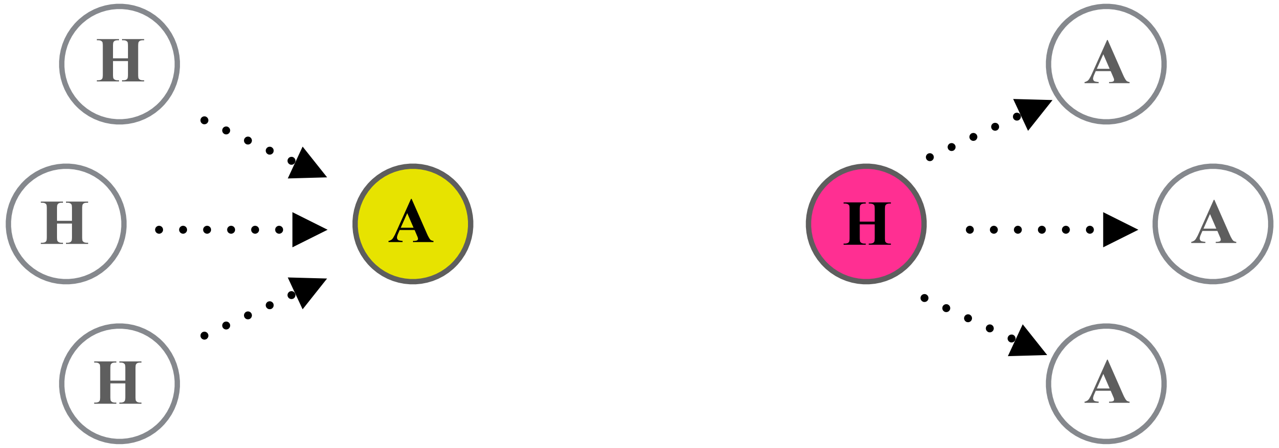
IV.3 HITS

- **Hyperlinked-Induced Topic Search** (HITS) identifies
  - **authorities** as good content sources (~high indegree)
  - **hubs** as good link sources (~high outdegree)
- **HITS** [Kleinberg '99] considers a web page
  - a **good authority** if many **good hubs link to it**
  - a **good hub** if it **links to many good authorities**

  ~ **mutual reinforcement** between hubs & authorities

Jon Kleinberg



HITS

- Given (partial) Web graph $G(V, E)$, let $a(v)$ and $h(v)$ denote the **authority score** and **hub score** of the web page v

$$a(v) \propto \sum_{(u,v) \in E} h(u)$$

$$h(v) \propto \sum_{(v,w) \in E} a(w)$$

- Authority and hub scores in **matrix notation**

$$\boldsymbol{a} = \alpha \, \boldsymbol{A}^T \boldsymbol{h}$$

$$\boldsymbol{h} = \beta \, \boldsymbol{A} \, \boldsymbol{a}$$

with adjacency matrix $\boldsymbol{A}$, hub & authority score vectors $\boldsymbol{a}$ & $\boldsymbol{h}$, and constants α and β

HITS as Eigenvector Computation

- Plugging authority and hub equations into each other produces

$$\boldsymbol{a} = \alpha\, \boldsymbol{A}^T \boldsymbol{h} = \boldsymbol{a} = \alpha\, \boldsymbol{A}^T \beta\, \boldsymbol{A}\, \boldsymbol{a} = \alpha\, \beta \boldsymbol{A}^T \boldsymbol{A}\, \boldsymbol{a}$$

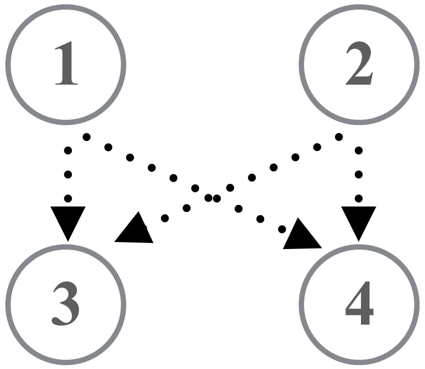
$$\boldsymbol{h} = \beta\, \boldsymbol{A}\, \boldsymbol{a} = \beta\, \boldsymbol{A}\, \alpha\, \boldsymbol{A}^T \boldsymbol{h} = \alpha\, \beta\, \boldsymbol{A}\, \boldsymbol{A}^T \boldsymbol{h}$$

  with $\boldsymbol{a}$ and $\boldsymbol{h}$ as **eigenvectors** of $\boldsymbol{A^T A}$ and $\boldsymbol{A A^T}$, respectively

