Information Networks: Hubs and Authorities

Computational Social Systems I (VU) (706.616)

Elisabeth Lex

ISDS, TU Graz

May 28, 2020

Information Networks

- Nodes: piece of information
- Links: join related pieces
- Most prominent example: Web

The Web as a graph

- Nodes: Web pages
- Links: hyperlinks ($< a\ href = "..." > ... < /a >$) that connect Web pages
- Links directed (point only in one direction)
- A technical side note: why directed links and not undirected (bidirectional) on the Web?

The Web as a graph

- Nodes: Web pages
- Links: hyperlinks ($< a\ href = "..." > ... < /a >$) that connect Web pages
- Links directed (point only in one direction)
- A technical side note: why directed links and not undirected (bidirectional) on the Web?
  - Links consistency does not scale (deleting a target doc would require to update all the source docs)
  - Tim Berners-Lee: "Let the links fail to make them scale"

Why node-link metaphor?

- Application of computer-aided authoring style known as *hypertext*
- Dates back to 1940'ies
- Idea: replace linear structure of text with graph structure
- Any portion of text can link to any other portion of text in an associative manner
- Tim Berners-Lee created the Web in early 90'ies by simplifying and combining this idea with distributed networked computer system (Internet)

Historical side note

- 1945 Vannevar Bush *"As We May Think"* [1]
- *Memex*: Digital library - knowledge management system, extends human brain
- Recording information with microphone, camera
- Index for later retrieval
- Would create trails of links connecting sequences of microfilm frames

[1] www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/

The Web as a directed graph

- Types of links: **navigational links** vs **transactional links**
- Navigational links connect a huge fraction of stable Web pages to other pages
- Transactional links: dynamic, related to particular transaction taking place on the Web, e.g., online payment
- Transactional content, however, linked together by navigational "backbone"

The Web as a directed graph

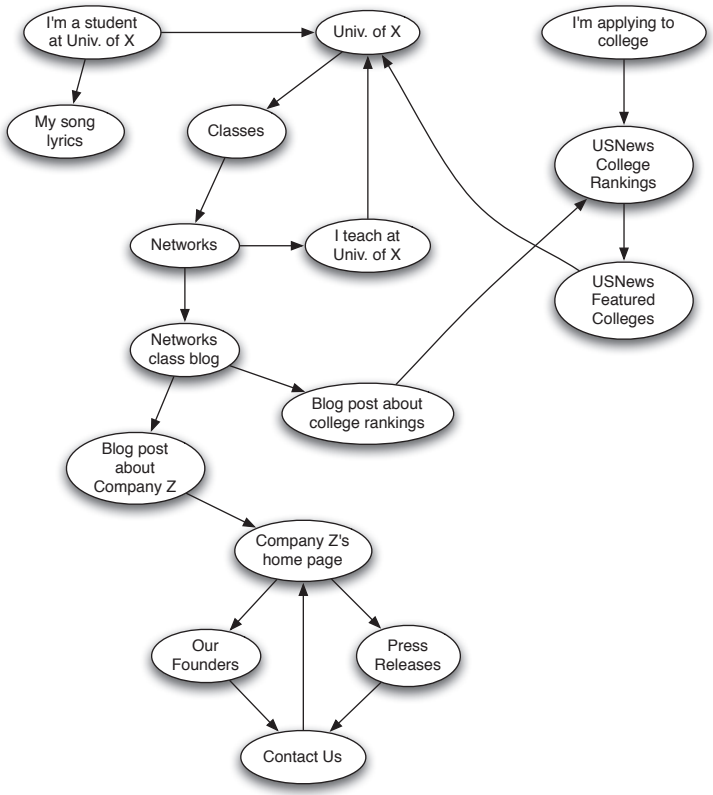
- Main difference between social networks and Web: Web's directed nature
- Directed graphs are asymmetric: links point from one node to another
- Analogy: friendship network versus follower network
- Follower network is asymmetric: e.g. celebrities are followed by millions of people, but they do not follow all of their fans

Paths: Directed graphs

- **Path**: sequence of nodes such that each consecutive pair in sequence connected by link
- If every node can reach every other node by path, graph is **connected**
- Otherwise: disconnected graph - breaks apart into set of connected **components**
- **Component**: subset of nodes such that
  - every node in subset has path to every other node in subset
  - subset is not part of some larger connected set

