Data Mining and Matrices

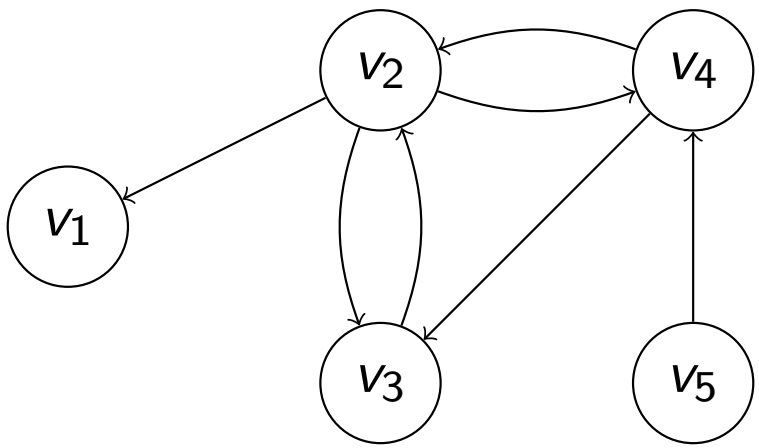
10 – Graphs II

Rainer Gemulla, Pauli Miettinen

Jul 4, 2013

Link analysis

- The web as a directed graph
  - Set of web pages with associated textual content
  - Hyperlinks between webpages (potentially with anchor text)
  - $\rightarrow$ Directed graph
- Our focus: Which pages are "relevant" (to a query)?
  - Analysis of link structure instrumental for web search
  - Assumption: incoming link is a quality signal (endorsement)
  - Page has high quality $\approx$ links from/to high-quality pages
  - (We are ignoring anchor text in this lecture.)
- Gives rise to HITS and PageRank algorithms
- Similarly: citations of scientific papers, social networks, ...



Outline

Eigenvectors and diagonalizable matrices

- Denote by $\mathbf{A}$ an $n \times n$ real matrix
- Recap eigenvectors
  - $\mathbf{v}$ is a **right eigenvector** with eigenvalue λ of $\mathbf{A}$ if $\mathbf{Av} = \lambda\mathbf{v}$
  - $\mathbf{v}$ is a **left eigenvector** with eigenvalue λ of $\mathbf{A}$ if $\mathbf{vA} = \lambda\mathbf{v}$
  - If $\mathbf{v}$ is a right eigenvector of $\mathbf{A}$, then $\mathbf{v}^T$ is a left eigenvector of $\mathbf{A}^T$ (and vice versa)
- $\mathbf{A}$ is **diagonalizable** if it has n linearly independent eigenvectors
  - Some matrices are not diagonalizable (called **defective**)
  - If $\mathbf{A}$ is symmetric (our focus), it is diagonalizable
  - If $\mathbf{A}$ is symmetric, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ can be chosen to be real and orthonormal $\rightarrow$ These eigenvectors then form an orthonormal basis of $\mathbb{R}^n$
  - Denote by $\lambda_1, \ldots, \lambda_n$ are the corresponding eigenvalues (potentially 0)
  - Then for every $\mathbf{x} \in \mathbb{R}^n$, there exist $c_1, \ldots, c_n$ such that

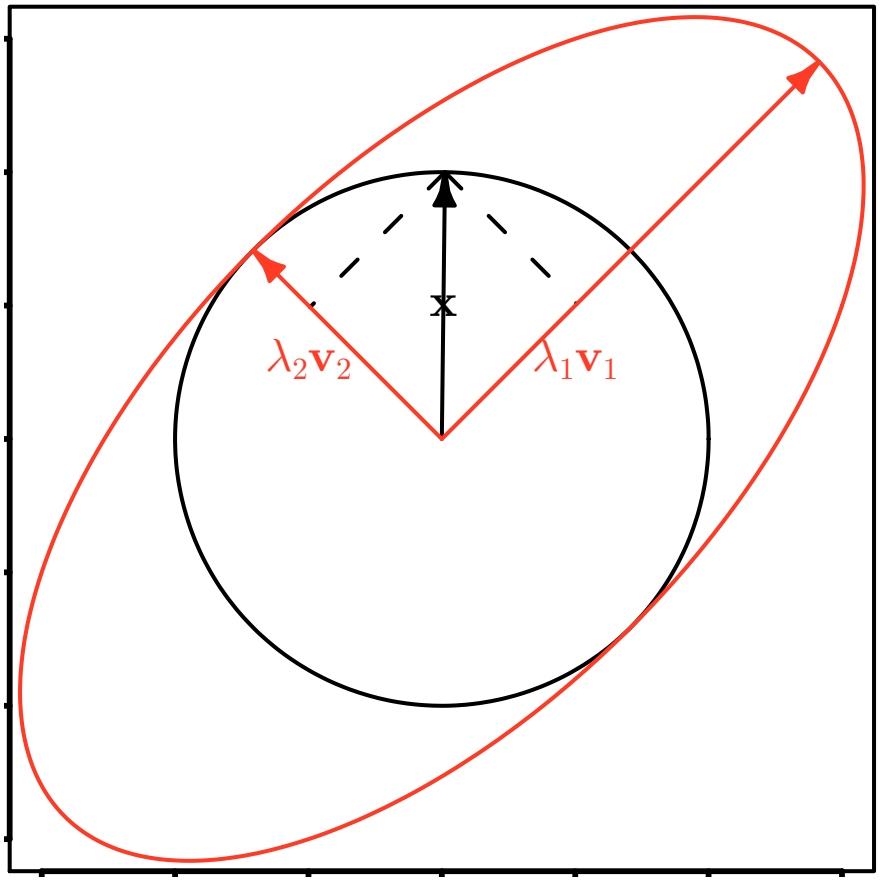
$$\mathbf{x} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$$

  - And therefore

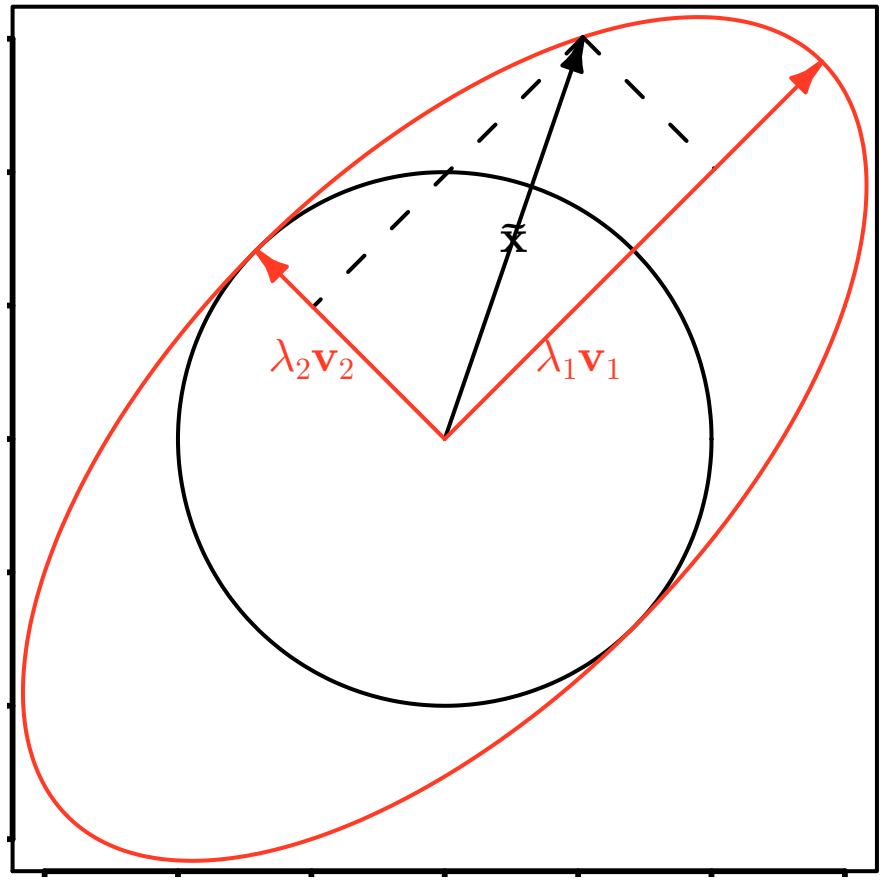
$$\mathbf{Ax} = \lambda_1 c_1\mathbf{v}_1 + \lambda_2 c_2\mathbf{v}_2 + \cdots + \lambda_n c_n\mathbf{v}_n$$

