Information Storage and Retrieval

CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

Link Analysis: Hubs and Authorities
9 February 2017

Hubs & Authorities

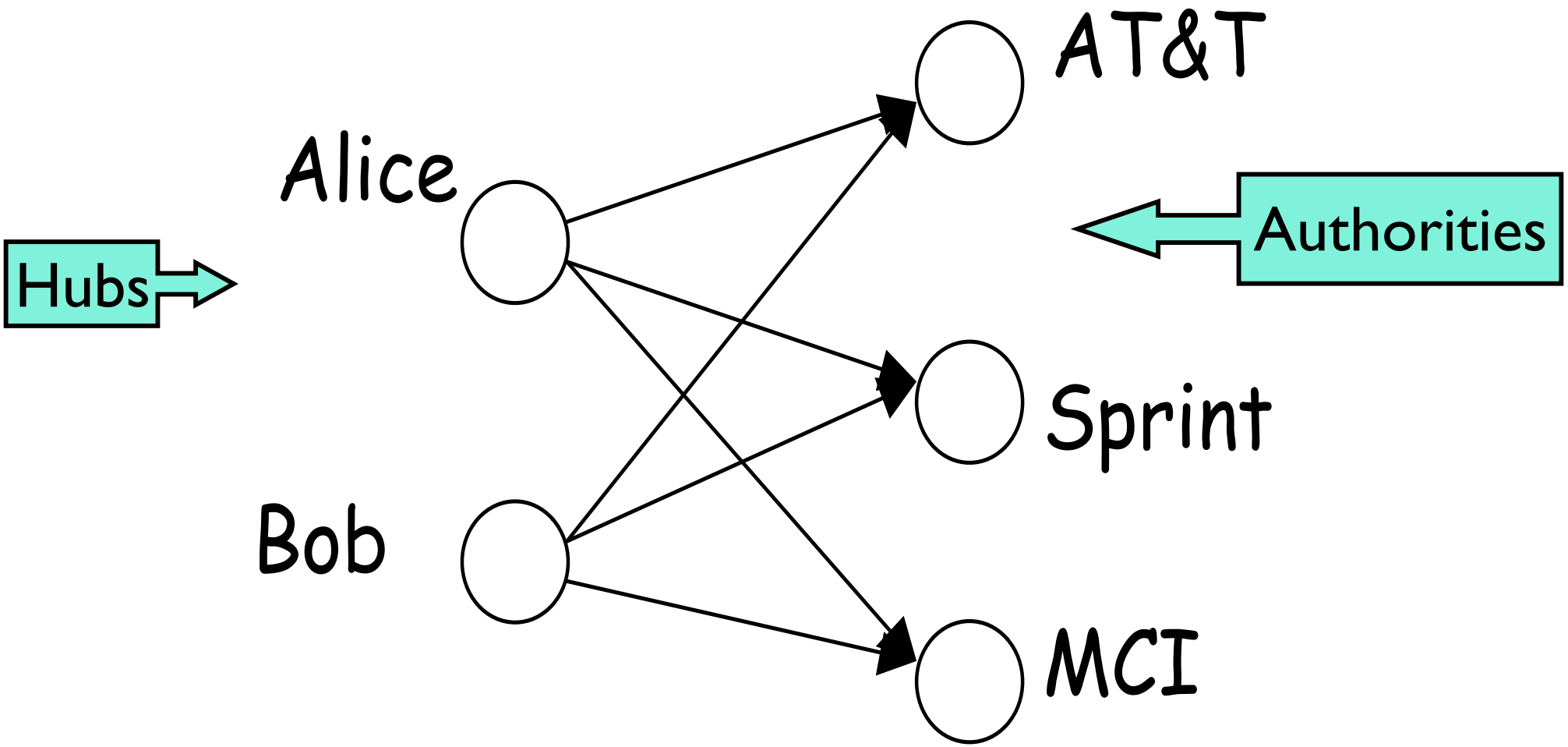
HITS - Hyperlink-Induced Topic Search

- Premise: there are two different types of relevance on the web.
- Relevance type 1: **Hubs.** A hub page is a good list of links to pages answering the information need.
  - Bob's list of recommended hotels in London
- Relevance type 2: **Authorities.** An authority page is a direct answer to the information need. Authority pages occur repeatedly on hub pages.
  - Home page of Four Seasons Hotel London
- Most approaches to search (including PageRank ranking) don't make the distinction between these two very different types of relevance.