- Intuitive Interpretation:
  - $\boldsymbol{A^T A}$ is the **cocitation matrix**,
    i.e., $\boldsymbol{A^T A}_{ij}$ is the number of web pages that link to both i and j
  - $\boldsymbol{A A^T}$ is the **coreference matrix**,
    i.e., $\boldsymbol{A A^T}_{ij}$ is the number of web pages to which both i and j link

Cocitation and Coreference Matrix

- **Adjacency matrix A**



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

- **Cocitation matrix A^TA**

$$A^T A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 \end{bmatrix}$$

- **Coreference matrix AA^T**

$$AA^T = \begin{bmatrix} 2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

HITS Algorithm

$\boldsymbol{a}^{(0)} = (1, \ldots, 1)^T$, $\boldsymbol{h}^{(0)} = (1, \ldots, 1)^T$

Repeat until convergence of $\boldsymbol{a}$ and $\boldsymbol{h}$:

$\boldsymbol{h}^{(i+1)} = \boldsymbol{A}\, \boldsymbol{a}^{(i)}$

$\boldsymbol{h}^{(i+1)} = \boldsymbol{h}^{(i+1)} / \mid \boldsymbol{h}^{(i+1)} \mid$ // re-normalize $\boldsymbol{h}$

$\boldsymbol{a}^{(i+1)} = \boldsymbol{A}^{\boldsymbol{T}}\, \boldsymbol{h}^{(i)}$

$\boldsymbol{a}^{(i+1)} = \boldsymbol{a}^{(i+1)} / \mid \boldsymbol{a}^{(i+1)} \mid$ // re-normalize $\boldsymbol{a}$

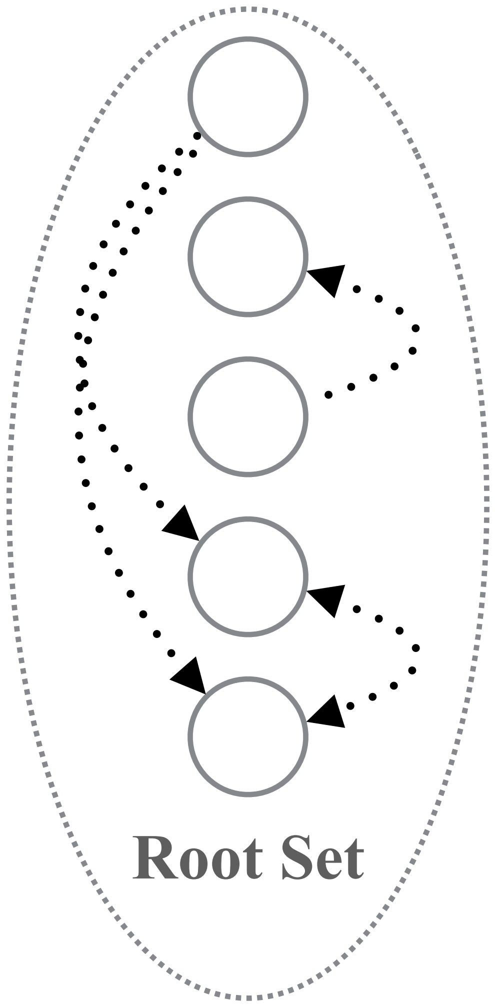
- **Convergence** is **guaranteed** under fairly general conditions:
  - For a symmetric n-by-n matrix $\boldsymbol{M}$ and a vector $\boldsymbol{v}$ that is not orthogonal to the principal eigenvector $w(\boldsymbol{M})$, the unit vector in the direction of $\boldsymbol{M}^k \boldsymbol{v}$ converges to $w(\boldsymbol{M})$ for $k \rightarrow \infty$

