Sections

Option C - Web Science (/topic/C-web-science.html)

# C.5 [HL] Analysing the web

*How the web can be represented and analysed through graph theory*

## C.5.1 Describe how the web can be represented as a directed graph

### What is a graph?

In graph theory, a graph is a set of nodes (also called vertices) that can be connected through edges. Graphs are used to model the relation between objects.

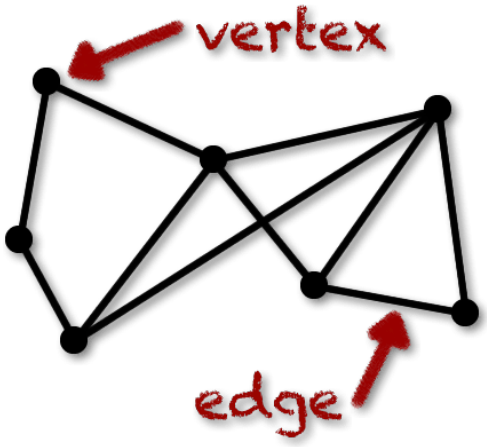


*Figure 1: A simple graph [1 (http://world.mathigon.org/Graph_Theory)]*

- Vertex: in a web graph each web page (say URL – Unique Resource Locator) is represented by a vertex
- Edge: in a web graph each hyperlink is represented by a directed edge

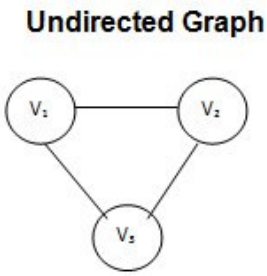


Figure 1: An Undirected Graph

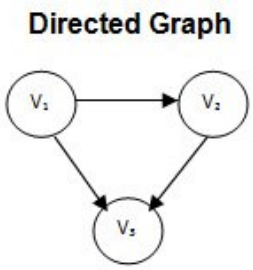


Figure 2: A Directed Graph

*Figure 2: Difference between undirected and undirected graph [2 (http://www.differencebetween.com/difference-between-directed-and-vs-undirected-graph/)]*

### Why is the web graph not complete?

A complete graph would mean that each vertex is connected with each other. However, not all web pages are hyperlinked to each other, which is why the web graph is not a complete graph.

## C.5.2 Outline the difference between the web graph and sub-graphs

### Web graph

- Web graph describes the directed links between web pages in the WWW.
- It is a directed graph with directed edges
    - Page A has a hyperlink to Page B, creating a directed edge from Page A to Page B

### Sub-Graph

- A set of pages that are part of the internet
- Can be a set of pages linked to a specific topic ex.: Wikipedia -> one topic but references(and hyperlinks) to other web pages
- Can be a set of pages that deal with part of an organization

## C.5.3 Describe the main features of the web graph such as bowtie structure, strongly connected core (SCC), diameter

### Bowtie structure

Computer scientists try to study the web graph in order to learn about the webs characteristics. One proposed structure for the web graph is the bowtie structure, which looks like this:

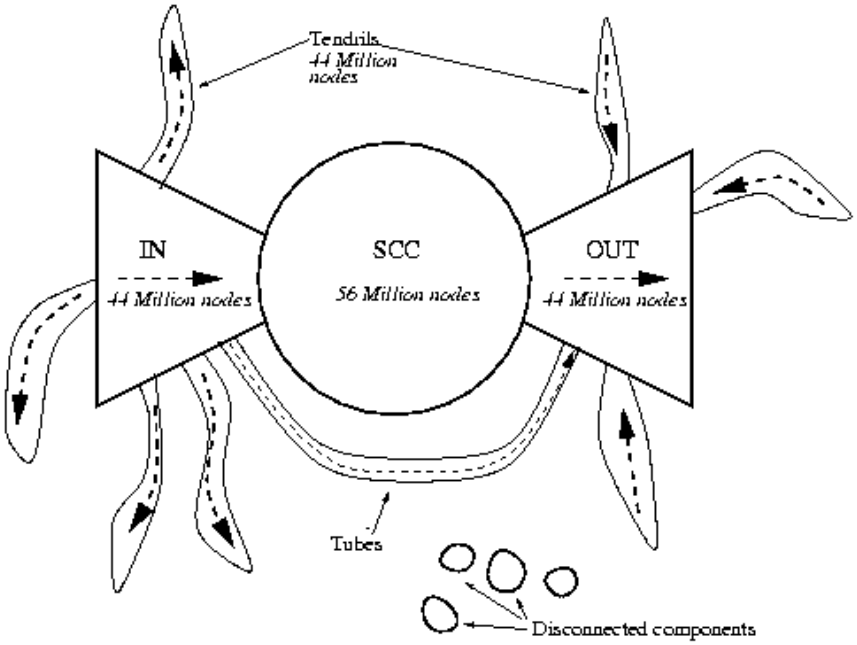


*Figure 3: Bowtie structure of the web graph (Broder et al., 2000) [3 (http://www9.org/w9cdrom/160/160.html)]*

In this model, different components are identified: SCC, IN, OUT, Tubes, Tendrils.

## SCC:

- strongly connected core from or to which many nodes lead to/from.
- Can reach all nodes in OUT
- Cannot reach nodes in IN

## IN-section:

- made up of nodes that can reach the SCC
- Cannot be reached by the SCC

## OUT-section:

- made up of nodes that can be reached by the SCC
- Cannot reach the SCC

## Tubes:

- Nodes not part of the SCC
- made up of nodes linking the IN- or the OUT-section.

## Tendrils:

- made up of nodes that are not connected to the SCC
- Connected to either the IN- or the OUT-sections

## Diameter:

- Different definitions
- Usually the average path length between two random nodes
- Usually considering individual parts (e.g. SCC) only, as connectivity between parts is limited (I.e. nodes in IN can usually not be reached from OUT) ***Further reading:***
- Mining the Inner Structure of the Web Graph (https://www.researchgate.net/publication/200111010_Mining_the_Inner_Structure_of_the_Web_Graph)
- TU-Darmstadt, Lecture on the Web Graph (https://www.ke.tu-darmstadt.de/lehre/archiv/ss05/web-mining/wm-graph.pdf)
- NYU, Lecture 9: Web Structure and Evolution (https://cs.nyu.edu/courses/fall07/G22.2580-001/lec9.html)

## C.5.4 Explain the role of graph theory in determining the connectivity of the web

### Connectivity

This is just a metric to discuss how well parts of a network connect to each other.

### Small world graph

This is a mathematical graph whereas not all nodes are directly neighbors, but any given pair of nodes can be reached by a small number of hops or better said with just a few links. This is due to nodes being interconnected through interconnected hubs.

- 2 Properties of the small world graph:
  - Mean shortest-path length will be small
    - Most pairs of nodes will be connected by at least one short path
  - many clusters (highly connected subgraphs)
- Analogy: airlines flight whereas you can reach any city most likely in just under three flights.
- Examples: network of our brain neurons
- Maximizes connectivity
- Minimizes # of connections

### 6 degrees of separation

This originates from the idea that any human in the world is related in some way over 6 or less connections (steps). This idea can be taken further in a more general perspective on a graph, whereas any given pair of nodes within the network can be reached with just a maximum of 6 steps.

The idea itself can be applied to the web graph, suggesting high connectivity regardless of big size.

- Not necessarily small world graph
  - High connectivity between all nodes

### Web diameter (and its importance)

The diameter of the web graph has no standard definition, but is usually considered to be the average distance (as each edge has the same path length, this would be steps) between two random nodes.

This is important because it is an indicator of how quickly one can reach some page from any starting page in average. This is of importance for crawler, which want to index as many pages as possible in the shortest path.

- average distance between two random nodes (pages)
- important for crawlers to have a guide of how many steps it should take to reach a page
- a factor to consider, is if the path is directed or undirected
    - often there is no direct path between nodes

### Importance of hubs and authorities (link to C.2.3)

Hubs and authorities have special characteristics:

- Hubs: have a large number of outgoing links
- Authorities: have a large number of incoming links

For connectivity, this means that a larger number of hubs improves connectivity, while authorities are more likely to decrease connectivity as they usually do not link to many other pages.

