COMP4650/6490 Document Analysis

Web Search Basics

ANU School of Computing

Australian National University

Outline

- Web Basics
  - The web & other networks
  - Nodes & edges
- Link Analysis
  - What is link analysis
  - Citation analysis
  - Authorities & Hubs
  - The HITS algorithm
  - PageRank

Australian National University

Outline

- Web Basics
  - The web & other networks
  - Nodes & edges
- Link Analysis
  - What is link analysis
  - Citation analysis
  - Authorities & Hubs
  - The HITS algorithm
  - PageRank

Web Basics

The Web & Other Networks

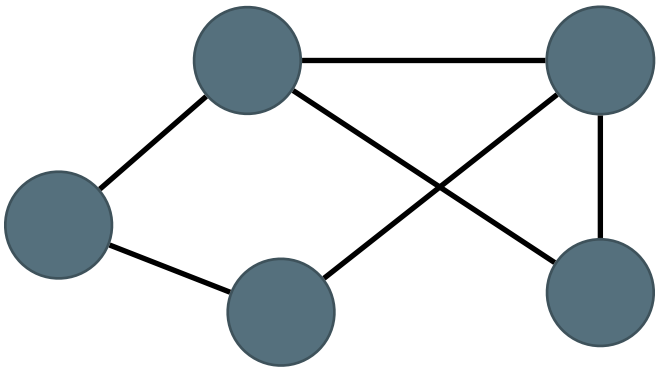
- Documents on the web are linked by hyperlinks
- Academic papers are linked by citations and co-authorship relations
- Legal documents are linked by citations
- Online users are linked by interactions or formal social network ties

Web Basics

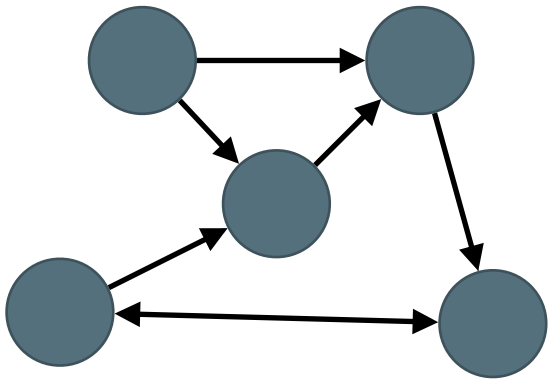
Nodes & Edges

- Nodes represent entities
  - e.g. documents, people, organisations
- Edges are relationships between nodes
  - e.g. hyperlinks, co-authorship relations
- Edges can be directed (go in a specific direction)

Undirected



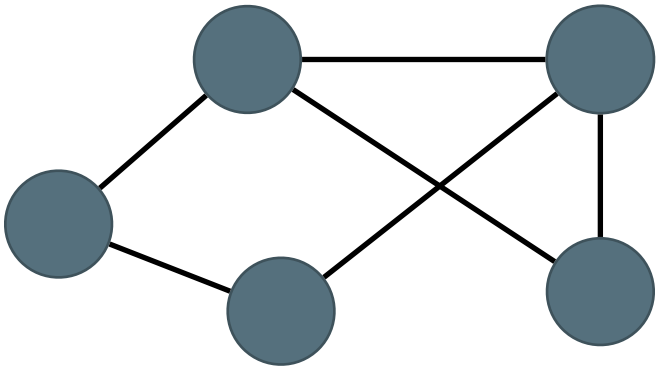
Directed

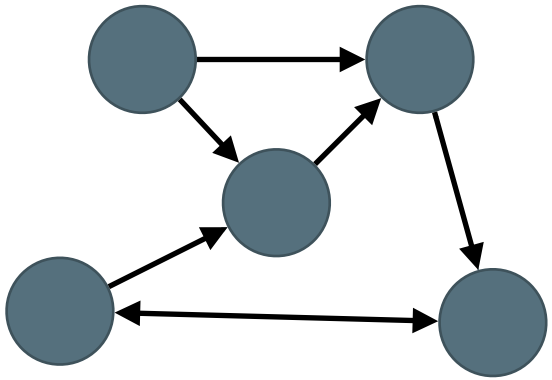


Web Basics

Degree of a Node

- Degree
  - The number of edges connected to a node
- In-degree
  - The number of edges going to a node
- Out-degree
  - The number of edges coming from a node