  - Eigenvectors "explain" effect of linear transformation $\mathbf{A}$

Example



$\lambda_1 = 2,\ \lambda_2 = 1$



$\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$

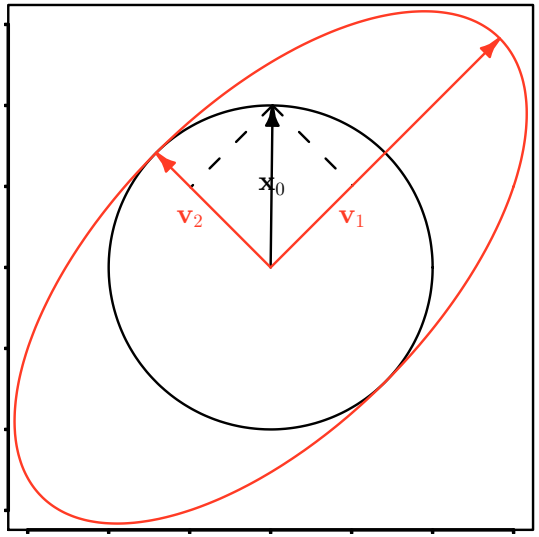
Power method

- Simple method to determine the largest eigenvalue λ_1 and the corresponding eigenvector $\mathbf{v}_1$
- Algorithm
  1. Start at some $\mathbf{x}_0$
  2. While not converged
     1. Set $\tilde{\mathbf{x}}_{t+1} \leftarrow \mathbf{A}\mathbf{x}_t$
     2. Normalize: $\mathbf{x}_{t+1} \leftarrow \tilde{\mathbf{x}}_{t+1}/\|\tilde{\mathbf{x}}_{t+1}\|$
- What happens here?
  - Observe that $\mathbf{x}_t = \mathbf{A}^t\mathbf{x}_0/C$, where $C = \|\mathbf{A}^t\mathbf{x}_0\|$
  - Assume that $\mathbf{A}$ is real symmetric
  - Then $\mathbf{x}_t = (\lambda_1^t c_1\mathbf{v}_1 + \lambda_2^t c_2\mathbf{v}_2 + \cdots + \lambda_n^t c_n\mathbf{v}_n)/C$
  - If $|\lambda_1| > |\lambda_2|$, then

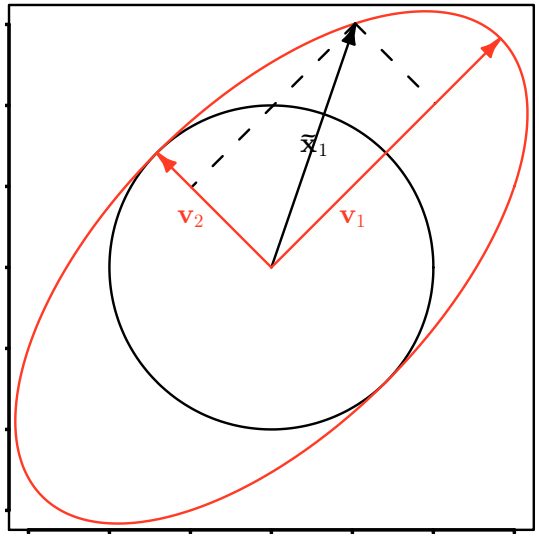
$$\lim_{t\to\infty} \frac{\lambda_2^t c_2}{\lambda_1^t c_1} = \lim_{t\to\infty} \left(\frac{\lambda_2}{\lambda_1}\right)^t \frac{c_2}{c_1} = 0$$

  - So as $t \to \infty$, $\mathbf{x}_t$ converges to $\mathbf{v}_1$

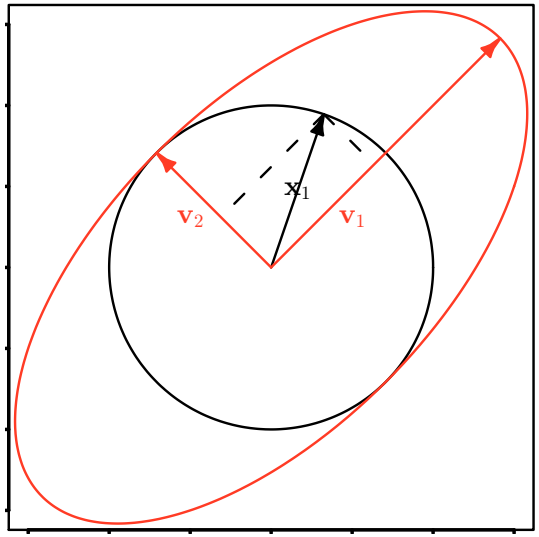
Power method (example)



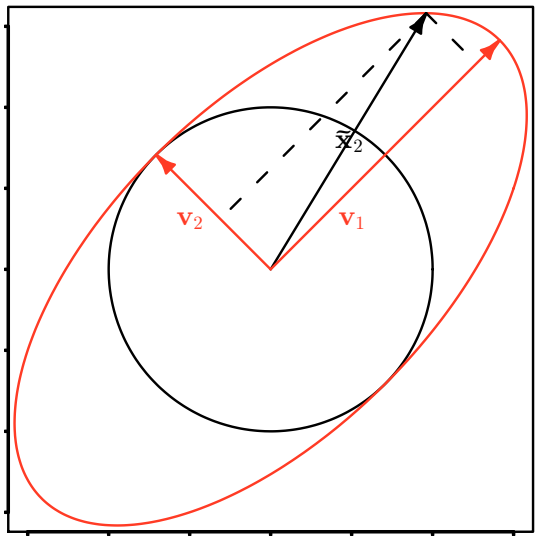
$n = 0$



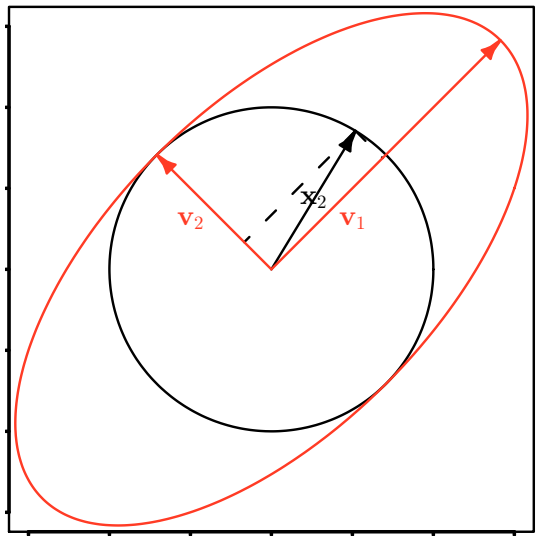
$n = 1$



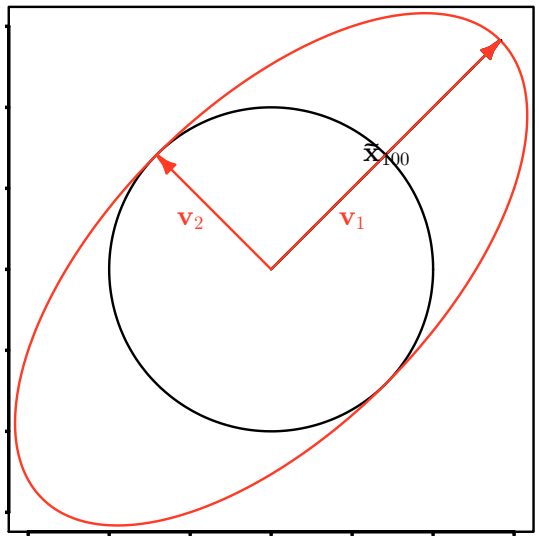
$n = 1$ (normalized)



