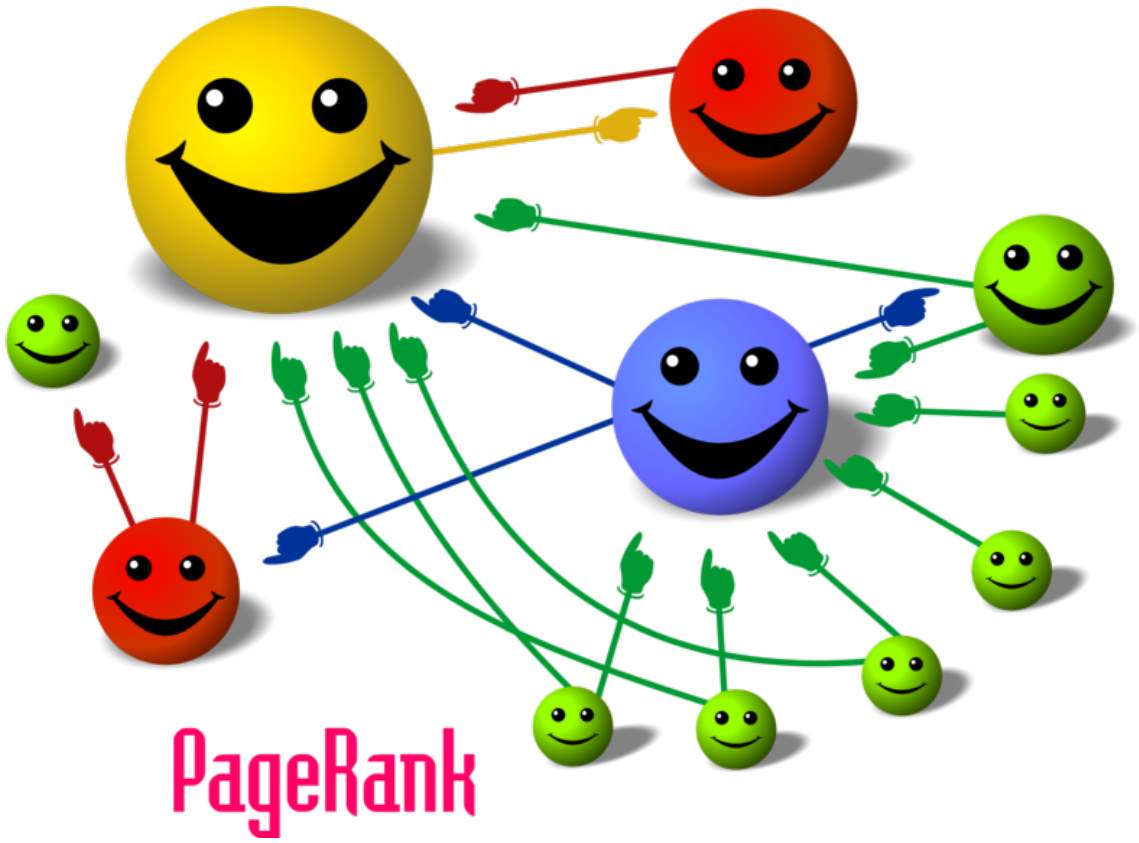
IR2: Link Analysis



Slides by Andrew DeOrio

Review

- Last time we used the words on a page to rank search results
  - Rank on document content
- Today we'll use the links between web pages to improve search results
  - Rank on document importance

Agenda

- Document importance
- Page Rank algorithm
- HITS algorithm
- Search engine optimization

Document importance

- Search for "Python threads"

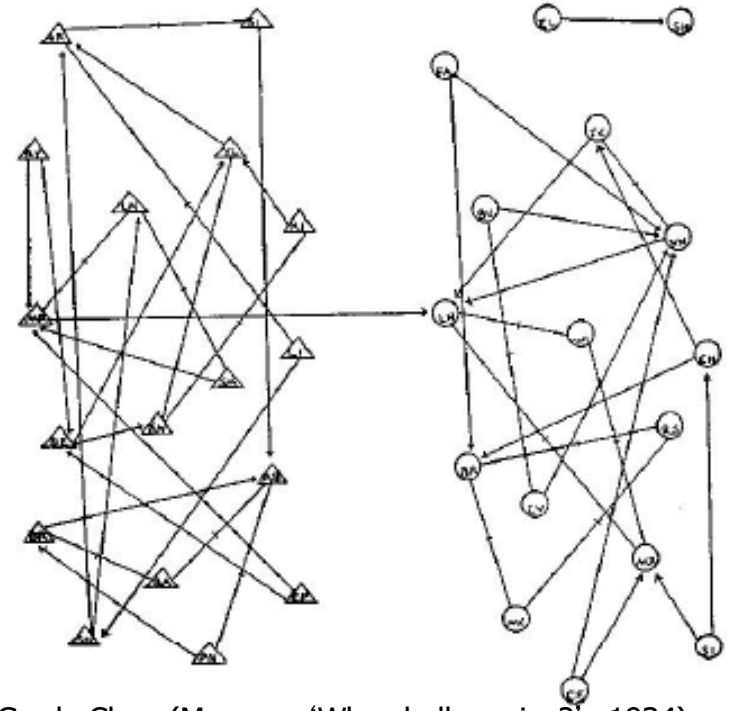
- Which document is a better result? Why?
  - docs.python.org
  - alex.oonutrition.ru

Importance in the real world

- How do people indicate something is important?
  - Talk about it
  - Reference it
  - Upvote it
  - Link to it

History of importance in graphs

- School kids – favorite (and captive) subjects of study
- These days much more difficult because need parental consent to gather social network data