Root Set & Expansion Set

- HITS operates on a **query-dependent subgraph of the Web**

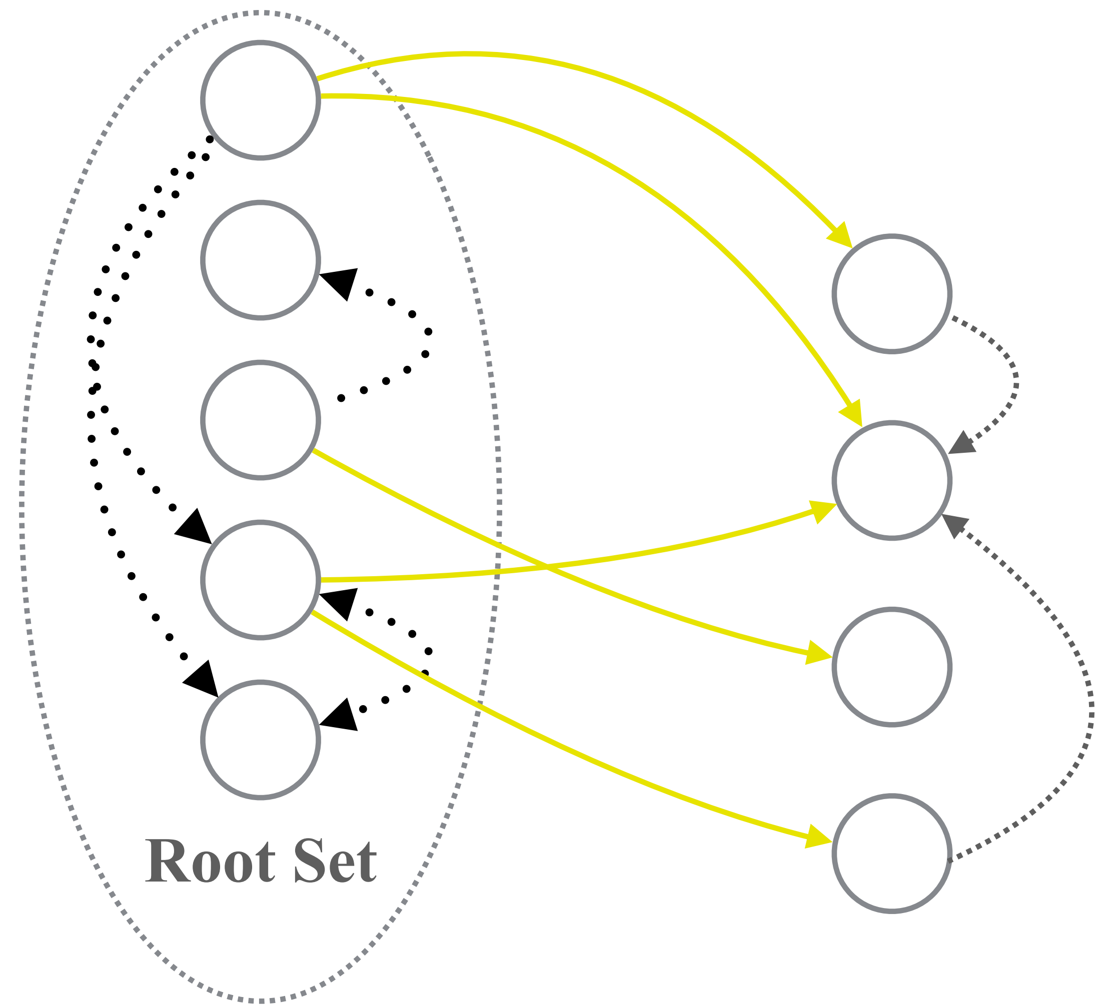
1. Determine sufficient number of **root pages** (e.g., 50-100 pages) based on relevance ranking for query (e.g., using TF*IDF)
2. For each root page, add **all of its successors**
3. For each root page, add **up to *d* predecessors**
4. Compute authority and hub scores on the **query-dependent subgraph** of the Web induced by this **expansion set** (typically: 1000-5000 pages)
5. Return **top-*k* authorities** and **top-*k* hubs**

Root Set & Expansion Set (Example)



