course:
**Searching the Web** (NDBI038)
**Searching the Web and Multimedia Databases** (BI-VWM)


lecture 4:

# Link analysis and the web page ranking

prof. RNDr. Tomáš Skopal, Ph.D.
Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague
Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague

# Today's lecture outline

- the Web graph
  - link analysis
  - discovering web communities
- web page ranking
  - motivation
  - PageRank

# Link analysis – motivation

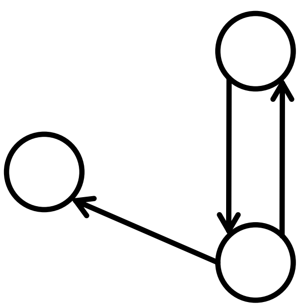
- in 1998, link analysis became popular for Web retrieval
- simple full-text search is not enough
  - all documents themselves are the same important
  - spammers (putting hidden text) betray
    the classic models (especially Boolean model)
- there is information in the web pages' links
  - significant information
    – only few links vs. lots of text
    – a link is very explicit and semantically rich information

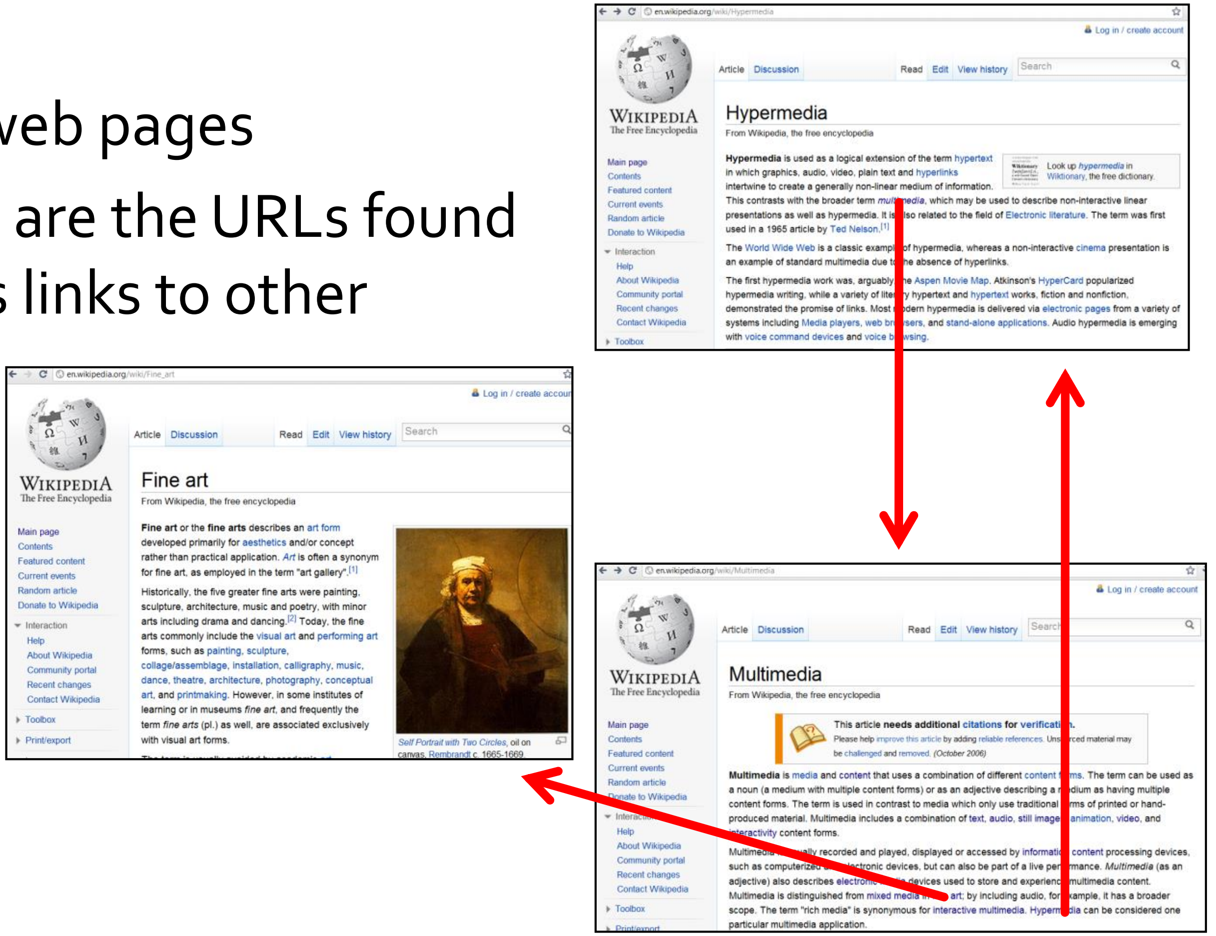
# Link analysis – motivation

- social context
  - the author of web page is also important
  - popularity of web page (many links from other pages)
  - recommendation of web page (who is linking?)
- link analysis could provide
  - direct discovery of web communities
  - augmentation of search engines
    - favoring trustworthy pages over garbage (spam) pages, even with the same full-text content
  - visualization and segmentation of the Web space

# Link analysis – the Web graph

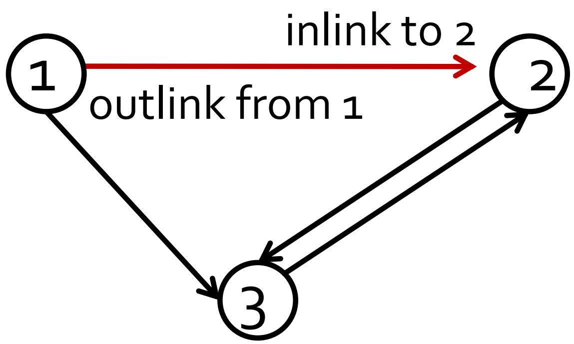
- Web graph
  - **nodes** are the web pages
  - **directed edges** are the URLs found in web pages as links to other web pages





# Link analysis – basic fragments

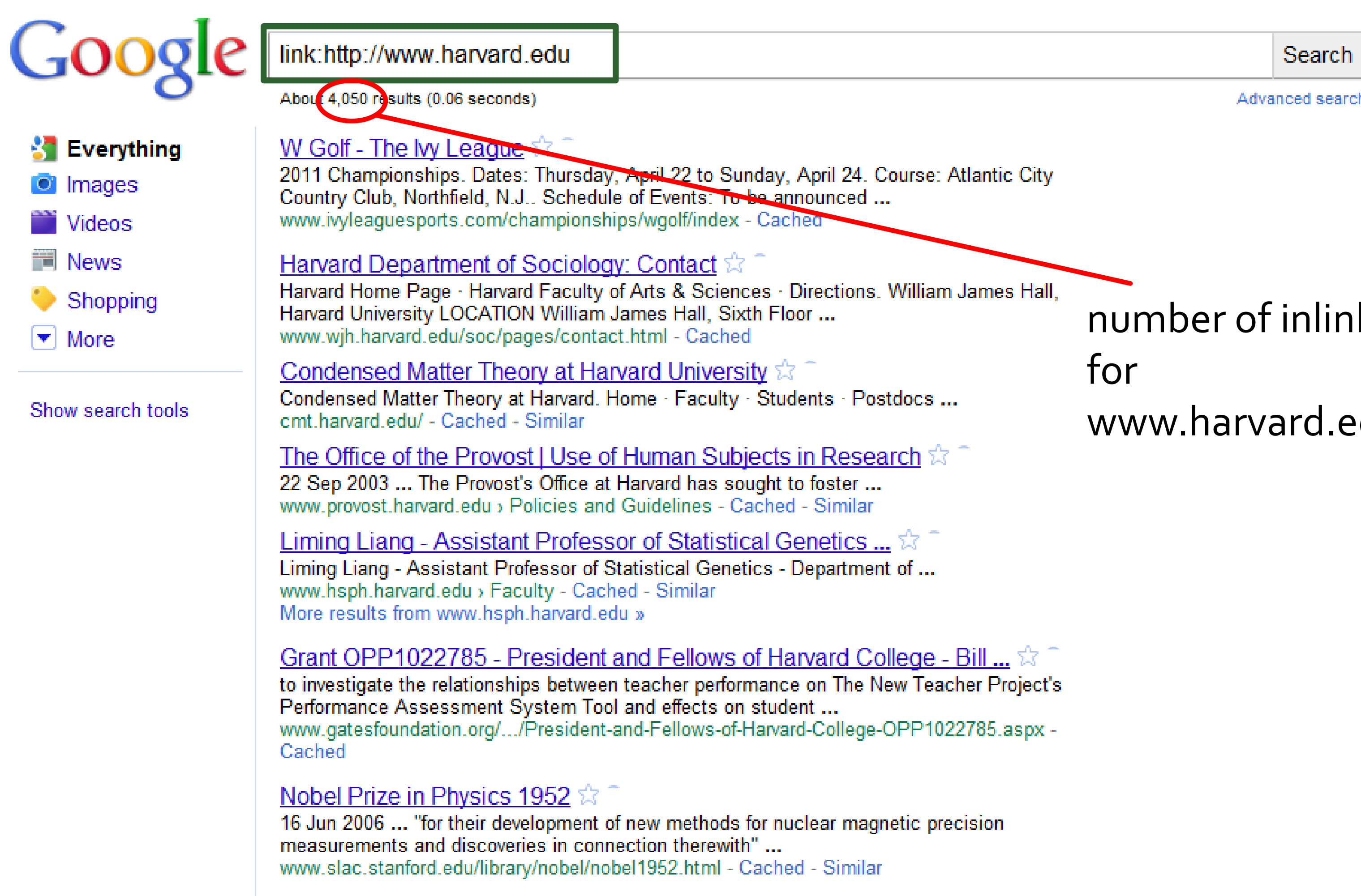
- directed edge = link
  - **inlink**, link from the perspective of **linked** page
  - **outlink**, link from the perspective of **linking** page



