Network Analysis: PageRank

Denis Helic

May 16, 2023

Outline

Network Analysis

Networks

- *Social networks*. Nodes are people and links are acquaintances, friendship, and so on.
- *Communication networks*. Internet: nodes are computers and links are cables connecting computers
- *Biological networks*. Metabolism: nodes are substances and links are metabolic reactions
- *Information networks*. Web: nodes are Web pages and links are hyperlinks connecting pages

Networks

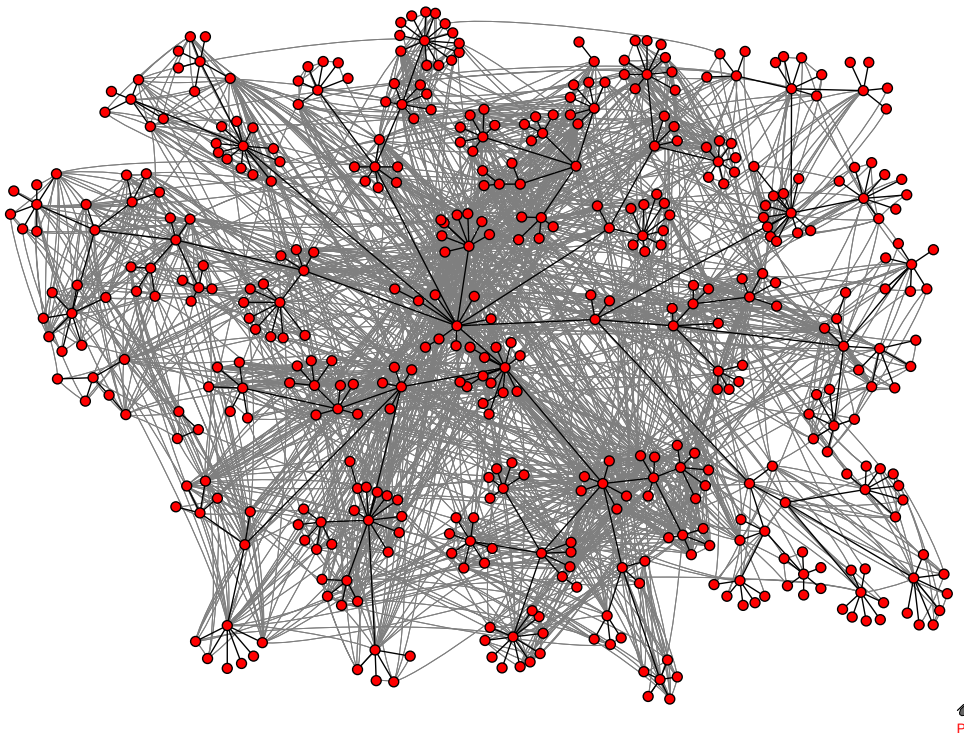


Figure: Social network of HP Labs constructed out of e-mail communication.
From: How to search a social network, Adamic, 2005.

Networks

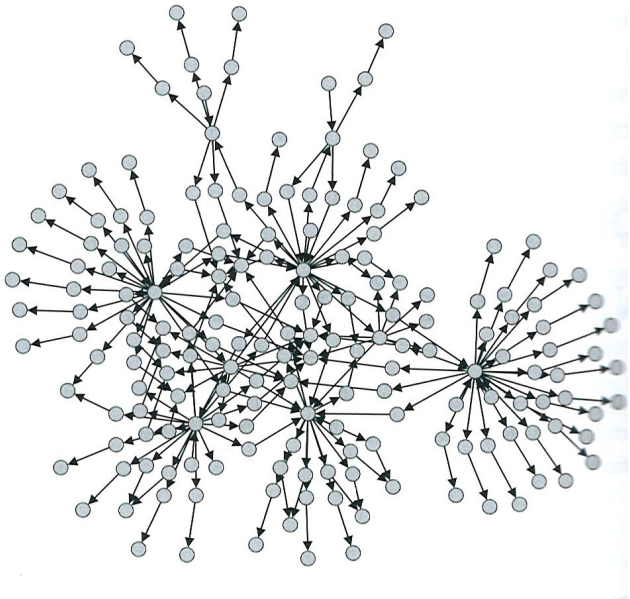


Figure: Network of pages and hyperlinks on a Website. From: Networks, Mark Newman, 2011.

Network analysis: what is it all about?

- Some phenomena are best understood through relations between individuals
- For example, human communication
- We do not investigate the individuals but the processes that emerge through their interaction with each other
- We are considering large-scale statistical properties of graphs
- Network analysis deals with the empirical analysis of large graphs (networks) that occur in different areas
- Using some of the graph algorithms that we discussed

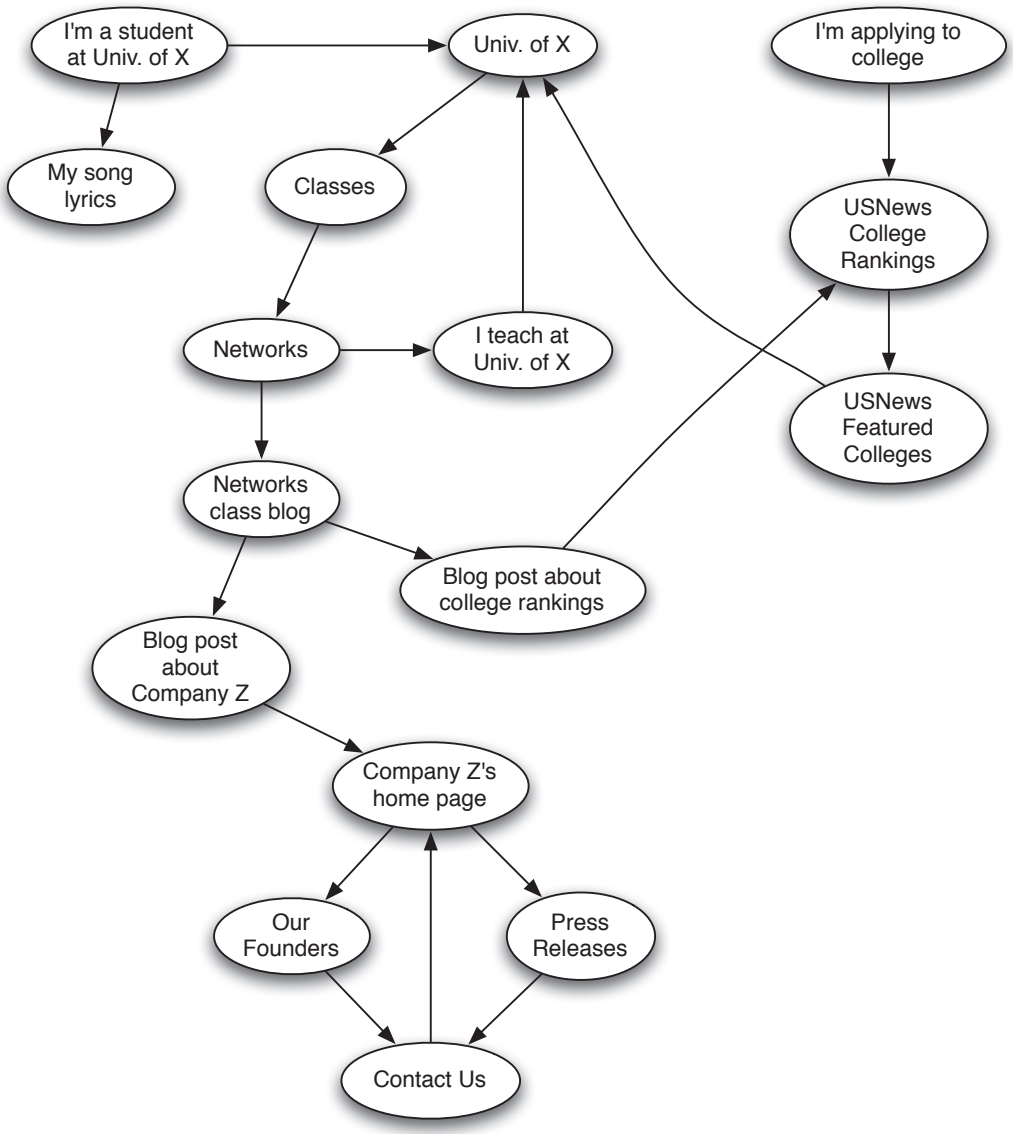
Information Networks

- Let us turn our attention to *information networks*
- I.e., Nodes are chunks of information and links join related chunks
- The most prominent example: the Web
- There are differences between various kinds of networks but we can use the same mathematical abstractions (graphs) to reason about them

