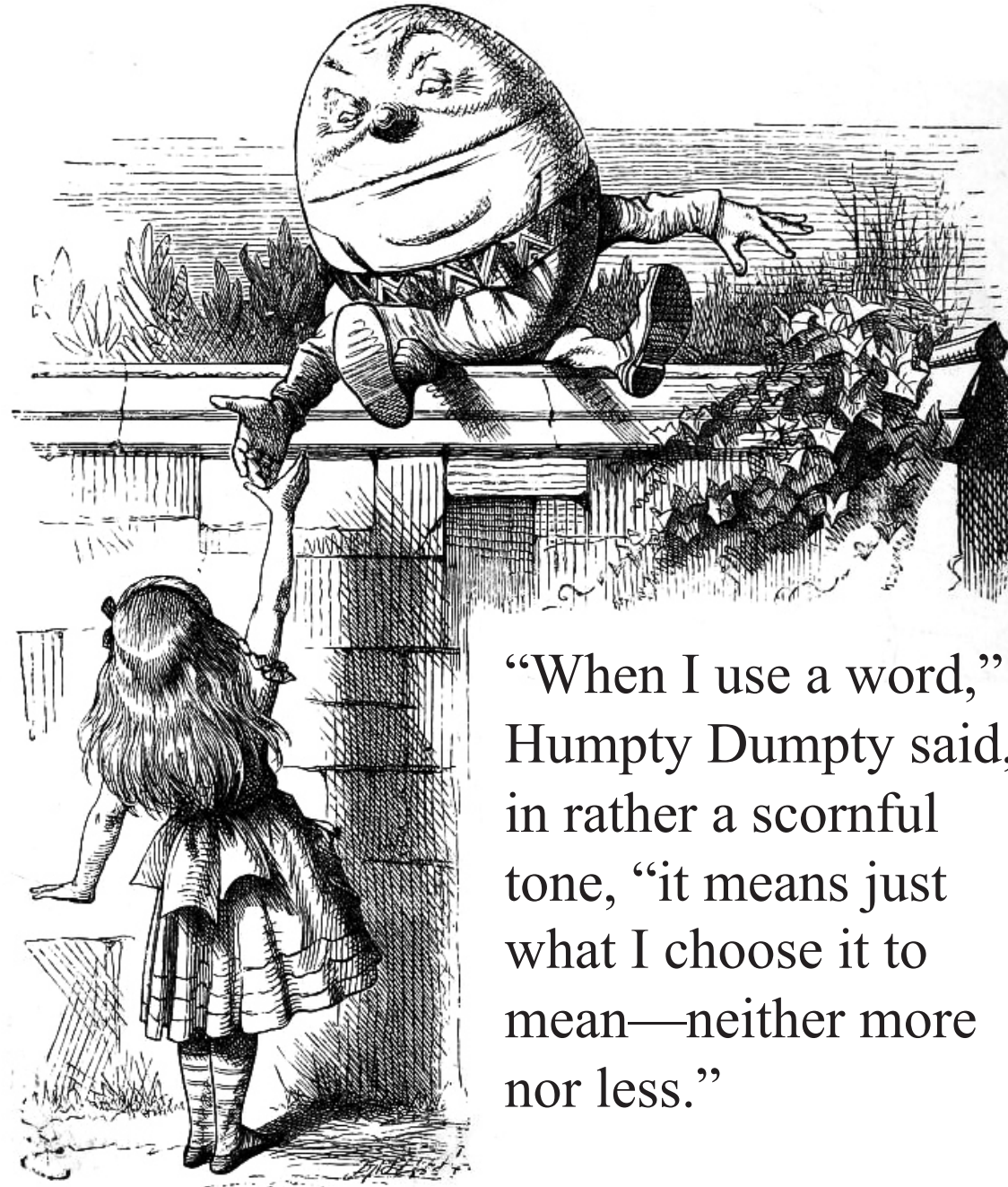UNIVERSITY OF
Southampton

The World Wide Web

COMP3220 Web Infrastructure

Dr Nicholas Gibbins – nmg@ecs.soton.ac.uk

What is the World Wide Web?

"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less."

