# Link Analysis in Document Networks

Compiled by: Cornelia Caragea, Sujatha Das

Credits: Giles, Manning, Hofmann, Mihalcea, Mobasher, Mooney, Schutze

August 22, 2014

# Searching the Web

Google cornell Florin Marines

Web Images Maps Shopping News More Search tools

About 58,800,000 results (0.38 seconds)

**Cornell** University
www.**cornell**.edu/
**Cornell** University contains seven undergraduate colleges plus the College of Veterinary Medicine, the Law School, the Samuel Curtis Johnson Graduate ...
Score: **25** / 30 · 41 Google reviews · Write a review

410 Thurston Ave Ithaca, NY 14850
(607) 255-5241

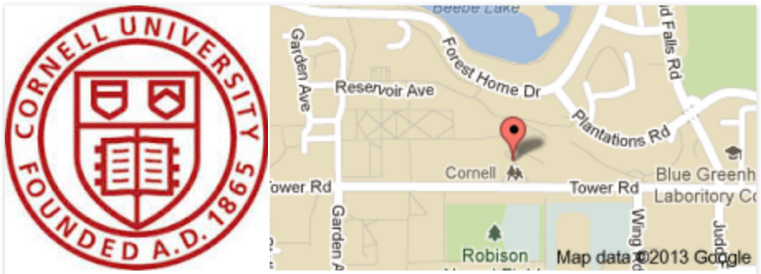
Admissions - Academics - CUinfo

**Cornell** University - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Cornell**_University
**Cornell** University is an American private Ivy League research university located in Ithaca, New York, United States. Founded in 1865 by Ezra **Cornell** and ...
History - Ithaca, New York - List of Cornell University alumni - Arts and Sciences

**Cornell** University Athletics
www.**cornell**bigred.com/
Official web site of Big Red athletics. Information about varsity sports, facilities, schedules, and the department, as well as an alumni section and Big Red Store.

**Cornell** University (**Cornell**) on Twitter
https://twitter.com/**Cornell**
The latest from **Cornell** University (@**Cornell**). **Cornell** University Twitter feed. Ithaca, NY.

**Cornell** Home
www.**cornell**college.edu/
Residential liberal arts college established in 1853. Operates under the distinctive One-Course-At-A-Time academic calendar.




## Cornell University

80,190 followers on Google+

Directions Follow

Cornell University is an American private Ivy League research university located in Ithaca, New York, United States. Wikipedia

**Address:** 410 Thurston Ave, Ithaca, NY 14850
**Acceptance rate:** 16.2% (2012)
**Mascot:** Big Red Bear
**Phone:** (607) 255-5241
**Colors:** White, Carnelian
**Founders:** Andrew Dickson White, Ezra Cornell

Recent posts

# Searching the Web

Google cornell Florin Marines

Web Images Maps Shopping News More Search tools

About 58,800,000 results (0.38 seconds)

**Cornell** University
www.**cornell**.edu/
**Cornell** University contains seven undergraduate colleges plus the College of Veterinary Medicine, the Law School, the Samuel Curtis Johnson Graduate ...
Score: **25** / 30 · 41 Google reviews · Write a review

410 Thurston Ave Ithaca, NY 14850
(607) 255-5241

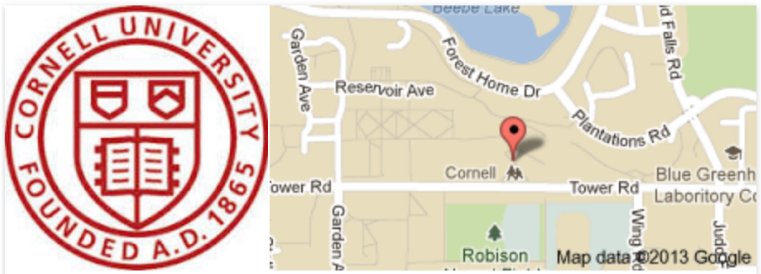
Admissions - Academics - CUinfo

**Cornell** University - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Cornell**_University
**Cornell** University is an American private Ivy League research university located in Ithaca, New York, United States. Founded in 1865 by Ezra **Cornell** and ...
History - Ithaca, New York - List of Cornell University alumni - Arts and Sciences

**Cornell** University Athletics
www.**cornell**bigred.com/
Official web site of Big Red athletics. Information about varsity sports, facilities, schedules, and the department, as well as an alumni section and Big Red Store.

