

# **Making PageRank Algorithm Robust to Collusion**

**Hui Zhang<sup>1</sup>, Ashish Goel<sup>2</sup>, Ramesh Govindan<sup>1</sup>, Kahn Mason<sup>2</sup>,  
Benjamin Van Roy<sup>2</sup>**

**<sup>1</sup>University of Southern California**

**<sup>2</sup>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.

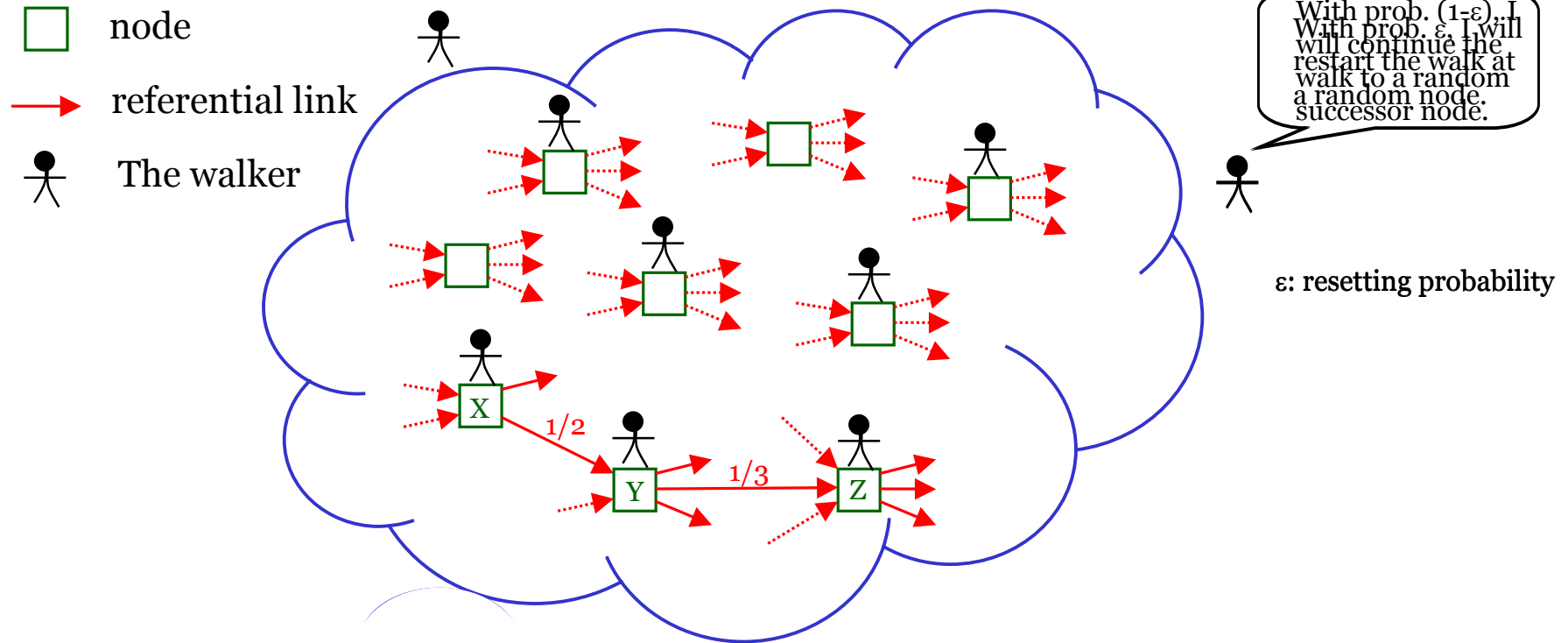
## **Research motivation**

- 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.