4

Before the Web

- A user typed a host address into a client program
- The client communicated with a file server using the File Transfer Protocol (FTP)
- The user typed commands into the client:
  - To navigate to the right directory
  - To specify whether the file being transferred was binary or ASCII
  - To get the right file
- The server sent a file back
- The client stored the file on the hard disk
- The user printed the file, or used a separate viewer

```
juno:~ nmg$
```

The Web experience

- A user clicks on a link in a browser
- (the browser talks to a web server)
- (the server sends a document back)
- The browser displays the document
- The user clicks on another link (etc)
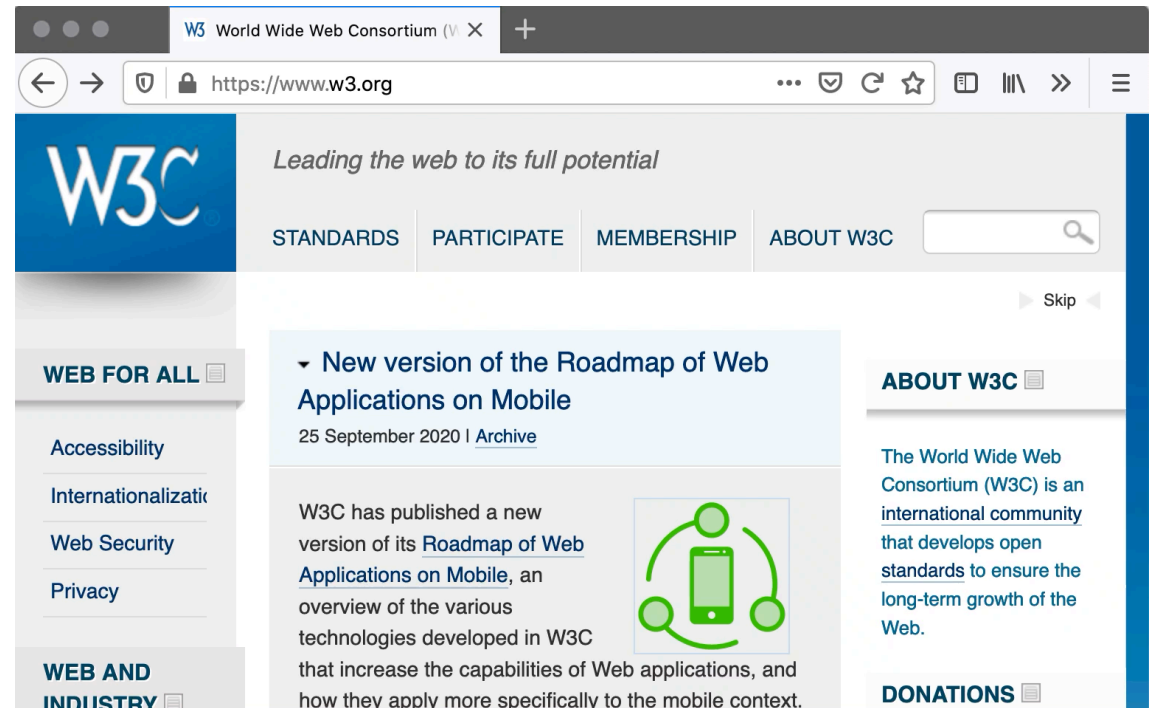
Web evolution

The Web is for scientists (1991-1995)
- Document-centric
- Limited interactivity

The Web is for commerce (1996-2000)
- Invention of Secure Sockets
- The dot-com bubble
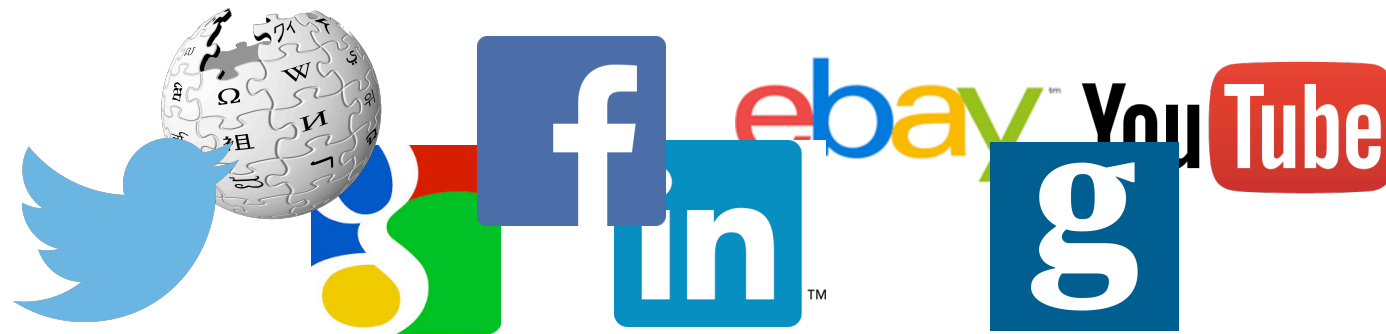
The Web is for users (2000-2005)
- Web 2.0 (just a marketing slogan?)
- Emphasis on user-generated content
- Web browser as rich client

The Web as application platform (2005-)

What is the World Wide Web?