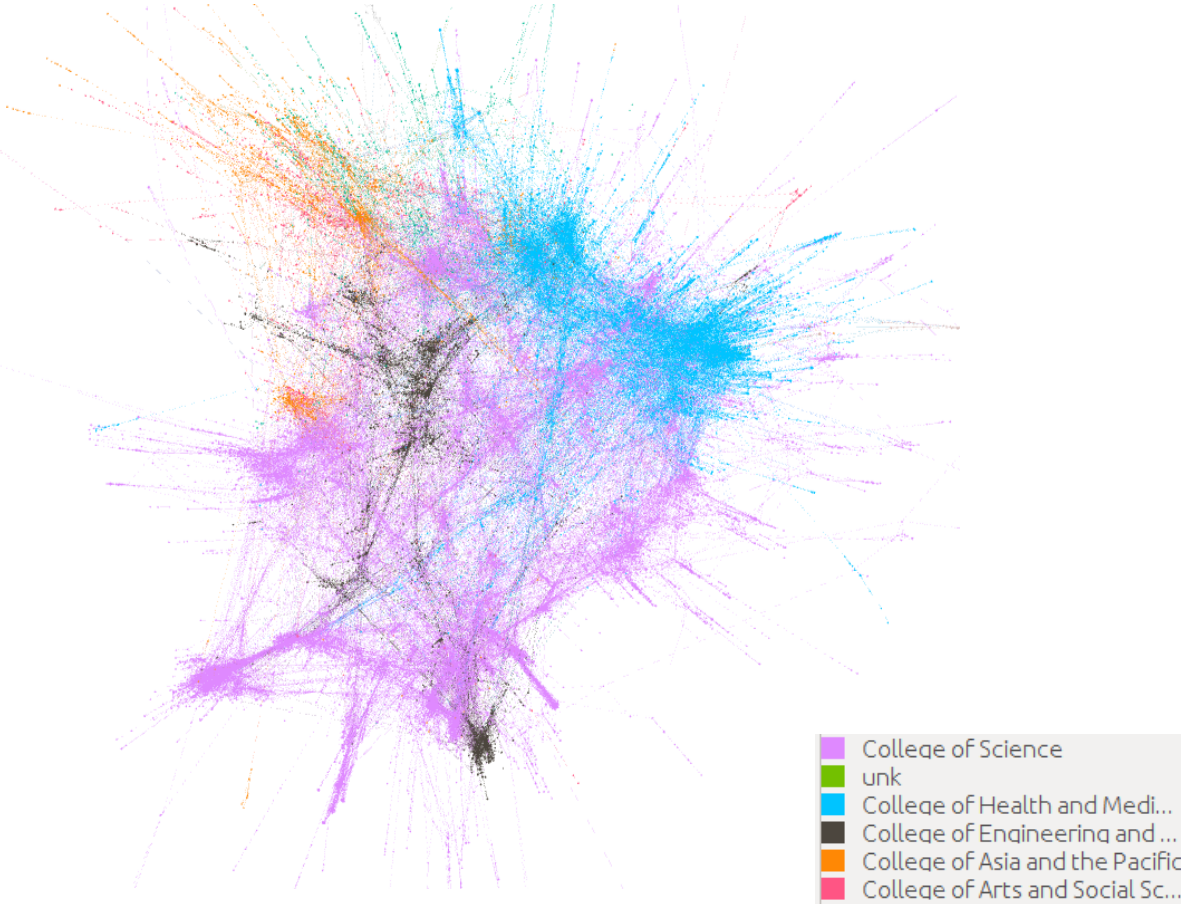




Web Basics

Example: Co-authorship at the ANU

- Nodes are people
- There is an edge between two people if they co-authored a paper
- Note: only part of the network is shown