The Web as a directed graph

- The basic distinction between social networks and the Web is the *directed* nature of the Web
- Directed graphs are asymmetric: links point from one node to another
- Analogy: friendship network versus name-recognition network (link from person A to person B if A has heard of B)
- Name-recognition network is asymmetric: e.g. celebrities are recognizable to millions of people, but they do not recognize all of their fans
- Facebook - Twitter distinction

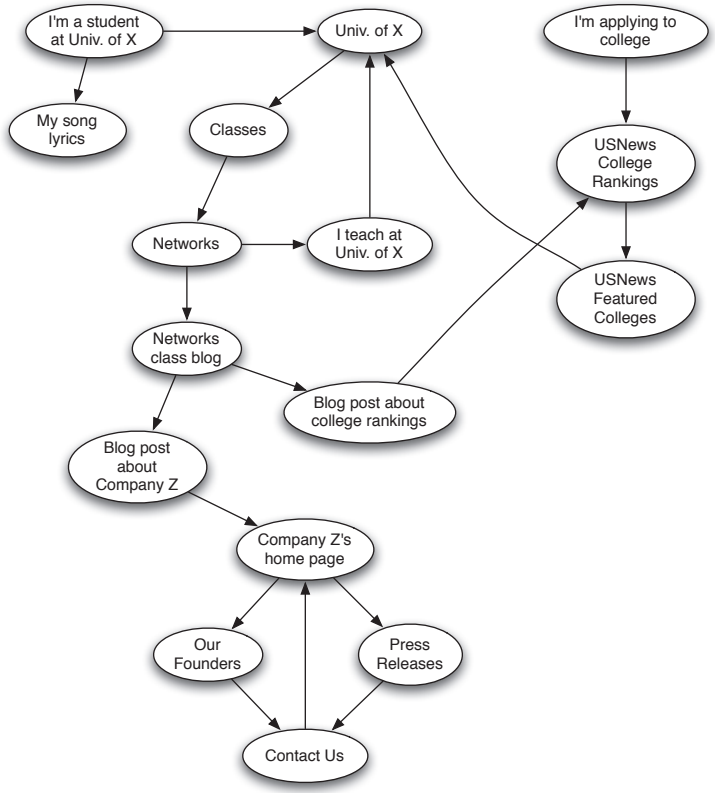
The Web as a directed graph



Strong Connectivity

Strong connectivity

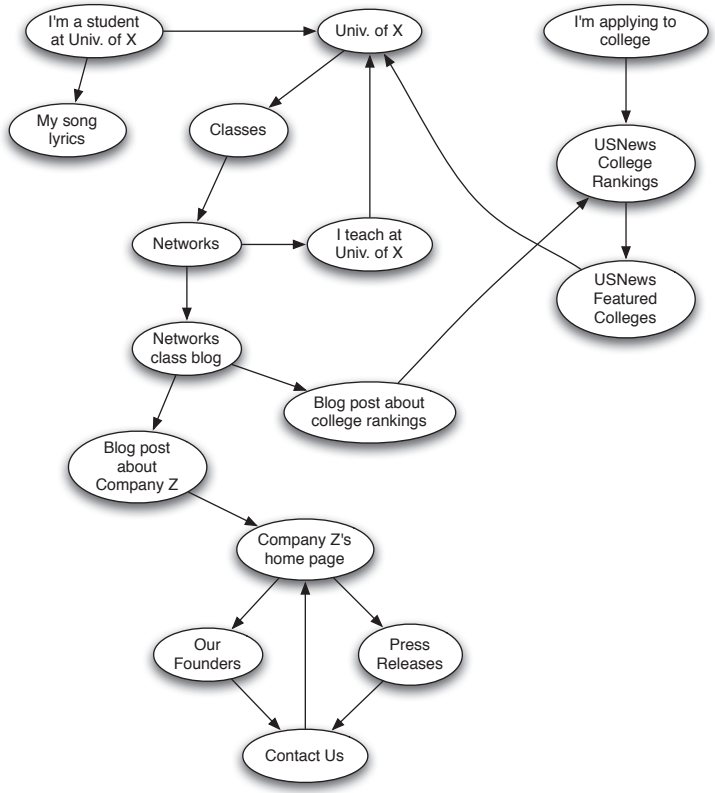
- If every node can reach every other node by a (directed) path the graph is **strongly connected**



Is the example graph strongly connected?

Strong connectivity

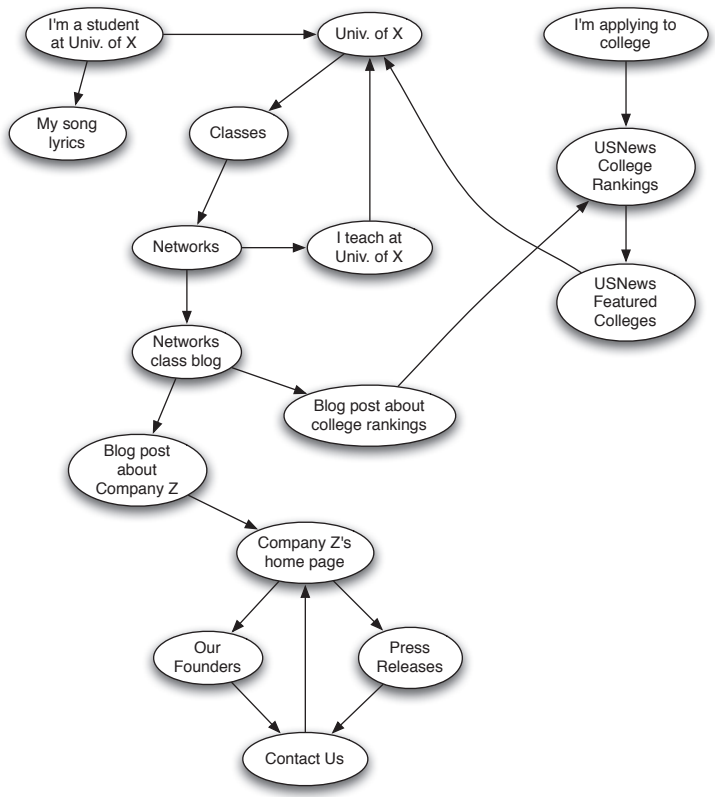
- If every node can reach every other node by a (directed) path the graph is **strongly connected**



