Link Analysis and Web Search

Part III: Dynamics — Episode 10

Baochun Li
Department of Electrical and Computer Engineering
University of Toronto

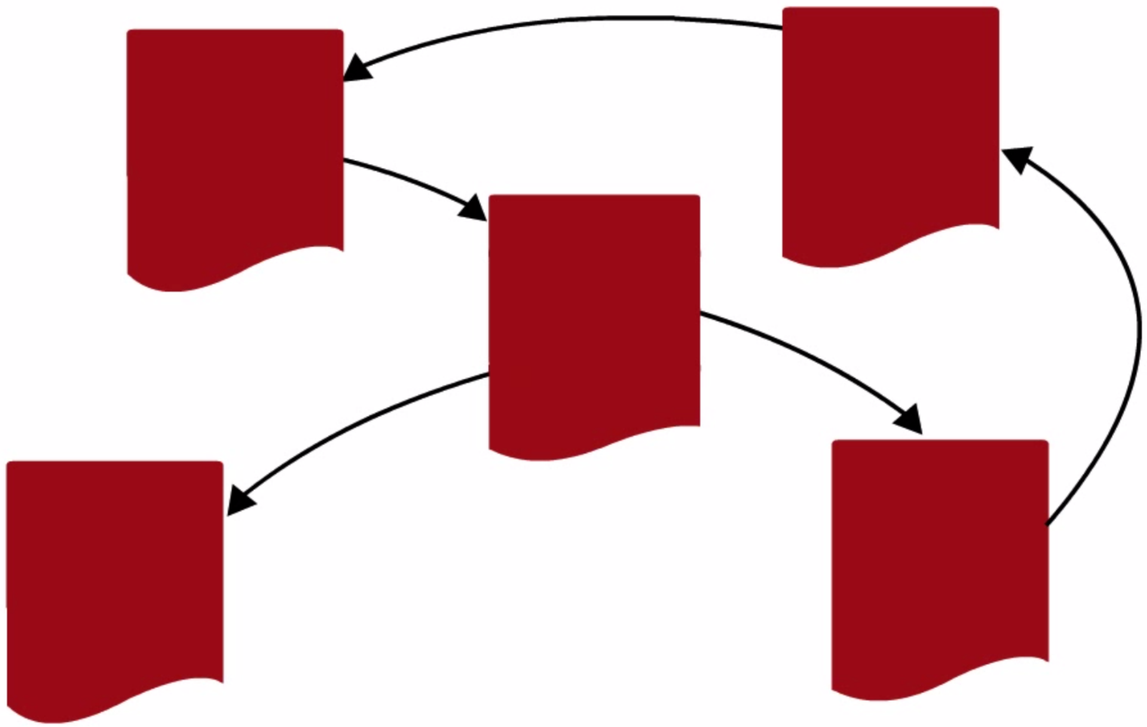
Required reading: “Networks, Crowds, and Markets,” Chapter 13.1 — 13.4, 14.1 — 14.3, 14.6

Information networks and the Web

- Logical relationships among pieces of information
  - Best example: the Web
- 1991: Tim Berners-Lee at CERN (Switzerland) created the Web
  - provided an easy way to make documents — web pages — for the world to see
  - view these pages using browsers
  - it is based on the idea of connecting these pages using links

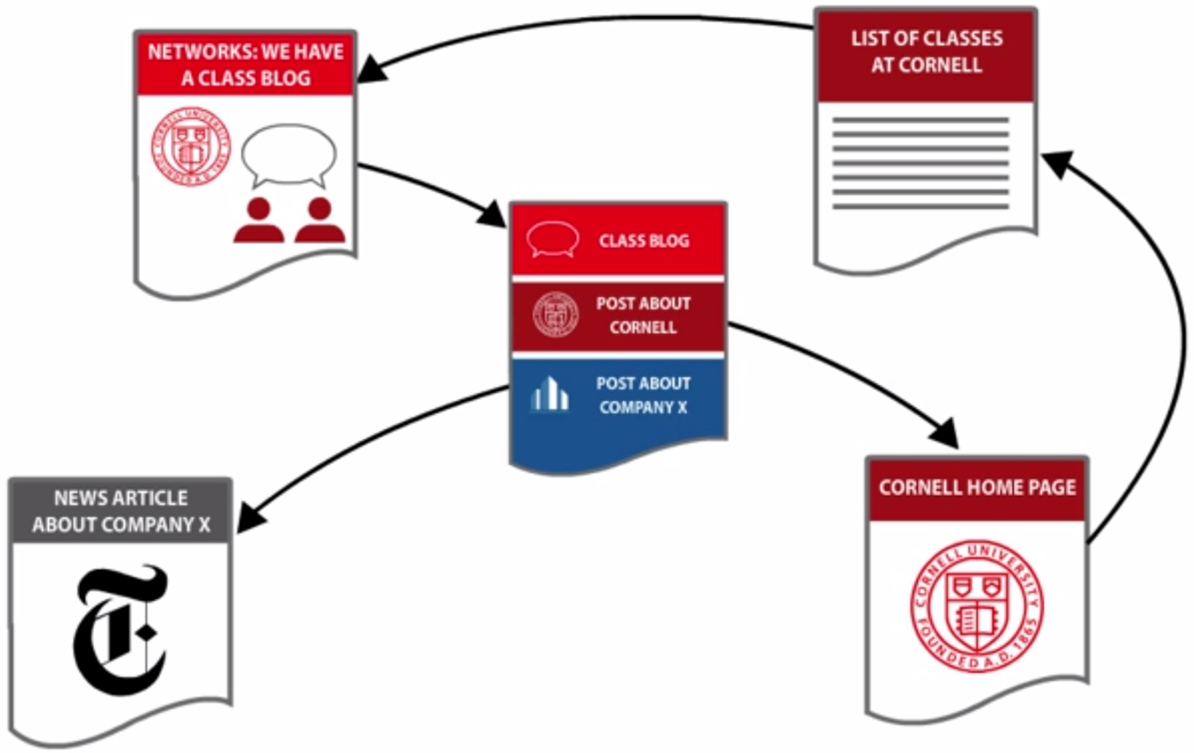
The idea of links is both inspired and non-obvious

There are many ways of organizing information: classification (library), series of folders (files), or just alphabetically (phone book)



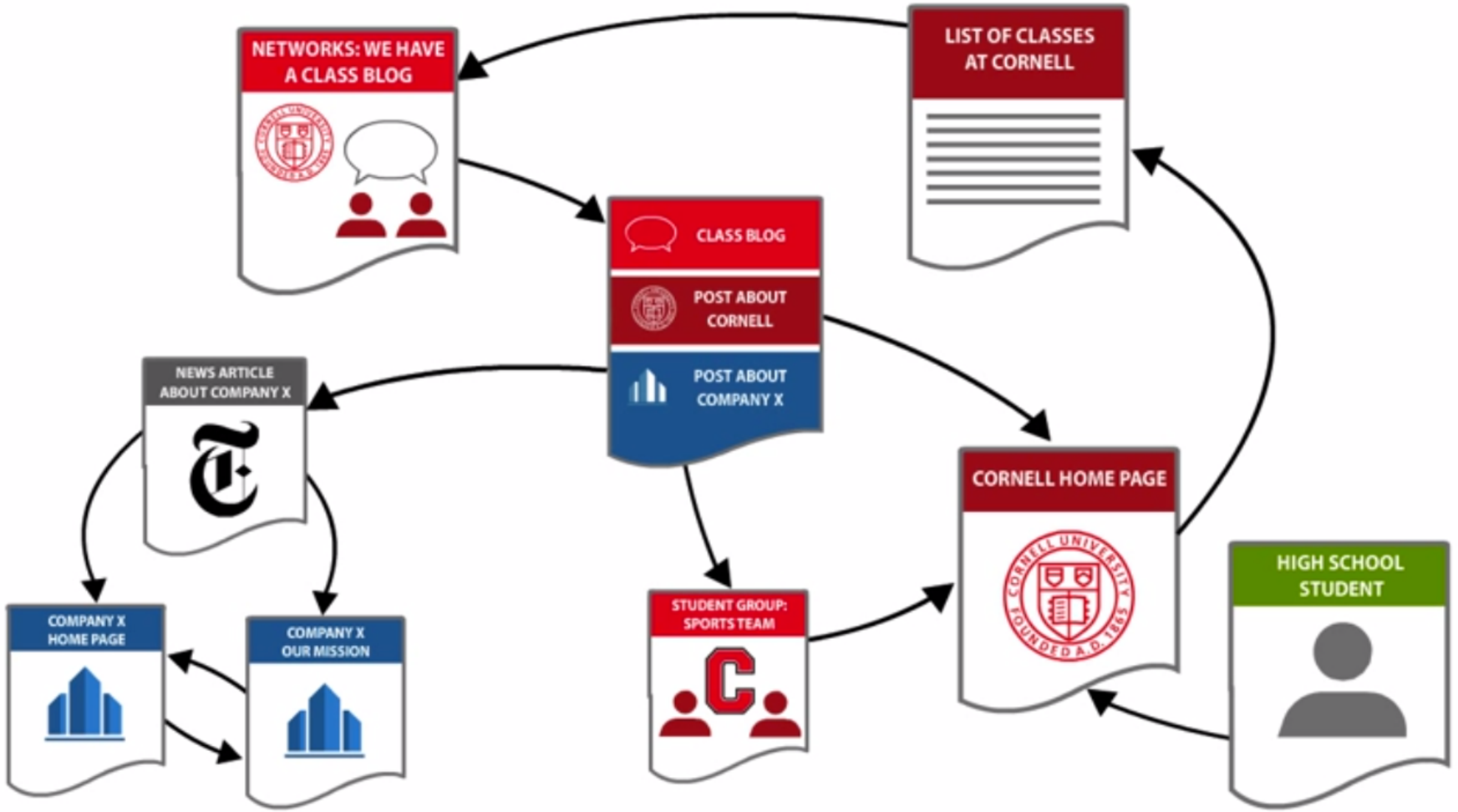
Modeling the web as a directed graph

Objective: create a "map" of the web
But how?

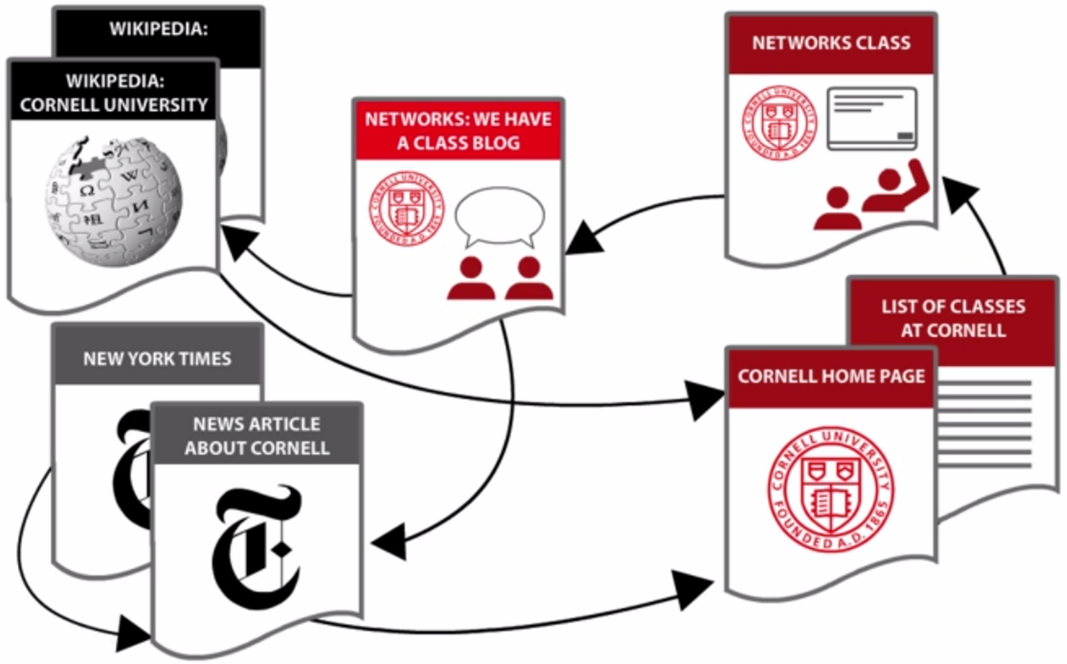


Strongly Connected Components

Strongly connected component: a subset of nodes such that (1) every node in the subset has a path to every other; and (2) the subset is not part of some larger set in which every node can reach every other.