# Link analysis – basic fragments

- outlinks – easy to count
  - included in the web page HTML code
  - e.g., <a href="http://www.goodpage.com"> Good page</a>
- inlinks – not that easy to count
  - not included in the **linked** page,
    but in the (possible) **linking** pages
  - services provided by search engines
    - e.g., google – search for **link:<URL of a page>**

# Link analysis – basic fragments

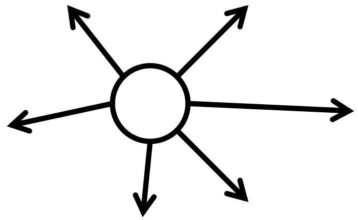



number of inlinks for www.harvard.edu

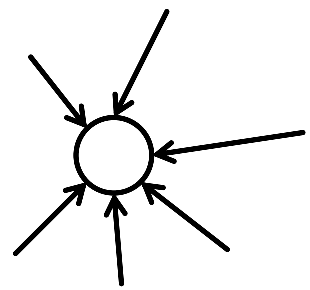
# Link analysis – basic fragments

- based on inlinks and outlinks,
  various subgraph patterns defined (named)

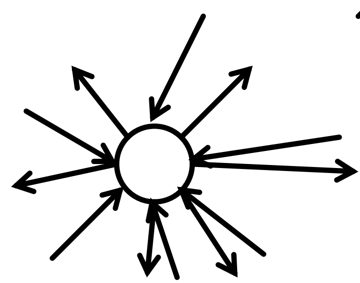
- hub
  - page with many outlinks



- authority
  - page with many inlinks

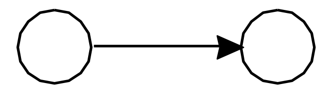


- a page could be both
  hub and authority

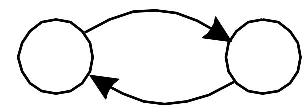


# Link analysis – basic fragments

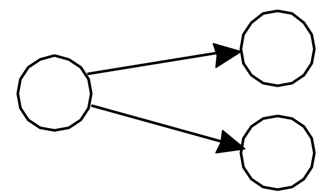
- page endorsement
  - web page refers to another page



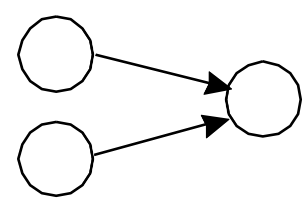
- relevant pages
  - pages refer to each other