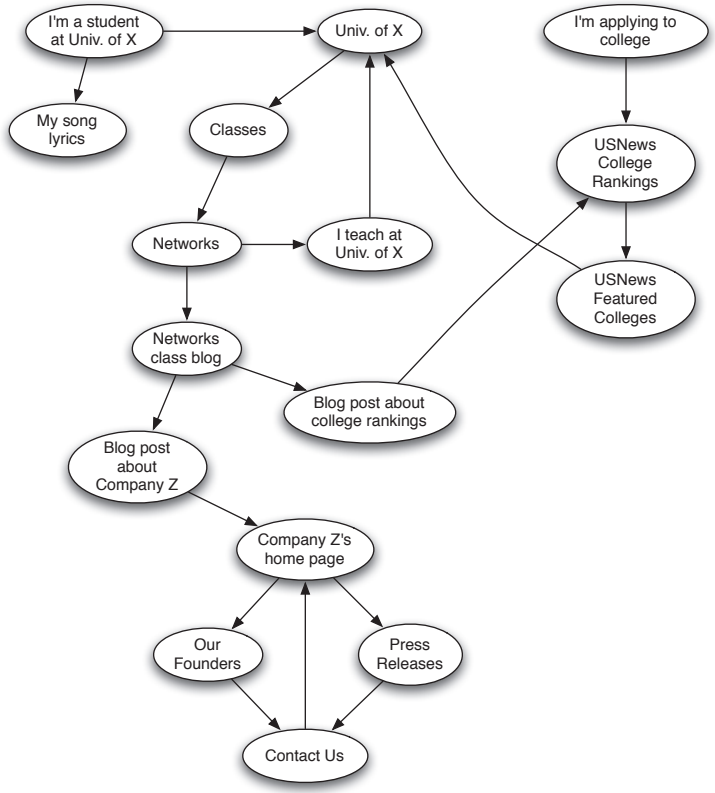
Paths: Directed Graphs

- **Path**: sequence of nodes such that each consecutive pair in sequence is connected by link in the forward direction
- E.g. sequence: Univ of X, Classes, Networks, Networks class blog, Blog about college rankings, USNews college rankings rankings



Strong connectivity

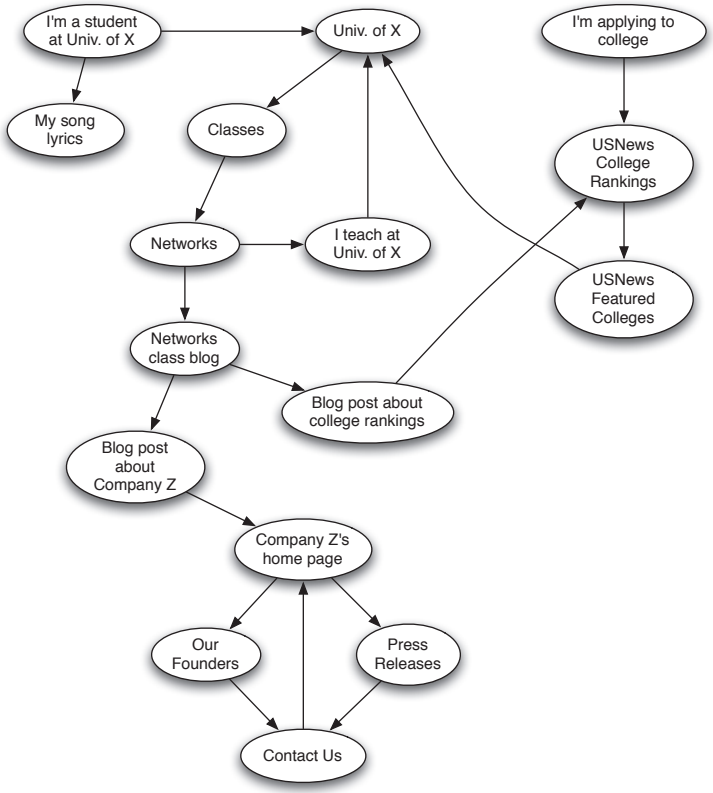
- If every node can reach every other node by (directed) path, graph is **strongly connected**



Is the example graph strongly connected?

Strong connectivity

- If every node can reach every other node by (directed) path, graph is **strongly connected**



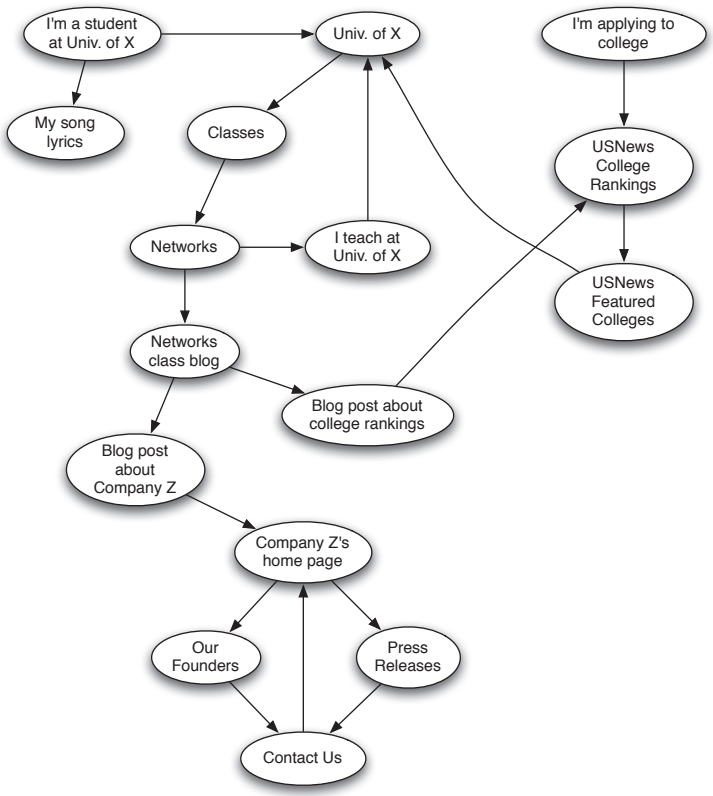
Is the example graph strongly connected? No! No directed path from e.g. Company Z's to USNews college rankings

Reachability

- Which nodes reachable from which other nodes using (directed) paths?
- Undirected graph: if two nodes in the same component - mutually reachable by a path
- Otherwise: if two nodes in different components, they cannot reach each other