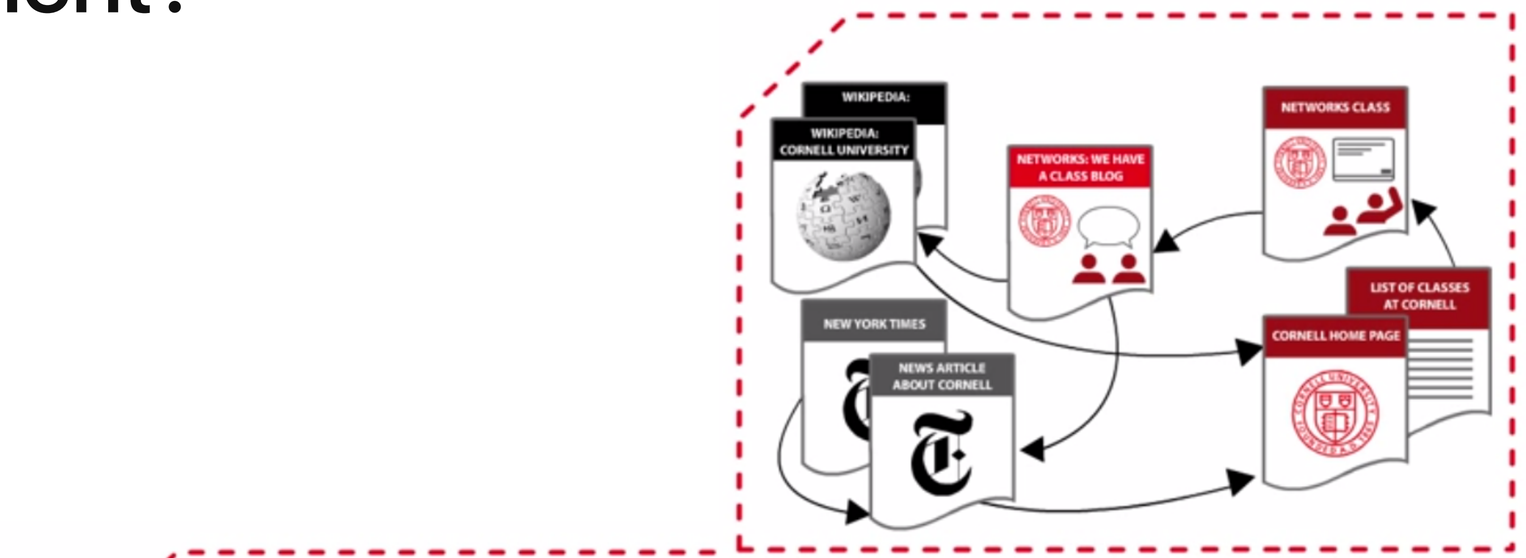

NETWORKS: WE HAVE A CLASS BLOG
LIST OF CLASSES AT CORNELL
CLASS BLOG
POST ABOUT CORNELL
POST ABOUT COMPANY X
NEWS ARTICLE ABOUT COMPANY X
CORNELL HOME PAGE
HIGH SCHOOL STUDENT
COMPANY X HOME PAGE
COMPANY X OUR MISSION
STUDENT GROUP: SPORTS TEAM

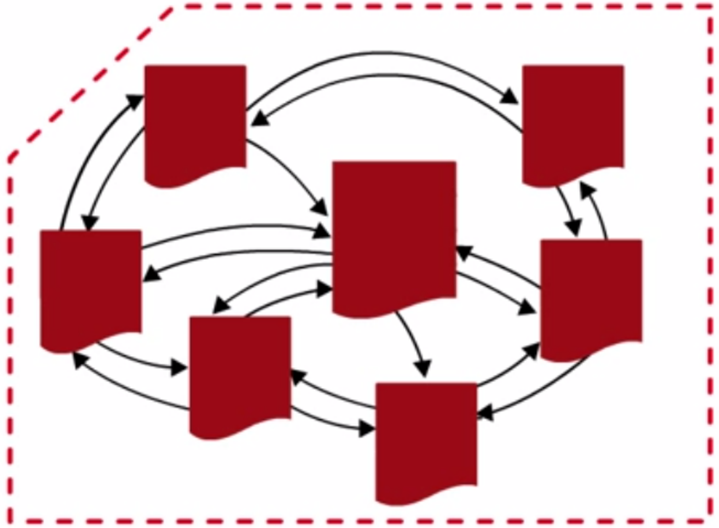
Now we can build a global map of the Web, using strongly connected components (Broder *et al.* [1999])



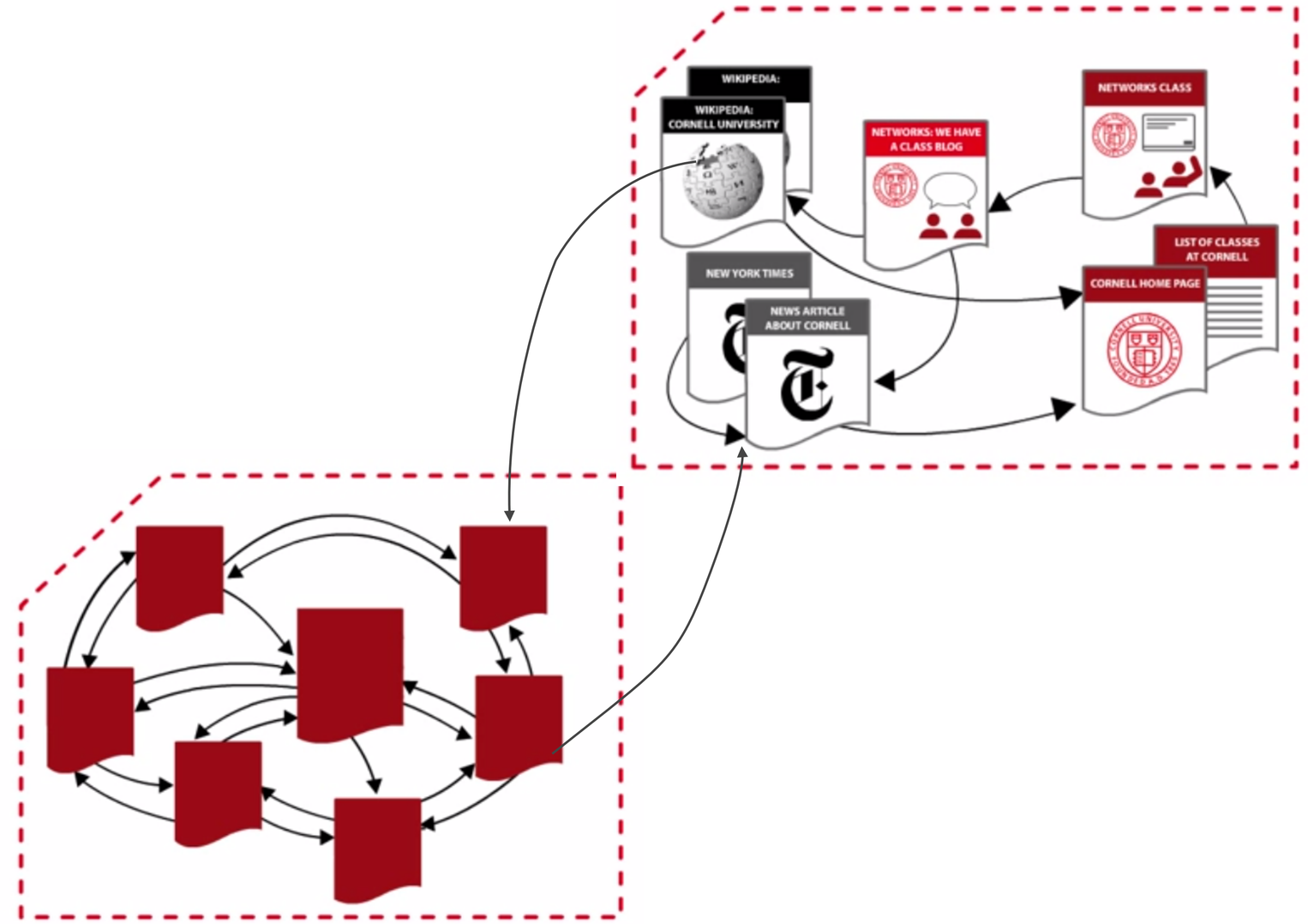
A giant strongly connected component

Can there be a second giant strongly connected component?

