Link Analysis

CS5154/6054

Yizong Cheng

11/22/2022

21 *Link analysis*

Introduction to

Information Retrieval

CS276

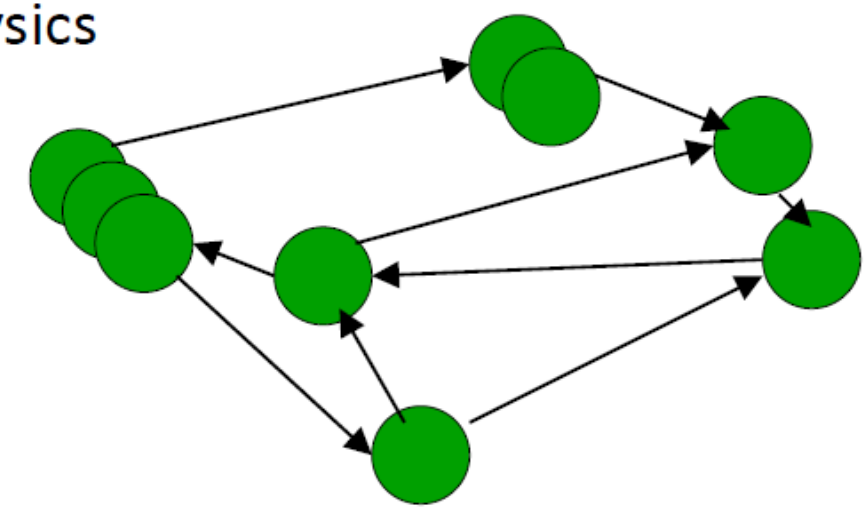
Information Retrieval and Web Search

Chris Manning and Pandu Nayak

Link analysis

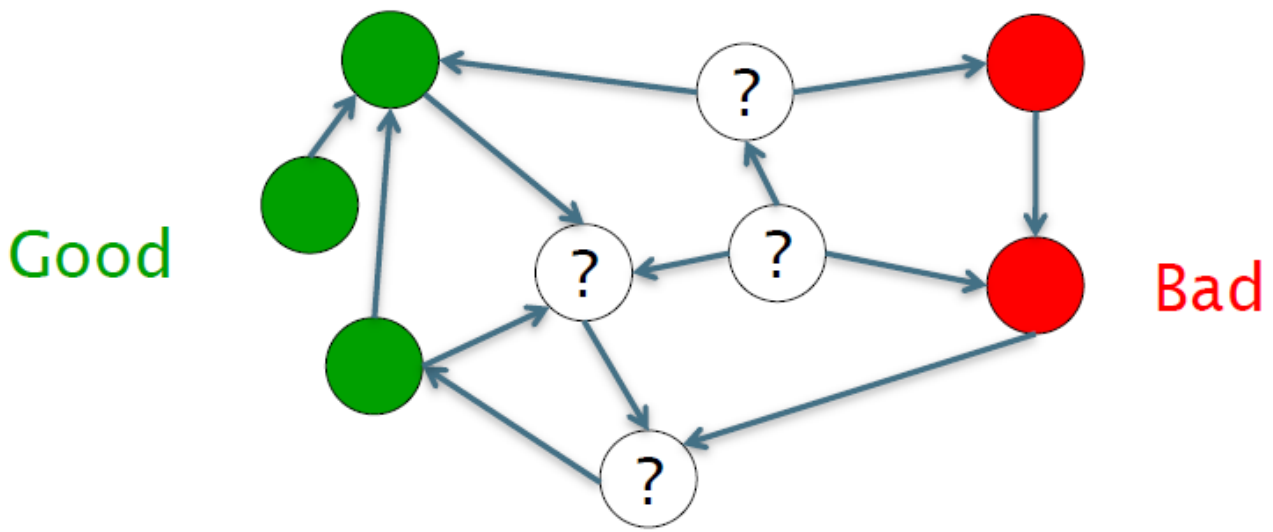
Today's lecture – hypertext and links

- We look beyond the *content* of documents
  - We begin to look at the hyperlinks between them
- Address questions like
  - Do the links represent a conferral of authority to some pages? Is this useful for ranking?
  - How likely is it that a page pointed to by the CERN home page is about high energy physics
- Big application areas
  - The Web
  - Email
  - Social networks



