Web Retrieval

Chapter 3: Authority Ranking
Chapter 1: Outline

Authority Ranking: Motivation
Collaboration environments as social networks
Authority ranking in citation networks
Authority ranking in the Web and Web 2.0
- PageRank, HITS & Co
Advanced authority ranking models
- content oriented
- user oriented
Information dissemination in social networks
Achievements of research: a latent variable – not directly measurable … but what about indicators?!
Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'
given: set of reviewers $V = \{v_1, \ldots, v_k\}$, confidence grades $res(v_i, d)$ for submission $d$

collective result (restrictivity by thresholds $t_1$ and $t_2$, tuning by weights $w(v_i)$):

$$
\text{decision}(d) = \begin{cases} 
+1 & \text{if } \sum_i res_i(d) \cdot w(v_i) > t_1 \\
-1 & \text{if } \sum_i res_i(d) \cdot w(v_i) < t_2 \\
0 & \text{otherwise}
\end{cases}
$$

Special cases:
- "Unanimous Decision"
- "Voting"
- "Weighted Average" (e.g., weighted by some quality estimator)
General idea: accurate restrictive decisions

\[
\text{reduction} = \frac{|A0| + |B0| + |J0|}{|U|}
\]

\[
\text{error} = \frac{|AR| + |RA| + |JA| + |JR|}{|AA| + |RR| + |AR| + |RA| + |JA| + |JR|}
\]

\[
\text{junkred} = \frac{|J0|}{|J0| + |JA| + |JR|}
\]

\[
\text{loss} = \frac{|A0| + |R0|}{|AA| + |RR| + |AR| + |RA| + |A0| + |R0|}
\]

tradeoff!
Peer review: a simple model

**Given:** set of reviewers $V = \{v_1, ..., v_L\}$, binary decision (accept/reject)

**Wanted:** Approximations for *loss* and *error* for „unanimous decision“

$$X_i = \begin{cases} 
1 & \text{if } v_i \text{ assigns paper correctly} \\
0 & \text{otherwise}
\end{cases}$$

Probability of correct collaborative decision:

$$P(X_1 = 1, X_2 = 1|\bar{J}) = P(X_1 = 1|\bar{J}) \cdot P(X_2 = 1|\bar{J}) + \text{cov}(X_1, X_2|\bar{J})$$

$$P(X_1 = 1, ..., X_k = 1|\bar{J}) =$$

$$P(X_1 = 1|\bar{J}) \cdot \prod_{i=1}^{k-1} \frac{P(X_i = 1|\bar{J})P(X_{i+1} = 1|\bar{J}) + \text{cov}(X_i, X_{i+1}|\bar{J})}{P(X_i = 1|\bar{J})}$$

.. analogously we obtain $P(X_1 = 0, ..., X_k = 0|\bar{J})$

$$\text{junkred}(Meta) = 1 - P(X_1 = ... = X_L|J)$$

$$= 1 - [P(X_1 = 1, ..., X_L = 1|J) + P(X_1 = 0, ..., X_L = 0|J)]$$

$$\text{loss}(Meta) = 1 - P(X_1 = ... = X_L|\bar{J})$$

$$= 1 - [P(X_1 = 1, ..., X_L = 1|\bar{J}) + P(X_1 = 0, ..., X_L = 0|\bar{J})]$$

$$\text{error}(Meta) = \frac{P(X_1 = 0, ...X_L = 0|\bar{J}) \cdot P(\bar{J}) + P(X_1 = ... = X_L|J) \cdot P(J)}{1 - \text{junkred} \cdot P(J) - \text{loss} \cdot P(\bar{J})}$$
Peer review: a simple model (2)

Example:
- Probability $p < 0.5$ to misassign doc from pos/neg for all $k$ methods
- 50% of docs are Junk
- JunkDoc is assigned to pos/neg with prob. 0.5
- $c < p(p-1)$ (no perfect correlation)

\[
\begin{align*}
\text{junkred} & = 1 - \left( \frac{c + \frac{1}{4}}{\frac{1}{2}} \right)^{L-1} \\
\text{loss} & = 1 - \left( (1-p) \cdot \left( \frac{c + (1-p)^2}{1-p} \right)^{L-1} + p \cdot \left( \frac{c + p^2}{p} \right)^{L-1} \right) \\
\text{error} & = \frac{1}{2} \cdot \frac{p \cdot \left( \frac{c + p^2}{p} \right)^{L-1} + \left( \frac{c + \frac{1}{4}}{\frac{1}{2}} \right)^{L-1}}{(1-p) \cdot \left( \frac{c + (1-p)^2}{1-p} \right)^{L-1} + p \cdot \left( \frac{c + p^2}{p} \right)^{L-1} + \left( \frac{c + \frac{1}{4}}{\frac{1}{2}} \right)^{L-1}}
\end{align*}
\]
Peer Reviewing is NOT Perfect!

Progress changes rules and ways of thinking!

Famous rejected papers:
- B-trees
- The first paper about the Web (Berners-Lee et al)
- The first paper (Hendler et al) and the second paper (Fensel et al) about Semantic Web

See also:
S. Santini:
We Are Sorry to Inform You..
How much damage could be caused by a peer reviewer having a bad day? IEEE Computer, Dec 2005, pp. 126-128
Where to submit?

Events with

- peer review
- high visibility and impact in the community
- high restrictivity (low acceptance rate, 5-15 %)
- good organizers and reviewers

Sources of recommendations

- your supervisor and colleagues
- Microsoft Libra
- Australian Ranking of ICT Conferences
- ...

Microsoft Libra

CORE Research & Education
Differences in publication culture

Computer Science

- Peer-reviewed conferences
- Top conferences have 5-15% acceptance rate
- Specialized and small conferences (attendance of 500+)
- Often value conferences > journals

Pure Sciences (eg, Math, Physics)

- Pre-print at Arxiv.org
- Rigorous reviews for journals
- Huge flagship conference (ICM 98 attracted ~4000)

Social Sciences

- Often value journals > conferences
- Conferences are mostly for gathering or short abstract
- based screening
- Rigorous reviews for journals
Where not to submit?

Bogus conferences

- Known conferences and journals of (very) dubious reputation
- Nagib-Callaos-Conferences, Khalid-Soliman-Conferences
- Blacklists are impossible to keep up (threats by organizers, e.g. fakeconferences.org)

Indicators: curious OC and PC, fake venues, missing or questionable reviewing process, paper presentation not required.

see also:
SCIgen - An Automatic CS Paper Generator
Bogus conferences

To name just a few:

- IMCSE: International Multiconference in Computer Science and Computer Engineering
- WMSCI or SCI: World Multiconference on Systemics, Cybernetics and Informatics
- ICCCT: International Conference on Computing, Communications and Control Technologies
- PISTA: Conference on Politics and Information Systems: Technologies and Applications
- SSCCII: Symposium of Santa Caterina on Challenges in the Internet and Interdisciplinary Research
- CITSA: International Conference on Cybernetics and Information Technologies, Systems and Applications
- ISAS: International Conference on Information Systems Analysis and Synthesis
- CISCI: Conferencia Iberoamericana en Sistemas, Cibernética e Informática
- SIECI: Simposium Iberoamericano de Educación, Cibernética e Informática
- WCAC: World Congress in Applied Computing
- Any IPSI International Conference or journal
- Any GESTS international conference or journal
- KCPR: International Conference on Knowledge Communication and Peer Reviewing (!!!)
- International e-Conference on Computer Science
- ...
**Publications: success indicators**

**visibility**: we publish on high-quality conferences (A+, A, B)
  - check DBLP profile against CORE

**impact**: our work is frequently cited by others
  - check Google scholar, Citeseer, ..
  - evaluate with Publish or Perish

**reproducibility**: method details, tuning parameters, evaluation datasets, libraries, etc. are documented and available to public

**reliability**: our models are correct, evaluation is consistent
Conference Ranking: Possible approaches

Automatic ranking (e.g. Microsoft Libra)
- Citations per paper (impact)
- Number of submissions from trusted academic institutions
- Acceptance rate (if known)

Semi-Automatic ranking (e.g. Cori)
- Pre-selection of candidate conferences, based on automatically computed success indicators
- Manual inspection of candidates, manual assignment of final scores
<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Title</th>
<th>Conference/Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>Michaela Götz, Christoph Koch, Wim Martens</td>
<td>Efficient Algorithms for the Tree Homeomorphism Problem. DRPL 2007: 17-31</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Christoph Koch, Dan Olteanu</td>
<td>MayBMS: Managing Incomplete Information with Probabilistic World-Set Decompositions. ICDE 2007: 1479-1480</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Michael Schmidt, Stefanie Scherzinger, Christoph Koch</td>
<td>Combined Static and Dynamic Analysis for Effective Buffer Minimization in Streaming XQuery Evaluation. ICDE 2007: 236-245</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Christoph Koch, Dan Olteanu</td>
<td>World-Set Decompositions: Expressiveness and Efficient Algorithms. ICOT 2007: 194-208</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Christoph Koch</td>
<td>XPath Leashed. PLAN-X 2007: 0-1</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Christoph Koch, Dan Olteanu</td>
<td>From complete to incomplete information and back. SIGMOD Conference 2007: 713-724</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Mingsheng Hong, Alan J. Demers, Johannes Gehrke, Christoph Koch, Mirek Riedewald, Walker M. White</td>
<td>Massively multi-query join processing in publish/subscribe systems. SIGMOD Conference 2007: 761-772</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Lyublana Antova, Christoph Koch, Dan Olteanu</td>
<td>Query language support for incomplete information in the MayBMS system. VLDB 2007: 1422-1425</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Lyublana Antova, Thomas Jansen, Christoph Koch, Dan Olteanu</td>
<td>Fast and Simple Relational Processing of Uncertain Data CoRR abs/0707.1644: (2007)</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Michael Benedikt, Christoph Koch</td>
<td>Interpreting Tree-to-Tree Queries. ICALP (2) 2006: 552-564</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Christoph Koch</td>
<td>Processing queries on tree-structured data efficiently. PODS 2006: 213-224</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Georg Gottlob, Christoph Koch</td>
<td>A Formal Comparison of Visual Web Wrapper Generators. SOFSEM 2006: 30-48</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Christoph Koch, Dan Olteanu, Stefanie Scherzinger</td>
<td>Building a Native XML-DBMS as a Term Project in a Database Systems Course. XIME-P 2006</td>
<td></td>
</tr>
</tbody>
</table>
Simple authority measures in citation networks (1)

- Total number of papers
- Total number of citations
- Average number of citations per paper
- Average number of citations per author
- Average number of papers per author
Citation metrics: example

- **h-Index**: $N_p$ papers have at least $h$ citations each, and the other $(N_p-h)$ papers have no more than $h$ citations each.


