Web Retrieval

PageRank

Frank Hopfgartner
Institute for Web Science and Technologies

Recapitulation

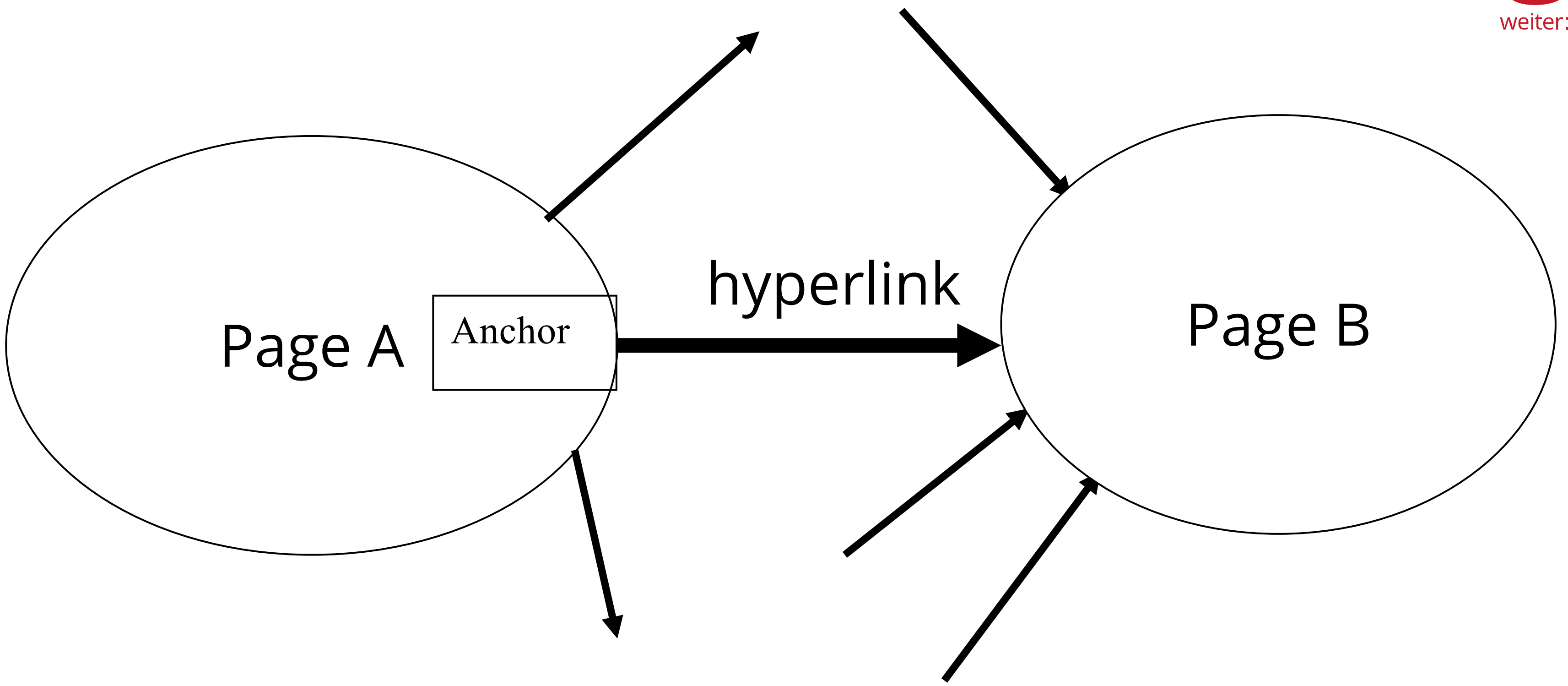
- Crawler
  - what is it?
    - features a crawler *must* provide
    - features a crawler *should* provide
  - crawler architecture
    - robots exclusion protocol
    - url normalization
  - why distributing the crawler
  - the URL frontier

Objectives of this lecture

- PageRank
  - Web graph
  - Origins
  - Motivation
  - Idea of PageRank
  - Recursive formalization
  - Random surfer
  - Formal Model

1. The Web as a graph

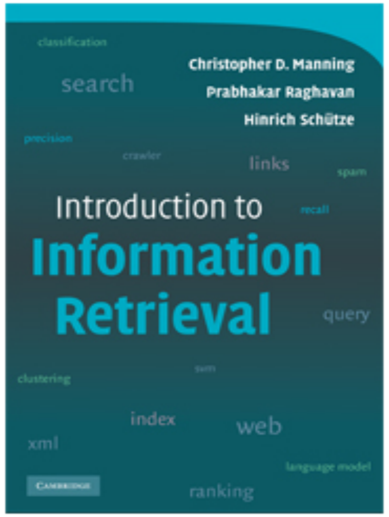
The Web as a directed graph



- **Hypothesis 1:** A hyperlink between pages denotes a conferral of authority (quality signal)
- **Hypothesis 2:** The text in the anchor of the hyperlink on page A describes the target page B

Assumption 1: reputed sites

Introduction to Information Retrieval



This is the companion website for the following book.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Informat*

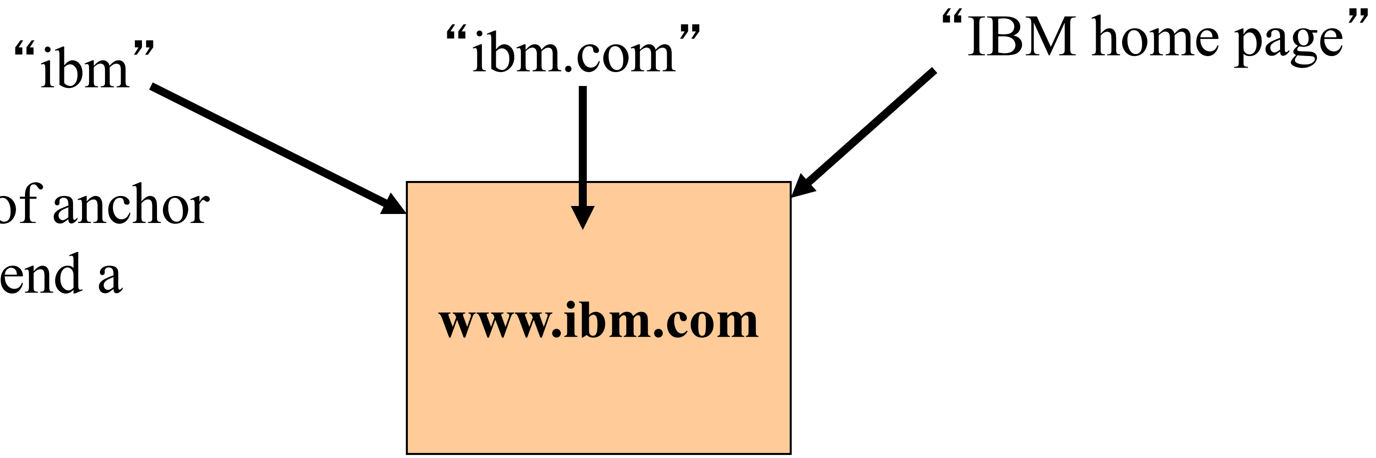
You can order this book at CUP, at your local bookstore or on the internet. The best search