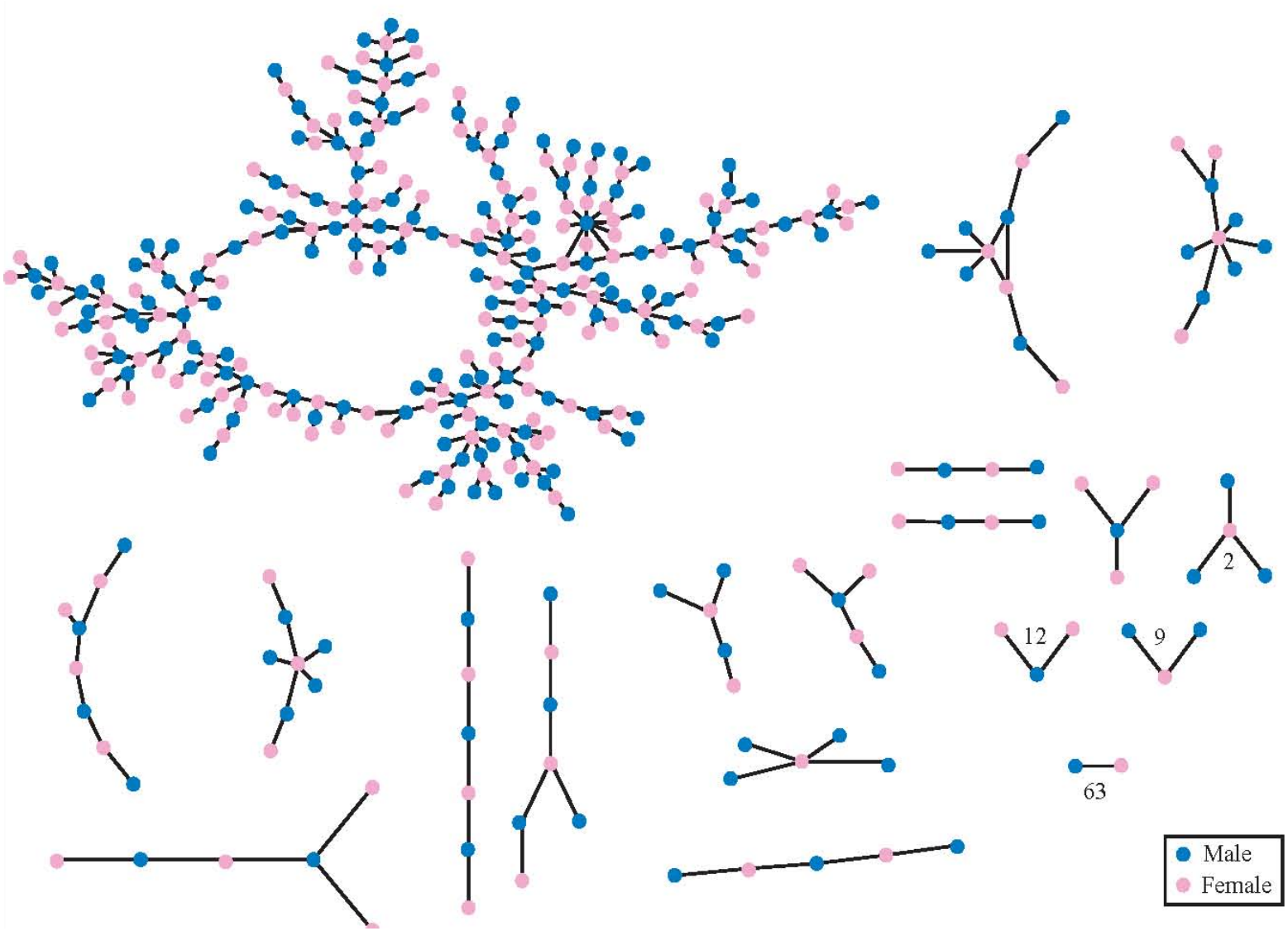


△ *boys*

○ *girls*

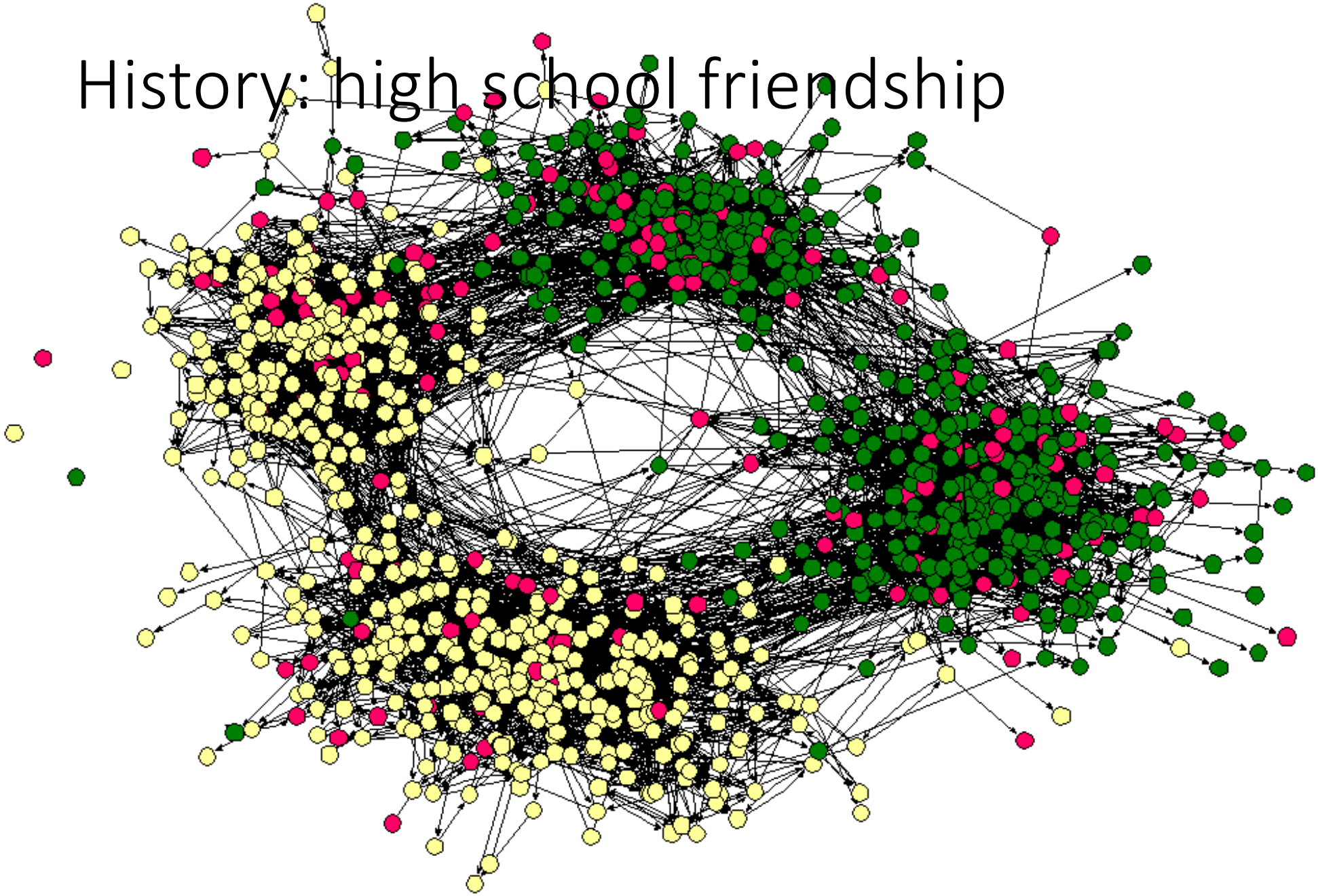
An Attraction Network in a Fourth Grade Class (Moreno, 'Who shall survive?', 1934).

History: high school dating



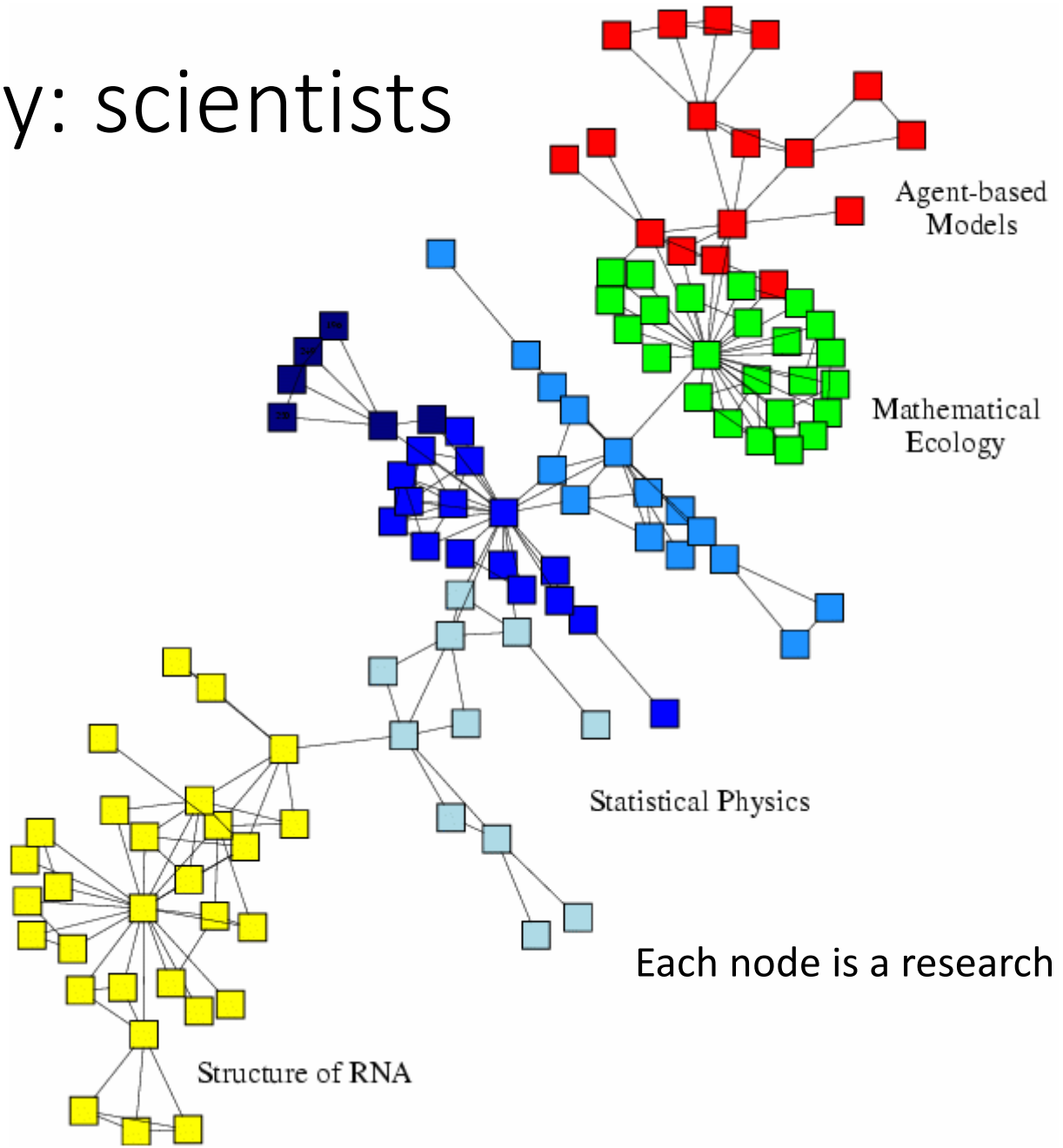
Chains of affection: The structure of adolescent romantic and sexual networks, Bearman, et al., American Journal of Sociology 110, 44-91 (2004)

History: high school friendship



High school friendship: James Moody, Race, school integration, and friendship segregation in America, American Journal of Sociology 107, 679-716 (2001).

History: scientists

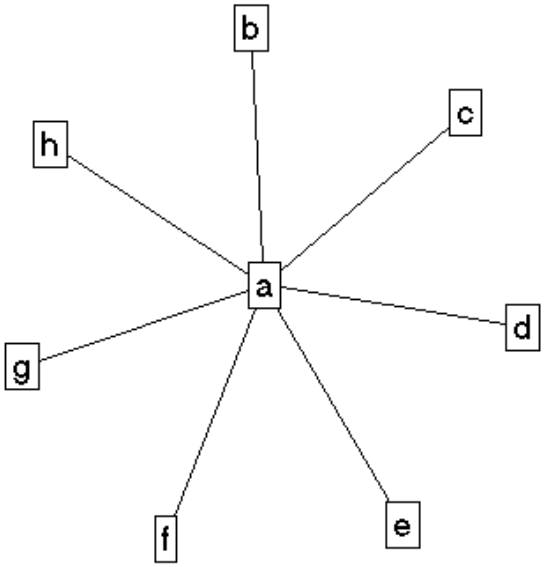


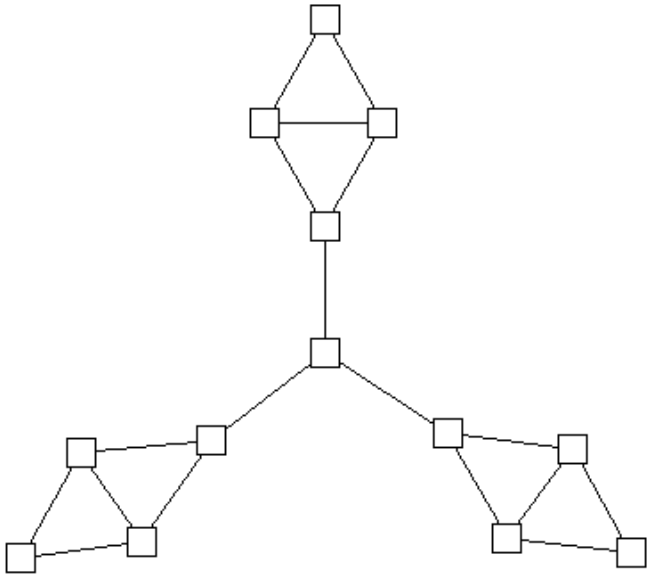
Each node is a research scientist

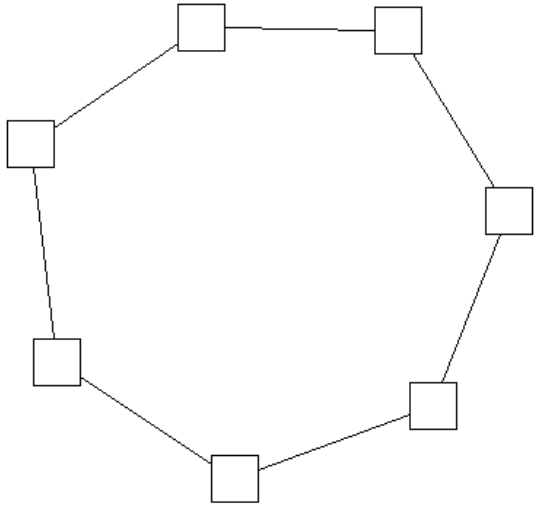
M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 8271-8276 (2002).

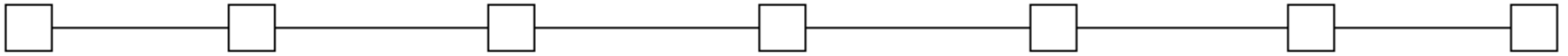
Importance in graphs

- Which node(s) are the most important?
- How would you measure it?



