A Large-Scale Study of Link Spam Detection by Graph Algorithms

Hiroo Saito
Masashi Toyoda
Masaru Kitsuregawa
Kazuyuki Aihara

University of Tokyo. JST, ERATO
University of Tokyo
University of Tokyo
University of Tokyo. JST, ERATO

AIRWEB’07, May 8, 2007
Outline

• Propose a link farm detection method using graph algorithms
• Distribution of detected link farms in the Web graph structure

1. SCC decomposition
   Around the largest SCC (CORE), large SCCs are link farms

2. Maximal clique enumeration
   Link farms in CORE can be found as maximal cliques

3. Minimum cut
   Link farms are expanded by min-cut. How many links for cutting them out?
Dataset

• Japanese Web archive crawled in May 2004
  – 96 million pages, 4.5 billion links
  – 60% pages in Japanese, 40% in other languages

• Site graph
  – Top of site: URL linked from 3 or more servers
  – A site is a set of URLs below the top URL
  – 5.9 million sites, 283 million links

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number</th>
<th>Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.com</td>
<td>2,711,588</td>
<td>46.2</td>
</tr>
<tr>
<td>.jp</td>
<td>1,353,842</td>
<td>23.1</td>
</tr>
<tr>
<td>.net</td>
<td>436,645</td>
<td>7.4</td>
</tr>
<tr>
<td>.org</td>
<td>211,983</td>
<td>3.6</td>
</tr>
<tr>
<td>.de</td>
<td>169,279</td>
<td>2.9</td>
</tr>
<tr>
<td>.info</td>
<td>144,483</td>
<td>2.5</td>
</tr>
<tr>
<td>.nl,.kr,.us,etc.</td>
<td>841,610</td>
<td>14.3</td>
</tr>
</tbody>
</table>

Degree

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>max. of indegree</td>
<td>61,006</td>
</tr>
<tr>
<td>avr. of indegree</td>
<td>48</td>
</tr>
<tr>
<td>max. of outdegree</td>
<td>70,294</td>
</tr>
<tr>
<td>avr. of outdegree</td>
<td>48</td>
</tr>
</tbody>
</table>
SCC decomposition

- Size distribution follows the power-law \((1 \leq n \leq 100)\) with a long and thick tail
- Large SCCs are spams \((100 < n)\)
  - 552 SCCs, 0.57M sites
  - 550 sample sites

Sampling results

<table>
<thead>
<tr>
<th></th>
<th>spam</th>
<th>suspicious</th>
<th>non-spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sites</td>
<td>527</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>ratio (%)</td>
<td>95.8</td>
<td>4.2</td>
<td>0</td>
</tr>
</tbody>
</table>
Distribution of SCCs in the bow tie

- Bow-tie structure [Broder et al. 2000]
  - CORE 30%
  - OUT 60%
  - TENDRILS 2%
  - OTHERS 7%

- Distribution of large SCCs
  - 450 / 552 (81%) SCCs in OUT
  - 385 / 450 (85%) SCCs directly connected to CORE

- CORE has many spam sites connecting to them

SCCs whose size are larger than 1,000
Maximal clique enumeration

- Use maximal cliques for extracting spam from CORE
  - Link farms tend to include cliques
- Maximal clique enumeration [Makino, Uno 2004]
  - Ignore nodes with high degree (80<d)
    - Because of $O(\text{max. degree}^4)$
  - Large cliques are link farms (40 < n)
    - 26,931 maximal cliques, 8,346 sites (many duplicates)
    - 165 sample sites

Sampling result

<table>
<thead>
<tr>
<th></th>
<th>spam</th>
<th>suspicious</th>
<th>non-spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>#sites</td>
<td>157</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>ratio (%)</td>
<td>95.2</td>
<td>4.8</td>
<td>0</td>
</tr>
</tbody>
</table>
Minimum cut

- How many spam sites around large SCCs and cliques?
- How many links for cutting off spam sites?

Apply max-flow / min-cut on the directed site graph

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>459</td>
<td>94.3</td>
</tr>
<tr>
<td>suspicious</td>
<td>27</td>
<td>5.5</td>
</tr>
<tr>
<td>non-spam</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Sampling result

CORE

Virtual source
210 white sites

Virtual sink
57,000 sites

57,000 sites

450,000 sites

8,000 sites

Min-cut: 18,000

SCCs

Cliques
Conclusions and future work

• An automatic link farm detection method
  – Based on graph algorithms
    • Seed extraction: SCC and maximal clique
    • Seed expansion: Max-flow / min-cut
  – High precision (95% ~ 99%)

• Distribution of link farms in the Web graph structure
  – Large SCCs around CORE, Maximal cliques in CORE
  – Only 18,000 links for cutting off 0.5M spam sites

Future work
  – Improving recall (small SCCs, large cliques in CORE)
  – Experiments on other datasets