The book aims to provide a modern approach to information retrieval from a computer scie
University and at the University of Stuttgart

We'd be pleased to get feedback about how this book works out as a textbook, what is m
comments to: informationretrieval (at) yahoogroups (dot) com

Anchor text

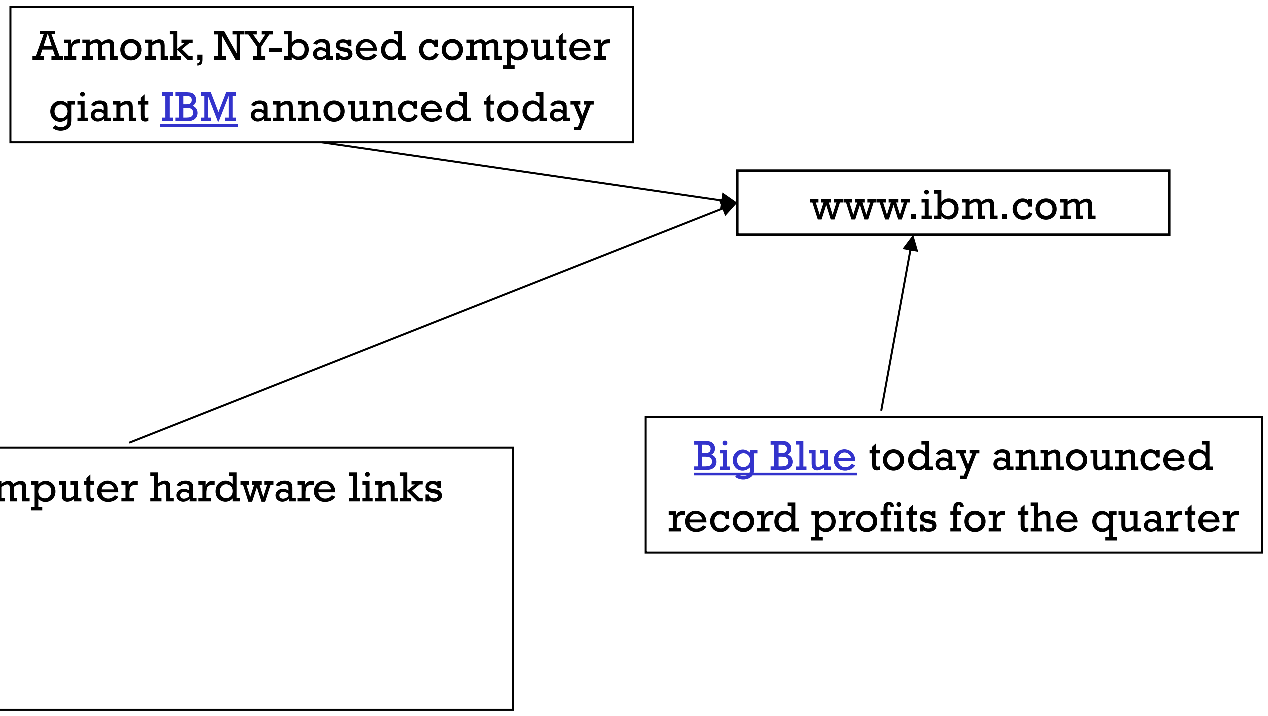
- For ***ibm*** how to distinguish between
  - IBM’s home page (mostly graphical)
  - IBM’s copyright page (high term freq. for ‘ibm’)
  - Rival’s spam page (arbitrarily high term freq.)



A million pieces of anchor text with “ibm” send a strong signal

Indexing anchor text

- When indexing a document *D*, include (with some weight) anchor text from links pointing to *D*



Indexing anchor text

- Thus: anchor text is often a better description of a page's content than the page itself
- Anchor text can be weighted more highly than document text

2. PageRank

Origins of PageRank: citation analysis

- Citation analysis: analysis of citations in the scientific literature
- Example citation: “Miller (2001) has shown that physical activity alters the metabolism of estrogens”
- We can view “Miller (2001)” as a hyperlink linking two scientific articles
- Application of these “hyperlinks” in the scientific literature
  - Measure the similarity of two articles by the overlap of other articles citing them
  - This is called cocitation similarity
  - Cocitation similarity on the web: Google’s “find pages like this” or “Similar” feature

Origins of PageRank: citation analysis

universität koblenz
weiter:denken

- Another application: citation frequency can be used to measure the impact of an article
  - Simplest measure: Each article gets one vote – not very accurate
- On the web: citation frequency = inlink count
  - A high inlink count does not necessarily mean high quality ...
  - ... mainly because of link spam
- Better measure: weighted citation frequency or citation rank
  - An article's vote is weighted according to its citation impact
  - Circular? No: can be formalized in a well-defined way