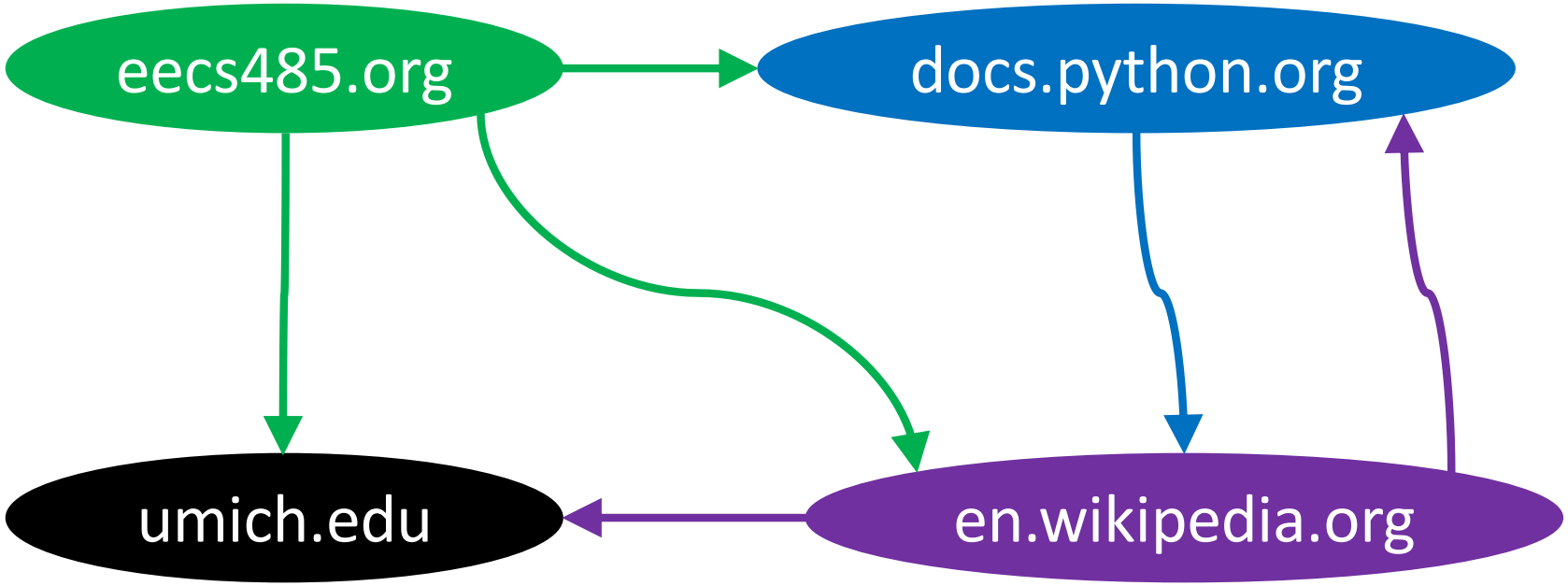






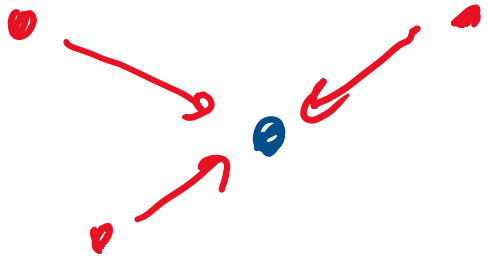
Link graph

- Web as a *graph* (AKA *network*)
  - Each web page is a vertex
  - Each hyperlink is a directed edge

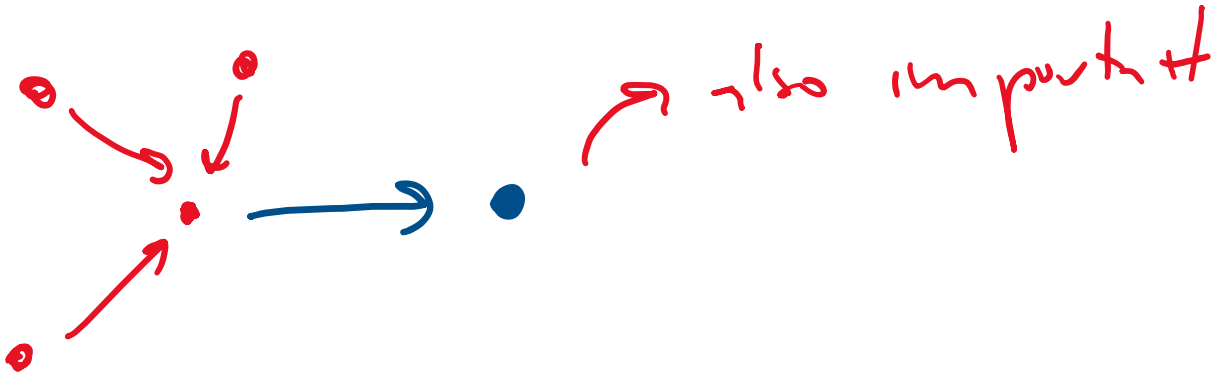


Using the link graph for importance

- If lots of pages link to a page, it is important



- If an important page links to a page, that page also is important



Web search history

- Search in the 1990s was pretty bad
  - Hand-curated list (e.g. Yahoo)
  - Search engines used only page content (e.g. AltaVista)
- Late 1990's: improve search using hyperlink graph
  - **PageRank**, Page
  - **HITS**, Kleinberg

More on the Wikipedia Timeline of web search engines

Page importance and search results

Goal:

1. Make a graph of the web
2. Figure out which sites are more important
3. Rank those sites higher in search results

Agenda

- Document importance
- **Page Rank algorithm**
- HITS algorithm
- Search engine optimization

PageRank

- Algorithm that made Google famous

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

Idea behind PageRank

- Humans know better than computers which pages are important

- Humans indicate importance through links
  - Like citations on an academic paper

PageRank intuition

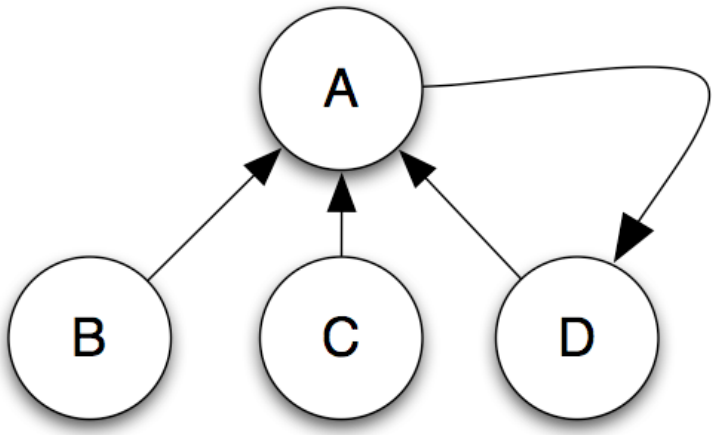
- Assume someone starts on a random web page
- Click a random link on that page
- Over and over
- Web pages with more visits are more important

PageRank

- A node with C links contributes 1/C of its PageRank to each target node

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

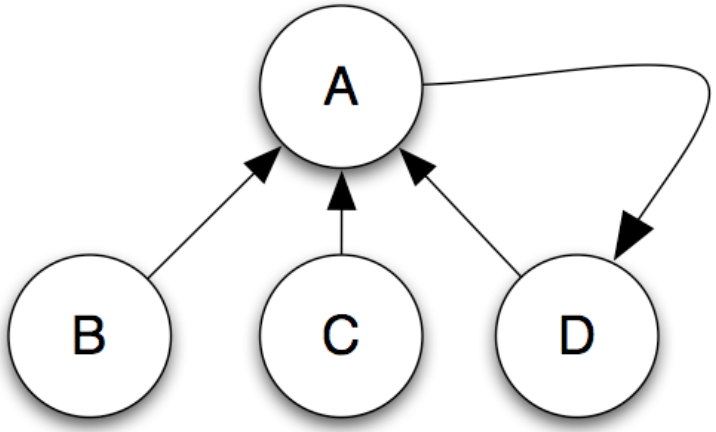
- Damping factor d is usually 0.85



PageRank example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

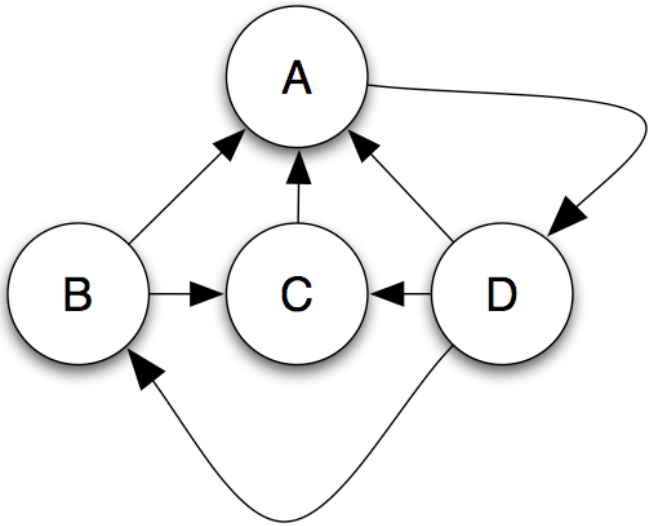
- Total PR = 1, so initialize each node to 0.25
- Set d = 0.85
- PR(A) = (1 - 0.85)/4 +
  0.85 * (0.25/1 + 0.25/1 + 0.25/1)
- PR(A) = 0.675



PageRank example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

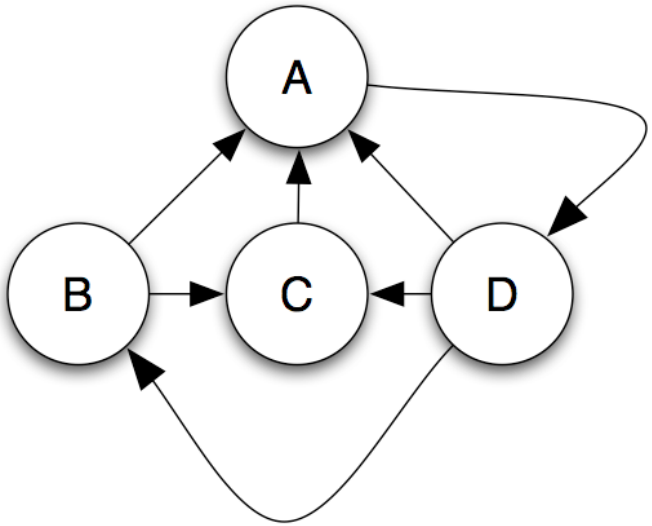
- Again, initialize all nodes to 0.25 and d=0.85
- PR(A) = *compute this*



PageRank example

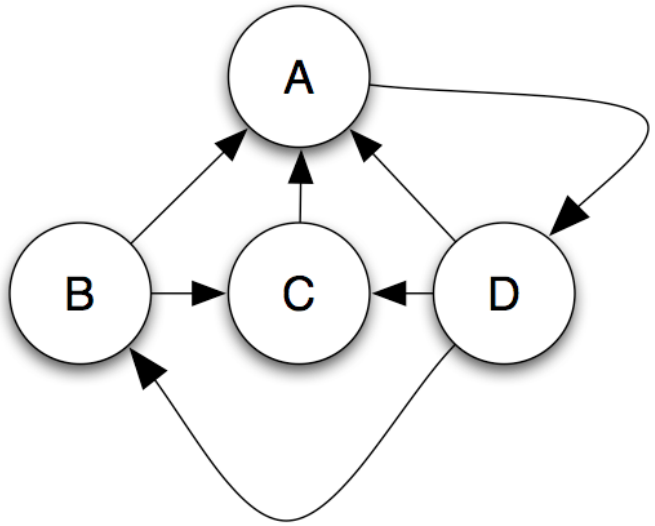
$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

- Again, initialize all nodes to 0.25 and d=0.85
- PR(A) = (1 - 0.85)/4 +
  0.85 * (0.25/2 + 0.25/1 + 0.25/3)
  = .0375+ .85*(0.125 + 0.25 + 0.083)
  = 0.4268