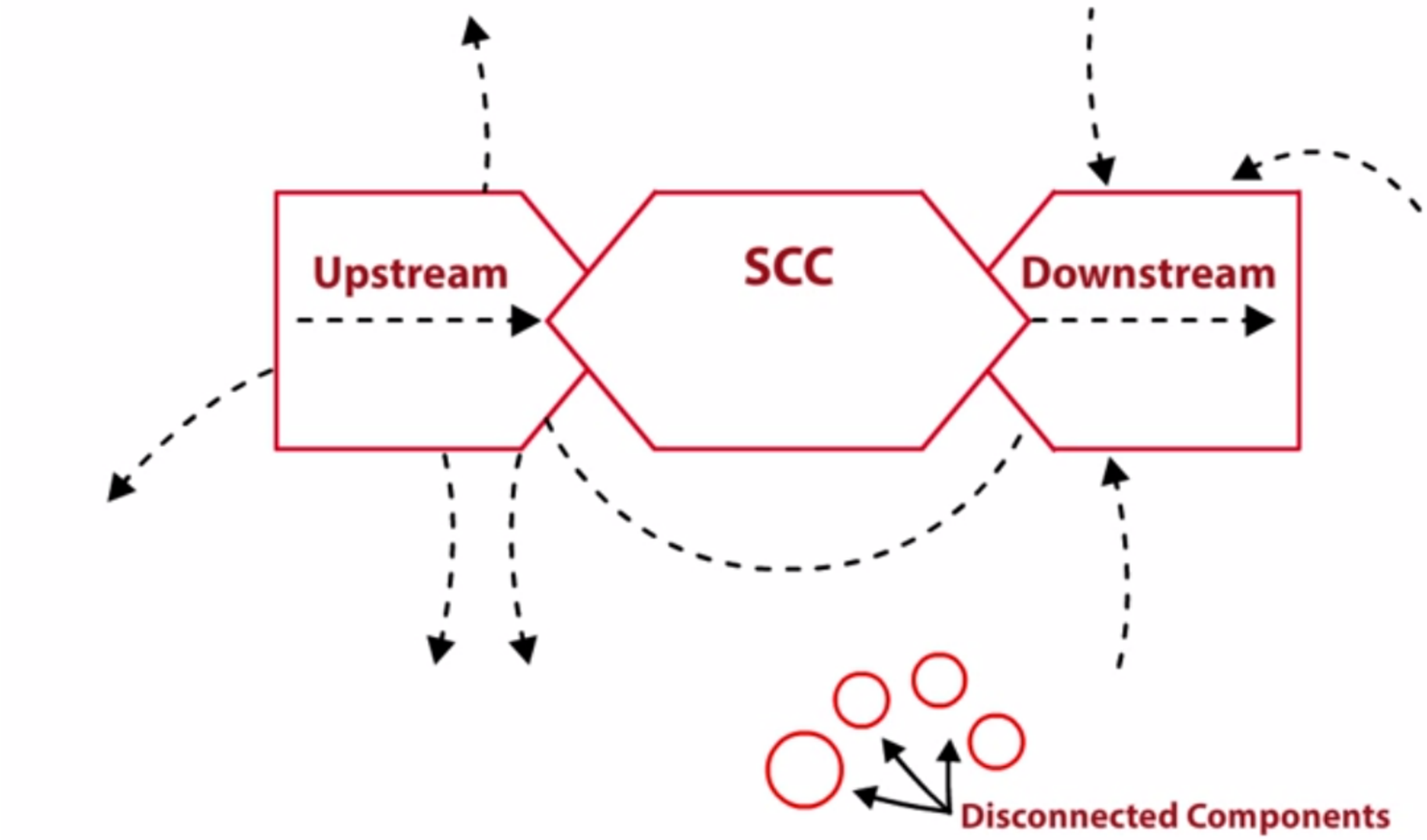


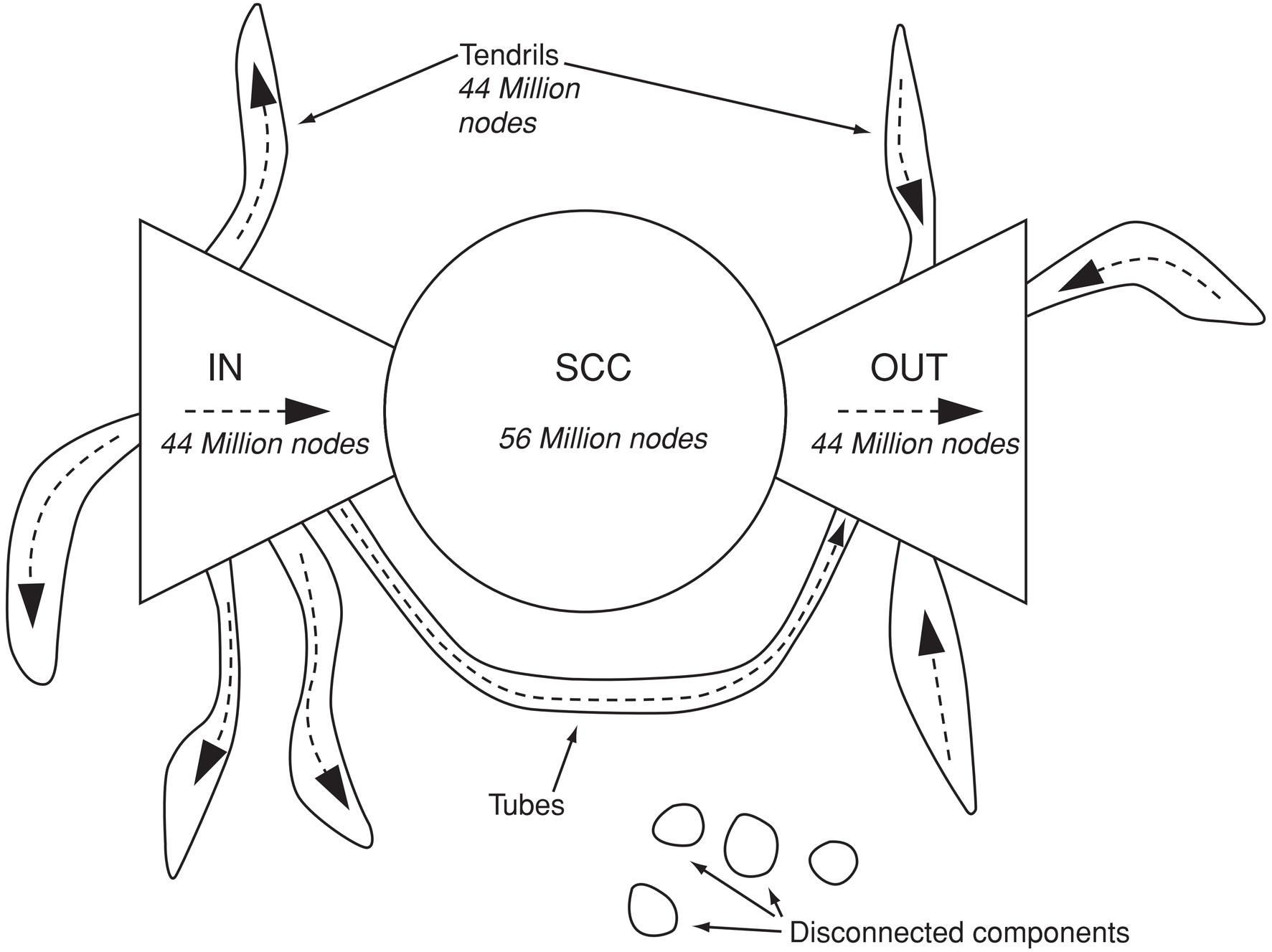


Not really — it’s too fragile



The bow-tie structure of the web




Tendrils
44 Million
nodes
IN
44 Million nodes
SCC
56 Million nodes
OUT
44 Million nodes
Tubes
Disconnected components

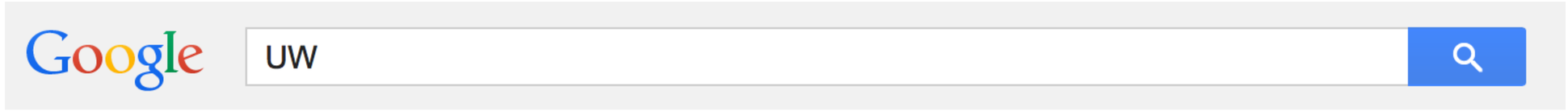
How do we find web pages using search?

- Up through the 1980s, very few people cared about information retrieval (search)
  - librarians
  - patent attorneys
- They are trained to formulate effective queries, and the documents they were searching for were written by professionals
  - research articles
  - court documents
  - U.S. patents

The Web is entirely different

- Both search users and web page authors are amateurs
- Scale is really large
- Highly dynamic nature of the content to be searched
  - Some of the authors may even optimize their content for a search engine
    - An industry called “Search Engine Optimization”
    - Millions of dollars on the line

When I search for a key phrase, what do I need?