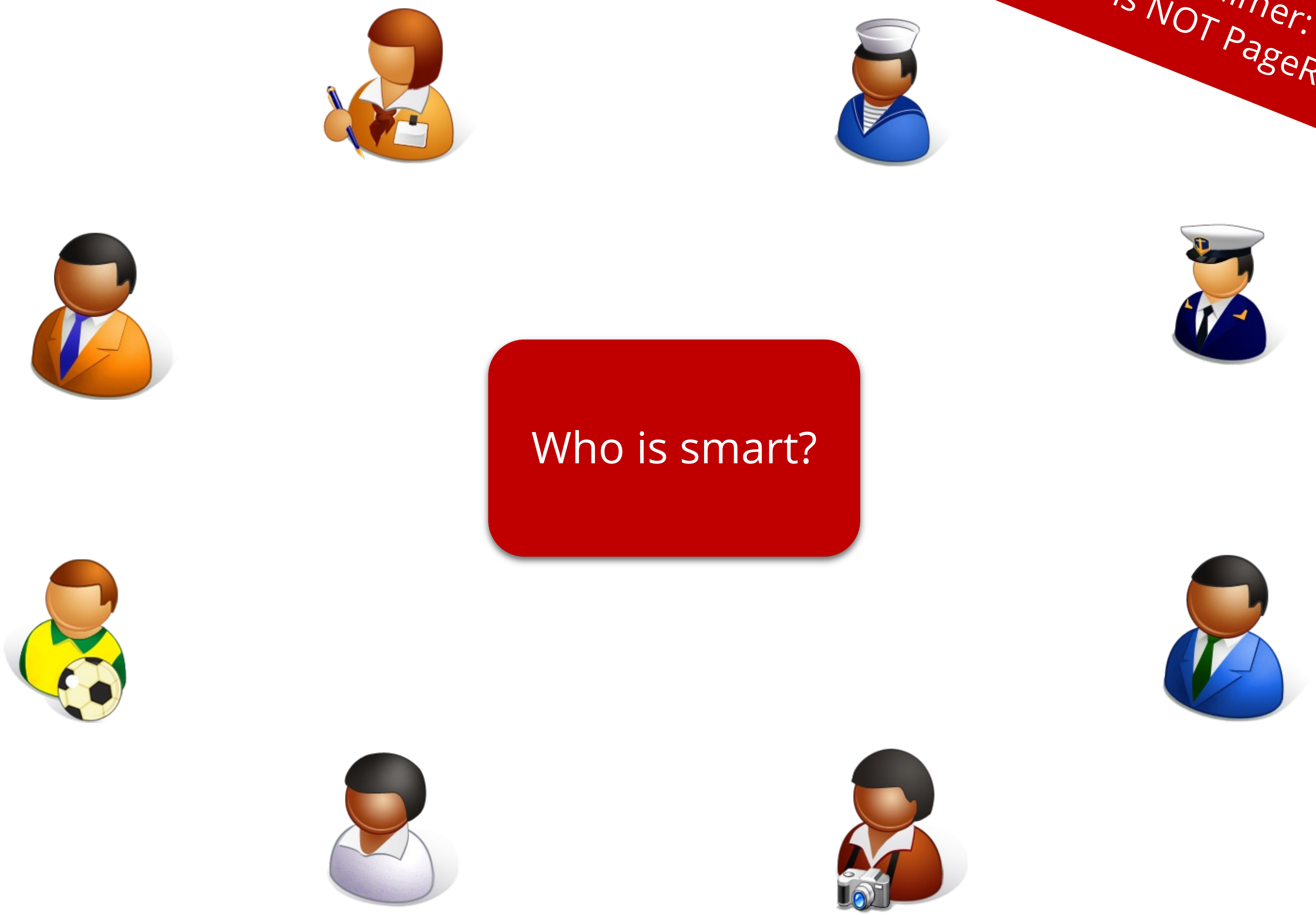
Origins of PageRank: citation analysis

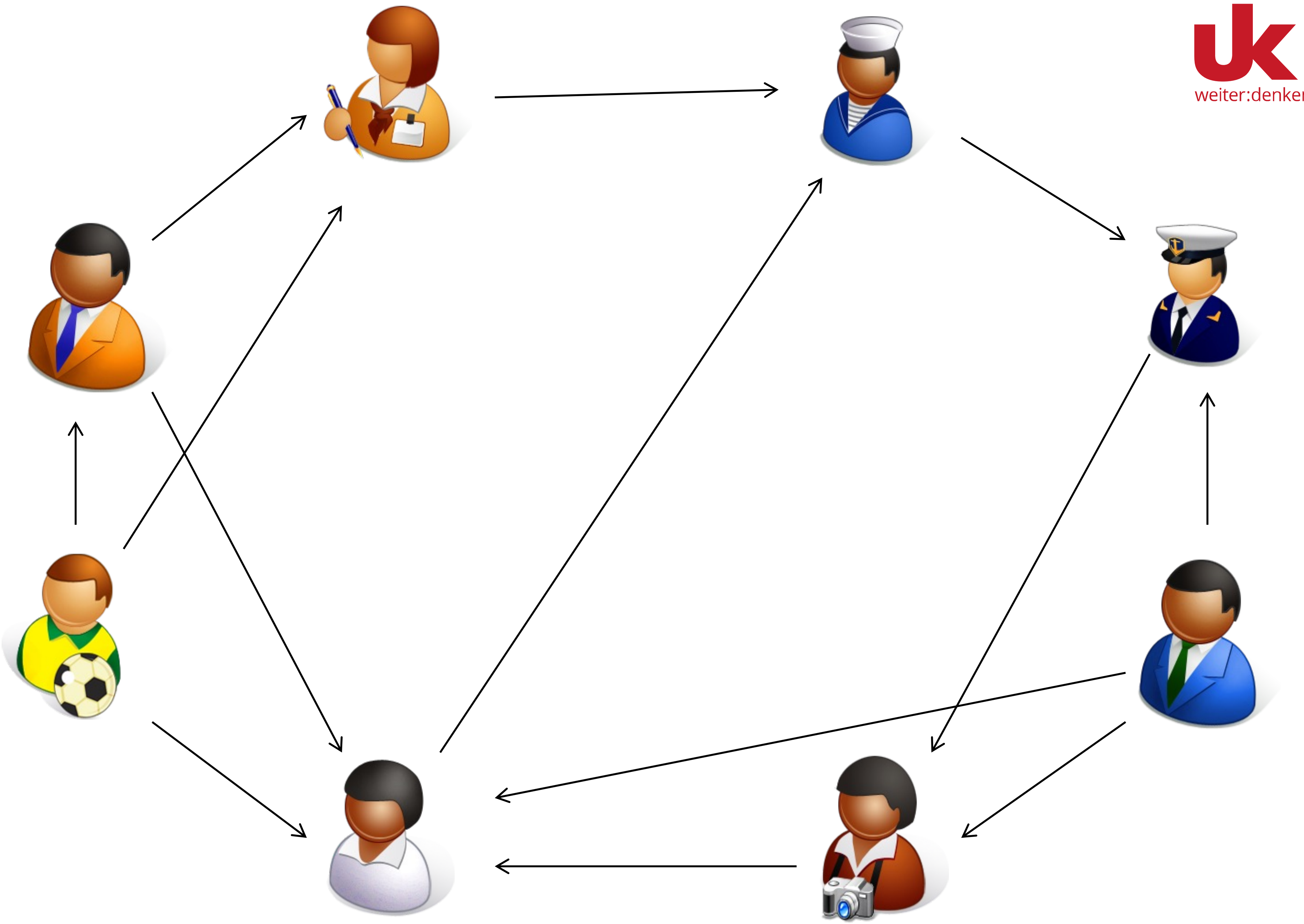
universität koblenz
weiter:denken

- Better measure: weighted citation frequency or citation rank
- This is basically PageRank
- PageRank was invented in the context of citation analysis by Pinsker and Narin in the 1960s
- Citation analysis is a big deal: The budget and salary of this lecturer are / will be determined by the impact of his publications