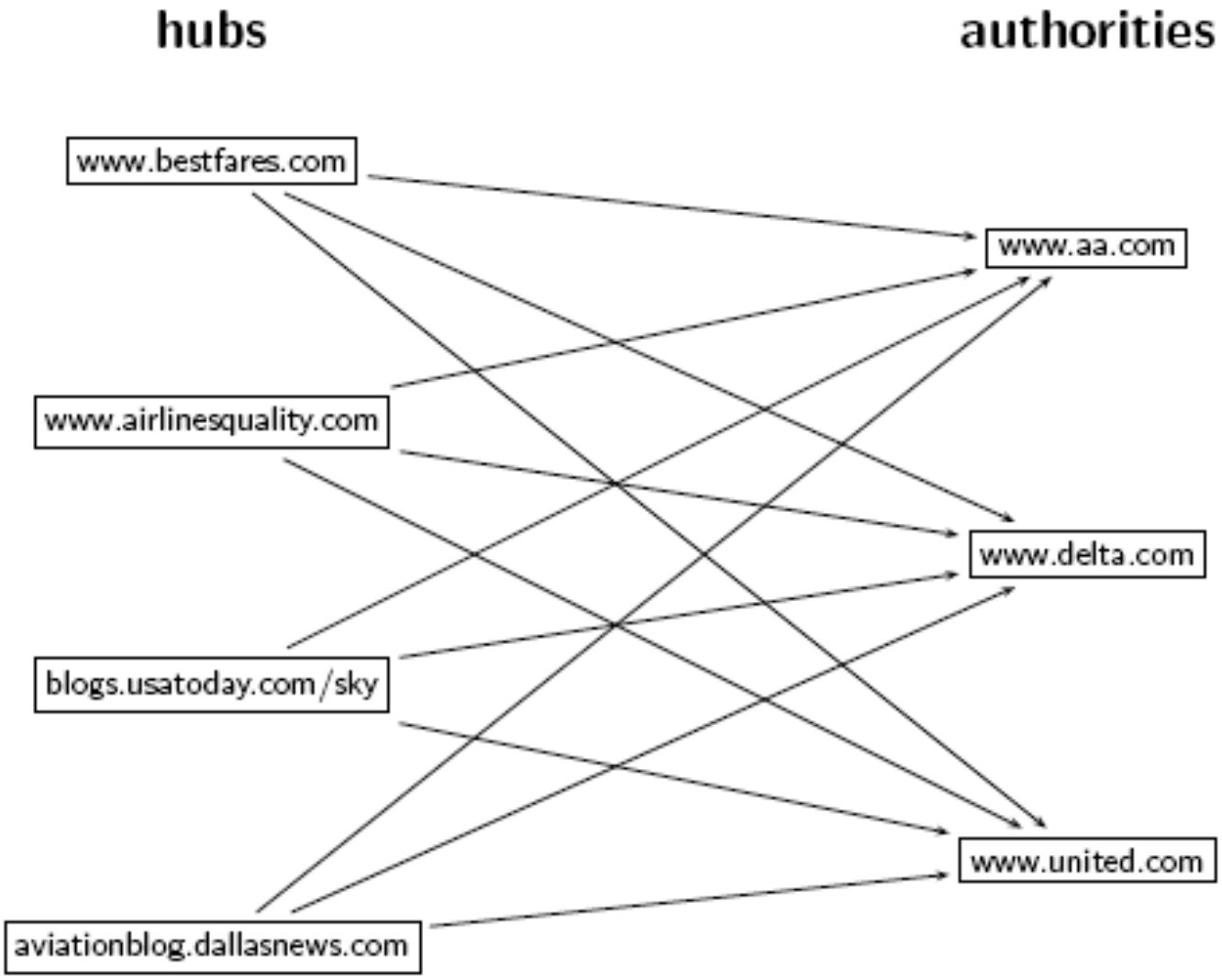
Hubs and Authorities

- Thus, a good hub page for a topic *points* to many authoritative pages for that topic.
- A good authority page for a topic is *pointed* to by many good hubs for that topic.
- Circular definition - will turn this into an iterative computation.