$n = 2$



$n = 2$ (normalized)



$n = 100$

Discussion

- Easy to implement and parallelize
- We will see: useful for understanding link analysis
- Convergence
  - Works if $\mathbf{A}$ is real symmetric, $|\lambda_1| > |\lambda_2|$, and $\mathbf{x}_0 \not\perp \mathbf{v}_1$ (i.e., $c_1 \neq 0$)
  - Speed depends on *eigengap* $|\lambda_1|/|\lambda_2|$
  - Also works in many other settings (but not always)

Power method and singular vectors

- Unit vectors $\mathbf{u}$ and $\mathbf{v}$ are **left and right singular vectors** of $\mathbf{A}$ if

$$\mathbf{A}^T\mathbf{u} = \sigma\mathbf{v} \qquad \text{and} \qquad \mathbf{A}\mathbf{v} = \sigma\mathbf{u}$$

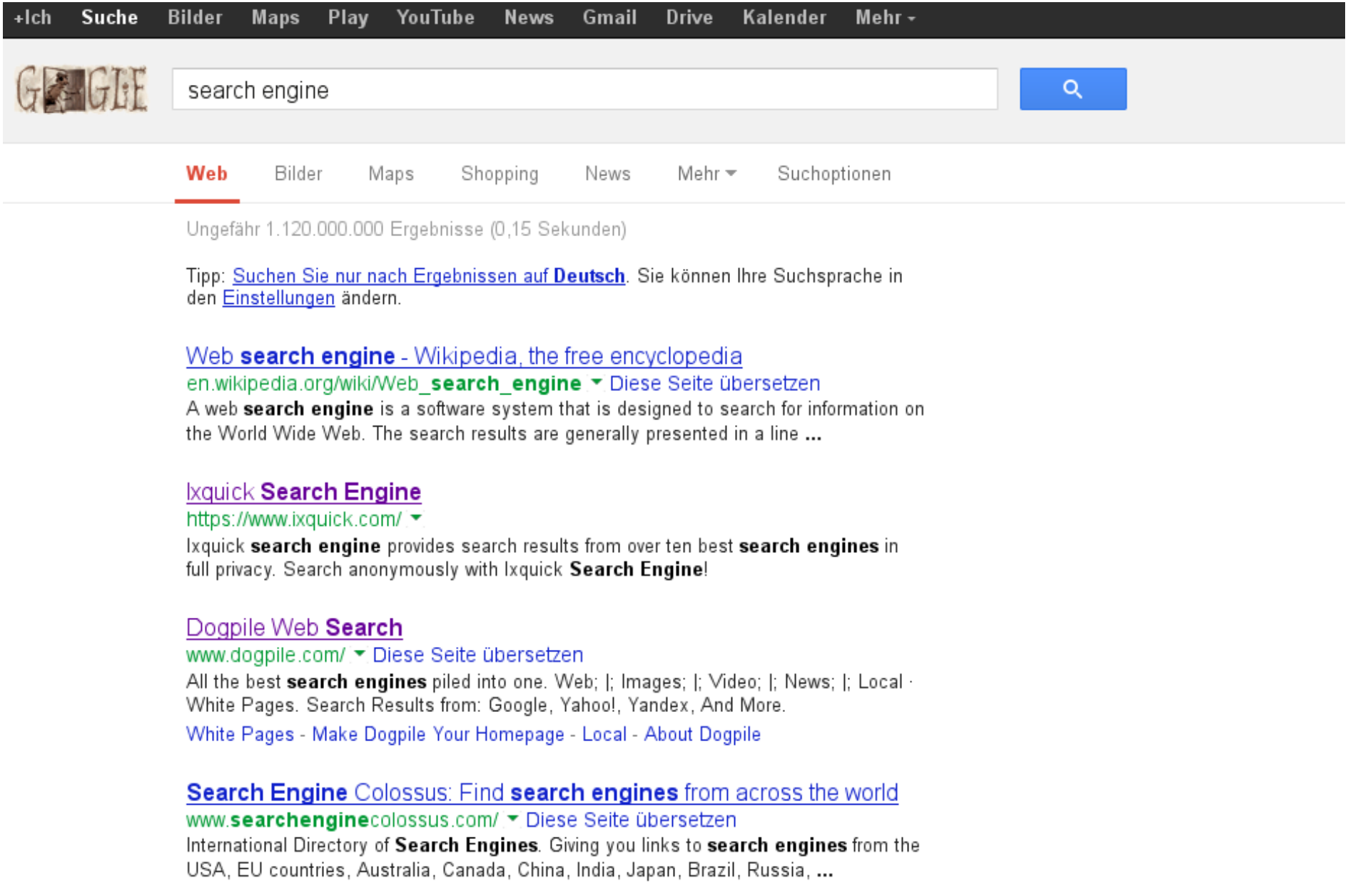
- σ is the corresponding **singular value**
- The SVD decomposition is formed of the singular values ($\mathbf{\Sigma}$) and corresponding left and right singular vectors (columns of $\mathbf{U}$ and $\mathbf{V}$)
- $\mathbf{u}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$ with eigenvalue σ^2 since

$$\mathbf{A}\mathbf{A}^T\mathbf{u} = \mathbf{A}\sigma\mathbf{v} = \sigma\mathbf{A}\mathbf{v} = \sigma^2\mathbf{u}$$