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

Reachability

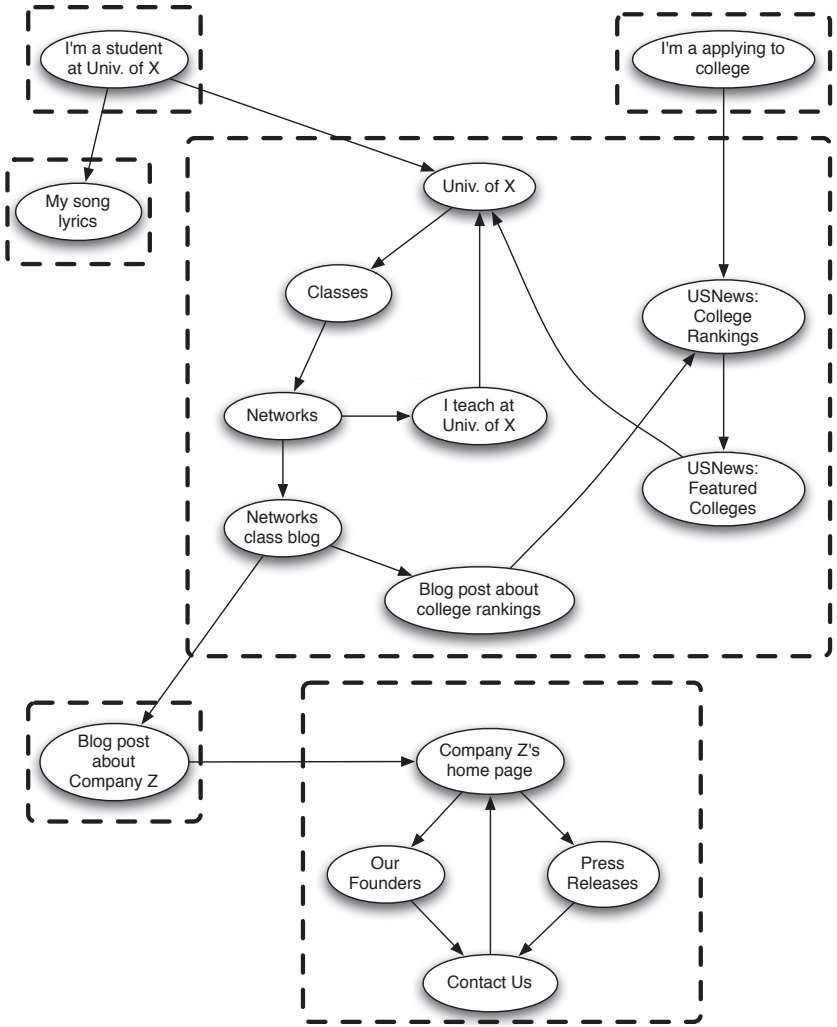
- Directed networks exhibit the following combinations:
  1. Pairs of nodes for which each can reach each other (Univ. of X and USNews: college rankings)
  2. Pairs for which one can reach the other but not vice versa (USNews college rankings and Company Z's home page)
  3. Pairs for which neither can reach the other (I'm a student and I'm applying to college)



Strongly connected components

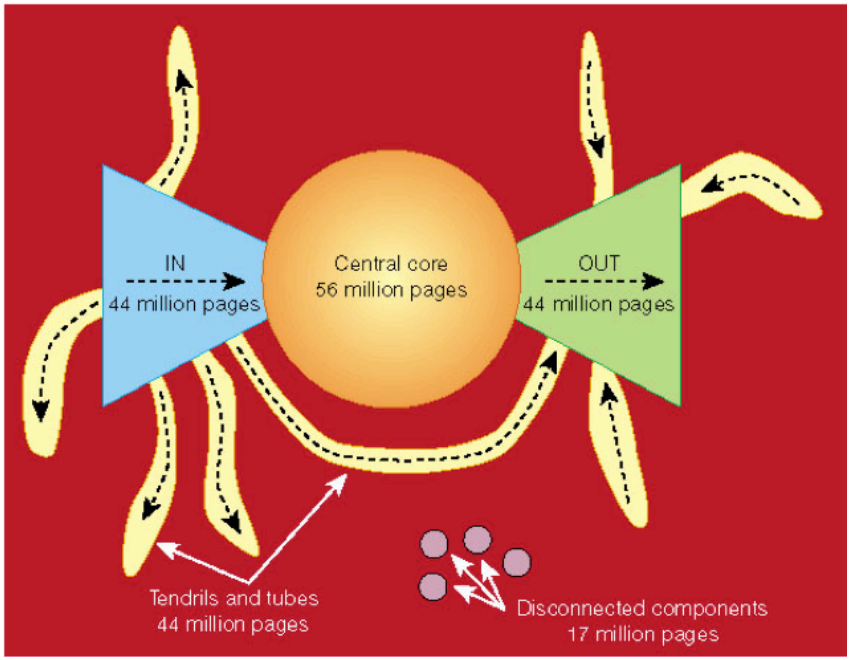
- **Strongly Connected Component (SCC)**: a subset of nodes such that
  1. every node in the subset hat a (directed) path to every other node in that subset (internally connected)
  2. the subset is not part of some larger set with the property that every node can reach every other
- Second part of definition differs from the undirected case
- SCC must not be completely isolated from the rest of the graph as in undirected case
- SCCs summarize a graph in a form of "super-nodes"

Strongly connected components



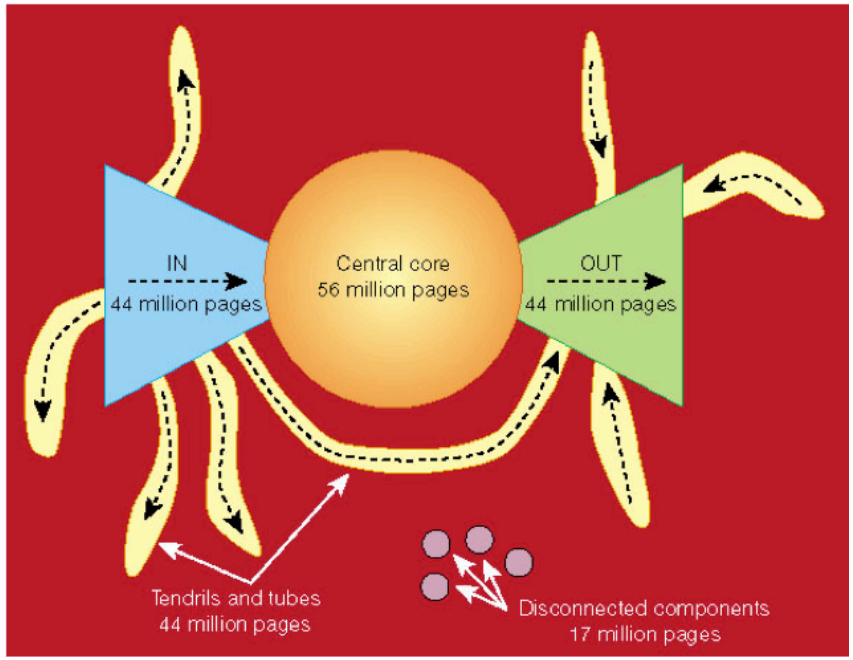
The Map of the Web

- Web **not** fully interconnected network!
- Web has a 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