The hope



Long distance telephone companies

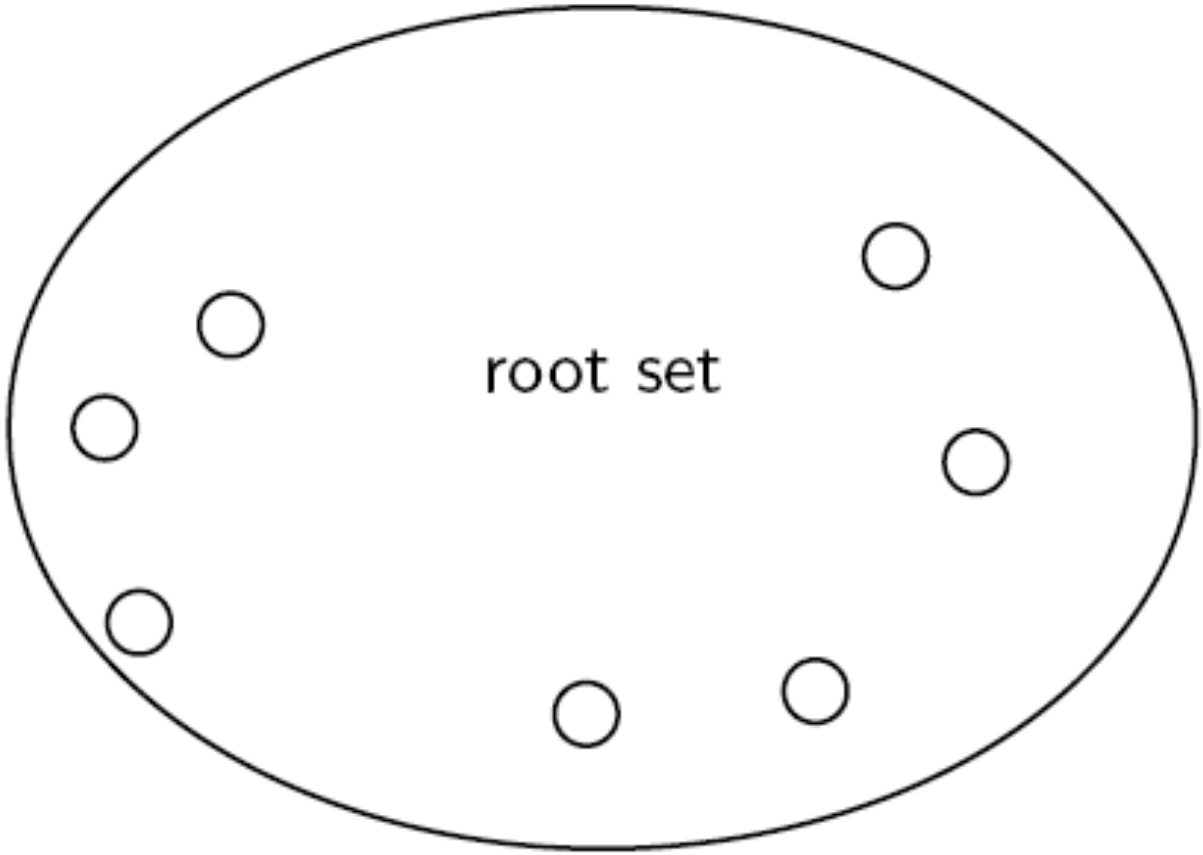

hubs
authorities
www.bestfares.com
www.airlinesquality.com
blogs.usatoday.com/sky
aviationblog.dallasnews.com
www.aa.com
www.delta.com
www.united.com