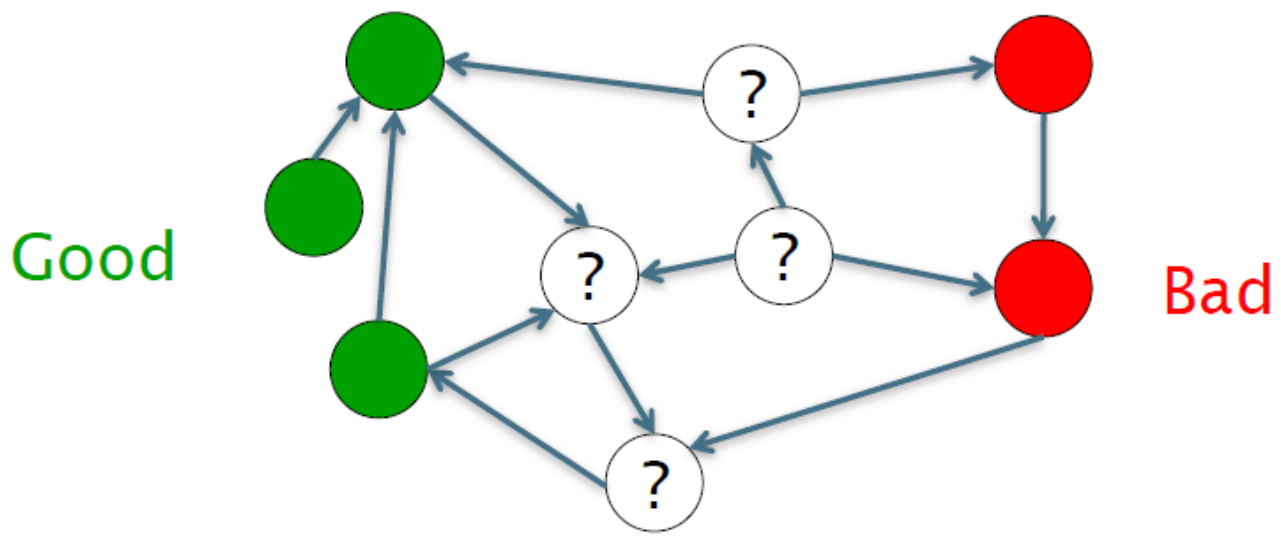
Links are everywhere

- Powerful sources of authenticity and authority
  - Mail spam – which email accounts are spammers?
  - Host quality – which hosts are "bad"?
  - Phone call logs
- The Good, The Bad and The Unknown



