Networks
What is a system: more formally

- \( S = (T, R) \)
  - \( S \): a System
  - \( T = \{A_1, A_2, ..., A_n\} \)
    - A family of sets of things: thinghood
  - Cartesian Product
    - Set of all possible associations of elements from each set, i.e. all \( n \)-tuples
    - \( \{A_1 \times A_2 \times ... \times A_n\} \)
  - \( R \): a (or set of) relation(s)
    - Subset of the Cartesian product of some set of sets: Systemhood
    - Many relations \( R \) can be defined on the same \( T \)

From Klir [2001]

jbollen@indiana.edu
http://informatics.indiana.edu/jbollen/IS01
Types of relations

- Equivalence: (~exact same features)
  - Reflexive,
  - Symmetric,
  - transitive

- Compatibility: (~synonyms)
  - Reflexive,
  - symmetric

- Partial orderings:
  - Reflective,
  - anti-symmetric,
  - transitive (t₁ >= t₂)

- Strict orderings:
  - anti-reflexive,
  - Antisymmetric,
  - transitive (t₁ > t₂)
"Any fact becomes important when it's connected to another." — Umberto Eco, Foucault's Pendulum

"We are all now connected by the Internet, like neurons in a giant brain." — Stephen Hawking

Image 2.1
The bridges of Königsberg.

From the contemporary map of Königsberg (now Kaliningrad, Russia) to Euler's graph. The graph constructed by Euler consists of four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. Euler showed in 1736 that there is no continuous path that would cross seven the bridges while never crossing the same bridge twice. The people of Königsberg agreed with him, gave up their fruitless search and in 1875 they built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links and it became rather straightforward to find the desired path.

jbollen@indiana.edu
http://informatics.indiana.edu/jbollen/I501

Barabasi (2012)
Barabasi (2012)

Bearman, Moody, and Stovel, 2004

Pablo Kaluza et al (2010)

<table>
<thead>
<tr>
<th>NETWORK NAME</th>
<th>NODES</th>
<th>LINKS</th>
<th>DIRECTED/UNDIRECTED</th>
<th>N</th>
<th>L</th>
<th>&lt;K&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>routers</td>
<td>Internet Connections</td>
<td>Undirected</td>
<td>192,244</td>
<td>609,066</td>
<td>2.67</td>
</tr>
<tr>
<td>WWW</td>
<td>webpages</td>
<td>links</td>
<td>Directed</td>
<td>325,729</td>
<td>1,497,134</td>
<td>4.60</td>
</tr>
<tr>
<td>Power Grid</td>
<td>power plants, transformers</td>
<td>cables</td>
<td>Undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
</tr>
<tr>
<td>Mobile-Phone Calls</td>
<td>subscribers</td>
<td>calls</td>
<td>Directed</td>
<td>36,595</td>
<td>91,826</td>
<td>2.51</td>
</tr>
<tr>
<td>Email</td>
<td>email addresses</td>
<td>emails</td>
<td>Directed</td>
<td>57,194</td>
<td>103,731</td>
<td>1.81</td>
</tr>
<tr>
<td>Science Collaboration</td>
<td>scientists</td>
<td>co-authorships</td>
<td>Undirected</td>
<td>23,133</td>
<td>186,936</td>
<td>16.16</td>
</tr>
<tr>
<td>Actor Network</td>
<td>actors</td>
<td>co-acting</td>
<td>Undirected</td>
<td>212,250</td>
<td>3,054,278</td>
<td>28.78</td>
</tr>
<tr>
<td>Citation Network</td>
<td>papers</td>
<td>citations</td>
<td>Directed</td>
<td>449,673</td>
<td>4,707,958</td>
<td>10.47</td>
</tr>
<tr>
<td>E. coli Metabolism</td>
<td>metabolites</td>
<td>chemical reactions</td>
<td>Directed</td>
<td>1,039</td>
<td>5,802</td>
<td>5.84</td>
</tr>
<tr>
<td>Yeast Protein Interactions</td>
<td>proteins</td>
<td>binding interactions</td>
<td>Undirected</td>
<td>2,018</td>
<td>2,930</td>
<td>2.90</td>
</tr>
</tbody>
</table>
The web graph...
As we may think

AS WE MAY THINK By VANNEVAR BUSH

The Atlantic Monthly, July 1945

“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......associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing.”

“Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them...”

jbollen@indiana.edu
http://informatics.indiana.edu/jbollen/1501
Douglas Engelbart

This session is entirely devoted to a presentation by Dr. Engelbart on a computer-based, interactive, multiconsole display system which is being developed at Stanford Research Institute under the sponsorship of ARPA, NASA and RADC. The system is being used as an experimental laboratory for investigating principles by which interactive computer aids can augment intellectual capability. The techniques which are being described will, themselves, be used to augment the presentation.

The session will use an on-line, closed circuit television hook-up to the SRI computing system in Menlo Park. Following the presentation remote terminals to the system, in operation, may be viewed during the remainder of the conference in a special room set aside for that purpose.

http://www.youtube.com/watch?v=yJDv.zdhzM

“mother of all demos” – dec 9, 1968
“longest-running vaporware project in the history of computing” – Wired Magazine

- XanaduSpace, 2008 (reached prototype)
- Gold and Green were fouled by politics and infighting.
Informatics: and computing

http://informatics.indiana.edu

Organisation Européenne pour la Recherche Nucléaire
CERN
European Organization for Nuclear Research

World Wide Web

Tim Berners-Lee (CERN): April 30, 1993

Hypertext Markup Language
The hypertext system to rule them all...

- 1991: Tim Berners-Lee deploys first Web Server
- 1994: World Wide Web Consortium (W3C) founded jointly between CERN and MIT
- 1995: Netscape IPO’s for $1.5 Billion, ushering in the “dot com” era
- 2004: Google IPO’s for $4 Billion
- 2004: Tim O’Reilly coins the term “Web 2.0”

April 2008: 165,000,000 Websites

Web Timeline and Number of Websites

figure: Jase Wilson
Modeling the web graph

- Straightforward graph model follows from page-level hypertext linking
- Directed graph: $G = (V,E)$
  - $V$ is set of $N$ web pages or sites
  - $E$ is subset of $V^2$ hyperlink
  - Sparse: $|E| \ll |V^2|$
Practically speaking...

- How do you study a growing, dynamic graph of $10^9$ nodes without a central agency that monitors, vets, approves, and translates contents into a registry?

Some early data

  - 200M pages, 1.5B links
- Yahoo/altavista: 2002 web graph snapshot
  - + 1.4 billion public web pages, 6.6B links
- SNAP Google web graph (2002)
  - 875k nodes, 5.1M edges
  - 3.5B, 1.7B pages and 128B, 64B hyperlinks, resp.
  - http://webdatacommons.org/hyperlinkgraph/
Web graph...

- **Directed:** $(v_i, v_j)$ does not imply $(v_j, v_i)$ and vice versa (hyperlinks)
- **Distance** $D(v_i, v_j)$: length of shortest path
  - No path: $D(v_i, v_j) = \infty$
  - Direct edge: $D(v_i, v_j) = 1$
- **Strongly connected component:** subgraph $G' = (V', E')$ of $G$ such that
  - $V'$ subset of $V$
  - For all pairs $v_i$ and $v_j$ in $V'$ there exists a path that connects $(v_i, v_j)$
- **Weakly connected component:** subgraph $G' = (V', E')$ of $G$ such that
  - $V'$ subset of $V$
  - For all pairs $v_i$ and $v_j$ in $V'$ there exists an undirected path that connects $(v_i, v_j)$
Graph properties

A few properties that we think are particularly interesting:

- **Diameter**: longest shortest path in graph
- **Average shortest path length (ASPL)**
- **Degree distributions**:
  - **Degree**: number of in and out-links adjacent to a vertex
  - **PDF or CDF**
- **Clustering coefficient**:
  - Average density of all node neighborhoods (Strogatz-Watts)
Figure 1 Distribution of links on the World-Wide Web. a, Outgoing links (URLs found on an HTML document); b, incoming links (URLs pointing to a certain HTML document). Data were obtained from the complete map of the nd.edu domain, which contains 325,729 documents and 1,469,680 links. Dotted lines represent analytical fits used as input distributions in constructing the topological model of the web; the tail of the distributions follows $P(k) \approx k^{-\gamma}$, with $\gamma_{\text{out}} = 2.45$ and $\gamma_{\text{in}} = 2.1$. c, Average of the shortest path between two documents as a function of system size, as predicted by the model. To check the validity of our predictions, we determined $d$ for documents in the domain nd.edu. The measured $\langle d_{\text{nd.edu}} \rangle = 11.2$ agrees well with the prediction $\langle d_{3 \times 10^5} \rangle = 11.6$ obtained from our model. To show that the power-law tail of $P(k)$ is a universal feature of the web, the inset shows $P_{\text{out}}(k)$ obtained by starting from whitehouse.gov (squares), yahoo.com (triangles) and snu.ac.kr (inverted triangles). The slope of the dashed line is $\gamma_{\text{out}} = 2.45$, as obtained from nd.edu in a.