**Cornell** University (**Cornell**) on Twitter
https://twitter.com/**Cornell**
The latest from **Cornell** University (@**Cornell**). **Cornell** University Twitter feed. Ithaca, NY.

**Cornell** Home
www.**cornell**college.edu/
Residential liberal arts college established in 1853. Operates under the distinctive One-Course-At-A-Time academic calendar.




## Cornell University

80,190 followers on Google+

Directions Follow

Cornell University is an American private Ivy League research university located in Ithaca, New York, United States. Wikipedia

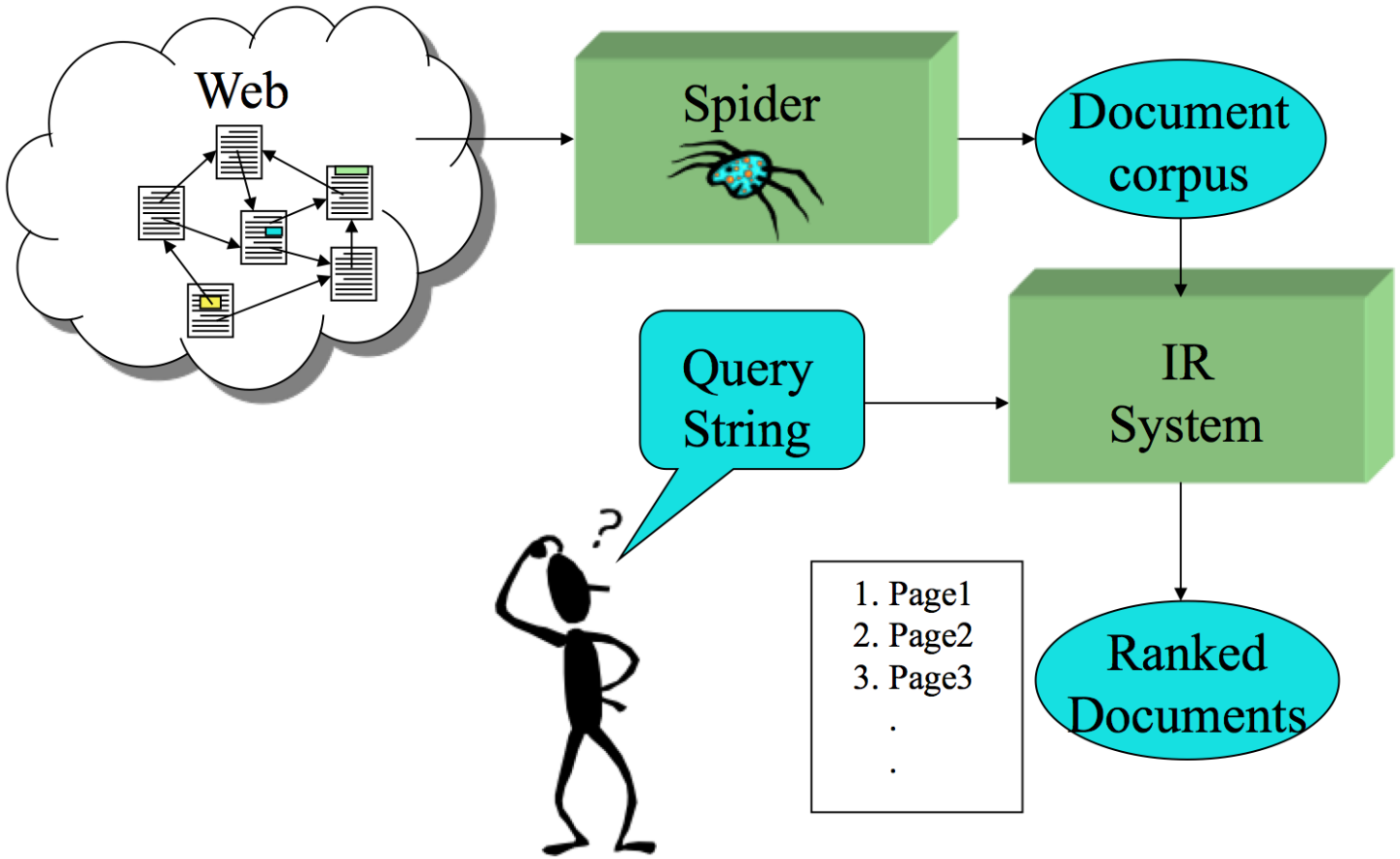
**Address:** 410 Thurston Ave, Ithaca, NY 14850
**Acceptance rate:** 16.2% (2012)
**Mascot:** Big Red Bear
**Phone:** (607) 255-5241
**Colors:** White, Carnelian
**Founders:** Andrew Dickson White, Ezra Cornell