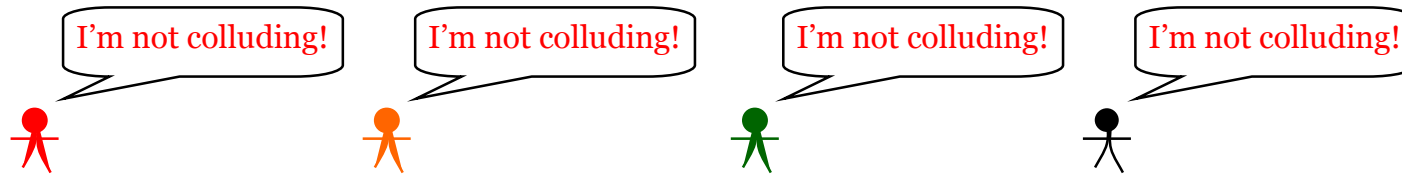
# PageRank: random walk model



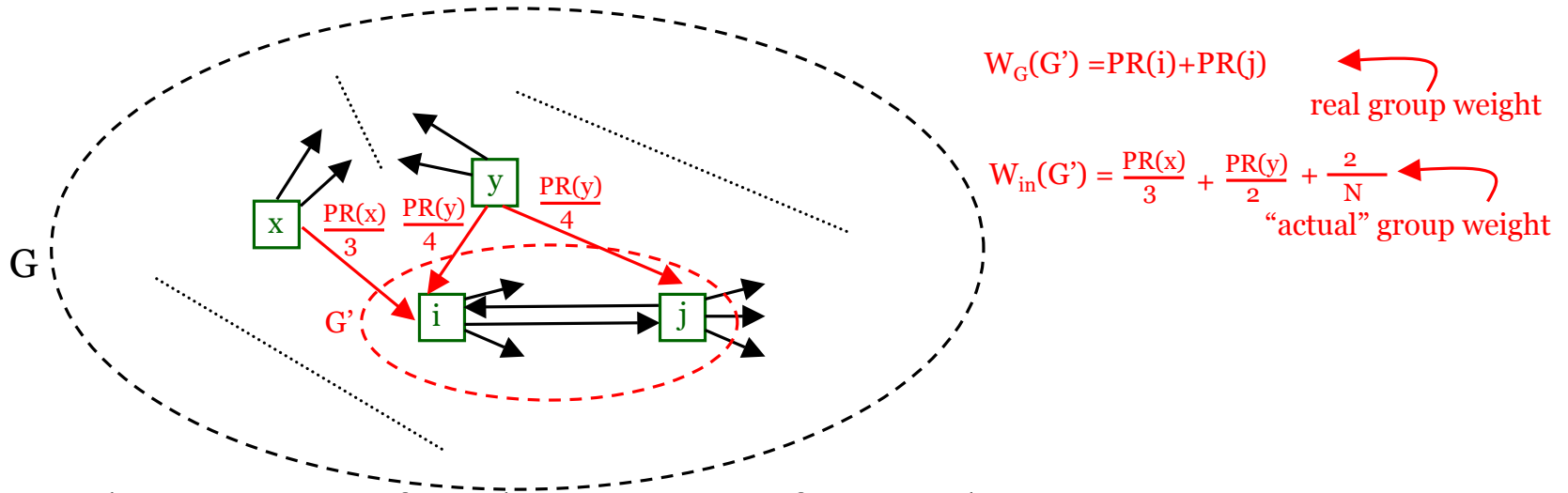
- As time goes on, the expected percentage of steps the walker is at each node  $v$  converges to the PageRank weight  $PR(v)$ .

## 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,

$$\forall G' \subseteq G, \text{Amp}(G') < \frac{1}{\varepsilon}$$

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<sup>th</sup>, 2000<sup>th</sup>, ..., 100000<sup>th</sup>.
  - $\varepsilon = 0.15$ .