A distributed information system that provides access to hypertext documents and other objects of interest

We have a general name for these objects of interest:
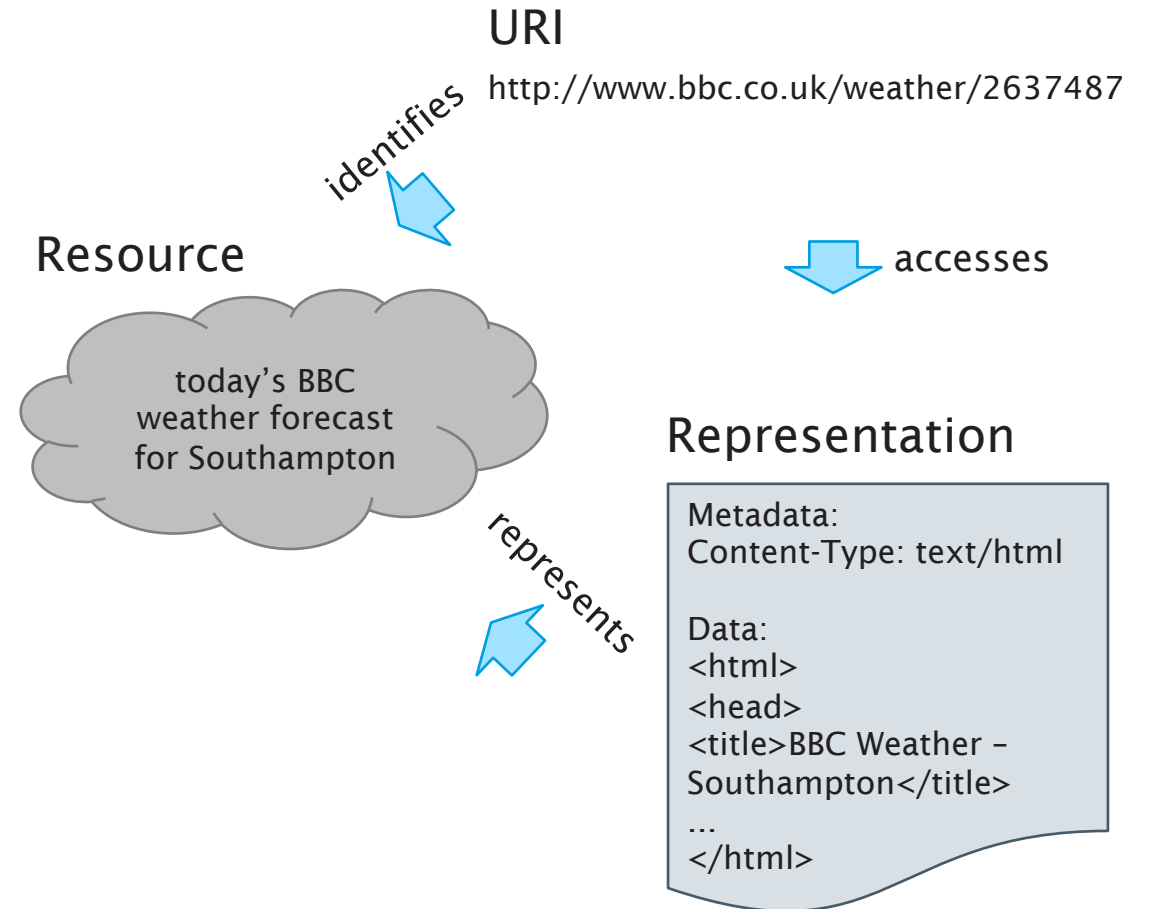
resources

What is a resource?

"Familiar examples [of resources] include an electronic document, an image, a source of information with a consistent purpose (e.g., 'today's weather report for Los Angeles'), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., 'parent' or 'employee'), or numeric values (e.g., zero, one, and infinity)."

Berners-Lee, T. et al (2005) *Uniform Resource Identifier (URI): Generic Syntax*. RFC3986.