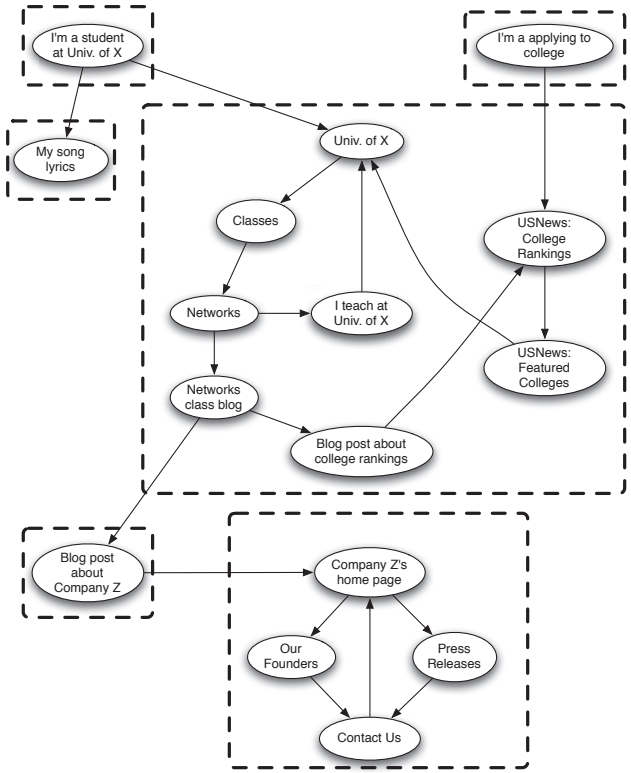


IN and OUT on the Web

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

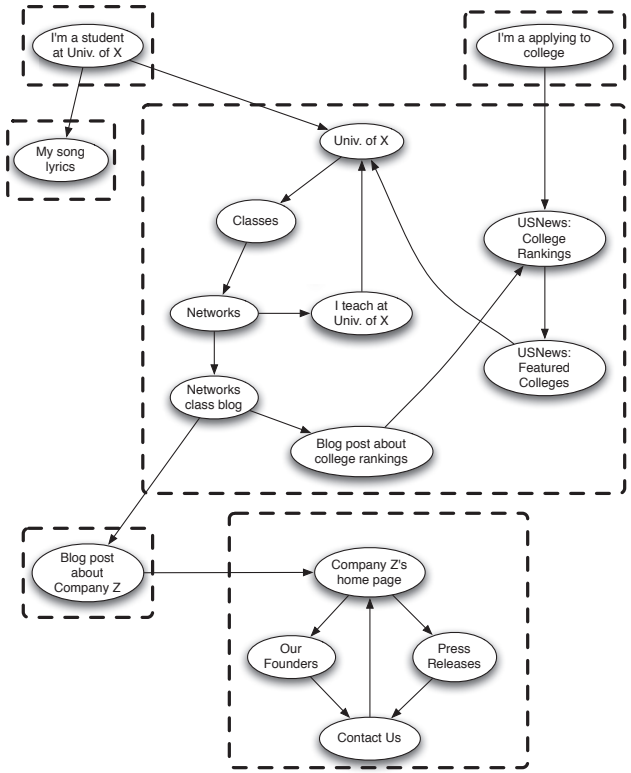


Example: IN and OUT



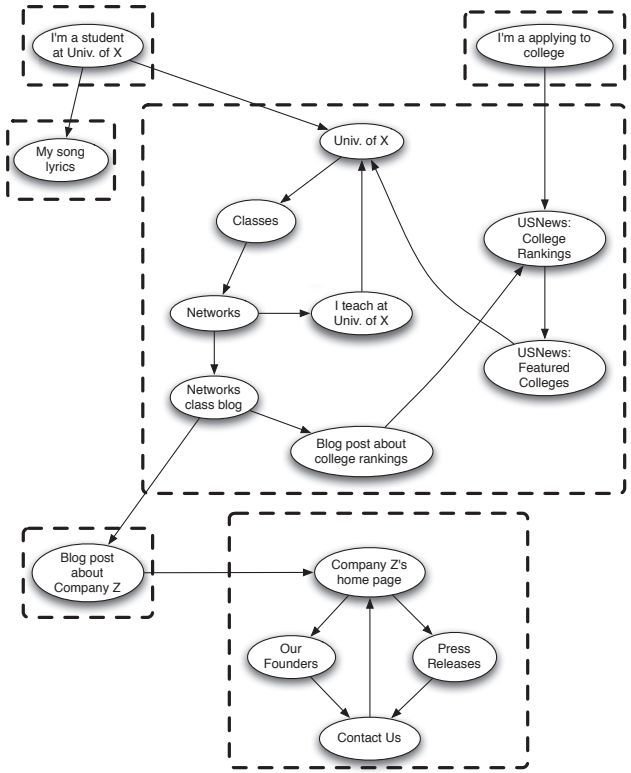
- SCC: largest in the middle of the network including e.g. Univ. of X
- I'm a student and I'm applying to college constitute

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

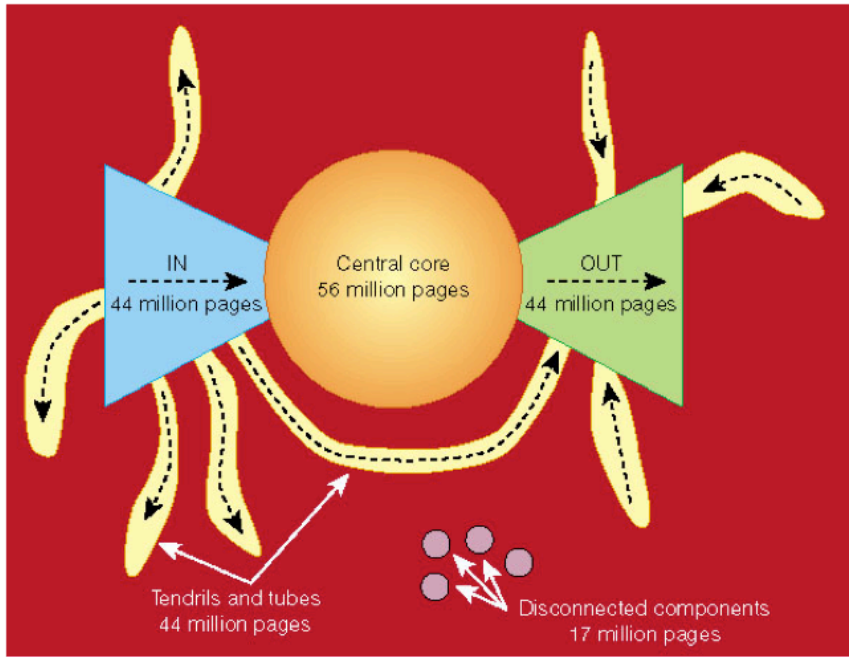
Example: IN and OUT



- 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

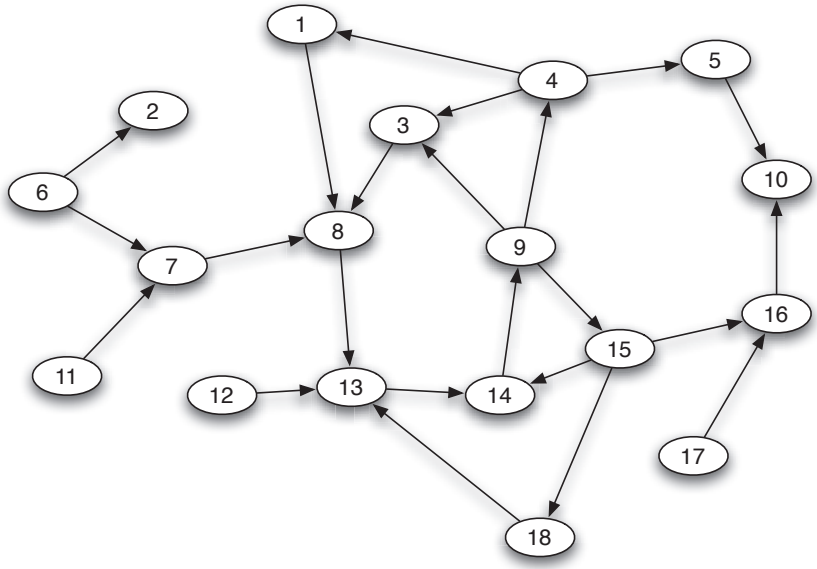
The Map of the Web

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



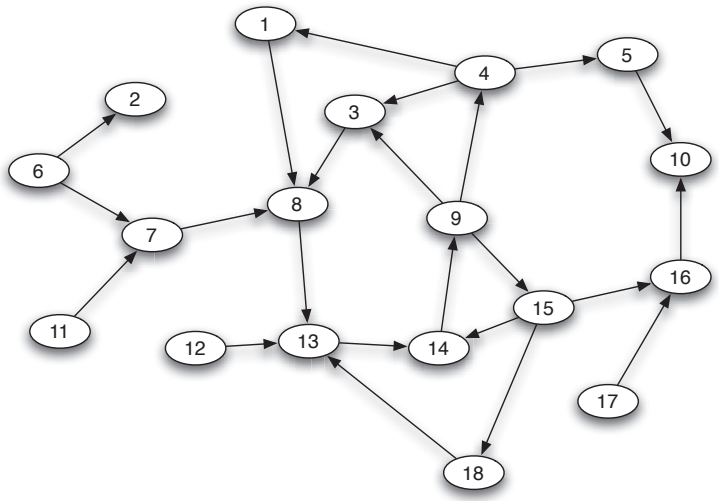
Example

- Exercise 1: Which nodes constitute the largest SCC? Which nodes belong to the IN of this SCC? Which to OUT?



