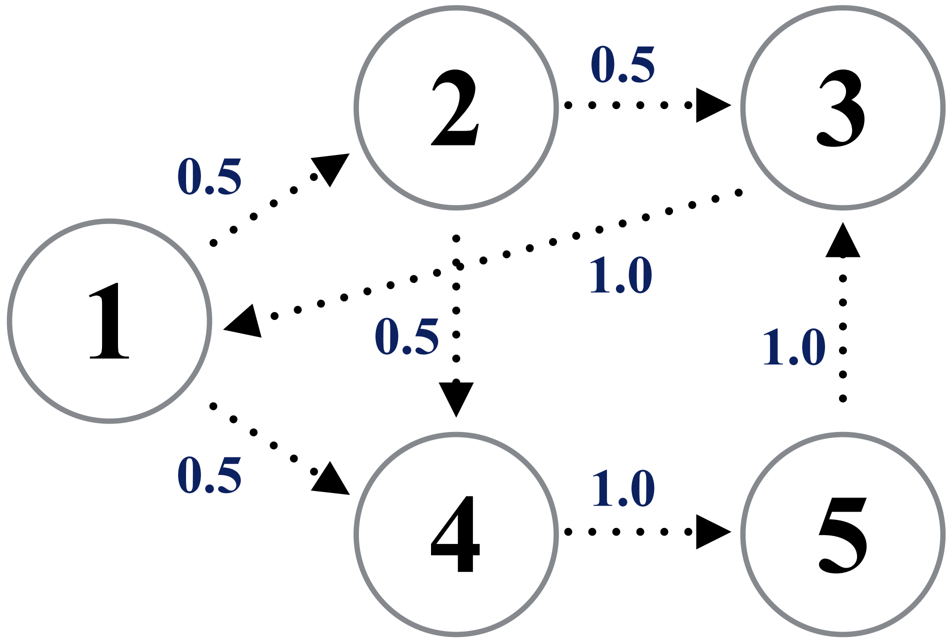


$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$S = \{1, \ldots, 5\}$$

Markov Chain (Example Revisited)



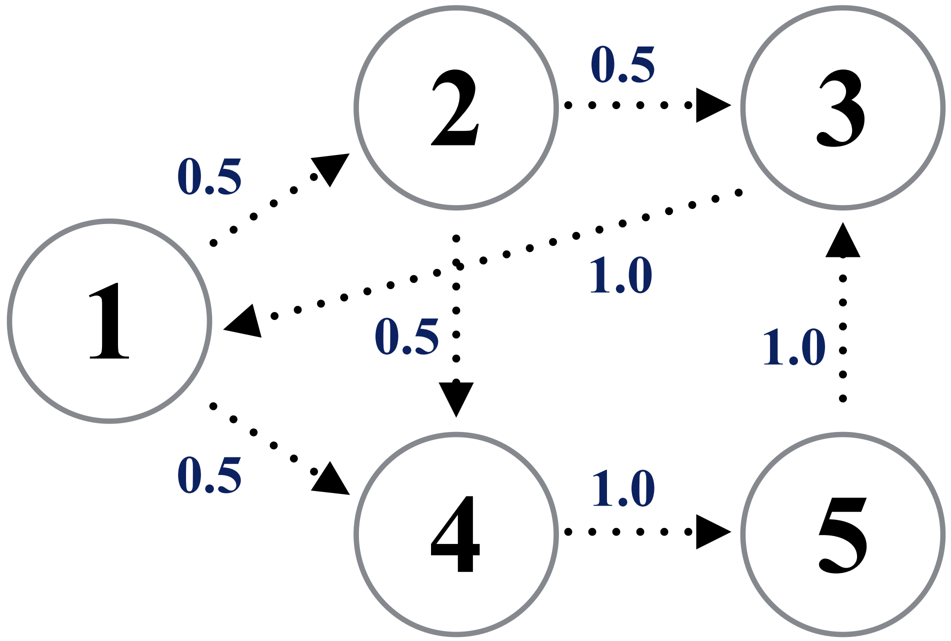
$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

$$S = \{1, \ldots, 5\}$$

Markov Chain (Example Revisited)



