CSC200: Lecture 23

Allan Borodin

This lecture and Announcements

Announcements

1. We will start this term with one tutorial section to be held in the lecture room SS 1069. If attendance warrants a second section, we can return to having two tutorial sections. Holding just one tutorial section allows us more time for grading.
2. Quiz 5 therefore will take place this Friday in SS 1069.
3. Scope of quiz is stable matching and the Gale-Shapley algorithm.

Todays lecture

- The world wide web WWW (Ch13)
- Perhaps starting today or next lecture: search engines (Ch 14)

New topic: Information networks

- Somewhat new topic but we will see that the topic soon becomes intertwined with social and economic networks.

- Will consider networks where the nodes are pieces of information (e.g. web pages) connected by links (e.g. hyperlinks) that indicate some relation between these pieces of information.

- This is in contrast to the bulk of our discussions to date where nodes have been people, organizations, intersections (in a traffic network) or agents (buyers and sellers) in a market.

- Of course when we restrict web pages to say personal home pages or facebook pages, then we again have a social network. But for this and the next one or two lectures we will be focusing on information networks.

- Information networks are also different in that now inherently the links are directed. (Of course, a social network can also be directed.)

A history of related "great ideas"

- There are a number of related developments relating to the world wide web WWW and some interesting history leading up to the actual deployment of this set of great ideas.
- The internet, the web and hypertext; search engines (to be discussed in Chapter 14).
- The internet was developed in the 1960s in response to the military and political need for a communications network that could be fault tolerant to equipment failures, and physical attacks on the network. One might now add "software attacks" on the network.
- The internet is based on decentralized packet routing in contrast to what was the existing telephone circuit routing.

Very brief history of the web

- The web is an application (with agreed upon protocols) developed and implemented (1989-1991) by Tim Berners-Lee for sharing information over the internet.

- Individual sites host publicly accessible web pages. (Of course, a given site can have limited access where the content is only accessible to authorized users.)

- Beyond that, the web utilizes the great idea of hypertext links that embody the relationships between web pages.

- As the text says, the idea to organize web pages as a network linked together by hypertext links was both inspired and non-obvious, and has had a profound impact. $\Rightarrow$ GREAT IDEA

- Why non-obvious?

Very brief history of the web

- The web is an application (with agreed upon protocols) developed and implemented (1989-1991) by Tim Berners-Lee for sharing information over the internet.

- Individual sites host publicly accessible web pages. (Of course, a given site can have limited access where the content is only accessible to authorized users.)

- Beyond that, the web utilizes the great idea of hypertext links that embody the relationships between web pages.

- As the text says, the idea to organize web pages as a network linked together by hypertext links was both inspired and non-obvious, and has had a profound impact. $\Rightarrow$ GREAT IDEA

- Why non-obvious? Could be organized alphabetically or as in say traditional libraries by taxonomy and then alphabetically.

The origins of hypertext

- The intellectual origins of hypertext can be found in
  - the prophetic 1945 article and Memex proposal by Vannevar Bush (see http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/)
  - Ted Nelsons pre-web 1965-1974 vision for a hypertext-enabled publishing network called Xanadu (see http://dc-mrg.english.ucsb.edu/conference/CNCSC/multimedia/documents/wardrip-fruin.pdf)

- Vannevar Bush: "Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. The human mind does not work that way. It operates by association."

The origins of hypertext

- The intellectual origins of hypertext can be found in
  - the prophetic 1945 article and Memex proposal by Vannevar Bush (see http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/)
  - Ted Nelsons pre-web 1965-1974 vision for a hypertext-enabled publishing network called Xanadu (see http://dc-mrg.english.ucsb.edu/conference/CNCSC/multimedia/documents/wardrip-fruin.pdf)

- Vannevar Bush: "Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. The human mind does not work that way. It operates by association."

- When you forget something (e.g. a name or place), how do you go about trying to remember (if you don't have others to ask)?

The web as a great idea

- Two essential aspects to the web:
  1. an agreed upon method for accessing publicly accessible pages on individual sites
  2. hypertext

- Before hypertext, the idea of embedding information chunks in a network occurs in citation indices. (See Fig 13.3 in textbook). But such networks are generally directed *acyclic* graphs (dags) due to the fixed time of creation.

- This is in contrast to the cross references in an encyclopedia. (See Fig 13.4 in textbook)

- As the text says, Bush's article foreshadowed the web as a universal encyclopedia, and as a giant socio-economic system.