- **g-Index**: the (unique) largest number such that the top $g$ articles received (together) at least $g^2$ citations (gives more weight to highly cited articles)

Social Networks: common metrics of interest
Small World Experiment (Six degrees of separation)

Stanley Milgram: 1967: Letters were handed out to people in Nebraska to be sent to a target (stock broker) in Boston
- People were instructed to pass on the letters to someone they knew on first-name basis
- The letters that reached the destination followed paths of avg length 5.2 (i.e. around 6)

Duncan Watts: 2001: Milgram's experiment recreated on the internet
- using an e-mail message as the "package" that needed to be delivered, with 48,000 senders and 19 targets (in 157 countries).
- the avg number of intermediaries was also around 6.

See also:
- The Kevin Bacon game
- The Erdös number
- The WeSTgame (DBLP based)
- etc.
Kevin Bacon Experiment

Craig Fass, Brian Turtle and Mike Ginelli: 1994: motivated by Bacon's most recent movie „The Air Up There” and his career discussion

Vertices: actors and actresses

Edge between u and v if they appeared in a movie together

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Average distance</th>
<th># of movies</th>
<th># of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rod Steiger</td>
<td>2.537527</td>
<td>112</td>
<td>2562</td>
</tr>
<tr>
<td>2</td>
<td>Donald Pleasence</td>
<td>2.542376</td>
<td>180</td>
<td>2874</td>
</tr>
<tr>
<td>3</td>
<td>Martin Sheen</td>
<td>2.551210</td>
<td>136</td>
<td>3501</td>
</tr>
<tr>
<td>4</td>
<td>Christopher Lee</td>
<td>2.552497</td>
<td>201</td>
<td>2993</td>
</tr>
<tr>
<td>5</td>
<td>Robert Mitchum</td>
<td>2.557181</td>
<td>136</td>
<td>2905</td>
</tr>
<tr>
<td>6</td>
<td>Charlton Heston</td>
<td>2.566284</td>
<td>104</td>
<td>2552</td>
</tr>
<tr>
<td>7</td>
<td>Eddie Albert</td>
<td>2.567036</td>
<td>112</td>
<td>3333</td>
</tr>
<tr>
<td>8</td>
<td>Robert Vaughn</td>
<td>2.570193</td>
<td>126</td>
<td>2761</td>
</tr>
<tr>
<td>9</td>
<td>Donald Sutherland</td>
<td>2.577880</td>
<td>107</td>
<td>2865</td>
</tr>
<tr>
<td>10</td>
<td>John Gielgud</td>
<td>2.578980</td>
<td>122</td>
<td>2942</td>
</tr>
<tr>
<td>11</td>
<td>Anthony Quinn</td>
<td>2.579750</td>
<td>146</td>
<td>2978</td>
</tr>
<tr>
<td>12</td>
<td>James Earl Jones</td>
<td>2.584440</td>
<td>112</td>
<td>3787</td>
</tr>
</tbody>
</table>

Kevin Bacon

No. of movies : 46
No. of actors : 1811
Average separation: 2.79

Is Kevin Bacon the most connected actor?

See also: http://oracleofbacon.org/

... 876 Kevin Bacon 2.786981 46 1811
Quantities of Interest

**Connected components:**
- how many, and how large?

**Network diameter:**
- maximum (worst-case) or average?
- exclude infinite distances? (disconnected components)
- the small-world phenomenon

**Clustering:**
- to what extent that links tend to cluster “locally”?
- what is the balance between local and long-distance connections?
- what roles do the two types of links play?

**Degree distribution:**
- what is the typical degree in the network?
- what is the overall distribution?
Graph theory: undirected graph notation

Graph $G=(V,E)$
- $V$ = set of vertices
- $E$ = set of edges

undirected graph
$E=\{(1,2),(1,3),(2,3),(3,4),(4,5)\}$
Graph theory: directed graph notation

Graph $G=(V,E)$
- $V$ = set of vertices
- $E$ = set of edges

Directed graph
$E=\{\langle 1,2 \rangle, \langle 2,1 \rangle, \langle 1,3 \rangle, \langle 3,2 \rangle, \langle 3,4 \rangle, \langle 4,5 \rangle \}$
degree $d(i)$ of node $i$
   - number of edges incident on node $i$

degree sequence
   - $[d(1), d(2), d(3), d(4), d(5)]$
   - $[2, 2, 3, 2, 1]$

degree distribution
   - $[(1, 1), (2, 3), (3, 1)]]$
in-degree $d_{in}(i)$ of node $i$
  – number of edges pointing to node $i$

out-degree $d_{out}(i)$ of node $i$
  – number of edges leaving node $i$

in-degree sequence
  – [1,2,1,1,1]

out-degree sequence
  – [2,1,2,1,0]
Degree distributions

Problem: find the probability distribution that best fits the observed data

$f_k = \frac{\text{fraction of nodes with degree } k}{\text{probability of a randomly selected node to have degree } k}$

Problem: find the probability distribution that best fits the observed data
Power-law distributions

The degree distributions of most real-life networks follow a power law

$$p(k) = Ck^{-\alpha}$$

Right-skewed/Heavy-tail distribution

- there is a non-negligible fraction of nodes that has very high degree (hubs)
- scale-free: no characteristic scale, average is not informative

- highly concentrated around the mean
- the probability of very high degree nodes is exponentially small
Power-law distribution gives a line in the log-log plot

\[ \log p(k) = -\alpha \log k + \log C \]

\[ p(k) = C k^{-\alpha} \]

\( \alpha \) : power-law exponent (typically \( 2 \leq \alpha \leq 3 \))
Power-law: Examples

(a) collaborations in mathematics
(b) citations
(c) World Wide Web
(d) Internet
(f) protein interactions

Taken from [Newman 2003]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>$-$</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>$-$</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,300</td>
<td>2.72</td>
<td>6.19</td>
<td>$-$</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>$-$</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>2.1</td>
<td>$-$</td>
<td>0.16</td>
<td>$-$</td>
<td>$-$</td>
<td>136</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>$-$</td>
<td>$-$</td>
<td>0.001</td>
<td>0.092</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>$-$</td>
<td>0.17</td>
<td>0.13</td>
<td>0.055</td>
<td>0.321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>$-$</td>
<td>0.005</td>
<td>0.001</td>
<td>$-$</td>
<td>0.029</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>3.2</td>
<td>$-$</td>
<td>$-$</td>
<td>$-$</td>
<td>$-$</td>
<td>$-$</td>
<td>$-$</td>
<td>45</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>$-$</td>
<td>-0.067</td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>16.1</td>
<td>21.2/1.7</td>
<td>$-$</td>
<td>0.11</td>
<td>74</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>3.9/3</td>
<td>$-$</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>$-$</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>2.7</td>
<td>$-$</td>
<td>0.44</td>
<td>0.44</td>
<td>119</td>
<td>157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>$-$</td>
<td>-0.189</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>$-$</td>
<td>0.10</td>
<td>0.080</td>
<td>0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>$-$</td>
<td>0.09</td>
<td>0.033</td>
<td>0.366</td>
<td>366</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>$-$</td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,977</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.003</td>
<td>-0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>-0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>$-$</td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>$-$</td>
<td>0.20</td>
<td>0.087</td>
<td>-0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>$-$</td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

**TABLE II** Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex–vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or “$-$” if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)

- power-law distributed degrees: $P[\text{degree}=k] \sim (1/k)^\alpha$
  with $\alpha \approx 2.1$ for indegrees and $\alpha \approx 2.7$ for outdegrees
Clustering (Transitivity) coefficient

Measures the density of triangles (local clusters) in the graph

Two different ways to measure it:

The ratio of the means

\[ C^{(1)} = \frac{\sum_i \text{triangles centered at node } i}{\sum_i \text{triples centered at node } i} \]
Example

\[ C^{(1)} = \frac{3}{1+1+6} = \frac{3}{8} \]
Clustering (Transitivity) coefficient

Clustering coefficient for node i

\[ C_i = \frac{\text{triangles centered at node } i}{\text{triples centered at node } i} \]

The mean of the ratios

\[ C^{(2)} = \frac{1}{n} C_i \]
The two clustering coefficients give different measures.