Example

- Exercise 2: Pages can move between different parts of the bow-tie structure as new links are created and old ones are removed
- 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

- Exercise 3: 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.

Centrality

Centrality

- One of the key topics in network analysis is *centrality*
- What are the most central nodes in a network?
- What are the most important nodes in a network?
- What are the most influential nodes in a network?

Centrality

- In different kind of networks different interpretation are possible
- E.g. in a social network the most central node might be the most popular person
- E.g. on the Web the most central node might be a page with the best quality of content in a specific field
- E.g. on the Internet the most central node might be a router with the highest bandwidth
- Thus, there are many possible definitions of importance and many possible interpretations and therefore there are many centrality measures

HITS Algorithm

Voting by in-links

- We can use links to assess the *authority* of a page on a topic
- Implicit endorsements through links of other pages
- Each individual link may have many possible meanings
- E.g. it may be off-topic, it may be a critique, it may be a paid advertisement, or it may be a real endorsement
- We assume that in aggregate if a page receives many links from other relevant pages then it receives a kind of *collective endorsement*

Voting by in-links

This can be operationalized as follows:

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

List-finding technique

We can further extend the voting mechanism

- 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

Voting by in-links

Let us apply voting mechanism on an example:

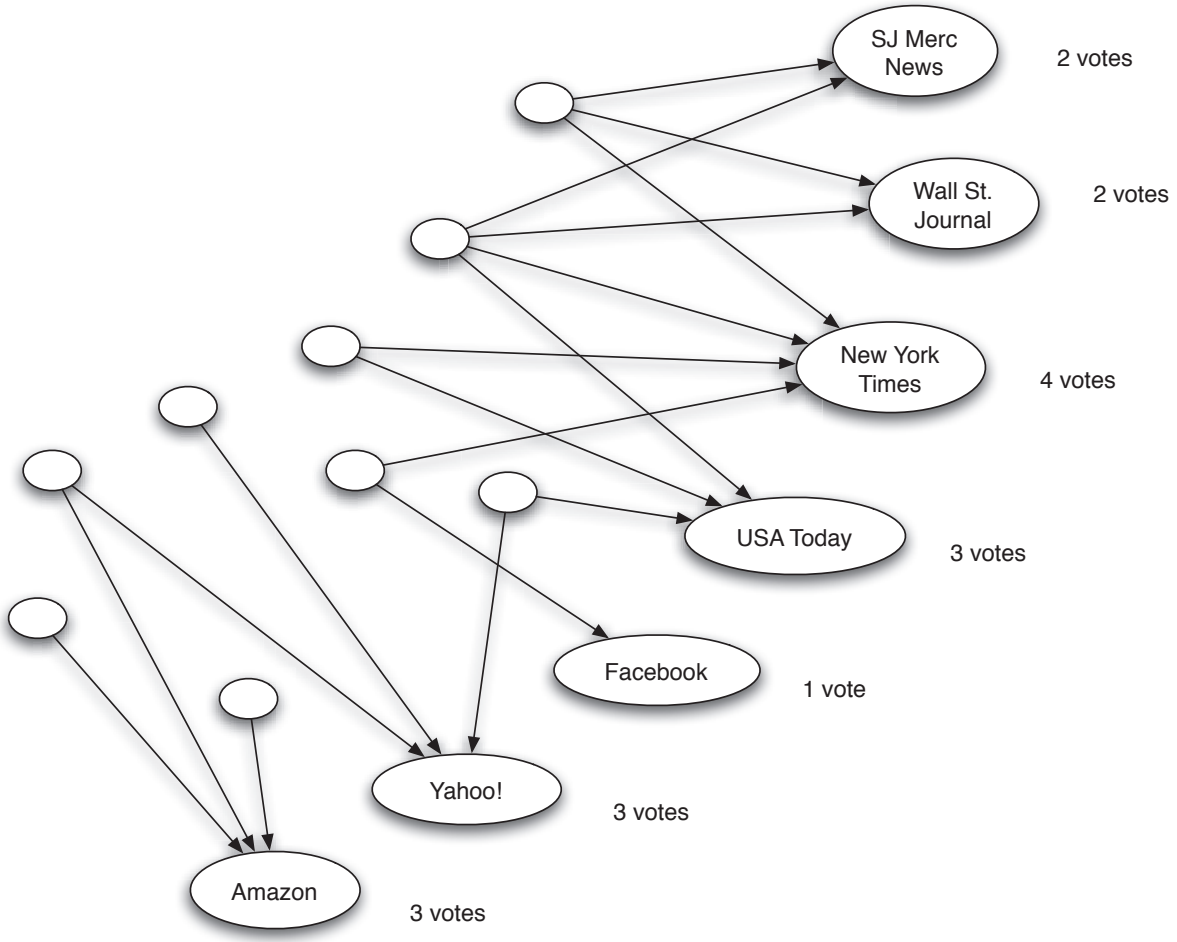
- First, collect a sample of pages relevant to the query “newspapers”
- Then: count votes to pages within this sample
- Typically, high scores for a mix of prominent newspapers
- Also, some prominent Web sites (e.g. Facebook) not directly related to newspapers will get a lot of votes
- Can you think of a reason why?

Voting by in-links

Let us apply voting mechanism on an example:

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

Newspapers example



List-finding technique

- 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 available
- E.g., pages that compile lists of resources relevant to the topic (lists of newspapers, etc.)
- Such lists exist for many broad enough queries (e.g. universities, hotels)