Small world networks


jbollen@indiana.edu
http://informatics.indiana.edu/jbollen/I5OL
We find that the average of $d$ over all pairs of vertices is $<d> = 0.35 + 2.06 \log(N)$ (Fig. 1c), indicating that the web forms a small-world network which characterizes social or biological systems. For $N=8 \times 10^8$, $<d_{\text{Web}}>=18.59$; that is, two randomly chosen documents on the web are on average 19 clicks away from each other.

The logarithmic dependence of $<d>$ on $N$ is important to the future potential of the web: we find that the expected 1,000% increase in the size of the web over the next few years will change $<d>$ very little, from 19 to only 21.

About power-law distributions

\[ P(k) \sim C k^{-a} \]

Usually \( 2 < a < 3 \)

On a log-log plot:
\[ \log(P(k)) = \log(C) - a \log(k) \]
Bell Curve
Most nodes have the same number of links
No highly connected nodes

Power Law Distribution
Very many nodes with only a few links
A few hubs with large number of links
Fig. 5. The degree distributions in the undirected associative network (a), the directed associative network (b), Roget’s thesaurus (c), and WordNet (d). All distributions are shown in log-log coordinates with the line showing the best fitting power law distribution. For Roget’s thesaurus and WordNet, the degree distributions shown are for the word level only.
Table 2
Summary statistics for semantic networks

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Associative Network</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Undirected</td>
<td>Directed</td>
<td>Roget</td>
<td>WordNet</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>words</td>
<td>5,018</td>
<td>5,018</td>
<td>29,381</td>
<td>122,005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>classes</td>
<td>—</td>
<td>—</td>
<td>1,000</td>
<td>99,642</td>
<td></td>
</tr>
<tr>
<td>&lt;k&gt;</td>
<td>words</td>
<td>22.0</td>
<td>12.7</td>
<td>1.7</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>classes</td>
<td>—</td>
<td>—</td>
<td>49.6</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td>3.04</td>
<td>4.27</td>
<td>5.60</td>
<td>10.56</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>.186</td>
<td>.186</td>
<td>.875</td>
<td>.0265</td>
<td></td>
</tr>
<tr>
<td>γ</td>
<td></td>
<td>3.01</td>
<td>1.79</td>
<td>3.19</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td>$L_{\text{random}}$</td>
<td>3.03</td>
<td>4.26</td>
<td>5.43</td>
<td>10.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_{\text{random}}$</td>
<td>4.35E-03</td>
<td>4.35E-03</td>
<td>.613</td>
<td>1.29E-04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $n$ = the number of nodes; $<k>$ = the average number of connections; $L$ = the average shortest path length; $D$ = the diameter of the network; $C$ = clustering coefficient; $\gamma$ = power law exponent for the distribution of the number of edges in undirected networks and incoming connections in directed networks; $L_{\text{random}}$ = the average shortest path length with random graph of same size and density; $C_{\text{random}}$ = the clustering coefficient for a random graph of same size and density.
Beware!

Diameters and tendrils

- 200 million pages and 1.5 billion links
- "In other recent work, [Albert, Jeong, and Barabasi 99] report the intriguing finding that most pairs of pages on the web are separated by a handful of links, almost always under 20, and suggest that this number will grow logarithmically with the size of the web. This is viewed by some as a "small world" phenomenon. Our experimental evidence reveals a rather more detailed and subtle picture: most ordered pairs of pages cannot be bridged at all and there are significant numbers of pairs that can be bridged, but only using paths going through hundreds of intermediate pages. Thus, the web is not the ball of highly-connected spaghetti we believed it to be; rather, the connectivity is strongly limited by a high-level global structure."
Power laws

“confirming previous reports on power laws; for instance, the fraction of web pages with \( i \) in-links is proportional to \( 1/i^{2.1} \). The constant 2.1 is in remarkable agreement with earlier studies at varying scales [Kumar et al. (1) 99, Barabasi and Albert 99].