$C^{(2)} = \frac{1}{5}(1 + 1 + 1/6) = \frac{13}{30}$

$C^{(1)} = \frac{3}{8}$

The diagram shows the network structure with nodes 1, 2, 3, 4, and 5, and the coefficients are calculated based on this structure.
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>$n$</th>
<th>$m$</th>
<th>$z$</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>--</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>--</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,099</td>
<td>245,300</td>
<td>9.27</td>
<td>6.19</td>
<td>--</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>--</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td></td>
<td>2.1</td>
<td>--</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>--</td>
<td></td>
<td>136</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>--</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>--</td>
<td>0.005</td>
<td>0.001</td>
<td>0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>2,695,044</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>0.607</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>2,035,496</td>
<td>2,130,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.11</td>
<td>0.29</td>
<td>0.607</td>
<td>14, 34</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td></td>
<td>3.0/3.5</td>
<td>--</td>
<td>0.12</td>
<td>0.15</td>
<td>0.157</td>
</tr>
<tr>
<td>Roget's Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>0.148</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,962</td>
<td>17,000,000</td>
<td>70.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>119, 157</td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>--</td>
<td>0.10</td>
<td>0.080</td>
<td>0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>--</td>
<td>0.69</td>
<td>0.033</td>
<td>0.366</td>
<td>366</td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>--</td>
<td>0.033</td>
<td>0.012</td>
<td>0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,977</td>
<td>53,249</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.66</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>--</td>
<td>0.16</td>
<td>0.23</td>
<td>0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>--</td>
<td>0.20</td>
<td>0.087</td>
<td>0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>--</td>
<td>0.18</td>
<td>0.28</td>
<td>0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex–vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or $\alpha$−2 if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Clustering coefficient for random graphs

The probability of two of your neighbors also being neighbors is \( p \), independent of local structure

- clustering coefficient \( C = p \)
- when \( z \) is fixed \( C = z/n = \mathcal{O}(1/n) \)

<table>
<thead>
<tr>
<th>Network</th>
<th>( n )</th>
<th>( z )</th>
<th>( C ) measured</th>
<th>( C ) for random graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet [153]</td>
<td>6,374</td>
<td>3.8</td>
<td>0.24</td>
<td>0.00060</td>
</tr>
<tr>
<td>World Wide Web (sites) [2]</td>
<td>153,127</td>
<td>35.2</td>
<td>0.11</td>
<td>0.00023</td>
</tr>
<tr>
<td>power grid [192]</td>
<td>4,941</td>
<td>2.7</td>
<td>0.080</td>
<td>0.00054</td>
</tr>
<tr>
<td>biology collaborations [140]</td>
<td>1,520,251</td>
<td>15.5</td>
<td>0.081</td>
<td>0.000010</td>
</tr>
<tr>
<td>mathematics collaborations [141]</td>
<td>253,339</td>
<td>3.9</td>
<td>0.15</td>
<td>0.000015</td>
</tr>
<tr>
<td>film actor collaborations [149]</td>
<td>449,913</td>
<td>113.4</td>
<td>0.20</td>
<td>0.00025</td>
</tr>
<tr>
<td>company directors [149]</td>
<td>7,673</td>
<td>14.4</td>
<td>0.59</td>
<td>0.0019</td>
</tr>
<tr>
<td>word co-occurrence [90]</td>
<td>460,902</td>
<td>70.1</td>
<td>0.44</td>
<td>0.00015</td>
</tr>
<tr>
<td>neural network [192]</td>
<td>282</td>
<td>14.0</td>
<td>0.28</td>
<td>0.049</td>
</tr>
<tr>
<td>metabolic network [69]</td>
<td>315</td>
<td>28.3</td>
<td>0.59</td>
<td>0.090</td>
</tr>
<tr>
<td>food web [138]</td>
<td>134</td>
<td>8.7</td>
<td>0.22</td>
<td>0.065</td>
</tr>
</tbody>
</table>
Graphs: paths

Path from node $i$ to node $j$: a sequence of edges (directed or undirected from node $i$ to node $j$)

- path length: number of edges on the path
- nodes $i$ and $j$ are connected
- cycle: a path that starts and ends at the same node
Graphs: shortest paths

Shortest Path from node i to node j

- also known as **BFS path**, or geodesic path
Measuring the small world phenomenon

d_{ij} = \text{shortest path between } i \text{ and } j

Diameter:
\[ d = \max_{i,j} d_{ij} \]

Characteristic path length:
\[ \ell = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij} \]

Harmonic mean
\[ \ell^{-1} = \frac{1}{n(n-1)/2} \sum_{i>j} d_{ij}^{-1} \]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>\ell</th>
<th>\alpha</th>
<th>C^{(1)}</th>
<th>C^{(2)}</th>
<th>r</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449,913</td>
<td>25,516,482</td>
<td>113.43</td>
<td>3.45</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7,673</td>
<td>55,392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253,339</td>
<td>496,489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52,909</td>
<td>245,900</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1,520,251</td>
<td>11,803,064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47,000,000</td>
<td>80,000,000</td>
<td>3.16</td>
<td>–</td>
<td>2.1</td>
<td>–</td>
<td>–</td>
<td>8, 9</td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59,912</td>
<td>86,300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td>–</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16,881</td>
<td>57,029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2,810</td>
<td>–</td>
<td>3.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>265, 266</td>
<td></td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269,504</td>
<td>1,497,135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>203,549,046</td>
<td>2,130,000,000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>74</td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783,339</td>
<td>6,716,198</td>
<td>8.57</td>
<td>–</td>
<td>3.0/2.0</td>
<td>–</td>
<td>–</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1,022</td>
<td>5,103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460,902</td>
<td>17,000,000</td>
<td>70.13</td>
<td>–</td>
<td>2.7</td>
<td>0.44</td>
<td>–</td>
<td>119, 157</td>
<td></td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10,697</td>
<td>31,992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4,941</td>
<td>6,594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19,603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>–0.033</td>
<td>366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1,439</td>
<td>1,723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1,377</td>
<td>2,213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24,997</td>
<td>53,248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1,296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>biological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>3,686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.240</td>
<td>214</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2,115</td>
<td>2,240</td>
<td>2.12</td>
<td>6.89</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2,359</td>
<td>7.68</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II: Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex-vertex distance \ell; exponent \alpha of degree distribution if the distribution follows a power law (or “\ldots” if not; in/out-degree exponents are given for directed graphs); clustering coefficient C^{(1)} from Eq. (3); clustering coefficient C^{(2)} from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Degree correlations

Do high degree nodes tend to link to high degree nodes?

Newman

- compute the correlation coefficient of the degrees $x_i, y_i$ of the two endpoints of an edge $i$

\[
Kor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}
\]

\[
r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \cdot \frac{1}{\sqrt{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \frac{1}{\sqrt{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]
<table>
<thead>
<tr>
<th>network</th>
<th>type</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>$\ell$</th>
<th>$\alpha$</th>
<th>$C^{(1)}$</th>
<th>$C^{(2)}$</th>
<th>$r$</th>
<th>Ref(s).</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449913</td>
<td>25516482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.20</td>
<td>0.78</td>
<td>0.208</td>
<td>20, 416</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55392</td>
<td>14.44</td>
<td>4.60</td>
<td>–</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
<td>105, 323</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253339</td>
<td>496489</td>
<td>3.92</td>
<td>7.57</td>
<td>–</td>
<td>0.15</td>
<td>0.34</td>
<td>0.120</td>
<td>107, 182</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52909</td>
<td>245300</td>
<td>9.27</td>
<td>6.19</td>
<td>–</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
<td>311, 313</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1520251</td>
<td>11803064</td>
<td>15.53</td>
<td>4.92</td>
<td>–</td>
<td>0.088</td>
<td>0.60</td>
<td>0.127</td>
<td>311, 313</td>
</tr>
<tr>
<td>computer science</td>
<td>undirected</td>
<td>47000000</td>
<td>80000000</td>
<td>3.16</td>
<td>2.1</td>
<td>8.9</td>
<td>–</td>
<td>0.16</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59912</td>
<td>86300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
<td>321</td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16881</td>
<td>57029</td>
<td>3.38</td>
<td>5.22</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td>–</td>
<td>0.005</td>
<td>0.001</td>
<td>–0.029</td>
<td>45</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td></td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>265, 266</td>
</tr>
<tr>
<td>information</td>
<td>undirected</td>
<td>269504</td>
<td>1497135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>–0.067</td>
<td>14, 34</td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>203549046</td>
<td>213000000</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.07</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>WWW AltaVista</td>
<td>directed</td>
<td>783399</td>
<td>6716198</td>
<td>8.57</td>
<td>3.0/–</td>
<td></td>
<td>0.08</td>
<td>0.44</td>
<td>119, 157</td>
<td></td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1022</td>
<td>5103</td>
<td>4.99</td>
<td>4.87</td>
<td>–</td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
<td>244</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460902</td>
<td>17000000</td>
<td>70.13</td>
<td>2.7</td>
<td></td>
<td>0.04</td>
<td>0.44</td>
<td>119, 157</td>
<td></td>
</tr>
<tr>
<td>technological</td>
<td>undirected</td>
<td>10697</td>
<td>31992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>–0.189</td>
<td>86, 148</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4941</td>
<td>6594</td>
<td>2.67</td>
<td>18.99</td>
<td>–</td>
<td>0.10</td>
<td>0.080</td>
<td>–0.003</td>
<td>416</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19603</td>
<td>66.79</td>
<td>2.16</td>
<td>–</td>
<td>0.69</td>
<td>0.003</td>
<td>0.366</td>
<td></td>
</tr>
<tr>
<td>software packages</td>
<td>directed</td>
<td>1439</td>
<td>1723</td>
<td>1.20</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.070</td>
<td>0.082</td>
<td>–0.016</td>
<td>318</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1377</td>
<td>2213</td>
<td>1.61</td>
<td>1.51</td>
<td>–</td>
<td>0.033</td>
<td>0.012</td>
<td>–0.119</td>
<td>395</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>undirected</td>
<td>24997</td>
<td>53248</td>
<td>4.34</td>
<td>11.05</td>
<td>3.0</td>
<td>0.010</td>
<td>0.030</td>
<td>–0.154</td>
<td>155</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>–0.366</td>
<td>6, 354</td>
</tr>
<tr>
<td>biological</td>
<td>undirected</td>
<td>765</td>
<td>3686</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.090</td>
<td>0.67</td>
<td>–0.260</td>
<td>214</td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>2115</td>
<td>2240</td>
<td>2.12</td>
<td>6.80</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>–0.156</td>
<td>212</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td>–</td>
<td>0.16</td>
<td>0.23</td>
<td>–0.263</td>
<td>204</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.90</td>
<td>–</td>
<td>0.20</td>
<td>0.087</td>
<td>–0.326</td>
<td>272</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>307</td>
<td>2359</td>
<td>7.64</td>
<td>3.97</td>
<td>–</td>
<td>0.18</td>
<td>0.28</td>
<td>–0.226</td>
<td>416, 421</td>
</tr>
</tbody>
</table>