Recent posts

# Searching the Web: Search Engines

- When issuing the single-word query "cornell," a search engine does not have very much to go on.
  - Did the searcher want information about the university?
  - The university's hockey team?
  - Cornell College in Iowa?
  - The Nobel-Prize-winning physicist Eric Cornell?
- The Problem of Ranking
  - Search engines *determine how to rank pages* using automated methods
  - There must be enough information *intrinsic* to the Web and its structure to figure out that "Cornell University" is the best answer.

# Web Search System




## Key issue for search engines:

- To filter, from among an enormous number of relevant documents, the few that are most important

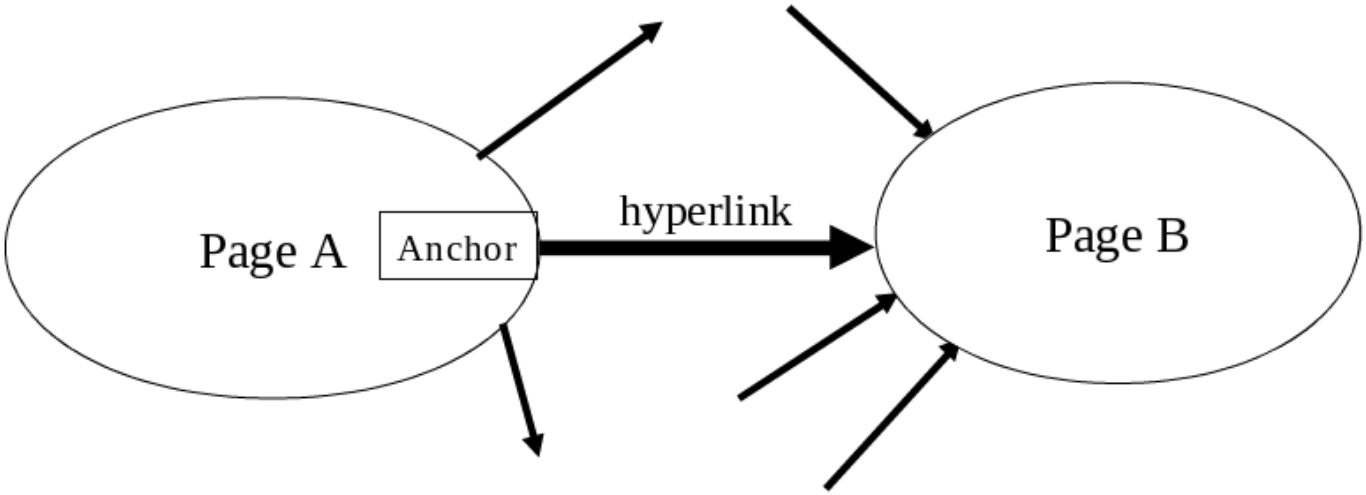
# Challenges in Web Search

- Volume of material – several billion items, growing steadily
- Items created dynamically or in databases
- Great variety – length, formats, quality control, purpose, etc.
- Inexperience of users – range of needs
- Economic models to pay for the service
- Location, personalization

