Optimal ranking in networks with community structure

Huafeng Xie

Dept of Physics, Brookhaven National Laboratory
New Media Lab, The Grad. Center, CUNY

NetSci 2006, Bloomington IN USA
World Wide Web

- Nodes (Vertices): Web pages in the WWW
- Links: Hyperlinks on the web pages
- Large size: $N \sim 10^{10}$
- Heterogeneous: community structure

Objective

Authority: Community Bias

PageRank

- Understand the interplay between the community structure and the average Google rank inside the community.
- As a search engine, how to reduce the undesired effects of community structure.

Authority and PageRank visualization of “java” query result

Definitions

- $E_{cw}, E_{wc}$
- $N_c, N_w$
- $<K_{in}>_c, <K_{out}>_c$
- $<K_{in}>_w, <K_{out}>_w$
- $1 << N_c << N_w$

$G_c$: the average Google rank value of the community member nodes

$G_w$: the average Google rank value of the outside world

Effects of community on $G_c$ is only determined by $E_{cw}, E_{wc}$ and a parameter in PageRank $\alpha$.

What’s the optimal value for $\alpha$
Google PageRank Algorithm

- **PageRank**: simulates random walks on the web
- **Rank Value** of a node $i$ is proportional to the number of random walkers on this node at stationary state

\[ G(i) = \alpha + \sum_{j \rightarrow i} (1 - \alpha) \frac{G(j)}{K_{out}(j)}. \]

Effects of Community Structure

Mean-field assumption:
the average Google ranks and out-degrees of community nodes sending links to the outside world are equal to the overall average values inside the community $G_c$. Assume the same for node sending links from the outside world to the community.

$$J_{cw} = (1 - \alpha) G_c E_{cw} / \langle K_{out} \rangle_c + \alpha G_c N_c$$
$$J_{wc} = (1 - \alpha) G_w E_{wc} / \langle K_{out} \rangle_w + \alpha G_w N_c$$
$$J_{cw} = J_{wc}$$

$$\frac{G_c}{G_w} = \frac{(1 - \alpha) \frac{E_{wc}}{\langle K_{out} \rangle_w N_c} + \alpha}{(1 - \alpha) \frac{E_{cw}}{\langle K_{out} \rangle_c N_c}} + \alpha$$

$G_w \approx 1$
Main Equation

In random networks with the same degree sequences,

Expected number of links from the outside world to the community:

$$E_{wc}^{(r)} = <K_{out}>_w N_c$$

Expected number of links from the community to outside world:

$$E_{cw}^{(r)} = <K_{out}>_c N_c N_w / (N_c + N_w) = <K_{out}>_c N_c$$

$$G_C = \frac{(1 - \alpha) \frac{E_{wc}}{E_{wc}^{(r)}} + \alpha}{1 - \alpha} \frac{E_{cw}^{(r)}}{E_{cw}} + \alpha$$

$$R_{wc} = \frac{E_{wc}}{E_{wc}^{(r)}}$$

$$R_{cw} = \frac{E_{cw}}{E_{cw}^{(r)}}$$

Provided that our mean-field assumption is valid

$$G_C = \frac{(1 - \alpha) R_{wc} + \alpha}{(1 - \alpha) R_{cw} + \alpha}$$
Empirical Study

UCLA: 31621 nodes, 353370 edges
LIU (Long Island University): 15471 nodes, 90111 edges

<table>
<thead>
<tr>
<th>Community</th>
<th>$N_c$</th>
<th>$E_{cc}$</th>
<th>$E_{cc}^{(r)}$</th>
<th>$E_{wc}$</th>
<th>$E_{cw}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCLA Library</td>
<td>2028</td>
<td>23062</td>
<td>1699</td>
<td>755</td>
<td>2141</td>
</tr>
<tr>
<td>UCLA School of Management</td>
<td>1340</td>
<td>15983</td>
<td>739</td>
<td>175</td>
<td>169</td>
</tr>
<tr>
<td>UCLA Academic Tech. Services</td>
<td>1907</td>
<td>26597</td>
<td>2248</td>
<td>139</td>
<td>3113</td>
</tr>
<tr>
<td>UCLA Social Science Division</td>
<td>626</td>
<td>3986</td>
<td>50</td>
<td>258</td>
<td>142</td>
</tr>
<tr>
<td>UCLA Humanity Division</td>
<td>864</td>
<td>4846</td>
<td>79</td>
<td>397</td>
<td>445</td>
</tr>
<tr>
<td>LIU CWP Campus</td>
<td>2756</td>
<td>18376</td>
<td>4105</td>
<td>336</td>
<td>1393</td>
</tr>
</tbody>
</table>
Empirical Study

\[ G_c = \frac{(1-\alpha)R_{wc} + \alpha}{(1-\alpha)R_{cw} + \alpha} \]

\[ R_{wc} = \frac{E_{wc}}{E_{wc}^{(r)}} \]

\[ R_{cw} = \frac{E_{cw}}{E_{cw}^{(r)}} \]

<table>
<thead>
<tr>
<th>Community</th>
<th>( R_{wc} )</th>
<th>( R_{cw} )</th>
<th>( R_{wc}^* )</th>
<th>( R_{cw}^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCLA Library</td>
<td>0.04</td>
<td>0.09</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>UCLA School of Management</td>
<td>0.01</td>
<td>0.01</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>UCLA Academic Tech. Services</td>
<td>0.007</td>
<td>0.1</td>
<td>0.003</td>
<td>0.07</td>
</tr>
<tr>
<td>UCLA Social Science Division</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>UCLA Humanity Division</td>
<td>0.04</td>
<td>0.08</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>LIU CWP Campus</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Optimal $\alpha$ for PageRank

$G_c = \frac{(1-\alpha)R_{wc}^* + \alpha}{(1-\alpha)R_{cw}^* + \alpha}$

- $\alpha$ should be as large as possible to avoid manipulations.
- $\alpha$ should be small enough to take into account network topology.

Indeed Google Uses a good value of $\alpha$, 0.15.
Acknowledgment

Sergei Maslov
- Department of Physics, Brookhaven National Laboratory

Koon-kiu Yan
- Department of Physics and Astronomy, Stony Brook University
- Department of Physics, Brookhaven National Laboratory

- Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy.
- Supports from The New Media Lab at the Grad Center of CUNY.