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices $n$; total number of edges $m$; mean degree $z$; mean vertex-vertex distance $\ell$; exponent $\alpha$ of degree distribution if the distribution follows a power law (or “$-$” if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (5); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient $r$, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.
Clique $K_n$
A graph that has all possible $n(n-1)/2$ edges
**Directed Graph**

**Strongly connected** graph: there exists a path from every i to every j

**Weakly connected** graph: If edges are made to be undirected the graph is connected
Web Structure: Connected Components

Study of Web Graph (Broder et al. 2000)

..strongly connected core tends to have small diameter

Source: A.Z. Broder et al., WWW 2000
Most nodes have only a small number of neighbors (degree), but there are some nodes with very high degree (\textit{power-law degree distribution})

- \textit{scale-free} networks

If a node \textit{x} is connected to \textit{y} and \textit{z}, then \textit{y} and \textit{z} are likely to be connected

- high \textit{clustering coefficient}

Most nodes are just a few edges away on average.

- \textit{small world} networks
A “Canonical” Natural Network has...

**Few** connected components:
- often only 1 or a small number, indep. of network size

**Small** diameter:
- often a constant independent of network size (like 6)
- or perhaps growing only logarithmically with network size or even shrink?
- typically exclude infinite distances

A **high** degree of clustering:
- considerably more so than for a random network
- in tension with small diameter

A **heavy-tailed** degree distribution:
- a small but reliable number of high-degree vertices
- often of *power law* form

**Is it possible that there is a unifying underlying generative process?**
Measuring and modelling network properties

Around 1999

- Watts and Strogatz, Dynamics and small-world phenomenon
- Faloutsos, On power-law relationships of the Internet Topology
- Kleinberg et al., The Web as a graph
- Barabasi and Albert, The emergence of scaling in real networks
All edges are *equally probable and appear independently*

NW size $N > 1$ and probability $p$: *distribution $G(N,p)$*

- each edge $(u,v)$ chosen to appear with probability $p$
- $N(N-1)/2$ trials of a biased coin flip

The usual *regime of interest* is when $p \sim 1/N$, $N$ is large

- e.g. $p = 1/2N$, $p = 1/N$, $p = 2/N$, $p = 10/N$, $p = \log(N)/N$, etc.
- in expectation, each vertex will have a “small” number of neighbors
- will then examine what happens when $N \rightarrow \infty$
- can thus study properties of *large networks with bounded degree*

*Degree distribution* of a typical $G$ drawn from $G(N,p)$:

- draw $G$ according to $G(N,p)$; look at a random vertex $u$ in $G$
- what is $\Pr[\deg(u) = k]$ for any fixed $k$?
- *Poisson distribution* with mean $l = p(N-1) \sim pN$
- Sharply concentrated; *not* heavy-tailed

Especially easy to *generate* NWs from $G(N,p)$
For any fixed $m \leq N(N-1)/2$, define distribution $G(N,m)$:

- choose *uniformly* at random from all graphs with *exactly* $m$ edges
- $G(N,m)$ is “like” $G(N,p)$ with $p = m/(N(N-1)/2) \sim 2m/N^2$
- this intuition can be made precise, and is correct
- if $m = cN$ then $p = 2c/(N-1) \sim 2c/N$
- mathematically trickier than $G(N,p)$
We have a large number $n$ of vertices

We start randomly adding edges one at a time

At what time $t$ will the network:

- have at least one "large" connected component?
- have a single connected component?
- have "small" diameter?
- have a "large" clique?

How gradually or suddenly do these properties appear?
Random graph models $G(N,p)$ and $G(N,m)$: Recap

Model $G(N,p)$:
- select each of the possible edges independently with prob. $p$
- expected total number of edges is $pN(N-1)/2$
- expected degree of a vertex is $p(N-1)$
- degree will obey a Poisson distribution ($not$ heavy-tailed)

Model $G(N,m)$:
- select $exactly$ $m$ of the $N(N-1)/2$ edges to appear
- all sets of $m$ edges equally likely
The $G(N,p)$ degree distribution follows a binomial

$$p(k) = B(n;k;p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Assuming $z=np$ is fixed, as $n\to\infty$ $B(n,k,p)$ is approximated by a Poisson distribution

$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

Highly concentrated around the mean, with a tail that drops exponentially
Monotone Network Properties

Threshold or tipping for (say) connectivity:
- fewer than \( m(N) \) edges \( \rightarrow \) graph almost certainly \( \textit{not} \) connected
- more than \( m(N) \) edges \( \rightarrow \) graph almost certainly \( \textit{is} \) connected
- made formal by examining limit as \( N \rightarrow \infty \)

Often interested in \textit{monotone} graph properties:
- let \( G \) have the property
- \textit{add edges} to \( G \) to obtain \( G' \)
- then \( G' \) \textit{must} have the property also

Examples:
- \( G \) is connected
- \( G \) has diameter \( \leq d \) (\textit{not} exactly \( d \))
- \( G \) has a clique of size \( \geq k \) (\textit{not} exactly \( k \))
- \( d, k \) may depend on NW size \( N \) (How?)

Difficult to study emergence of non-monotone properties as the number of edges is increased - what would it mean?
Consider Erdos-Renyi $G(N,m)$ model

- select $m$ edges at random to include in $G$

Let $P$ be some *monotone* property of graphs

- $P(G) = 1 \rightarrow G$ has the property
- $P(G) = 0 \rightarrow G$ does not have the property

Let $m(N)$ be some function of NW size $N$

- formalize idea that property $P$ appears "suddenly" at $m(N)$ edges

Say that $m(N)$ is a *threshold function for $P$* if:

- let $m'(N)$ be any function of $N$
- look at *ratio* $r(N) = m'(N)/m(N)$ as $N \rightarrow \infty$
- if $r(N) \rightarrow 0$: probability that $P(G) = 1$ in $G(N,m'(N))$: $\rightarrow 0$
- if $r(N) \rightarrow \infty$: probability that $P(G) = 1$ in $G(N,m'(N))$: $\rightarrow 1$

A *purely structural* definition of tipping

- tipping results from incremental increase in *connectivity*
Random Graphs: Which Properties Tip (2)?

Connected component of size > \(N/2\):
- threshold function is \(m(N) = N/2\) (or \(p \sim 1/N\))
- note: full connectivity impossible

Fully connected:
- threshold function is \(m(N) = (N/2)\log(N)\) (or \(p \sim \log(N)/N\))
- NW remains extremely sparse: only \(\sim \log(N)\) edges per vertex

Small diameter:
- threshold is \(m(N) \sim N^{3/2}\) for diameter 2 (or \(p \sim 2/\sqrt{N}\))
- fraction of possible edges still \(\sim 2/\sqrt{N} \rightarrow 0\)
- generate very small worlds
A model in which all connections are equally likely
  - each of the $N(N-1)/2$ edges chosen randomly & independently
As we add edges, a precise sequence of events unfolds:
  - graph acquires a giant component
  - graph becomes connected
  - graph acquires small diameter
Many properties appear very suddenly (tipping, thresholds)
All statements are mathematically precise
But is this how natural networks form?
If not, which aspects are unrealistic?
  - may all edges are not equally likely!
Erdos-Renyi: Clustering Coefficient

Generate a network $G$ according to $G(N,p)$
Examine a “typical” vertex $u$ in $G$
  - choose $u$ at random among all vertices in $G$
  - what do we expect $c(u)$ to be?
Answer: exactly $p$!
In $G(N,m)$, expect $c(u)$ to be $2m/N(N-1)$
Both cases: $c(u)$ entirely determined by overall density
Baseline for comparison with “more clustered” models
  - Erdos-Renyi has no bias towards clustered or local edges
Watt’s models: Caveman and Solaria

Erdos-Renyi:
- sharing a common neighbor makes two vertices **no more likely** to be directly connected than two very “distant” vertices
- every edge appears entirely **independently** of existing structure

But in many settings, the **opposite** is true:
- you tend to meet new friends through your old friends
- two web pages pointing to a third might share a topic
- two companies selling goods to a third are in related industries

Watts’ **Caveman** world:
- **overall** density of edges is low
- but two vertices with a common neighbor are likely connected

Watts’ **Solaria** world
- overall density of edges low; no special bias towards local edges
- “like” Erdos-Renyi
Making it (Somewhat) Precise: the $\alpha$-model

The $\alpha$-model has the following parameters or “knobs”:

- **$N$:** size of the network to be generated
- **$k$:** the *average degree* of a vertex in the network to be generated
- **$p$:** the *default probability* two vertices are connected
- **$\alpha$:** adjustable parameter dictating bias towards local connections

For any vertices $u$ and $v$:
- define $m(u,v)$ to be the number of common neighbors (so far)

Key quantity: the *propensity* $R(u,v)$ of $u$ to connect to $v$

- if $m(u,v) \geq k$, $R(u,v) = 1$ (share too many friends *not* to connect)
- if $m(u,v) = 0$, $R(u,v) = p$ (no mutual friends $\rightarrow$ no bias to connect)
- else, $R(u,v) = p + \left(\frac{m(u,v)}{k}\right)^\alpha (1-p)$

Generate NW incrementally
- using $R(u,v)$ as the edge probability; details omitted

Note: $\alpha = \infty$ is “like” Erdos-Renyi (but not exactly)
Watts-Strogatz Model

$C(p)$ : clustering coeff.
$L(p)$ : average path length

(Watts and Strogatz, Nature 393, 440 (1998))
For small $\alpha$, should generate large clustering coefficients
- we “programmed” the model to do so
- Watts claims that proving precise statements is hard...