$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

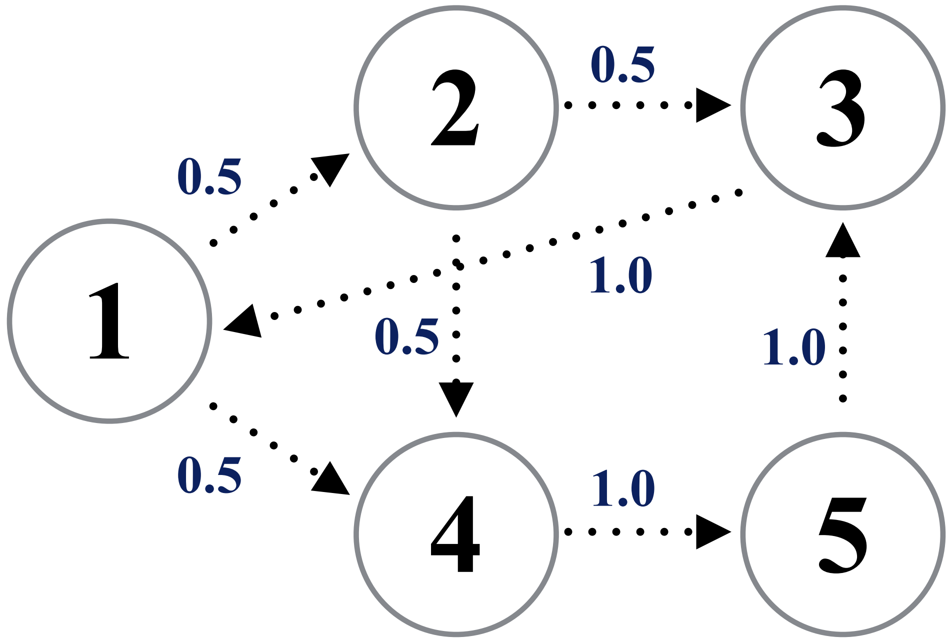
$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^2 = \begin{bmatrix} 0.0 & 0.0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$S = \{1, \ldots, 5\}$$

Markov Chain (Example Revisited)



$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

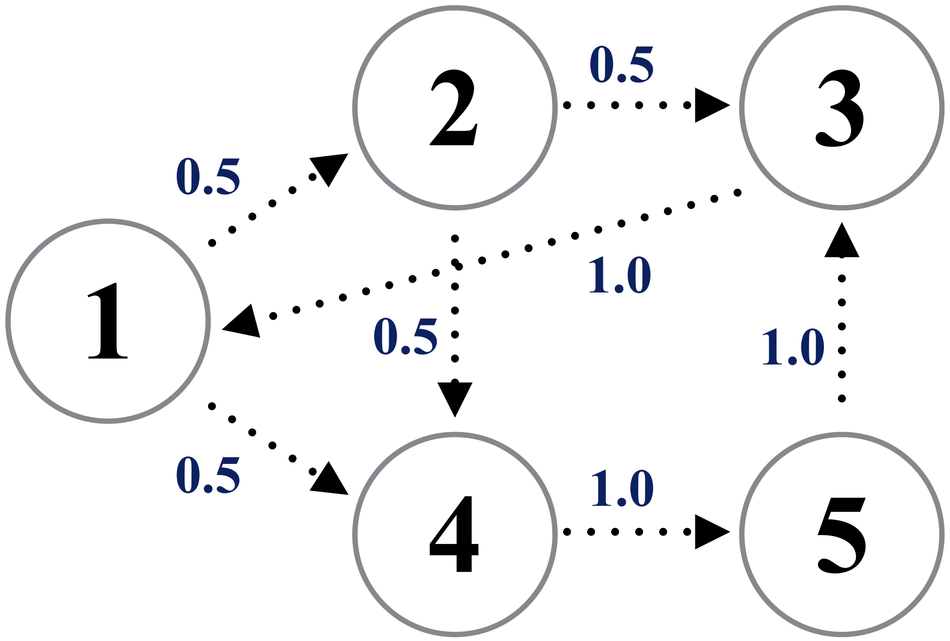
$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^2 = \begin{bmatrix} 0.0 & 0.0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$\boldsymbol{\pi}^3 = \begin{bmatrix} 0.25 & 0.0 & 0.5 & 0.0 & 0.25 \end{bmatrix}$$

$$S = \{1, \ldots, 5\}$$

Markov Chain (Example Revisited)