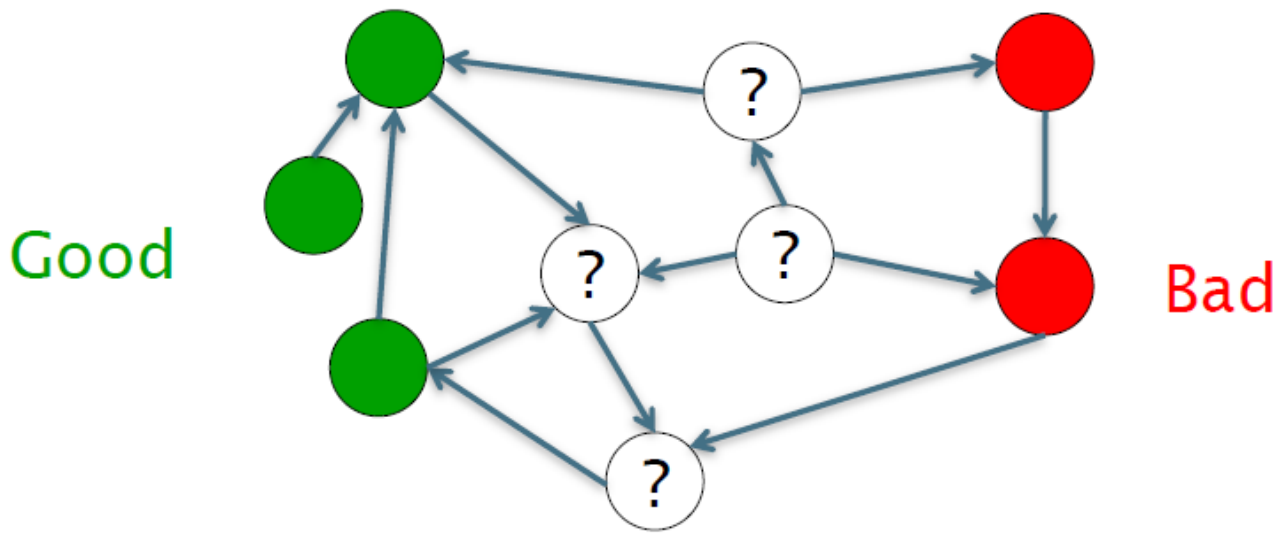
Example 1: Good/Bad/Unknown

- The Good, The Bad and The Unknown
  - Good nodes won't point to Bad nodes
  - All other combinations plausible



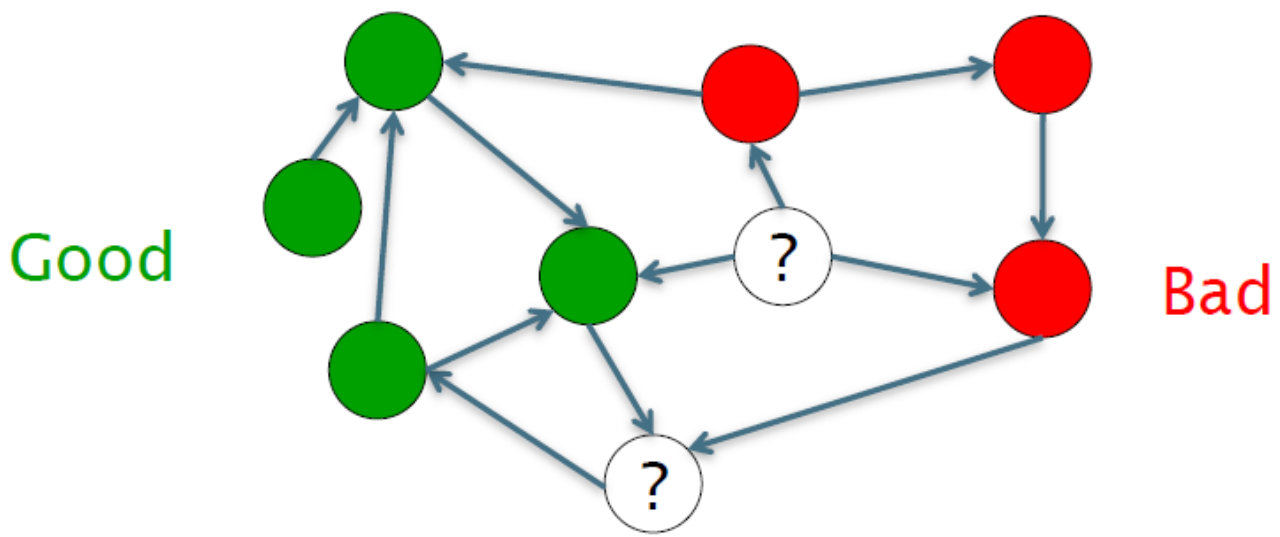
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