But we do not want a new model for every little property
- Erdos-Renyi $\rightarrow$ small diameter
- $\alpha$-model $\rightarrow$ high clustering coefficient

In the interests of Occam’s Razor, we would like to find
- a single, simple model of network generation...
- ... that simultaneously captures many properties

Watt’s small world: small diameter and high clustering
Watts examines three real networks as case studies:
- the Kevin Bacon graph
- the Western states power grid
- the nervous system
For each of these networks, he:
- computes its size, diameter, and clustering coefficient
- compares diameter and clustering to best Erdos-Renyi approximate
- shows that the best $\alpha$-model approximation is better
- important to be “fair” to each model by finding best fit
Overall moral:
- if we care only about diameter and clustering, $\alpha$ is better than $p$
\[ \langle k \rangle \sim 6 \]
\[ P(k=500) \sim 10^{-99} \]
\[ N_{\text{WWW}} \sim 10^9 \]
\[ \Rightarrow N(k=500) \sim 10^{-90} \]

\[ P_{\text{out}}(k) \sim k^{-\gamma_{\text{out}}} \]
\[ \gamma_{\text{out}} = 2.45 \]

\[ P_{\text{in}}(k) \sim k^{-\gamma_{\text{in}}} \]
\[ \gamma_{\text{in}} = 2.1 \]

What does that mean?

**Poisson distribution**

\[ P(k) \]

\[ \langle k \rangle \]

\[ k \]

**Power-law distribution**

\[ P(k) \]

\[ k \]

**Exponential Network**

**Scale-free Network**
The number of nodes (N) is not fixed

- Networks continuously expand by additional new nodes
  - WWW: addition of new nodes
  - Citation: publication of new papers

The attachment is not uniform

- A node is linked with higher probability to a node that already has a large number of links
  - WWW: new documents link to well known sites (CNN, Yahoo, Google)
  - Citation: Well cited papers are more likely to be cited again
Scale-Free Networks: the Barabasi-Albert model

**Growth:** Starting with a small number \((m_0)\) of nodes, at every time step, we add a new node with \(m(<< m_0)\) edges that link the new node to \(m\) different nodes already present in the system.

**Preferential attachment:** The probability \(\Pi\) that a new node will be connected to node \(i\) depends on the degree \(k_i\) of node \(i\), such that

\[
\Pi(k_i) = \frac{k_i}{\sum_i k_i}
\]

After \(t\) time-steps the network has \(N = t + m_0\) nodes and \(m \cdot t\) edges. Vertices \(j\) with high degree are likely to get more links ("Rich get richer")

The network evolves into a stationary scale-free state with

\[
P(k) \sim 2m^2k^{-\gamma} \quad \gamma_{BA} = 3
\]
Preferential attachment explains

- heavy-tailed degree distributions
- small diameter (~\(\log(N)\), via “hubs”)

Small average path length: \(\ell \sim \frac{\ln(N)}{\ln\ln(N)}\)

Node degree correlations: The dynamical process that creates a scale free network builds up nontrivial correlations between the degrees of connected nodes.

Clustering coefficient is \(~5\) times larger than that of a random graph. Will not generate high clustering coefficient as with Watts-Strogatz model. Clustering coefficient decreases with network size, following approximately a power law \(C \sim N^{-0.75}\) i.e. no bias towards local connectivity, but towards hubs
Authority ranking in social networks: common models
The Random Surfer Model

Idea: incoming links are endorsements & increase page authority, authority is higher if links come from high-authority pages

$$PR(q) = \sum_{p \in IN(q)} PR(p) \times t(p, q)$$

Authority (page q) = stationary prob. of visiting q

random walk: uniformly random choice of links + random jumps
A stochastic matrix is an $n \times n$ matrix $M$ with row sum $\sum_{j=1..n} M_{ij} = 1$ for each row $i$.

Random surfer follows a stochastic matrix.

**Theorem:**
For every stochastic matrix $M$ all Eigenvalues $\lambda$ have the property $|\lambda| \leq 1$ and there is an Eigenvector $x$ with Eigenvalue 1 s.t. $x \geq 0$ and $\|x\|_1 = 1$.

Suggests power iteration $x^{(i+1)} = M^T x^{(i)}$.

But: real Web graph has sinks, may be periodic, is not strongly connected.
“Rank Sink” Problem
- In general, many Web pages have no inlinks/outlinks
- It results in dangling edges in the graph

E.g.
- no parent $\rightarrow$ rank 0
  - $M^T$ converges to a matrix whose last column is all zero
- no children $\rightarrow$ no solution
  - $M^T$ converges to zero matrix
Google’s PageRank

Idea: incoming links are endorsements & increase page authority, authority is higher if links come from high-authority pages

\[ PR(q) = \varepsilon \cdot j(q) + (1 - \varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p, q) \]

Authority (page q) = stationary prob. of visiting q

random walk: uniformly random choice of links + random jumps
A **stochastic process** is a family of random variables \( \{X(t) \mid t \in T\} \).

T is called parameter space, and the domain \( M \) of \( X(t) \) is called state space. T and M can be discrete or continuous.

A stochastic process is called **Markov process** if for every choice of \( t_1, \ldots, t_{n+1} \) from the parameter space and every choice of \( x_1, \ldots, x_{n+1} \) from the state space the following holds:

\[
P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 \land X(t_2) = x_2 \land \ldots \land X(t_n) = x_n \right] \\
= P \left[ X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n \right]
\]

A Markov process with discrete state space is called **Markov chain**. A canonical choice of the state space are the natural numbers. Notation for Markov chains with discrete parameter space: \( X_n \) rather than \( X(t_n) \) with \( n = 0, 1, 2, \ldots \).
The Markov chain $X_n$ with discrete parameter space is

**homogeneous** if the transition probabilities $p_{ij} := P[X_{n+1} = j \mid X_n = i]$ are independent of $n$

**irreducible** if every state is reachable from every other state with positive probability:

$$\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0 \quad \text{for all } i, j$$

**aperiodic** if every state $i$ has period 1, where the period of $i$ is the greatest common divisor of all (recurrence) values $n$ for which

$$P[X_n = i \land X_k \neq i \text{ for } k = 1, \ldots, n-1 \mid X_0 = i] > 0$$
The Markov chain $X_n$ with discrete parameter space is

**positive recurrent** if for every state $i$ the recurrence probability is 1 and the mean recurrence time is finite:

$$\sum_{n=1}^{\infty} P[X_n = i \land X_k \neq i \text{ for } k = 1, ..., n-1 \mid X_0 = i] = 1$$

$$\sum_{n=1}^{\infty} n P[X_n = i \land X_k \neq i \text{ for } k = 1, ..., n-1 \mid X_0 = i] < \infty$$

**ergodic** if it is homogeneous, irreducible, aperiodic, and positive recurrent.
For the \textbf{n-step transition probabilities} 
\[
p_{ij}^{(n)} := P[ X_n = j | X_0 = i ] \quad \text{the following holds:} \]
\[
p_{ij}^{(n)} = \sum_{k} p_{ik}^{(n-1)} p_{kj} \quad \text{with} \quad p_{ij}^{(1)} := p_{ik}
\]
\[
= \sum_{k} p_{ik}^{(n-l)} p_{kj}^{(l)} \quad \text{for} \ 1 \leq l \leq n-1
\]
in matrix notation: \[ P^{(n)} = P^{n} \]

For the \textbf{state probabilities after n steps} 
\[
\pi_j^{(n)} := P[ X_n = j ] \quad \text{the following holds:} \]
\[
\pi_j^{(n)} = \sum_{i} \pi_i^{(0)} p_{ij}^{(n)} \quad \text{with initial state probabilities} \quad \pi_i^{(0)}
\]
in matrix notation: \[ \Pi^{(n)} = \Pi^{(0)} P^{(n)} \quad \text{(Chapman-Kolmogorov equation)} \]
Every homogeneous, irreducible, aperiodic Markov chain with a finite number of states is positive recurrent and ergodic.

For every ergodic Markov chain there exist stationary state probabilities

\[ \pi_j := \lim_{n \to \infty} \pi_j^{(n)} \]

These are independent of \( \Pi^{(0)} \) and are the solutions of the following system of linear equations:

\[ \pi_j = \sum_i \pi_i \, p_{ij} \quad \text{for all } j \]
\[ \sum_j \pi_j = 1 \]

(balance equations)

In matrix notation:

\[ \Pi = \Pi \, P \]
\[ \Pi \vec{1} = 1 \]
Model a random walk of a Web surfer as follows:
• follow outgoing hyperlinks with uniform probabilities
• perform „random jump“ with probability $\varepsilon$

→ ergodic Markov chain

The PageRank of a URL is the stationary visiting probability of URL in the above Markov chain.