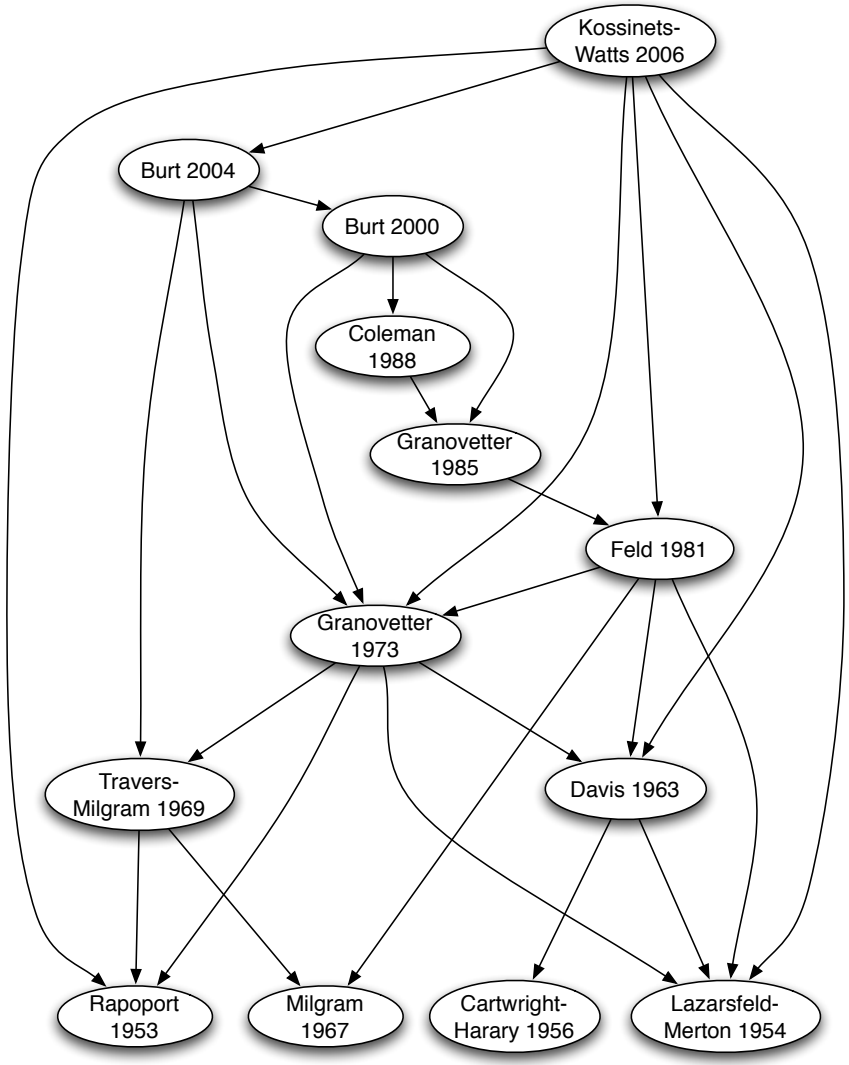


Figure : The network of citations among a set of research papers forms a directed acyclic graph that, like the Web, is a kind of information network. In contrast to the Web, the passage of time is much more evident in citation networks, since their links tend to point strictly backward in time. [Fig 13.3, textbook]