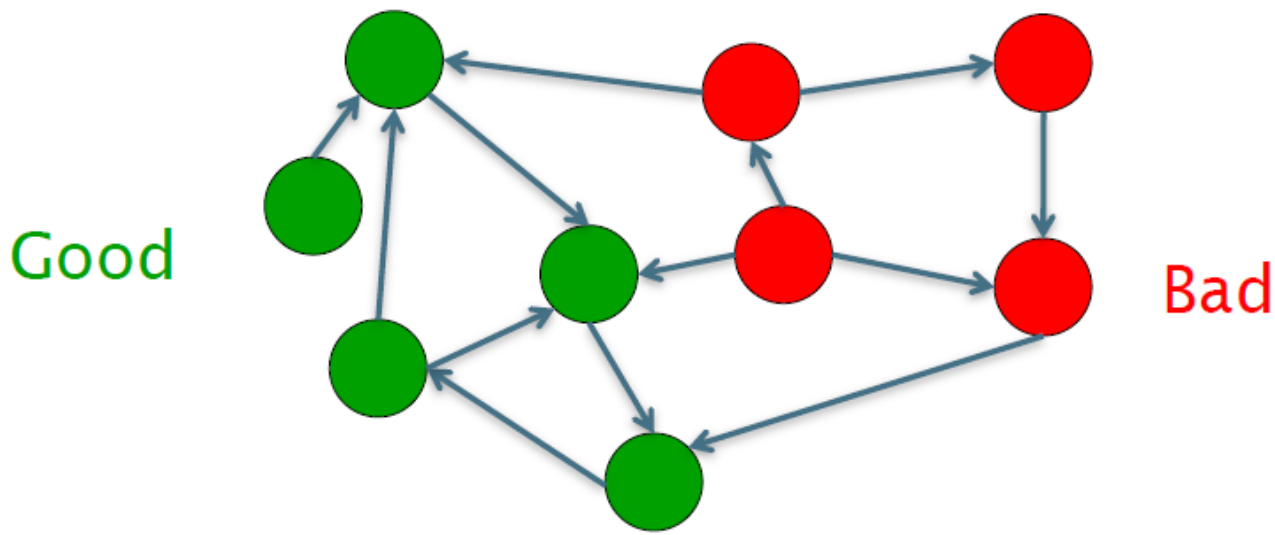
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



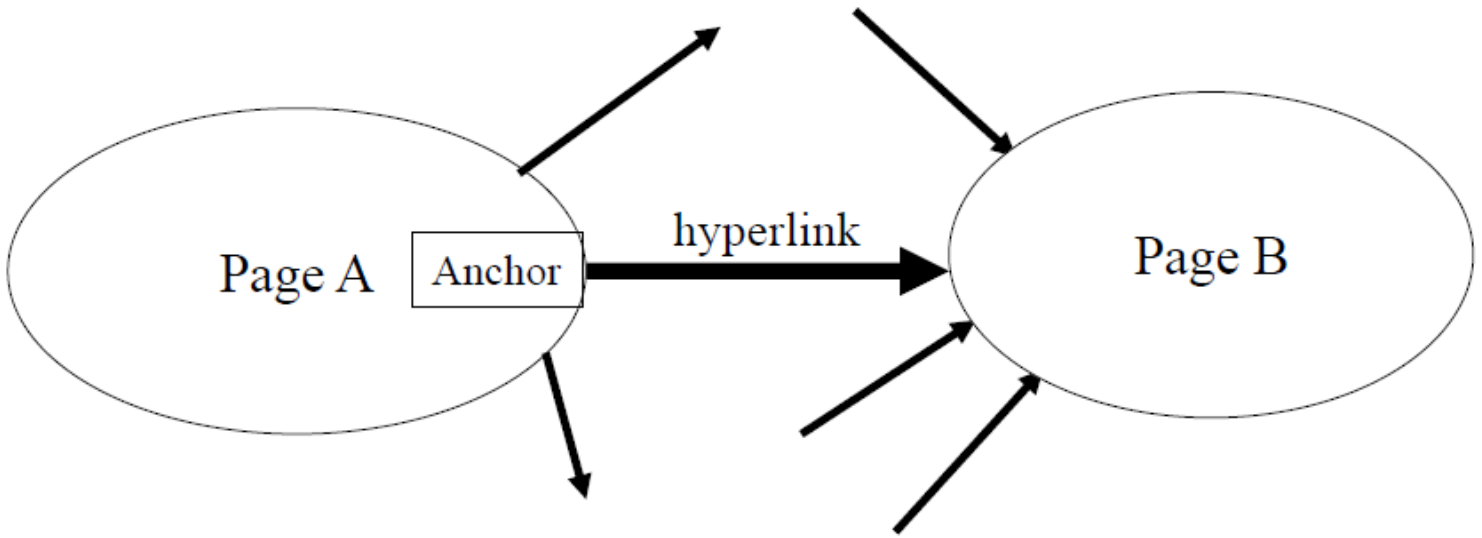
Simple iterative logic

- Good nodes won't point to Bad nodes
  - If you point to a Bad node, you're Bad
  - If a Good node points to you, you're Good