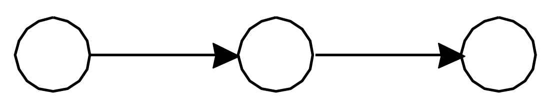
- co-citation
  - page refers to several pages



- social choice
  - a page is referred by several pages

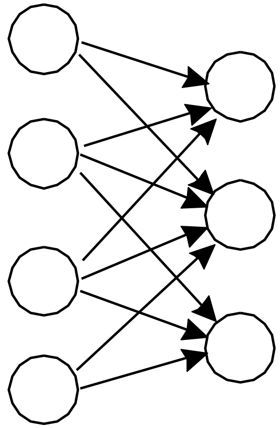


- transitive endorsement
  - $p_1$ refers to $p_3$ through $p_2$

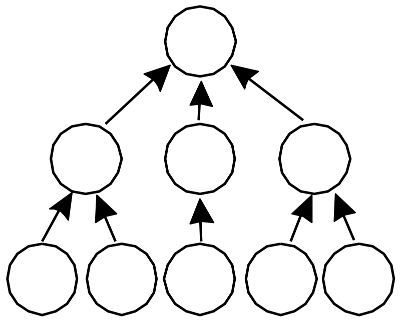


# Web communities

More complex shapes represent web communities (clusters) of pages (or multimedia documents) within the Web graph.



- **bipartite graph**
  two sets of pages, each page of the first set refers to the second one
  meaning – web community sharing interests defined by the second part of the graph



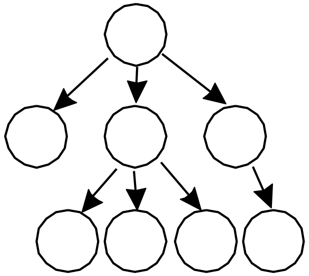
- **in-tree**
  generalization of "social choice"
  meaning – nodes of its upper levels serve as authoritative information sources due to the high number of (transitive) endorsements

# Web communities

- **out-tree**

  generalization of "co-citation"

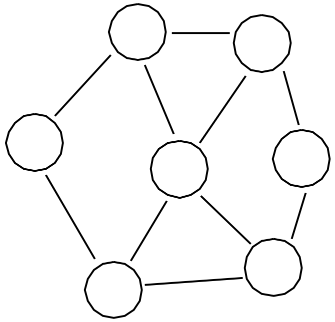
  meaning – its nodes (pages) serve as
  hubs to relevant pages of some monothematic content



- **2-connected component**

  a component of graph which remains
  connected after removal of an arbitrary node

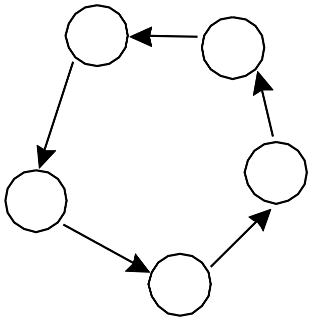
  meaning – represents tightly interconnected
  "peer-to-peer" web community