Australian National University

Outline

- Web Basics
  - The web & other networks
  - Nodes & edges
- Link Analysis
  - What is link analysis
  - Citation analysis
  - Authorities & Hubs
  - The HITS algorithm
  - PageRank

Link Analysis

What is Link Analysis

- Link analysis uses information about the *structure of the web graph* to aid search
- It is one of the *major innovations* in web search
- It was one of the primary reasons for Google's initial success

Link Analysis

Citation Analysis

Bibliometrics

- Many documents include *bibliographies* with citations to previously published documents
- Using citations as edges, a collection of documents can be viewed as a graph
- The structure of this graph can provide interesting information such as the *similarity of documents* even when document content is ignored

Link Analysis

Citation Analysis

Impact Factor of a Scientific Journal

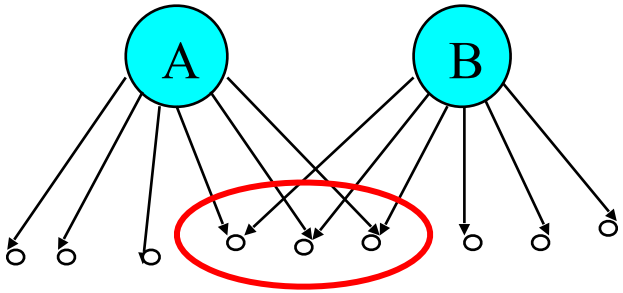
- Developed by Garfield in 1972 to measure the importance (e.g. quality, influence) of scientific journals
- A measure of how often papers in a journal are cited
- Computed and published annually by the Institute for Scientific Information (ISI)
- The *impact factor* of a journal *J* in year *Y* is the average number of citations (from indexed documents published in year *Y*) to a paper published in *J* in year *Y−1* or *Y−2*
- Does not account for the quality of the citing article

Link Analysis

Citation Analysis

Bibliographic Coupling

- Measure of similarity of documents introduced by Kessler in 1963
- The *bibliographic coupling* of two documents *A* and *B* is the number of documents cited by both *A* and *B,* i.e. size of the *intersection* of their bibliographies
- Maybe want to normalise by size of bibliographies?

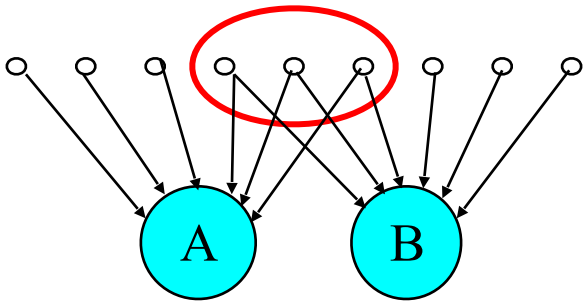


Link Analysis

Citation Analysis

Co-Citation

- An alternative citation-based measure of similarity introduced by Small in 1973
- Co-citation is the number of documents that cite both *A* and *B*
- Maybe want to normalise by the total number of documents citing either *A* or *B*?



Link Analysis

Citation Analysis

Web links are a bit different than citations:

- Many links are *navigational*
- Many pages with high in-degree are portals not content providers
- Not all links are endorsements
- Company websites don't point to their competitors
- But citations to relevant literature is enforced by peer-review