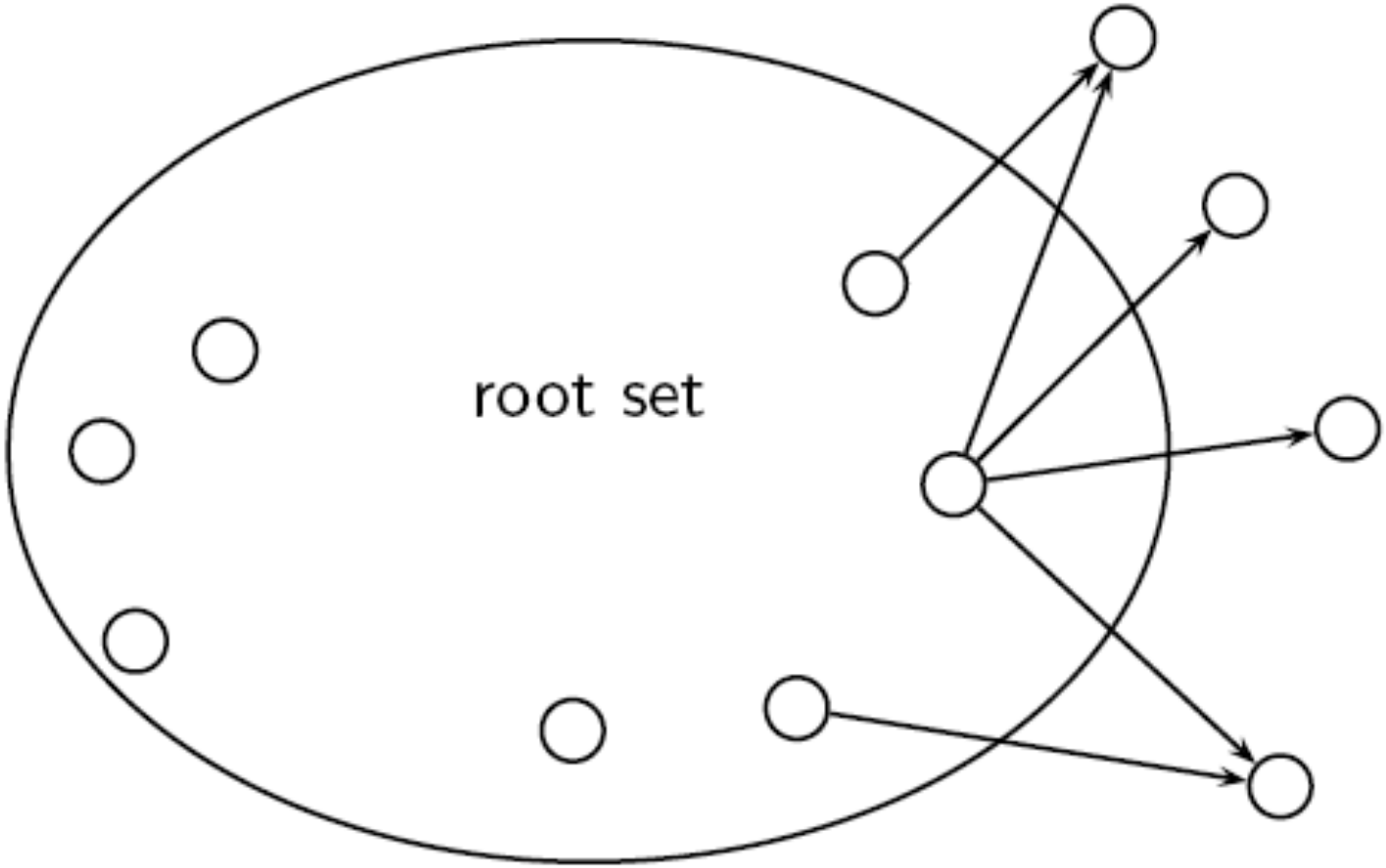
High-level scheme

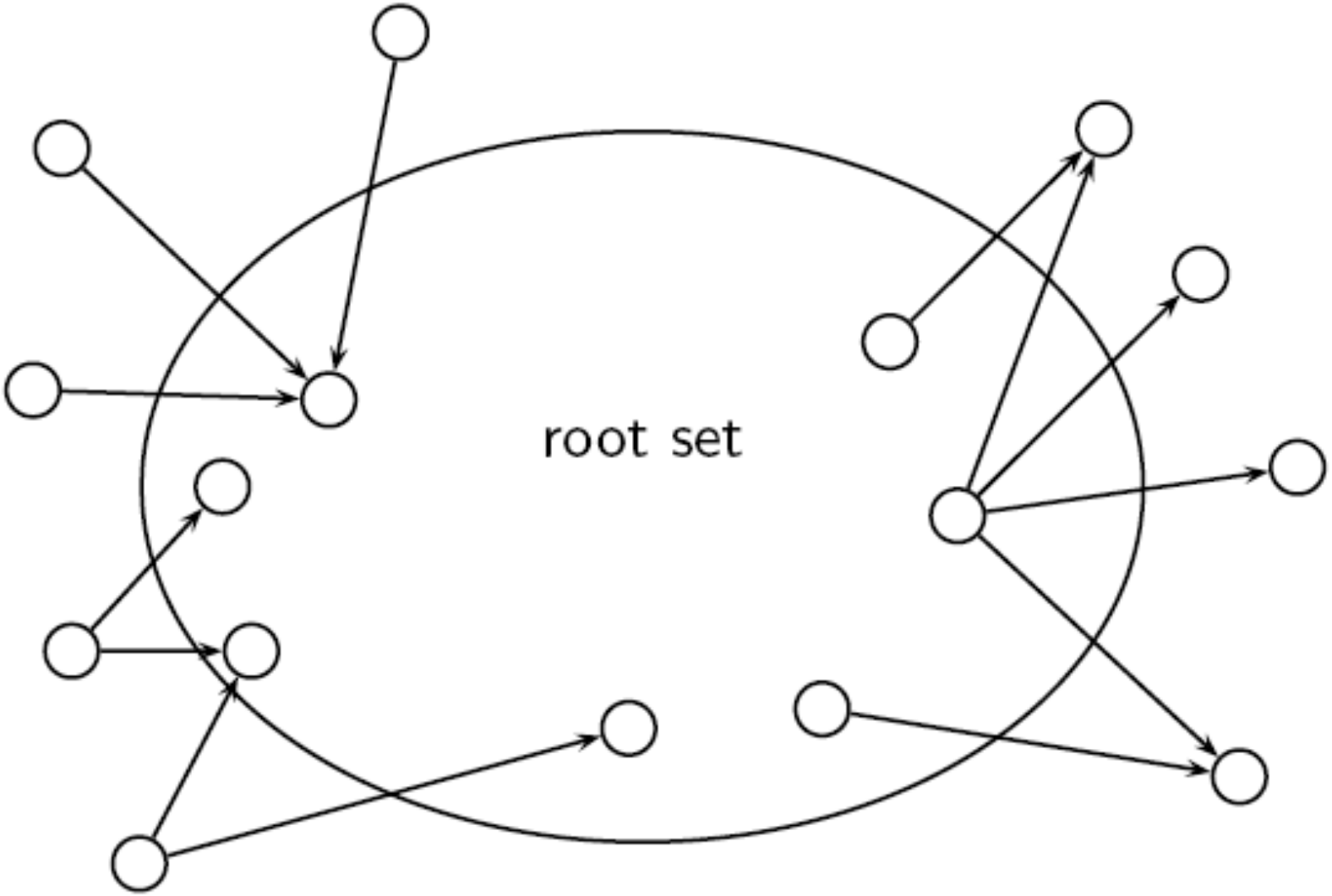
- Extract from the web a base set of pages that *could* be good hubs or authorities.
- From these, identify a small set of top hub and authority pages;
- → iterative algorithm.