# Motivation for Link Analysis

- Early search engines mainly compare content similarity of the query and the indexed pages. I.e.,
  - They use information retrieval methods, cosine, TF-IDF, ...
- From mid 90's, it became clear that content similarity alone was no longer sufficient.
  - The number of pages grew rapidly in the mid-late 1990's.
    - The "cornell" query, Google estimates: millions of relevant pages.
    - How to rank the result pages suitably to present to the user?
  - Content similarity is easily spammed.
    - A page owner can repeat some words and add many related words to boost the rankings of his pages and/or to make the pages relevant to a large number of queries.

# The Web as a Directed Graph




**Assumption 1:** A hyperlink between pages denotes author perceived relevance (quality signal)

**Assumption 2:** The anchor of the hyperlink describes the target page (textual context)

# Links are essential to Ranking!

- Web pages are connected through hyperlinks (Manually-made social network)
  - Some hyperlinks: organize information at the same site.
  - Other hyperlinks: point to pages from other Web sites. Such out-going hyperlinks often indicate an implicit conveyance of authority to the pages being pointed to.
  - Pages that are pointed to by many other pages are likely to contain authoritative information.
- Hyperlinks can be used to assess the authority of a page on a topic, through implicit endorsements via pages that point to it

# Concept of Relevance

- Relevance, as conventionally defined, is binary (relevant or not relevant). It is usually estimated by the similarity between the terms in the query and each document.
- Importance measures documents by their likelihood of being useful to a variety of users. It is usually estimated by some measure of popularity.
- Web search engines rank documents by combination of relevance and importance. The goal is to present an average user with the most important of the relevant documents.

# Using hyperlinks for the query "Cornell"

- Collect pages that are relevant to "Cornell' using IR (text-only) techniques.
- Let these pages "vote" through their links for pages on the Web.
- Which page on the Web receives the greatest number of in-links from pages that are relevant to Cornell?
  - Answer: www.cornell.edu

# Hyperlink algorithms

- During 1997-1998, the two most influential hyperlink based search algorithms PageRank and HITS were reported.
- Both algorithms are related to using the link structure in social networks.
  - Exploit the hyperlinks of the Web to rank pages according to their levels of "prestige" or "authority".
  - HITS: Jon Kleinberg (Cornell University), at Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, January 1998
  - PageRank: Sergey Brin and Larry Page, PhD students from Stanford University, at Seventh International World Wide Web Conference (WWW7) in April, 1998.
- PageRank powers the Google search engine!

# Authorities and Hubs

- Authorities for a query are pages that are recognized as providing significant, trustworthy, and useful information on a topic
  - In-degree (number of pointers to a page) is one simple measure of authority
  - However in-degree treats all links as equal
  - Links from pages that are themselves authoritative should count more
- Hubs for a query are index pages that provide lots of useful links to relevant content pages (topic authorities)

# Authorities and Hubs - Examples

- Authorities:
  - Newspaper home pages
  - Course home pages
  - Home pages of auto manufacturers
- Hubs
  - List of newspapers
  - Course bulletin
  - List of US auto manufacturers
- HITS uses both Hubs & Authorities whereas PageRank uses Authorities

# Other applications of Link Analysis

- Apart from search ranking, hyperlinks are also useful for finding Web communities.
  - A Web community is a cluster of densely linked pages representing a group of people with a special interest.
- Beyond explicit hyperlinks on the Web, links in other contexts are useful too, e.g.,
  - for discovering communities of named entities (e.g., people and organizations) in free text documents, and for analyzing social phenomena in emails..