# Experiment result of *Collusion200* (I)

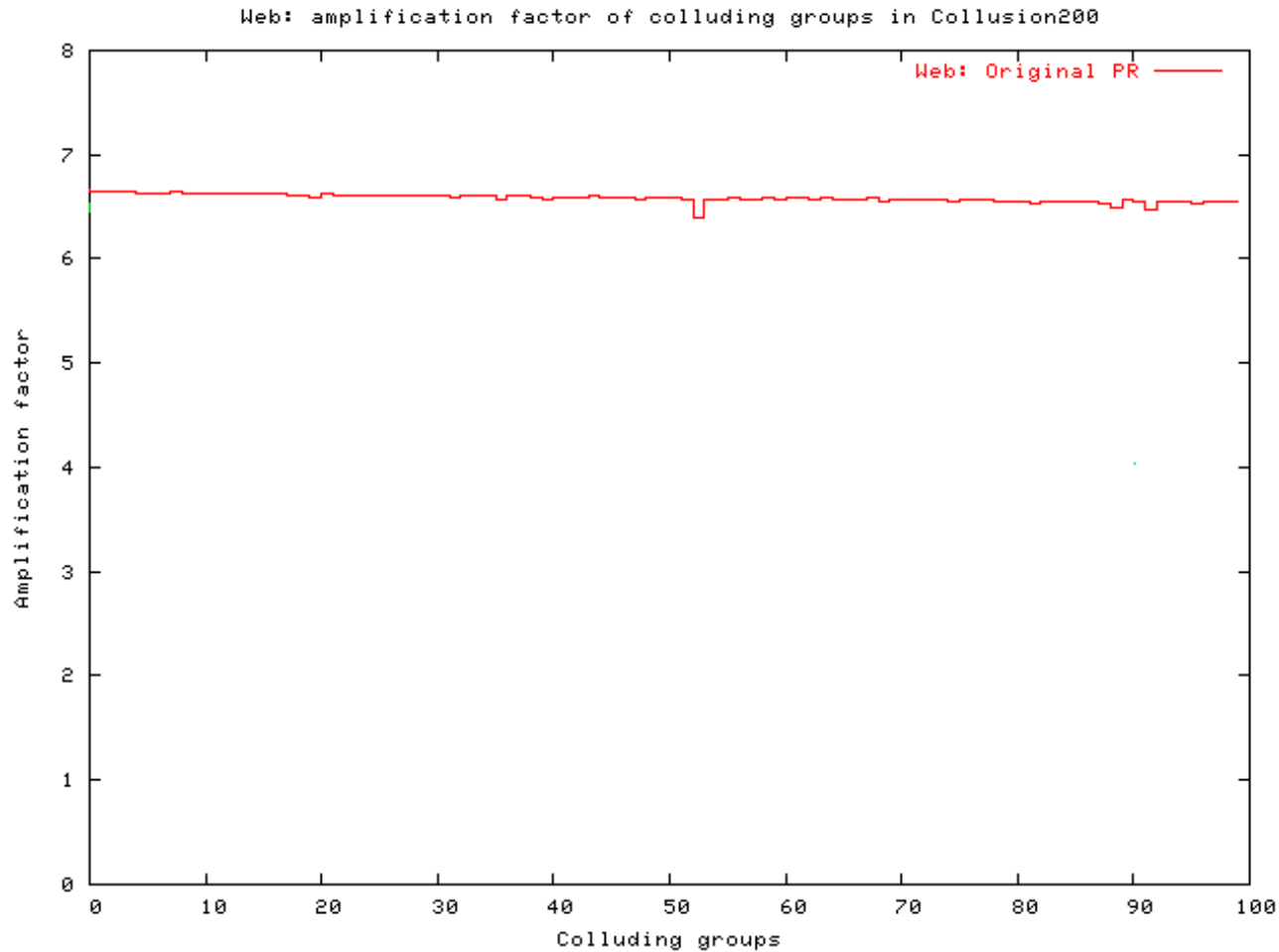


Figure 1:  $W$  - Amplification factors of the 100 colluding groups in *Collusion200*.

# Experiment result of *Collusion200* (III)

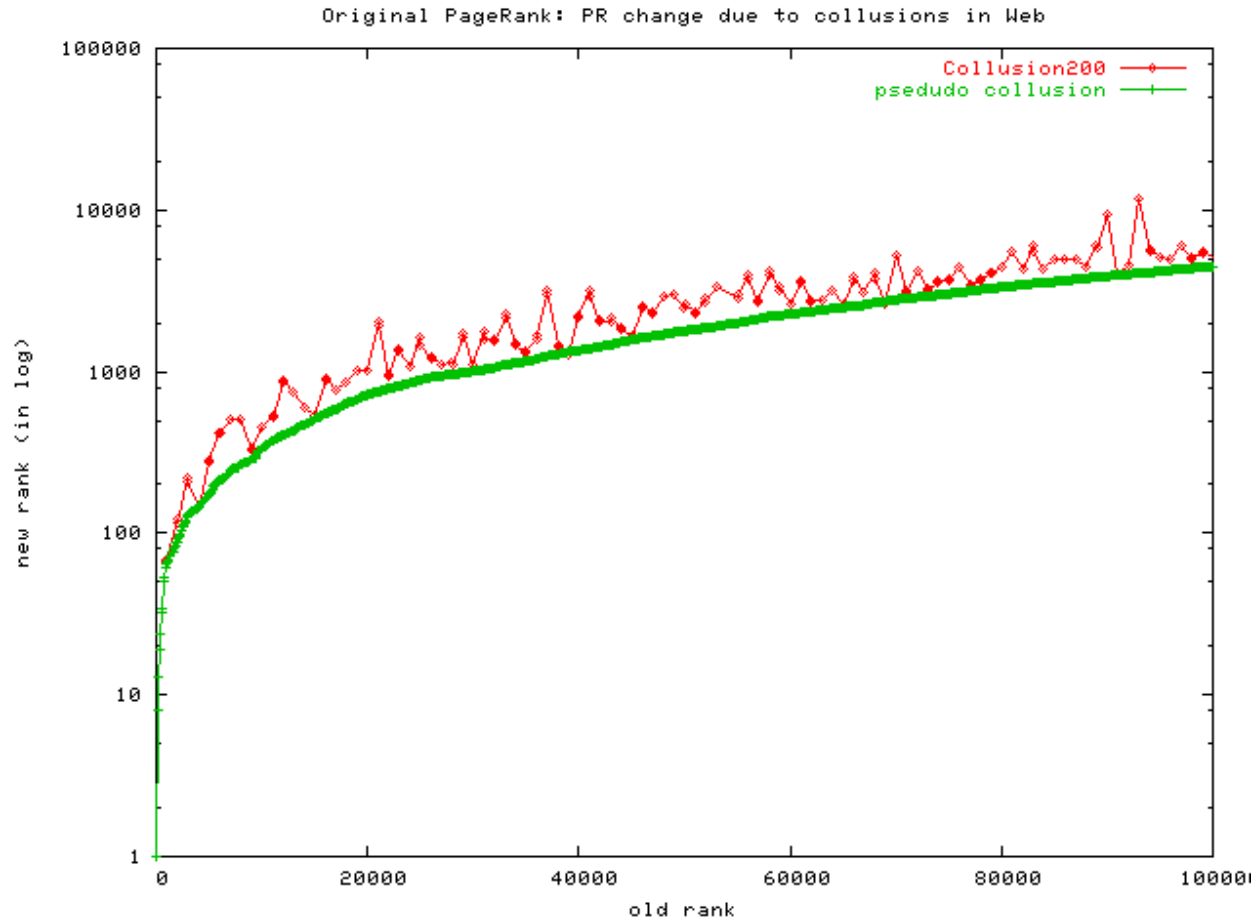


Figure 2:  $W$  – new PR rank after *Collusion200*.