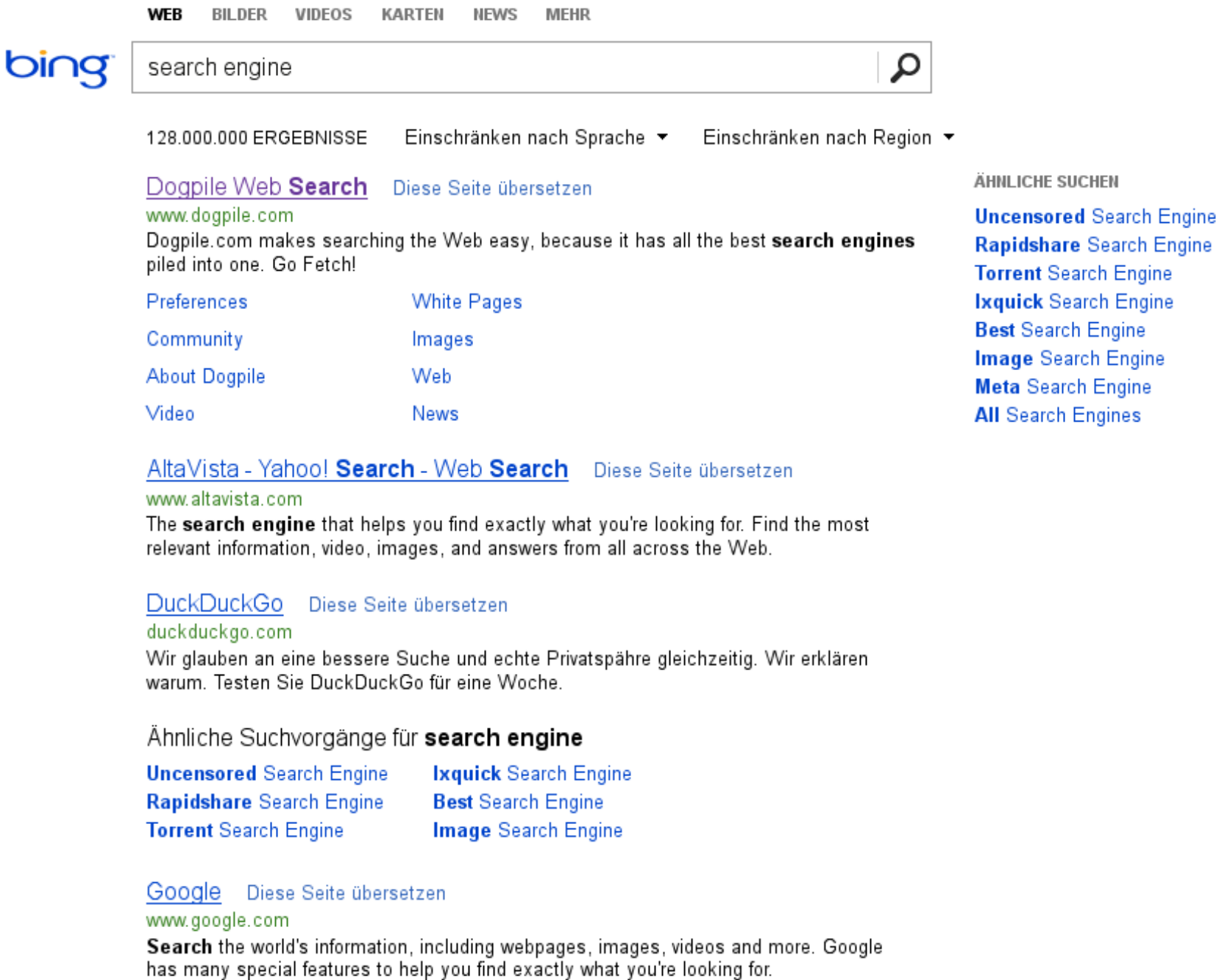
- Similarly $\mathbf{v}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$ with eigenvalue σ^2
- Power method for principal singular vectors
  1. $\mathbf{u}_{t+1} \leftarrow \mathbf{A}\mathbf{v}_t \,/\, \|\mathbf{A}\mathbf{v}_t\|$
  2. $\mathbf{v}_{t+1} \leftarrow \mathbf{A}^T\mathbf{u}_{t+1} \,/\, \|\mathbf{A}^T\mathbf{u}_{t+1}\|$
- Why does it work?
  - $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric (and positive semi-definite)
  - $\mathbf{u}_{t+2} = \mathbf{A}\mathbf{v}_{t+1} \,/\|\mathbf{A}\mathbf{v}_{t+1}\| \; = \mathbf{A}\mathbf{A}^T\mathbf{u}_{t+1} \,/\, \|\mathbf{A}\mathbf{A}^T\mathbf{u}_{t+1}\|$

Outline

Asking Google for search engines



Asking Bing for search engines



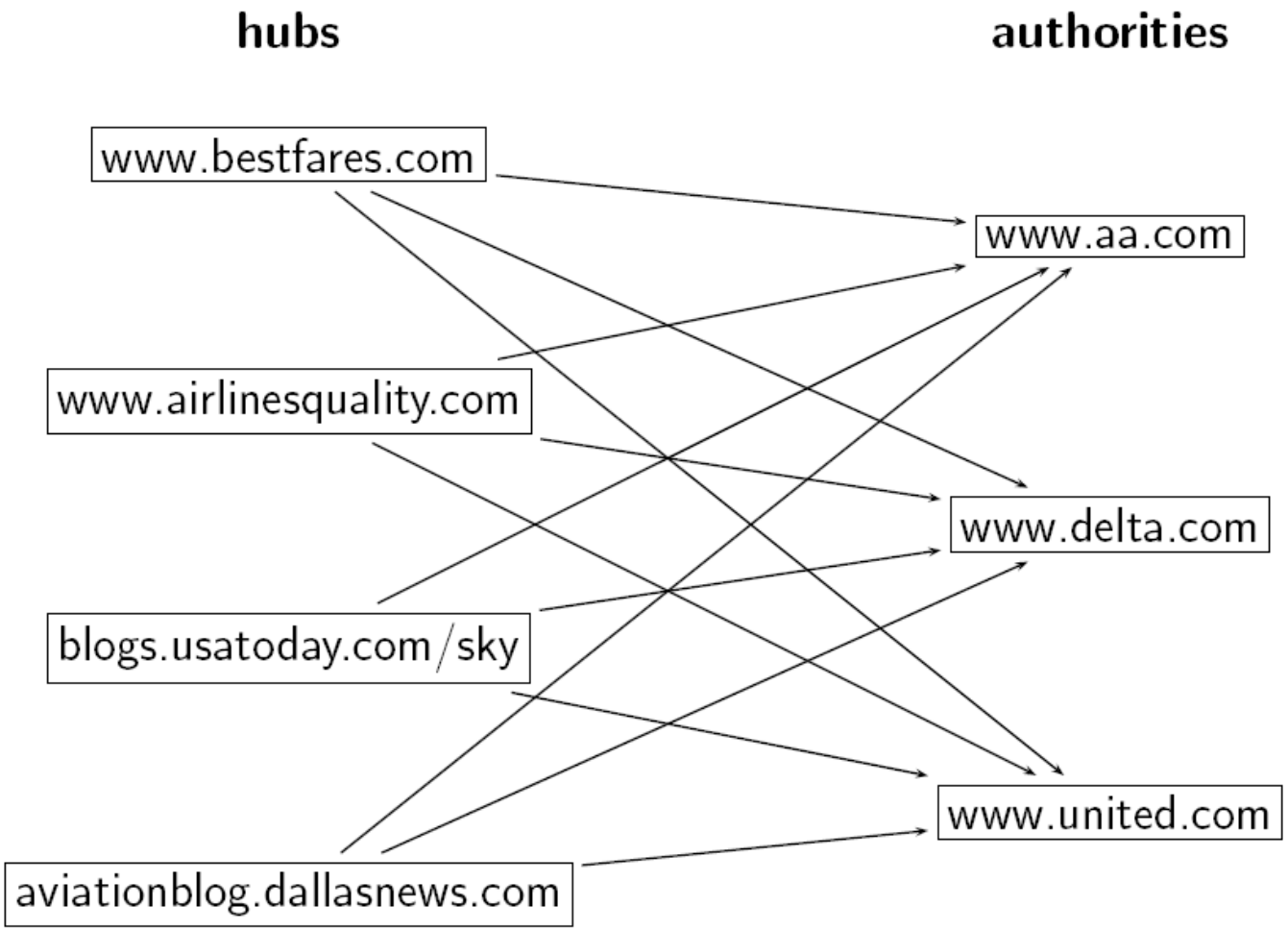
Searching the WWW

- Some difficulties in web search
  - "search engine": many of the search engines do not contain phrase "search engine"
  - "Harvard": millions of pages contain "Harvard", but `www.harvard.edu` may not contain it most often
  - "lucky": there is an "I'm feeling lucky" button on `google.com`, but `google.com` is (probably) not relevant (popularity)
  - "automobile": some pages say "car" instead (synonymy)
  - "jaguar": the car or the animal? (polysemy)
- Query types
  1. Specific queries ("name of Michael Jackson's dog")
     → Scarcity problem: few pages contain required information
  2. Broad-topic queries ("Java")
     → Abundance problem: large number of relevant pages
  3. Similar-page queries ("Pages similar to `java.com`")
- Our focus: broad-topic queries
  - Goal is to find "most relevant" pages

Hyperlink Induced Topic Search (HITS)

- HITS analyzes the link structure to mitigate these challenges
  - Uses links as source of exogenous information
  - Key idea: If p links to q, p confers "authority" on q
    $\rightarrow$ Try to find authorities through links that point to them
  - HITS aims to balance between relevance to a query (content) and popularity (in-links)
- HITS uses two notions of relevance
  - **Authority** page directly answers information need
    $\rightarrow$ Page pointed to by many hubs for the query
  - **Hub** page contains link to pages that answer information need
    $\rightarrow$ Points to many authorities for the query
  - Note: circular definition
- Algorithm
  1. Create a focused subgraph of the WWW based on the query
  2. Score each page w.r.t. to authority and hub
  3. Return the pages with the largest authority scores