Further generalizations have been studied
(e.g. random walk with back button etc.)

Drawback of Page rank method:
Page rank is query-independent and orthogonal to relevance
Example: Page Rank Computation

$\varepsilon = 0.2$

$P = \begin{pmatrix}
0.0 & 0.5 & 0.5 \\
0.1 & 0.0 & 0.9 \\
0.9 & 0.1 & 0.0 \\
\end{pmatrix}$

\[\begin{align*}
\Pi^{(0)} & \approx \begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \end{pmatrix}^T \\
\Pi^{(1)} & \approx \begin{pmatrix} 0.333 \\ 0.200 \\ 0.466 \end{pmatrix}^T \\
\Pi^{(2)} & \approx \begin{pmatrix} 0.439 \\ 0.212 \\ 0.346 \end{pmatrix}^T \\
\Pi^{(3)} & \approx \begin{pmatrix} 0.332 \\ 0.253 \\ 0.401 \end{pmatrix}^T \\
\Pi^{(4)} & \approx \begin{pmatrix} 0.385 \\ 0.176 \\ 0.527 \end{pmatrix}^T \\
\Pi^{(5)} & \approx \begin{pmatrix} 0.491 \\ 0.244 \\ 0.350 \end{pmatrix}^T
\end{align*}\]

$\pi_1 = 0.1 \pi_2 + 0.9 \pi_3$

$\pi_2 = 0.5 \pi_1 + 0.1 \pi_3$

$\pi_3 = 0.5 \pi_1 + 0.9 \pi_2$

$\pi_1 + \pi_2 + \pi_3 = 1$

$\Rightarrow \pi_1 \approx 0.3776, \pi_2 \approx 0.2282, \pi_3 \approx 0.3942$
HITS Algorithm: Hyperlink-Induced Topic Search (1)

Idea:
Determine • good content sources: Authorities (high indegree)
• good link sources: Hubs (high outdegree)

Find • better authorities that have good hubs as predecessors
• better hubs that have good authorities as successors

For Web graph $G=(V,E)$ define for nodes $p, q \in V$

**authority score**

$$x_q = \sum_{(p,q) \in E} y_p$$

**hub score**

$$y_p = \sum_{(p,q) \in E} x_q$$
Authority and hub scores in matrix notation:

\[
\begin{align*}
\vec{x} & = A^T \vec{y} \\
\vec{y} & = A \vec{x}
\end{align*}
\]

Iteration with adjacency matrix A:

\[
\begin{align*}
\vec{x} & := A^T \vec{y} := A^T A \vec{x} \\
\vec{y} & := A \vec{x} := A A^T \vec{y}
\end{align*}
\]

Using Linear Algebra, we can prove:

- \(x\) and \(y\) converge
- \(x\) and \(y\) are \textit{Eigenvectors} of \(A^T A\) and \(A A^T\), resp.

Intuitive interpretation:

\[
\begin{align*}
M^{(\text{auth})} & := A^T A \\
M^{(\text{hub})} & := AA^T
\end{align*}
\]

is the cocitation matrix: \(M^{(\text{auth})}_{ij}\) is the number of nodes that point to both \(i\) and \(j\)

is the bibliographic-coupling matrix: \(M^{(\text{hub})}_{ij}\) is the number of nodes to which both \(i\) and \(j\) point
Implementation of the HITS Algorithm

1) Determine sufficient number (e.g. 50-200) of „root pages“ via relevance ranking (e.g. using tf*idf ranking)
2) Add all successors of root pages
3) For each root page add up to d predecessors
4) Compute iteratively
   the authority and hub scores of this „base set“ (of typically 1000-5000 pages)
   with initialization $x_q := y_p := 1 / |\text{base set}|$
   and normalization after each iteration
   $\rightarrow$ converges to principal Eigenvector (Eigenvector with largest Eigenvalue (in the case of multiplicity 1))
5) Return pages in descending order of authority scores (e.g. the 10 largest elements of vector $x$)

Drawback of HITS algorithm:
relevance ranking within root set is not considered
Example: HITS Algorithm

root set

base set
(Bharat and Henzinger, 1998)

• HITS problems
  1) The document can contain many *identical* links to the same document in another host
  2) Links are generated automatically (e.g. messages posted on newsgroups)

• Solutions
  1) Assign weight to *identical* multiple edges, which are inversely proportional to their multiplicity
  2) Prune irrelevant nodes or regulating the influence of a node with a relevance weight
Improved HITS Algorithm

Potential weakness of the HITS algorithm:
• irritating links (automatically generated links, spam, etc.)
• topic drift (e.g. from „Jaguar car“ to „car“ in general)

Improvement:
• Introduce edge weights:
  0 for links within the same host,
  1/k with k links from k URLs of the same host to 1 URL (xweight)
  1/m with m links from 1 URL to m URLs on the same host (yweight)
• Consider relevance weights w.r.t. query topic (e.g. tf*idf)

→ Iterative computation of

authority score  \[ x_q = \sum_{(p,q) \in E} y_p \cdot \text{topic score}(p) \cdot x\text{weight}(p,q) \]

hub score  \[ y_p = \sum_{(p,q) \in E} x_q \cdot \text{topic score}(q) \cdot y\text{weight}(p,q) \]
Typed graphs: data items, users, friends, groups, postings, ratings, queries, clicks, ... with weighted edges $\rightarrow$ spectral analysis of various graphs
Simplified and cast into relational schema:

- **Users** \((UId, \text{Nickname, } \ldots)\)
- **Docs** \((DId, \text{Author, PostingDate, } \ldots)\)
- **Tags** \((TId, \text{String})\)
- **Friendship** \((UId1, UId2, FScore)\)
- **Content** \((DId, TId, Score)\)
- **Rating** \((UId, DId, RScore)\)
- **Tagging** \((UId, TId, DId, TScore)\)
- **TagSim** \((TId1, TId2, TSim)\)

- Actually several kinds of “Friends”: same group, fan & star, true friend, etc.
- Tags could be typed or explicitly organized in hierarchies
- Numeric values for FScore, RScore, TScore, TSim may be explicitly specified or derived from co-occurrence statistics
Tagging relation is central:

- ternary relationship between users, tags, docs
- could be represented as hypergraph or tensor
- or (lossfully) decomposed into 3 binary projections (graphs):

**UsersTags** \((\text{UId}, \text{TId}, \text{UTscore})\)

\[ x.\text{UTscore} := \sum_d \{s \mid (x.\text{UId}, x.\text{TId}, d, s) \in \text{Ratings}\} \]

**TagsDocs** \((\text{TId}, \text{Did}, \text{TDscore})\)

\[ x.\text{TDscore} := \sum_u \{s \mid (u, x.\text{TId}, x.\text{Did}, s) \in \text{Ratings}\} \]

**DocsUsers** \((\text{Did}, \text{UId}, \text{DUscore})\)

\[ x.\text{DUscore} := \sum_t \{s \mid (x.\text{UId}, t, x.\text{Did}, s) \in \text{Ratings}\} \]
Apply link analysis (PR etc.) to appropriately defined matrices

- **SocialPageRank** [Bao et al.: WWW 2007]:
  Let $\mathbf{M}_{UT}$, $\mathbf{M}_{TD}$, $\mathbf{M}_{DU}$ be the matrices corresponding to relations UsersTags, TagsDocs, DocsUsers
  Compute iteratively:
  \[
  \tilde{\mathbf{r}}_U = \mathbf{M}'_{DU} \times \tilde{\mathbf{r}}_D \\
  \tilde{\mathbf{r}}_D = \mathbf{M}'_{TD} \times \tilde{\mathbf{r}}_T \\
  \tilde{\mathbf{r}}_T = \mathbf{M}'_{UT} \times \tilde{\mathbf{r}}_U
  \]

- **FolkRank** [Hothen et al.: ESWC 2006]:
  Define **graph G as union of graphs** UsersTags, TagsDocs, DocsUsers
  Assume each user has personal preference vector $\mathbf{p}$
  Compute iteratively:
  \[
  \tilde{\mathbf{r}}_D = \alpha \tilde{\mathbf{r}}_D + \beta \mathbf{M}_G \times \tilde{\mathbf{r}}_D + \gamma \mathbf{p}
  \]
Social PageRank: General Idea

High quality web pages are usually popularly annotated and popular web pages, up-to-date web users and hot social annotations have the following relations:

popular web pages are bookmarked by many up-to-date users and annotated by hot annotations;

up-to-date users like to bookmark popular pages and use hot annotations;

hot annotations are used to annotate popular web pages and used by up-to-date users.
Social PageRank: Inputs

Assume that there are $N_A$ annotations, $N_P$ web pages and $N_U$ web users.