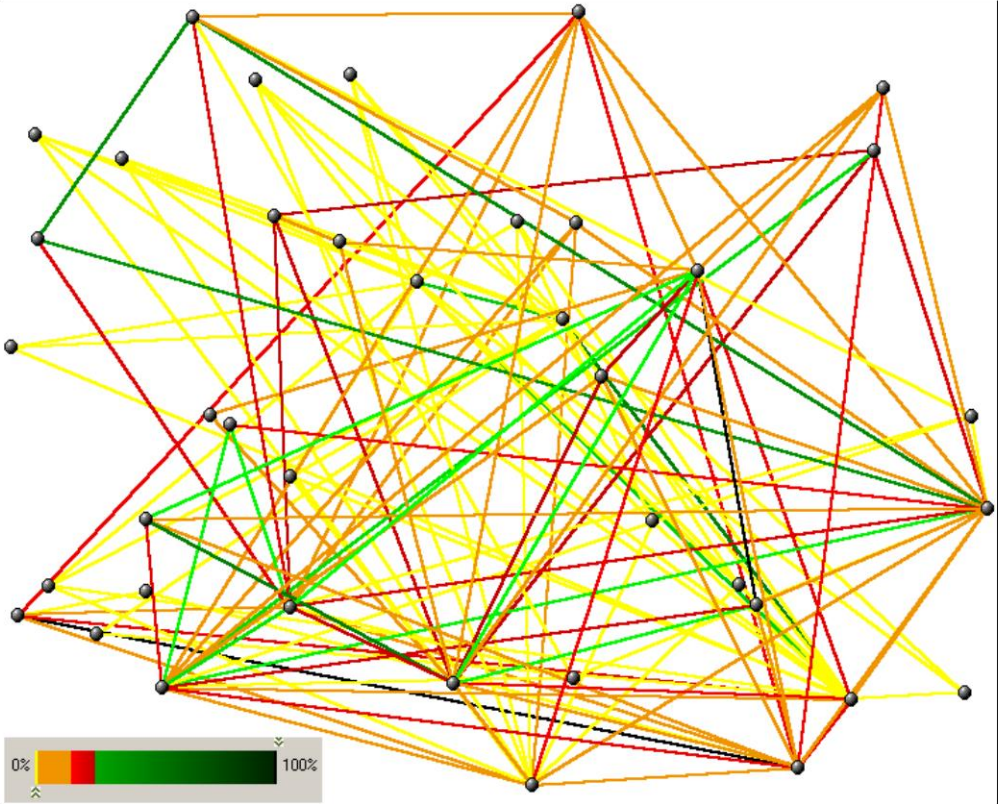
- **cycle**

  a chain of page endorsements where
  the last page in the chain endorses the first page.

  meaning – "weaker" web community,
  forming a web ring



# Web communities – example




- real web subgraph
- 2-connected component with 134 edges
- color scale – used similarity measure (cosine) between web pages
  - criterion used for full-text classification of web pages relationship
  - additional confirmation of community relevancy

# Link analysis – ranking web pages

- ranking of web pages = measuring “popularity”
  - based on inlink statistics
  - inspiration in bibliometry (citations of scientific articles)
- page rank = non-negative real number
  - computed from just the structural information, i.e., query-independent, fulltext content independent
- two ranking algorithms introduced in 1995-1998
  - **PageRank** by Sergey Brin and Larry Page, later evolved into the giant Google (US patent granted 2001)
  - **HITS** by Jon Kleinberg, an extension of which is used in Teoma search engine

# Link analysis – ranking web pages

- application of page ranking in search engines
  - page rank is combined with query-dependent content rank
  - at query time – three steps
    1) ranking by content (e.g., vector space query and cosine similarity)
    2) get page ranks for the pages returned as a query result
    3) re-ranking of the result by aggregations with page rank scores

**1) ranking by content**

| | P3 | P2 | P4 | P6 | P1 | P5 |
|---|---|---|---|---|---|---|
| content scores: | 0.92 | 0.86 | 0.81 | 0.73 | 0.55 | 0.32 |

**2) get page ranks**

| | P5 | P6 | P1 | P3 | P4 | P2 |
|---|---|---|---|---|---|---|
| page rank scores: | 0.62 | 0.55 | 0.48 | 0.42 | 0.37 | 0.35 |

**3) aggregation and re-ranking**

| | P6 | P3 | P2 | P4 | P1 | P5 |
|---|---|---|---|---|---|---|
| final scores: | 0.401 | 0.386 | 0.301 | 0.299 | 0.264 | 0.198 |

# Link analysis – ranking web pages

- PageRank's thesis

  *"A web page is important if it is pointed to by other important pages."*

  - a bit circular definition, but could be well formalized

- HITS' thesis

  *"A page is a good hub if it points to good authorities, and a page is a good authority if it is pointed to by good hubs."*

  - i.e., two ranks (hub rank, authority rank)
  - also circular definition, also could be well formalized

# PageRank – original formula

- original summation formula
  where $r(P_i)$ is the PageRank score of page $P_i$, $|P_j|$ is the number of $P_j$'s outlinks, $B_{P_i}$ is the set of pages linking/pointing to $P_i$