Example

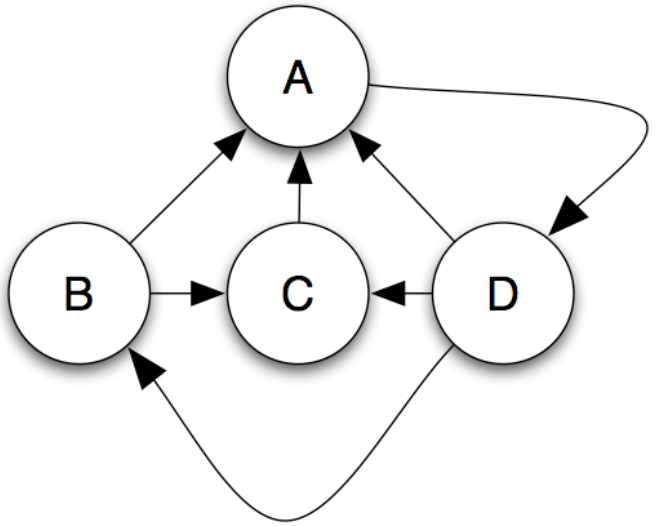
$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



A	**B**	**C**	**D**
0.25	0.25	0.25	0.25

Example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

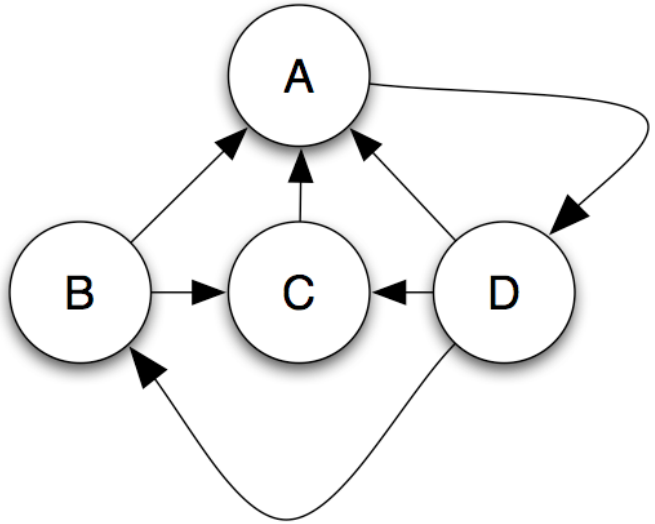


PR(A) = 0.0375 + 0.85(0.25/2 + 0.25/1 + 0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428			

Example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

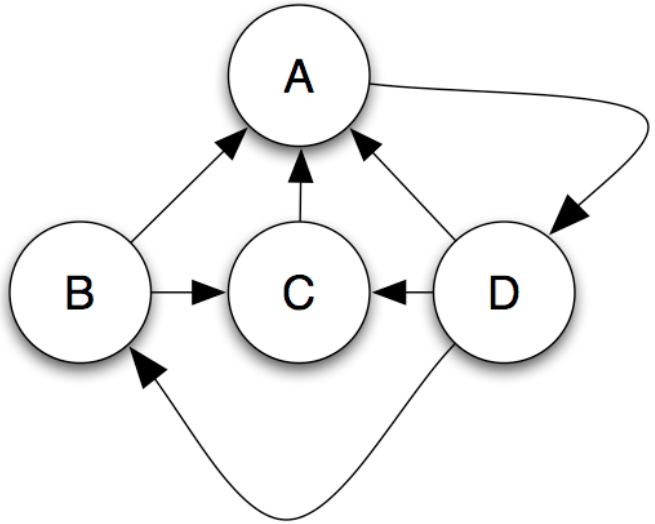


PR(B) = 0.0375 + 0.85(0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428	0.109		

Example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

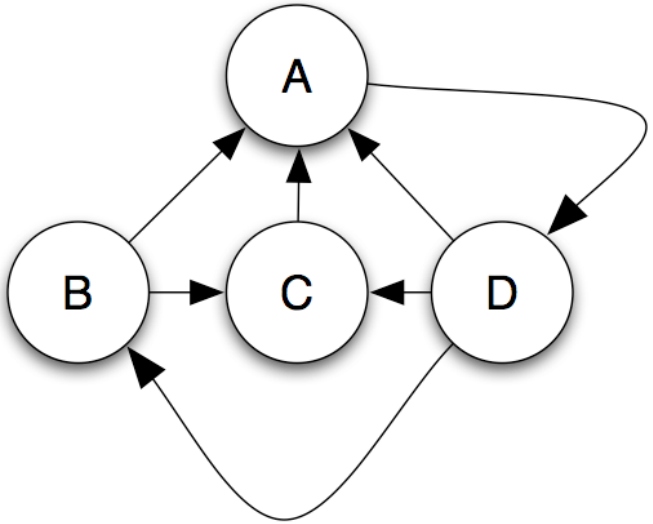


PR(C) = 0.0375 + 0.85(0.25/2 + 0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428	0.109	0.215	

Example (normalization)

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

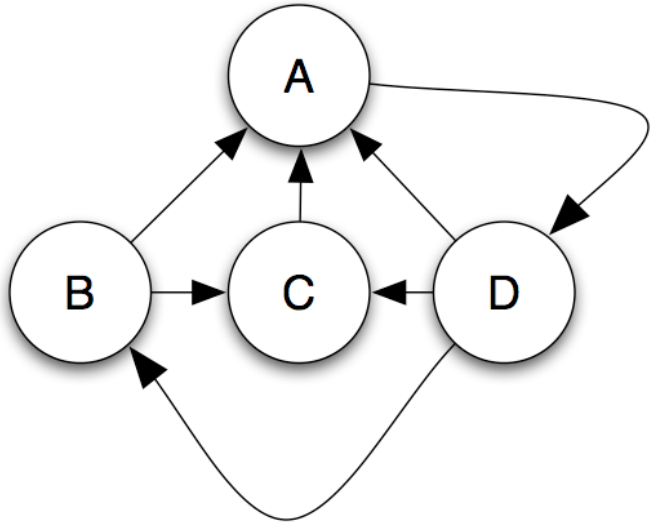


A	B	C	D
0.25	0.25	0.25	0.25
0.428	**0.109**	**0.215**	**0.25**

Sum > 1. Normalize before next iteration.

Example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

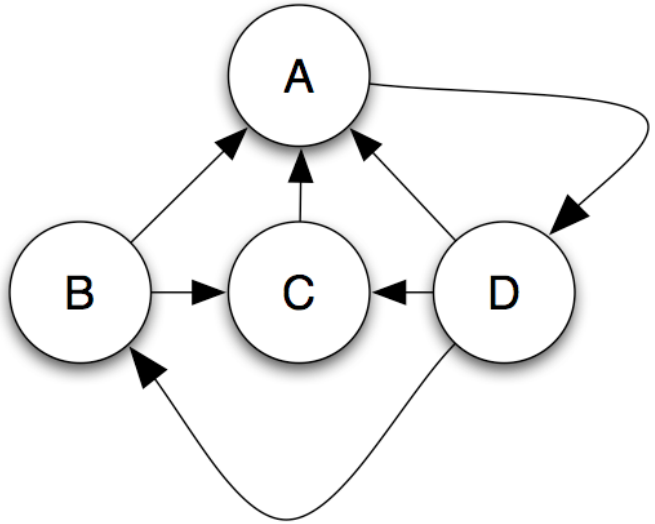


PR(D) = 0.0375 + 0.85(0.25/1)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.427	0.108	0.215	0.25

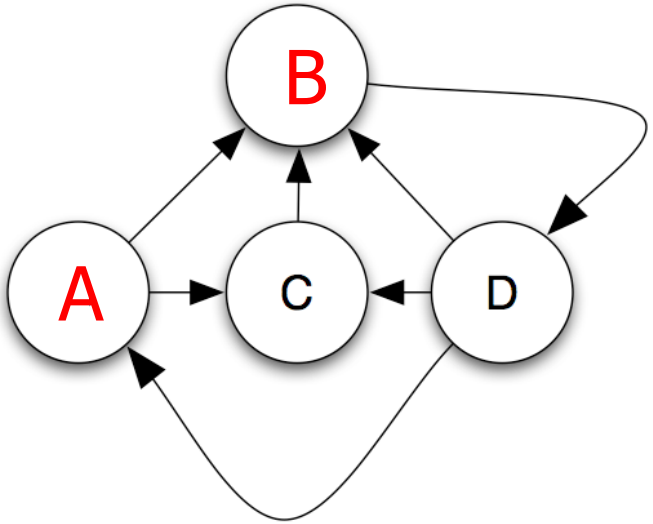
Example

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



A	B	C	D
0.25	0.25	0.25	0.25
0.427	0.108	0.215	0.25
0.337	0.108	0.154	0.401
0.328	0.151	0.197	0.324
0.361	0.129	0.193	0.317

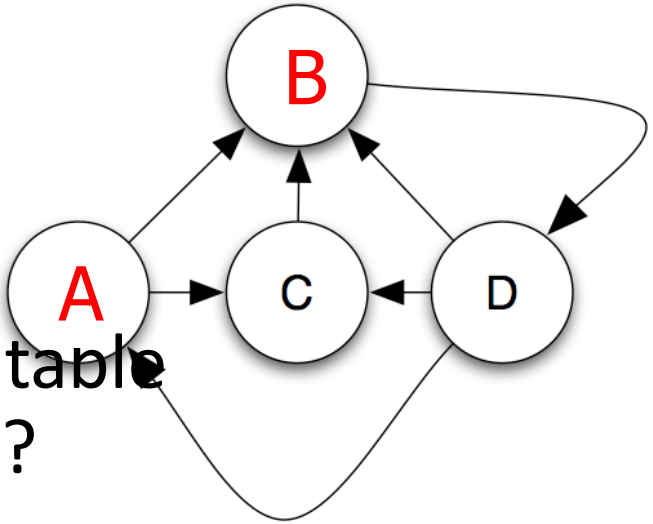
Exercise



- If I change the order of nodes in my table (or graph) will the page rank change?

~~A~~ **B**	~~B~~ **A**	**C**	**D**
0.25	0.25	0.25	0.25

Exercise



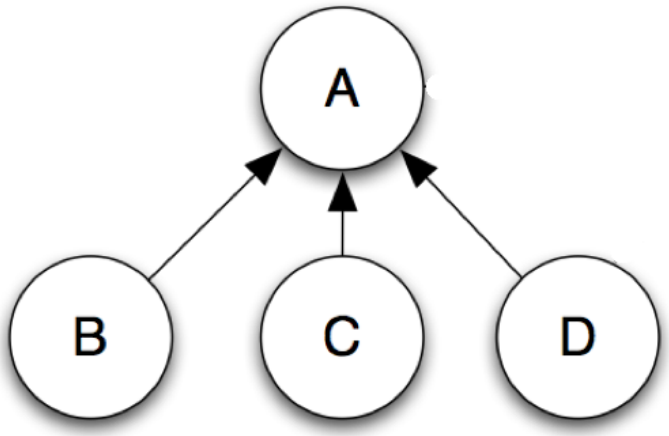
- If I change the order of nodes in my table (or graph) will the page rank change?
- No: labels don't matter, structure of the graph does

~~A~~ **B**	~~B~~ **A**	**C**	**D**
0.25	0.25	0.25	0.25
0.427	0.108	0.215	0.25
0.337	0.108	0.154	0.401
0.328	0.151	0.197	0.324
0.361	0.129	0.193	0.317

Sink nodes

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

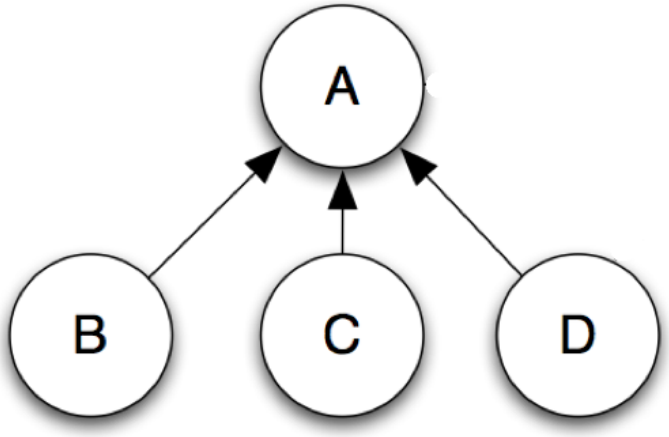
- What happens after many iterations?