# There is a long flat portion...

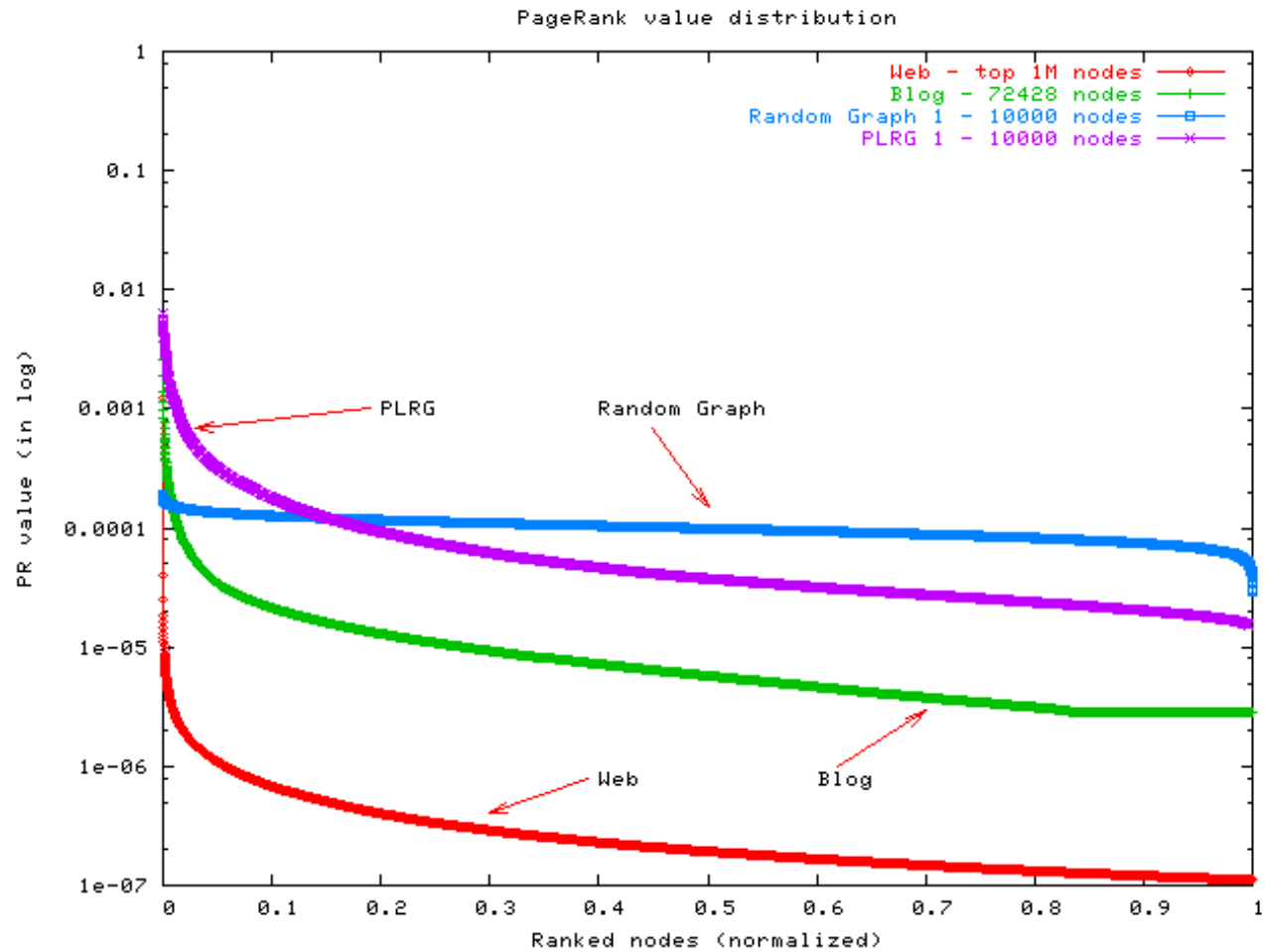


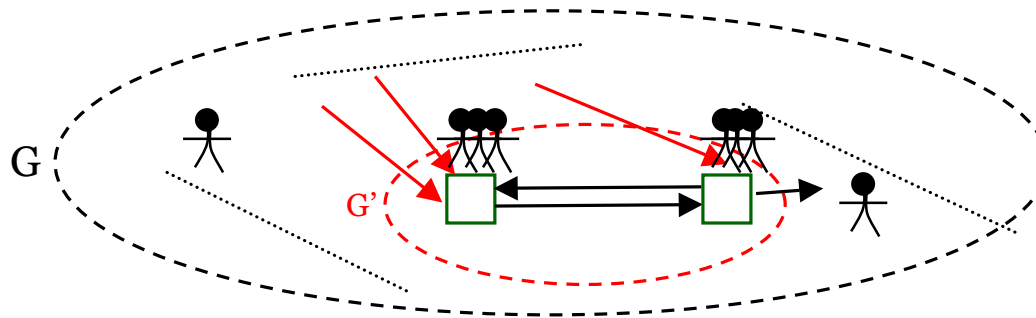
Figure 3: The PR weight distribution of 4 topologies.