A small Wikipedia cross-reference network

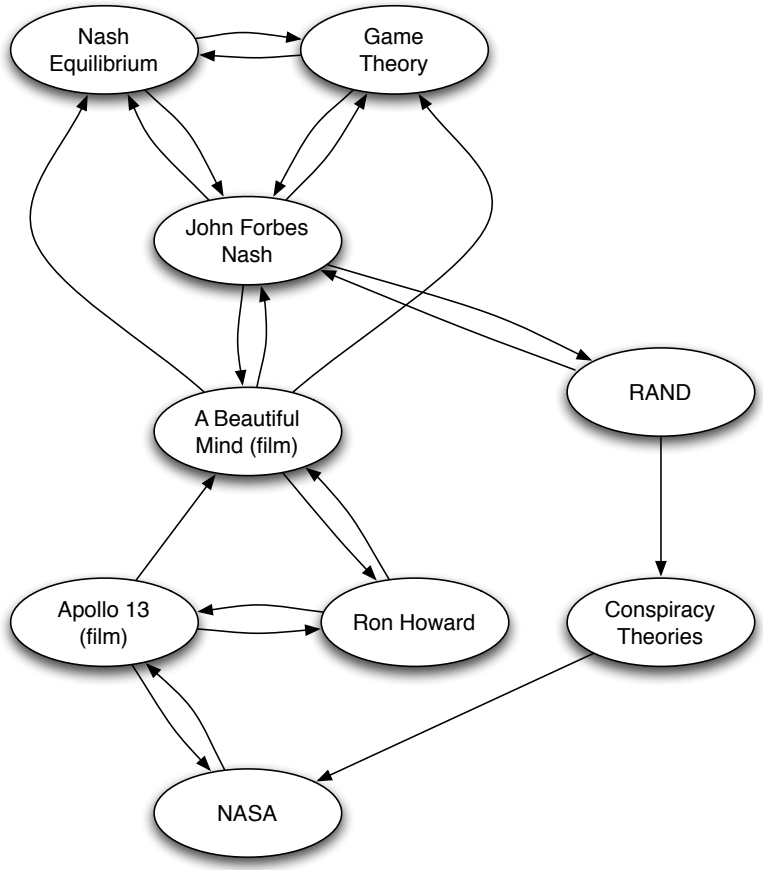


Figure : This is another kind of information network that can be represented as a directed graph. The figure shows the cross-references among a set of Wikipedia articles on topics in game theory, and their connections to "related topics" including popular culture and government agencies. [Fig 13.4, textbook]

The evolution of the web

- In the first decade of the web most links were navigational links, enabling the movement between document pages.
- As the web has evolved it has now become a "web of data and social networks" [Hall and Tiropanis 2012]
- Furthermore, many links now are used to trigger programs to be executed on the computer hosting the page. For example "add to shopping cart". These are called transactional links.
- Our emphasis is on the navigational links that define the web as a directed graph. Links induce a directed graph because my linking to a web page X on my page Y does not imply that X knows about me and X may or may not wish to cross reference my page Y even if they do know about me.

What does the web look like as a directed graph?