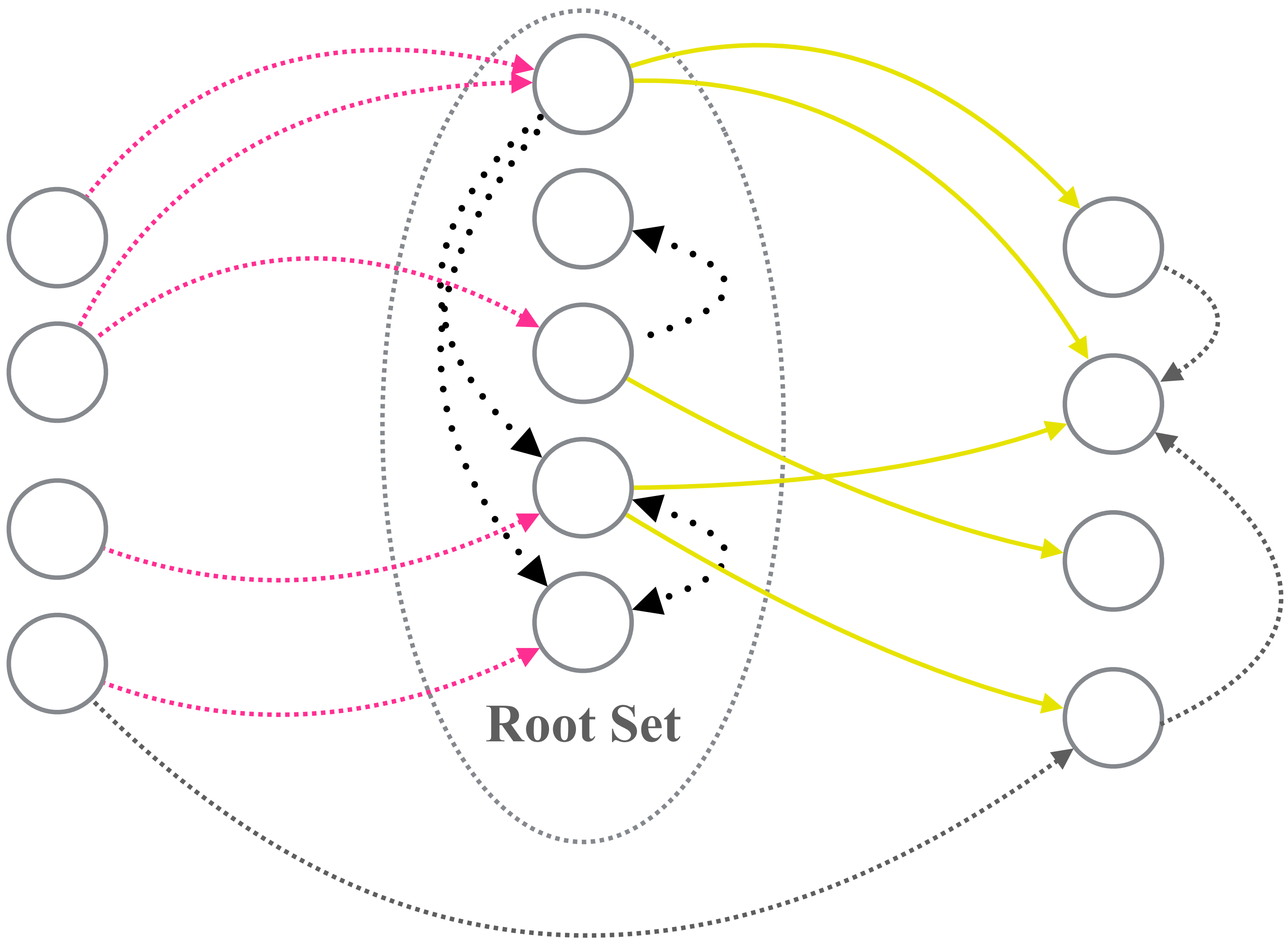
- Shortcoming: **Relevance scores** within root set **not considered**

Root Set & Expansion Set (Example)



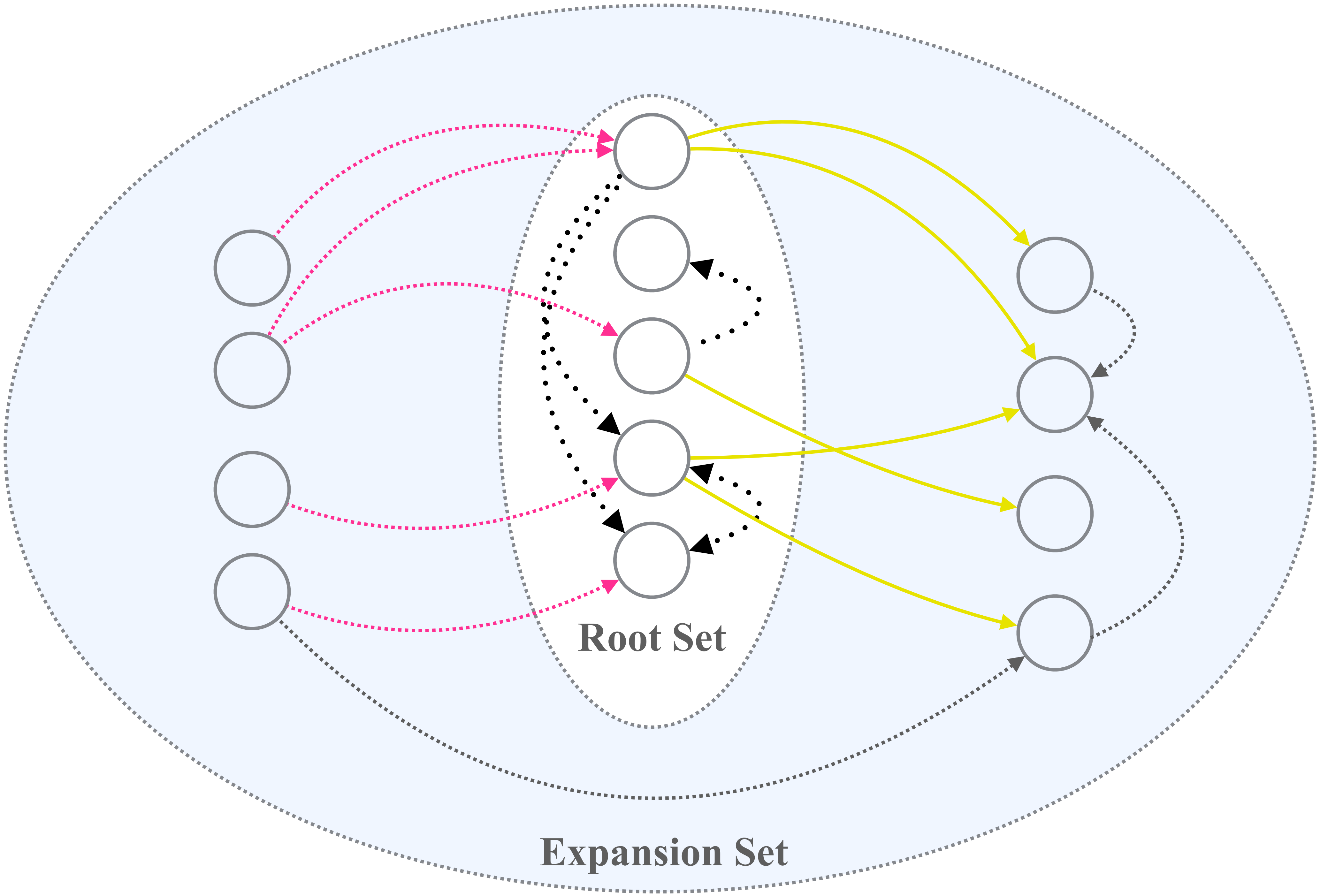
- <u>Shortcoming</u>: **Relevance scores** within root set **not considered**