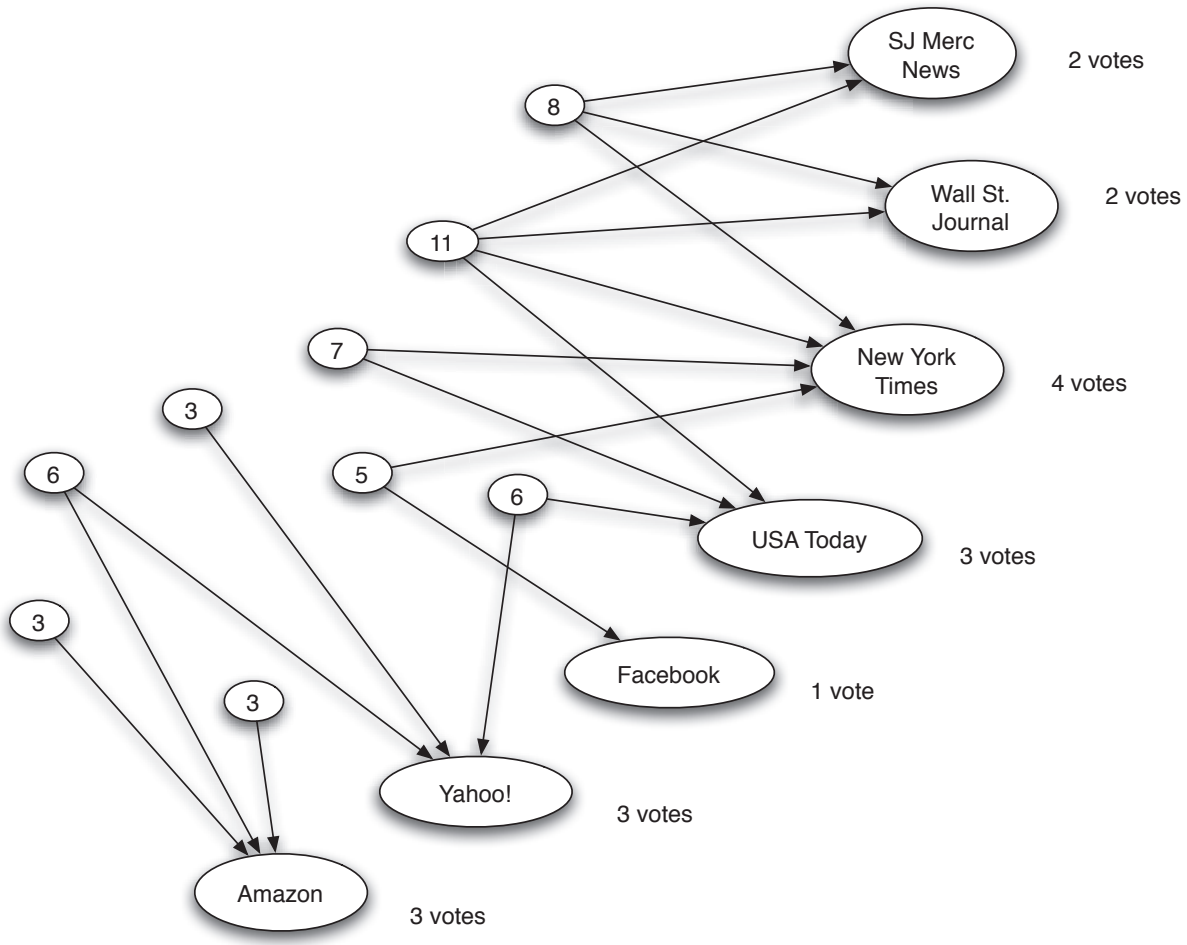
List-finding technique

We will discuss a useful technique for finding good lists:

- Among the pages casting votes, only a few of them exist that 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

Newspapers example

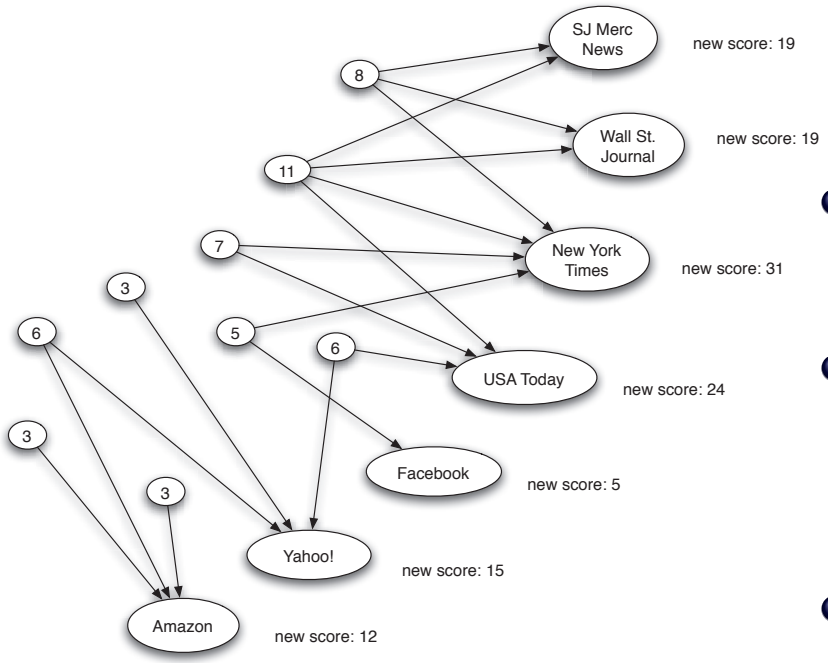
Page's value as a list: equal to the sum of the votes received by all pages that it voted for



The Principle of Repeated Improvement

- 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
- This time we give each page's vote a weight equal to its value as a list

Newspapers example



- Now the other newspapers (i.e., SJ Merc News, Wall Str. Journal) have surpassed the initially high-scoring Yahoo and Amazon
- Reason: they were endorsed by pages that were estimated to be good lists
- Suppose that you want to buy a new sci-fi book and get a lot of recommendations from your friends
- But there are some friends that you will trust more on this issue because you know that they are e.g. dedicated sci-fi fans (i.e. there are good lists)

Repeated improvement

- Finally, we don't need to stop after one step of reweighting but could continue
- We use now refined votes to refine list scores, then again refine votes/scores, etc.
- We can repeat this process for as many steps as we want
- *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 and will not change with new refinements

Hubs and Authorities

Hubs and Authorities

Let us specify this ranking procedure more precisely:

- We call pages prominent and highly endorsed for a query *authorities*
- The high value lists are called *hubs*
- For each page p we try to estimate its value as a potential authority $auth(p)$ and its value as a potential hub $hub(p)$
- Each of these is initially 1, as we don't know the best in either of these categories yet

Voting

- Voting procedure: Here, we use the quality of hubs to refine estimates for the quality of authorities
- *Authority Update Rule*: For each page p, update $auth(p)$ to be the sum of the $hub(q)$ scores of all pages q that point to it

List-finding technique

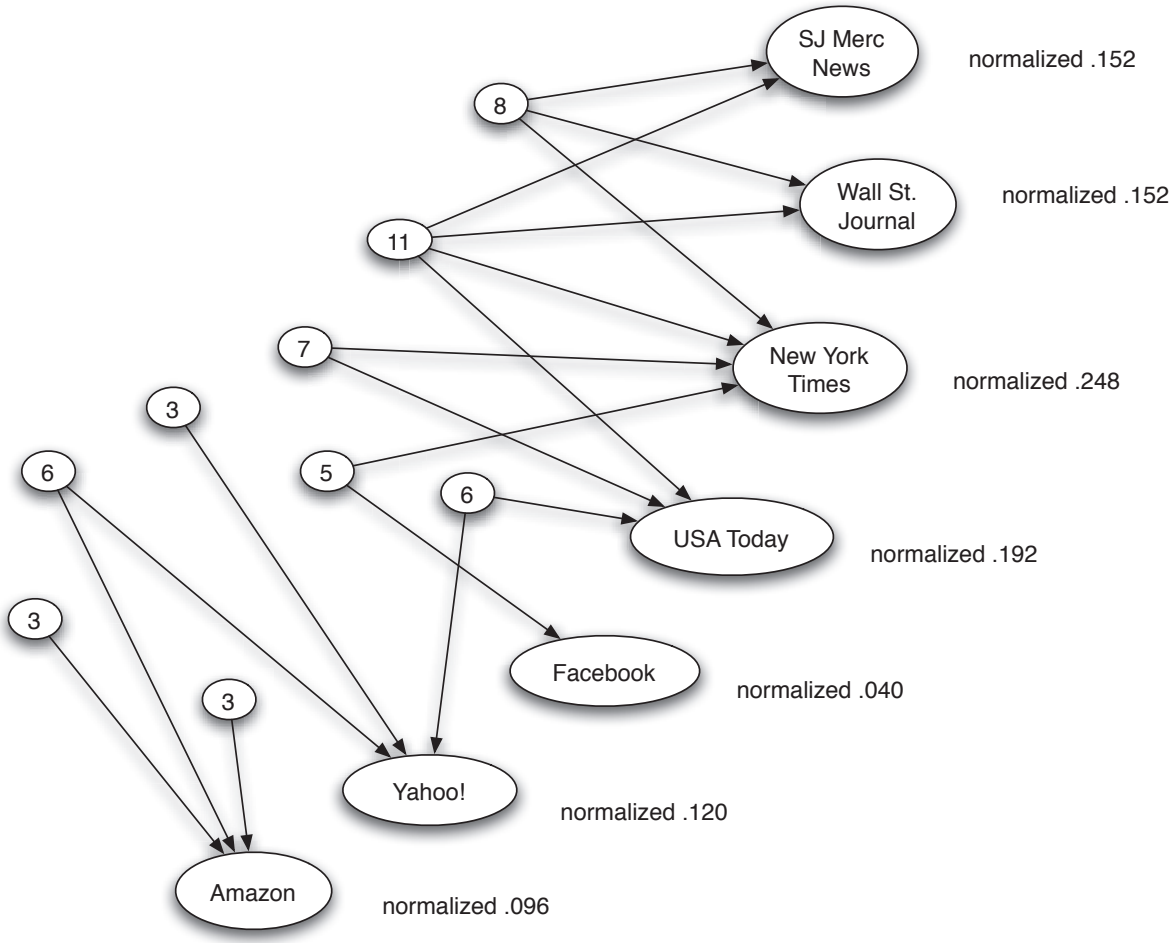
- List-finding procedure: in which we use the quality of authorities to refine estimates for the quality of hubs
- *Hub Update Rule*: For each page p, update $hub(p)$ to be the sum of the $auth(q)$ scores of all pages q that it points to