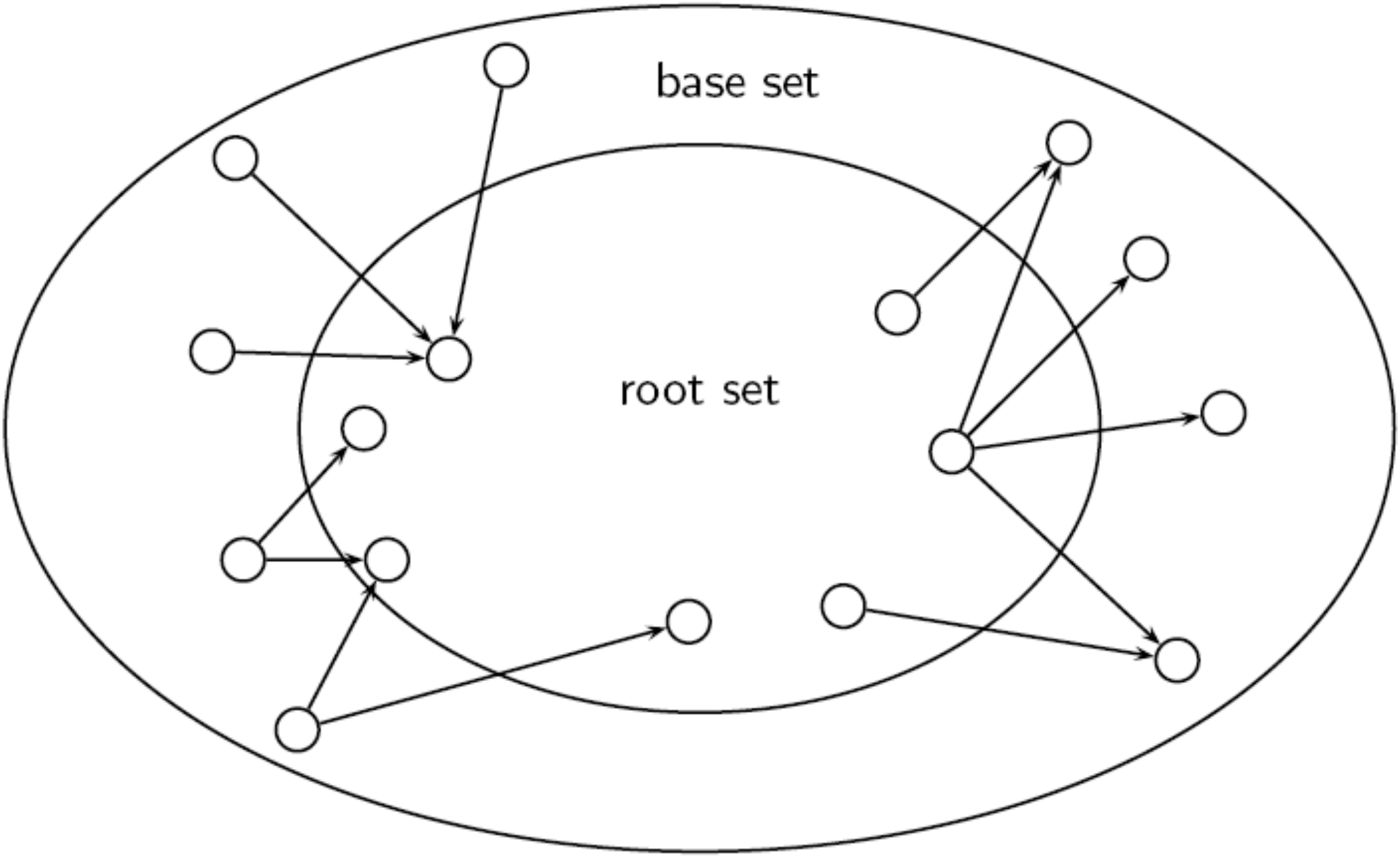
Root set and base set

- Do a regular web search first
- Call the search result the **root set**
- Find all pages that are linked to or link to pages in the root set
- Call this larger set the **base set**
- Finally, compute hubs and authorities for this (small) web graph


root set


root set


root set


base set
root set

Root set and base set

- Root set typically 200-1000 nodes.
- Base set may have up to 5000 nodes.
- How do you find the base set nodes?
  - Follow out-links by parsing root set pages.
  - Find d's in-links by searching for all pages containing a link to d
  - This assumes our inverted index supports search for links (in addition to terms)

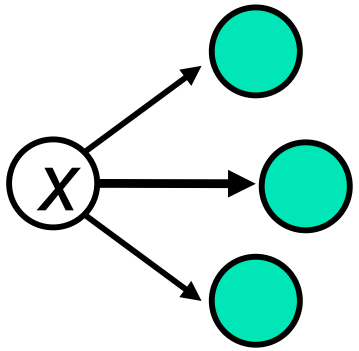
Hub and authority scores

- Compute, for each page *x* in the base set, a hub score *h(x)* and an authority score *a(x).*
- Initialize: for all *x*, *h(x)*← *1; a(x)* ← *1*;
- Iteratively update all *h(x), a(x)*;
- After convergence
  - output pages with highest *h()* scores as top hubs
  - highest *a()* scores as top authorities.

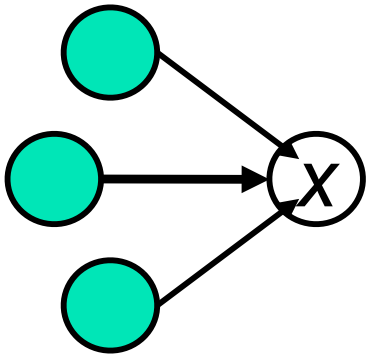
Iterative update

- Repeat the following updates, for all x:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



Scaling

- To prevent the a() and h() values from getting too big, can scale down after each iteration.
- Scaling factor doesn't really matter.
- We only care about the **relative** values of the scores

How many iterations?

- Claim: relative values of scores will converge after a few iterations:
  - in fact, suitably scaled, *h()* and *a()* scores settle into a steady state!
- We only require the relative orders of the *h()* and *a()* scores - not their absolute values.
- In practice, ~5 iterations get you close to stability.