Root Set & Expansion Set (Example)



- Shortcoming: **Relevance scores** within root set **not considered**

Root Set & Expansion Set (Example)



- Shortcoming: **Relevance scores** within root set **not considered**

Improved HITS

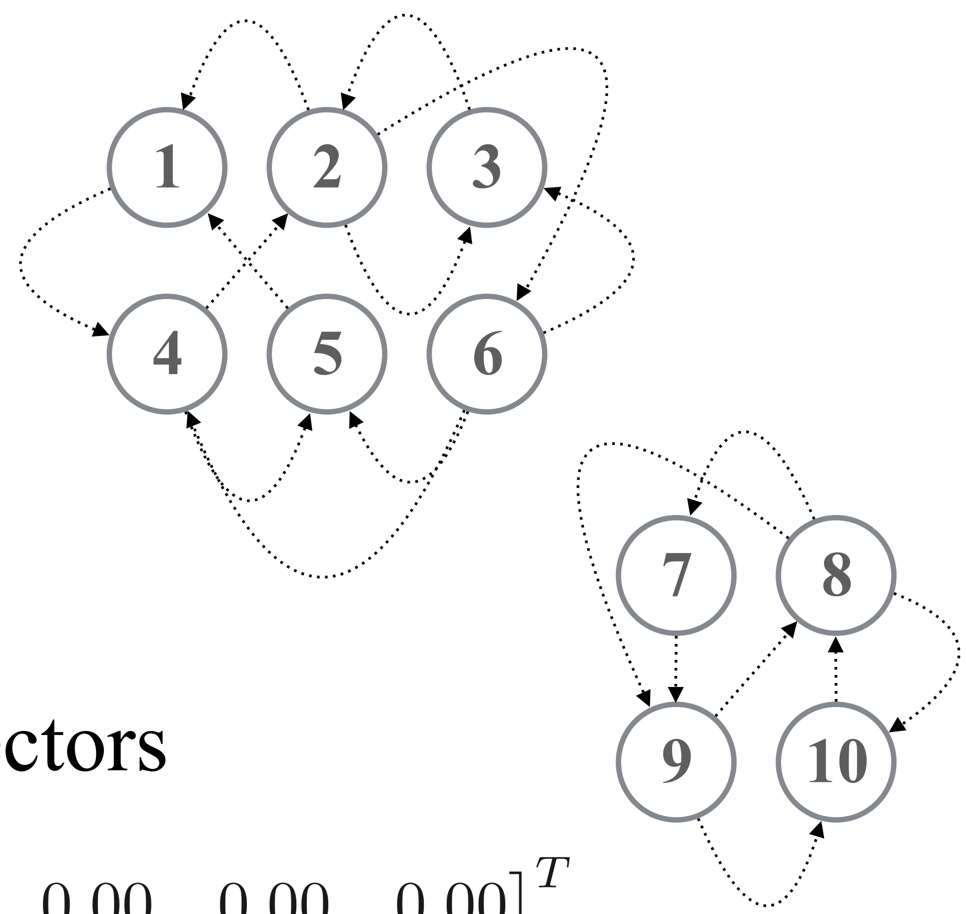
- Potential weaknesses of the HITS algorithm:
  - **irritating links** (e.g., automatically generated links, spam, etc.)
  - **topic drift** (e.g., from *jaguar car* to *car*)
- [Bharat and Henzinger '98] introduce **edge weights**
  - 0 for links within the same host
  - $1/k$ with k links from k URLs of the same host to 1 URL (*aweight*)
  - $1/m$ with m links from 1 URL to m URLs on the same host (*hweight*)
- Consider **relevance weights** $rel(v)$ w.r.t. query (e.g., TF*IDF)