Sink nodes

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

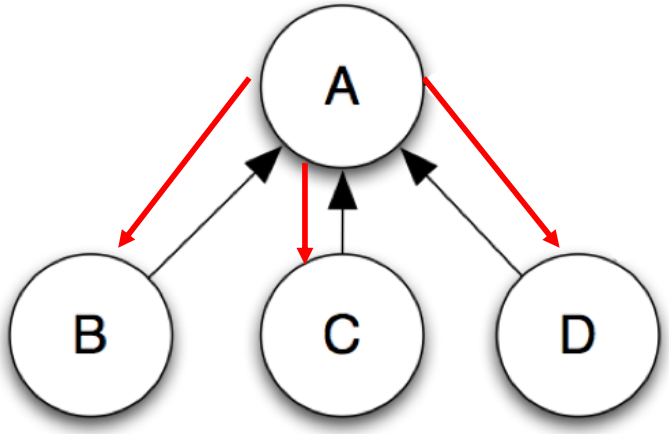
- What happens after many iterations?
  - PR(A) keeps increasing
  - PR(B) = PR(C) = PR(D) decreasing



Sink nodes

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

- Nodes with no outlinks are disallowed
- Can "drain rank" from rest of system
- Solution: Add edge from sink=>every node



Sink regions

- Must have non-zero probability of reaching every node from every other node
- Solution: with prob (1-d), random surfer types in a random URL instead of clicking a link

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

Adding PageRank to a search engine

- Weighted sum of page importance and query-similarity
- Score(query, doc)=
  - w*sim(q, p) + (1-w) * PR(p)
    - If sim(q, p) > 0
    - Otherwise, 0
- Where:
  - 0 < w < 1
  - Values sim(q,p) and PR(p) are normalized

Agenda

- Document importance
- Page Rank algorithm
- **HITS algorithm**
- Search engine optimization

Hubs and authorities

- Due to Kleinberg, 1997
- Unlike PageRank, is query-dependent

- A page is a good ***authority*** if it is pointed-to by many good ***hubs***
- A page is a good ***hub*** if it points to many good ***authorities***
- Good hubs and authorities reinforce each other

HITS algorithm

$$auth(p) = \sum_{i=1}^{n} hub(i)$$

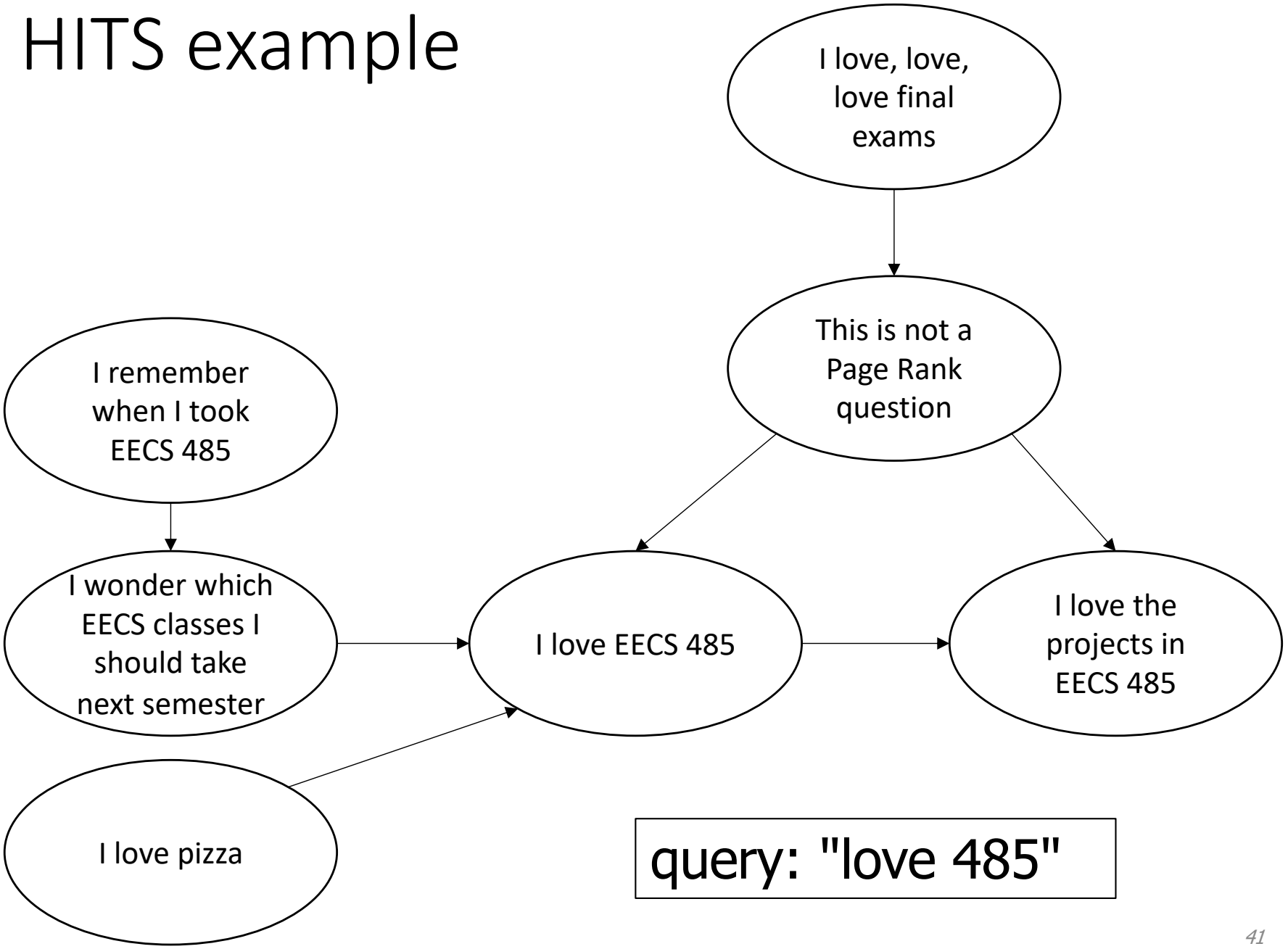
$$hub(p) = \sum_{i=1}^{n} auth(i)$$

1. Obtain *root set* using input query
2. Expand the root set by radius 1. This is called the *base set*
3. Iteratively compute hub and authority scores for each node in graph

More HITS

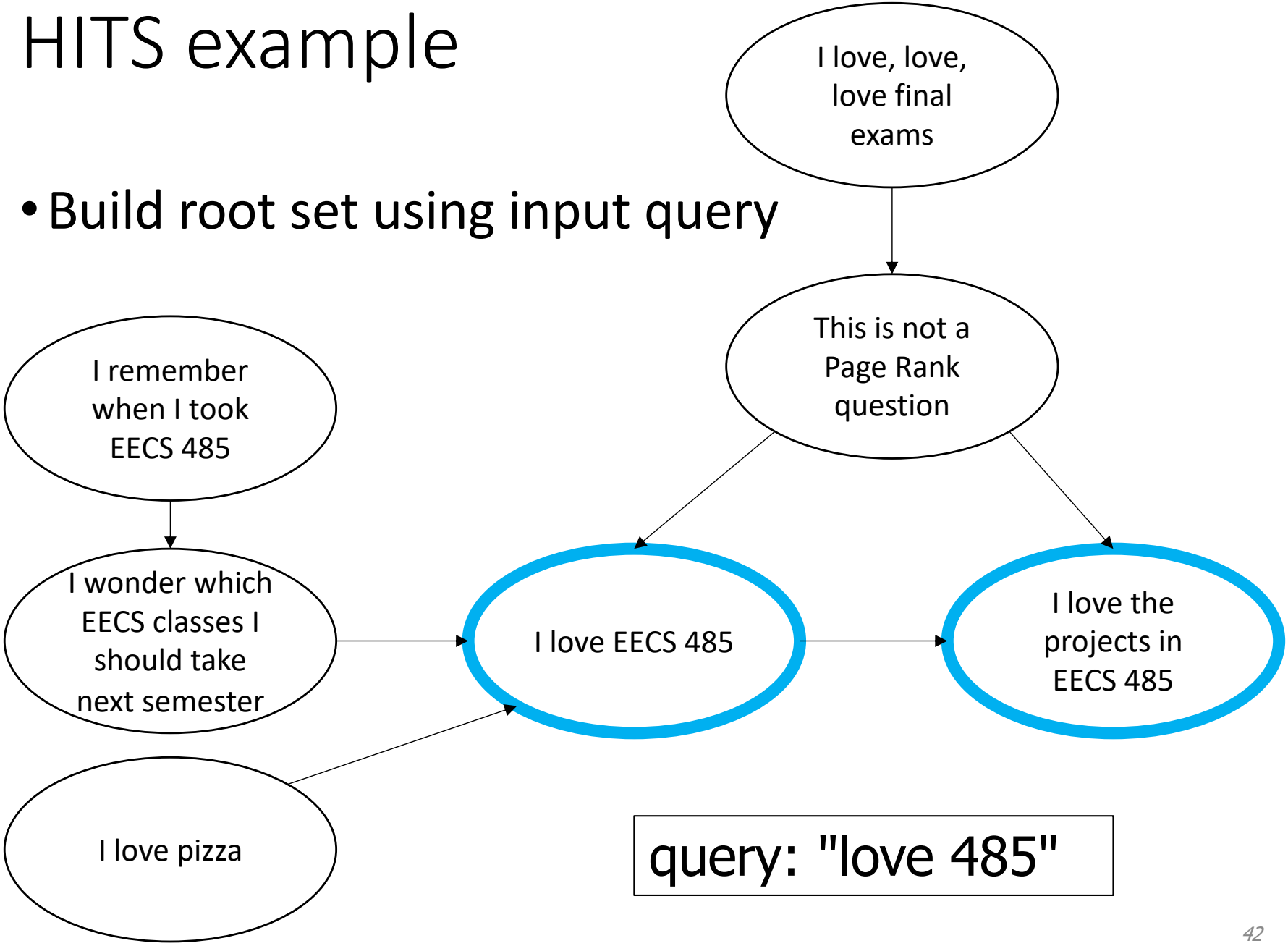
1. Initialize all auth and hub scores to 1
2. For all nodes, update authority scores
3. For all nodes, update hub scores
4. Normalize scores
   - Divide each auth by sqrt(sum($auth^2$))
   - Divide each hub by sqrt(sum(hub^2))
5. If converges, terminate; else goto 2