### Link to power laws (C.5.6)

See degree distribution C.5.6 (bottom of the document) __

## C.5.5 Explain that search engines and web crawling use the web graph to access information

_Algorithmically describe how a web spider crawls through the web _

### Simple algorithm

In its basic form the algorithm will start from a seed (a number of pages from which to start) and put all the URLs in a queue. It will then loop through the queue until it is empty, each time dequeuing an URL, requesting its document, indexing this document while also collecting links from it. These links will be added to the queue, if they haven't been visited yet.

```
queue = LoadSeed();
while (queue is not empty){
    dequeue url
    request document
    store document for later processing
    parse document for links
    add unseen links to queue
}
```

### Adaptive crawling

A more advanced crawler algorithm will prioritize on what to crawl and adapt the queue live so that more relevant information is indexed first. There is an additional stage in the algorithm, where the document is analyzed for relevance, so that the queue is reorganized accordingly.

```
queue = LoadSeed();
while (queue is not empty){
    dequeue url
    request document
    store document
    // process document
    analyze document to determine if it's relevant
    update prioritizer with information about this document
    // extract and queue links
    parse document for links
    eliminate document for links
    prioritize links
    add prioritized links to queue
}
```

Credits for the pseudo code go to Jim's Blog!

*Source:*

Jim's Random Notes Blog, 2011 (http://blog.mischel.com/2011/12/16/writing-a-web-crawler-crawling-models/)

## C.5.6 Discuss whether power laws are appropriate to predict the development of the web

### Degree distribution

Another factor to look at when considering connectivity is the degree distribution of a network. The degree of a page is the number of connections (links) it has, which can further categorized into incoming and outgoing links.

It is very obvious that the number of pages with a higher degree decreases. The distribution suggests a power law, although it does not exactly fit the model.

- Degree of a page is the number of links it has (in/out)
- Degree distribution is how many pages there are with an increasing degree

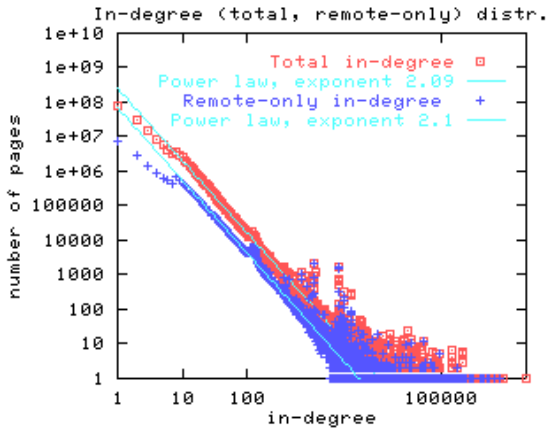


*Figure 4: In-degree distribution (Broder et al., 2000) [3 (http://www9.org/w9cdrom/160/160.html)]*

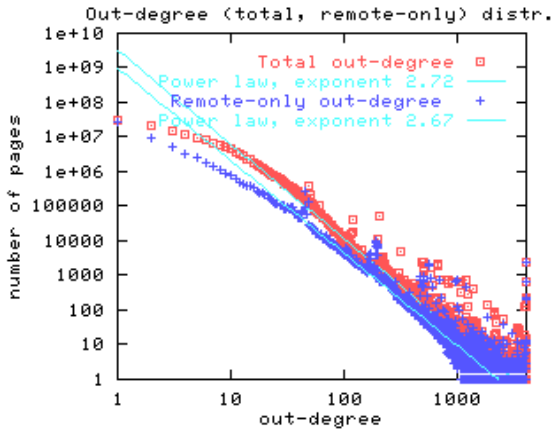


*Figure 5: Out-degree distribution (Broder et al., 2000) [3 (http://www9.org/w9cdrom/160/160.html)]*

---

*Sources:*

[1] Mathigon.org (http://world.mathigon.org/Graph_Theory)

[2] Differencebetween.com (http://www.differencebetween.com/difference-between-directed-and-vs-undirected-graph/)

[3] Broder et al., 2000 (http://www9.org/w9cdrom/160/160.html)