$$a(v) \propto \sum_{(u,v) \in E} h(u) \cdot rel(v) \cdot aweight(u,v)$$

$$h(v) \propto \sum_{(v,w) \in E} a(w) \cdot rel(v) \cdot hweight(v,w)$$

Dominant Subtopics in HITS

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



- HITS returns the authority and hub vectors

$$\boldsymbol{a} = \begin{bmatrix} 0.15 & 0.08 & 0.26 & 0.18 & 0.21 & 0.12 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}^T$$

$$\boldsymbol{h} = \begin{bmatrix} 0.10 & 0.28 & 0.04 & 0.15 & 0.08 & 0.35 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}^T$$

- Observation: Only the nodes {1, …, 6} in the dominant subtopic have a **non-zero authority and hub score**

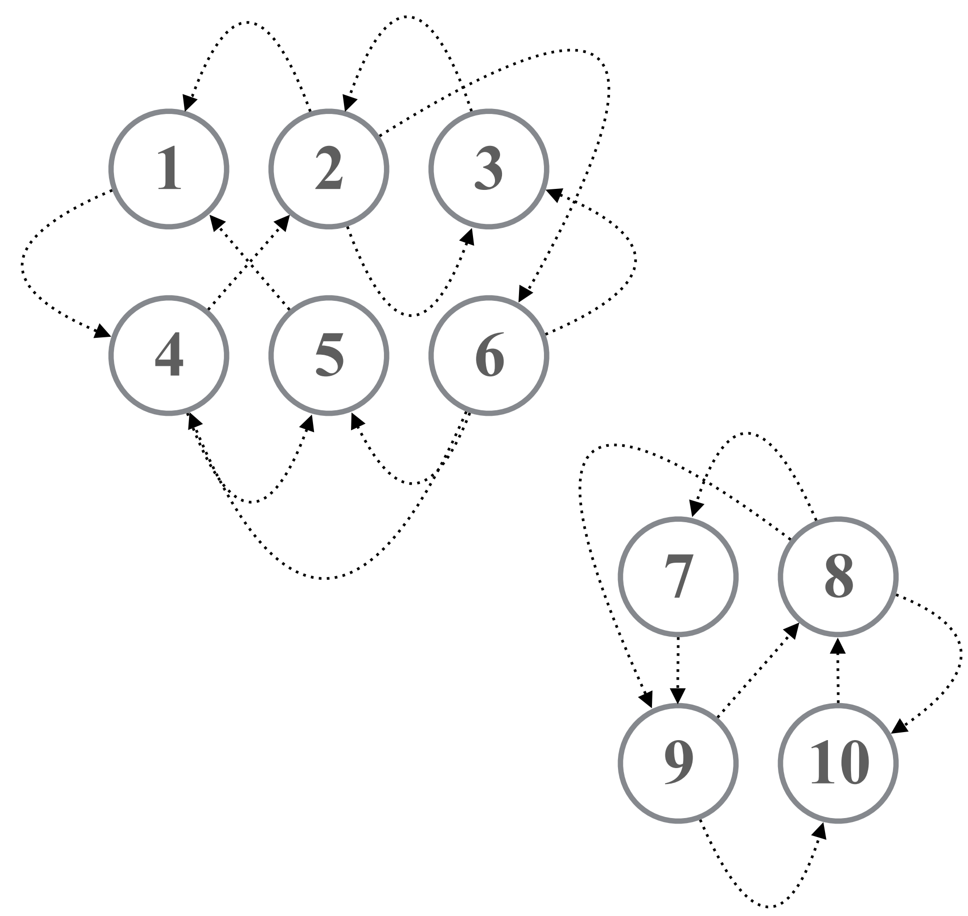
HITS & SVD

- The authority vector $\boldsymbol{a}$ and hub vector $\boldsymbol{h}$ determined by HITS are **eigenvectors** of the matrices $\boldsymbol{AA^T}$ and $\boldsymbol{A^TA}$, respectively

- For $\boldsymbol{A} = \boldsymbol{U\Sigma V^T}$ as the SVD of the adjacency matrix $\boldsymbol{A}$
  - $\boldsymbol{U}$ contains the eigenvectors of $\boldsymbol{AA^T}$ as its columns (with $\boldsymbol{U_1}$ corresponding to the hub vector $\boldsymbol{h}$)
  - $\boldsymbol{V}$ contains the eigenvectors of $\boldsymbol{A^TA}$ as its columns (with $\boldsymbol{V_1}$ corresponding to the authority vector $\boldsymbol{a}$)