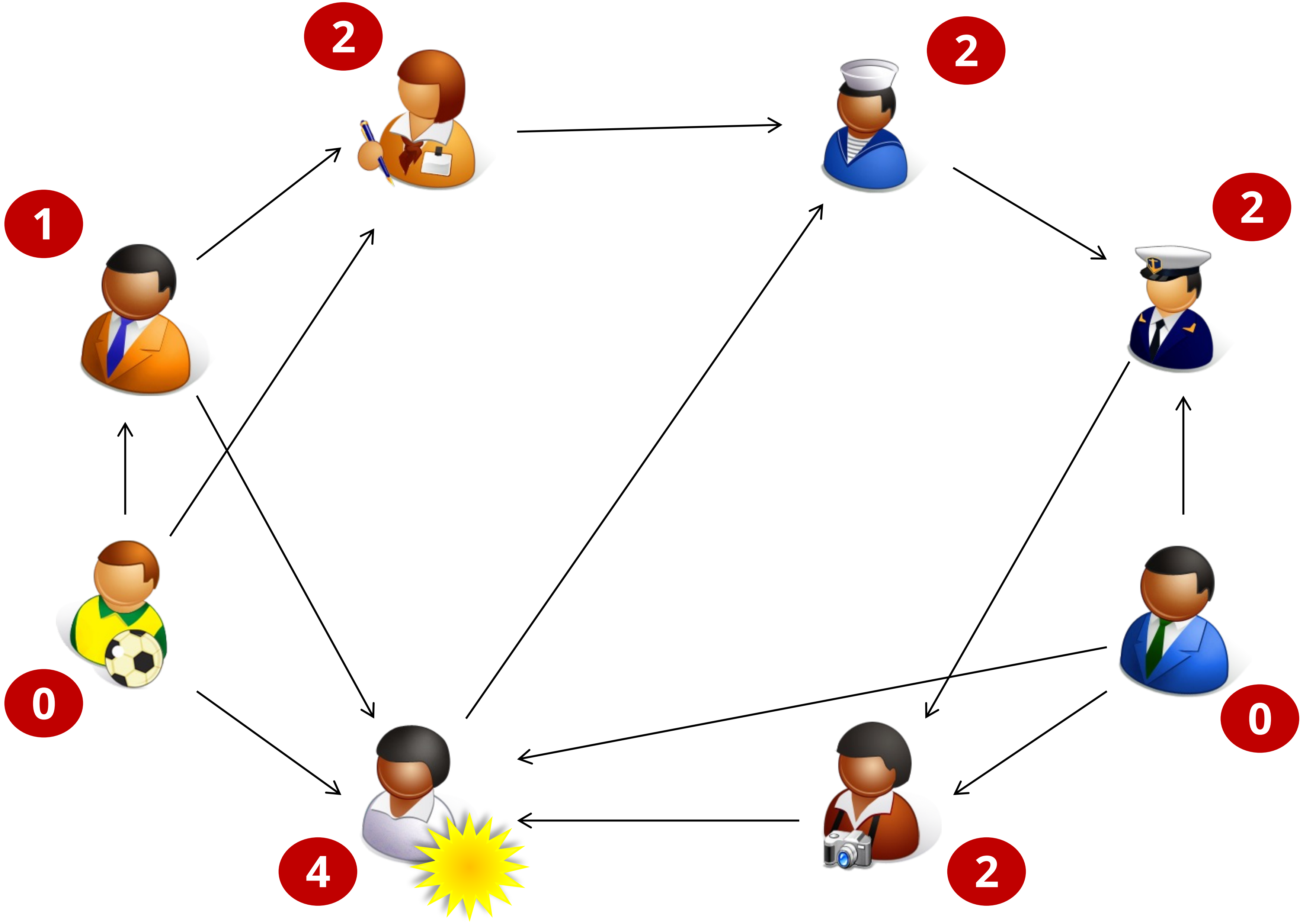
Motivation

Disclaimer:
This is NOT PageRank



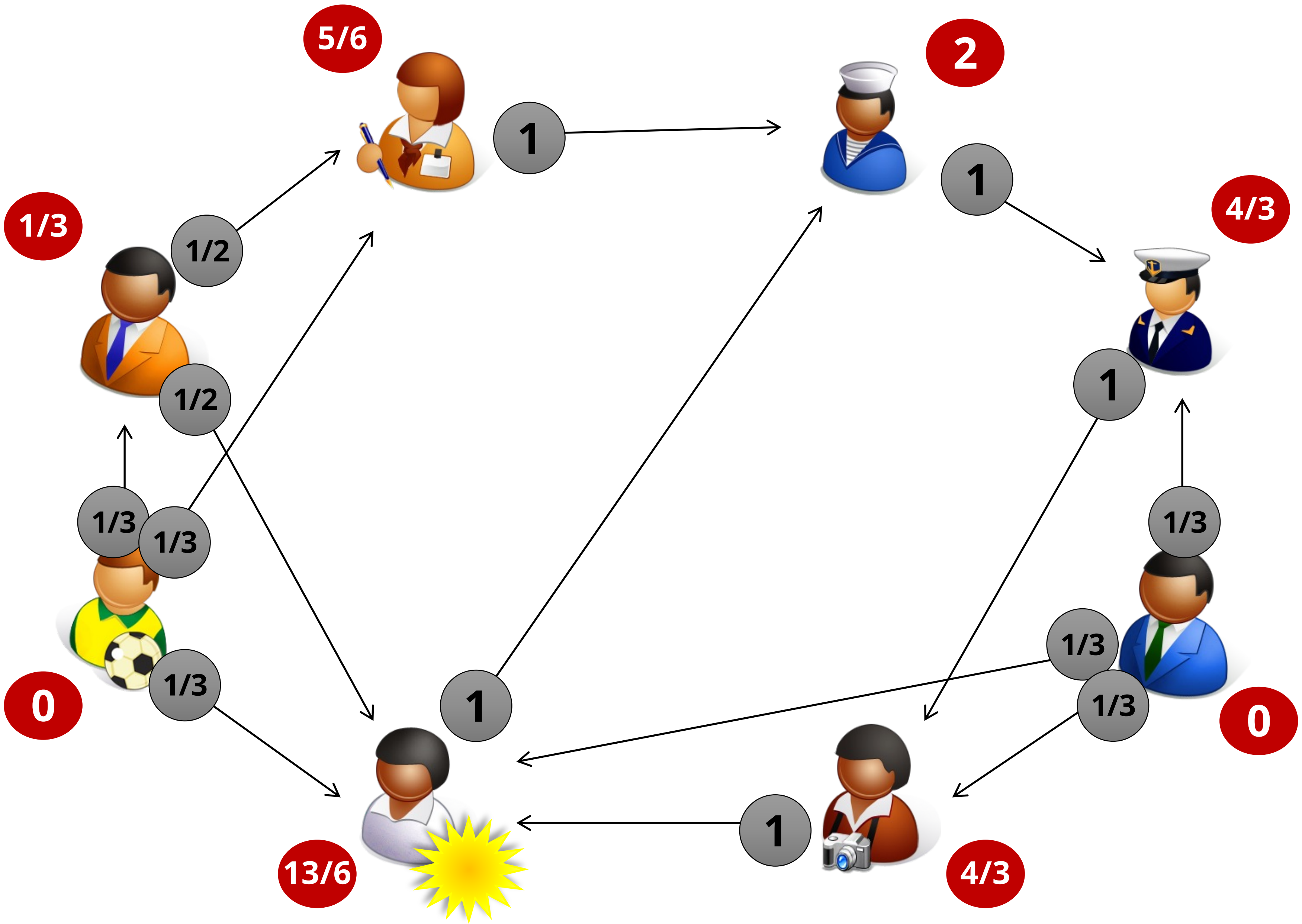


universität koblenz
weiter:denken



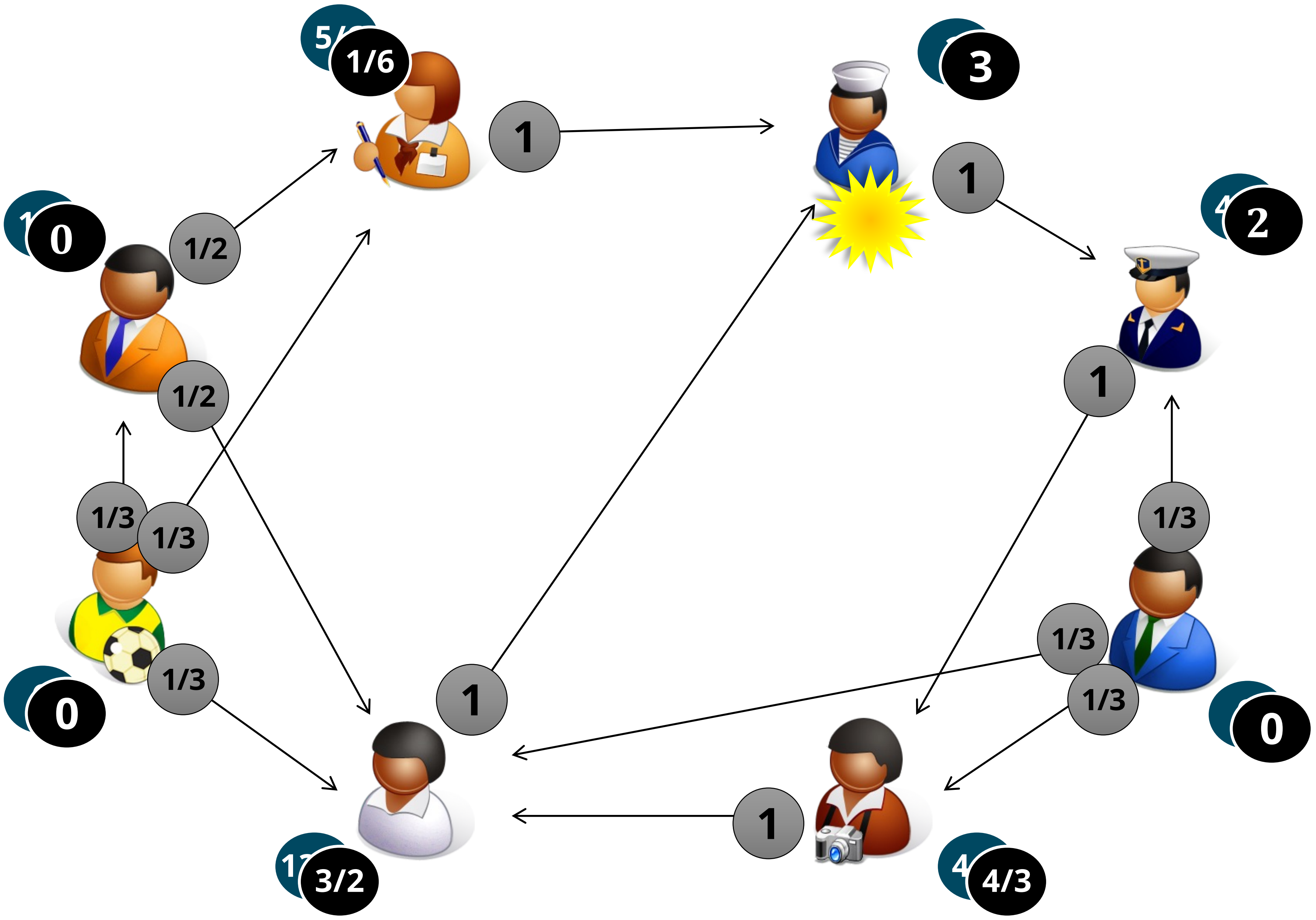
Count votes
(in-degree)

universität koblenz
weiter:denken



Split the votes

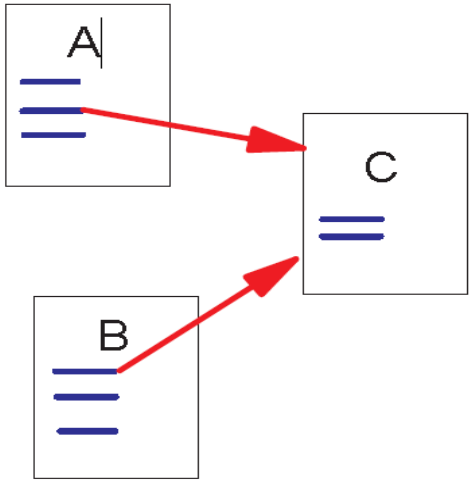
universität koblenz
weiter:denken



Weight the expert votes

Link structure of the Web

- 4.2 billion web pages → 25.2 billion links



Backlinks and Forward links

- ➢A and B are C's backlinks
- ➢C is A and B's forward link

- Intuitively, a webpage is important if it has a lot of backlinks.
- What if a webpage has only one link off www.yahoo.com?