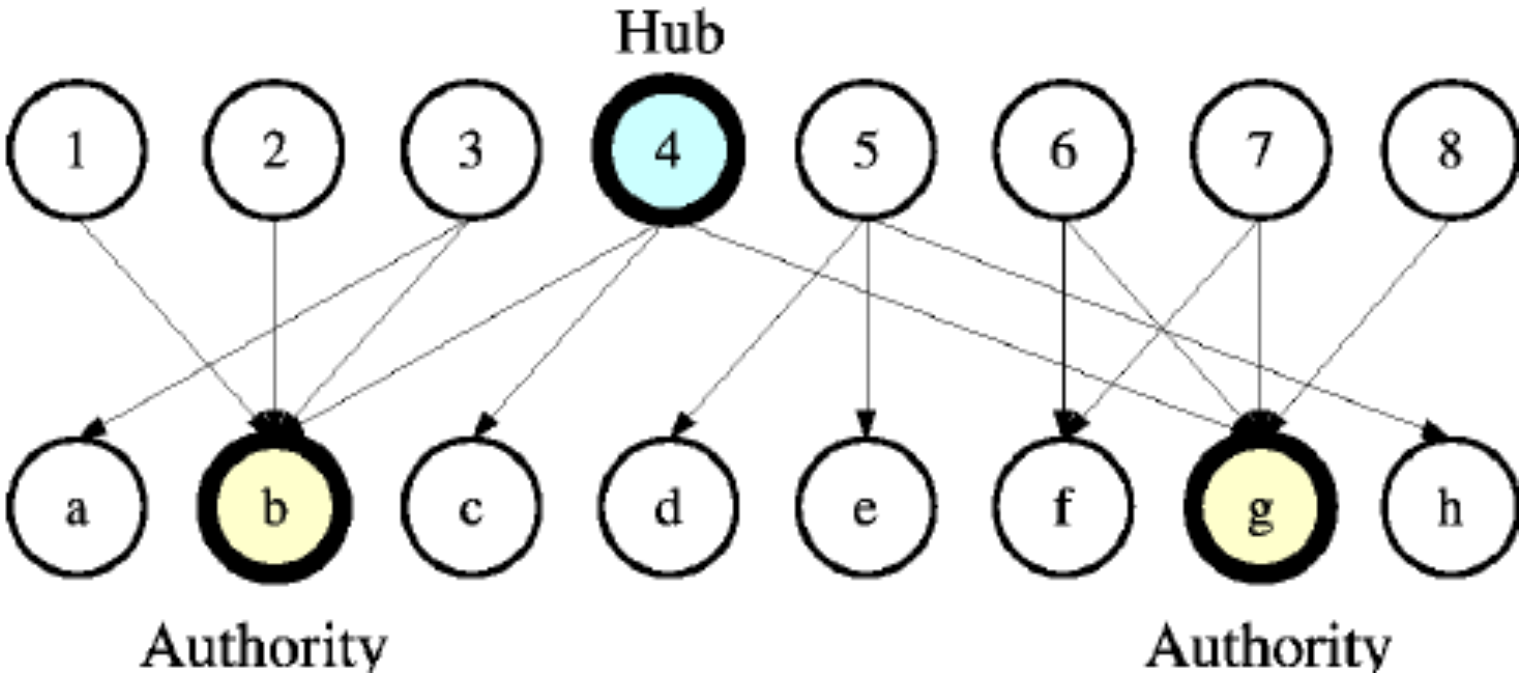
Link Analysis

Authorities & Hubs

- *Authorities* are pages that are recognised as providing significant, trustworthy, and useful information on a topic
  - In-degree (number of pointers to a page) is one simple measure of authority, but in-degree treats all links equally
  - Should links from pages that are themselves authoritative count more?
- *Hubs* are index pages that provide lots of useful links to relevant content pages (i.e. Authorities)