HITS & SVD (Example)

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



$$
\boldsymbol{U} = \begin{bmatrix}
-0.20 & 0.00 & -0.14 & 0.00 & -0.39 & 0.70 & 0.00 & 0.29 & 0.00 & -0.43 \\
-0.56 & 0.00 & 0.66 & 0.00 & 0.24 & -0.16 & 0.00 & 0.32 & 0.00 & -0.22 \\
-0.08 & 0.00 & -0.25 & 0.00 & 0.49 & 0.31 & 0.00 & 0.53 & 0.00 & 0.54 \\
-0.31 & 0.00 & -0.53 & 0.00 & 0.54 & -0.08 & 0.00 & -0.25 & 0.00 & -0.49 \\
-0.16 & 0.00 & 0.32 & 0.00 & 0.22 & 0.56 & 0.00 & -0.66 & 0.00 & 0.24 \\
-0.70 & 0.00 & -0.29 & 0.00 & -0.43 & -0.20 & 0.00 & -0.14 & 0.00 & 0.39 \\
0.00 & -0.27 & 0.00 & 0.33 & 0.00 & 0.00 & 0.80 & 0.00 & 0.40 & 0.00 \\
0.00 & -0.80 & 0.00 & 0.40 & 0.00 & 0.00 & -0.27 & 0.00 & -0.33 & 0.00 \\
0.00 & -0.49 & 0.00 & -0.65 & 0.00 & 0.00 & -0.16 & 0.00 & 0.54 & 0.00 \\
0.00 & -0.16 & 0.00 & -0.54 & 0.00 & 0.00 & 0.49 & 0.00 & -0.65 & 0.00
\end{bmatrix}
$$

$$
\boldsymbol{\Sigma} = \begin{bmatrix}
2.12 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.98 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.74 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.48 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.45 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.84 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.81 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.41 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.30
\end{bmatrix}
\quad
\boldsymbol{V} = \begin{bmatrix}
-0.34 & 0.00 & 0.56 & 0.00 & 0.31 & 0.48 & 0.00 & -0.47 & 0.00 & 0.07 \\
-0.19 & 0.00 & -0.45 & 0.00 & 0.71 & 0.26 & 0.00 & 0.37 & 0.00 & 0.16 \\
-0.60 & 0.00 & 0.21 & 0.00 & -0.13 & -0.42 & 0.00 & 0.25 & 0.00 & 0.57 \\
-0.42 & 0.00 & -0.25 & 0.00 & -0.57 & 0.60 & 0.00 & 0.21 & 0.00 & -0.13 \\
-0.48 & 0.00 & -0.47 & 0.00 & 0.07 & -0.34 & 0.00 & -0.56 & 0.00 & -0.31 \\
-0.26 & 0.00 & 0.37 & 0.00 & 0.16 & -0.19 & 0.00 & 0.45 & 0.00 & -0.71 \\
-0.00 & -0.40 & 0.00 & 0.27 & 0.00 & 0.00 & -0.33 & 0.00 & -0.80 & 0.00 \\
-0.00 & -0.33 & 0.00 & -0.80 & 0.00 & 0.00 & 0.40 & 0.00 & -0.27 & 0.00 \\
-0.00 & -0.54 & 0.00 & 0.49 & 0.00 & 0.00 & 0.65 & 0.00 & 0.16 & 0.00 \\
-0.00 & -0.65 & 0.00 & -0.16 & 0.00 & 0.00 & -0.54 & 0.00 & 0.49 & 0.00
\end{bmatrix}
$$

HITS for Community Detection

- Problem: Root set may contain **multiple subtopics or communities** (e.g., for ambiguous queries like *jaguar* or *java*) and HITS may favor only the dominant subtopic

- Approach:

  - Consider the *k* eigenvectors of $A^T A$ associated with the *k* largest eigenvalues (e.g., using SVD on A)

  - For each of these *k* eigenvectors, the largest authority scores indicate a densely connected "community"