University of Washington

University of Waterloo

University of Wisconsin

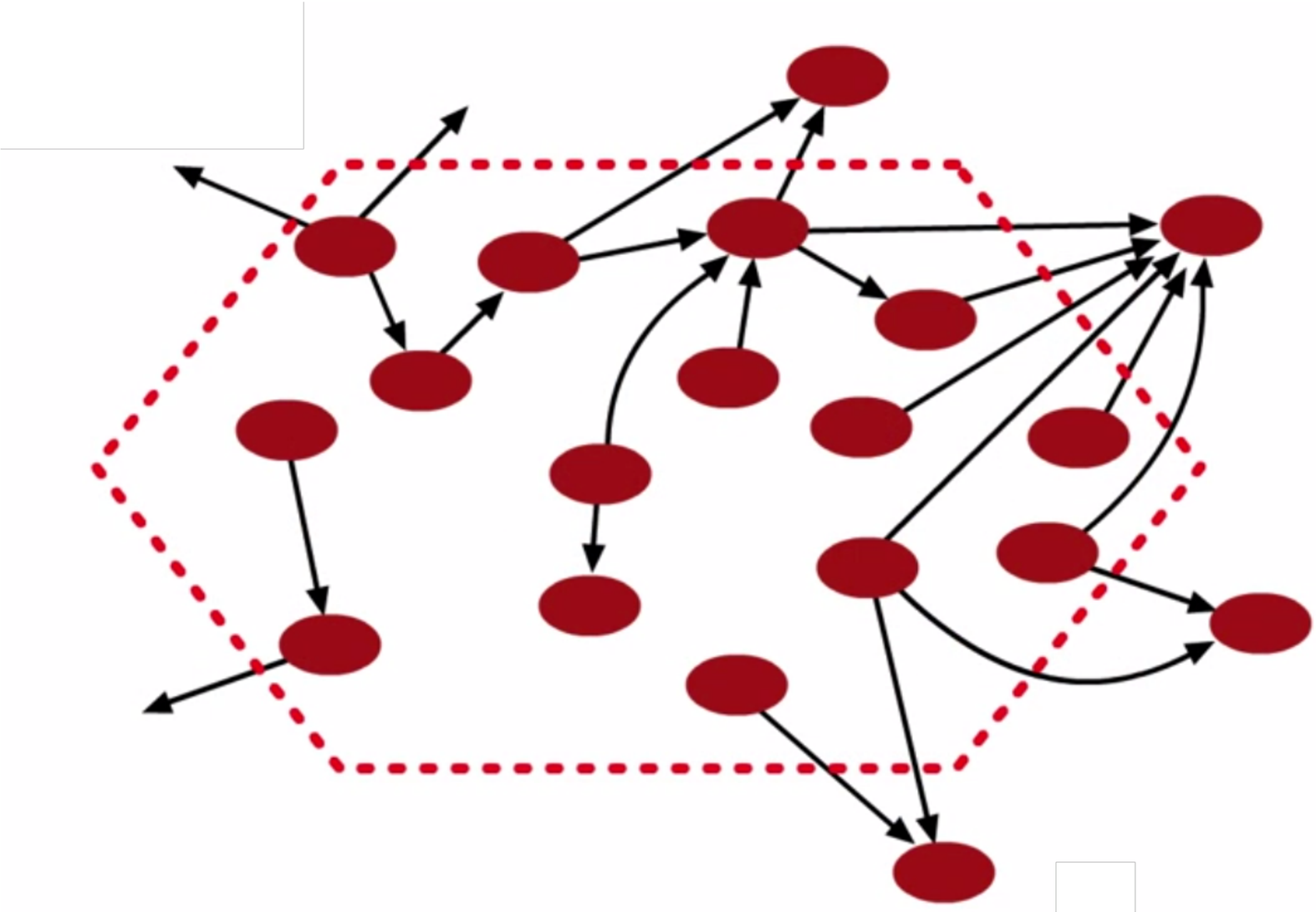
University of Windsor

University of Winnipeg

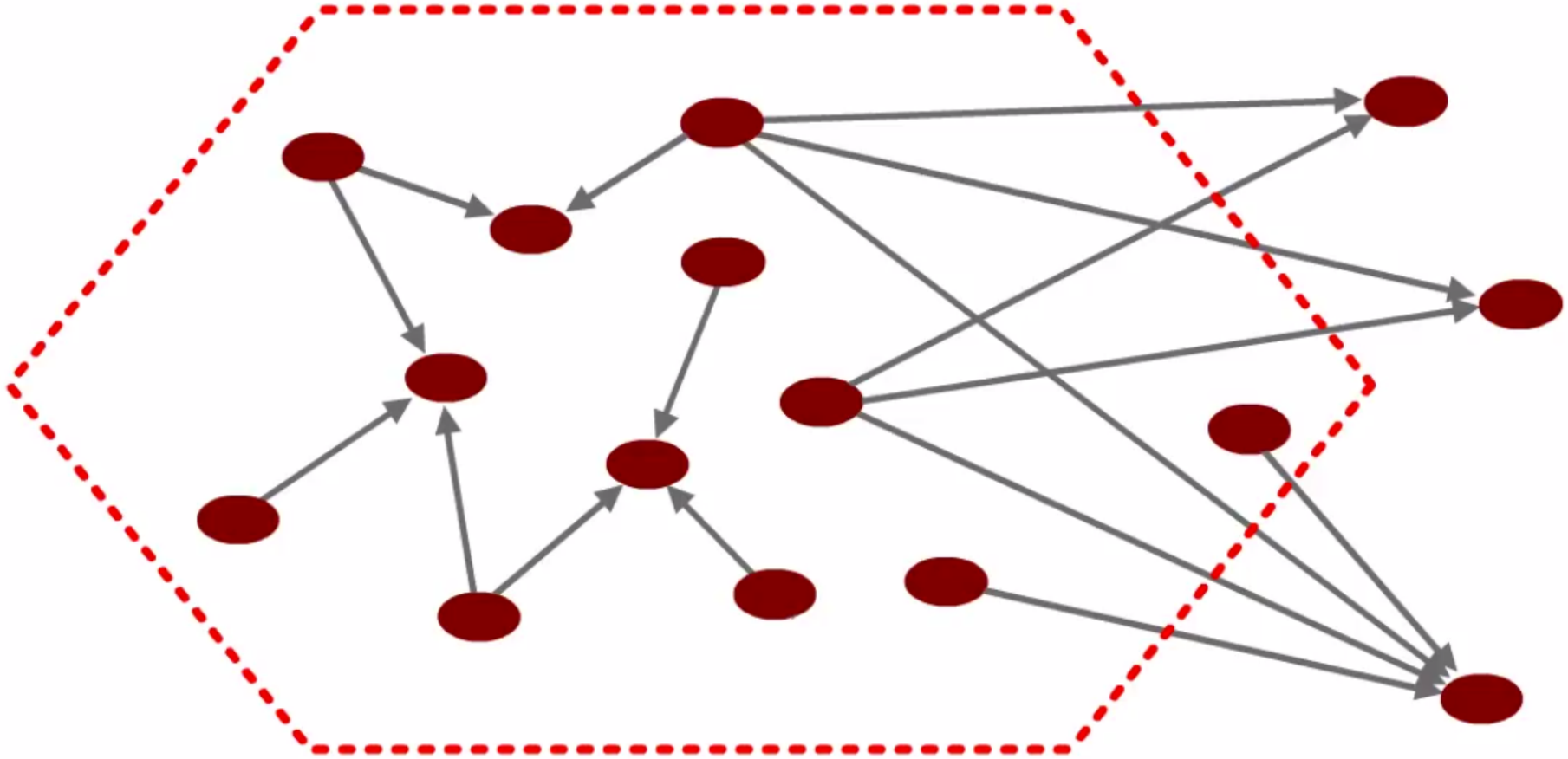
University of Wyoming

?

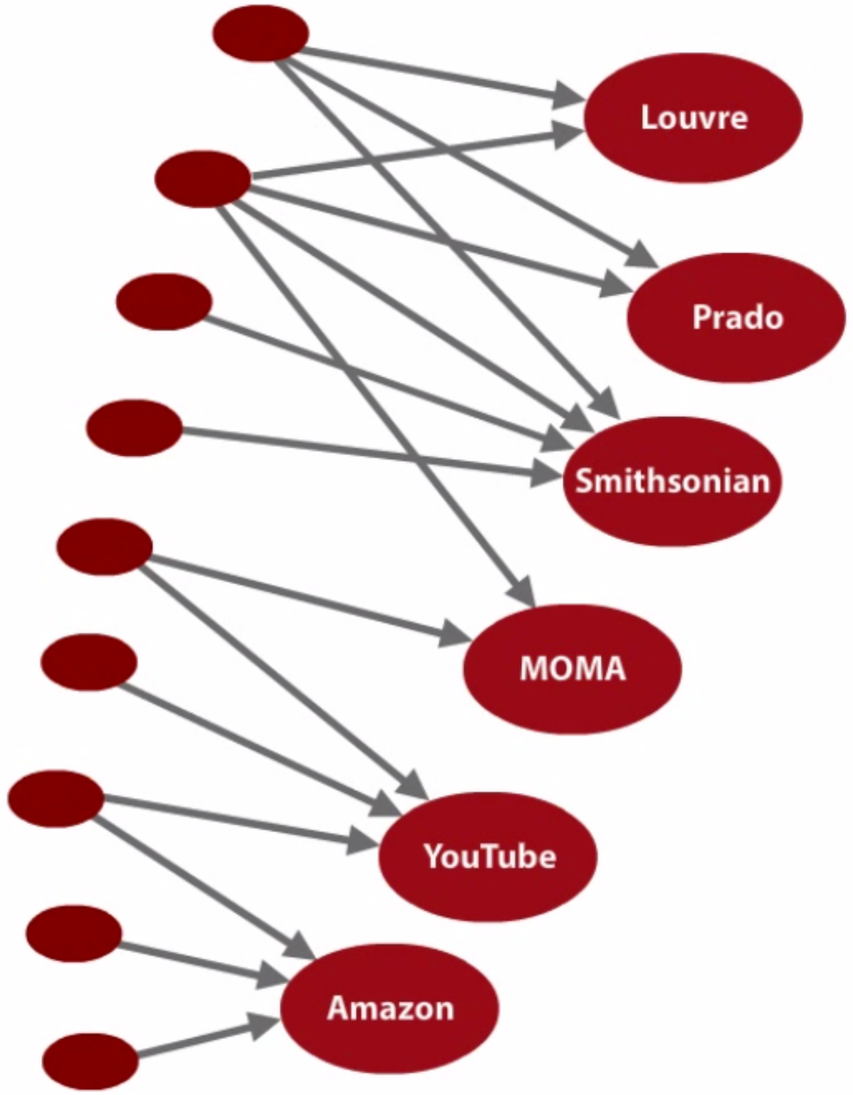
Basic idea: let the links “vote”



Using links as more than simple “votes”

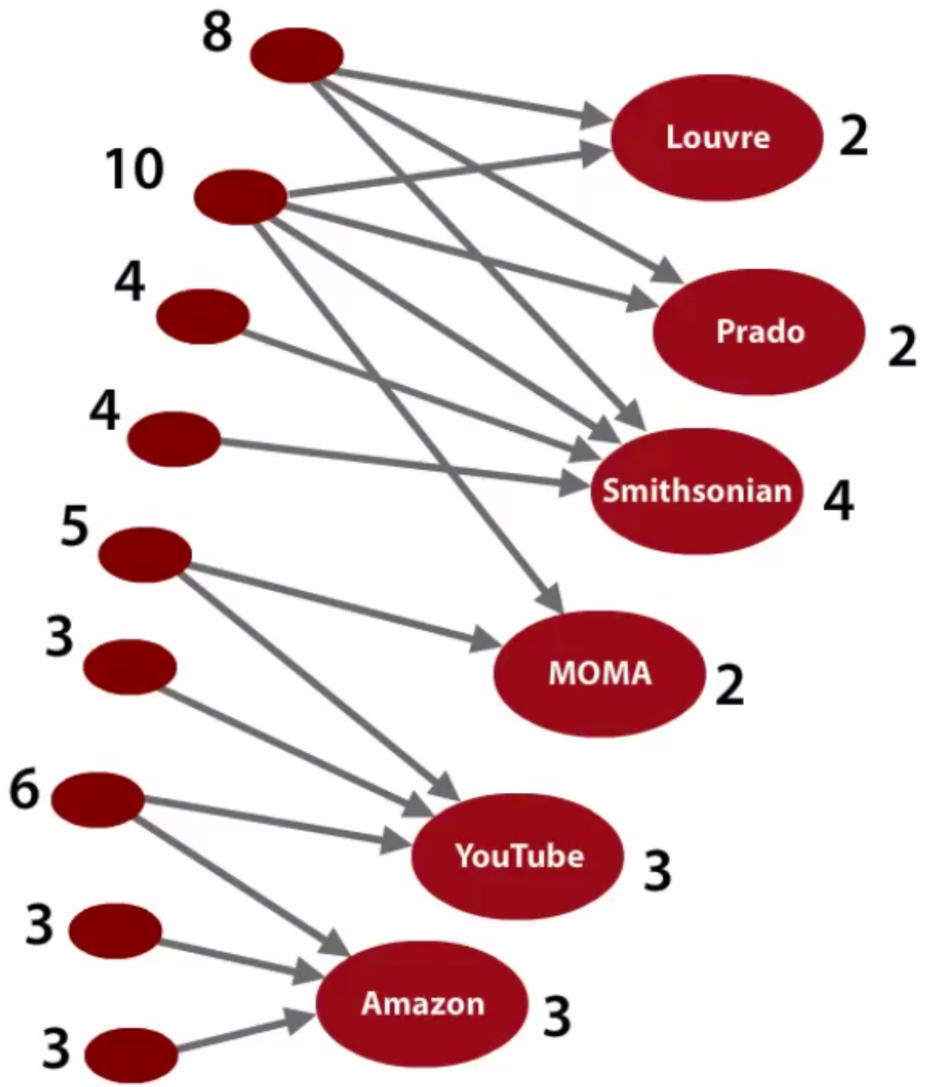


Searching for “good museum”