Link Analysis

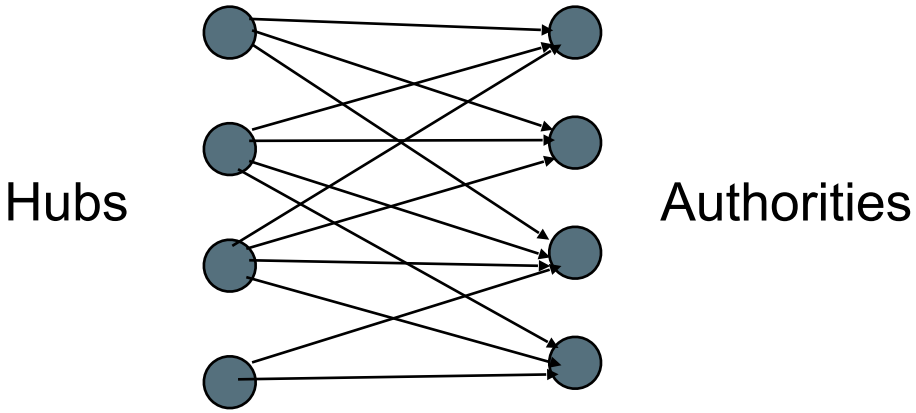
Authorities & Hubs



Link Analysis

HITS Algorithm

- Hyperlink-Induced Topic Search (HITS) algorithm determines *Hubs* and *Authorities* for a particular topic through analysis of a *relevant subgraph*
- Algorithm developed by Jon M. Kleinberg in 1998
- Based on mutually recursive assumptions:
  - *Hubs* point to lots of *Authorities*
  - *Authorities* are pointed to by lots of *Hubs*

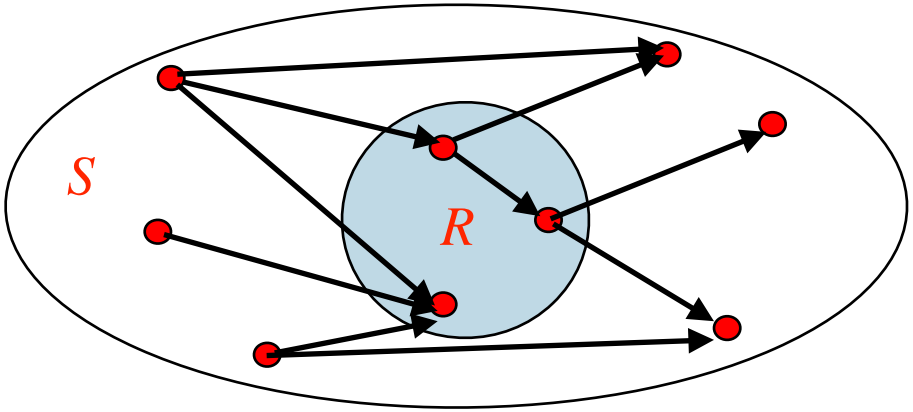


Link Analysis

HITS Algorithm

Constructing a relevant subgraph S

- For a specific *query* which specifies the *topic*, let the *root set* R be the set of pages returned by a standard search engine
- Initialise the subgraph S using all pages in R
- Add page p to S if $\exists q \in R,\ p \rightarrow q$
- Add page p' to S if $\exists q \in R,\ q \rightarrow p'$



Link Analysis

HITS Algorithm

Use an *iterative* algorithm to slowly converge on a mutually reinforcing set of *Hubs* and *Authorities*:

- Maintain for each page $p \in S$
  - Authority score a_p
  - Hub score h_p
- Initialise $a_p = h_p = 1, \forall p \in S$
- Maintain normalised scores

$$\sum_{p \in S} a_p^2 = 1, \quad \sum_{p \in S} h_p^2 = 1$$

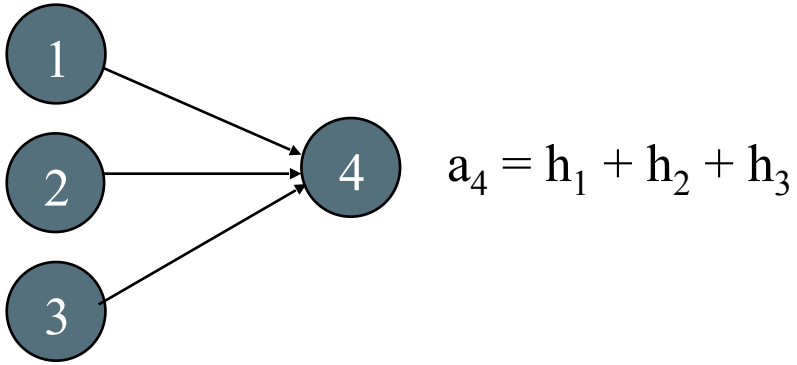
Link Analysis

HITS Algorithm

Update Rules:

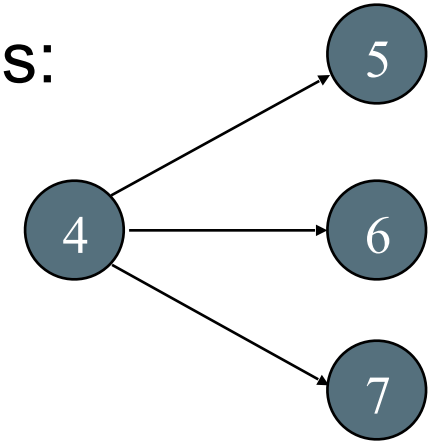
- Good Authorities are pointed to by good Hubs:

$$a_q = \sum_{p:\ p \to q} h_p$$



- Good Hubs point to good Authorities:

$$h_p = \sum_{q:\ p \to q} a_q$$



Link Analysis

HITS Algorithm

The Iterative Algorithm

- Initialise: $a_p = h_p = 1, \forall p \in S$
- For *t* = *1* to *T*
  - Update authority/hub scores

$$a_q = \sum_{p:\, p \to q} h_p, \forall q \in S$$

$$h_p = \sum_{q:\, p \to q} a_q, \forall p \in S$$

  - Normalise authority/hub scores

$$a_q = \frac{a_q}{\sqrt{\sum_{q \in S} a_q^2}}, \forall q \in S$$

$$h_p = \frac{h_p}{\sqrt{\sum_{p \in S} h_p^2}}, \forall p \in S$$

Link Analysis

HITS Algorithm

Convergence

- The algorithm converges to a *fix-point* if iterated indefinitely
- Let A be the adjacency matrix for the subgraph
  i.e. $A_{ij} = \begin{cases} 1, & \text{if } i \rightarrow j \text{ for } i, j \in S \\ 0, & \text{otherwise} \end{cases}$
- Authority vector $\mathbf{a}$ converges to the principal eigenvector of $A^\top A$ (scaled by a constant)
- Hub vector $\mathbf{h}$ converges to the principal eigenvector of $AA^\top$ (scaled by a constant)
- In practice, 20 iterations produces fairly stable results

Link Analysis

HITS Algorithm

- Example: Authorities for query “Java”
  - java.sun.com
  - comp.lang.java FAQ
- Example: Authorities for query “search engine”
  - Yahoo.com
  - Excite.com
  - Lycos.com
  - Altavista.com
- Example: Authorities for query “Gates”
  - Microsoft.com
  - roadahead.com

Link Analysis

HITS Algorithm

Finding Similar Pages Using Link Structure

- Given a page q, let the root set R be the k (e.g. 200) pages that point to q
- Add page p to the subgraph if $\exists q \in R,\ p \rightarrow q$
- Add page p' to the subgraph if $\exists q \in R,\ q \rightarrow p'$
- Run the HITS algorithm on the subgraph
- Return the best *Authorities* in the subgraph as the best similar-pages for q

This approach finds *Authorities* in the "link neighbourhood" of page q

Link Analysis

HITS Algorithm

Finding Similar Pages Using Link Structure

- Example: Given “honda.com”
  - toyota.com
  - ford.com
  - bmwusa.com
  - saturncars.com
  - nissanmotors.com
  - audi.com
  - volvocars.com

Link Analysis

PageRank

- Alternative link analysis method used by Google (Brin & Page, 1998).
- Does not attempt to capture the distinction between hubs and authorities
- Gives each page a score which measures authority
- Applied to the *entire* document corpus rather than a local neighbourhood of pages surrounding the results of a query
- May be understood through Markov theory (random walk on web graph) or network analysis