Hubs and authorities (example)

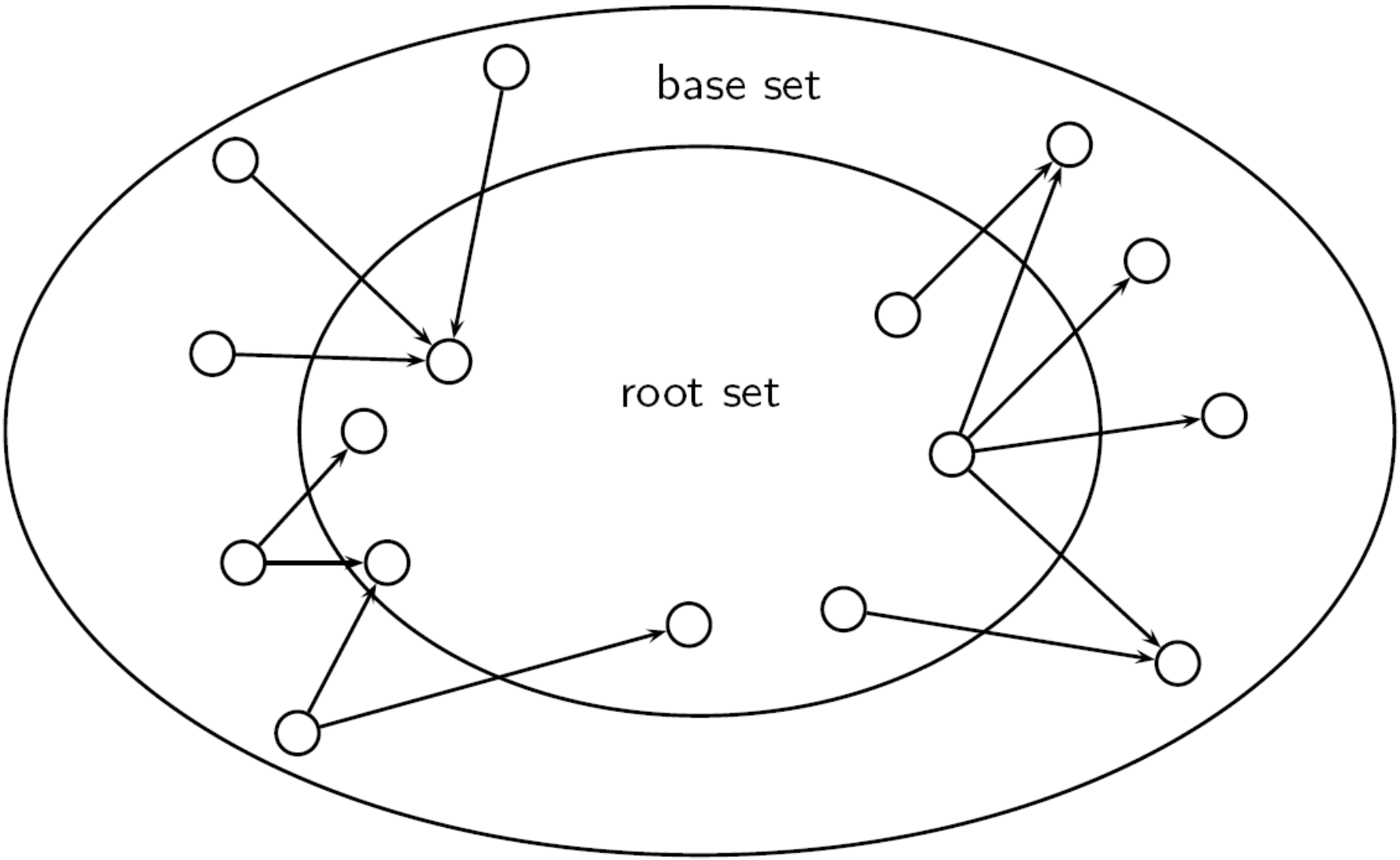


Manning et al., 2008

Creating a focused subgraph

- Desiderata
  1. Should be small (for efficiency)
  2. Should contain most (or many) of the strongest authorities (for recall)
  3. Should be rich in relevant pages (for precision)
- Using all pages that contain query may violate (1) and (2)
- Construction
  - **Root set**: the highest-ranked pages for the query (regular web search)
    $\rightarrow$ Satisfies (1) and (3), but often not (2)
  - **Base set**: pages that point to or are pointed to from the root set
    $\rightarrow$ Increases number of authorities, addressing (2)
  - **Focused subgraph** = induced subgraph of base set
    $\rightarrow$ Consider all links between pages in the base set

Root set and base set



Heuristics

- Retain efficiency
  - Focus on t highest ranked pages for the query (e.g., $t = 200$)
    $\rightarrow$ Small root set
  - Allow each page to bring in at most d pages pointing to it (e.g., $d = 50$)
    $\rightarrow$ Small base set ($\approx$ 5000 pages)
- Try to avoid links that serve a purely navigational function
  - E.g., link to homepage
  - Keep **transverse** links (to different domain)
  - Ignore **intrinsic** links (to same domain)
- Try to avoid links that indicate collusion/advertisement
  - E.g., "This site is designed by..."
  - Allow each page to be pointed to at most m times from each domain ($m \approx$ 4–8)

Hubs and authorities

- Simple approach: rank pages by in-degrees in focused subgraph
  - Works better than on whole web
  - Still problematic: some pages are "universally popular" regardless of underlying query topic
- Key idea: weight links from different pages differently
  - Authoritative pages have high in-degree and a common topic
    $\rightarrow$ Considerable overlap in sets of pages that point to authorities
  - Hub pages "pull together" authorities on a common topic
    $\rightarrow$ Considerable overlap in sets of pages that are pointed to by hubs
  - **Mutual reinforcment**
    - ★ Good hub points to many good authorities
    - ★ Good authority is pointed to by many good hubs

Hub and authority scores

- Denote by $G = (V, E)$ the focused subgraph
- Assign to page p
  - A non-negative **hub weight** u_p
  - A non-negative **authority weight** v_p
- Larger means "better"
- Authority weight = sum of weights of hubs pointing to the page