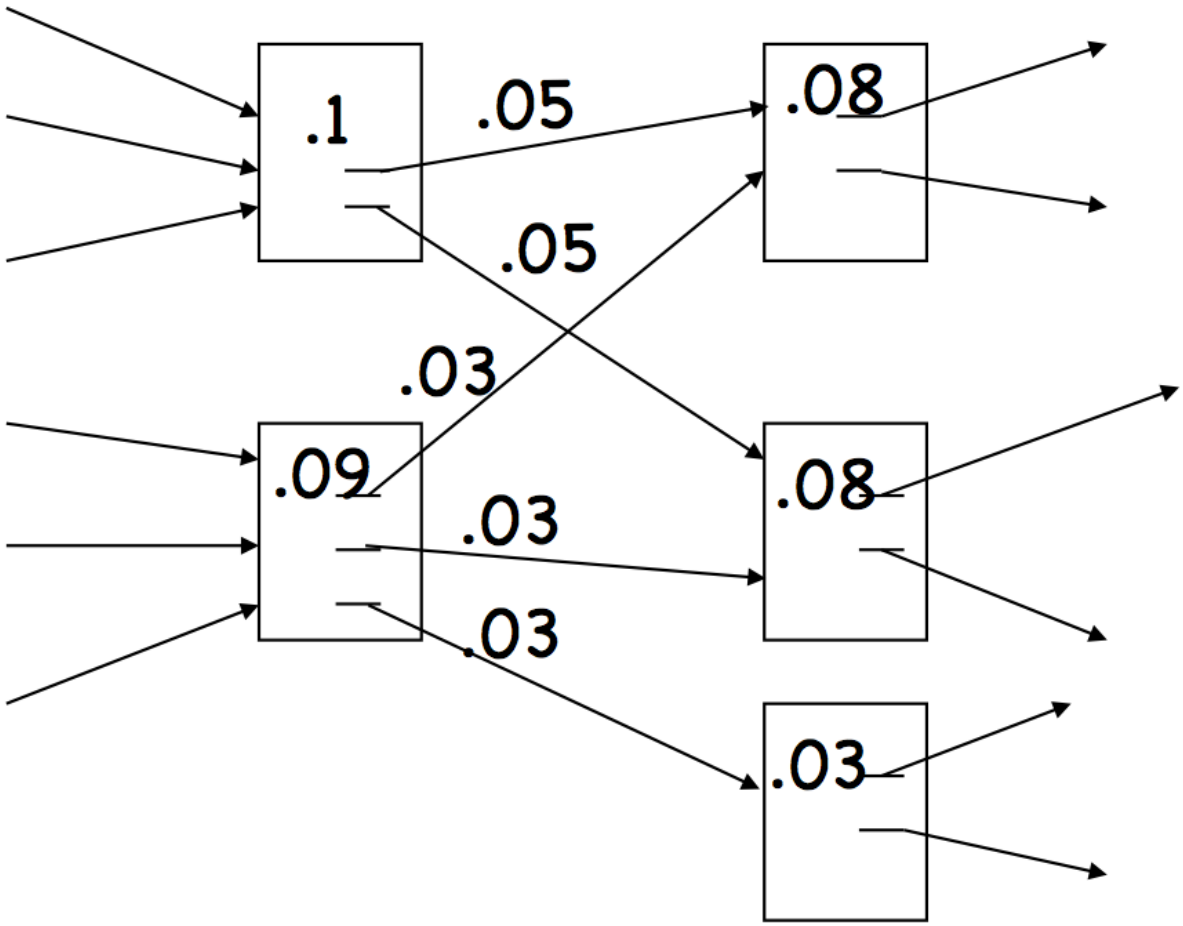
# The PageRank Algorithm

# Idea behind PageRank

- Query-independent reputation for a webpage based on link structure of the Web
- PageRank starts with the simple "voting" based on in-links
- Nodes repeatedly pass endorsements across their out-going links, with the weight of a node's endorsement based on the current estimate of its PageRank
  - More important nodes make stronger endorsements

# Computing PageRank

- Can view it as a process of PageRank "flowing" from pages to the pages they link to.