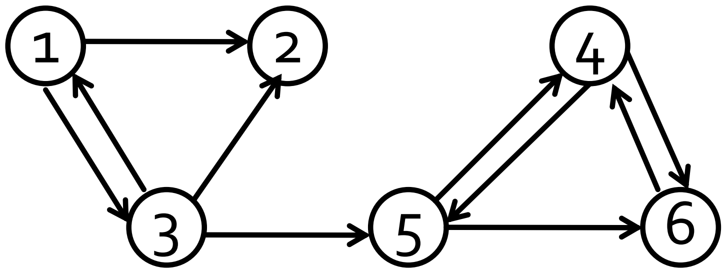
$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

- obvious problem: *$r(P_j)$ values unknown*
  - ok, let's make it an **iterative process**
  - initialize all PageRanks to 1/*n* (*n* = number of all web pages)
  - applying the above equation in multiple iterations
  - the updated formula for a (k+1)$^{th}$ iteration could be rewritten as:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

# PageRank – original formula

- example,
  6 web pages
  $P_1 – P_6$




$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

| Iteration 0 | Iteration 1 | Iteration 2 | Rank after Iteration 2 |
|---|---|---|---|
| $r_0(P_1) = 1/6$ | $r_1(P_1) = 1/18$ | $r_2(P_1) = 1/36$ | 5 |
| $r_0(P_2) = 1/6$ | $r_1(P_2) = 5/36$ | $r_2(P_2) = 1/18$ | 4 |
| $r_0(P_3) = 1/6$ | $r_1(P_3) = 1/12$ | $r_2(P_3) = 1/36$ | 5 |
| $r_0(P_4) = 1/6$ | $r_1(P_4) = 1/4$ | $r_2(P_4) = 17/72$ | 1 |
| $r_0(P_5) = 1/6$ | $r_1(P_5) = 5/36$ | $r_2(P_5) = 11/72$ | 3 |
| $r_0(P_6) = 1/6$ | $r_1(P_6) = 1/6$ | $r_2(P_6) = 14/72$ | 2 |

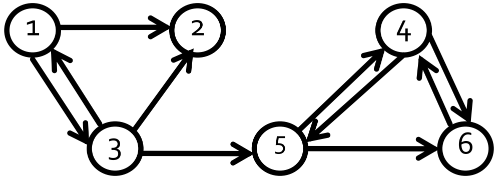
# PageRank – original formula

- matrix representation better, instead of $r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$

  we get $$\pi^{(k+1)T} = \pi^{(k)T} H$$

  where $\pi^{(k)T}$ is the PageRank vector at iteration *k* (contains ranks for all pages),
  **H** is a link matrix; $\mathbf{H}_{ij} = 1/|P_i|$ if there is a link from $P_i$ to $P_j$, otherwise $\mathbf{H}_{ij} = 0$

- for the previous example



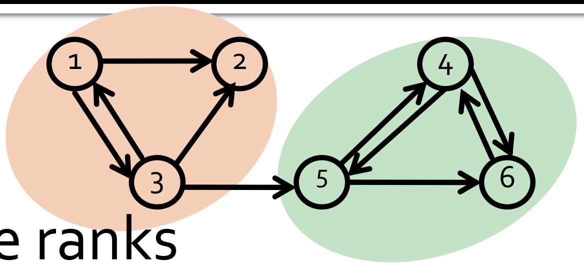

$$H = \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

<u>Note 1:</u> **H** is generally the *adjacency matrix* of the Web graph (but not just binary)
<u>Note 2:</u> **H** is very sparse.
In practice, up to 10 nonzeros per row.

# PageRank – original formula

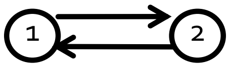
- the original formula gives to rise natural questions
  - Will this iterative process run indefinitely or will it converge?
  - For which **H** is it guaranteed to converge?
  - Will it converge to ranks that follow the PageRank thesis?
  - Will it converge to one PageRank vector, or to multiple?
  - Does the convergence depend on the starting vector $\pi^{(0)T}$ ?
  - If converging, how long it takes?
- the answer is: The original version is **not sufficient**.