Web Architecture

Resources are *identified* by URIs (Uniform Resource Identifiers)

Resources have *representations* in different formats (HTML, text, PDF)

Resources can be *interacted* with using network protocols (HTTP)

URI

http://www.bbc.co.uk/weather/2637487

identifies

accesses

Resource

today's BBC weather forecast for Southampton

Representation

Metadata:
Content-Type: text/html

Data:
<html>
<head>
<title>BBC Weather –
Southampton</title>
...
</html>

represents

Web principles

- All entities of interest should be identified by URIs
- All URIs should be resolvable (i.e. you can use them to fetch something)
- When you resolve a URI, you get some data about the identified resource
- Data should be provided using standard formats
- Data should be linked with other data

5 Stars of Linked Data (2010)

★   Available on the Web (in whatever format) under an open licence

★★   As above, but as machine-readable structured data
            (e.g. Excel instead of an image of a table)

★★★   As above, but in a non-proprietary format
            (e.g. CSV instead of Excel)

★★★★   As above, but using W3C standards (RDF, SPARQL) to identify things,
            so that others can point at your data

★★★★★   As above, but linked to other people's data to provide context

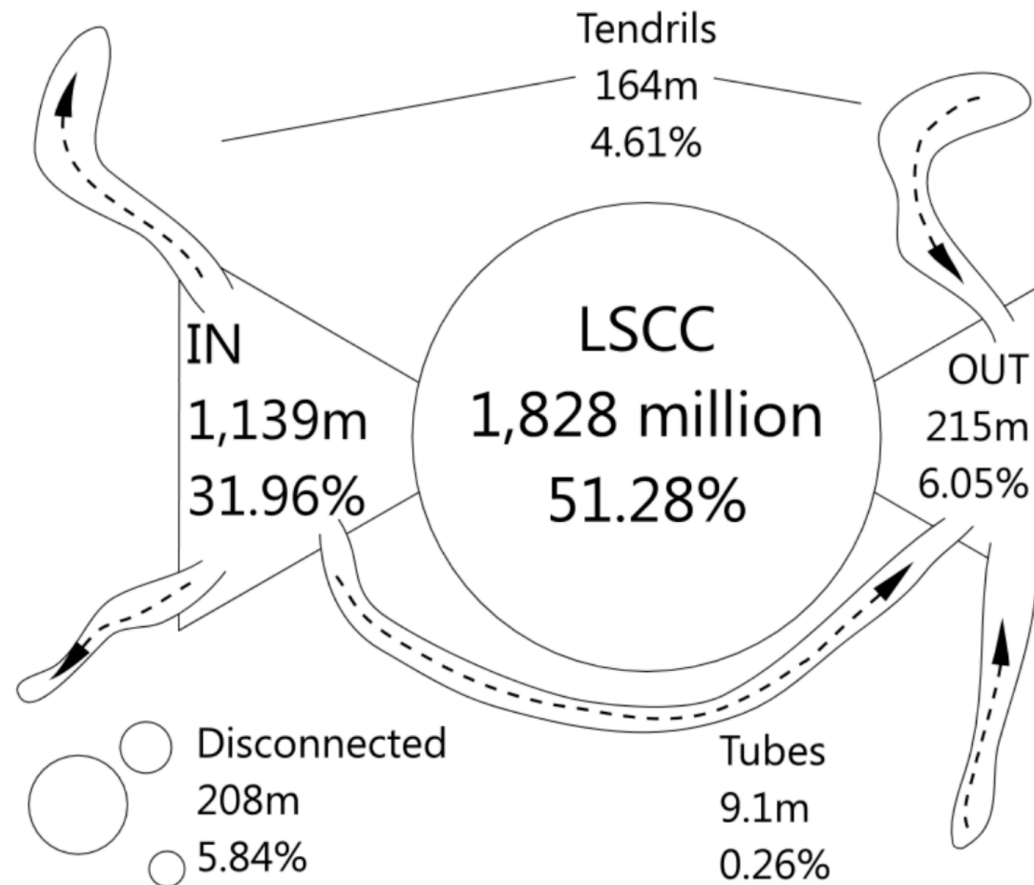# What is the World Wide Web?