Japan Elementary Schools

Hubs

schools
LINK Page-13
"ú–{‚ÌŠw□Z
□a‰„□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
100 Schools Home Pages (English)
K-12 from Japan 10/...rnet and Education)
http://www...iglobe.ne.jp/~IKESAN
‚l‚f‚j□¬Šw□Z‚U"N‚P'g•¨Œê
□ÒŠ—'¬—§□ÒŠ—"Œ□¬Šw□Z
Koulutus ja oppilaitokset
TOYODA HOMEPAGE
Education
Cay's Homepage(Japanese)
–y"ì□¬Šw□Z‚Ìƒz□[ƒ□ƒy□[ƒW
UNIVERSITY
‰J—³□¬Šw□Z DRAGON97-TOP
□Â‰³□¬Šw□Z‚T"N‚P'gƒz□[ƒ□ƒy□[ƒW

Authorities

The American School in Japan
The Link Page
‰³□è□s—§ˆä"c□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
Kids' Space
ˆÀ□é□s—§ˆÀ□é□¼•"□¬Šw□Z
‹{□é‹³ç'åŠw•□'®□¬Šw□Z
KEIMEI GAKUEN Home Page (Japanese)
Shiranuma Home Page
fuzoku-es.fukui-u.ac.jp
welcome to Miasa E&J school
□_"Þ□ìŒ§□E‰¡•l□s—§'†□ì□¼□¬Šw□Z‚Ìƒy
http://www...p/~m_maru/index.html
fukui haruyama-es HomePage
Torisu primary school
goo
Yakumo Elementary,Hokkaido,Japan
FUZOKU Home Page

Things to note

- Pulled together good pages regardless of language of page content.
- Use *only* link analysis after base set assembled
  - iterative scoring is query-independent.
- Iterative computation after text index retrieval - significant overhead.

Hub/authority vectors

- View the hub scores *h()* and the authority scores *a()* as vectors with *n* components.
- Recall the iterative updates

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$

$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

Rewrite in matrix form

- $\mathbf{h}=A\mathbf{a}$.
- $\mathbf{a}=A^t\mathbf{h}$.

Recall A^t is the transpose of A.

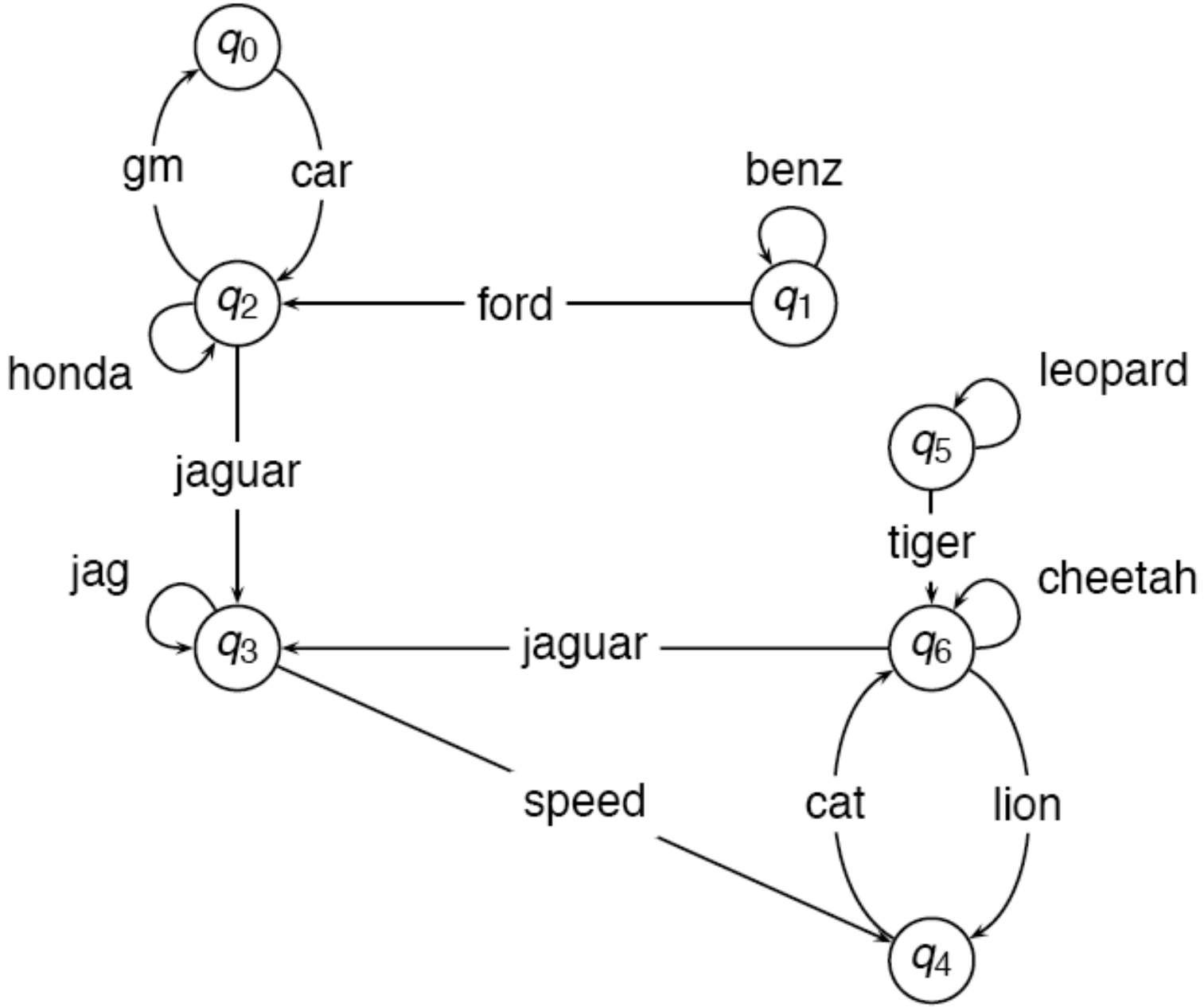
Substituting, $\mathbf{h}=AA^t\mathbf{h}$ and $\mathbf{a}=A^tA\mathbf{a}$.