- $M_{PU}$ is the $N_P \times N_U$ association matrix between pages and users
- $M_{AP}$ is the $N_A \times N_P$ association matrix between annotations and pages
- $M_{UA}$ is the $N_U \times N_A$ association matrix between users and annotations
- Element $M_{PU}(p_i, u_j)$ is assigned with the count of annotations used by user $u_j$ to annotate page $p_i$.
- Elements of $M_{AP}$ and $M_{UA}$ are initialized similarly.
- $P_0$ be the vector containing randomly initialized SocialPageRank scores.
Algorithm 2: SocialPageRank (SPR)

Step 1

Input:
Association matrices $M_{PU}$, $M_{AP}$, and $M_{UA}$ and the random initial SocialPageRank score $P_0$

Step 2

Do:

\[
U_i = M_{PU}^r \cdot P_i \quad (5.1)
\]

\[
A_i = M_{UA}^r \cdot U_i \quad (5.2)
\]

\[
P_i^{\prime} = M_{AP}^r \cdot A_i \quad (5.3)
\]

\[
A_i^{\prime} = M_{AP} \cdot P_i^{\prime} \quad (5.4)
\]

\[
U_i^{\prime} = M_{UA} \cdot A_i^{\prime} \quad (5.5)
\]

\[
P_{i+1} = M_{PU} \cdot U_i^{\prime} \quad (5.6)
\]

Until $P_i$ converges.

Step 3

Output:

$P^*$: the converged SocialPageRank score.
FolkRank: Formalism

Entities of a Folksonomy

- Users $U$
- Tags $T$
- Resources $R$
- Assignments $Y$

Representation

- Tripartite undirected hypergraph
- $G=(V,E)$, $V=U \cup T \cup R$, $E=\{ (u,t,r) \mid (u,t,r) \in Y \}$
FolkRank: Graph Preparation

Flatten the Folksonomy graph.
Apply PageRank.
A resource tagged with important tags by important users becomes important. Symmetrically for tags and users.
FolkRank: Adaptation problems

Important! The flat Folksonomy graph is undirected.
Part of the weight that goes through an edge at time $t$, will flow back at time $t+1$.
Results are similar to an edge degree ranking. They are identical for $d=1$. 
A topic is defined through preference vector $\vec{p}$.

A topic can be defined through tags, resources or users.

Let $\vec{w}_0$ be the Adapted PageRank vector for $d=1$.

Let $\vec{w}_1$ be the Adapted PageRank vector for $d<1$ and a specified preference vector.

The FolkRank vector is $\vec{w} := \vec{w}_1 - \vec{w}_0$

$$\vec{w}_{t+1} = dA\vec{w}_t + (1-d)\vec{p}$$
**Singular Value Decomposition (SVD)**

**Theorem:**
Each real-valued $m \times n$ matrix $A$ with rank $r$ can be decomposed into the form $A = U \times \Delta \times V^T$ with an $m \times r$ matrix $U$ with orthonormal column vectors, an $r \times r$ diagonal matrix $\Delta$, and an $n \times r$ matrix $V$ with orthonormal column vectors. This decomposition is called *singular value decomposition* and is unique when the elements of $\Delta$ are sorted.

**Theorem:**
In the singular value decomposition $A = U \times \Delta \times V^T$ of matrix $A$ the matrices $U$, $\Delta$, and $V$ can be derived as follows:
- $\Delta$ consists of the singular values of $A$, i.e. the positive roots of the Eigenvalues of $A^T \times A$,
- the columns of $U$ are the Eigenvectors of $A \times A^T$,
- the columns of $V$ are the Eigenvectors of $A^T \times A$. 
SVD for Regression

**Theorem:**
Let $A$ be an $m \times n$ matrix with rank $r$, and let $A_k = U_k \times \Delta_k \times V_k^T$, where the $k \times k$ diagonal matrix $\Delta_k$ contains the $k$ largest singular values of $A$ and the $m \times k$ matrix $U_k$ and the $n \times k$ matrix $V_k$ contain the corresponding Eigenvectors from the SVD of $A$.

Among all $m \times n$ matrices $C$ with rank at most $k$, $A_k$ is the matrix that minimizes the Frobenius norm

$$
\|A - C\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} - C_{ij})^2
$$
Reminder: SVD

\[ A \approx U \Sigma V^T = \sum_i \sigma_i u_i \circ v_i \]

- Best rank-k approximation in L2
\[ A \approx U \Sigma V^T = \sum_i \sigma_i u_i \circ v_i \]

- Best rank-k approximation in L2
Specially Structured Tensors

**Tucker Tensor**

\[ x = \mathcal{G} \times_1 U \times_2 V \times_3 W \]
\[ = \sum_{r} \sum_{s} \sum_{t} g_{rst} u_r \circ v_s \circ w_t \]
\[ \equiv [\mathcal{G} ; U, V, W] \]

**Kruskal Tensor**

\[ x = \sum_{r} \lambda_r u_r \circ v_r \circ w_r \]
\[ \equiv [\lambda ; U, V, W] \]
Tucker Decomposition - intuition

author x keyword x conference
A: author x author-group
B: keyword x keyword-group
C: conf. x conf-group
G: how groups relate to each other
Goal: extension to $\geq 3$ modes

$$\chi \approx [\lambda; A, B, C] = \sum_{r} \lambda_r \ a_r \circ b_r \circ c_r$$
HITS and TOPHITS for Web Retrieval

Sparse adjacency matrix and its SVD:

\[ x_{ij} = \begin{cases} 
1 & \text{if page } i \text{ links to page } j \\
0 & \text{otherwise} 
\end{cases} \]

\[ X \approx \sum_{\tau} \sigma_{\tau} \cdot h_{\tau} \cdot a_{\tau} \]

Kleinberg, JACM, 1999
HITS Authorities on Sample Data

We started our crawl from http://www-neos.mcs.anl.gov/neos, and crawled 4700 pages, resulting in 560 cross-linked hosts.
Three-Dimensional View of the Web

Observe that this tensor is very sparse!

$$x_{ijk} = \begin{cases} 1 & \text{if page } i \rightarrow \text{ page } j \\ 0 & \text{otherwise} \end{cases}$$

Kolda, Bader, Kenny, ICDM05
Three-Dimensional View of the Web

Observe that this tensor is very sparse!

\[ x_{ijk} = \begin{cases} 
1 & \text{if page } i \rightarrow \text{ page } j \\
0 & \text{otherwise} 
\end{cases} \]

Kolda, Bader, Kenny, ICDM05
Endangered Species
Animals today are being threatened by a variety of environmental pressures. For example, the jaguar is losing prime habitat in the world. Zoos are trying to raise awareness of their plight.

Rain Forest Zoo
We have a new exhibit opening next month highlighting the endangered species of the Americas, including the jaguar.

Jaguar FAQ
Jaguars are an endangered species that live in the tropical rain forests of Central and South America. They live about 11 years in the wild and up to 22 years at a zoo.

Online Atlas
View maps of animal habitats from around the world, including those of endangered animals in North, South, and Central America.

\[ x_{ijk} = \begin{cases} 
1 & \text{if page } i \rightarrow \text{ page } j \\
0 & \text{otherwise} 
\end{cases} \]

Observe that this tensor is very sparse!

Kolda, Bader, Kenny, ICDM05
**Main Idea:** Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

\[
\chi \approx \sum_{r=1}^{R} \lambda_r h_r \odot a_r
\]

- Hub scores for 1st topic
- Authority scores for 1st topic
- Hub scores for 2nd topic
- Authority scores for 2nd topic

...
Main Idea: Extend the idea behind the HITS model to incorporate term (i.e., topical) information.

\[
\chi \approx \sum_{r=1}^{R} \lambda_r \ h_r \circ a_r \circ t_r
\]
TOPYHITS uses 3D analysis to find the dominant groupings of web pages and terms. For a given term $k$, the similarity score $x_{ijk}$ between pages $i$ and $j$ is defined as:

$$x_{ijk} = \begin{cases} \frac{1}{\log(w_k) + 1} & \text{if } i \rightarrow j \text{ with term } k \\ 0 & \text{otherwise} \end{cases}$$

where $w_k$ is the number of unique links using term $k$. The Tensor PARAFAC model can be used to decompose the matrix of term scores into factors representing terms, authorities, and hubs.
Lots of **linked** information

**Easy-to-use** browsers

**Freedom** of browsing:
- No pre-determined presentation and aggregation of results
- No dictation of a browsing path, you’re the navigator
Which links to follow?
Where to go next?
What information is there?
What are the „good“ links ...
... leading to „good“ RDF resources?
Solution: Relevance Ranking for Linked Open Data

Who is loved/hated most?
Who loves/hates most?

Score Resource

<table>
<thead>
<tr>
<th>Score</th>
<th>Predicate</th>
<th>Score Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>hates</td>
<td>0.71 Bob</td>
</tr>
<tr>
<td>1.00</td>
<td>loves</td>
<td>0.70 Chris</td>
</tr>
<tr>
<td>0.001</td>
<td>hates</td>
<td>0.70 Don</td>
</tr>
<tr>
<td>0.001</td>
<td>hates</td>
<td>0.70 Alex</td>
</tr>
<tr>
<td>0.10</td>
<td></td>
<td>0.10 Elly</td>
</tr>
</tbody>
</table>

TripleRank

Who is loved/hated most?
Who loves/hates most?
Breadth-first search

Limits:

– Overall statements
– Exploration Depth
– Links per property
– Links per statements
Linked Data as Tensor

Transformation to 3-D-Tensor

Alex, Elly, Don, Chris, Bob

Preprocessing

A, B, C, D, E

Loves, Hates
Pre-Processing

Wheighting

Remove properties occurring in >40% of all triples

\[ T(x, y, z) = \begin{cases} 
1 + \log_{\text{links}(z)}, & \text{if } x \text{ points to } y \text{ using property } z \\
0, & \text{else}
\end{cases} \]
Reminder: Hubs and Authorities

- Good authorities are linked by good hubs:
  \[
  auth(r) = \sum_{s:(r_s,p,r) \in S} hub(r_s)
  \]