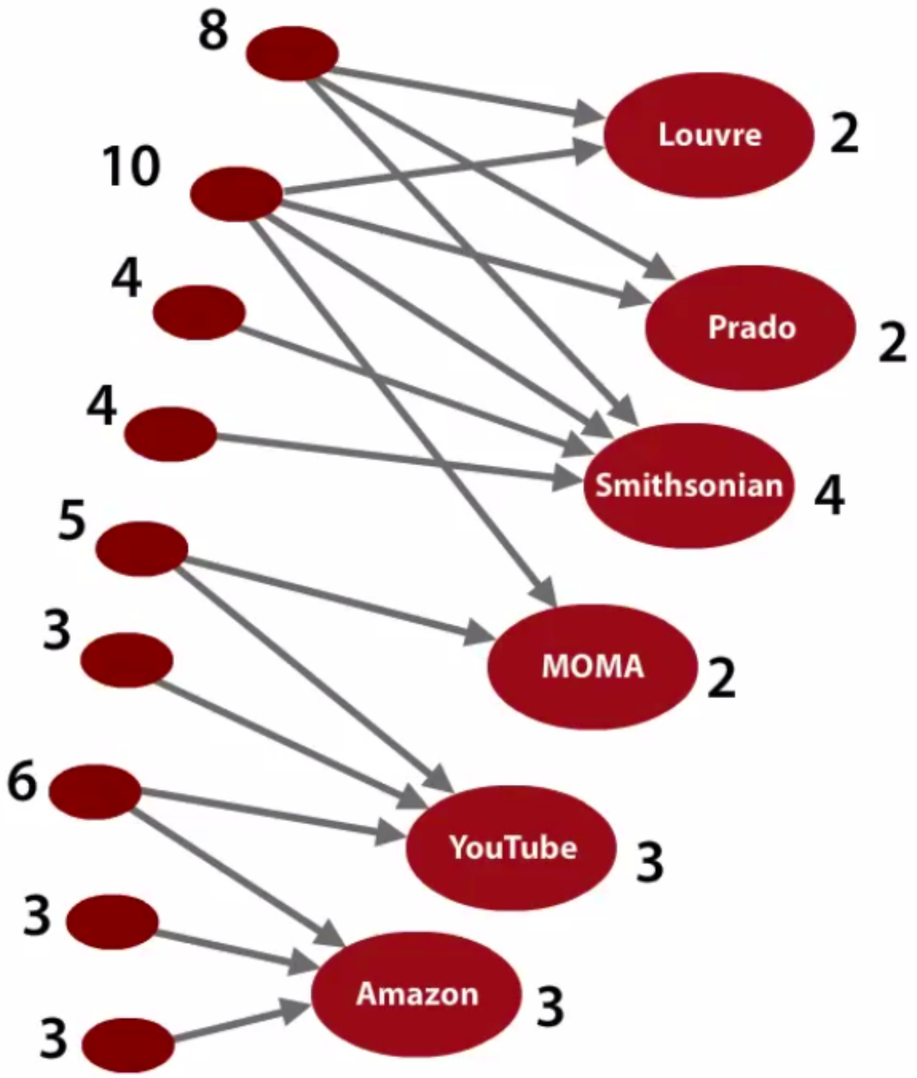
Link-based ranking with hubs and authorities

Key idea: voting again and again

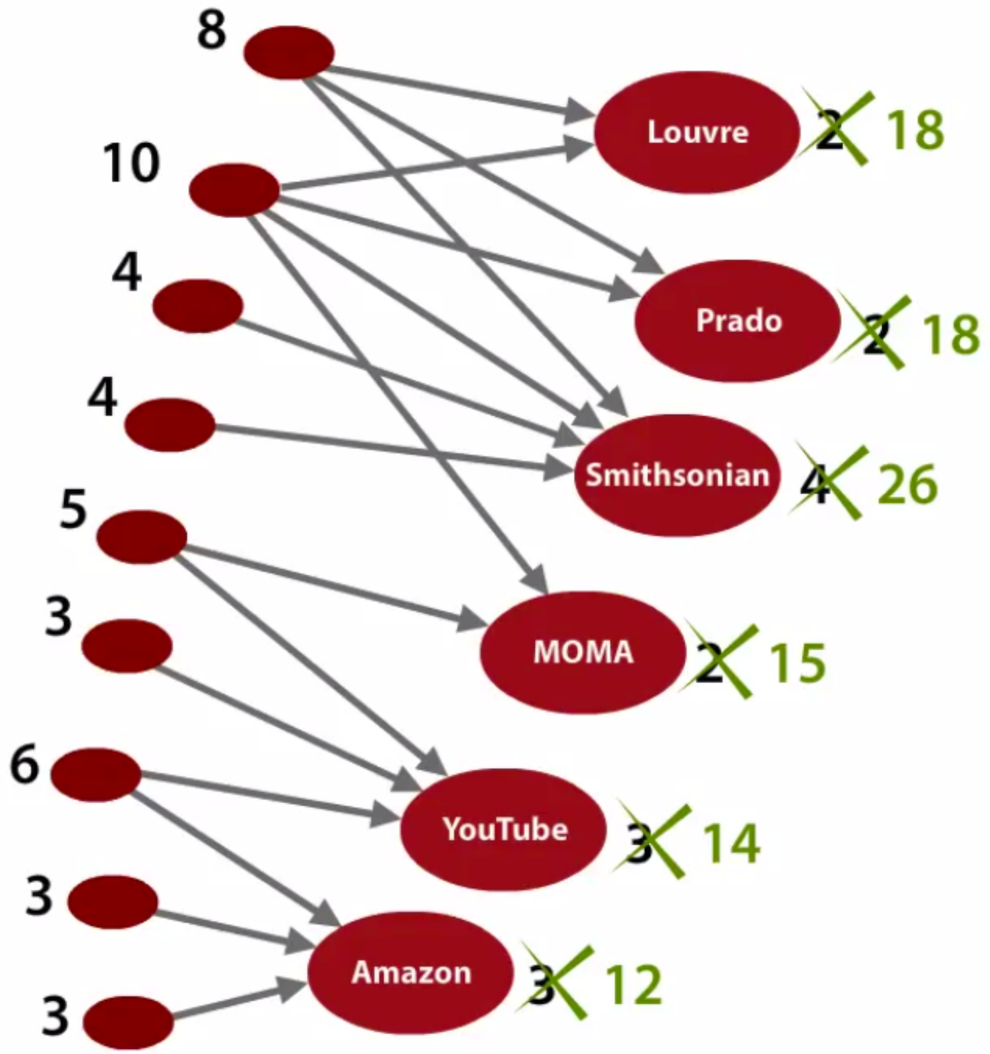


Link-based ranking with hubs and authorities

Key idea: principle of repeated improvement



But why stop here?



Let’s make it more formal

- Two kinds of quality measures for web pages
  - Authority score — Auth(p): level of endorsement
  - Hub score — Hub(p) : quality as a list

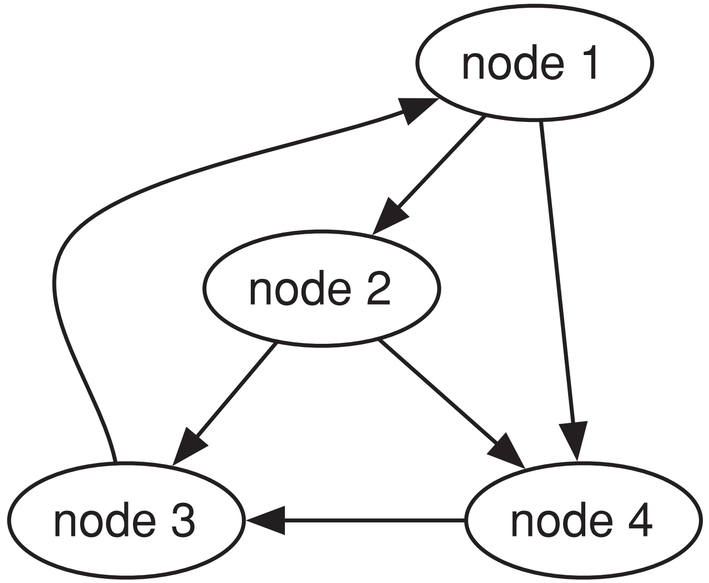
Authority update rule: Auth(p) = sum of hub scores of all pages that link to p.

Hub update rule: Hub(p) = sum of authority scores of all pages that p points to.

Divide all scores so that they add to 1.

Using adjacency matrices to represent a graph

- view a set of n pages, 1, 2, ...n, as a set of nodes in a directed graph
- n x n matrix M: M_{ij} is equal to 1 if thee is a link from node i to node j

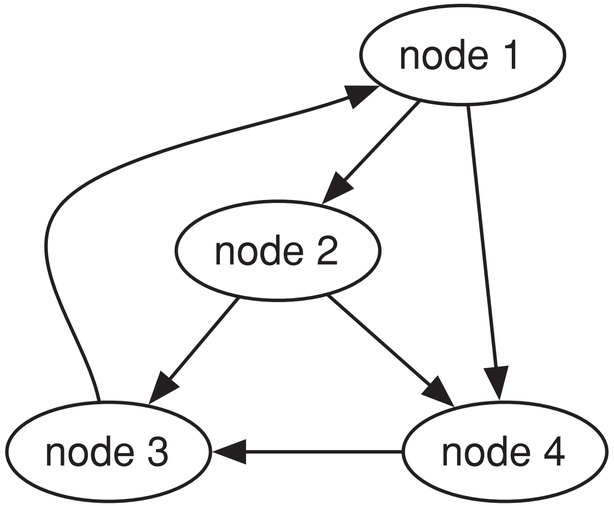


$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Hub update rule as matrix multiplication

$$h_i \leftarrow M_{i1}a_1 + M_{i2}a_2 + \cdots + M_{in}a_n$$

$$h \leftarrow Ma$$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 2 \\ 4 \end{bmatrix}$$

The authority update rule as matrix multiplication

$$a_i \leftarrow M_{1i}h_1 + M_{2i}h_2 + \cdots + M_{ni}h_n$$

$$a \leftarrow M^T h$$

Unwinding the k-step hub-authority updates

$$a^{\langle 1 \rangle} = M^T h^{\langle 0 \rangle}$$

$$h^{\langle 1 \rangle} = M a^{\langle 1 \rangle} = M M^T h^{\langle 0 \rangle}$$

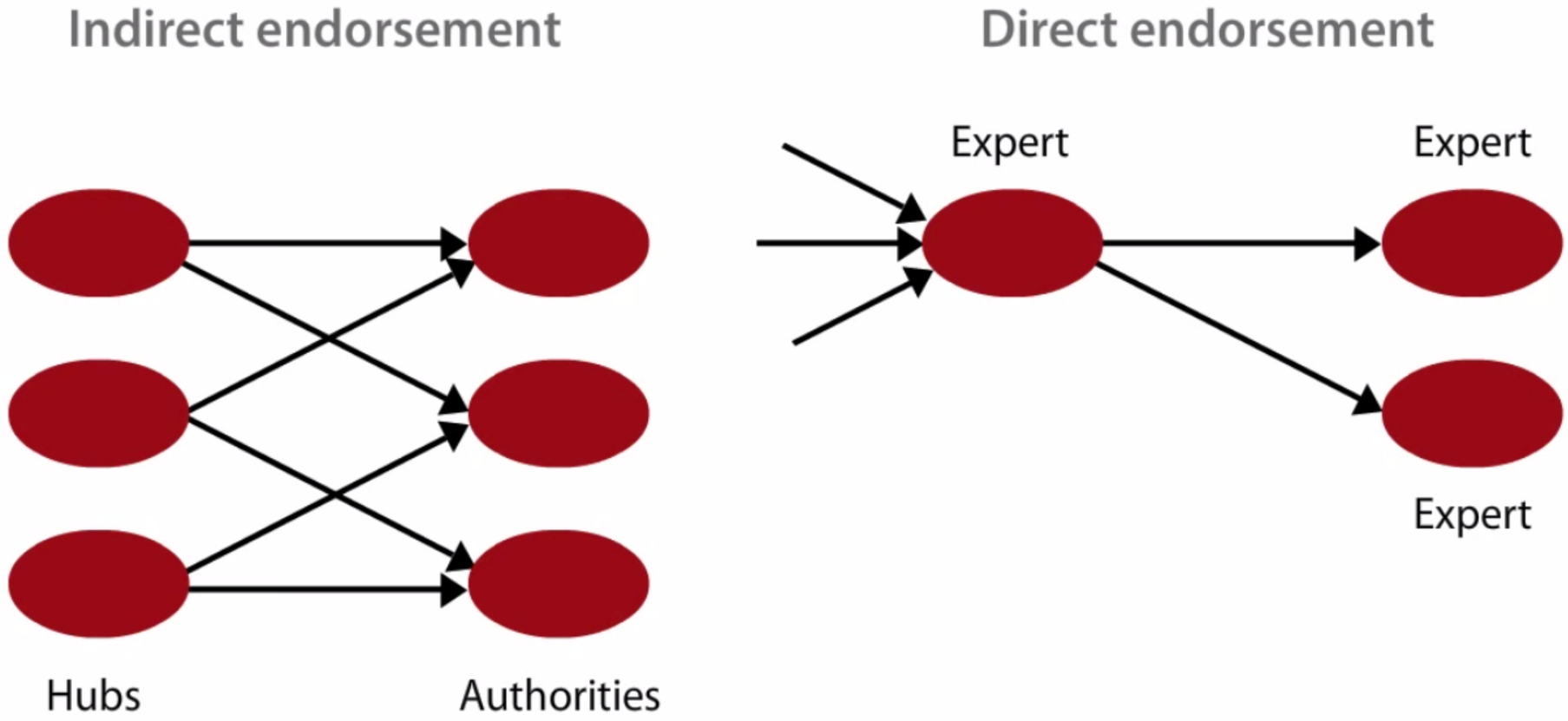
$$a^{\langle 2 \rangle} = M^T h^{\langle 1 \rangle} = M^T M M^T h^{\langle 0 \rangle}$$

