WEB STRUCTURE

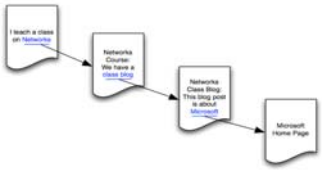
CS1305 COMPUTATION AND CULTURE IN A DIGITAL AGE

ORGANIZING WEB PAGES

Brainstorming: How would you organize thousands of web pages?
What about ones that share relating topics?

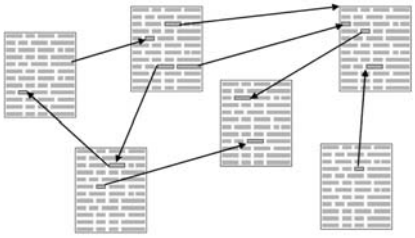
ORGANIZING WEB PAGES AS A NETWORK

How do we go from one web page to another? How are web pages related to one another?



The use of a network structure truly brings forth the globalizing power of the Web by allowing anyone authoring a Web page to highlight a relationship with any other existing page, anywhere in the world

HYPERTEXT



Hypertext is text displayed on a computer display or other electronic devices with references (hyperlinks) to other text that the reader can immediately access, or where text can be revealed progressively at multiple levels of detail

HISTORY OF HYPERTEXT

Vannevar Bush's described in his essay "As We May Think", published in July 1945, a hypothetical machine called the Memex: a hypertext-like device capable of allowing its users to comb through a large set of documents stored on microfilm, connected via a network of "links" and "associative trails" that anticipated the hyperlinked structure of today's Web.

However, Bush was not the first person to imagine the web.

By the 1930s, Otlet started imagining a global network that would easily store and retrieve information. Left is a partial illustration.

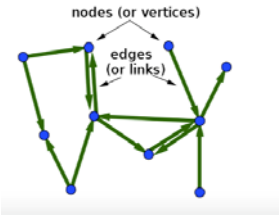
© MUSÉE MAE CENTRE D'ARCHIVES (BELGIUM)

HYPERTEXT STRUCTURE

Think about the structure: does it have to be linear, or non-linear? (what about network?)
Under what circumstances is the structure non-linear?

DIRECTED GRAPH

A directed graph is graph, i.e., a set of objects (called vertices or nodes) that are connected together, where all the edges are directed from one vertex to another.

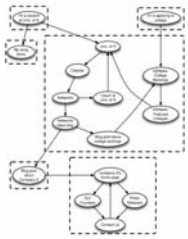


STRONGLY CONNECTED COMPONENTS

Definition: A Strongly Connected Component (SCC) of a directed graph is the maximal subset of a graph with a directed path between any two vertices.

Why is SCC important? What are some of SCC's application?

Can SCC merge into each other?



WEB 2.0

Reading: https://www.cbsnews.com/news/what-is-web-20/
How Web 2.0 is different from Web 1.0? Watch this video:
https://www.youtube.com/watch?v=NLlGopyXT_g

EXERCISE

List the nodes in the largest strongly connected component of this graph.

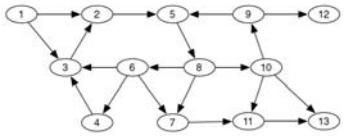


Figure 1: The network of Web pages for Question (1).

EXERCISE

Suppose you are allowed to add one link to the graph in Figure 1, going from one node to another; which link would you add if you wanted to increase the size of the largest strongly connected component by as much as possible? Give an explanation for your answer.

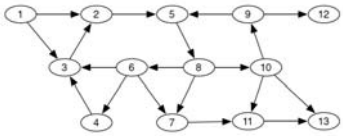


Figure 1: The network of Web pages for Question (1).

EXERCISE

Which link would you reverse in this way if you wanted to increase the size of the largest strongly connected component by as much as possible? Give an explanation for your answer.

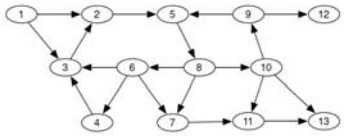


Figure 1: The network of Web pages for Question (1).

EXERCISE

Can you remove exactly one link in this way so that after the removal, the resulting graph contains no strongly connected component with more than one node?