- Good hubs link to good authorities:
  \[
  hub(r) = \sum_{s:(r,p,r_o) \in S} auth(r_o)
  \]
PARAFAC Analysis

Higher-order equivalent to matrix decompositions like SVD

Differs from applying SVD on each matrix constituting the tensor (problem of many sparse matrices)

Analysis considers the tensor as a whole
Evaluation

16 test persons
Faceted browsing scenario
What are the most interesting, most related, most useful resources (objects)?
10 queries
Comparison against Baseline

Baseline method:
- Uses pre-processed tensor
- Selects one slice corresponding to predicate (facet) in question
- Ranks on in-degree
- Corresponds to HITS

Ex.: look for authorities wrt predicate d
1387 answers

Overall inter-rater agreement: 0.7

(0 ≤ agreement ≤ 1)

nearly doubled recall without significant loss of precision

<table>
<thead>
<tr>
<th></th>
<th>Avg Number of Results</th>
<th>Avg Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Positive</td>
</tr>
<tr>
<td>BaselineII–TripleRank</td>
<td>1.075</td>
<td>0.273</td>
</tr>
<tr>
<td>TripleRank–BaselineII</td>
<td>4.731</td>
<td>1.913</td>
</tr>
<tr>
<td>BaselineII∩TripleRank</td>
<td>2.863</td>
<td>1.338</td>
</tr>
<tr>
<td>BaselineII</td>
<td>3.948</td>
<td>1.626</td>
</tr>
<tr>
<td>TripleRank</td>
<td>7.594</td>
<td>3.251</td>
</tr>
</tbody>
</table>
Advantages of TripleRank

Identification of similar properties

Indicators for the contribution of properties to authority scores

Both enable to compute “best” authorities across the tensor, i.e. across properties
**Example: Twitter users**

<table>
<thead>
<tr>
<th>@timberners_lee</th>
<th>@parishilton</th>
</tr>
</thead>
<tbody>
<tr>
<td>following: 59</td>
<td>following: 272</td>
</tr>
<tr>
<td>followers: 20,692</td>
<td>followers: 1,709,116</td>
</tr>
</tbody>
</table>

**RT @janl: Apple:**

"Preparing Your Web Content for iPad:
2. Use W3C standard web technologies."  
9:46 AM Mar 21st via TweetDeck

**I Love my new I-Pad.**

So much fun!

Technology rocks!

about 2 hours ago via UberTwitter
<table>
<thead>
<tr>
<th>POS</th>
<th>WeFollow recommends</th>
<th>h-set recommends</th>
<th>#followers: total in h-set</th>
<th>H-set recommends</th>
<th>#followers: total in H-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tommyh</td>
<td>PaulMiller</td>
<td>2215</td>
<td>timberners_lee</td>
<td>20.694</td>
</tr>
<tr>
<td>2</td>
<td>jahendler</td>
<td>jahendler</td>
<td>909</td>
<td>timoreilly</td>
<td>1.428.425</td>
</tr>
<tr>
<td>3</td>
<td>ivan_herman</td>
<td>tommyh</td>
<td>738</td>
<td>jahendler</td>
<td>910</td>
</tr>
<tr>
<td>4</td>
<td>PaulMiller</td>
<td>ivan_herman</td>
<td>680</td>
<td>LeeFeigenbaum</td>
<td>273</td>
</tr>
<tr>
<td>5</td>
<td>opencalais</td>
<td>opencalais</td>
<td>1797</td>
<td>danbri</td>
<td>1.781</td>
</tr>
<tr>
<td>6</td>
<td>danja</td>
<td>danja</td>
<td>1313</td>
<td>kidhehnen</td>
<td>1.806</td>
</tr>
<tr>
<td>7</td>
<td>CaptSolo</td>
<td>juansequeda</td>
<td>988</td>
<td>ivan_herman</td>
<td>680</td>
</tr>
<tr>
<td>8</td>
<td>juansequeda</td>
<td>CaptSolo</td>
<td>1224</td>
<td>PaulMiller</td>
<td>2.216</td>
</tr>
<tr>
<td>9</td>
<td>sclopit</td>
<td>gothwin</td>
<td>685</td>
<td>tommyh</td>
<td>737</td>
</tr>
<tr>
<td>10</td>
<td>gothwin</td>
<td>robocrunch</td>
<td>3679</td>
<td>novas pivack</td>
<td>7.896</td>
</tr>
<tr>
<td>11</td>
<td>robocrunch</td>
<td>alexiskold</td>
<td>4321</td>
<td>johnbreslin</td>
<td>1.933</td>
</tr>
<tr>
<td>12</td>
<td>kristathomas</td>
<td>kristathomas</td>
<td>1499</td>
<td>w3c</td>
<td>10.672</td>
</tr>
<tr>
<td>13</td>
<td>kendall</td>
<td>kendall</td>
<td>1694</td>
<td>mimasnews</td>
<td>198</td>
</tr>
<tr>
<td>14</td>
<td>bobdc</td>
<td>andraz</td>
<td>2065</td>
<td>iand</td>
<td>1.094</td>
</tr>
<tr>
<td>15</td>
<td>phclouin</td>
<td>sclopit</td>
<td>513</td>
<td>rww</td>
<td>1.045.511</td>
</tr>
<tr>
<td>16</td>
<td>brown2020</td>
<td>cjmcconnors</td>
<td>466</td>
<td>terraces</td>
<td>632</td>
</tr>
<tr>
<td>17</td>
<td>alexiskold</td>
<td>g kob</td>
<td>399</td>
<td>mhausenblas</td>
<td>449</td>
</tr>
<tr>
<td>18</td>
<td>andraz</td>
<td>dorait</td>
<td>2541</td>
<td>opencalais</td>
<td>1.799</td>
</tr>
<tr>
<td>19</td>
<td>cjmcconnors</td>
<td>phclouin</td>
<td>266</td>
<td>ltdodds</td>
<td>621</td>
</tr>
<tr>
<td>20</td>
<td>ontoligent</td>
<td>openamplify</td>
<td>1387</td>
<td>semanticnews</td>
<td>843</td>
</tr>
</tbody>
</table>
## Example: Twitter Users (3)

<table>
<thead>
<tr>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.147</td>
<td>semantics</td>
<td>0.238</td>
<td>timberners_lee</td>
<td>0.111</td>
<td>programming</td>
<td>0.183</td>
<td>cjmconnors</td>
</tr>
<tr>
<td>0.125</td>
<td>business</td>
<td>0.143</td>
<td>PEPublishing</td>
<td>0.053</td>
<td>analytics</td>
<td>0.178</td>
<td>DublinCore</td>
</tr>
<tr>
<td>0.106</td>
<td>lod</td>
<td>0.142</td>
<td>jahendler</td>
<td>0.047</td>
<td>semantic</td>
<td>0.108</td>
<td>spirinet</td>
</tr>
<tr>
<td>0.091</td>
<td>semweb</td>
<td>0.138</td>
<td>timoreilly</td>
<td>0.046</td>
<td>microdata</td>
<td>0.098</td>
<td>AskAaronLee</td>
</tr>
<tr>
<td>0.054</td>
<td>semanticweb</td>
<td>0.097</td>
<td>semanticnews</td>
<td>0.040</td>
<td>startups</td>
<td>0.096</td>
<td>GeoffWigz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.266</td>
<td>java</td>
<td>0.419</td>
<td>SCMagazine</td>
<td>0.163</td>
<td>security</td>
<td>0.285</td>
<td>BLSocSci</td>
</tr>
<tr>
<td>0.251</td>
<td>php</td>
<td>0.143</td>
<td>HTML5watcher</td>
<td>0.130</td>
<td>web20</td>
<td>0.278</td>
<td>socialwendy</td>
</tr>
<tr>
<td>0.220</td>
<td>http</td>
<td>0.128</td>
<td>opencalais</td>
<td>0.083</td>
<td>privacy</td>
<td>0.162</td>
<td>pedantic_web</td>
</tr>
<tr>
<td>0.214</td>
<td>joomla</td>
<td>0.102</td>
<td>hadoop</td>
<td>0.067</td>
<td>apps</td>
<td>0.154</td>
<td>drthinkmore</td>
</tr>
<tr>
<td>0.199</td>
<td>javascript</td>
<td>0.097</td>
<td>LS1storange</td>
<td>0.054</td>
<td>china</td>
<td>0.136</td>
<td>linuxhoundhost</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
<th>Score</th>
<th>Hashtag</th>
<th>Score</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.049</td>
<td>music</td>
<td>0.265</td>
<td>Beyond15</td>
<td>0.212</td>
<td>foaf</td>
<td>0.484</td>
<td>jwolfnbaa</td>
</tr>
<tr>
<td>0.044</td>
<td>video</td>
<td>0.185</td>
<td>junglejar</td>
<td>0.172</td>
<td>socialmedia</td>
<td>0.165</td>
<td>rdfQuery</td>
</tr>
<tr>
<td>0.032</td>
<td>semweb</td>
<td>0.172</td>
<td>CSS3</td>
<td>0.118</td>
<td>facebook</td>
<td>0.130</td>
<td>CreativeCustoms</td>
</tr>
<tr>
<td>0.023</td>
<td>iphone</td>
<td>0.134</td>
<td>davidstack</td>
<td>0.109</td>
<td>rdfa</td>
<td>0.078</td>
<td>websciencetrust</td>
</tr>
<tr>
<td>0.021</td>
<td>innovation</td>
<td>0.125</td>
<td>emtacl</td>
<td>0.054</td>
<td>webscience</td>
<td>0.064</td>
<td>virtualrooms</td>
</tr>
</tbody>
</table>