## **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

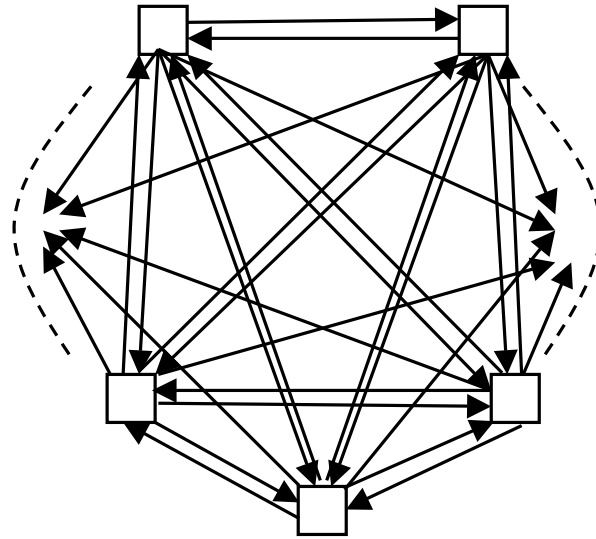
- To increase their PR weight, i.e., the stationary weight in the random walk, the colluding nodes will stall the random walk.



- When the resetting probability  $\varepsilon$  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  $1/\varepsilon$  (the average walk length), while that of non-colluding nodes is relatively insensitive to the change in  $\varepsilon$ .