$$h^{\langle 2 \rangle} = M a^{\langle 2 \rangle} = M M^T M M^T h^{\langle 0 \rangle} = (M M^T)^2 h^{\langle 0 \rangle}$$

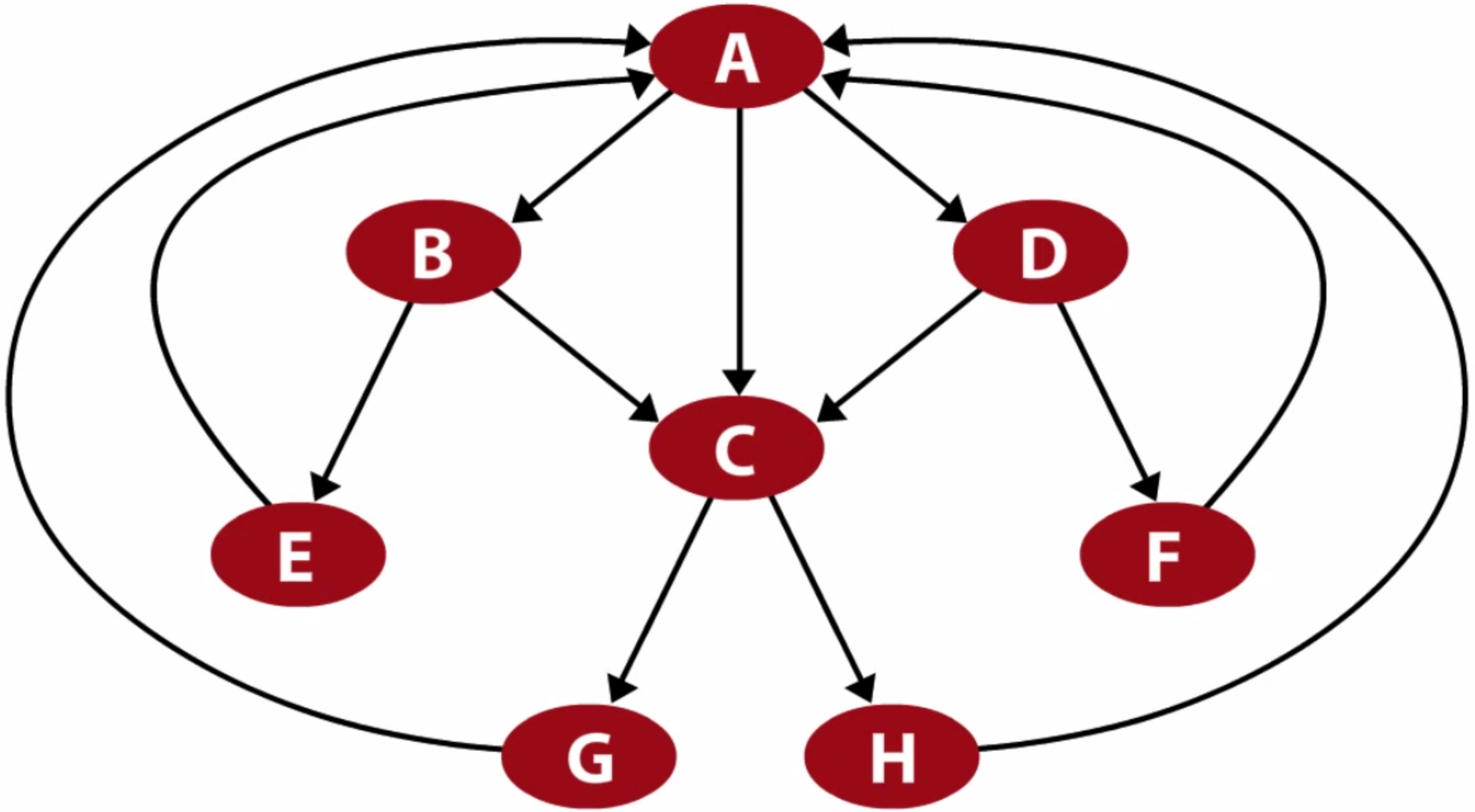
$$a^{\langle k \rangle} = (M^T M)^{k-1} M^T h^{\langle 0 \rangle}$$

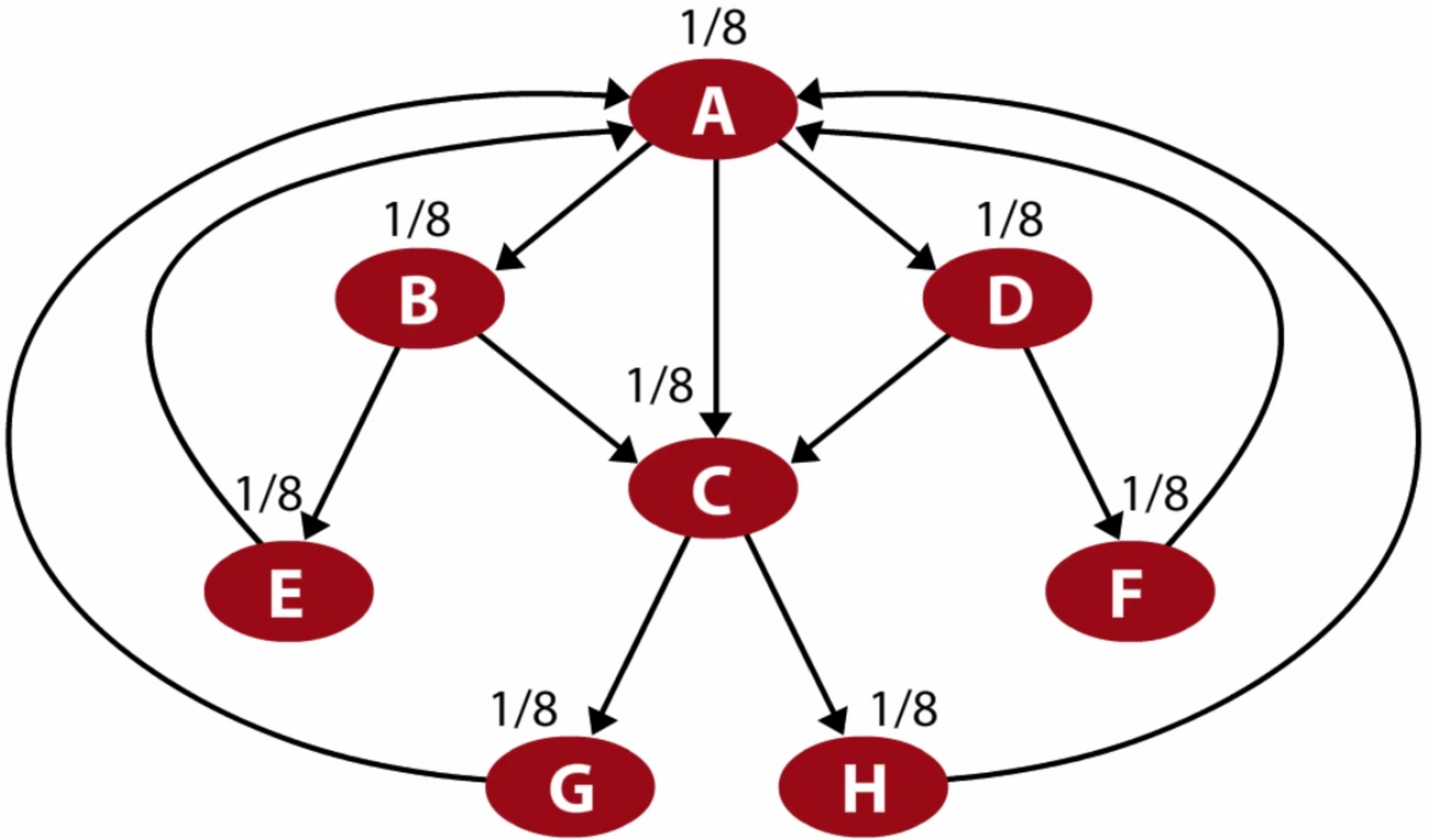
$$h^{\langle k \rangle} = (M M^T)^k h^{\langle 0 \rangle}$$