Web growth
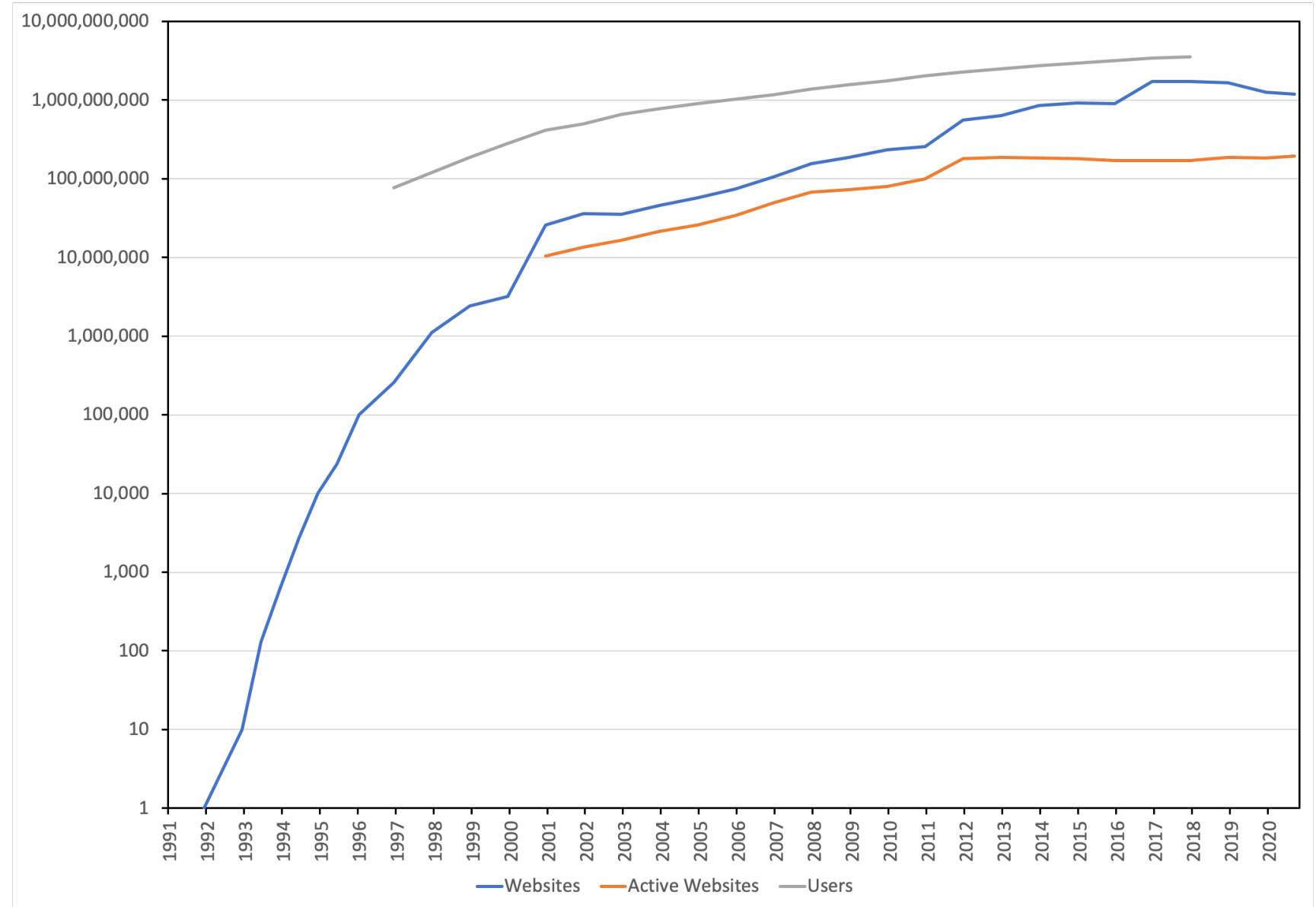
Measuring the Web

How many webpages are there?
- Harder to count than websites (deep Web – not linked)
- Estimate size of the indexed Web using search engines
- Around 800 million in 1999 (compare with ~3 million websites)
- At least 10 billion in 2005 (compare with ~34 million active websites)
- At least 1 trillion in 2016 (compare with ~170 billion websites)

What is the diameter of the Web?
- How many links do you need to follow to travel between an arbitrary pair of webpages?
- Longest shortest finite path
- Even harder to measure…
- Estimated at 19 in 1999

Lawrence, S. and Giles, C.L. (1999) Accessibility of information on the Web. Nature, 400, pp. 107-109.
Albert, R. and Barabási, A-L. (1999) Diameter of the World-Wide Web. Nature, 401, pp.130-133.

The shape of the Web

Tendrils
164m
4.61%

IN
1,139m
31.96%

LSCC
1,828 million
51.28%

OUT
215m
6.05%

Disconnected
208m
5.84%

Tubes
9.1m
0.26%

Broder, A. et al (2000) *Graph Structure in the Web*. Computer Networks 33, pp. 309-320.
Meusel, R. et al (2015) *Graph Structure in the Web – Revisited*. In Proceedings of the 14th International World Wide Web Conference. pp. 427–432.

Next Lecture: Hypertext