PageRank----idea

- Backlinks coming from important pages convey more importance to a page
    - For example, if a web page has a link from the yahoo home page, it may be just one link but it is a very important one

PageRank: a recursive formalization

universität koblenz
weiter:denken

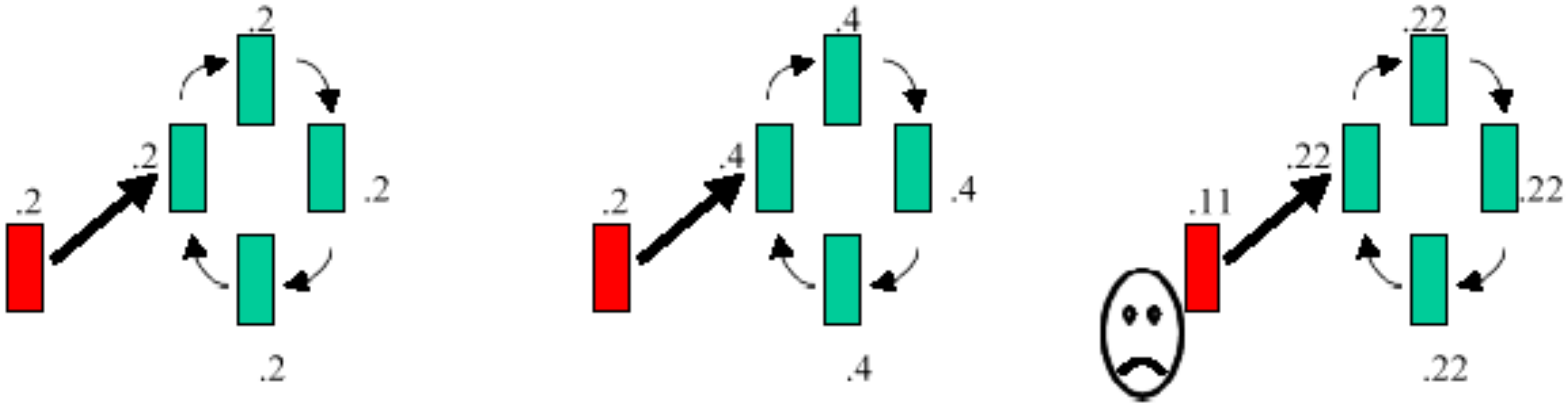
$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

- u: a web page
- B_u: the set of u's backlinks
- N_v: the number of forward links of page v
- c: the normalization factor

The equation is recursive, but it may be computed by starting with any set of ranks and iterating the computation until it converges

PageRank: a recursive formalization

- A problem with such definition: *rank sink*
- If two web pages point to each other but to no other page, during the iteration, this loop will accumulate rank but never distribute any rank



PageRank: a recursive formalization

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u)$$

- $E(u)$ is some vector over the web pages (for example uniform, favorite page, etc.) that corresponds to a source of rank
- $E(u)$ is a user designed parameter

Google PageRank – idea

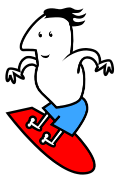
- Intention
  - Identify good sources for information
- Static quality measure
  - Independent of query (who is smart?)
- Idea
  - Good sources are well linked
    - Good information is referenced more often
    - A reference from a good source is worth more
  - → simply counting in-degree is not enough
- How to calculate?
  - Thought experiment: The Random Surfer

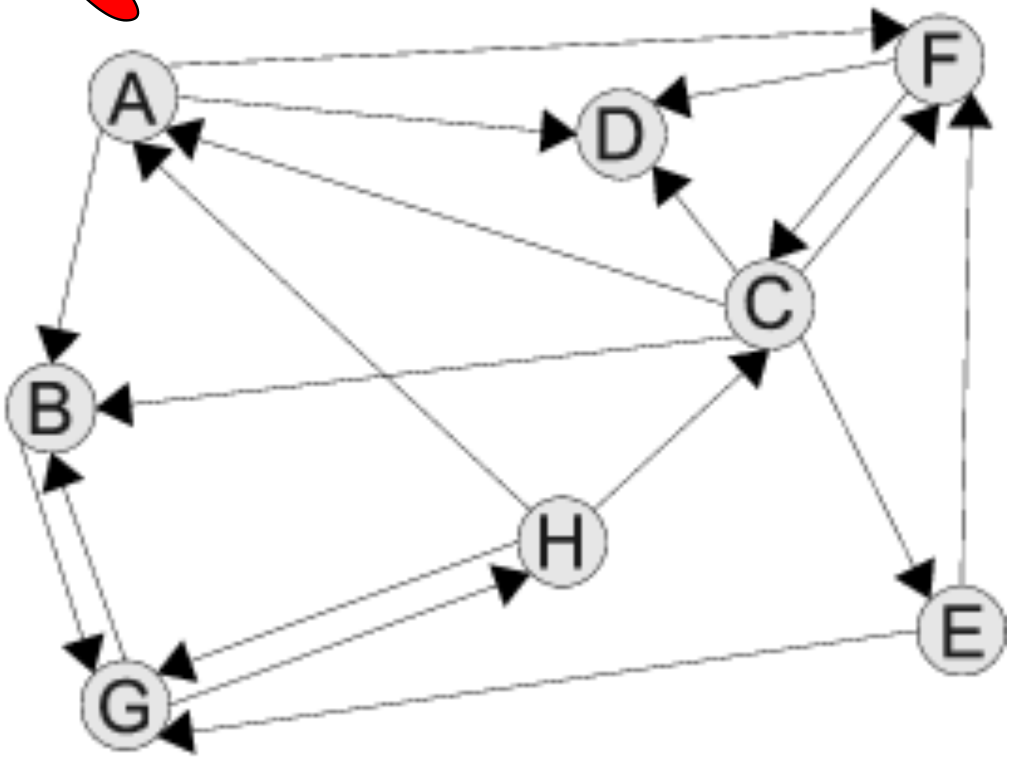
Brin, Page

3. Random surfer

Random Surfer

- User surfing the web
  - Randomly follows links
  - Well linked pages are visited more often
- Count how often documents are visited.
- Example:
  - Random walk on the graph

