The Structure of the Web

Moreno Marzolla
Dip. di Informatica—Scienza e Ingegneria (DISI)
Università di Bologna

http://www.moreno.marzolla.name/
A Brief History of the Web
Vannevar Bush and the Memex

- Memex was an *hypotetical* electro-mechanical system capable of storing information and links between items on (punched) microfilms
  - Vannevar Bush, “As We May Think”, The Atlantic Monthly, July 1945
- Memex was never implemented
  - Nevertheless, all basic ideas on hypertexts were there

Slide credit: Fabio Vitali

http://www.unc.edu/~elese/diglib/index.htm

http://www.ibiblio.org/pioneers/bush.html
A Brief History of the Web
Ted Nelson and Project Xanadu(R)

• In the '60s, Ted Nelson proposed an integrated system (Project Xanadu®) for publishing documents.
  – Bi-directional, “unbreakable” links
  – Transclusion (embedding) of arbitrary pieces of documents within other documents
  – Copyright management
  – Everyone can publish, mix, annotate any document

• Some concepts from Xanadu® were later used, in a very simplified form, for the World Wide Web
A Brief History of the Web
Douglas Engelbart

- Douglas Engelbart built an innovative system for text editing, video conferencing, collaborative text editing based on mouse, windows and most of the things we know (and use) today
- The system was demonstrated during a presentation in 1968 (“the mother of all demos”)
  - [https://www.youtube.com/watch?v=VScVgXM7lQQ&list=PL76DBC8D6718B8FD3](https://www.youtube.com/watch?v=VScVgXM7lQQ&list=PL76DBC8D6718B8FD3)

Slide credit: Fabio Vitali
What is the Web today?

- A way for sharing (publishing) information
  - A collection of independent Web Servers
  - Each Web Server hosts a number of resources (mostly, but not only, HTML pages)
  - Each resource is identified by a URI

- A way for browsing the shared information
  - Web browsers can render HTML pages and their embedded content
  - Browsers allow users to navigate the Web following hyperlinks
A set of Web resources

- List of my publications and my CV
- My CV: I work at the University of Bologna
- UniBO Web Page: People
- Publications: Optimized training of support vector machines for the Cell processor
- Cell processor resources at IBM research
- Other resource
- HTML page
- people at UniBO: ... Moreno Marzolla
  Complex Systems
- cellsvm.pdf
A set of Web resources

List of my publications and my CV

My CV: I work at the University of Bologna

UniBO Web Page: People

Cell processor resources at IBM research

People at UniBO: Moreno Marzolla

Optimized training of support vector machines for the Cell processor

cellsvm.pdf
Hyperlinks

- Elements of HTML pages can be annotated with hyperlinks pointing to other resources (including other HTML pages)
- This organization is given for granted today, but is both inspired and non-obvious
  - Many different ways or organizing data catalogs: by name, by topic, hierarchical...
  - Hyperlinks allow non-linear knowledge structuring, such that logical relationships in the text become first-class citizens
Precursors of hypertext: Citation networks

- The network of citations among scientific papers form a directed graph
- Unlike the Web, links tend to point strictly backward in time
Precursors of hypertest: cross-references in a encyclopedias

Some cross-references from wikipedia
Back to the Web

- Our simple idea of the Web made of HTML pages and hyperlinks was mostly correct when the Web was born, but is too simplistic today.
- In the beginning, Web servers had a passive role, just providing static content.
- Now some links can be associated to dynamic actions:
  - “Buy it now”
  - “Add to shopping cart”
  - “Update my calendar”
  - “Upload my photo”
Two kind of links

- **Navigational**
  - Allow navigation from one page to another

- **Transactional**
  - Perform actions
  - E.g., “buy it now” will
    - charge your credit cart
    - start shipping of the product to your home address
    - move you to a separate Web page containing a purchase receipt
    - The goal of “buy it now” is not just to move you to the receipt page

- A lot of content on the Web has a transactional nature, but is linked together by a navigational “backbone”
  - It is reachable via relatively stable Web pages connected to each other by more traditional navigational links.
The Web as a directed graph
The Web as a **directed** graph