Sometimes need probabilistic analogs – e.g., mail spam

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 a hyperlink on page A describes the target page B

21.3 Hubs and Authorities

Link analysis: HITS

Kleinberg (1999)

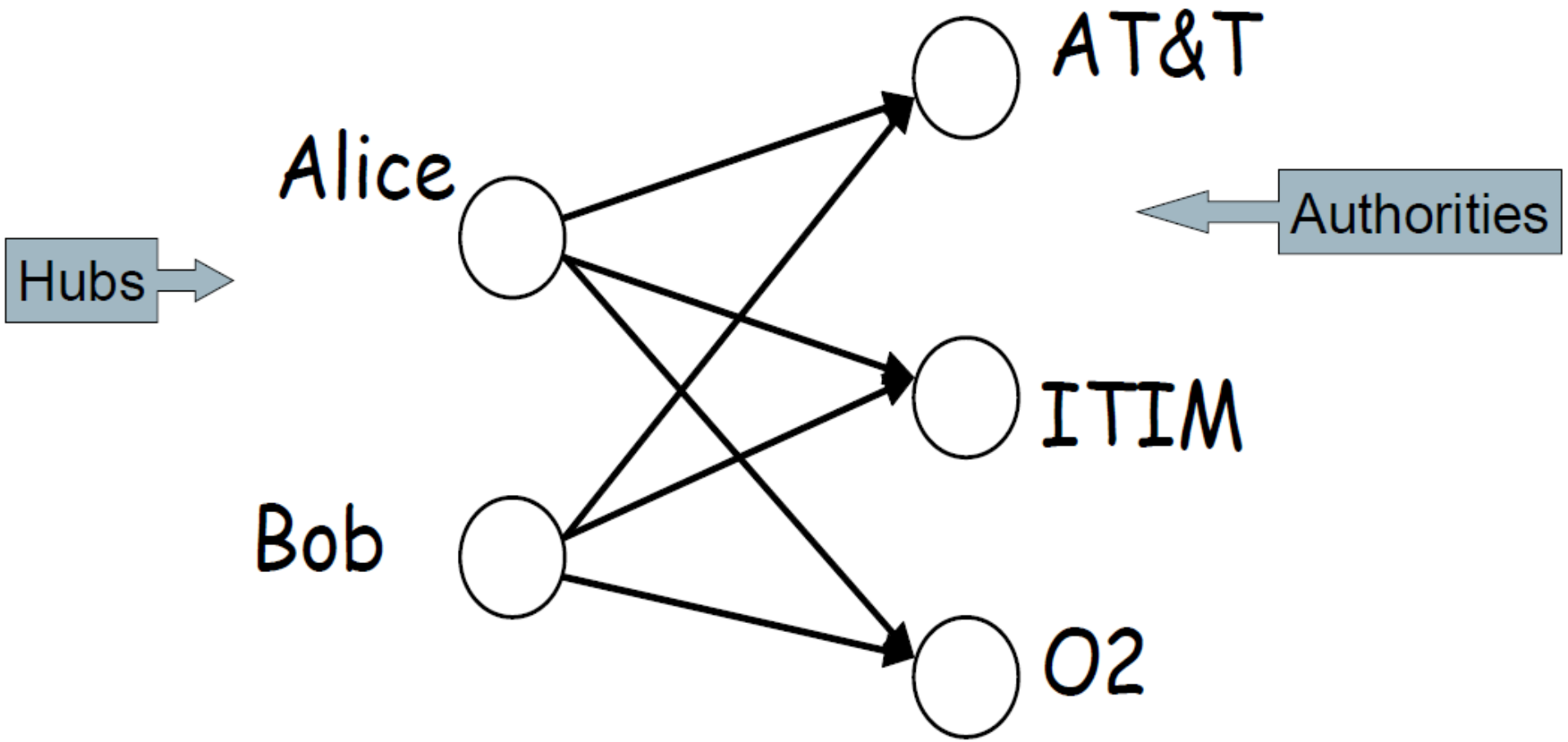
Hyperlink-Induced Topic Search (HITS)

