Making PageRank Algorithm Robust to Collusion

Hui Zhang\textsuperscript{1}, Ashish Goel\textsuperscript{2}, Ramesh Govindan\textsuperscript{1}, Kahn Mason\textsuperscript{2}, Benjamin Van Roy\textsuperscript{2}

\textsuperscript{1}University of Southern California
\textsuperscript{2}Stanford University
Outline

- Research motivation.
- PageRank algorithm: a brief introduction.
- Study of PageRank’s robustness to collusion.
- Adaptive-resetting: make PageRank robust to collusion.
- Conclusion & future works.
Research motivation

- Build reputation in large-scale systems
  - P2P file sharing systems
  - Blogging communities
  - Networked gaming, ..., etc.

- Collusion-proofness is an essential criterion in evaluating a rating scheme.
PageRank [Brin1998]

- A rating scheme to rank hypertext documents on the WWW.

- An iterative algorithm to calculate the importance of a web page based on the importance of its parent pages.

- Can be applied to other systems than WWW.
PageRank: random walk model

- As time goes on, the expected percentage of steps the walker is at each node $v$ converges to the PageRank weight $PR(v)$.

4/21/2005
PageRank: is it collusion-proof?

- Can a node easily boost its rank by manipulating its outgoing links with others’?
Amp(G): a metric on group collusion

- In the system of node group G, for a subgroup G', the amplification factor Amp(G') = \[ \frac{W_G(G')}{W_{in}(G')} \]

\[ W_G(G') = \sum_{i: \ i \in G'} PR(i) \]

\[ W_{in}(G') = \sum_{(i,j): \ i \notin G', \ j \in G', \ \exists i \rightarrow j} \frac{PR(i)}{out(i)} + \frac{|G'|}{|G|} \]
Theorem on Amp

- In the original PageRank system,

\[ \forall G' \subseteq G, Amp(G') < \frac{1}{\varepsilon} \]

where \( \varepsilon \) is the resetting probability.
Two experimental topologies

• $W$, a Web link topology
  ▪ Contains the link structure of upwards of 80 million URLs.
  ▪ Source: the Stanford WebBase.

• $B$, a weblog blogrolling topology
  ▪ Contains the blogrolling structure of upwards of 72,000 blogs.
  ▪ Source: www.blogstreet.com - the XML -RPC weblog service.
Experiment 1: Collusion200

• Model a small number of web pages *simultaneously* colluding.

• Methodology:
  
  • 100 colluding groups;
  
  • Each colluding group has the circle topology consisting of two nodes with adjacent ranks;
  
  • Arbitrarily chose nodes originally ranked around 1000\(^{th}\), 2000\(^{th}\), …, 100000\(^{th}\).
  
  • \(\varepsilon = 0.15\).
Experiment result of **Collusion200 (I)**

Figure 1: $W$ - Amplification factors of the 100 colluding groups in *Collusion200*. 

4/21/2005
Experiment result of **Collusion200 (III)**

Figure 2: $W$ – new PR rank after **Collusion200**.
There is a long flat portion...

Figure 3: The PR weight distribution of 4 topologies.

4/21/2005
Next step: how to detect collusions?

- Identifying colluding groups is unlikely to be computationally tractable.
  - The densest k-subgraph problem [Feige et al. 1997].
  - The classical CLIQUE problem.
  - The problem of finding hiding large cliques in random graphs [Juels 1998].
An observation on collusion behaviors

- To increase their PR weight, i.e., the stationary weight in the random walk, the colluding nodes will stall the random walk.

- When the resetting probability $\epsilon$ increases, the colluding nodes must suffer a significant drop in PR weight.