# PageRank – original formula




- some of the problems
  - **rank sinks** – some pages lose the ranks during iterations, while some others accumulate them all
    - for the previous example, consider $13^{th}$ iteration

| Iteration 0 | Iteration 1 | Iteration 2 | Iteration 13 | Rank after Iter. 2 | Rank after Iter. 13 |
|---|---|---|---|---|---|
| $r_0(P_1) = 1/6$ | $r_1(P_1) = 1/18$ | $r_2(P_1) = 1/36$ | $\mathbf{r_{13}(P_1) = 0}$ | 5 | **4** |
| $r_0(P_2) = 1/6$ | $r_1(P_2) = 5/36$ | $r_2(P_2) = 1/18$ | $\mathbf{r_{13}(P_2) = 0}$ | 4 | **4** |
| $r_0(P_3) = 1/6$ | $r_1(P_3) = 1/12$ | $r_2(P_3) = 1/36$ | $\mathbf{r_{13}(P_3) = 0}$ | 5 | **4** |
| $r_0(P_4) = 1/6$ | $r_1(P_4) = 1/4$ | $r_2(P_4) = 17/72$ | $\mathbf{r_{13}(P_4) = 2/3}$ | 1 | **1** |
| $r_0(P_5) = 1/6$ | $r_1(P_5) = 5/36$ | $r_2(P_5) = 11/72$ | $\mathbf{r_{13}(P_5) = 1/3}$ | 3 | **2** |
| $r_0(P_6) = 1/6$ | $r_1(P_6) = 1/6$ | $r_2(P_6) = 14/72$ | $\mathbf{r_{13}(P_6) = 1/5}$ | 2 | **3** |

  - **cycles** – the iterative process may not stop (rank flipping)
    - consider 

 and the initialization $\pi^{(0)T}=(1\ 0)$, which leads to $\pi^{(1)T}=(0\ 1)$, $\pi^{(2)T}=(1\ 0)$, and so on…

# PageRank – advanced formula

- the original equation $\pi^{(k+1)T} = \pi^{(k)T}H$ resembles the **power method** applied to a **Markov chain** with **transition probability matrix H**
  - don't worry, we are not going to stuck in too much math ☺
  - mentioned because Markov theory is well-studied (over one hundred years)
- the important is that if **H** is **stochastic**, **irreducible** and **aperiodic** and if using the iteration process as described by the equation, then
  - there exist unique (just one) PageRank vector $\pi$
  - does not matter whatever the initialization $\pi^{(0)T}$ is
  - the iteration process converges
  - the problems with cycles and rank sinks are removed
- gives positive answers to the questions from the previous slides

# PageRank – advanced formula

- to obtain the desired matrix, **H** must be modified into the resulting matrix that is stochastic, irreducible and aperiodic

- a matrix is **stochastic** if the rows sum to 1
  - in the matrix **H**, this is true for pages with **at least one outlink** (remember, then $\mathbf{H}_{ij} = 1/|P_i|$ )
  - but it is not true for so-called **dangling nodes** (pages without outlinks) e.g., pdfs, images, or web pages containing just text
    - could be fixed by replacing the zero rows in **H** by $1/n\mathbf{e}^T$ (note that $n$ is the number of all web pages and $\mathbf{e}^T$ is $n$-dimensional vector of 1s)

# PageRank – advanced formula

- hence, we obtain a stochastic matrix

$$S = H + a\left(\frac{1}{n}e^T\right)$$

where **a** is a **binary dangling vector**, such that
$a_i = 1$ if $P_i$ has no outlinks (dangling page) or $a_i = 0$ otherwise

- considering the previous example,

$$H = \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$S = \begin{array}{c} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

# PageRank – advanced formula

- finally, **S** must be made **irreducible** and **aperiodic**
  - this could be done by making the matrix **primitive** (**A** is primitive if for some $k$ $\forall_{i,j}$ $\mathbf{A}^k_{ij} > 0$)
  - primitive matrix implies its irreducibility and aperiodicity

- Brin and Page defined the **Google matrix** as

  $$G = \alpha S + (1-\alpha)\frac{1}{n}ee^T$$ or simply $$G = \alpha S + (1-\alpha)E$$

  where $E = \frac{1}{n}ee^T$

  where $\alpha \in (0,1)$ is a parameter
  - the second component in the formula ensures the matrix **G** is primitive – completely dense and positive (**no zero inside**)