Reachability

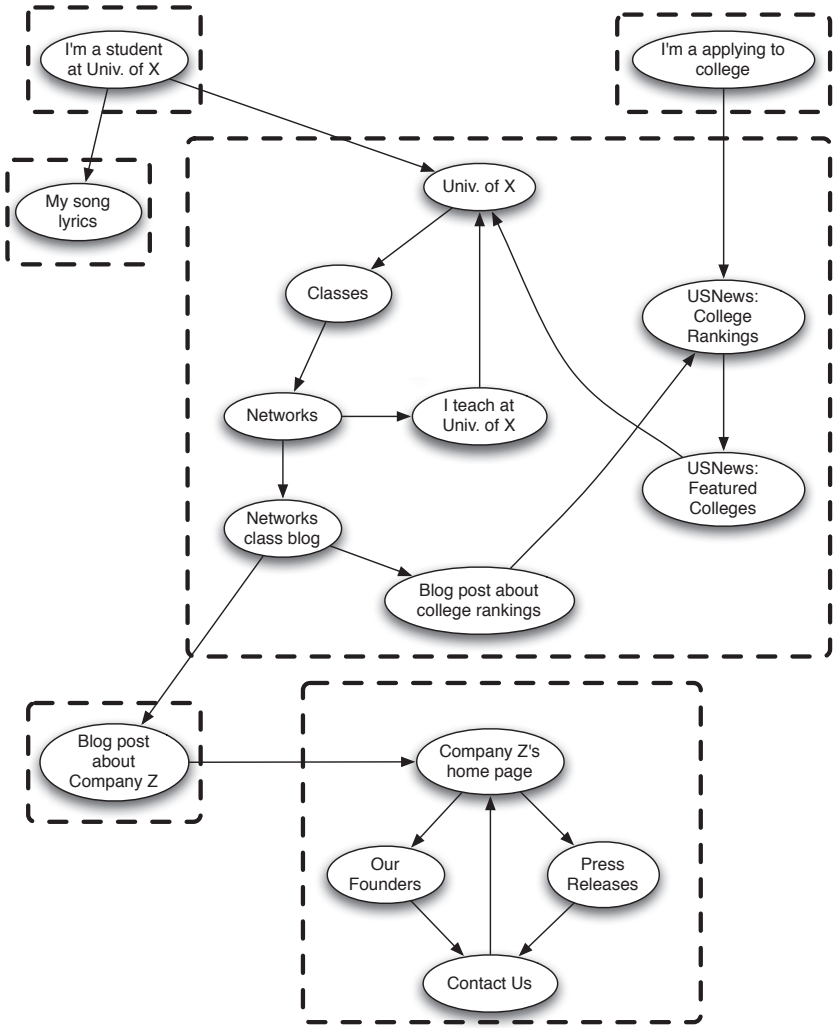
- 3 cases in directed networks:
  1. Pairs of nodes, which can reach each other (Univ. of X and USNews: college rankings)
  2. Pairs of nodes, where one can reach the other but not vice versa (USNews college rankings and Company Z's home page)
  3. Pairs of nodes, which cannot reach each other (I'm a student and I'm applying to college)



Strongly connected components

- **Strongly Connected Component (SCC)**: subset of nodes such that
  - every node in subset has (directed) path to every other node in subset
  - subset not part of some larger set with property that every node can reach every other
- Reachability is key
- Directed graph: SCC does not need to be completely isolated from rest of graph
- SCCs summarize a graph in form of "super-nodes"

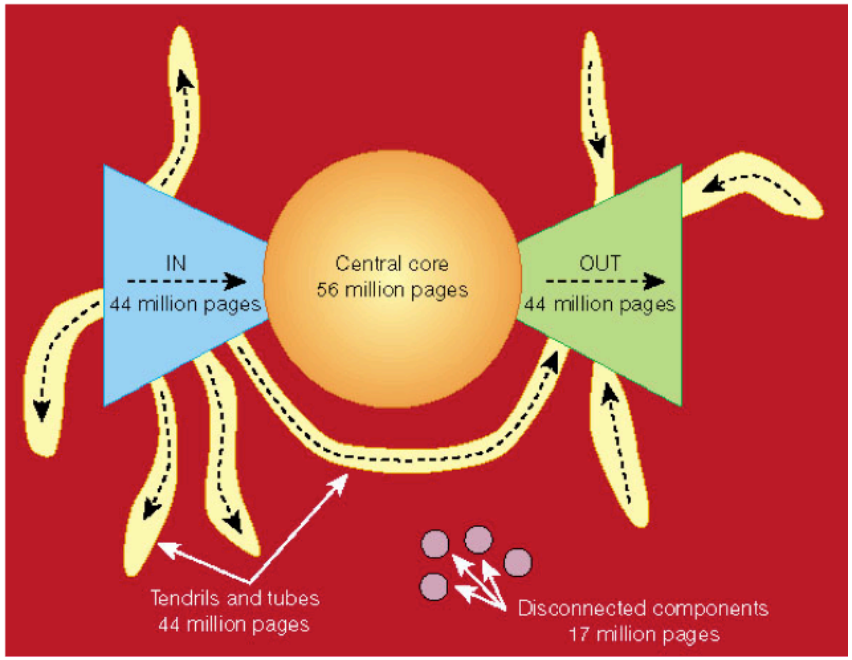
Strongly connected components



Structure of the Web

- Web growth rate in the 90's was very quick
- (Broder et al., 1999) set out to build global map of the Web using strongly connected components (SCC) as basic building blocks
- Dataset: index of pages and links from AltaVista (largest search engine at that time)
- Since then, study has been replicated many times, e.g. larger datasets, subsets of the Web such as Wikipedia, etc.