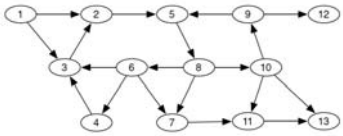


Figure 1: The network of Web pages for Question (1).

LINK BASED ANALYSIS

CS1305 COMPUTATION AND CULTURE IN A DIGITAL AGE

READING ASSIGNMENT

How to appear #1 on Google?
https://moz.com/beginners-guide-to-seo

Write a blog post about the information you learnt from the chapter(s) of your choice, make it interesting

EXERCISE 1

Try this on the network of Web pages shown in Figure 2. In particular, say whether the indicated set of numbers forms an equilibrium set of PageRank values under the Basic PageRank Update Rule. Also, provide an explanation for your answer: specify either why they form an equilibrium, or how they fail to form an equilibrium.

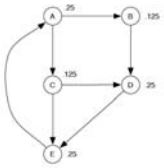


Figure 2: The network of Web pages for Question (2).

EXERCISE 2

For the network of Web pages shown in Figure 3, determine the equilibrium PageRank values under the Basic PageRank Update Rule. Give an explanation for your answer.
(Hint: You can use the approach described in class; let x denote the (unknown) PageRank value of node A, then work out the other PageRank values in terms of x, and then determine a value for x.)

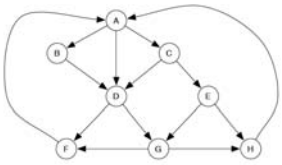
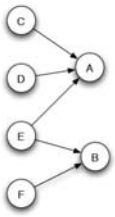


Figure 3: The network of Web pages for Question (3).

EXERCISE 3

Show the values that you get if you run two rounds of computing hub and authority values on the following network

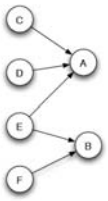


Hint 1: One round of updates for the hub and authority rule applies the authority update rule using the current hub scores and then applies the hub update rule to the resulting authority scores; so a single round consists of two updates, first to update, the authority scores, and then to update the hub scores.

Hint 2: Recall also that computing hub and authority values involves a nal normalization step, in which at the end we divide each authority score by the sum of all authority scores, and divide each hub score by the sum of all hub scores.

EXERCISE 3 (CONTINUED)

Now we come to the issue of creating pages so as to achieve large authority scores, given an existing hyperlink structure.



In particular, suppose you wanted to create a new Web page X, and add it to the network in Figure 4, so that it could achieve a (normalized) authority score that is as large as possible.

One thing you might try is to create a second page Y as well, so that Y links to X and thus confers authority on it. In doing this, it's natural to wonder whether it helps or hurts X's authority to have Y link to other nodes as well.

First option, Y links only to X, while in the second option, Y links to other strong authorities in addition to X.

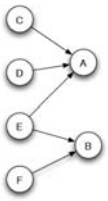
Option 1: Add new nodes X and Y to Figure 4; create a single link from Y to X; create no links out of X.

Option 2: Add new nodes X and Y to Figure 4; create links from Y to each of A, B, and X; create no links out of X.

For which of Options 1 or 2 does page X get a higher authority score (taking normalization into account)? Give a brief explanation in which you provide some intuition for why this option gives X a higher score.

.

EXERCISE 3 (OPTIONAL)



Suppose instead of creating two pages, you create three pages X, Y , and Z, and again try to strategically create links out of them so that X gets ranked as well as possible.

Describe a strategy for adding three nodes X, Y , and Z to the network in Figure 4, with choices of links out of each, so that when you run the 2-step hub-authority computation (as in parts (a) and (b)), and then rank all pages by their authority score, node X shows up in second place. (You can have the links from X, Y , and Z point to any nodes you want, including others among X, Y , and Z, and/or the existing nodes in Figure 4. Note that you are not allowed to remove any of the nodes or links that were already present in Figure 4, nor add any links out of the existing nodes in that figure; the only modification you can make to the network in Figure 4 is to add links out of the three new nodes X, Y , and Z.)

Show the hub and authority scores in the new network you create, to demonstrate that you've succeeded in getting node X into second place. Give the values both before and after this final normalization step.