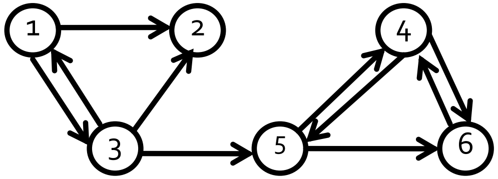
# PageRank – advanced formula

- is the adjustment from **H** to **G** natural?
  - it is, if we extend the model of links between pages
  - consider a random surfer, that visits pages based on outlinks from the current page (the original intuition on **H**), moreover
    - a dangling page gets outlinks to all pages (intuition on **S**)
    - from time to time, following just the outlinks is "boring", so the surfer "teleports" randomly anywhere (intuition on **G**)
      - the $\alpha$ parameter controls the proportion of "following" and "teleporting" of the surfer
- since **G** is the desired matrix, the famous PageRank formula is: $\pi^{(k+1)T} = \pi^{(k)T} G$ (short version)

$$\pi^{(k+1)T} = \pi^{(k)T} (\alpha S + (1-\alpha) E)$$ (expanded version)

# PageRank – advanced formula

- summarizing the advanced formula
  - **G** very dense, unlike **H** or **S**, which is bad for computation
  - the iterative process converges (50-100 iterations enough)
  - there is just one PageRank vector $\pi$, regardless of initialization
  - the PageRank vector is positive, so no ties caused by zeros
- the $\alpha$ parameter
  - observed that $\alpha = 0.85$ works the best (used by Google)
- again, consider our example




| | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ |
|---|---|---|---|---|---|---|
| $\pi^T$ = | (0.03721, | 0.05396, | 0.04151, | 0.3751, | 0.206, | 0.2862) |
| ranks | 6th | 4th | 5th | 1st | 3rd | 2nd |

# PageRank – the computation

- since the matrices are huge, the formulas "materializing" **G** or **S** below cannot be used directly

$$\pi^{(k+1)T} = \pi^{(k)T} G$$

$$\pi^{(k+1)T} = \pi^{(k)T} (\alpha S + (1-\alpha) E)$$

- generally, direct methods cannot be applied
  - due to storing intermediate matrices which is impossible at Google scale (in July 2008, the Google matrix size was expected $10^{12} \times 10^{12}$)
- only matrix-free methods are feasible
  - vector-sparse matrix multiplications, storing just the $\pi^{(k)T}$ and **a** vectors, and the very sparse matrix **H** (compact storage schema, e.g., 10 O(*n*))
  - note that even the $\pi^{(k)T}$ alone is huge amount of data, a few terabytes

# PageRank – the computation

- the power method is matrix-free

$$\pi^{(k+1)T} = \pi^{(k)T}G$$
$$= \alpha\pi^{(k)T}S + \frac{1-\alpha}{n}\pi^{(k)T}ee^T$$
$$= \alpha\pi^{(k)T}H + (\alpha\,\pi^{(k)T}a + 1 - \alpha)e^T/n$$

  - note that vector-matrix multiplication $\pi^{(k)T}\mathbf{H}$ is O($n$) because **H** is extremely sparse and has about 10 nonzeros per row (10 outlinks per page on average)
  - also, only 50 iterations is sufficient, so the whole PageRank vector computation takes just 50 O($n$) time!
    - -$t$/$\log_{10}\alpha$ iterations is needed to get PageRank's accuracy to $t$ digits
    - in practice, for $\alpha$ = 0.85 and 50 iterations we get 2-3 digits accuracy, which is enough if the PageRank is combined with content scores at query time (e.g., the cosine similarity scores for vector model)

# PageRank – other topics

- analyzing the PageRank parameters
  - the $\alpha$ factor, the link matrix **H**, the teleportation matrix **E**
- PageRank sensitivity
- issues in large-scale implementation of PageRank
- accelerating the computation of PageRank
- updating the PageRank vector

further reading:

A.N. Langville, C.D. Meyer, **Google's PageRank and Beyond**,
Princeton University Press, 2006

# Check your PageRank!

- at www.prchecker.info