HITS example



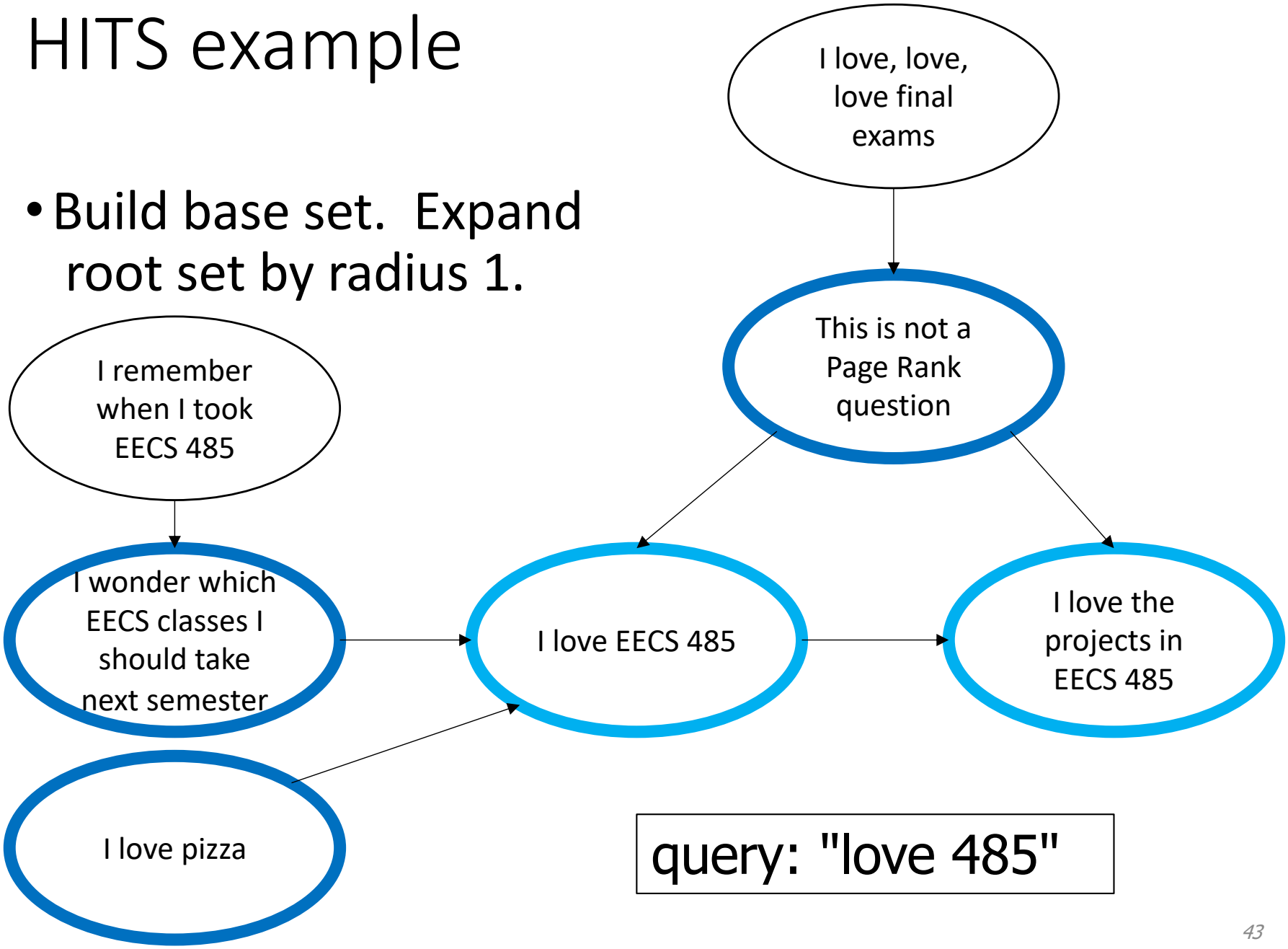
HITS example

- Build root set using input query



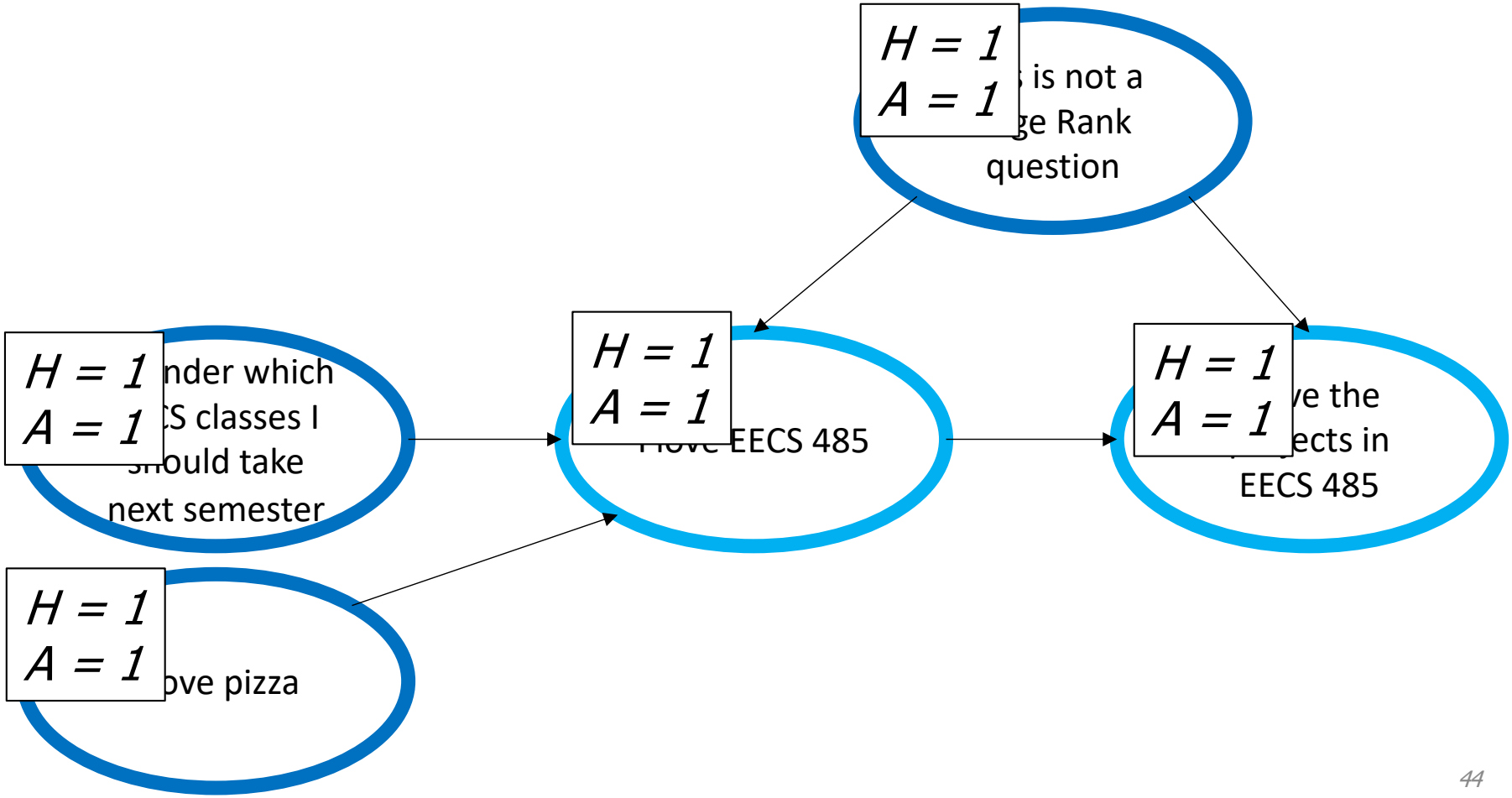
HITS example

- Build base set. Expand root set by radius 1.



HITS example

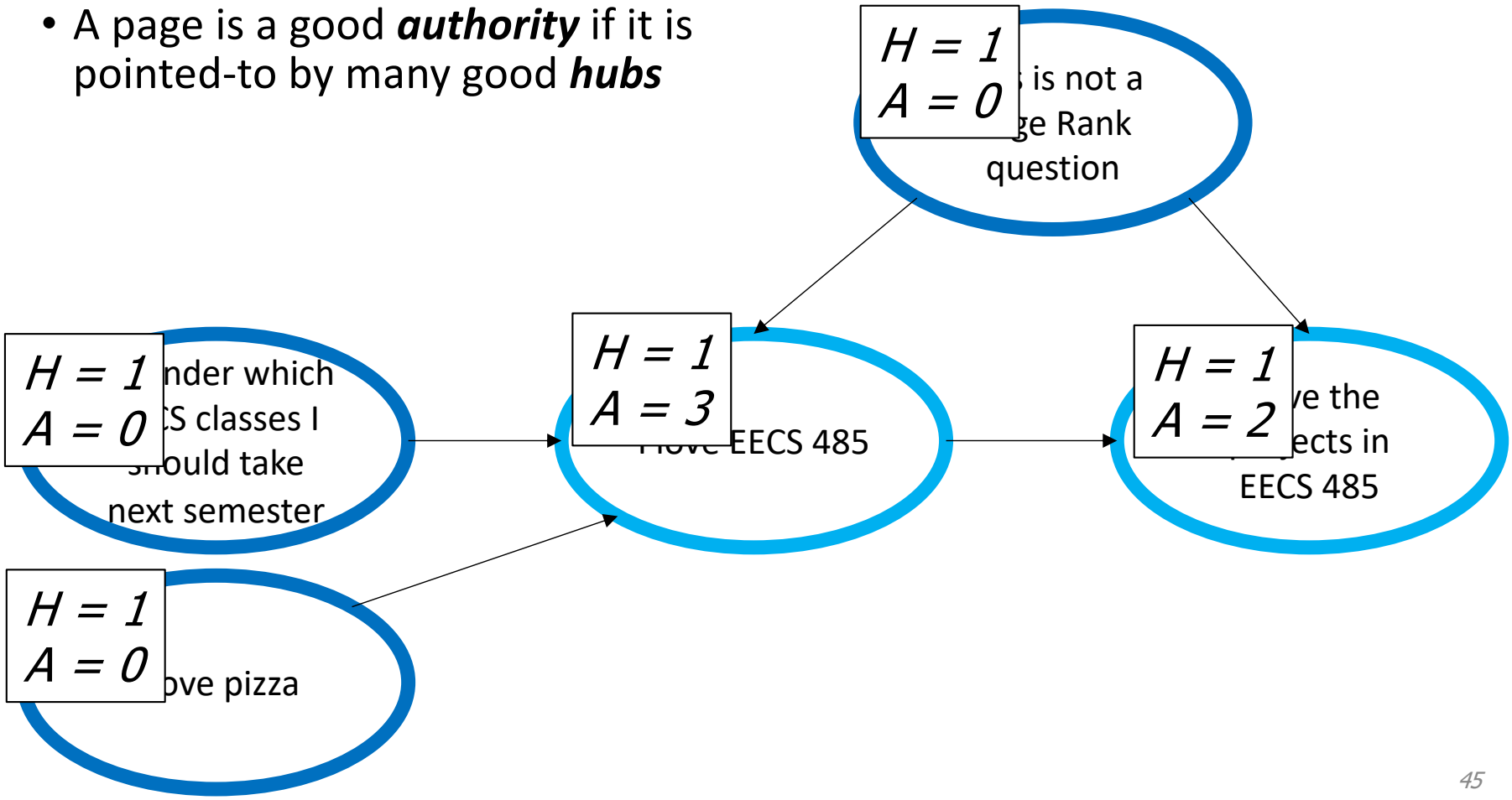
- Initialize all hub and authority scores to 1