$$S = \{1, \ldots, 5\}$$

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

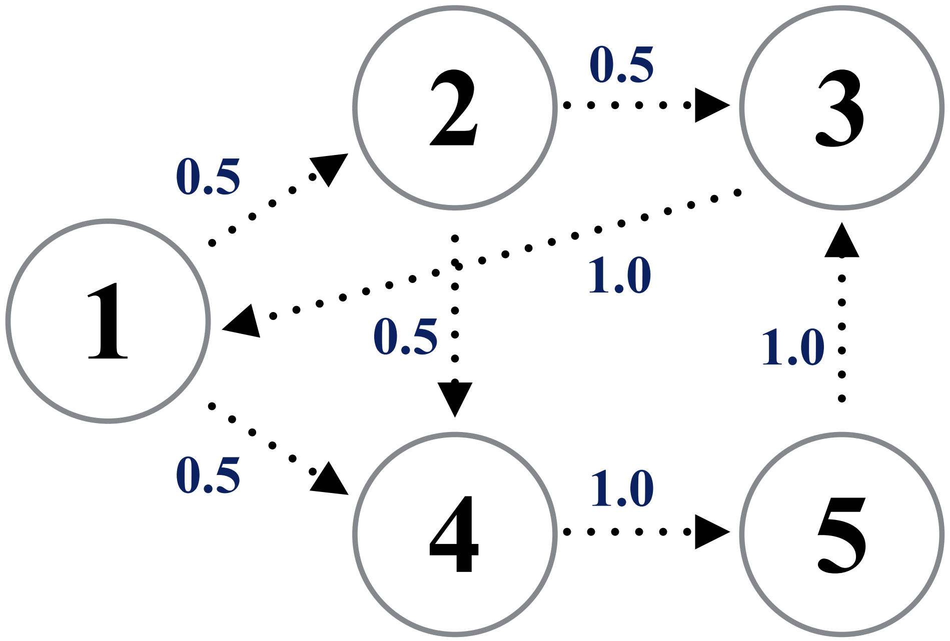
$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^2 = \begin{bmatrix} 0.0 & 0.0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$\boldsymbol{\pi}^3 = \begin{bmatrix} 0.25 & 0.0 & 0.5 & 0.0 & 0.25 \end{bmatrix}$$

$$\boldsymbol{\pi}^4 = \begin{bmatrix} 0.5 & 0.125 & 0.25 & 0.125 & 0 \end{bmatrix}$$

Markov Chain (Example Revisited)



$$S = \{1, \ldots, 5\}$$

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

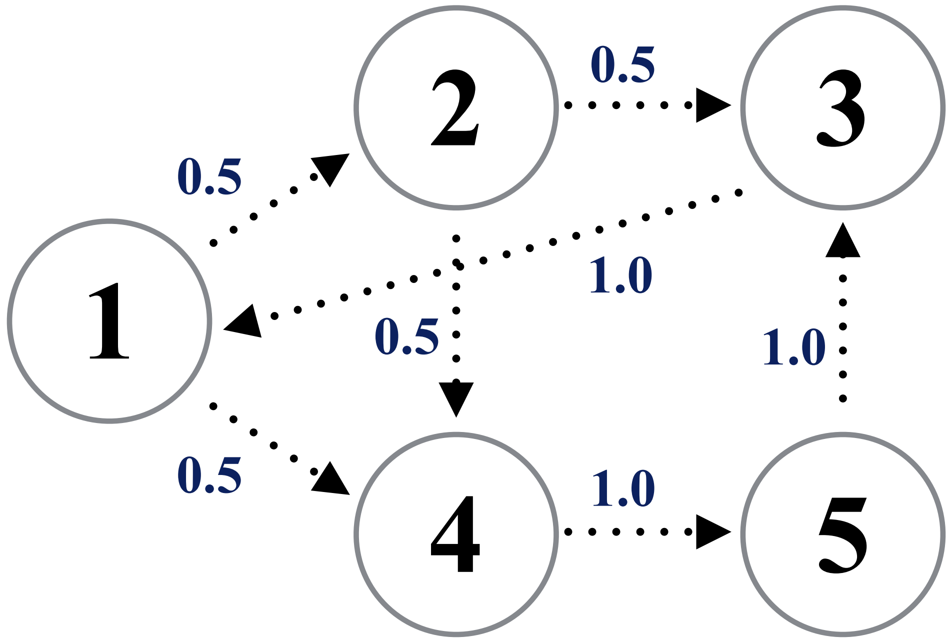
$$\boldsymbol{\pi}^2 = \begin{bmatrix} 0.0 & 0.0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$\boldsymbol{\pi}^3 = \begin{bmatrix} 0.25 & 0.0 & 0.5 & 0.0 & 0.25 \end{bmatrix}$$

$$\boldsymbol{\pi}^4 = \begin{bmatrix} 0.5 & 0.125 & 0.25 & 0.125 & 0 \end{bmatrix}$$

$$\boldsymbol{\pi}^5 = \begin{bmatrix} 0.25 & 0.25 & 0.0625 & 0.3125 & 0.125 \end{bmatrix}$$