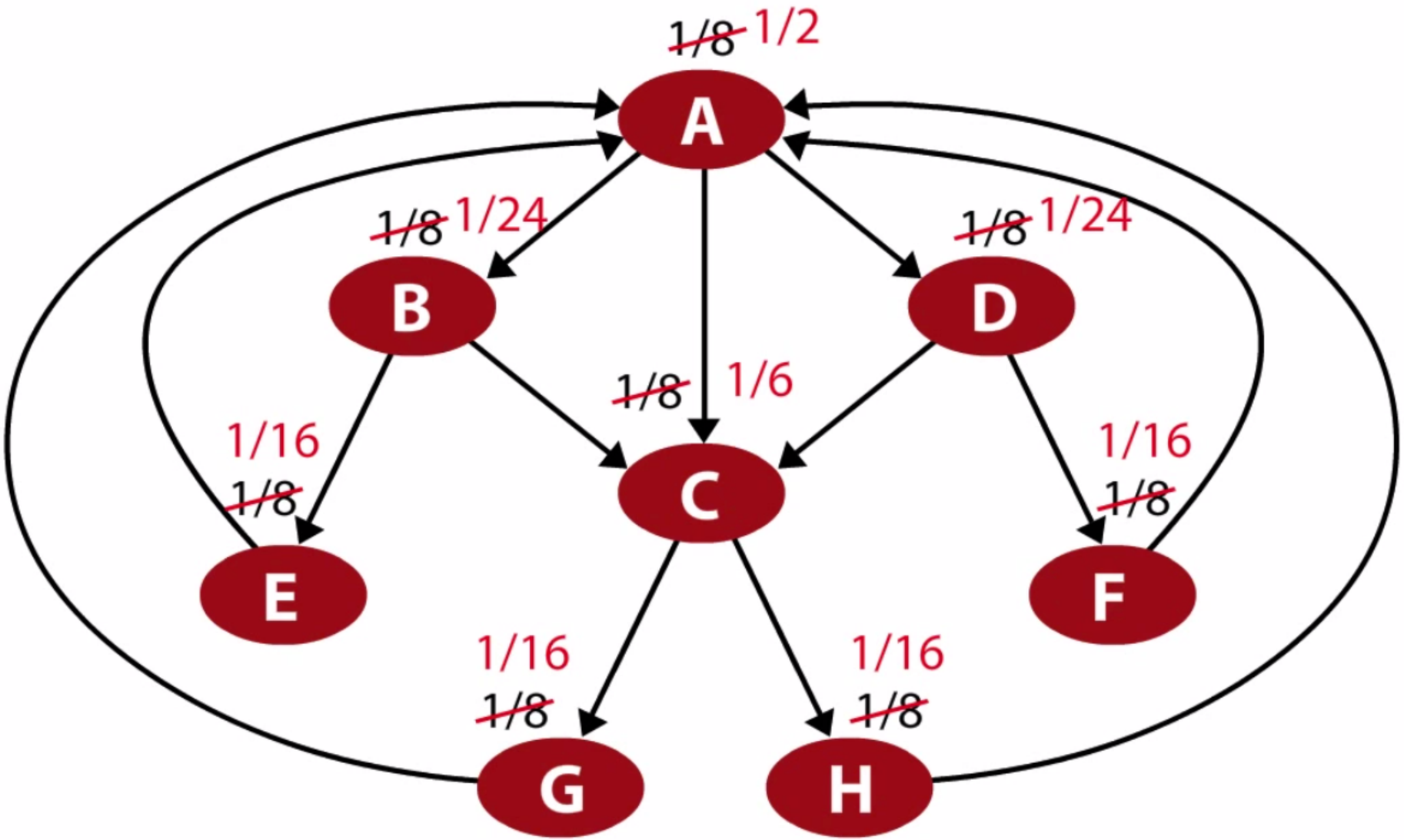
PageRank: the core of Google search



Basic PageRank update rule



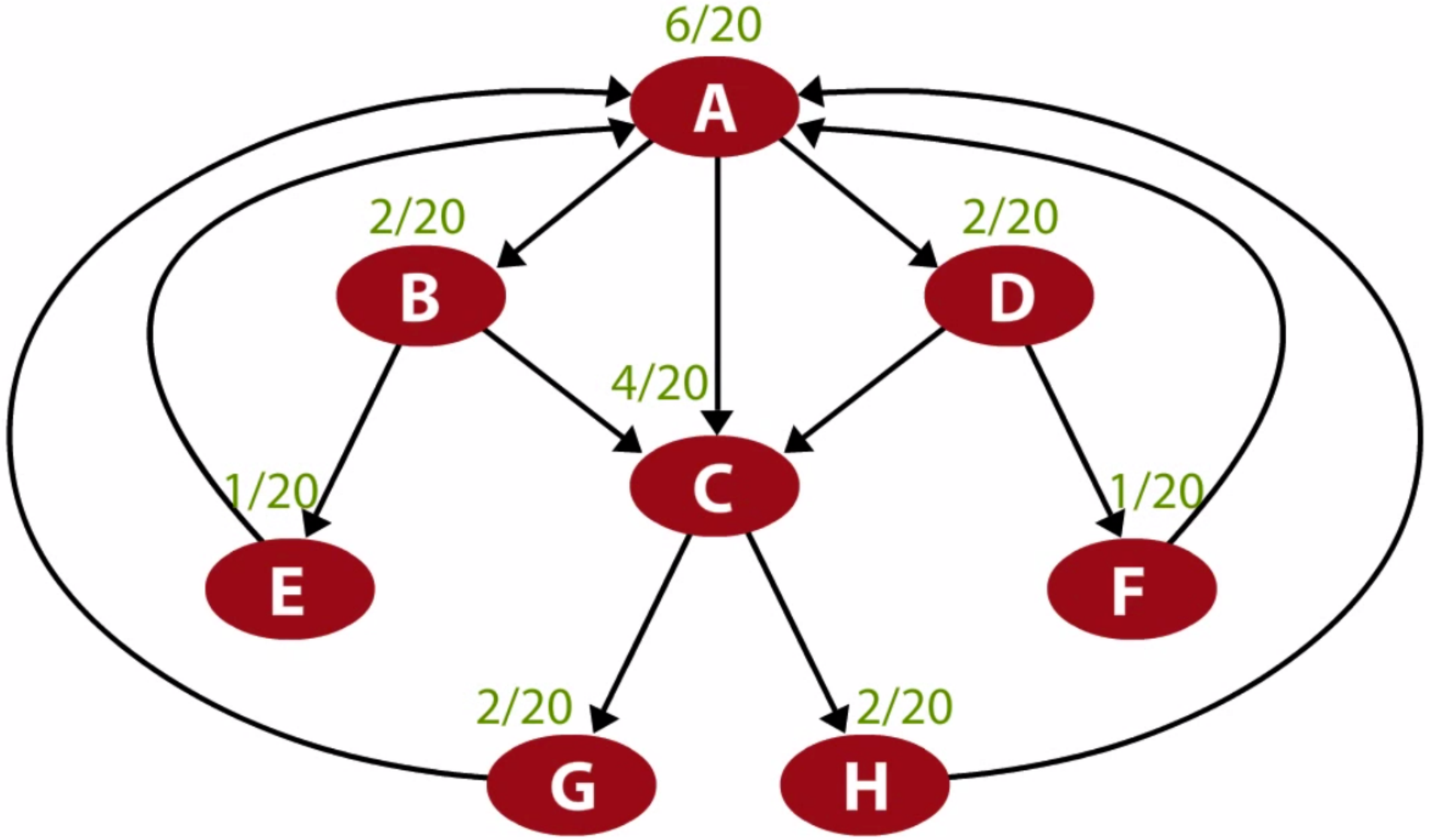

1/8
A
1/8
B
1/8
D
1/8
C
1/8
E
1/8
F
1/8
G
1/8
H


1/8
1/2
A
1/8
1/24
B
1/8
1/24
D
1/8
1/6
C
1/16
1/8
E
1/16
1/8
F
1/16
1/8
G
1/16
1/8
H

Basic PageRank update rule

- Assign each page p a PageRank value
  - start with 1/n, n is the number of pages
- Basic PageRank update rule
  - Each node divides its current PageRank into equal shares, and then pass them across outbound links
  - And then use the principle of repeated improvement
- **Fact**: If the network is strongly connected, then there are unique equilibrium PageRank values

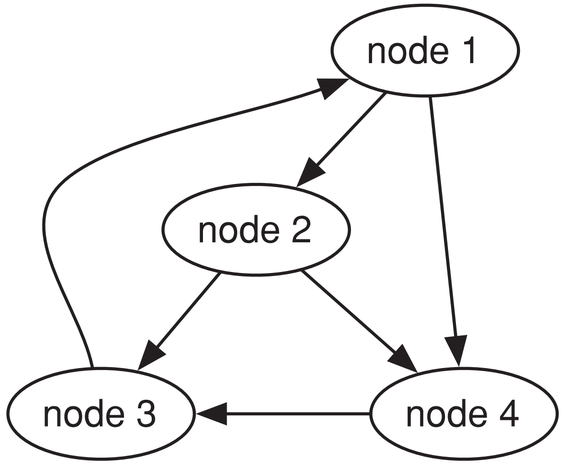
Equilibrium PageRank values



Basic PageRank updates as matrix multiplication

- Each entry in the adjacency matrix N_{ij} specifies the portion of i's PageRank that should be passed to j in one single step

  start with $1/L_i$, where L_i is the number of links out of i



$$\begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

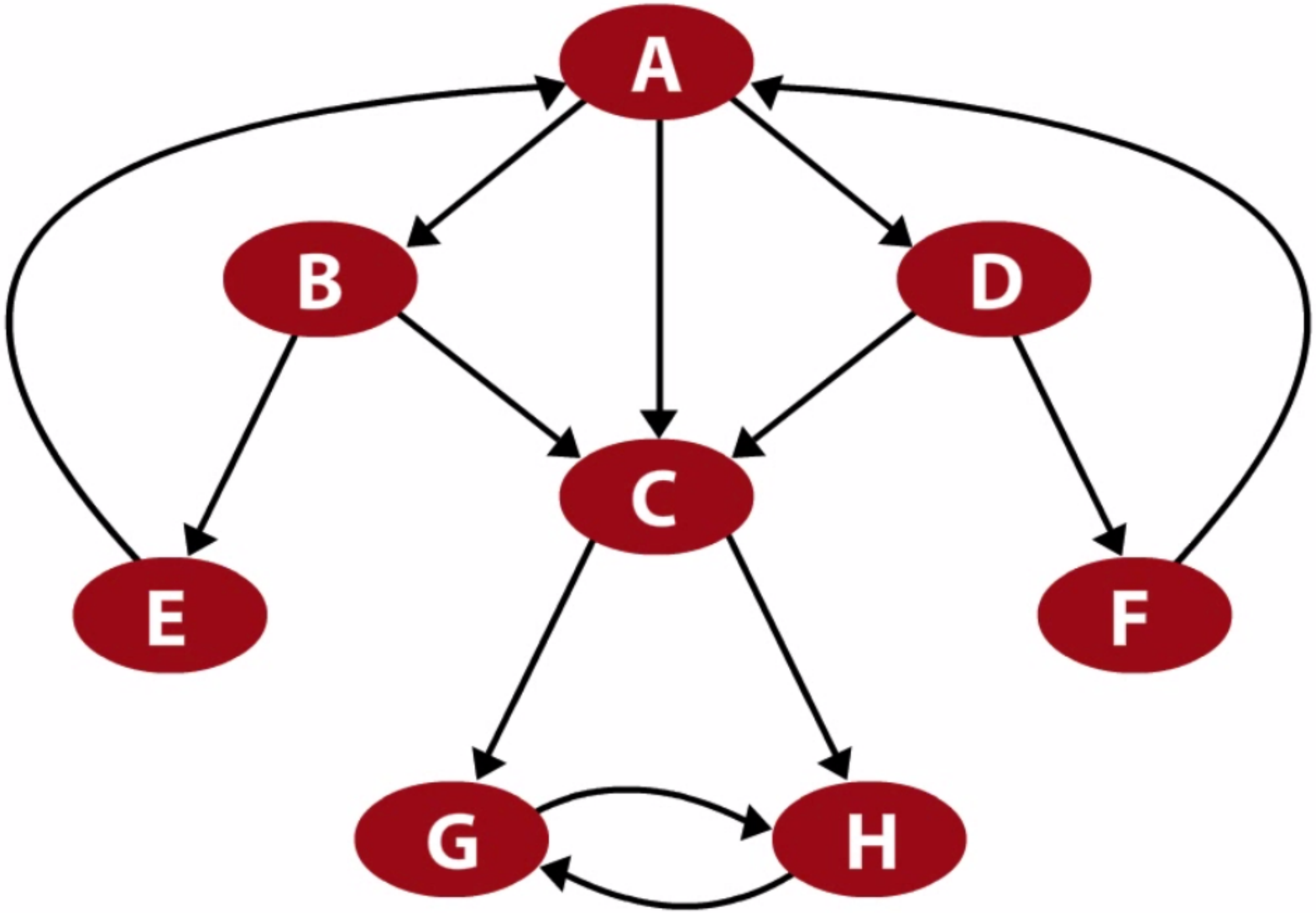
Basic PageRank updates as matrix multiplication