Link Analysis

PageRank

Initial PageRank Idea

- Initial try for page q:

$$r_q = C \sum_{p:\ p \to q} \frac{r_p}{N_p}$$

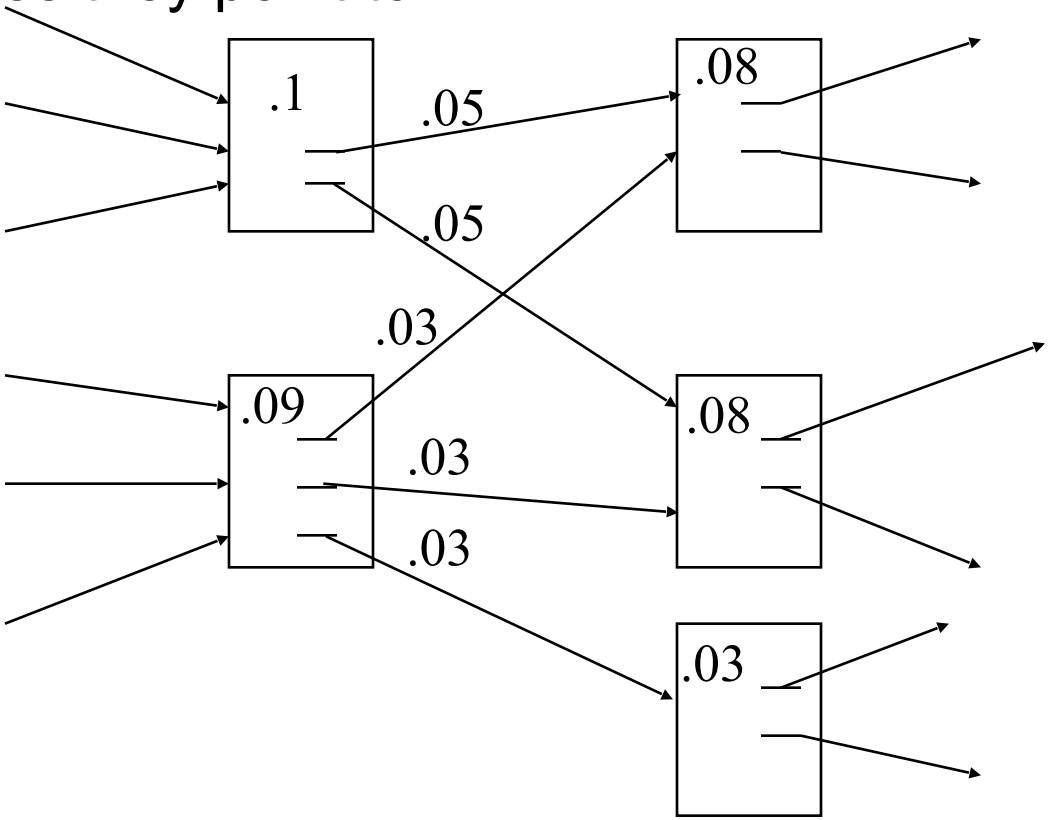
- N_p is the total number of out-links from page p
- A page p "gives" an equal fraction of its authority to all the pages it points to (e.g. q)
- C is a normalising constant such that the rank of all pages always sums to 1

Link Analysis

PageRank

Initial PageRank Idea

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



Link Analysis

PageRank

Initial PageRank Idea

- Initial Algorithm: Iterate rank-flowing process until convergence
- Let S be the set of pages
- Initialise: $r_q = 1/|S|, \forall q \in S$
- Until ranks do not change (much) (i.e. convergence)

Update: $r_q = \sum_{p:\ p \to q} \frac{r_p}{N_p}, \forall q \in S$

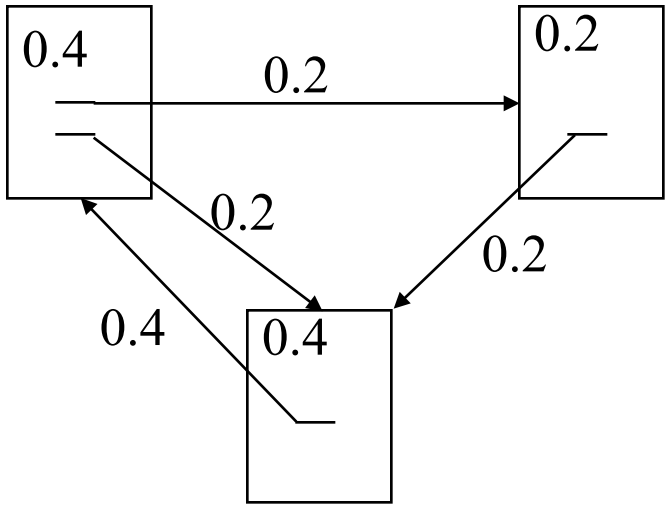
Normalise: $r_q = r_q / (\sum_{q \in S} r_q), \forall q \in S$