Markov Chain (Example Revisited)



$$S = \{1, \ldots, 5\}$$

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^1 = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\pi}^2 = \begin{bmatrix} 0.0 & 0.0 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

$$\boldsymbol{\pi}^3 = \begin{bmatrix} 0.25 & 0.0 & 0.5 & 0.0 & 0.25 \end{bmatrix}$$

$$\boldsymbol{\pi}^4 = \begin{bmatrix} 0.5 & 0.125 & 0.25 & 0.125 & 0 \end{bmatrix}$$

$$\boldsymbol{\pi}^5 = \begin{bmatrix} 0.25 & 0.25 & 0.0625 & 0.3125 & 0.125 \end{bmatrix}$$

$$\vdots$$

$$\boldsymbol{\pi} = \begin{bmatrix} 0.25 & 0.125 & 0.25 & 0.1875 & 0.1875 \end{bmatrix}$$

Computing $\boldsymbol{\pi}$ (Method 1)

- Stationary state distribution is the limit distribution
- Idea: Compute k-step state probabilities $\boldsymbol{\pi}^{\mathrm{k}}$ until they converge

- **Power (iteration) method**
  - select arbitrary initial state probability distribution $\boldsymbol{\pi}^0$
  - compute $\boldsymbol{\pi}^k = \boldsymbol{\pi}^{k\text{-}1}\, \mathbf{P}$ until they converge (e.g., $|\, \boldsymbol{\pi}^k - \boldsymbol{\pi}^{k\text{-}1} \,| < \varepsilon$)
  - report last $\boldsymbol{\pi}^k$ as stationary state distribution $\boldsymbol{\pi}$

- Power (iteration) method basically **simulates the Markov chain** and is the **method of choice in practice** when dealing with huge state spaces, exploiting that **matrix-vector multiplication** is **easy to parallelize**

Computing $\boldsymbol{\pi}$ (Method 2)

- Stationary state distribution $\boldsymbol{\pi}$ fulfills $\boldsymbol{\pi} = \boldsymbol{\pi}\,\mathbf{P}$, which can be cast into a **system of linear equations**

$$\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{\pi}_1 &= 0.0 \times \boldsymbol{\pi}_1 + 0.0 \times \boldsymbol{\pi}_2 + 1.0 \times \boldsymbol{\pi}_3 + 0.0 \times \boldsymbol{\pi}_4 + 0.0 \times \boldsymbol{\pi}_5 \\
\boldsymbol{\pi}_2 &= 0.5 \times \boldsymbol{\pi}_1 + 0.0 \times \boldsymbol{\pi}_2 + 0.0 \times \boldsymbol{\pi}_3 + 0.0 \times \boldsymbol{\pi}_4 + 0.0 \times \boldsymbol{\pi}_5 \\
\boldsymbol{\pi}_3 &= 0.0 \times \boldsymbol{\pi}_1 + 0.5 \times \boldsymbol{\pi}_2 + 0.0 \times \boldsymbol{\pi}_3 + 0.0 \times \boldsymbol{\pi}_4 + 1.0 \times \boldsymbol{\pi}_5 \\
\boldsymbol{\pi}_4 &= 0.5 \times \boldsymbol{\pi}_1 + 0.5 \times \boldsymbol{\pi}_2 + 0.0 \times \boldsymbol{\pi}_3 + 0.0 \times \boldsymbol{\pi}_4 + 0.0 \times \boldsymbol{\pi}_5 \\
\boldsymbol{\pi}_5 &= 0.0 \times \boldsymbol{\pi}_1 + 0.0 \times \boldsymbol{\pi}_2 + 0.0 \times \boldsymbol{\pi}_3 + 1.0 \times \boldsymbol{\pi}_4 + 0.0 \times \boldsymbol{\pi}_5 \\
1 &= 1.0 \times \boldsymbol{\pi}_1 + 1.0 \times \boldsymbol{\pi}_2 + 1.0 \times \boldsymbol{\pi}_3 + 1.0 \times \boldsymbol{\pi}_4 + 1.0 \times \boldsymbol{\pi}_5
\end{aligned}$$

- Solutions can be found, e.g., using **Gauss elimination**

Computing $\boldsymbol{\pi}$ (Method 3)