- Therefore, we expect the PR weight of colluding nodes to be highly correlated with $1/\epsilon$ (the average walk length), while that of non-colluding nodes is relatively insensitive to the change in $\epsilon$. 
An intuitive example

node

referential link
An intuitive example

.node

referential link

A colluding group
An intuitive example

- A colluding node $x$: $\text{PR}(x) = \frac{1}{K + (N - K)\varepsilon} \approx \frac{1}{N\varepsilon}$, and $\text{co-co}($PR$(x)$, $1/\varepsilon$) $\approx 1$. ($\text{co-co}$: correlation coefficient)

- A non-colluding node $y$: $\text{PR}(x) = \frac{\varepsilon}{K + (N - K)\varepsilon} \approx \frac{1}{N}$, and $\text{co-co}($PR$(y)$, $1/\varepsilon$) $\approx 0.$

4/21/2005

$N$: the system size; $K$: the colluding group size; $K << N.$
Co-co distribution in real-world graphs

Figure 4: the co-co PDF distribution in $W$ and $B$: the $[0, 0.1]$ range actually corresponds to $[-1, 0.1]$ range.
Adaptive-resetting scheme

• Part I – collusion detection:
  ▪ Given the topology, calculate the PR vector under different $\varepsilon$ values.
    - $\{\varepsilon\} = \{0.0375, 0.05, 0.075, 0.15, 0.3, 0.45, 0.6\}$, $\varepsilon_{\text{default}} = 0.15$.
  ▪ Calculate the correlation coefficient between the curve of each node $x$'s PR weight and the curve of $1/\varepsilon$. Label it as $co-co(x)$.

• Part II – $\varepsilon$ personalization:
  ▪ Calculate each node $x$'s out-link personalized-\(\varepsilon = F(\varepsilon_{\text{default}}, co-co(x))\).
    - Exponential function $F_{\text{Exp}} = \varepsilon_{\text{default}}^{(1.0-co-co(x))}$
    - Linear function $F_{\text{Linear}} = \varepsilon_{\text{default}} + (0.5-\varepsilon_{\text{default}})*co-co(x)$
  ▪ The final PR weight vector is calculated with these personalized resetting values.
Experiment result of **Collusion200 (IV)**

Figure 5: \( W \) - Amplification factors of the 100 colluding groups in *Collusion200*.  

4/21/2005
Experiment result of *Collusion200 (V)*

Figure 6: $W$ – new PR weight after *Collusion200*. 

4/21/2005
Experiment result of **Collusion200 (VI)**

Figure 7: $W$ – new PR rank after **Collusion200**.
**Experiment 2: Collusion22**

- Model various colluding subgraphs.
- Methodology:
  - 3 colluding groups:
    - **G1: 10-node ring**
    - **G2: 10-node star topology**
    - **G3: 2-node ring**

4/21/2005
Experiment result of **Collusion22 (I)**

**Figure 8: Amplification factors of the 3 colluding groups in Collusion22.**

4/21/2005
Experiment result of **Collusion22 (II)**

Figure 9: $W$ – new PR weight after **Collusion22**.
### New top-25 URL list in $W$

<table>
<thead>
<tr>
<th>Rank</th>
<th>Old list</th>
<th>New list</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td><a href="http://www.tucows.com/">http://www.tucows.com/</a></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td><a href="http://www.gendex.com/ged2html">http://www.gendex.com/ged2html</a></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td><a href="http://www.acme.com/">http://www.acme.com/</a></td>
</tr>
<tr>
<td>13</td>
<td><a href="http://www.acme.com/software/thttpd">http://www.acme.com/software/thttpd</a></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td><a href="http://upload.tucows.com/contactus.html">http://upload.tucows.com/contactus.html</a></td>
</tr>
<tr>
<td>15</td>
<td><a href="http://upload.tucows.com/contactus.html">http://upload.tucows.com/contactus.html</a></td>
<td><a href="http://www.w3.org/">http://www.w3.org/</a></td>
</tr>
<tr>
<td>18</td>
<td><a href="http://www.w3.org/">http://www.w3.org/</a></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td><a href="http://www.adobe.com/">http://www.adobe.com/</a></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td><a href="http://www.tucows.com/privacy.html">http://www.tucows.com/privacy.html</a></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td><a href="http://www.gendex.com/ged2html">http://www.gendex.com/ged2html</a></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The old and new top-25 list of $W$

4/21/2005
Conclusion & future works

• A collusion-proof rating scheme based on PageRank algorithm.

• Future works:
  ▪ Optimum analysis of the adaptive-resetting scheme.

  ▪ Study of Web link structure evolution under PageRank within the framework of game theory.