A (directed) path
The Web as a **directed** graph

- I'm a student at Univ. of X
- My song lyrics
- Classes
- Networks
class blog
- Networks
- I teach at Univ. of X
- Blog post about college rankings
- Company Z's home page
- Our Founders
- Press Releases
- Contact Us
- USNews College Rankings
- USNews Featured Colleges
- I'm applying to college

**Strongly Connected Components (SCCs)**
Strongly Connected Components

- We say that a *strongly connected component* (SCC) in a directed graph is a subset of the nodes such that:
  - (i) every node in the subset has a path to every other node; and
  - (ii) the subset is not part of some larger set with the property that every node can reach every other.
Computing SCCs

**Algorithm** \( \text{SCC}(\text{Graph } G, \text{ node } x) \rightarrow \text{list of nodes} \)

1. \( L := \text{empty list of nodes} \)
2. Execute \( \text{DFS}(G, x) \) marking visited nodes
3. Compute the transposed graph \( G^T \) (invert direction of all edges in \( G \))
4. Execute \( \text{DFS}(G^T, x) \), inserting in \( L \) all visited nodes which were marked during step (1)

\[ \text{return } L \]

- Algorithm \( \text{SCC}(G,x) \) identifies the strongly connected component containing node \( x \) in time \( O(n+m) \)
- It is possible to identify all strongly connected components in time \( O(nm+n^2) \)
  - Actually you can do that in \( O(n+m) \) using a more complex algorithm (e.g., Tarjan or Gabow algorithms)
A Picture of the Web

  - Based on a Web crawl performed by AltaVista on may 1999
  - The crawl is based on a large set of starting points accumulated over time from various sources, including voluntary submissions
  - The crawl proceeds in roughly a BFS manner
  - **203 million** URLs
  - **1466 million** links
A “picture” of the Web (in 2000)

Single “giant” SCC

- It contains about 56 million nodes
- Why a SCC component?
  - Many search engines contain a large directory of Web sites
  - From most Web sites there are links back to the search engines
Why a *single* SCC?

- Suppose there are two SCC, say A and B
  - It would be possible to merge A and B into a single SCC if there is just a link from A to B, and another from B to A
  - If A and B are sufficiently large, there is a high probability that such links exist
Graph structure of the Web

• From the “bow tie” picture alone we have a high-level view of the structure of the Web

• We now dig a bit deeper into the Web graph
  – In/Out degree distribution
  – Connectivity distribution
  – Diameter
  – Resiliency to edge removal
Degree distribution

- In- and Out-degrees follow a power law

\[ P[X = k] \sim \frac{1}{k^\alpha} \]

for some \( \alpha > 1 \)

\( \alpha \approx 2.1 \)

\( \alpha \approx 2.72 \)

Why power law?

- One might ask why in/out degree distribution obeys the power law $P[X = k] \sim 1/k^\alpha$

- *Rich-gets-richer effect*
  - A new node $X$ joins the network
  - $X$ connects with $L$ of existing nodes
    - Node $X$ connects to node $Y$ with probability proportional to the degree of $Y$
  - Thus, nodes with more edges will likely “attract” new connections

- The resulting network is called *scale-free network*
Connectivity analysis

Weakly connected (undirected) components

Strongly connected components

α ≈ 2.5

WCC containing 91% of Web pages

Giant SCC 56 million nodes, 28% of Web pages
Other properties

<table>
<thead>
<tr>
<th>$k$</th>
<th>1000</th>
<th>100</th>
<th>10</th>
<th>5</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (millions)</td>
<td>177</td>
<td>167</td>
<td>105</td>
<td>59</td>
<td>41</td>
<td>15</td>
</tr>
</tbody>
</table>

Size of the largest surviving weak component when links to pages with in-degree at least $k$ are removed from the graph.

- (Direct) diameter of the SCC: >28
- Maximum finite shortest path length: >503
- With 75% probability, there is no directed path from a random start node to a random end node
- The average directed path length is 16
Exponential vs scale-free network

Exponential networks are homogeneous: all nodes have the same average number of links

Scale-free networks are not homogeneous: few nodes are highly connected, most nodes have just few links

Change in diameter $d$ as a function of the fraction $f$ of removed nodes.
Cluster size distribution for various $f$

$f = \text{fraction of failed/attacked nodes}$