- Stationary state probability distribution $\boldsymbol{\pi}$ is the left **eigenvector** of the **transition probability matrix** **P** for the **eigenvalue** $\lambda = 1$

$$\boldsymbol{\pi}\,\mathbf{P} = \boldsymbol{\lambda}\,\boldsymbol{\pi}$$

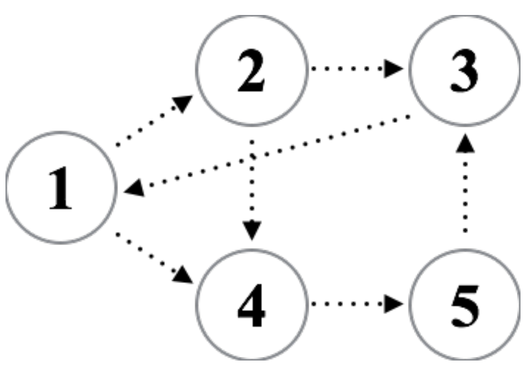
- Can be computed using the **characteristic polynomial**

$$(\mathbf{P} - \boldsymbol{\lambda}\,\mathbf{I})\,\boldsymbol{\pi} = \mathbf{0}$$

PageRank as a Markov Chain

- **Random surfer model**
  - follows a uniform random outgoing link with probability (1-ε)
  - jumps to a uniform random web page with probability ε
- Let **A** be the **adjacency matrix** of the Web graph, matrix **T** captures following of a uniform random outgoing link

$$\mathbf{T}_{ij} = \begin{cases} 1/out(i) & : \quad (i,j) \in E \\ 0 & : \quad \text{otherwise} \end{cases}$$

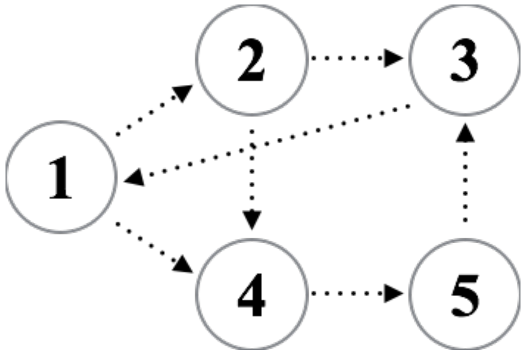


$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{T} = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 1/1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/1 \\ 0 & 0 & 1/1 & 0 & 0 \end{bmatrix}$$

PageRank as a Markov Chain

- **Random surfer model**
  - follows a uniform random outgoing link with probability (1-ε)
  - jumps to a uniform random web page with probability ε
- Vector **j** captures jumping to a uniform random web page

$$\mathbf{j}_i = 1/|V|$$



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

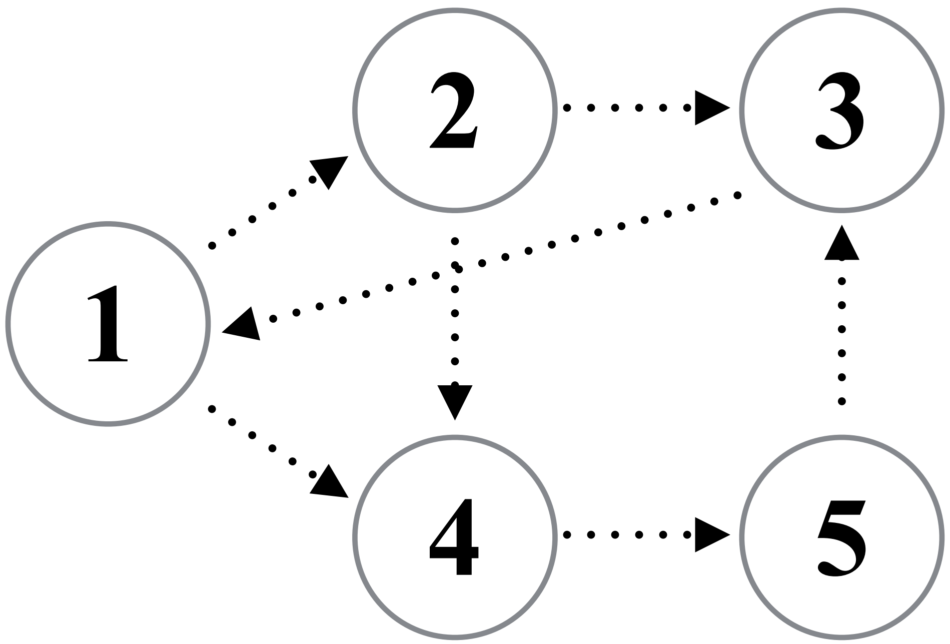
$$\mathbf{j} = \begin{bmatrix} 1/5 & \ldots & 1/5 \end{bmatrix}$$