$$v_p \leftarrow \sum_{(q,p) \in E} u_q$$

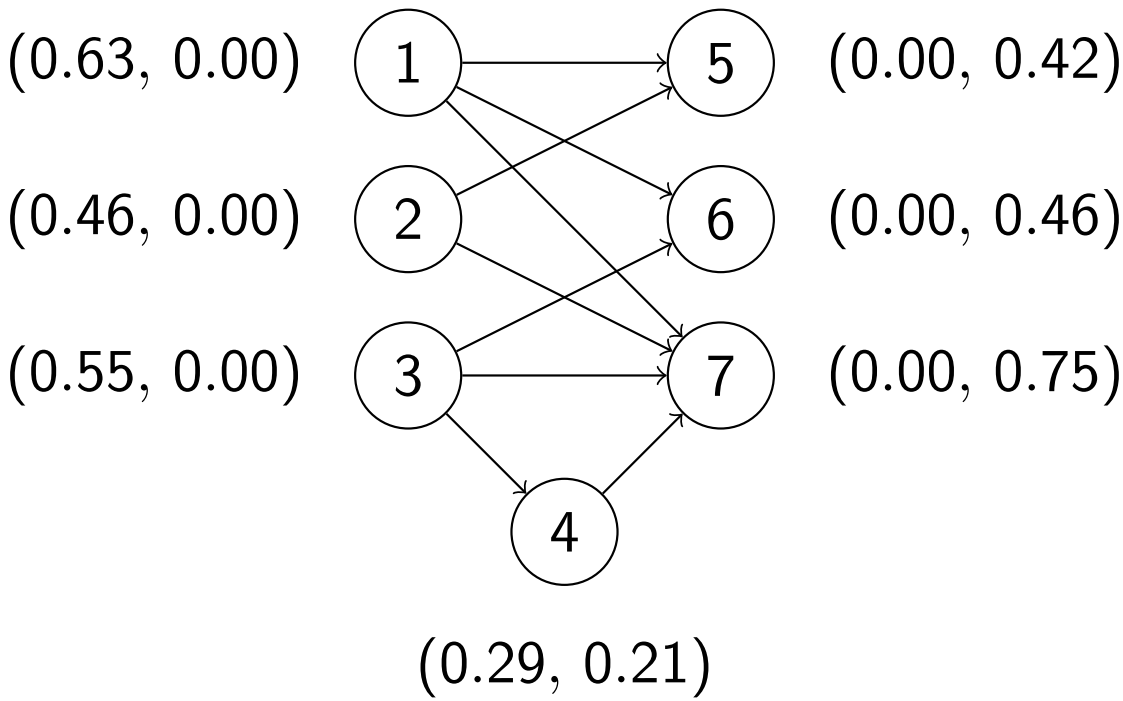
- Hub weight = sum of weights of authorities pointed to by the page

$$u_p \leftarrow \sum_{(p,q) \in E} v_p$$

- HITS iterates until it reaches a fixed point
  - Normalize vectors to length 1 after every iteration (does not affect ranking)

Example

- $\mathbf{u} = \begin{pmatrix} 0.63 & 0.46 & 0.55 & 0.29 & 0.00 & 0.00 & 0.00 \end{pmatrix}^T$ (hubs)
- $\mathbf{v} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.21 & 0.42 & 0.46 & 0.75 \end{pmatrix}^T$ (authorities)



Authorities for Chicago Bulls

0.85 www.nba.com/bulls

0.25 www.essex1.com/people/jmiller/bulls.htm
"da Bulls"

0.20 www.nando.net/SportServer/basketball/nba/chi.html
"The Chicago Bulls"

0.15 users.aol.com/rynocub/bulls.htm
"The Chicago Bulls Home Page"

0.13 www.geocities.com/Colosseum/6095
"Chicago Bulls"

Top-authority for Chicago Bulls

Manning et al., 2008

Hubs for Chicago Bulls

1.62 www.geocities.com/Colosseum/1778
"Unbelieveabulls!!!!!"

1.24 www.webring.org/cgi-bin/webring?ring=chbulls
"Erin's Chicago Bulls Page"

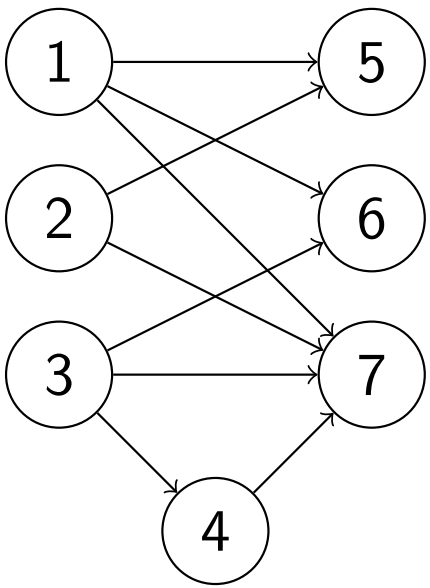
0.74 www.geocities.com/Hollywood/Lot/3330/Bulls.html
"Chicago Bulls"

0.52 www.nobull.net/web_position/kw-search-15-M2.htm
"Excite Search Results: bulls"

0.52 www.halcyon.com/wordsltd/bball/bulls.htm
"Chicago Bulls Links"

What happens here?

- Adjacency matrix $\mathbf{A}$ ($\mathbf{A}_{pq} = 1$ if p links to q)
  - $v_p \leftarrow \sum_{(q,p)\in E} u_q = (\mathbf{A}_{*p})^T \mathbf{u}$
  - Thus: $\mathbf{v} \leftarrow \mathbf{A}^T \mathbf{u}$
  - Similarly $\mathbf{u} \leftarrow \mathbf{A}\mathbf{v}$
- This is the power method for principal singular vectors
  - $\mathbf{u}$ and $\mathbf{v}$ correspond to principal left and right singular vectors of $\mathbf{A}$
  - $\mathbf{u}$ is principal eigenvector of $\mathbf{A}\mathbf{A}^T$ (co-citation matrix)
  - $\mathbf{v}$ is principal eigenvector of $\mathbf{A}^T\mathbf{A}$ (bibliographic coupling matrix)



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

Discussion

- Hub and authority weights depend on query
  $\rightarrow$ Scores need to be computed online
- HITS can find relevant pages regardless of content
  - Pages in base set often do not contain query keywords
  - Once base set is constructed, we only do link analysis
- Potential topic drift
  - Pages in base set may not be relevant to the topic
  - May also return Japanese pages for English query (if appropriately connected)
- Sensitive to manipulation
  - E.g., adversaries can create densely coupled hub and authority pages

Outline

Markov chains

- A **stochastic process** is family of random variables $\{ X_t : t \in T \}$
  - Here: $T = \{ 1, 2, \dots \}$ and t is called *time*
  - Thus we get sequence $X_1, X_2, \dots$
  - Instance of a discrete-time stochastic process
- $\{ X_t \}$ is **Markov chain** if it is memory-less

$$P(X_{t+1} = j \mid X_1 = i_1, \dots, X_{t-1} = i_{t-1}, X_t = i) = P(X_{t+1} = j \mid X_t = i)$$

- If $X_t = i$, we say that Markov chain is in **state** i at time t

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Properties
Coin flips	0	1	1	0	1	1	1	0	MC
Invert	0	1	0	1	0	1	0	1	MC
First-one	0	1	1	1	1	1	1	1	MC
1 on odd time	0	0	1	0	1	0	1	0	MC
Sum	0	1	2	2	3	4	5	5	MC
Sum (2-window)	0	1	2	1	1	2	2	1	¬MC

Finiteness and time-homogeneity

- Markov chain is **finite** if it has a finite number of states
- Markov chain is **time-homogeneous** if

$$P(X_{t+1} = j \mid X_t = i) = P(X_1 = j \mid X_0 = i)$$

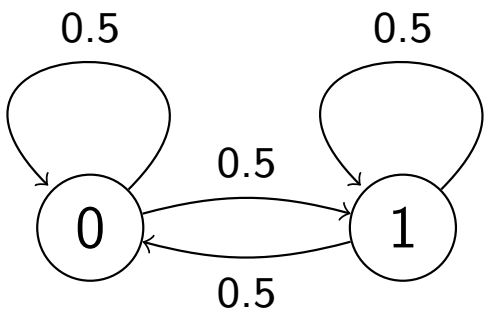
- We assume finite, time-homogeneous Markov chains from now on

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Properties
Coin flips	0	1	1	0	1	1	1	0	MC, F, TH
Invert	0	1	0	1	0	1	0	1	MC, F, TH
First-one	0	1	1	1	1	1	1	1	MC, F, TH
1 on odd time	0	0	1	0	1	0	1	0	MC, F, ¬TH
Sum	0	1	2	2	3	4	5	5	MC, ¬F, TH
Sum (2-window)	0	1	2	1	1	2	2	1	¬MC