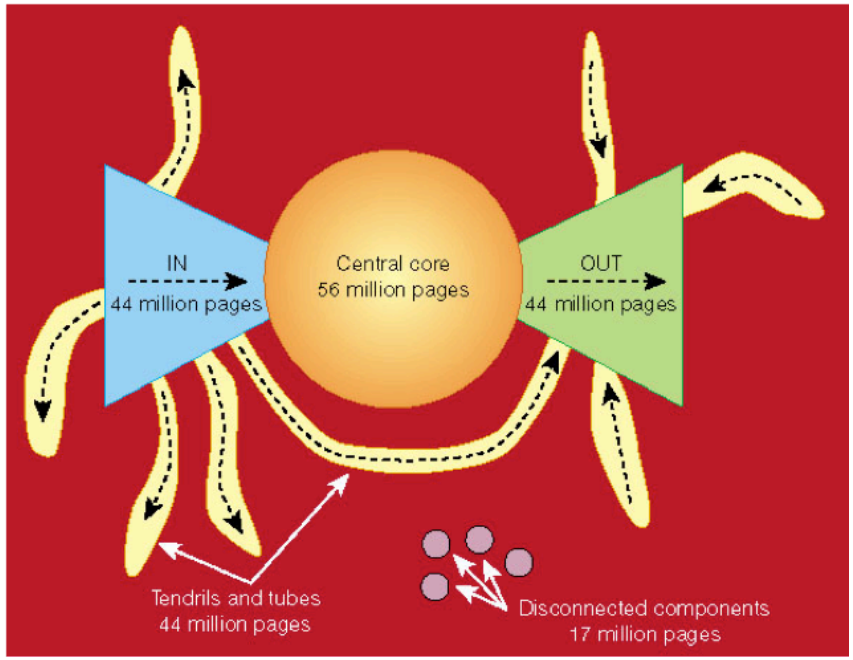
The Map of the Web



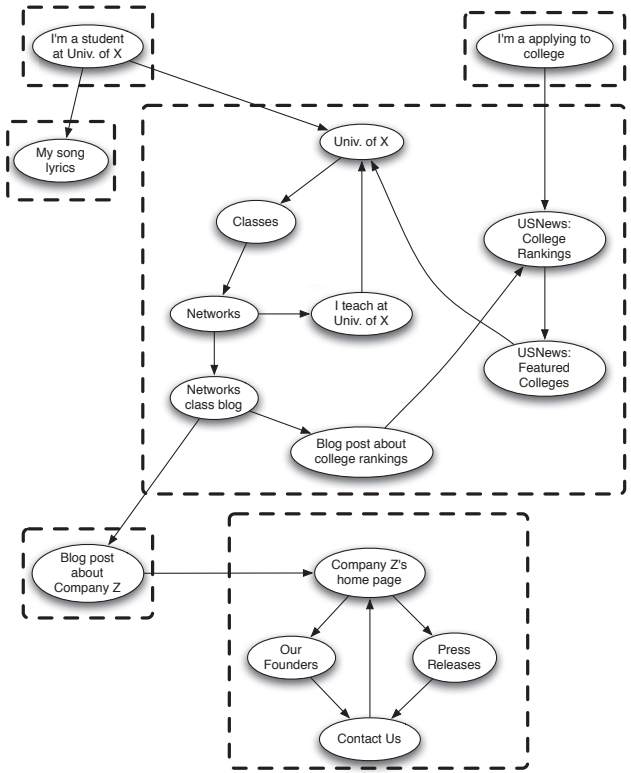
- Web **not** fully interconnected network!
- Web has giant SCC that contains interconnected pages
- Why? Major search engines and "start pages " (e.g. directories) keep links to core
- Reachability within core is good
- Some pages from core link back to search engines
- Suppose two giant SCCs X and Y: single link from X to Y and vice versa would turn X and Y into one single SCC

IN and OUT on the Web

- Now: Position all other components in relation to giant SCC
- Classify nodes by their ability to reach and be reached from giant SCC
- IN: nodes that can reach giant SCC but not vice versa (e.g. new pages that have not yet been linked to)
- OUT: nodes that can be reached from giant SCC but not vice versa (e.g. as corporate websites containing only internal links)