- SVD useful as a general tool to **detect communities in graphs**

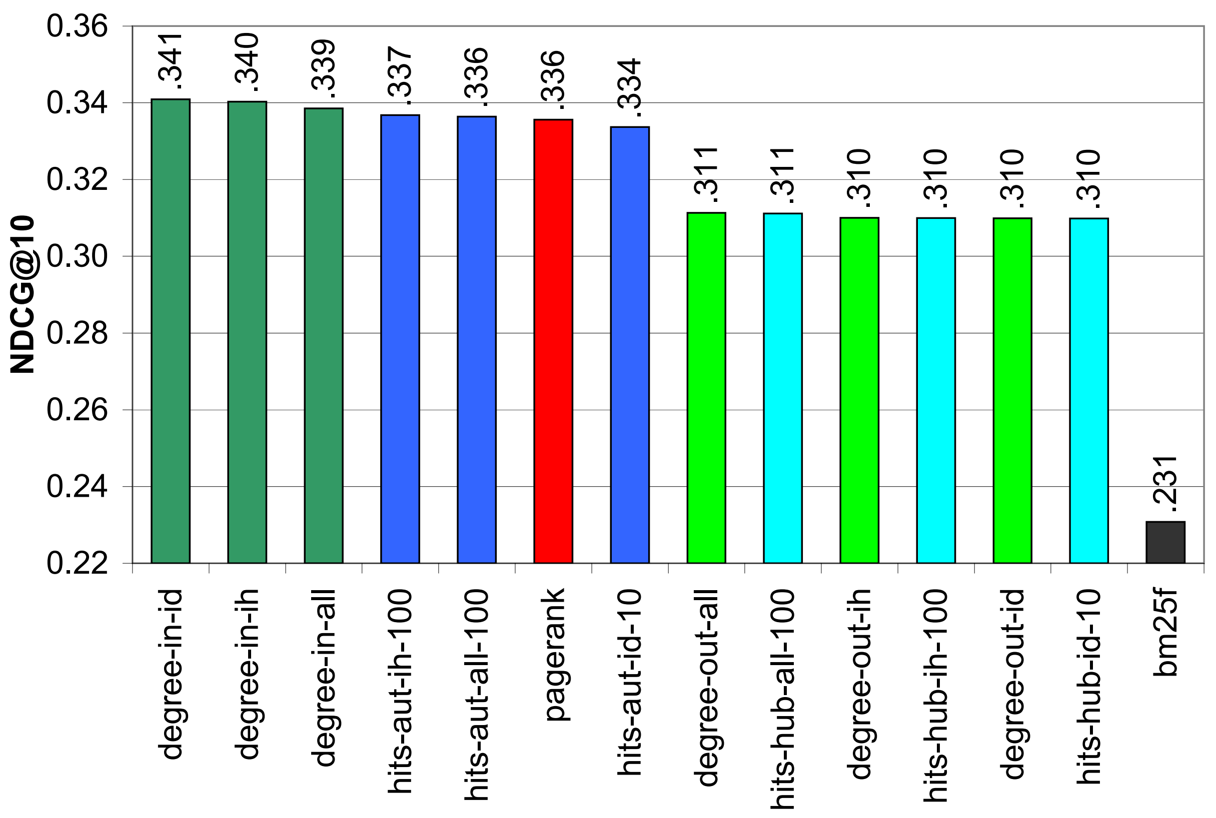
HITS vs. PageRank

	PageRank	HITS
Matrix construction	static	query time
Matrix size	huge	moderate
Stochastic matrix	yes	no
Dampening by random jumps	yes	no
Outdegree normalization	yes	no
Score stability to perturbations	yes	no
Resilience to topic drift	n/a	no
Resilience to spam	no	no

- But: PageRank features (e.g., random jump) could be incorporated into HITS; HITS could be applied to the entire Web; PageRank could also be applied to a query-dependent subgraph

HITS vs. PageRank

- [Najork et al. '07] compare HITS, PageRank, etc. in terms of their **retrieval effectiveness** when combined with Okapi BM25F
- Dataset: Web crawl consisting of **463 M web pages** containing **17.6 M hyperlinks** and referencing **2.9 B distinct URLs**; **28 K queries** sampled from a query log
- Methods:
  - **PageRank**
  - **HITS** (auth / hub)
  - **Degree** (in / out)
    - **all** (all links considered)
    - **id** (only inter-domain links)
    - **in** (only inter-host links)



Summary of IV.3

- **Hubs**
  as web pages that link to good authorities
- **Authorities**
  as web pages that are linked to by good hubs
- **HITS**
  operates on a query-dependent subgraph of the Web
  determines eigenvectors of the matrices AA^T and A^TA
- **SVD**
  helps to circumvent the dominant subtopic problem in HITS
  can be used as a general tool to identify communities in graphs

Additional Literature for IV.3

- **K. Bharat and M. Henzinger:** *Improved Algorithms for Topic Distillation in a Hyperlinked Environment*, SIGIR 1998
- **A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas**: *Link analysis ranking: algorithms, theory, and experiments*. ACM TOIT 5(1), 2005
- **J. Dean and M. Henzinger**: *Finding Related Pages in the World Wide Web*, Computer Networks 31:1467-1479, 1999
- **J. Kleinberg**: *Authoritative sources in a hyperlinked environment*, Journal of the ACM 46:604-632, 1999
- **M. Najork, H. Zaragoza, and M. Taylor**: *HITS on the Web: How does it Compare?*, SIGIR 2007