Repeated improvement

- Repeated improvement step in which we start with all hub scores and all authority scores equal to 1
- We choose a number of steps k
- We then perform a sequence of k hub-authority updates, where in each update:
  1. First: apply the Authority Update Rule to the current set of scores
  2. Then: apply the Hub Update Rule to the resulting set of scores
- At the end, the hub and authority scores may have very large numbers - therefore we *normalize* them to a probability distribution[1]

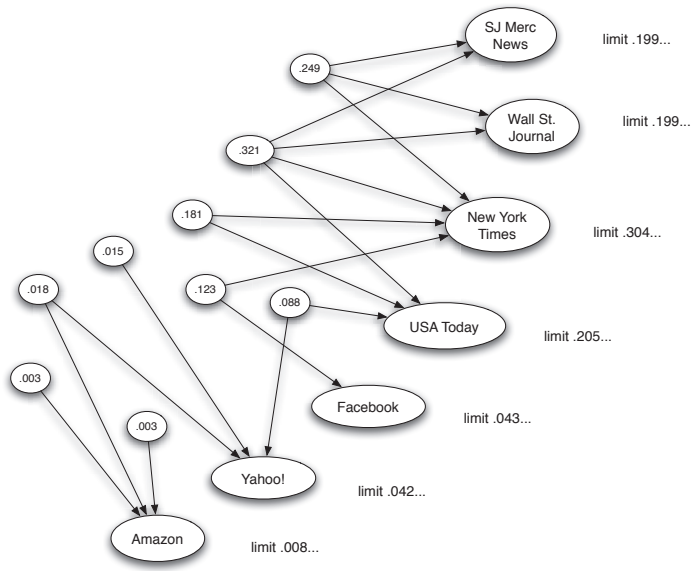
[1] divide each authority score by sum of all authority score, same with hub score

Newspapers example



Result of normalizing the 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 vals
- Kind of equilibrium: relative sizes remain unchanged if authority update rule or hub update rule is applied
- A page's authority score proportional to hub scores of pages that point to the page and vice versa

Hubs and authorities: Recap

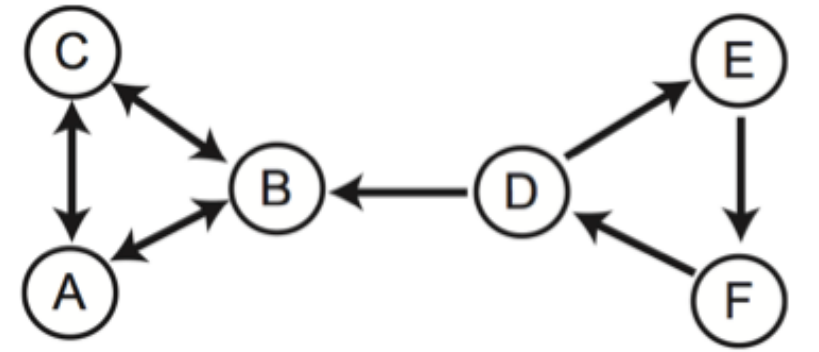
- Intuition behind hubs and authorities is based on idea that pages play multiple roles in the network
- In particular, pages (hubs) can strongly endorse other pages without themselves being heavily endorsed
- Real-life example: E.g. competing firms will not link to each other
- The only way to pull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

HITS Algorithm

- The procedure we discussed corresponds to the **Hypertext Induced Topic Selection (HITS)** algorithm (Kleinberg, 1999)
- Used for rating and ranking websites based on the link information when identifying topic areas (search query dependent)
- Mutual reinforcement between Web pages: "A better hub points to many good authorities, and a better authority is pointed to by many good hubs"
- 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

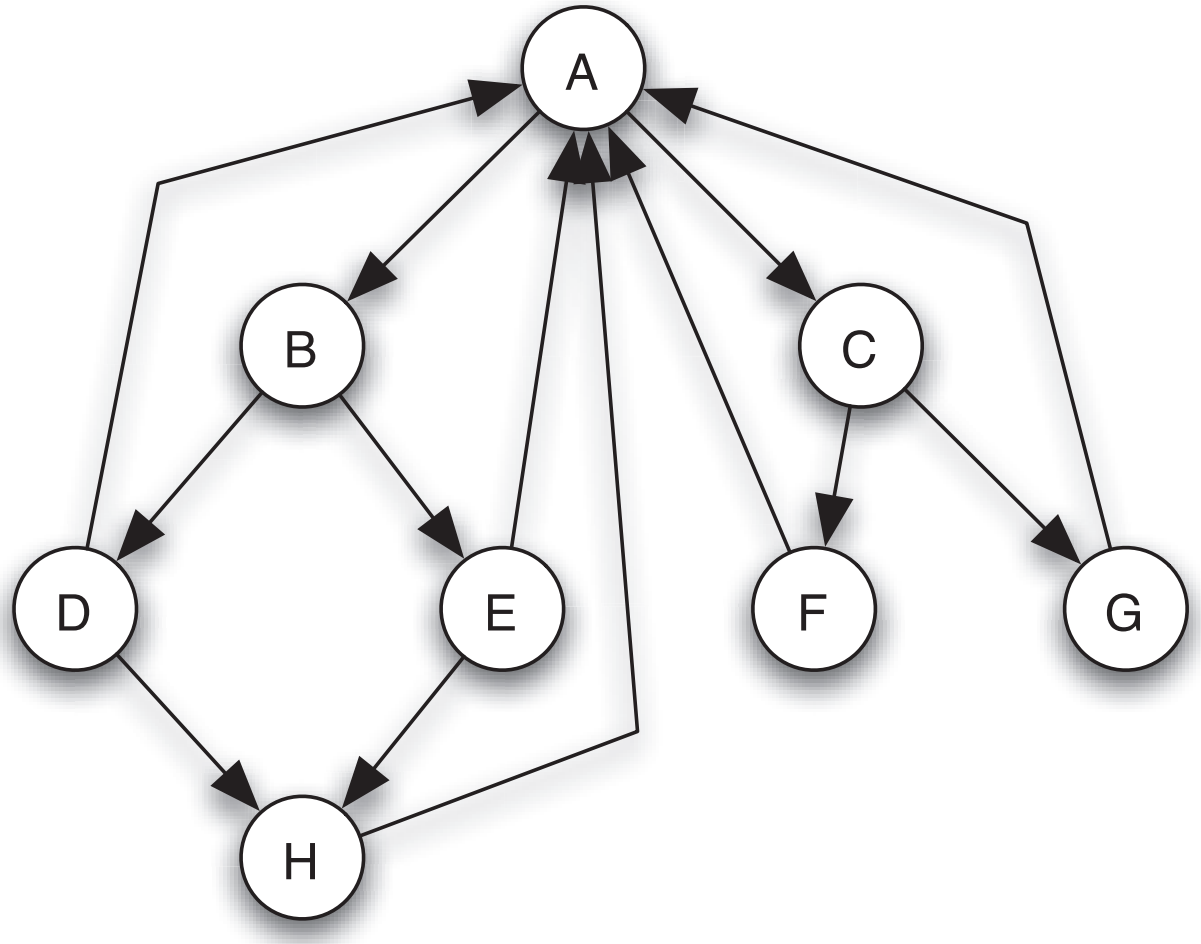
Example

- Exercise 4: Apply HITS update rules and calculate two steps of algorithm.