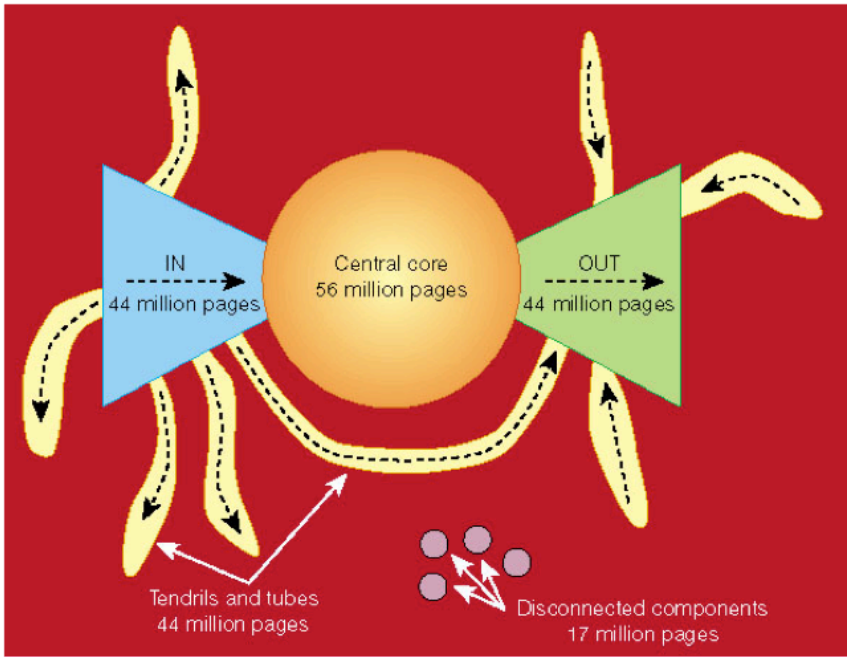
Example: IN and OUT



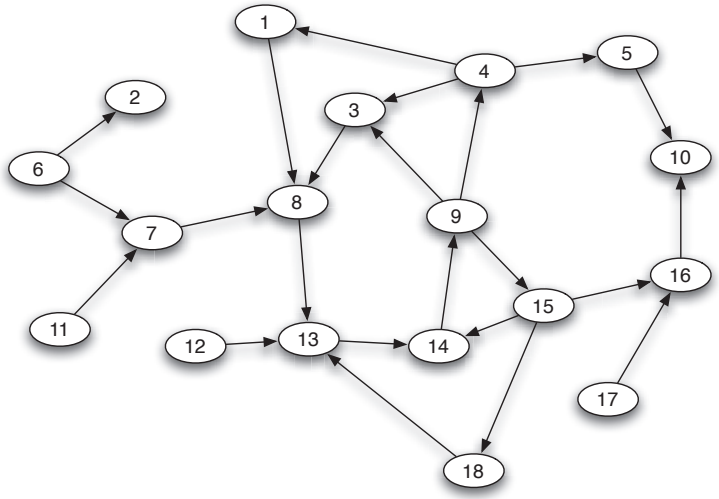
- SCC: largest in the middle of the network including e.g. Univ. of X
- I'm a student and I'm applying to college constitute IN
- Blog post about... and the whole component involving Company Z constitute OUT

Structure of the Web

- Besides: pages that belong to none of IN, OUT, or giant SCC
- *Tendrils and Tubes*: Connect to either IN or OUT, or both, but not to core
- *Disconnected*: Nodes that are disconnected from the SCC even if we ignore link direction



Example



- Name a link (add or delete) so as to increase the size of the largest SCC.
- Name a link (add or delete) so as to increase the size of IN
- Name a link (add or delete) so as to increase the size of OUT

Example

- Describe an example of a graph where removing a single link can reduce the size of the largest SCC by a large number, e.g. 1000 nodes.

- Describe an example of a graph where adding a single link can reduce the size of the OUT by a large number, e.g. 1000 nodes.

Search on the Web and Ranking

- Goal of search: given a query / keywords, find e.g. documents that are relevant and rank them
- Web search is a hard problem - why?
  - Keywords are limited way to express complex information needs
  - Synonymy and polysemy
  - Diversity of Web content and queries
  - Search engine can find and index millions of documents that are relevant to a one-word query
  - However: humans perform the search - can look at only a few of the results

Ranking of search results

- Compute a relevance score
- Score can be a combination of (i) Content score (TF*IDF) and (ii) Popularity score (e.g., HITS, PageRank)

Link Analysis and Ranking

In response to word query "Graz" at Google www.tugraz.at is retrieved in the Top 10 results

- What are the clues that suggest that www.tugraz.at is a good answer to query "Graz" (or a much better answer than 82.7 million of other Web pages about Graz)
- All 82.7 million pages have Graz in the text
- Link perspective: www.tugraz.at stands out
- Very often other pages relevant to Graz link to www.tugraz.at

Link analysis and Ranking

- Links to assess *authority* of page on topic
- Implicit endorsements through links of other pages
- Each individual link may have many possible meanings
- E.g. it may be real endorsement, paid advertisement, off-topic, critique
- Assumption: in aggregate, if page receives many links from other relevant pages, it receives a kind of **collective endorsement**
- I.e., in-links = endorsements, can be exploited for ranking