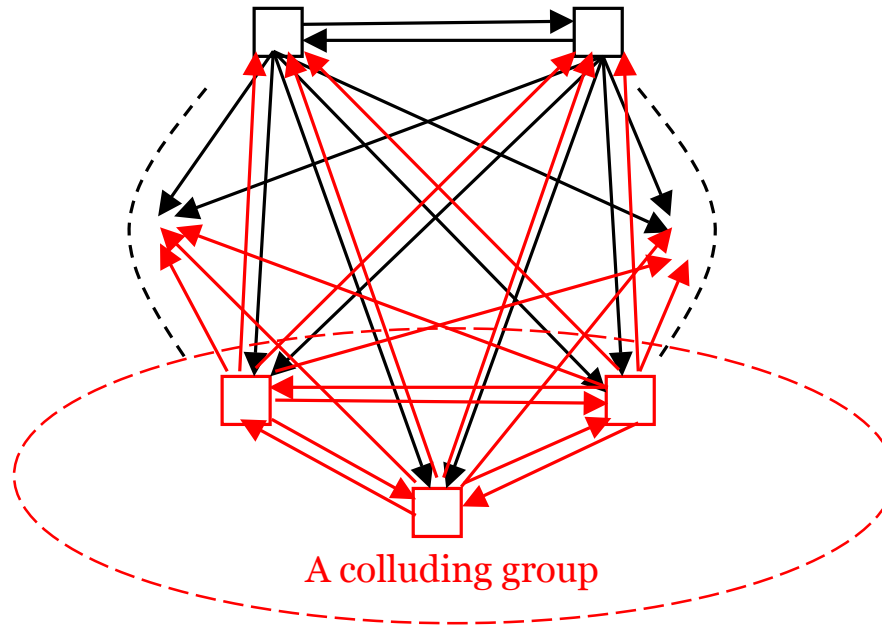
# An intuitive example

- node
- referential link

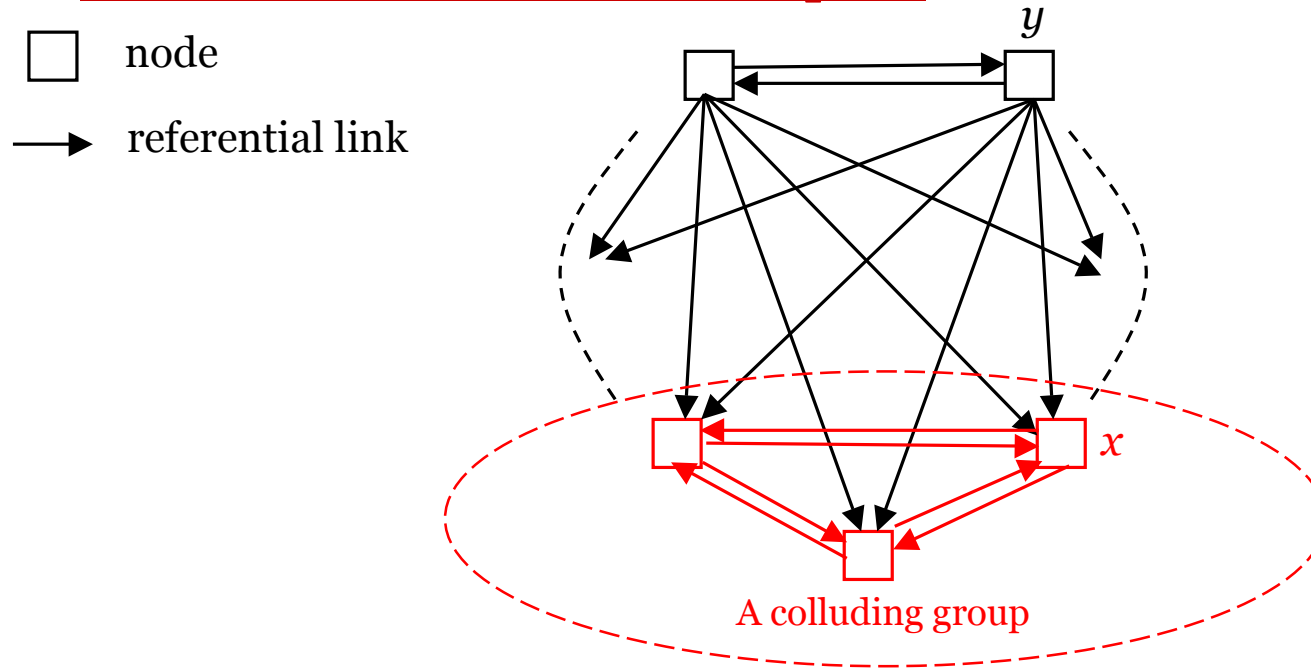


# An intuitive example

- node
- referential link



## An intuitive example



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

# Co-co distribution in real-world graphs

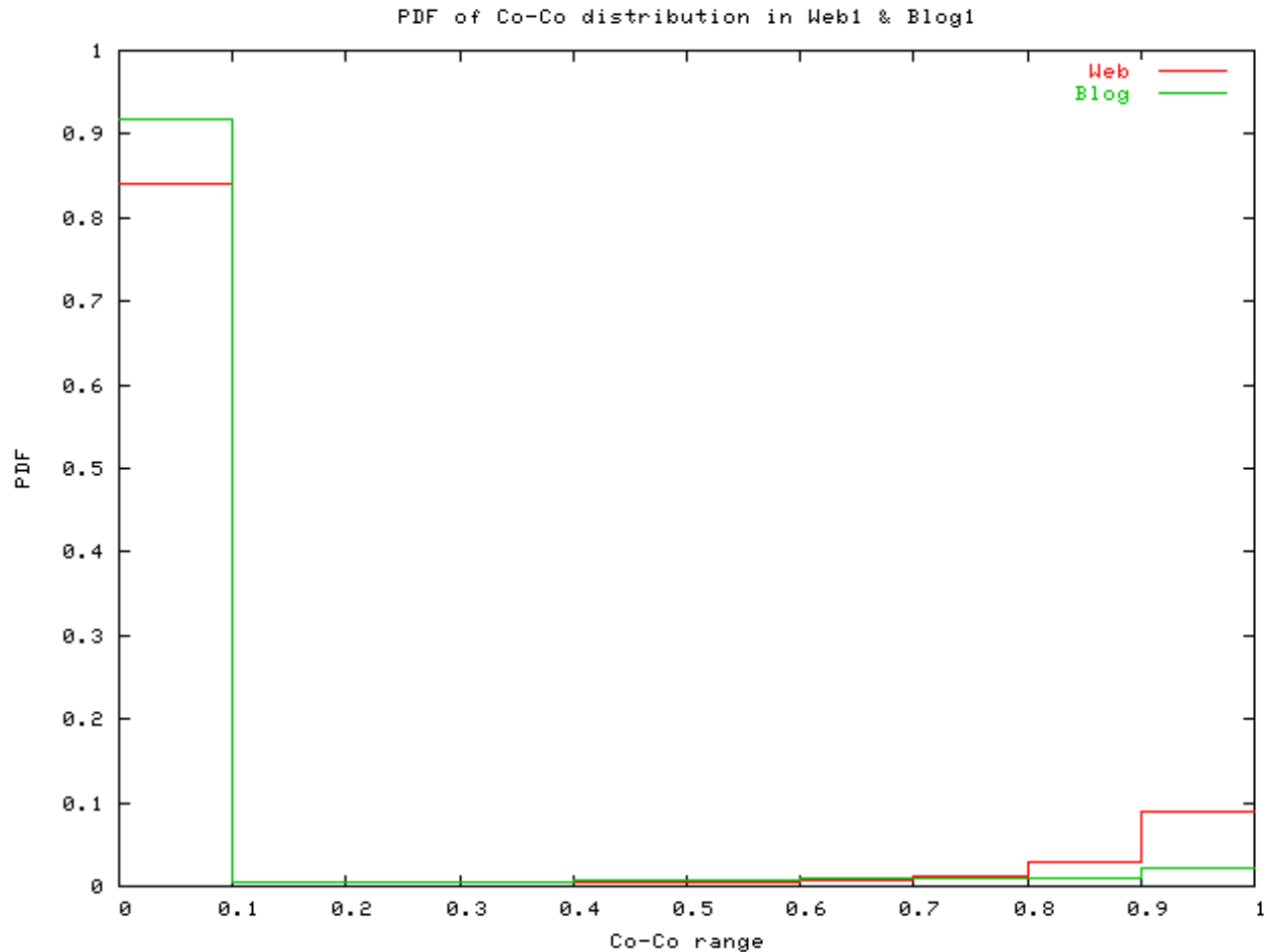


Figure 4: the co-co PDF distribution in  $W$  and  $B$  : the  $[0, 0.1]$  range actually corresponds to  $[-1, 0.1]$  range.

# Adaptive-resetting scheme

- Part I – collusion detection:
  - Given the topology, calculate the PR vector under different  $\varepsilon$  values.
    - $\{\varepsilon\} = \{0.0375, 0.05, 0.075, 0.15, 0.3, 0.45, 0.6\}$ ,  $\varepsilon_{default} = 0.15$ .
  - 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- $\varepsilon = F(\varepsilon_{default}, co-co(x))$ .
    - Exponential function  $F_{Exp} = \varepsilon_{default}^{(1.0-co-co(x))}$ .
    - Linear function  $F_{Linear} = \varepsilon_{default} + (0.5-\varepsilon_{default}) * co-co(x)$
  - The final PR weight vector is calculated with these personalized resetting values.