- **Transition probability matrix** of Markov chain then obtained as

$$\mathbf{P} = (1 - \epsilon)\,\mathbf{T} + \epsilon \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^T \mathbf{j}$$

PageRank as a Markov Chain

- With ε = 0.15 we obtain



$$\mathbf{P} = \begin{bmatrix} 0.030 & 0.455 & 0.030 & 0.455 & 0.030 \\ 0.030 & 0.030 & 0.455 & 0.455 & 0.030 \\ 0.880 & 0.030 & 0.030 & 0.030 & 0.030 \\ 0.030 & 0.030 & 0.030 & 0.030 & 0.880 \\ 0.030 & 0.030 & 0.880 & 0.030 & 0.030 \end{bmatrix}$$

$$\boldsymbol{\pi} = \begin{bmatrix} 0.24079 & 0.13234 & 0.24799 & 0.18858 & 0.19029 \end{bmatrix}$$

PageRank as a Markov Chain

- **Transition probability matrix** of Markov chain then obtained as

$$\mathbf{P} = (1-\epsilon)\,\mathbf{T} + \epsilon \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}^T \mathbf{j}$$

$$\boldsymbol{\pi}_i = (1-\epsilon) \sum_{(j,i) \in E} \frac{\boldsymbol{\pi}_j}{out(j)} + \frac{\epsilon}{|V|}$$

- We need to deal with **dangling nodes** (having out-degree zero)
  - **Re-normalize** $\boldsymbol{\pi}^k$ such that $|\,\boldsymbol{\pi}^k\,| = 1$ after every iteration of power method
  - Make **P** truly right stochastic by defining matrix **T** as

$$\mathbf{T}_{ij} = \begin{cases} 1/out(i) & : \quad (i,j) \in E \\ 1/|V| & : \quad out(i) = 0 \\ 0 & : \quad \text{otherwise} \end{cases}$$

4

PageRank as a Markov Chain (Is It Ergodic?)

- Markov chain defined by transition probability matrix **T** is
  - **finite** (only finite number of web pages)
  - **time-homogeneous** (by design)
  - **irreducible** (random surfer can jump from every state *i* to every state *j*)
  - **positive recurrent** (random surfer can "jump up" on state *i*)
  - **aperiodic** (period of every state is 1 because of "jump up" on state *i*)

  …it is thus **ergodic** and unique stationary state probabilities $\boldsymbol{\pi}$ exist

- **Random jump** is **essential** to make the Markov chain ergodic

PageRank & Queries

- **Random jump probability** typically set as ε = 0.15 (i.e., random surfer follows on average about seven links in a row)

- PageRank determines a **static global ranking** of web pages, is **query-independent**, and **orthogonal to textual relevance**

- Combination of PageRank score and retrieval models, e.g., as
  - **linear combination** of cosine similarity and PageRank score

$$\alpha \times sim(q,d) + (1-\alpha) \times pr(d)$$

  - **document prior** in a query-likelihood language model

$$P(q|d) \times P(d)$$

  - together with many other features in **machine-learned ranking model**

Summary of IV.2

- **Markov chains**
  as a kind of stochastic process useful to describe random walks
- **Stationary state distribution**
  is guaranteed to exist if the Markov chain is finite and ergodic
  can be computed using (i) power iteration (ii) solving a system of linear equations or (iii) determining an eigenvector of a matrix
- **PageRank**
  as Google's initial secret of success
  is based on a random surfer model
  can be described as a finite and ergodic Markov chain
  yields a static query-independent importance score

Additional Literature for IV.2

- **S. Brin and L. Page**: *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks 30:107-117, 1998
- **M. Bianchini, M. Gori, and F. Scarselli**: *Inside PageRank*, ACM TOIT 5(1):92-128, 2005
- **M. Franceschet**: *PageRank: Standing on the Shoulders of Giants*, CACM 54(6):92-101, 2011
- **A. N. Meyer and C. D. Meyer**: *Survey: Deeper Inside PageRank*, Internet Mathematics 1(3):335-380, 2003
- **L. Page, S. Brin, R. Motwani, and T. Winograd**: *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford University, 1999