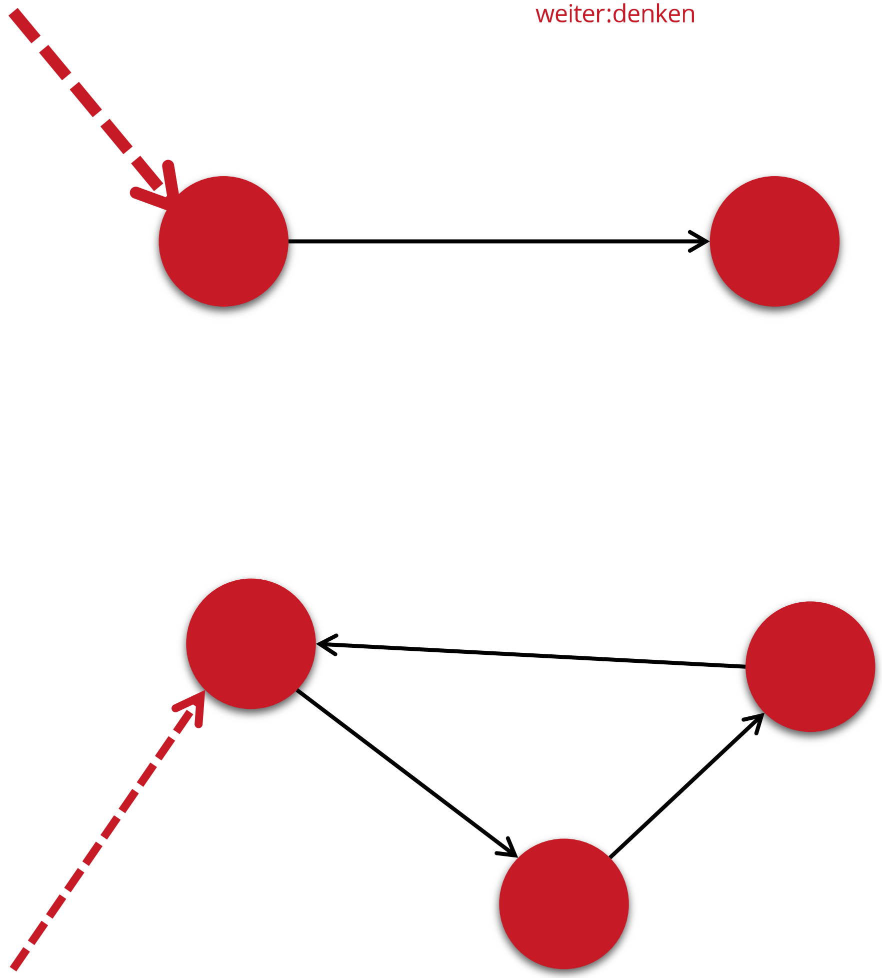




A	B	C	D	E	F	G	H
3	3	1	0	1	1	4	2

Random Surfer

- Problems
  - „Dead ends“
  - Graph not connected
  - Circles
- Solution
  - Teleports
    - Surfer jumps to a random page on the web
  - Use in dead ends
  - Use randomly at all other nodes (with low probability)



Formal model

- Markov Chain
  - States: web pages
  - Transitions: hyperlinks
  - Transition probabilities: uniform distribution
    - Teleports need to be incorporated
- Represented as stochastic matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n1} & \cdots & p_{nn} \end{pmatrix}$$

  - $0 \leq p_{ij} \leq 1$: transition probability from state *i* to state *j*
  - $\sum_j p_{ij} = 1 \;\; \forall i$

Setting the transition probabilities

- For node i
  - If „dead end“ (out-degree of zero)

$$p_{ij} = {}^{1}/_{n}$$

  - Otherwise
    - Link to node j

$$p_{ij} = \frac{\alpha}{n} + (1 - \alpha)\frac{1}{O(i)}$$

    - No link to node j

$$p_{ij} = {}^{\alpha}/_{n}$$

  - α : Probability of teleport
  - $O(i)$: out-degree of node i

Example

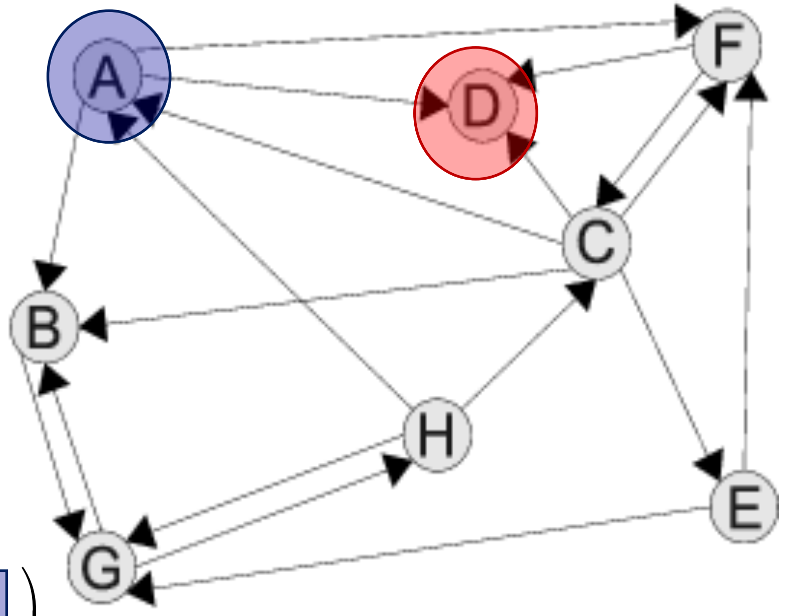
Set $\alpha = 0.1$

$$p_{1,2} = P(A \to B) = \frac{0.1}{8} + 0.9\frac{1}{3} = 0.3125$$

$$p_{1,3} = P(A \to C) = \frac{0.1}{8} = 0.0125$$

$$p_{4,1} = P(D \to A) = \frac{1}{8} = 0.125$$

$$P = \begin{pmatrix} 0.0125 & 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.0125 \\ 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.9125 & 0.0125 \\ 0.1925 & 0.1925 & 0.0125 & 0.1925 & 0.1925 & 0.0125 & 0.1925 & 0.0125 \\ 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.4625 & 0.4625 & 0.0125 \\ 0.0125 & 0.0125 & 0.4625 & 0.4625 & 0.0125 & 0.0125 & 0.0125 & 0.0125 \\ 0.0125 & 0.4625 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.4625 \\ 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.0125 & 0.0125 & 0.3125 & 0.0125 \end{pmatrix}$$



Computing PageRank

- PageRank value
  - Probability of random surfer to be in particular state (node) after infinitely many moves

$$\pi(i) = \lim_{t \to \infty} \frac{v(i,t)}{t}$$

    - $v(i,t)$: number of visits in node i after t steps
  - π as vector of PageRank values
- Algebraic approach
  - Design of matrix P (stochastic)
  - π is left eigenvector for the largest (principal) eigenvalue (1) of P

Example

- Matrix P

$$P = \begin{bmatrix} 0.0125 & 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.0125 \\ 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.9125 & 0.0125 \\ 0.1925 & 0.1925 & 0.0125 & 0.1925 & 0.1925 & 0.0125 & 0.1925 & 0.0125 \\ 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.4625 & 0.4625 & 0.0125 \\ 0.0125 & 0.0125 & 0.4625 & 0.4625 & 0.0125 & 0.0125 & 0.0125 & 0.0125 \\ 0.0125 & 0.4625 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.0125 & 0.4625 \\ 0.3125 & 0.0125 & 0.3125 & 0.0125 & 0.0125 & 0.0125 & 0.3125 & 0.0125 \end{bmatrix}$$