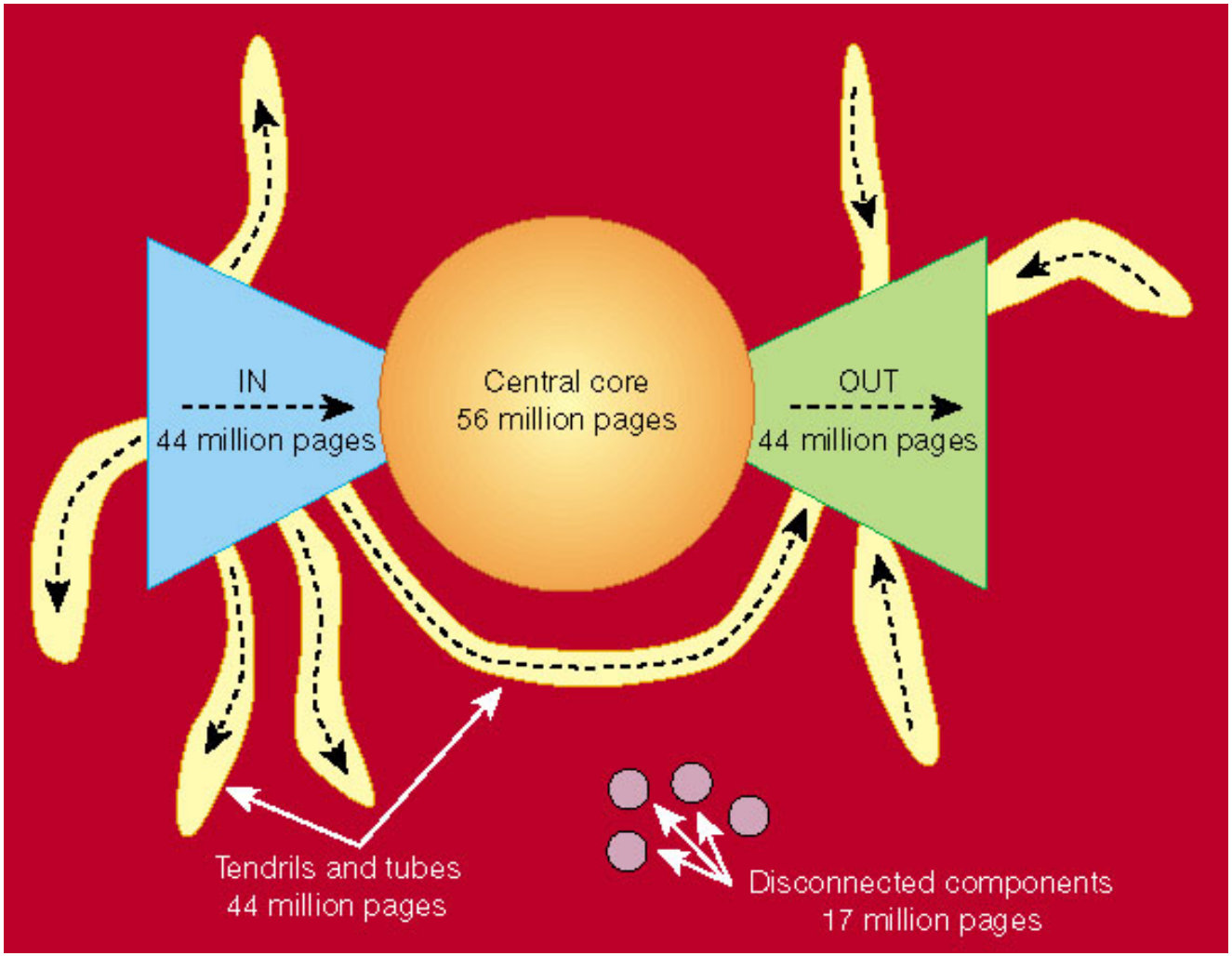
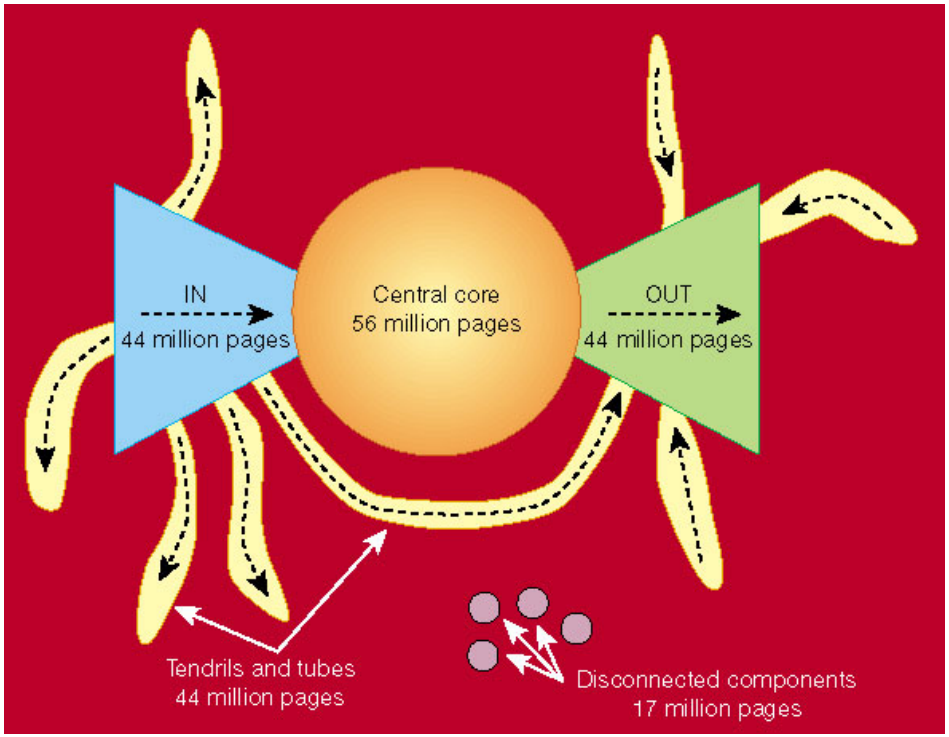


Figure : A schematic picture of the bow tie structure of the Web. Although the numbers are now outdated, the structure has persisted. [Fig 13.7, textbook]

The bow tie structure persists?



- The web of ∼1999 as depicted by Broder (now at Google).
- This bow tie structure has
  - one giant strongly connected component; this is the knot in a bow tie.
  - set of nodes in "IN" and "OUT", likely made up of many small connected components.
  - tendrils, tubes, and disconnected nodes.
- Claim in some studies: The bow tie structure persists even though the web is always growing dynamically in size and diversity of content.