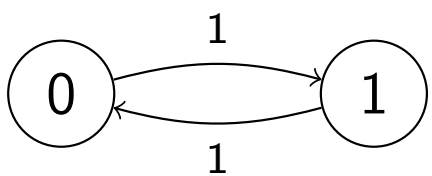
Markov chains and graphs

Markov chains can be represented as graph

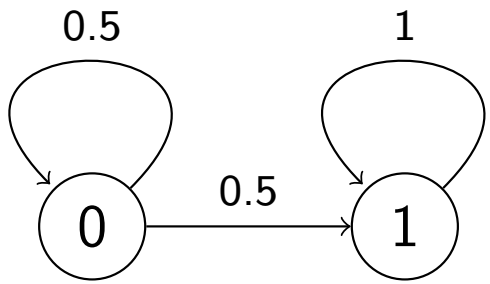
- V = set of states
- $(i, j) \in E$ if $P(X_1 = j \mid X_0 = i) > 0$
- $w_{ij} = P(X_1 = j \mid X_0 = i)$



Coin flips



Invert



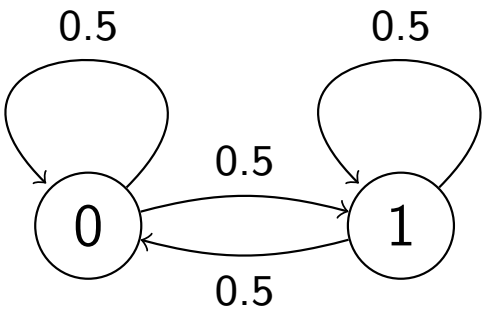
First-one

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Properties
Coin flips	0	1	1	0	1	1	1	0	MC, F, TH
Invert	0	1	0	1	0	1	0	1	MC, F, TH
First-one	0	1	1	1	1	1	1	1	MC, F, TH

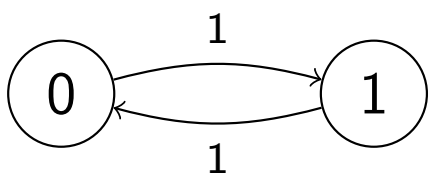
Irreducibility and aperiodicity

A Markov chain is

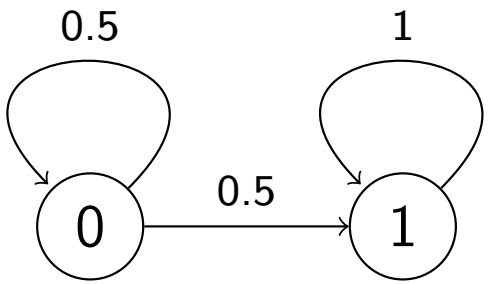
- **irreducible**: for all $i, j \in V$, there is a path from i to j
- **aperiodic**: for all i, $\gcd\{ t : P(X_t = i \mid X_0 = i) > 0 \} = 1$



Coin flips



Invert

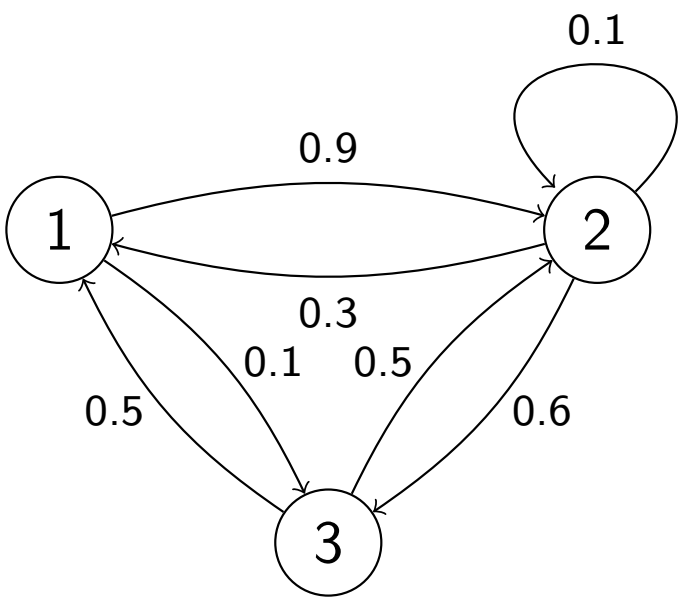


First-one

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Properties
Coin flips	0	1	1	0	1	1	1	0	MC, F, TH, I, A
Invert	0	1	0	1	0	1	0	1	MC, F, TH, I, ¬A
First-one	0	1	1	1	1	1	1	1	MC, F, TH, ¬I

Transition matrix

- Consider the graph of a Markov chain
- Associated adjacency matrix $\mathbf{P}$ is called **transition matrix**
  - $\mathbf{P}$ is row-stochastic (rows sum to 1)



$$\mathbf{P} = \begin{pmatrix} 0 & 0.9 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

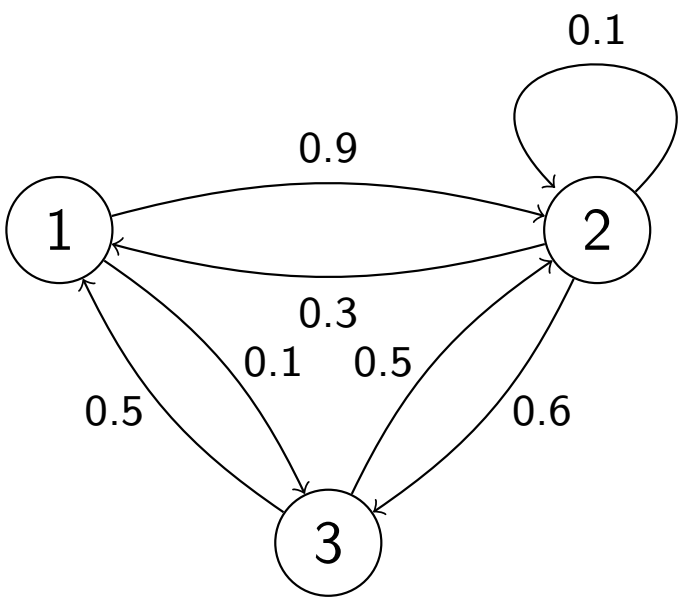
Surfing the chain

- $\mathbf{p}_t = \begin{pmatrix} p_{t,1} & \cdots & p_{t,n} \end{pmatrix}$ = distribution of states after t steps
  - I.e., $p_{t,i} = P(X_t = i)$
  - $\mathbf{p}_0$ is initial distribution
- After one step, we have

$$\mathbf{p}_{t+1,j} = \sum_i P(X_t = i) P(X_{t+1} = j \mid X_t = i) = \sum_i p_{t,i} \mathbf{P}_{ij} = \mathbf{p}_t \mathbf{P}_{*j}$$

$$\mathbf{p}_{t+1} = \mathbf{p}_t \mathbf{P}$$

- After k steps, we have $\mathbf{p}_{t+k} = \mathbf{p}_t \mathbf{P}^k$

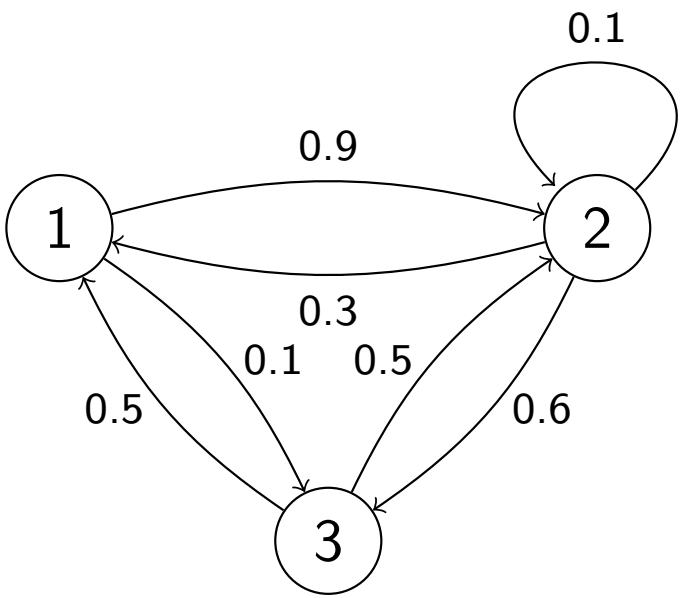


$$\mathbf{P} = \begin{pmatrix} 0 & 0.9 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

$$\mathbf{p}_0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$
$$\mathbf{p}_1 = \begin{pmatrix} 0 & 0.9 & 0.1 \end{pmatrix}$$
$$\mathbf{p}_2 = \begin{pmatrix} 0.32 & 0.13 & 0.54 \end{pmatrix}$$
$$\mathbf{p}_3 = \begin{pmatrix} 0.31 & 0.57 & 0.12 \end{pmatrix}$$