Voting by in-links

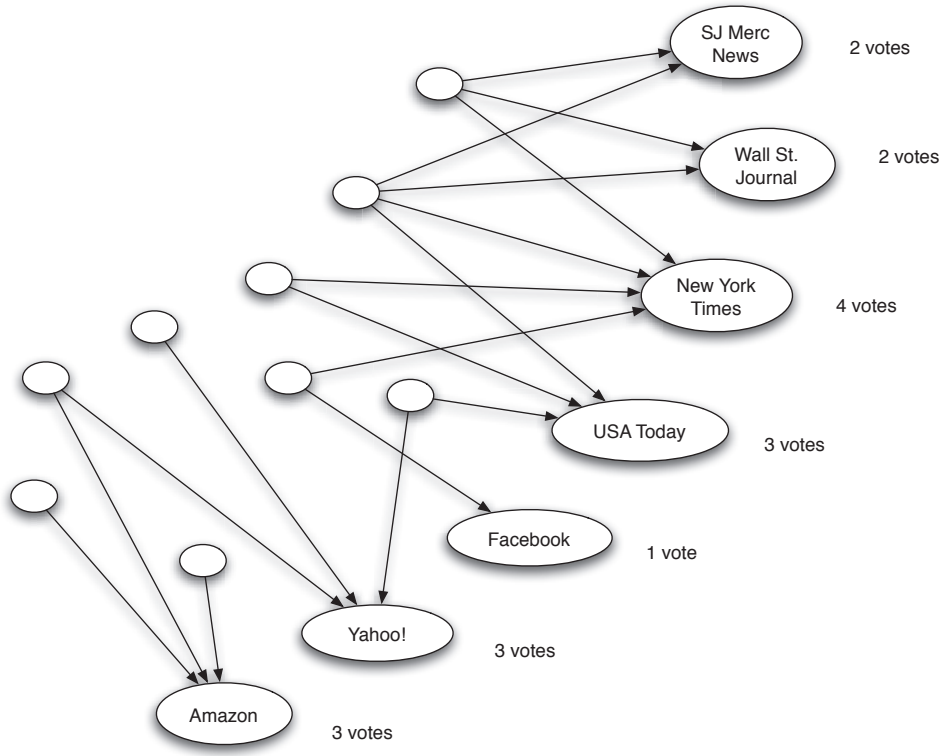
- First: collect a large sample of pages relevant to a query, e.g. Graz
- E.g. by means of classical text-based information retrieval
- Then: let pages in this sample vote through their links
- Then: rank the pages according to the number of votes that they receive

Example: Newspapers

- Consider typical example of a one-word query: newspapers
- No single best answer here but a number of them
- E.g. all prominent newspapers on the Web
- Ideal answer: list of these newspapers

Newspapers example

- First, collect a sample of pages relevant to the query "newspapers"
- Then: count votes to pages within this sample



Newspapers example

- Typically, high scores for a mix of prominent newspapers
- Also, some prominent Web sites (e.g. Facebook) not directly related to newspapers will get a lot of votes
- Can you think of a reason why?

Newspapers example

- Typically, high scores for a mix of prominent newspapers
- Also, some prominent Web sites (e.g. Facebook) not directly related to newspapers will get a lot of votes
- Can you think of a reason why?
- Reason: such pages have a lot of in-links no matter what the query is

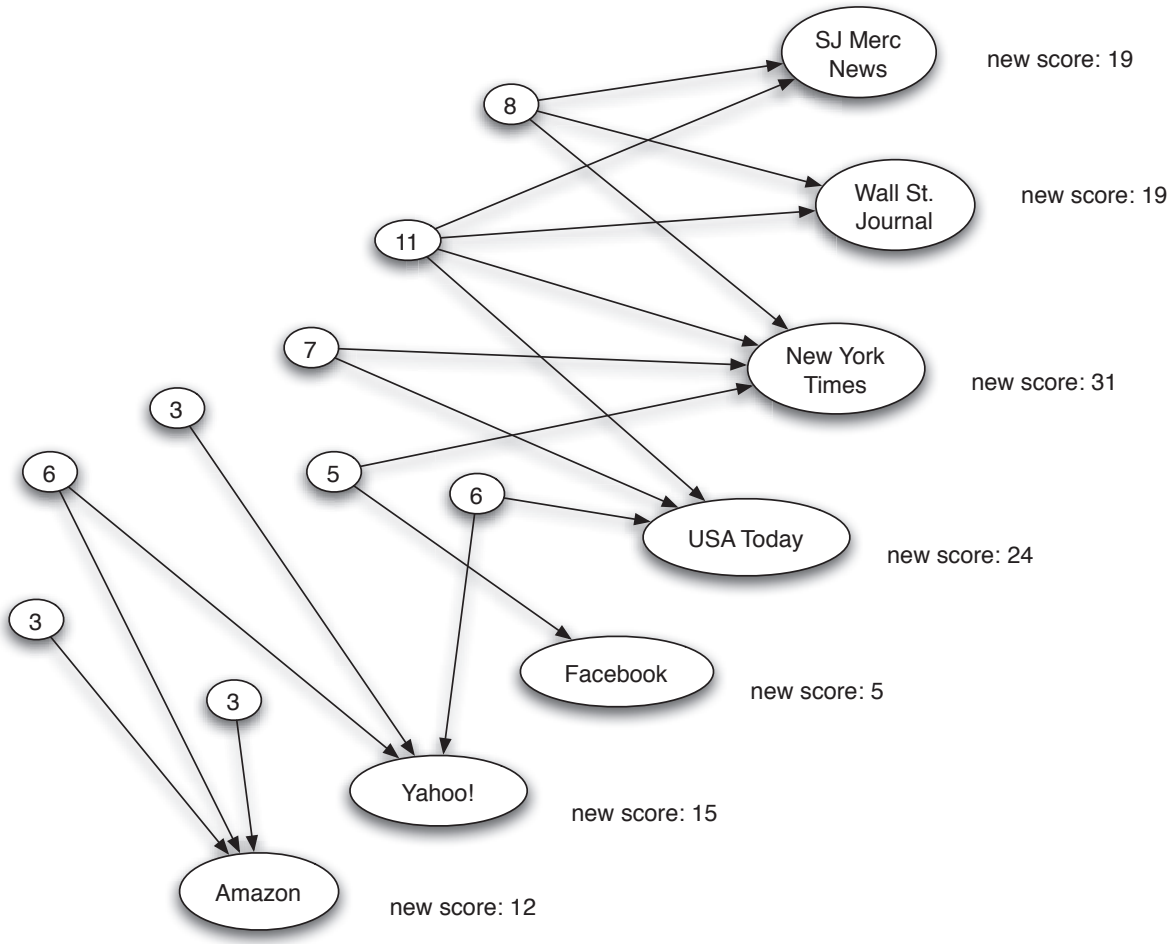
Finding good lists

- Voting by in-links: only a very simple kind of measure
- In addition to most prominent newspapers, also other kinds of useful answers to query
- E.g., pages that compile lists of resources relevant to the topic
- Such lists exist for many broad enough queries (e.g. universities, hotels, newspapers)