- In response to a query, instead of an ordered list of pages each meeting the query, find two sets of inter-related pages:
  - *Hub pages* are good lists of links on a subject
    - e.g., "Bob's list of cancer-related links."
  - *Authority pages* occur recurrently on good hubs for the subject
- Best suited for "broad topic" queries rather than for page-finding queries
- Gets at a broader slice of common *opinion*

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

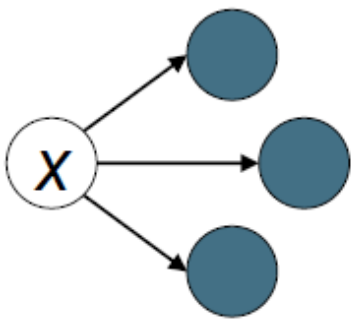


Mobile telecom companies

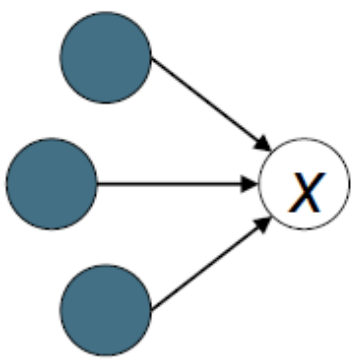
Iterative update

- 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 *h()* and *a()* 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.

Hub/authority vectors

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

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

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

For a web page v in our subset of the web, we use $h(v)$ to denote its hub score and $a(v)$ its authority score. Initially, we set $h(v) = a(v) = 1$ for all nodes v. We also denote by $v \mapsto y$ the existence of a hyperlink from v to y. The core of the iterative algorithm is a pair of updates to the hub and authority scores of all pages given by Equation 21.8, which capture the intuitive notions that good hubs point to good authorities and that good authorities are pointed to by good hubs.

$$\begin{aligned} h(v) &\leftarrow \sum_{v \mapsto y} a(y) \\ a(v) &\leftarrow \sum_{y \mapsto v} h(y). \end{aligned} \tag{21.8}$$

Thus, the first line of Equation (21.8) sets the hub score of page v to the sum of the authority scores of the pages it links to. In other words, if v links to pages with high authority scores, its hub score increases. The second line plays the reverse role; if page v is linked to by good hubs, its authority score increases.

What happens as we perform these updates iteratively, recomputing hub scores, then new authority scores based on the recomputed hub scores, and so on? Let us recast the equations Equation (21.8) into matrix-vector form. Let $\vec{h}$ and $\vec{a}$ denote the vectors of all hub and all authority scores respectively, for the pages in our subset of the web graph. Let A denote the adjacency matrix of the subset of the web graph that we are dealing with: A is a square matrix with one row and one column for each page in the subset. The entry A_{ij} is 1 if there is a hyperlink from page i to page j, and 0 otherwise. Then, we may write Equation (21.8)

$$\begin{aligned} \vec{h} &\leftarrow A\vec{a} \\ \vec{a} &\leftarrow A^T\vec{h}, \end{aligned} \tag{21.9}$$

where A^T denotes the transpose of the matrix A. Now the right hand side of each line of Equation (21.9) is a vector that is the left hand side of the other line of Equation (21.9). Substituting these into one another, we may rewrite Equation (21.9) as

$$\begin{aligned} \vec{h} &\leftarrow AA^T\vec{h} \\ \vec{a} &\leftarrow A^TA\vec{a}. \end{aligned} \tag{21.10}$$

Now, Equation (21.10) bears an uncanny resemblance to a pair of eigenvector equations (Section 18.1); indeed, if we replace the $\leftarrow$ symbols by $=$ symbols and introduce the (unknown) eigenvalue, the first line of Equation (21.10) becomes the equation for the eigenvectors of AA^T, while the second becomes the equation for the eigenvectors of A^TA:

$$\begin{aligned} \vec{h} &= (1/\lambda_h)AA^T\vec{h} \\ \vec{a} &= (1/\lambda_a)A^TA\vec{a}. \end{aligned} \tag{21.11}$$

Here we have used λ_h to denote the eigenvalue of AA^T and λ_a to denote the eigenvalue of A^TA.