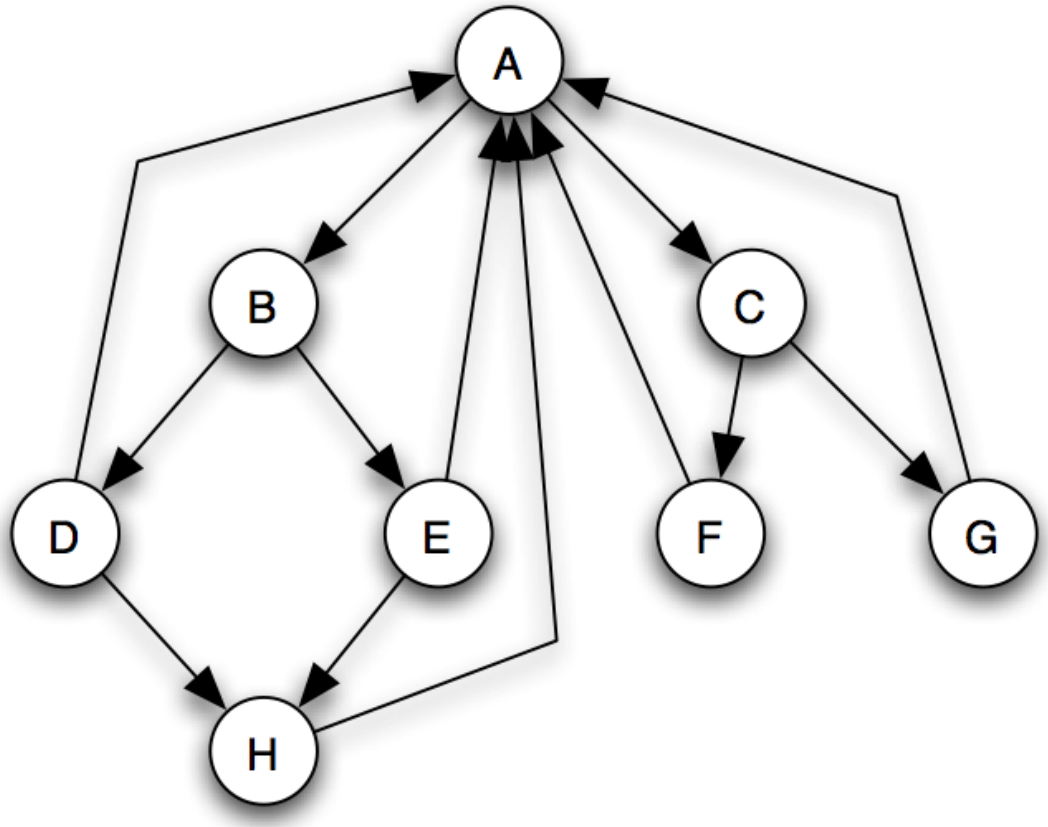


# Basic PageRank computation

- Iterate rank-flowing process until convergence:
- Let $S$ be the set of pages
- Initialize $R(A) = \frac{1}{|S|} = \frac{1}{n}$ for all $A \in S$
- Until ranks do not change (convergence)
  - For each $A \in S$ :

$$R(A) = \sum_{B \to A} \frac{R(B)}{out(B)}$$

# PageRank Algorithm - Example




What are the PageRank values after the first two updates?