Finding good lists

- Some pages vote for many of the pages that receive votes
- Such pages have some sense where good answers are
- Thus, should be scored high as lists
- Page's value as list: equal to sum of votes received by all pages that it voted for

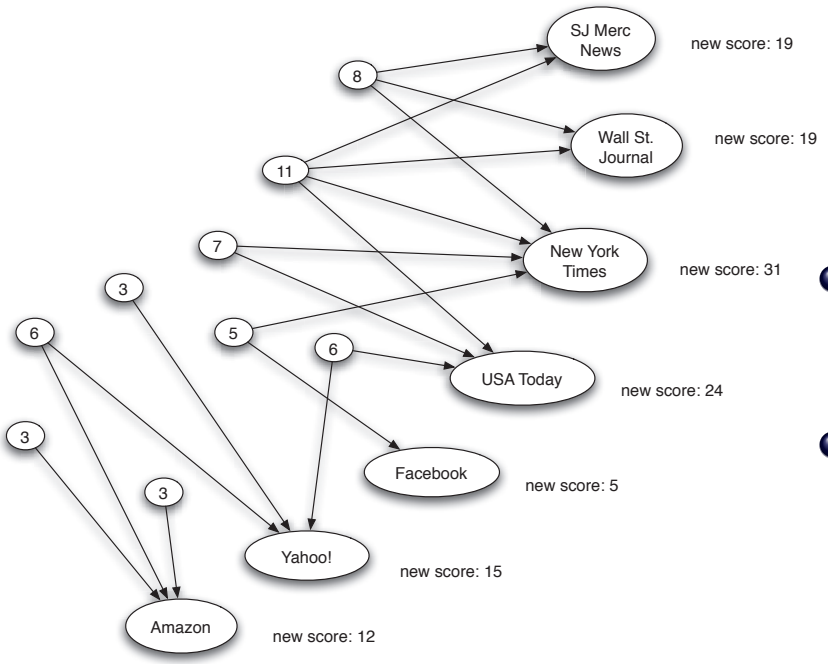
Newspapers example



Repeated Improvement Principle

- Pages scoring well as lists have a better sense for where the good results are
- We should weight their votes more heavily
- Thus, count the votes again
- Give each page's vote a weight equal to its value as a list

Newspapers example



- SJ Merc News, Wall Str. Journal surpass Yahoo and Amazon
- Reason: endorsement by good lists

Repeated improvement

- Re-weighting can continue: use refined votes to refine list scores, then again refine votes/scores, etc.
- Repeat this process for as many steps wanted
- **Principle of Repeated Improvement**: Each refinement to one side of the figure enables a further refinement to the other
- In typical case, all numbers will converge

Hubs and Authorities

Let's formalize:

- **Authorities**: pages that are highly endorsed for a query
- *Hubs*: high value lists
- For each page p, we its estimate authority score $auth(p)$ and hub $hub(p)$ score
- Initially: $auth(p) = hub(p) = 1$

Update authority values step

- Voting procedure: use scores of hubs to refine estimates for scores of authorities

- **Authority Update Rule**: For each page p, update $auth(p)$ score to be the sum of the $hub(q)$ scores of all pages q that point to it

Update hub values step

- List-finding procedure: use scores of authorities to refine estimates for the scores of hubs