Link Analysis

PageRank

Initial PageRank Idea

- Sample stable fix-point

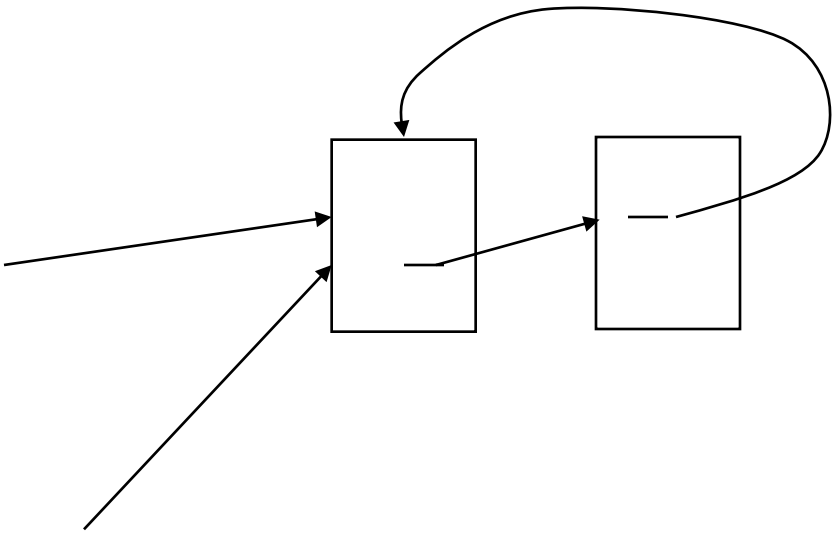


Link Analysis

PageRank

Initial PageRank Idea

- Problem:
  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.



Rank flows into
cycle and can't get out

Link Analysis

PageRank

Rank Source

- Solution:
  Introduce a "*rank source*" that continually replenishes the rank of each page q by a fixed amount e_q

$$r_q = C\left(\sum_{p:\ p\to q}\frac{r_p}{N_p} + e_q\right)$$

Link Analysis

PageRank

PageRank Algorithm

- Let S be the set of pages
- Let $e_q = \alpha / |S|, \forall q \in S$ (for some $0 < \alpha < 1$, e.g. 0.15)
- Initialise: $r_q = 1/|S|, \forall q \in S$
- Until ranks do not change (much) (i.e. convergence)

Update: $r_q = e_q + (1-\alpha) \sum_{p:\ p \to q} \frac{r_p}{N_p}, \forall q \in S$

Normalise: $r_q = r_q / (\sum_{q \in S} r_q), \forall q \in S$

Link Analysis

PageRank

Speed of Convergence

- Early experiments on Google used 322 million links
- PageRank algorithm converged (with a small tolerance) in about 52 iterations
- Number of iterations required for convergence is empirically *O(log n)* (where *n* is the number of links)
- Therefore, calculation is efficient.

Link Analysis

PageRank

Google Ranking

- Complete Google ranking includes (based on university publications prior to commercialisation)
  - Vector-space similarity component
  - Keyword proximity component
  - HTML-tag weight component (e.g. title preference)
  - PageRank component
- Details of current commercial ranking functions are trade secrets

Summary

- Web Basics
  - The web & other networks
  - Nodes & edges
- Link Analysis
  - What is link analysis
  - Citation analysis
  - Authorities & Hubs
  - The HITS algorithm
  - PageRank

References

- Chapter 21, Introduction to Information Retrieval
- Some slides are from
  - Raymond J. Mooney, Information Retrieval and Web Search, University of Texas
  - Davide Mottin, Konstantina Lazaridou, Hasso Plattner Institute, Graph Mining course