- Vector r: the PageRanks of all the nodes

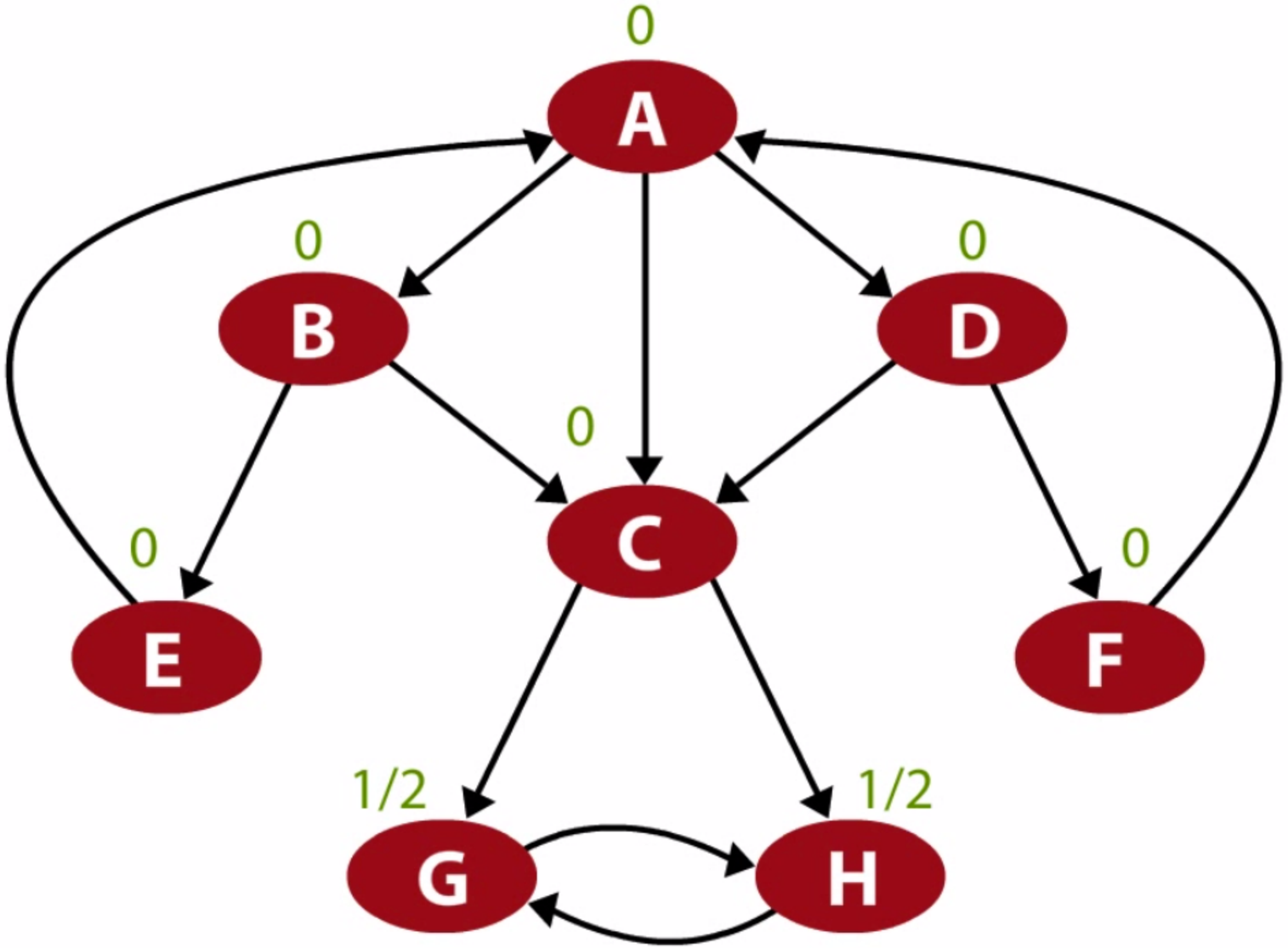
$$r_i \leftarrow N_{1i} r_1 + N_{2i} r_2 + \cdots + N_{ni} r_n.$$

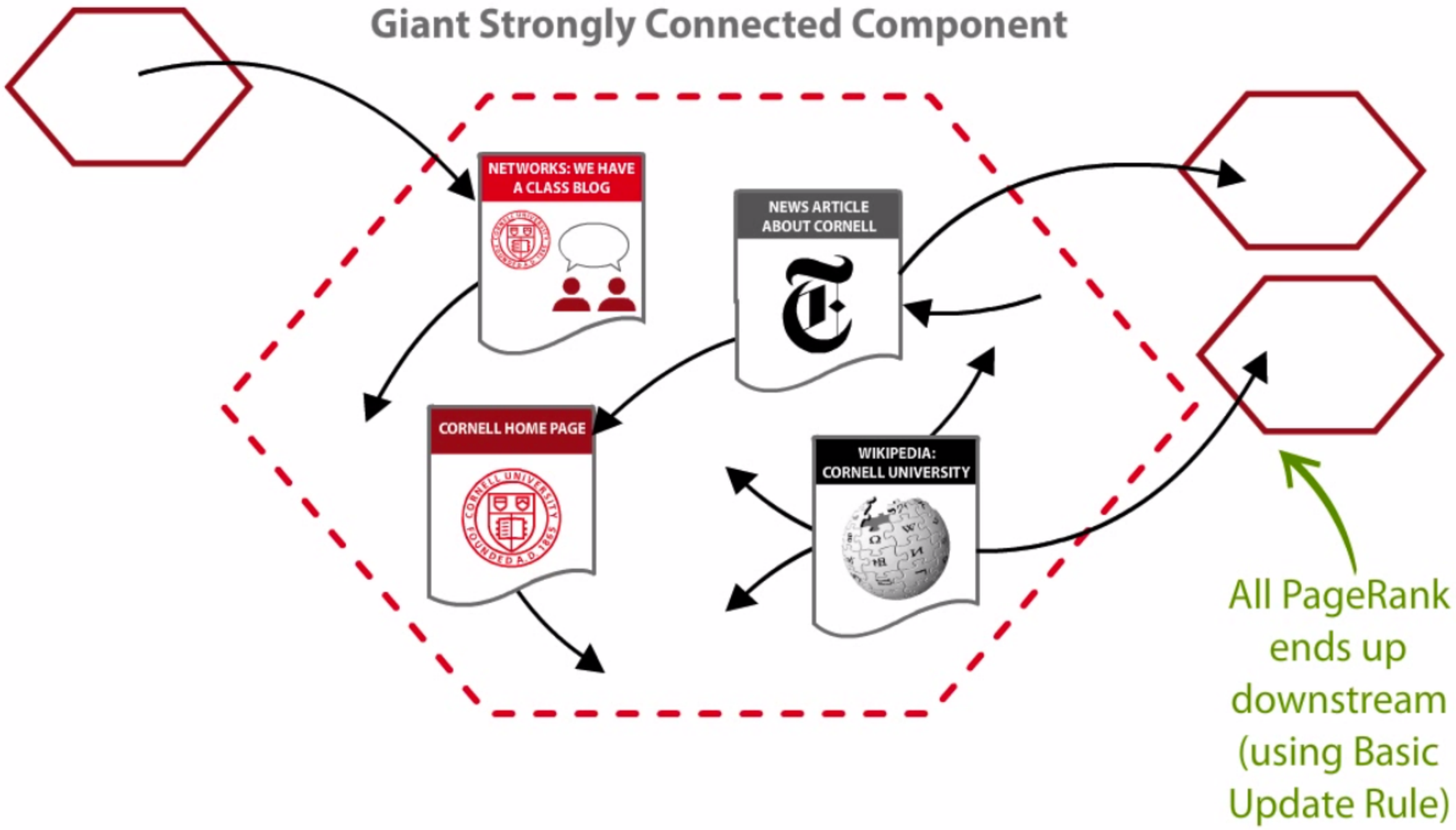
$$r \leftarrow N^T r.$$

One major problem with the basic update rule



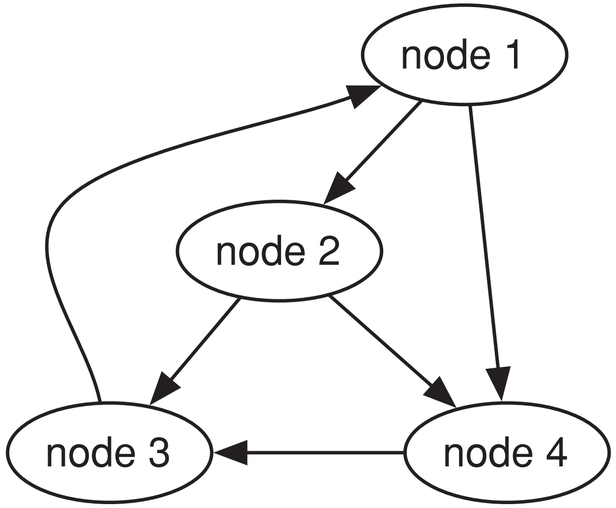
One major problem with the basic update rule




Giant Strongly Connected Component
NETWORKS: WE HAVE A CLASS BLOG
NEWS ARTICLE ABOUT CORNELL
CORNELL HOME PAGE
WIKIPEDIA: CORNELL UNIVERSITY
All PageRank ends up downstream (using Basic Update Rule)

Scaled PageRank update rule

- Fact: Repeatedly applying the scaled PageRank update rule converges to a unique equilibrium set of PageRank values for all networks.



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

$$\tilde{N}_{ij} \leftarrow sN_{ij} + (1-s)/n$$

$$r_i \leftarrow \tilde{N}_{1i}r_1 + \tilde{N}_{2i}r_2 + \cdots + \tilde{N}_{ni}r_n$$

$$r \leftarrow \tilde{N}^T r. \quad r^{\langle k \rangle} = (\tilde{N}^T)^k r^{\langle 0 \rangle}.$$

Required reading: “Networks, Crowds, and Markets,” Chapter 13.1 — 13.4, 14.1 — 14.3, 14.6

Final Examination

- December 15, 2023, BA 1180, 1:10pm - 3pm, 110 minutes
- Covers all the lectures, but not the critique papers
- Sample examination questions on the course website
  - Length: a bit longer than the sample exams
  - Format: the same — a large question divided into 2-3 small ones
- **Special Office Hour:** December 11, Monday, 2-5pm, BA 4118

Preparing for the final examination

- **Strongly recommended**: read the corresponding chapters in the textbook (in the course website)
  - Exercises at the end of each chapter will be very helpful
- If this is not feasible, understand the examples in the lecture slides
- Video recordings may help as well (though two were not available)

Sneak preview: my new graduate course next Fall

- Tentatively titled “Performant Software Systems”
- Covers how modern and high-performance software systems are built using the Rust programming language
- Still pending review and approval

Course Evaluations