References


Sizes of LARGE graphs
Already outdated but gives some idea

- Google (2008): indexed over $10^{12}$ URLs
- Facebook (2012): $721 \times 10^6$ individuals and $137 \times 10^9$ links
- Phone companies (2013) process a few trillion calls a year
- The human brain (2011) has around $100 \times 10^9$ and $100 \times 10^{12}$ neuronal connections
Graph representations

- Edge list
- Adjacency list
Graph mining tasks

- Random walk steps (e.g. to extract a massive graph nearby the seed) $O(1)$
- Connected components $O(n)$
- PageRank — determine importance of nodes $O(n)$
- Effective diameter (mean shortest path or longest shortest path to connect 90% of possible node pairs) $O(n)$
- Extremal eigenvalues of graph Laplacian (the first eigenvector helps to split the graph) $O(n \log n)$
- Triangle counting (detect interesting groups) $O(n^{3/2})$
- All-pairs problems $O(n^3)$ time and $O(n^2)$ memory
Classification of large graphs

Graphs are sparse: the number of edges $O(n)$

- Small graphs ($< 10^4$ vertices) — all algorithms are feasible
- Large Small graphs ($10^4$—$10^6$ vertices) — $O(n^2)$ in time is fine but $O(n^2)$ in memory may be prohibitive
- Small Large graphs ($10^6$—$10^8$ vertices) — $O(n^2)$ is prohibitive without specialized computing resources
- Large graphs ($10^8$—$10^{10}$ vertices)
- LARGE graphs ($> 10^{10}$ vertices)
Sources for graph data

- https://snap.stanford.edu/data/index.html
- http://law.di.unimi.it/datasets.php
- https://sparse.tamu.edu
We are interested at backlinks of the page
**The PageRank**

**Definition.** Let $E(u)$ be some vector over the Web pages that corresponds to a source of rank. Then, the PageRank of a set of Web pages is an assignment, $R'$, to the Web pages which satisfies

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

Such that $c$ is maximized and $\| R' \|_1 = 1$.

Eigenvalue problem:

$$R' = c(A + E1^\top)R'$$

Loop that acts as a rank sink
Computing PageRank

\[ R_0 = \text{an initial guess for the rank vector} \]

\[
\textbf{while } \delta > \epsilon \rightarrow \\
R_{j+1} = AR_j \\
d = \|R_j\|_1 - \|R_{j+1}\|_1 \\
R_{j+1} = R_{j+1} + dE \\
\delta = \|R_{j+1} - R_j\|_1 \\
\textbf{end} \]
Convergence of PageRank Computation

- Red squares represent 322 Million Links.
- Green diamonds represent 161 Million Links.

The graph shows the total difference from the previous iteration against the number of iterations.
Choosing source vector $E$

- $E$ is uniform. Problem: sites like highly interlinked mailing lists archives receive overly high ranking.

- $E$ is personalized: $E(u) > 0$ only for one page, e.g. user’s personal web page.

- A compromise: $E$ consists of all root-level pages of all web servers.
<table>
<thead>
<tr>
<th>Web Page</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download Netscape Software</td>
<td>11589.00</td>
</tr>
<tr>
<td><a href="http://www.w3.org/">http://www.w3.org/</a></td>
<td>10717.70</td>
</tr>
<tr>
<td>Welcome to Netscape</td>
<td>8673.51</td>
</tr>
<tr>
<td>Point: It’s What You’re Searching For</td>
<td>7930.92</td>
</tr>
<tr>
<td>Web-Counter Home Page</td>
<td>7254.97</td>
</tr>
<tr>
<td>The Blue Ribbon Campaign for Online Free Speech</td>
<td>7010.39</td>
</tr>
<tr>
<td>CERN Welcome</td>
<td>6562.49</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>6561.80</td>
</tr>
<tr>
<td>Welcome to Netscape</td>
<td>6203.47</td>
</tr>
<tr>
<td>Wusage 4.1: A Usage Statistics System For Web Servers</td>
<td>5963.27</td>
</tr>
<tr>
<td>The World Wide Web Consortium (W3C)</td>
<td>5672.21</td>
</tr>
<tr>
<td>Lycos, Inc. Home Page</td>
<td>4683.31</td>
</tr>
<tr>
<td>Starting Point</td>
<td>4501.98</td>
</tr>
<tr>
<td>Welcome to Magellan!</td>
<td>3866.82</td>
</tr>
<tr>
<td>Oracle Corporation</td>
<td>3587.63</td>
</tr>
</tbody>
</table>

Table 1: Top 15 Page Ranks: July 1996