Kumar (1999) The Web as a graph
Figures 1 and 2: In-degree and out-degree distributions subscribe to the power law. The law also holds if only off-site (or "remote-only") edges are considered.
Figures 3 and 4: In- and out-degree distributions show a remarkable similarity over two crawls, run in May and October 1999. Each crawl counts well over 1 billion distinct edges of the web graph.
Figures 5 and 6: Distribution of weakly connected components and strongly connected components on the web. The sizes of these components also follow a power law.
Fig. 9. Connectivity of the Web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE: i.e., a passage from a portion of IN to a portion of OUT without touching SCC.
“As we discussed above, the directed diameter of the SCC is at least 28. Likewise, the maximum finite shortest path length is at least 503, but is probably substantially more than this: unless a short tube connects the most distant page of IN to the most distant page of OUT without passing through the SCC, the maximum finite shortest path length is likely to be close to $475 + 430 = 905$.”

These results are interesting contrast to those of [Albert, Jeong, and Barabasi 99], who predict an average distance of 19 for the web based on their crawl of the nd.edu site; it is unclear whether their calculations consider directed or undirected distances. Our results on the other hand show that over 75% of time there is no directed path from a random start node to a random finish node; when there is a path, the figure is roughly 16. However, if links can be traversed in either direction, the distance between random pairs of nodes can be much smaller, around 7, on average.
New web graph crawl data

http://webdatacommons.org/hyperlinkgraph/topology.html
Of random graphs...

Erdos-Renyi random graph model

- $G(n, M)$: chose uniformly at random from all possible graphs with $n$ nodes and $m$ edges
- $G(n, p)$: generate graph by taking $n$ nodes, connect nodes randomly, each edge exists with uniform probability $p$
Of random graphs...

Notable characteristics:
$P(\text{deg}(v) = d)$

out of $n$ vertices, $d$ have vertices, with probability $p$, so

$$\binom{n}{d} p^d$$

And $n-d$ nodes have no edges, so

$$(1-p)^{n-d}$$

therefore

$$p(D(v) = d)) = \binom{n}{d} p^d (1-p)^{n-d}$$

Expected mean degree is $np$!
Barabasi-Albert's model

- Web growth is clearly not random
  - Cf. Empirical degree distributions
  - If designed by people to represent semantics and associative relations, page to page links are not likely to be random nor independent.
  - Semantics: very few things make sense, very many make little sense
Matthew effect and preferential attachment

For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath.
—Matthew 25:29, King James Version.

“Rich get richer, poor get poorer.”

Sociological feedback effect: features of certain items or people are accentuated because presence of features influences future probability of acquiring features.
Mathematics: drawing from history of past draws...
Preferential attachment

- Graph growth model
- Probability of each new node $i$ connecting to node $j$ depends on degree($j$), or

"linked get more linked"

Scale-free networks with exponent $\gamma = 3$

\[
P(i, j) = \frac{D(j)}{\sum_k D(k)}
\]
Applications

- Knowledge modeling: semantic web
- Knowledge extraction:
  - Link prediction
  - Inference
  - See our recent fact-checking paper
- Community detection
- Multi-layer, multi-scale networks
- Resource allocation: see our paper about scientific funding
Life in the network


“We live life in the network... Each of these transactions leaves digital breadcrumbs which, when pulled together, offer increasingly comprehensive pictures of both individuals and groups, with the potential of transforming our understanding of our lives, organizations, and societies in a fashion that was barely conceivable just a few years ago.”
Computational Social Science

- Today, President Obama signed an Executive Order that directs Federal agencies to use behavioral science insights to better serve the American people. The Executive Order directs Federal agencies to identify programs in which applying behavioral science insights can yield substantial improvements; develop strategies for applying behavioral science insights to programs; and, where possible, for rigorously testing and evaluating the impact of these insights; recruit behavioral science experts to join the Federal Government; and strengthen agency relationships with the research community.