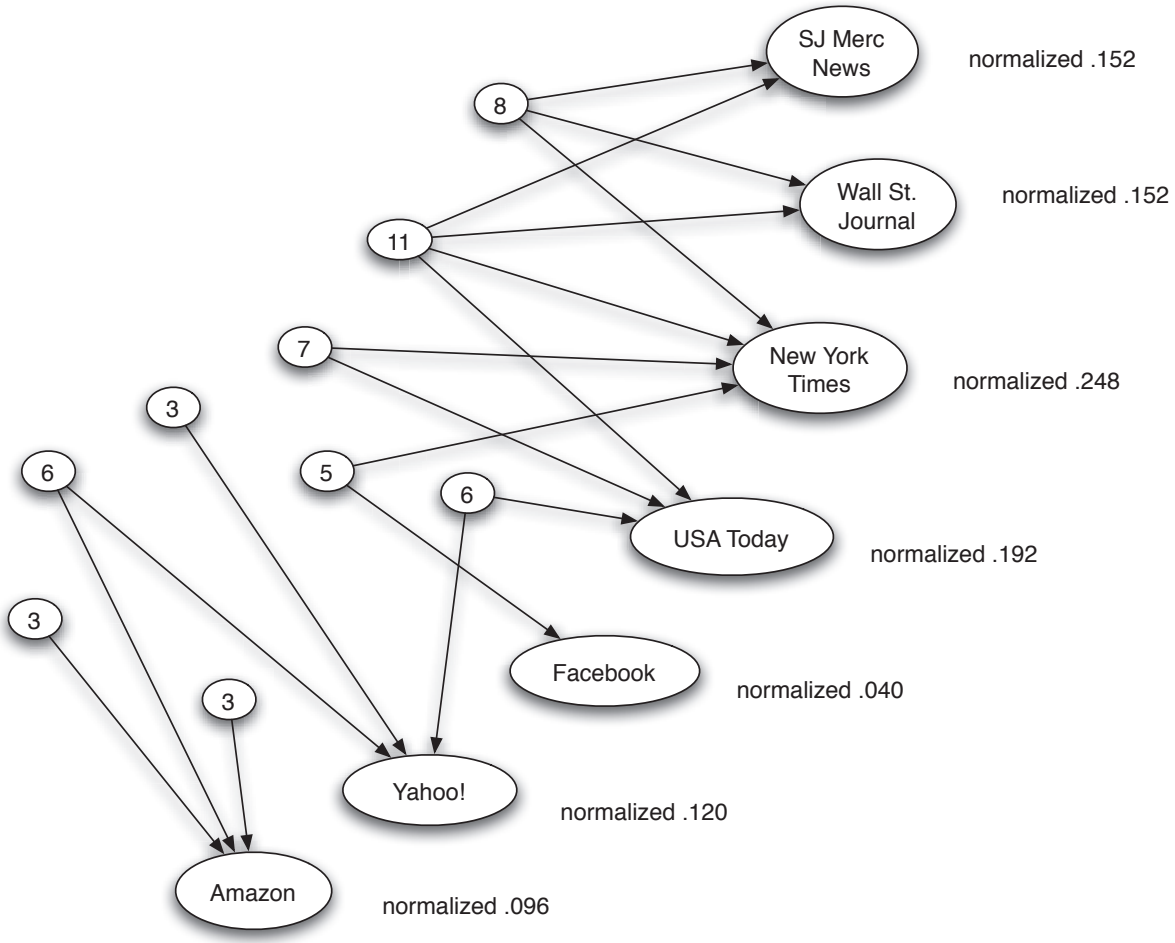
- *Hub Update Rule*: For each page p, update $hub(p)$ score to be the sum of the $auth(q)$ scores of all pages q that it points to

Repeated improvement step

- Choose number of steps k
- Perform sequence of k hub-authority updates, where in each update:
  1. Apply Authority Update Rule to current set of scores
  2. Apply Hub Update Rule to resulting set of scores
- At the end, since hub and authority scores may have very large numbers - normalize them to probability distribution[2]

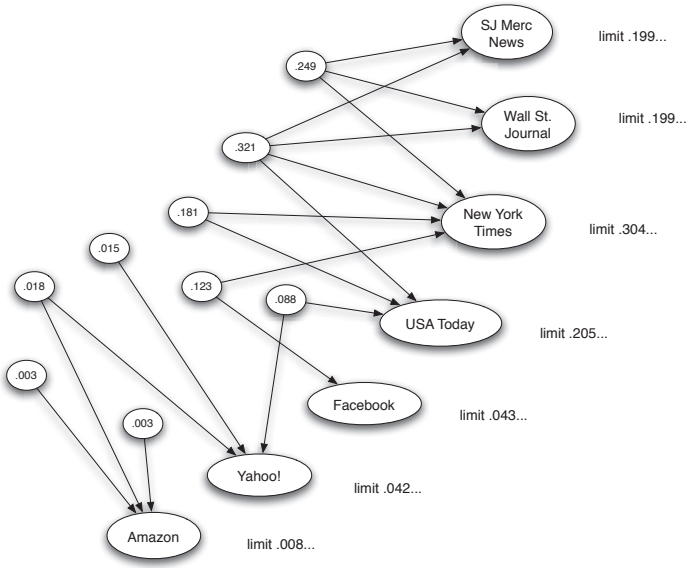
[2] divide each authority score by sum of all authority scores, same with hub score

Newspapers example



Result of normalizing authority scores (dividing by sum of authority scores), e.g. first page: 19/125=0.152

Newspapers example



- Normalized values converge to limits as k goes infinity
- Same limits even if other initial hub/authority values
- Page's authority score proportional to hub scores of pages that point to the page and vice versa

Hubs and authorities: Recap

- Intuition behind hubs and authorities based on idea that pages play multiple roles on the Web
- Pages (hubs) can strongly endorse other pages without themselves being heavily endorsed
- Real-life example: E.g. competing firms will not link to each other
- The only way to pull them together is through hubs that link to all of them at once

Hypertext Induced Topic Selection (HITS) Algorithm (Kleinberg, 1999)

What we discussed so far is the basis for the HITS algorithm:

- Idea: web page has two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic
- 2 ways to categorize web page: authority on the subject and delivers good links to other good authorities
- Authority value: Sum of scaled hub values that point to that page
- Hub value: Sum of scaled authority values of the pages it points to
- Both are defined recursively

Summary

- Information Networks
- How to rank by importance of pages
- HITS algorithm

Take Away

On the Web, there is intrinsic information available that we can exploit to determine how to rank pages using automated methods that look at the Web itself. This can happen without external sources of information.

Thanks for your attention

elisabeth.lex@tugraz.at

Slides use figures from Chapter 13 and Chapter 14 of Networks, Crowds and Markets by Easley and Kleinberg (2010)