HITS example

$$auth(p) = \sum_{i=1}^{n} hub(i)$$

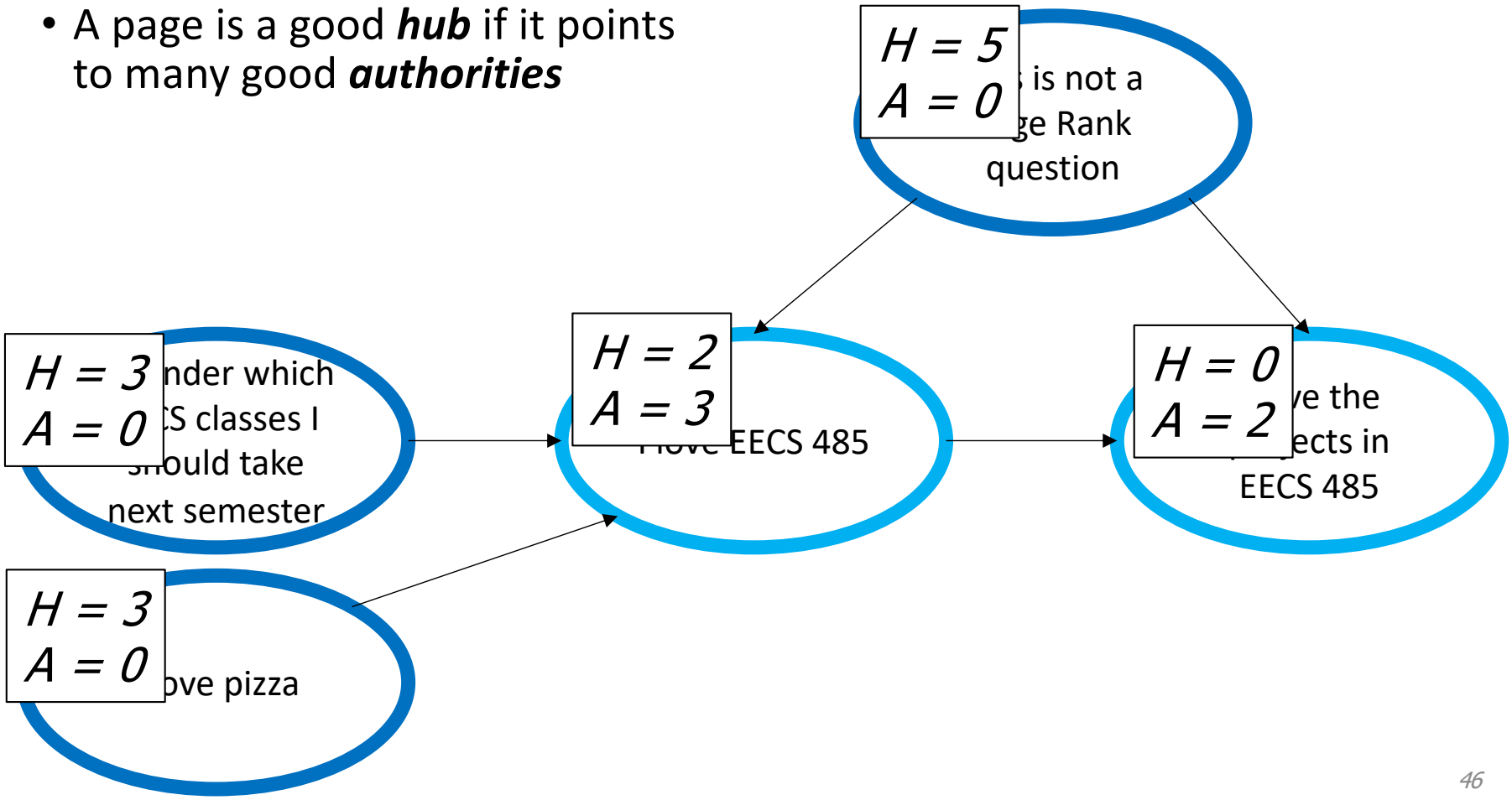
- Update authority scores
- A page is a good ***authority*** if it is pointed-to by many good ***hubs***



HITS example

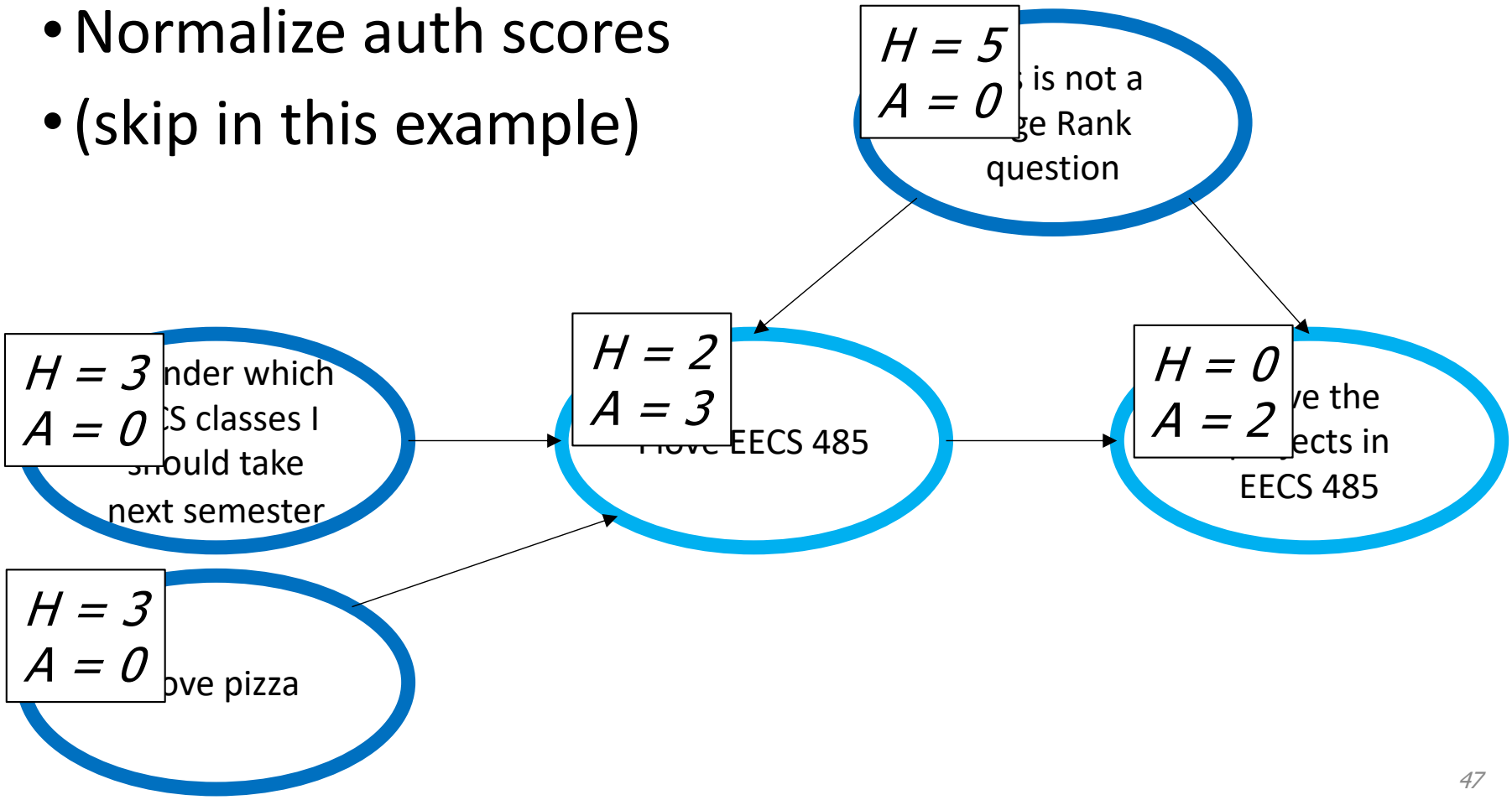
$$hub(p) = \sum_{i=1}^{n} auth(i)$$

- Update hub scores
- A page is a good ***hub*** if it points to many good ***authorities***



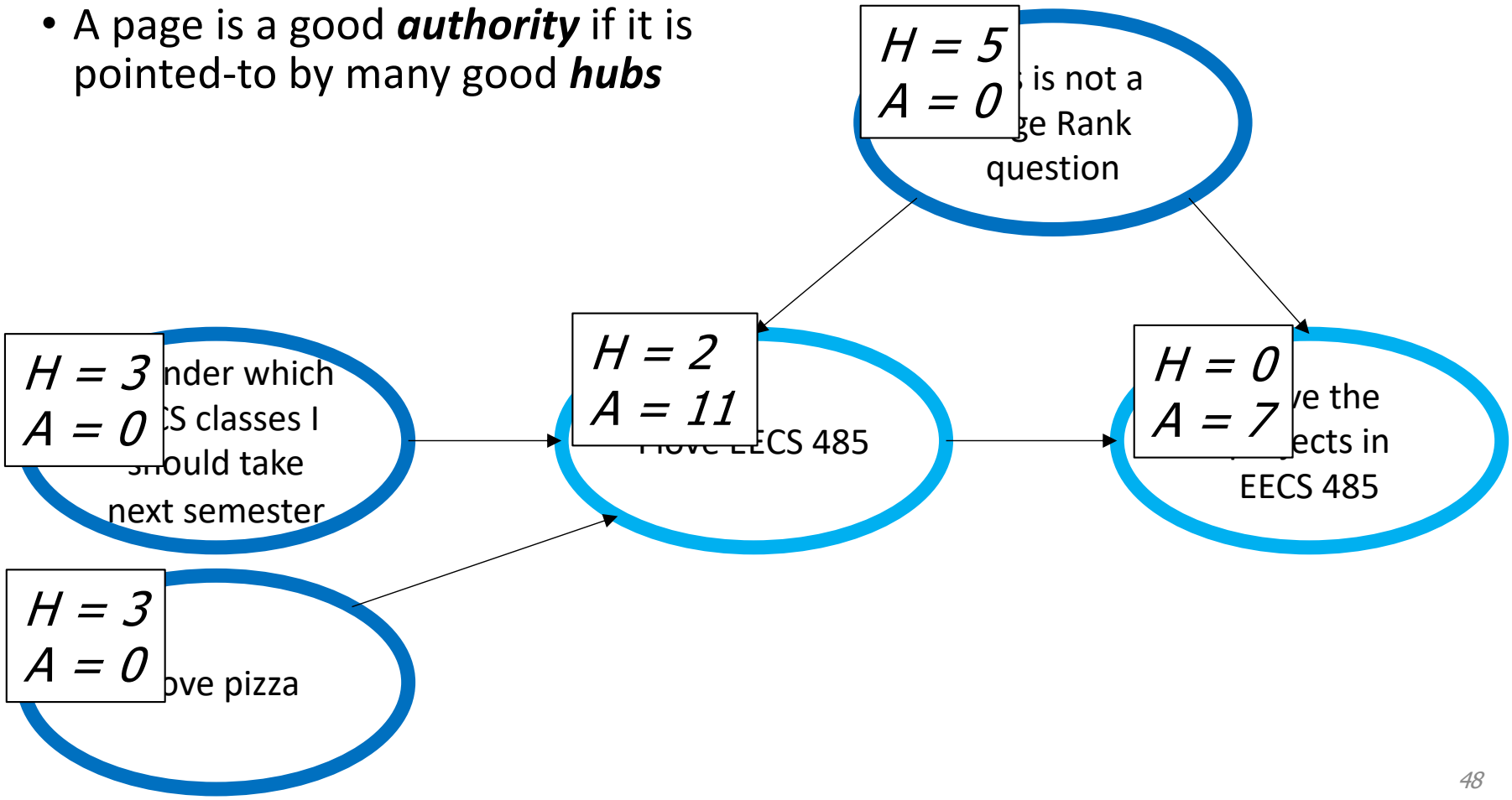
HITS example

- Normalize hub scores
- Normalize auth scores
- (skip in this example)