Alternative visualizations of the web graph

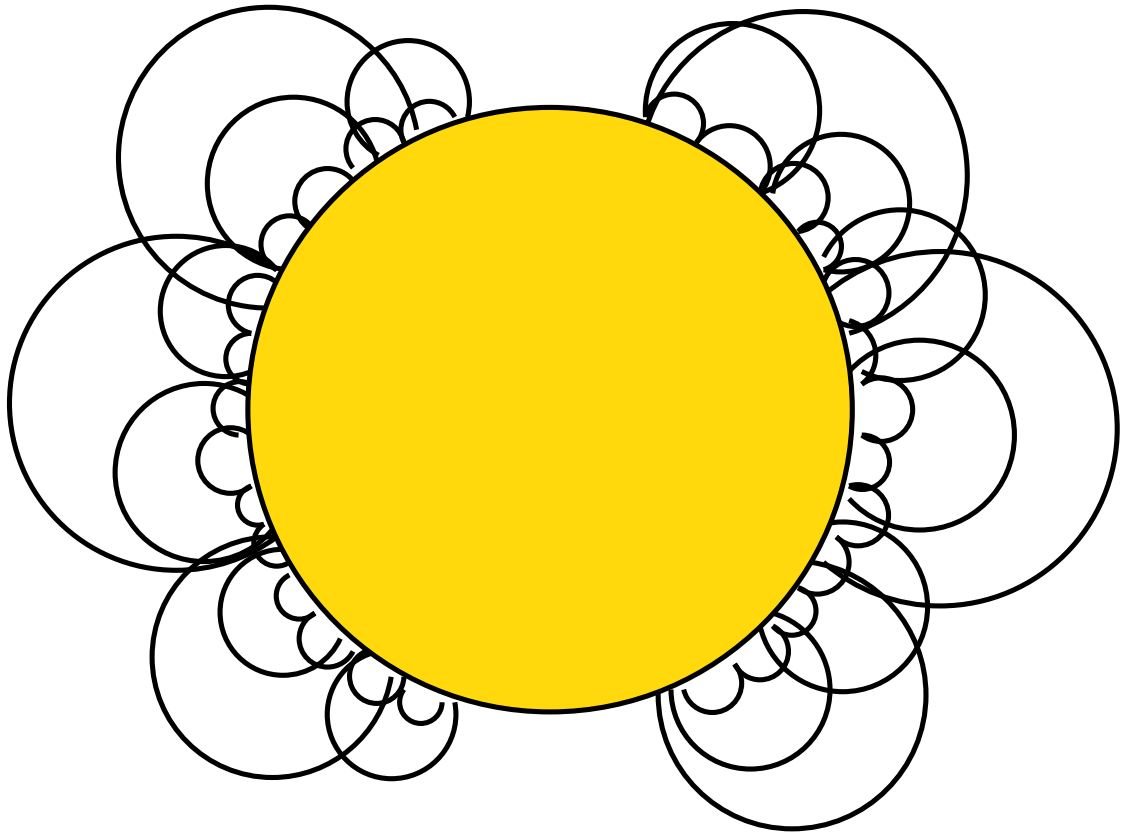


Figure 4: The daisy structure of the Web

Figure : The daisy view of the Web. [Donato et al, WebDB 2005]