The resulting computation thus takes the following form:

1. Assemble the target subset of web pages, form the graph induced by their hyperlinks and compute AA^T and A^TA.
2. Compute the principal eigenvectors of AA^T and A^TA to form the vector of hub scores $\vec{h}$ and authority scores $\vec{a}$.
3. Output the top-scoring hubs and the top-scoring authorities.

HITS

This method of link analysis is known as *HITS*, which is an acronym for *Hyperlink-Induced Topic Search*.

AA^T and A^TA are Symmetric Matrices

- Matrix B is symmetric if and only if $B = B^T$.
- $(CD)^T = D^TC^T$.
- $(AA^T)^T = (A^T)^TA^T = AA^T$.
- Similarly, A^TA is symmetric because $(A^TA)^T = A^T(A^T)^T = A^TA$.

Rewrite in matrix form

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

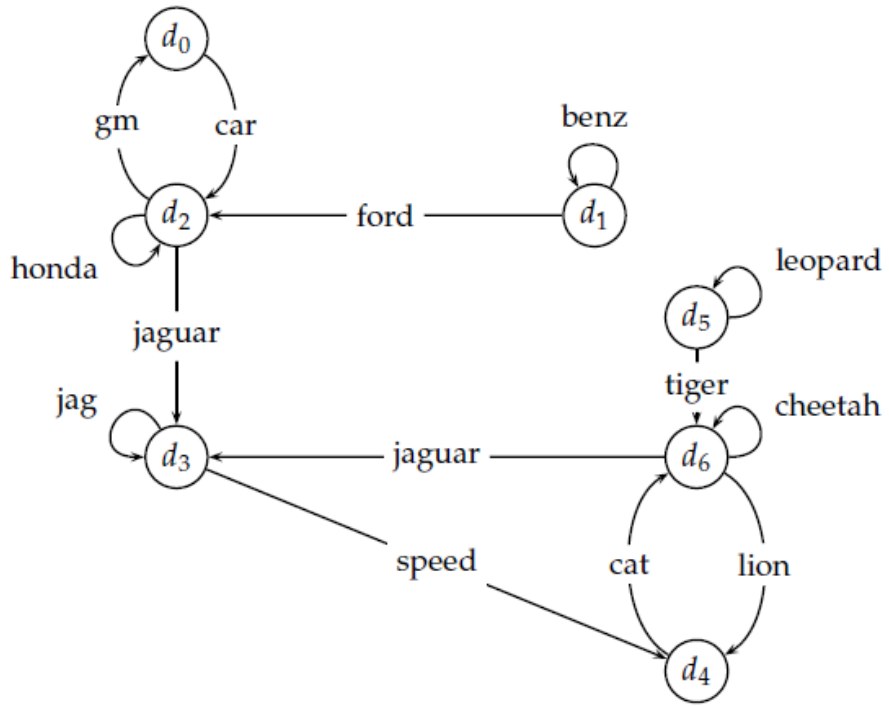
Recall A^T is the transpose of A.

Substituting, $\mathbf{h}=\mathbf{AA}^T\mathbf{h}$ and $\mathbf{a}=\mathbf{A}^T\mathbf{Aa}$.

Thus, **h** is an eigenvector of $\mathbf{AA}^T$ and **a** is an eigenvector of $\mathbf{A}^T\mathbf{A}$.

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

Guaranteed to converge.



► **Figure 21.4** A small web graph. Arcs are annotated with the word that occurs in the anchor text of the corresponding link.

Example 21.2: Assuming the query jaguar and double-weighting of links whose anchors contain the query word, the matrix A for Figure 21.4 is as follows:

$$\begin{array}{ccccccc} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 \end{array}$$

The hub and authority vectors are:

$$\vec{h} = (0.03 \quad 0.04 \quad 0.33 \quad 0.18 \quad 0.04 \quad 0.04 \quad 0.35)$$

$$\vec{a} = (0.10 \quad 0.01 \quad 0.12 \quad 0.47 \quad 0.16 \quad 0.01 \quad 0.13)$$

Here, q_3 is the main authority – two hubs (q_2 and q_6) are pointing to it via highly weighted jaguar links.