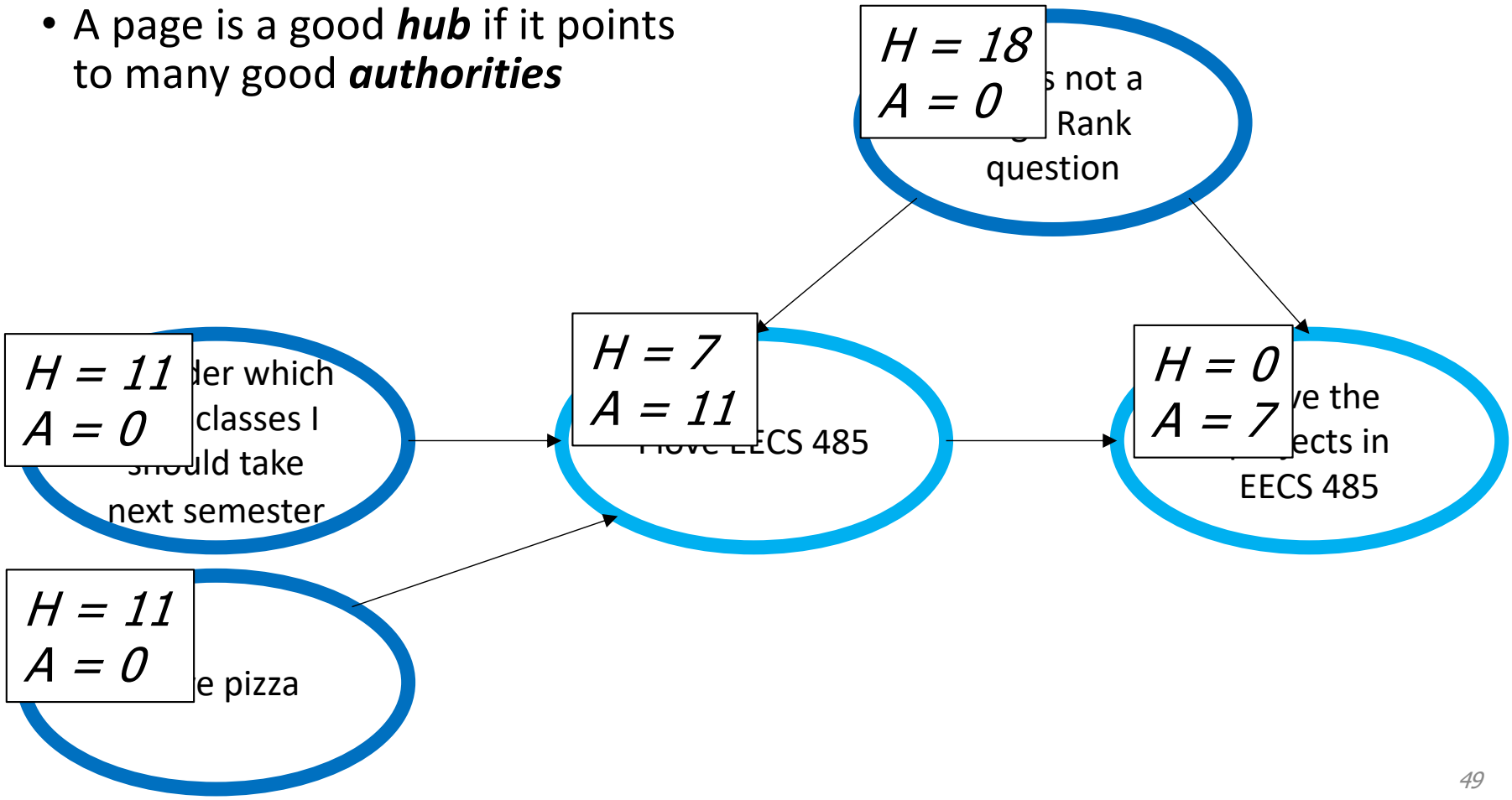
HITS example

- Update authority scores
- A page is a good ***authority*** if it is pointed-to by many good ***hubs***



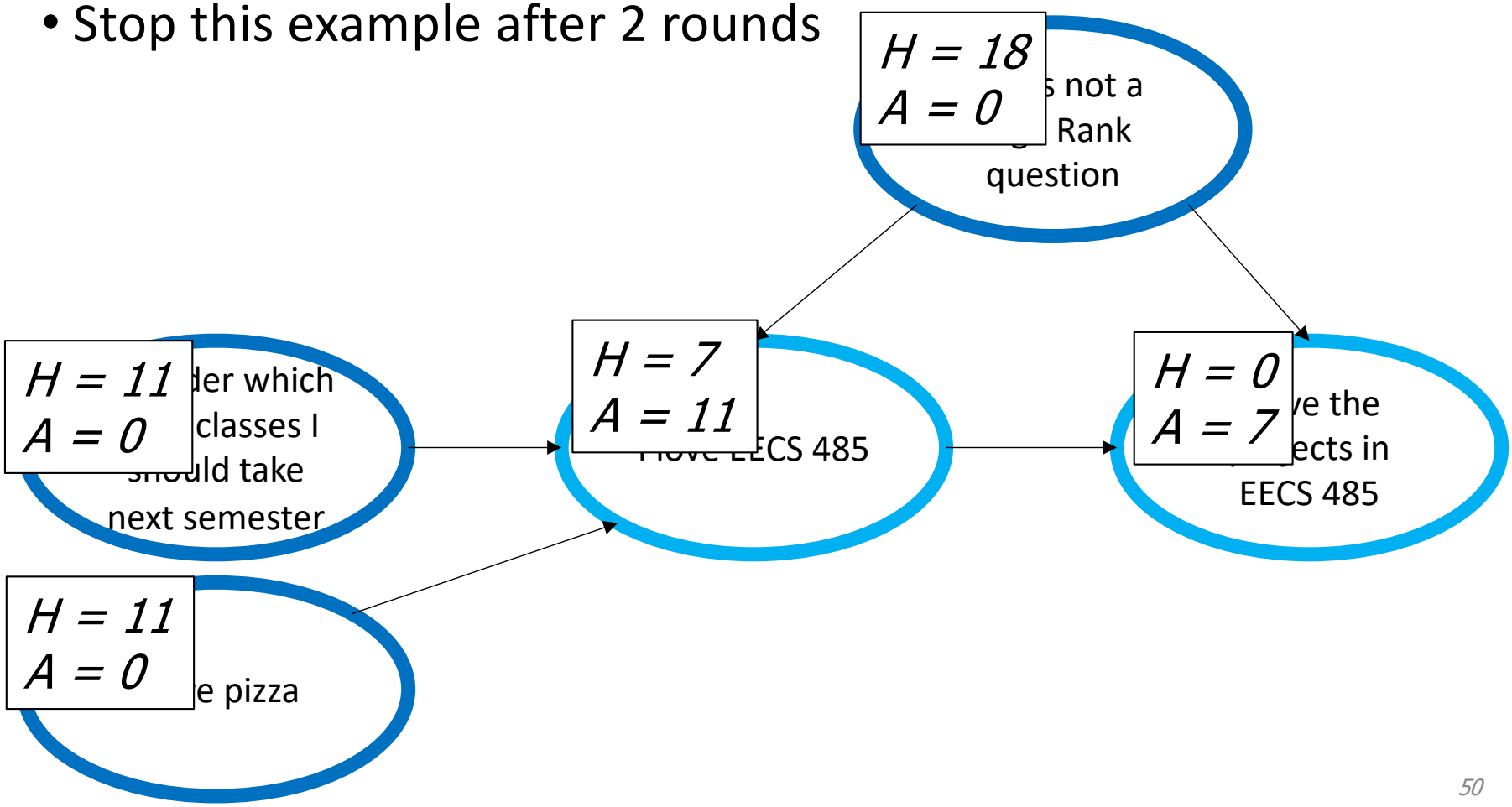
HITS example

- Update hub scores
- A page is a good ***hub*** if it points to many good ***authorities***



HITS example

- HITS continues until convergence
- Stop this example after 2 rounds



HITS vs. PageRank

- HITS is dependent on the query
  - You need the query to build the root set and base set
- PageRank is independent of a query
  - One considers the links between pages, nothing else
  - Compute it once and save the result
- Scores from either/both algorithms can be combined with Vector model tf-idf similarity score

Agenda

- Document importance
- Page Rank algorithm
- HITS algorithm
- **Search engine optimization**

Search engine optimization

- *Search engine optimization:* techniques to increase the visibility of your page in search engine results

- Example 1: You have a personal web page. You want your web page to be the top hit when someone searches for your name.
- Example 2: Most users don't look beyond first page of search results. If your business appears on the first page, more clicks. More clicks -> more business.

Keywords

- TF x IDF is one factor in search rankings
- What can you do to influence TF x IDF?

Keywords

- TF x IDF is one factor in search rankings
- What can you do to influence TF x IDF?
  - Pages focus on a particular topic
  - You'll probably naturally use important (relevant) words multiple times
  - Length won't help you (remember normalization?)
  - Create a page with many repetitions of keywords
    - Search engines hate this. They "spam" filter for it.

Keyword technical considerations

- Search engine needs text to extract keywords
- Use text rather than images for important content
  - Is your name in a banner image? Is it also in the text?
- Site should work with JavaScript disabled
  - Some search engine use a JavaScript rendering engine on some pages, but don't count on it.

Links

- PageRank is another factor in search rankings
- What can you do to influence Page Rank?

Link spam

- Bots add comments to blogs. Comments link back to your site.
- Blog owner should use `rel="nofollow"` attribute for links in comments. Search engines then ignore these links.
- Search engines hate this. They "spam" filter for it.

Link technical considerations

- Avoid links that look like form queries
  - http://www.mysite.com/info?about
- Use different links for different pages
  - Think about how wolverine access works. You click a link, and the URL doesn't change! Don't do this.

Other relevance factors

- Google says, “Relevancy is determined by over 200 factors, one of which is the PageRank for a given page.”

Next time

- Last time we used the words on a page to rank search results
  - Rank on document content
- Today we used the links between web pages to improve search results
  - Rank on document importance
  - Next time we'll cover speed and scaling