A visualization of the Chinese web graph

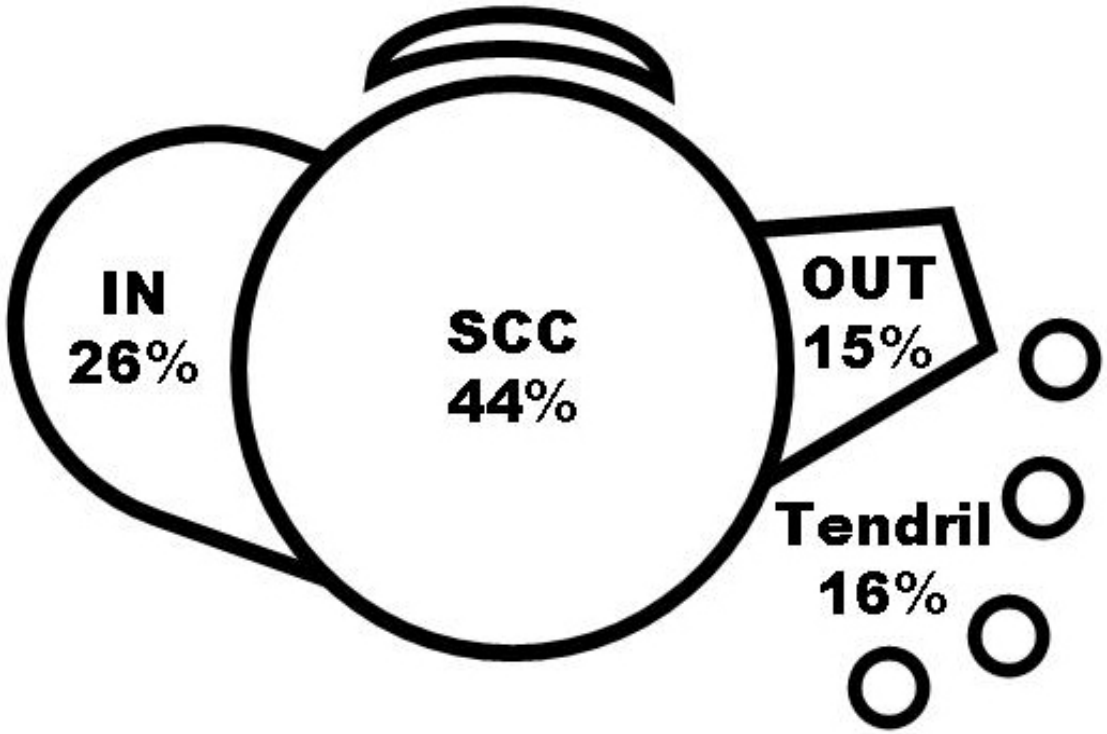


Figure 1. A Teapot Graph of Chinese Web

Figure : The teapot view of the Web. [Zhu et al, WWW 2008]

Recent obervations on web graph structure

Meusel et al [WWW 2014 conference] analyze a 2012 crawl of the web containing over 3.5 billion web pages and 128.7 billion links. They observe:

- Some observed features of the web graph depend on the crawling process and hence cannot be called structural graph properties.
- The existence of a giant strongly connected component (i.e. the central core in Broder's bow-tie) persists and hence can be understood to be a structural property.
- The proportion of nodes that can reach and be reached from the central core (i.e. the "IN" and "OUT" nodes in the bow tie) depends on the particular crawl.
- The distributions for node in-degrees, out-degrees and sizes of strongly connected compnents do not obey *power law distributions* (as observed in smaller crawls) but still appear to be "heavy-tailed". **Note:** We will discuss power law distributions in Chapters 18 and 20.

How big is the web?

- The first Google index (1998) had 26 million pages.
- In a July 2008 blog, Google claimed that there are 1 trillion unique URLs (Unique Resoure Identifers).
  http://googleblog.blogspot.ca/2008/07/we-knew-web-was-big.html
- More recently (2013) [see
  http://www.google.com/insidesearch/howsearchworks/thestory/] the claim is that there are 60 trillion individual pages.
- But Google does not index all of these pages. And amongst these pages there are duplicates or near duplicates.
- Statistics Brain Reearch Insititute
  [http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/] claims that Google indexes 30 trillion
  pages.

Some more conservative estimates

There are many (often inconsistent) estimates and some report a much smaller size of the web.

- One December 2013 estimate (http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html) states that Google indexes about 43 billion pages, Bing about 16.5 billion pages. These are often very rough estinates based on word occurences. And amongst these pages there are duplicates or near duplicates.
- In [http://www.worldwidewebsize.com/], the claim is that "The Indexed Web contains at least 4.82 billion pages (Tuesday, January 12, 2016)" Is this a typo? They provide a graph for the size of the web (from Oct 15, 2015 to January 12, 2016) that shows a mostly constant web size of approximately 47 billion pages.
- Note that these two estimates are off by almost three orders of magnitude from the 30 trillion estimate.

How is the web size estimated

- Note that in principle there are inflinitely many web pages since, for example, online calendars that have a next day or next month link can generate an infinite number of pages.
- How is the indexed web size estimated? According to Ask.com: "The size of the index of a search engine is estimated on the basis of a method that combines word frequencies obtained in a large offline text collection (corpus), and search counts returned by the engines. Each day 50 words are sent to all four search engines. The number of webpages found for these words are recorded; with their relative frequencies in the background corpus, multiple extrapolated estimations are made of the size of the engine's index, which are subsequently averaged."

Continuing evolution of the web: web 2.0

- Major forces:
  - More common authoring styles (wikipedia) $\Rightarrow$ more shared content
  - The improved quality of search engines (Chapter 14)
  - Personal online data (gmail) hosted by large companies; one aspect of the cloud
  - Links between people (facebook) rather than documents
  - Graphical data and searching on pictures vs. text
  - Crowd sourcing

- The web as a major source of advertising (Chapter 15)

- The "semantic web", with "meta data enabling automated agents (i.e. programs) to access data more intelligently and perform tasks on behalf of humans".
  - Ex: "I need a non business flight...". What does such an automated travel agent need to know?
  - Question: What would be your examples of automated agents?