Stationary distribution

- Distribution π satisfying $\pi = \pi\mathbf{P}$ is called **stationary distribution**
  $\rightarrow$ Distribution does not change if we make more steps
- Unique stationary distribution exists if chain is irreducible
- If additionally aperiodic, $\lim_{k\to\infty} \mathbf{p}_0\mathbf{P}^k = \pi$ for any distribution $\mathbf{p}_0$
  - This is just the power method
  - π is the principal eigenvector of $\mathbf{P}$
  - Corresponding eigenvalue is 1 and has multiplicity 1



$$\mathbf{P} = \begin{pmatrix} 0 & 0.9 & 0.1 \\ 0.3 & 0.1 & 0.6 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

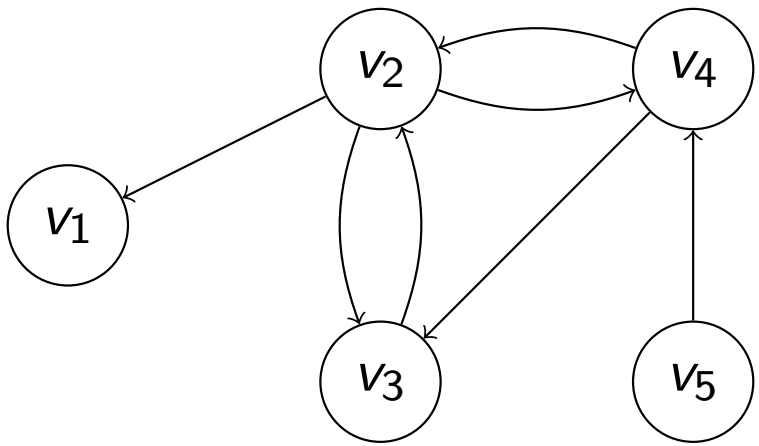
$$\begin{aligned} \mathbf{p}_0 &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \\ \mathbf{p}_1 &= \begin{pmatrix} 0 & 0.9 & 0.1 \end{pmatrix} \\ \mathbf{p}_2 &= \begin{pmatrix} 0.32 & 0.13 & 0.54 \end{pmatrix} \\ \mathbf{p}_3 &= \begin{pmatrix} 0.31 & 0.57 & 0.12 \end{pmatrix} \\ \pi &= \begin{pmatrix} 0.27 & 0.44 & 0.29 \end{pmatrix} \end{aligned}$$

Outline

A surfer

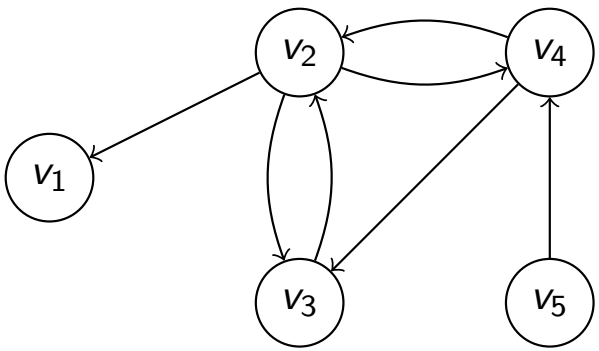
Random surfer (1)

- Consider a random surfer who
  1. Starts at a random web page
  2. Repeatedly clicks a random link to move to next web page
- PageRank is steady-state distribution of the random surfer
  - High PageRank = page frequently visited
  - Low PageRank = page infrequently visited
  - PageRank thus captures the "importance" of each webpage
- When is a page frequently visited?
  - When it has many in-links from frequently visited pages
- Still a circular definition, but now well-defined



Random surfer (2)

- Random surfer as a Markov chain
  - States = web pages
  - Transitions = normalized adjacency matrix (s.t. rows sum to 1)
  - Called **walk matrix** $= \mathbf{D}^{-1}\mathbf{W}$
  - Note $\mathbf{L}_{\text{rw}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$
- Pitfalls
  - How to handle dead ends? (there are many of them on the web)
  - How to avoid getting stuck in subgraphs?



$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

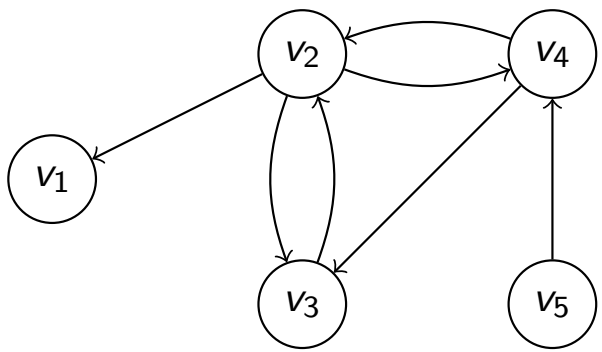
$$\mathbf{D}^{-1}\mathbf{W} = \begin{pmatrix} ? & ? & ? & ? & ? \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

A surfer with a problem

A surfer without a problem

Teleportation

- A teleporting surfer
  - If no outgoing links, go to random site (handles dead ends)
  - With probability α, teleport to a random site (handles subgraphs)
    $\rightarrow$ Can be thought of as typing URL into address bar
  - With probability $1 - \alpha$, follow random link
- Teleportation ensures irreducibility and aperiodicity
- PageRank of page $i = \pi_i$



$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{P}_{0.1} = \begin{pmatrix} 0.20 & 0.20 & 0.20 & 0.20 & 0.20 \\ 0.32 & 0.02 & 0.32 & 0.32 & 0.02 \\ 0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.47 & 0.47 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.92 & 0.02 \end{pmatrix}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 0.15 \\ 0.36 \\ 0.24 \\ 0.20 \\ 0.05 \end{pmatrix}^T$$

Discussion

- PageRank is query-independent
  $\rightarrow$ Static, global ordering
- For web search, PageRank is one component of many
  - E.g., only pages satisfying the query are of interest
- Walks and teleportation can be done *non-uniformly*
  - Topic-specific PageRank
  - Personalized PageRank
  - Do not teleport to "dubious" websites (e.g., link farms)

Outline

Lessons learned

- Link analysis exploits links structure for relevance assessment
  $\rightarrow$ We discussed HITS and PageRank
- Relevance score related to principal eigenvectors
  - HITS: of co-citation and bibliographic coupling matrix
  - PageRank: of walk matrix of a random, teleporting surfer
  - Power method is simple way to compute these eigenvectors

HITS	PageRank
Distinguishes hubs and authorities	Single relevance score
Query dependent	Query independent
Computed online	Computed offline
Mutual reinforcement	Random surfer
No normalization	Out-degree normalization
(Was?) used by `ask.com`	`google.com` etc.

Suggested reading

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze
  Introduction to Information Retrieval (Chapter 21)
  Cambridge University Press, 2008
  `http://nlp.stanford.edu/IR-book/`

- Jon Kleinberg
  Authoritative sources in a hyperlinked environment
  Journal of the ACM, 46(5), pp. 604–632, 1999
  `http://www.cs.cornell.edu/home/kleinber/auth.pdf`

- Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd
  The PageRank Citation Ranking: Bringing Order to the Web.
  Technical Report, Stanford InfoLab, 1999
  `http://ilpubs.stanford.edu:8090/422/`