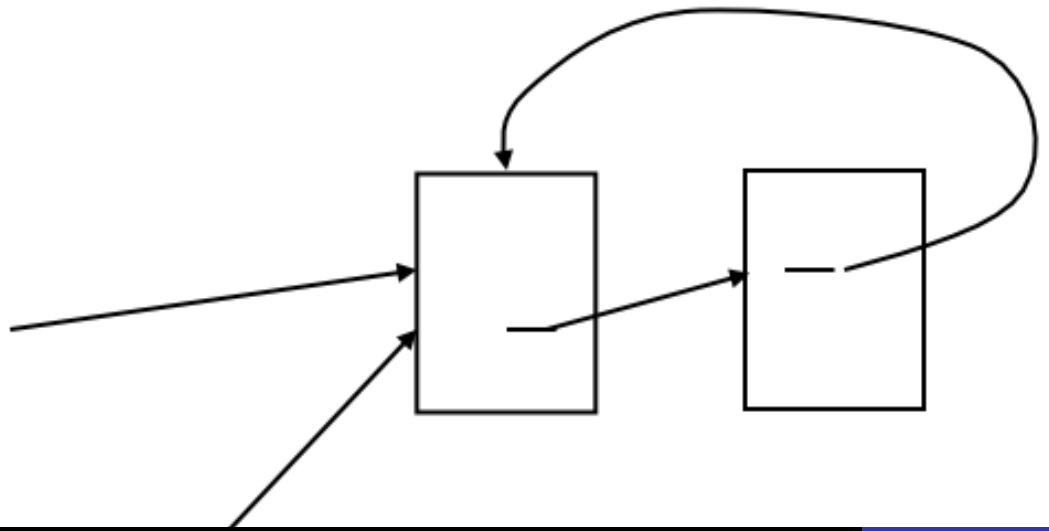
# PageRank Algorithm - Example

## Result

| Step | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 1/2 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/8 |
| 2 | 3/16 | 1/4 | 1/4 | 1/32 | 1/32 | 1/32 | 1/32 | 1/16 |

# Computational problems

- The web is full of dead-ends
  - A group of pages that only point to themselves but are pointed to by other pages act as a "rank sink" and absorb all the rank in the system
  - Pages with no outgoing links
- The underlying graph is modified by adding probabilistic jumps to nodes to handle these problems
  - Can show convergence of the iterative PageRank computation process on this graph



Rank flows into cycle and can't get out

# Teleporting

- At a dead end, jump to a random web page
- At any non-dead end, with probability 15%, jump to a random web page
- With remaining probability (85%), go out on a random link
  - 15% - the $\epsilon$ parameter

$$R(A) = \frac{\epsilon}{n} + (1 - \epsilon) \sum_{(B,A) \in G} \frac{R(B)}{out(B)}$$

- Result of teleporting: it cannot get stuck locally

# The PageRank Algorithm

- Let $S$ be the total set of pages and $n = |S|$
- Choose $\epsilon$ s.t. $0 < \epsilon < 1$, e.g., 0.15
- Initialize $R(A) = \frac{1}{n}$ for all $A \in S$
- Until ranks do not change (convergence)
  - For each $A \in S$ :

$$R(A) = \left[ (1 - \epsilon) \sum_{B \to A} \frac{R(B)}{out(B)} \right] + \frac{\epsilon}{n}$$

# The Random Surfer Model

- PageRank can be seen as modeling a "random surfer" that starts on a random page and then at each point:
  - With probability $\frac{\epsilon}{n}$ randomly jumps to page A
  - Otherwise, randomly follows a link on the current page
- R(A) models the probability that this random surfer will be on page A at any given time
- "Jumps" are needed to prevent the random surfer from getting "trapped" in web sinks with no outgoing links

# Speed of Convergence

- Early experiments on Google used 322 million links
- PageRank algorithm converged (within small tolerance) in about 52 iterations
- Number of iterations required for convergence is empirically $O(logn)$ (where $n$ is the number of links)
- Therefore calculation is quite efficient

# PageRank vs. HITS

- Computation
  - Once for all documents and queries (offline)
- Query-independent
  - Requires combination with query-dependent criteria

- Computation
  - Requires computation for each query
- Query-dependent
- Quality depends on quality of start set
- Gives hubs as well as authorities

# Ranking considerations in Search Engines

- Paid advertisers
- Manually-created classification
- Vector-space ranking with corrections for document length
- Feedback from Query Logs
- Extra weighting for specific fields, e.g., title, anchors, etc.
- Popularity or importance, e.g., PageRank
  Many of these factors are NOT made public.

# Conclusions

- Link analysis uses information about the structure of the web graph to aid search
- Using link structure can be viewed as one of the major innovations in web search
- It is the primary reason for Google's success