Thus, **h** is an eigenvector of AA^t and **a** is an eigenvector of A^tA.

Further, our algorithm is a particular, known algorithm for computing eigenvectors: the *power iteration* method.

Guaranteed to converge.

Web graph example



Raw matrix H

	q_0	q_1	q_2	q_3	q_4	q_5	q_6
q_0	0	0	1	0	0	0	0
q_1	0	1	1	0	0	0	0
q_2	1	0	1	2	0	0	0
q_3	0	0	0	1	1	0	0
q_4	0	0	0	0	0	0	1
q_5	0	0	0	0	0	1	1
q_6	0	0	0	2	1	0	1

Hub vectors

	$\vec{h}_0$	$\vec{h}_1$	$\vec{h}_2$	$\vec{h}_3$	$\vec{h}_4$	$\vec{h}_5$
q_0	0.14	0.06	0.04	0.04	0.03	0.03
q_1	0.14	0.08	0.05	0.04	0.04	0.04
q_2	0.14	0.28	0.32	0.33	0.33	0.33
q_3	0.14	0.14	0.17	0.18	0.18	0.18
q_4	0.14	0.06	0.04	0.04	0.04	0.04
q_5	0.14	0.08	0.05	0.04	0.04	0.04
q_6	0.14	0.30	0.33	0.34	0.35	0.35

Authority vectors

	$\vec{a}_1$	$\vec{a}_2$	$\vec{a}_3$	$\vec{a}_4$	$\vec{a}_5$	$\vec{a}_6$	$\vec{a}_7$
q_0	0.06	0.09	0.10	0.10	0.10	0.10	0.10
q_1	0.06	0.03	0.01	0.01	0.01	0.01	0.01
q_2	0.19	0.14	0.13	0.12	0.12	0.12	0.12
q_3	0.31	0.43	0.46	0.46	0.46	0.47	0.47
q_4	0.13	0.14	0.16	0.16	0.16	0.16	0.16
q_5	0.06	0.03	0.02	0.01	0.01	0.01	0.01
q_6	0.19	0.14	0.13	0.13	0.13	0.13	0.13

Top-ranked pages

- Pages with highest indegree: q2, q3, q6
- Pages with highest outdegree: q2, q6
- Pages with highest Pagerank: q6
- Pages with highest hub score: q6 (close: q2)
- Pages with highest authority score: q3

PageRank vs. HITS

- PageRank can be precomputed, HITS has to be computed at query time.
  - HITS is too expensive in most application scenarios.
- The PageRank and HITS make two different design choices concerning (i) the eigenproblem formalization (ii) the set of pages to apply the formalization to.
- These two are orthogonal.
  - We could also apply HITS to the entire web and PageRank to a small base set.
- On the web, a good hub almost always is also a good authority.
- Why?
- The actual difference between PageRank ranking and HITS ranking is therefore not as large as one might expect.

Issues

- Topic Drift
  - Off-topic pages can cause off-topic "authorities" to be returned
    - E.g., the neighborhood graph can be about a "super topic"
- Mutually Reinforcing Affiliates
  - Affiliated pages/sites can boost each others' scores
    - Linkage between affiliated pages is not a useful signal