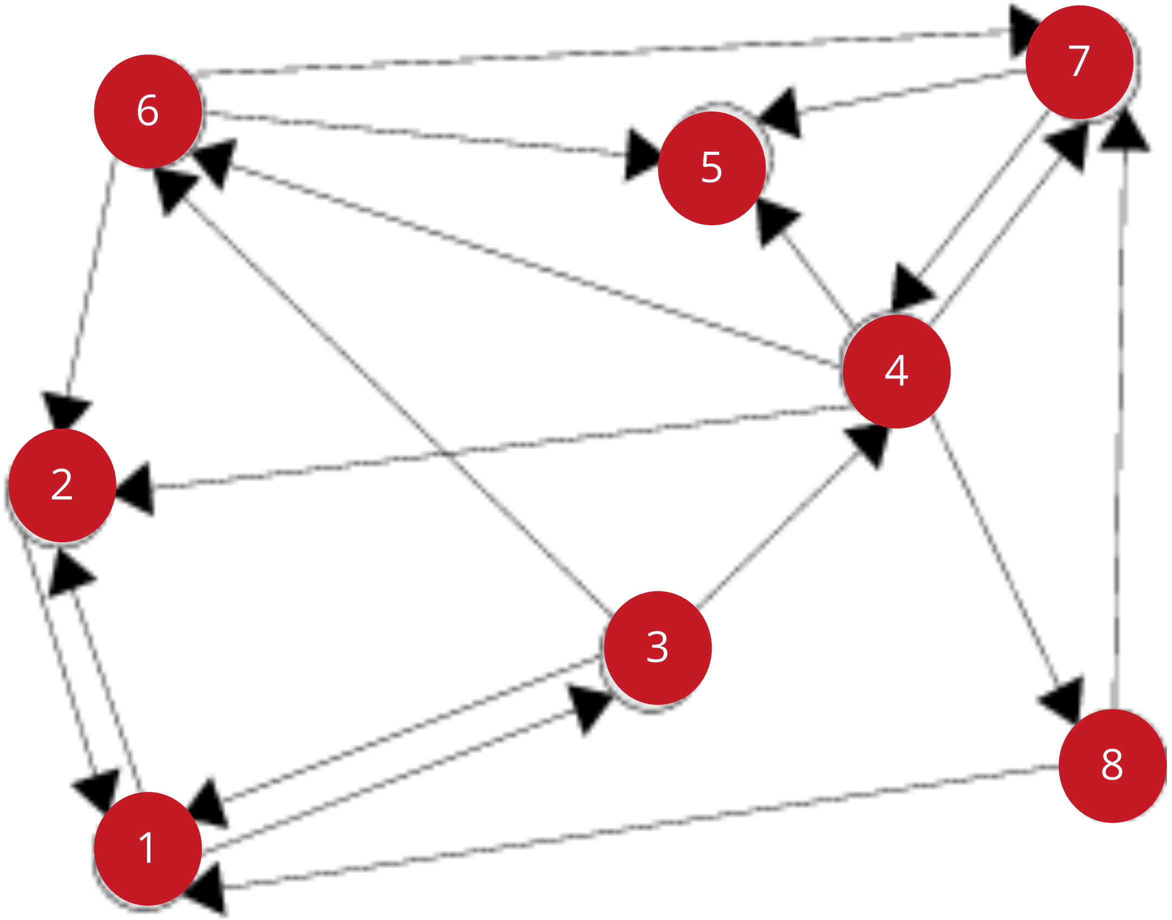
- Using eigen-decomposition of P
- Left principal eigenvector (normalized to represent a distribution)

$$\pi = \begin{pmatrix} 0.0851 & 0.1901 & 0.0978 & 0.0969 & 0.0410 & 0.0674 & 0.2747 & 0.1470 \end{pmatrix}$$

6 2 4 5 8 7 1 3

Example

- Assigning ranks to graph nodes



Power method

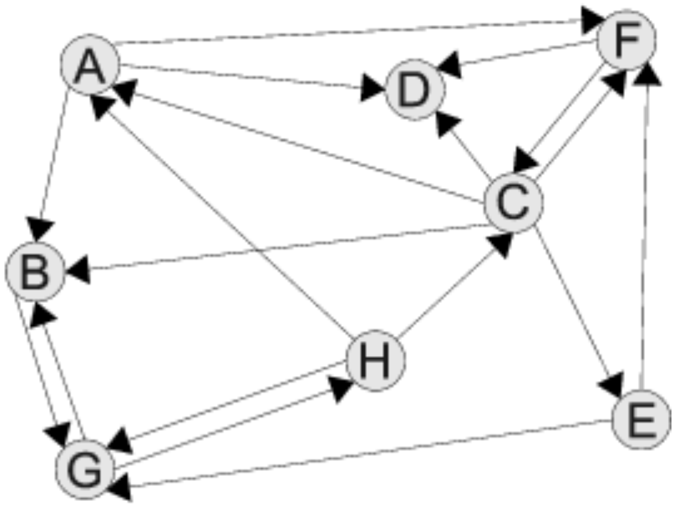
- Computation in practice
  - P is VERY large ($n \times n$, where n is number of nodes)
  - Parallel and distributed execution needed
- Power Method
  - Start vector X_0
  - Iteration

$$X_{k+1} = X_k \cdot P$$

  - Converges against principal eigenvector
- Drawback
  - Slow in convergence (not needed, stable ranking enough)
- Advantage
  - Computation of one entry requires two n-dim vectors
  - Suitable for distributed processing (MapReduce)

Example

- Matrix as before
- Iterations



$$x_0 = \begin{pmatrix} 0.1250 & 0.1250 & 0.1250 & 0.1250 & 0.1250 & 0.1250 & 0.1250 & 0.1250 \end{pmatrix}$$

$$x_1 = \begin{pmatrix} 0.0886 & 0.1428 & 0.1203 & 0.1428 & 0.0491 & 0.1203 & 0.2553 & 0.0828 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} 0.0751 & 0.1911 & 0.1076 & 0.1303 & 0.0502 & 0.0767 & 0.2257 & 0.1435 \end{pmatrix}$$

$$\ldots$$

$$x_{10} = \begin{pmatrix} 0.0845 & 0.1924 & 0.0970 & 0.0972 & 0.0412 & 0.0675 & 0.2714 & 0.1488 \end{pmatrix}$$

6 2 5 4 8 7 1 3

6 2 4 5 8 7 1 3

$$\pi = \begin{pmatrix} 0.0851 & 0.1901 & 0.0978 & 0.0969 & 0.0410 & 0.0674 & 0.2747 & 0.1470 \end{pmatrix}$$

Remarks

- Web graph constantly changing
- PageRank independent of query
  - Compute offline
  - Once per week
- Link Spam
  - Link farming
    - Mark subset of nodes as good/bad
    - See how good/bad PageRank flows through network
- Topic PageRank
  - Teleport only to nodes belonging to topic
- Today used by all large scale web search engines
- Applied in other fields (different network types)

4. Summary

Summary

- PageRank
  - Web graph
  - Origins
  - Motivation
  - Idea of PageRank
  - Recursive formalization
  - Random surfer
  - Formal Model