Example

- Exercise 5: Apply HITS update rules and calculate two steps of algorithm.



PageRank Algorithm

PageRank

- Intuition from last time: links as votes (HITS algorithm)
- Page more important if it has many in-links
- Do you think that all in-links are equal?

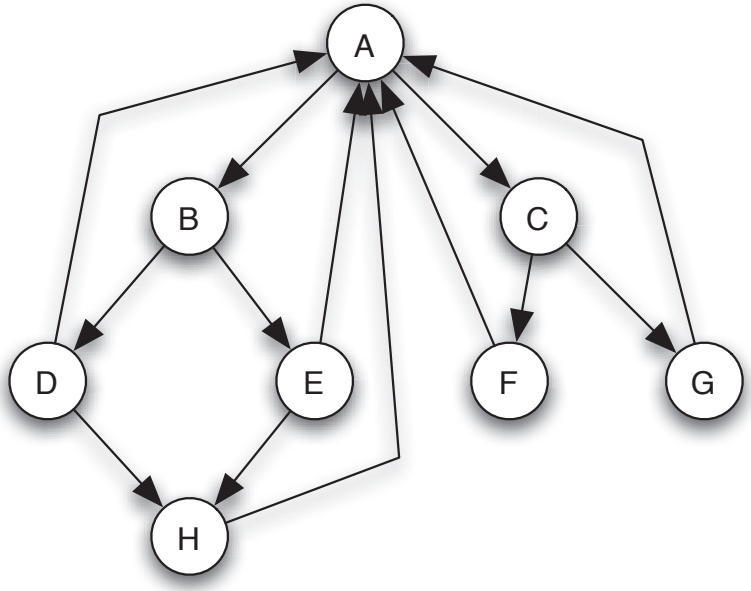
PageRank

- Intuition from last time: links as votes (HITS algorithm)
- Page more important if it has many in-links
- Do you think that all in-links are equal?
- No! Links from important pages are more important

The basic definition of PageRank

- We compute PageRank in the following way:
  1. Given a Web graph with n nodes assign each node initial PageRank $1/n$
  2. Choose a number of steps k
  3. Perform a sequence of k updates where we calculate rank of each node: each page divides and passes its current PageRank equally across its out-going links. Each page updates its new PageRank to be the sum of the shares it receives.

PageRank Example: First two steps



All pages start out with a PageRank of 1/8. PR(A) =1/2. It gets all of F, G, H and half of D and E. What about B and C?
PR(B) and PR(C): get half of A's PR, so only 1/16

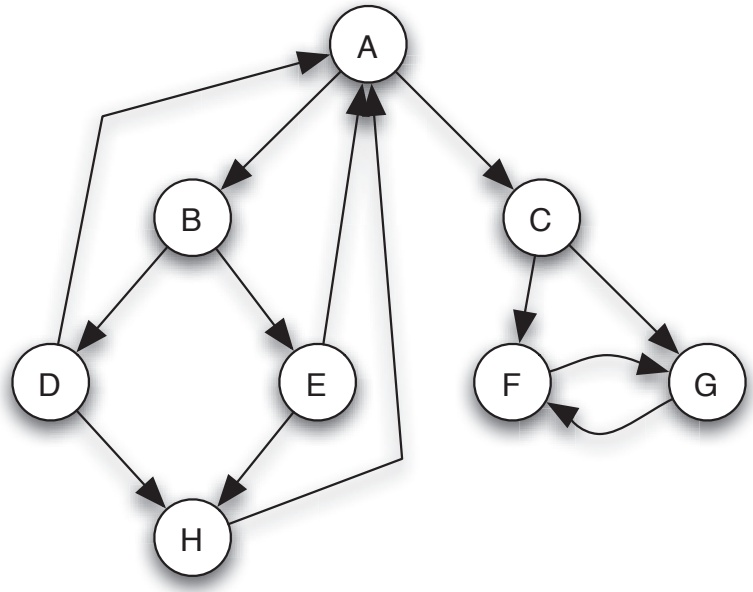
Step \ Page	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	5/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

Equilibrium Values of PageRank

- Similarly as with hub-authority computation if we increase the number of iteration steps k the values will become stable and will not change anymore
- The calculation converges and we reach an equilibrium
- One can prove this convergence
- One can also prove that for a strongly connected network the equilibrium values are unique

A basic problem with PageRank: Example

- Now, F and G point to each other and not to A
- PageRank that flows from C to F and G can never flow back to the network
- Links out of C - “slow leak”, all the PageRank ends up at F and G



- Results in convergence to PageRank of 1/2 for each of F and G
- Others have PageRank of 0

A basic problem with PageRank

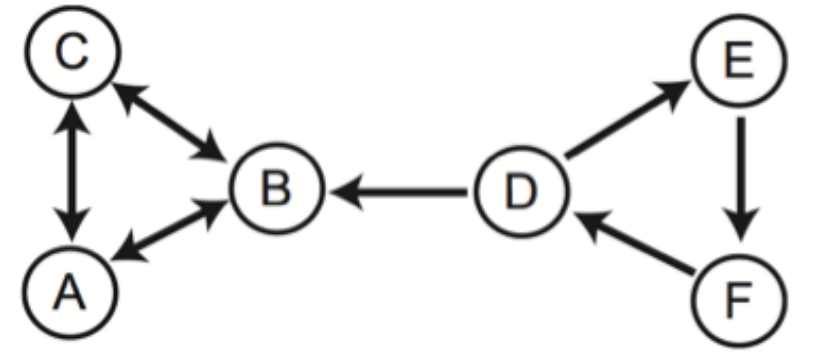
- Wrong nodes may end up with "all" the PageRank
- If graph is not strongly connected - complete PageRank will leak to nodes in OUT
- Therefore: scale down all values by a scaling factor s (strictly between 0 and 1)
- Divide the residual $1 - s$ equally over all nodes giving $(1 - s)/n$ to each
- Preserves the total PageRank in the network - based on redistribution
- Can be shown that this rule converges and that no PageRank is leaking

Limit of the Scaled Update

- Repeated application converges to set of limiting PageRank values as number of updates k goes to infinity
- These limiting values form the unique equilibrium: unique set of values that remains unchanged under application of update rule
- Depend on choice of scaling factor s
- In practice: scaling factor s usually between 0.8 and 0.9

Example

- Exercise 6: Apply Basis PageRank Update Rule and calculate two steps of PageRank algorithm.



Example

- Exercise 7: Suppose we run the scaled PageRank algorithm with a scaling factor s = 0.8. In which cluster (A, B, C) or (D, E, F) is the node with the highest PageRank situated? In equilibrium, which node A or B has a higher PageRank?