# Experiment result of *Collusion200* (IV)

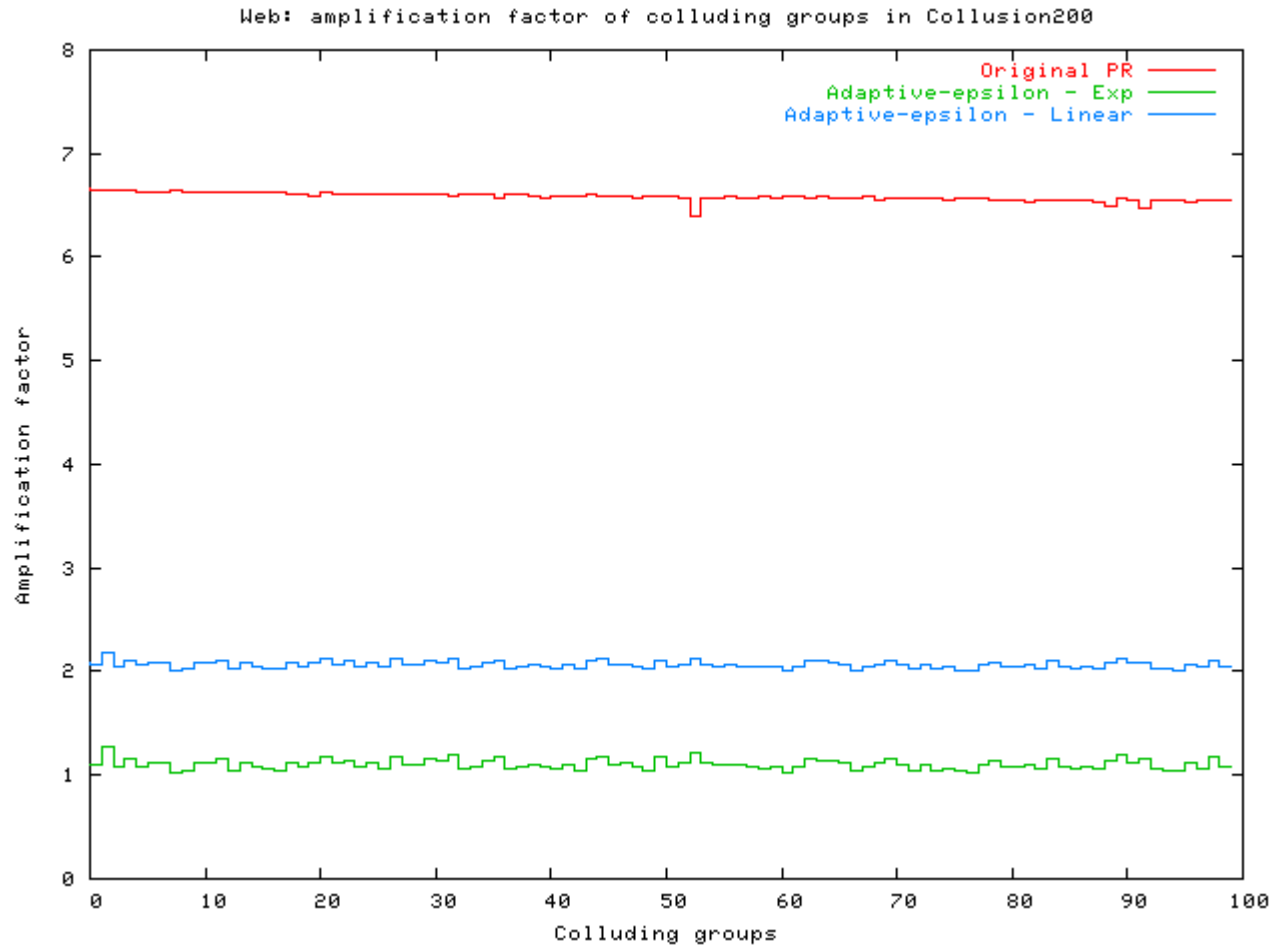


Figure 5:  $W$  - Amplification factors of the 100 colluding groups in *Collusion200*.

# Experiment result of *Collusion200* (V)

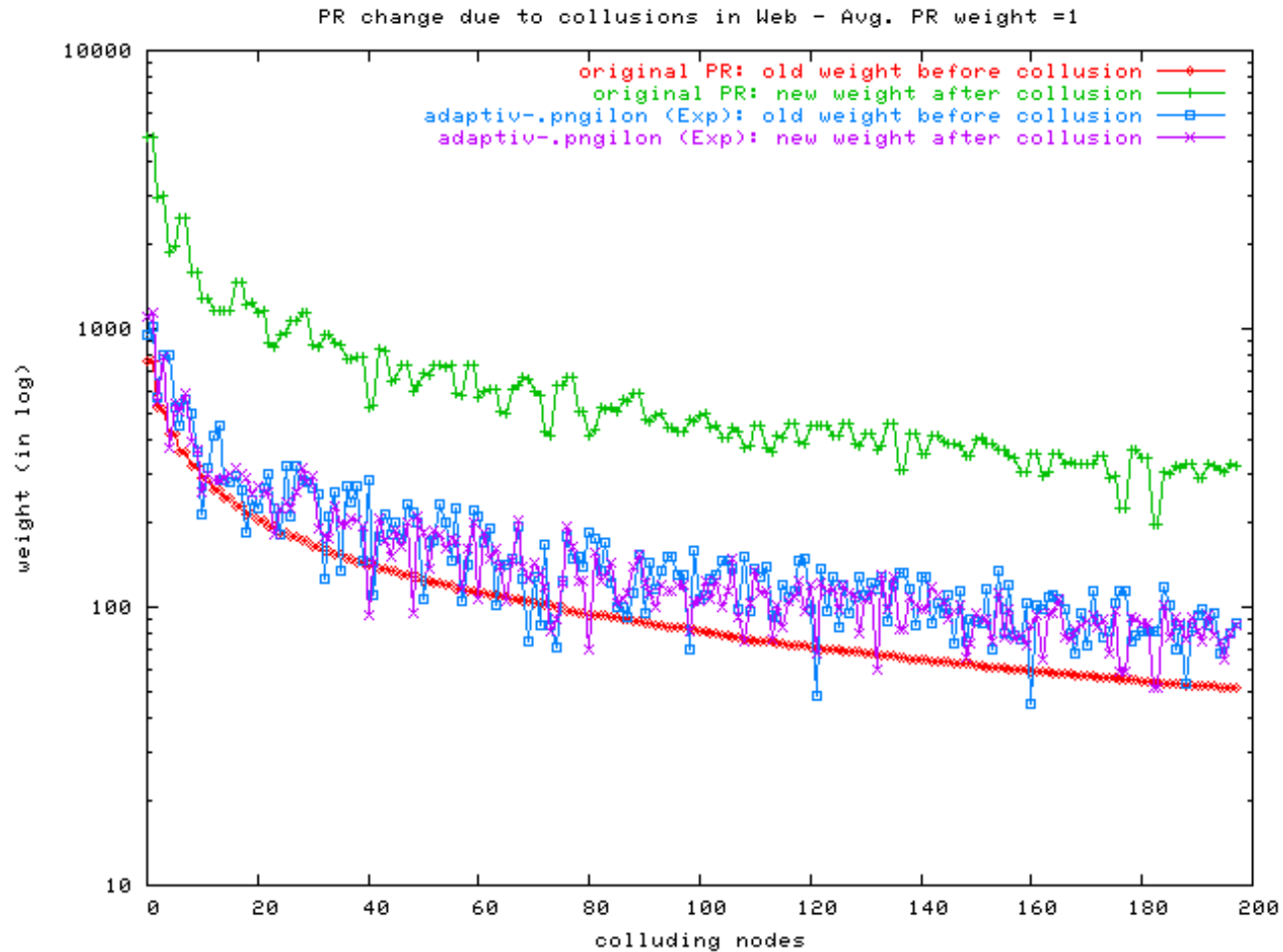


Figure 6:  $W$  – new PR weight after *Collusion200*.

# Experiment result of *Collusion200* (VI)

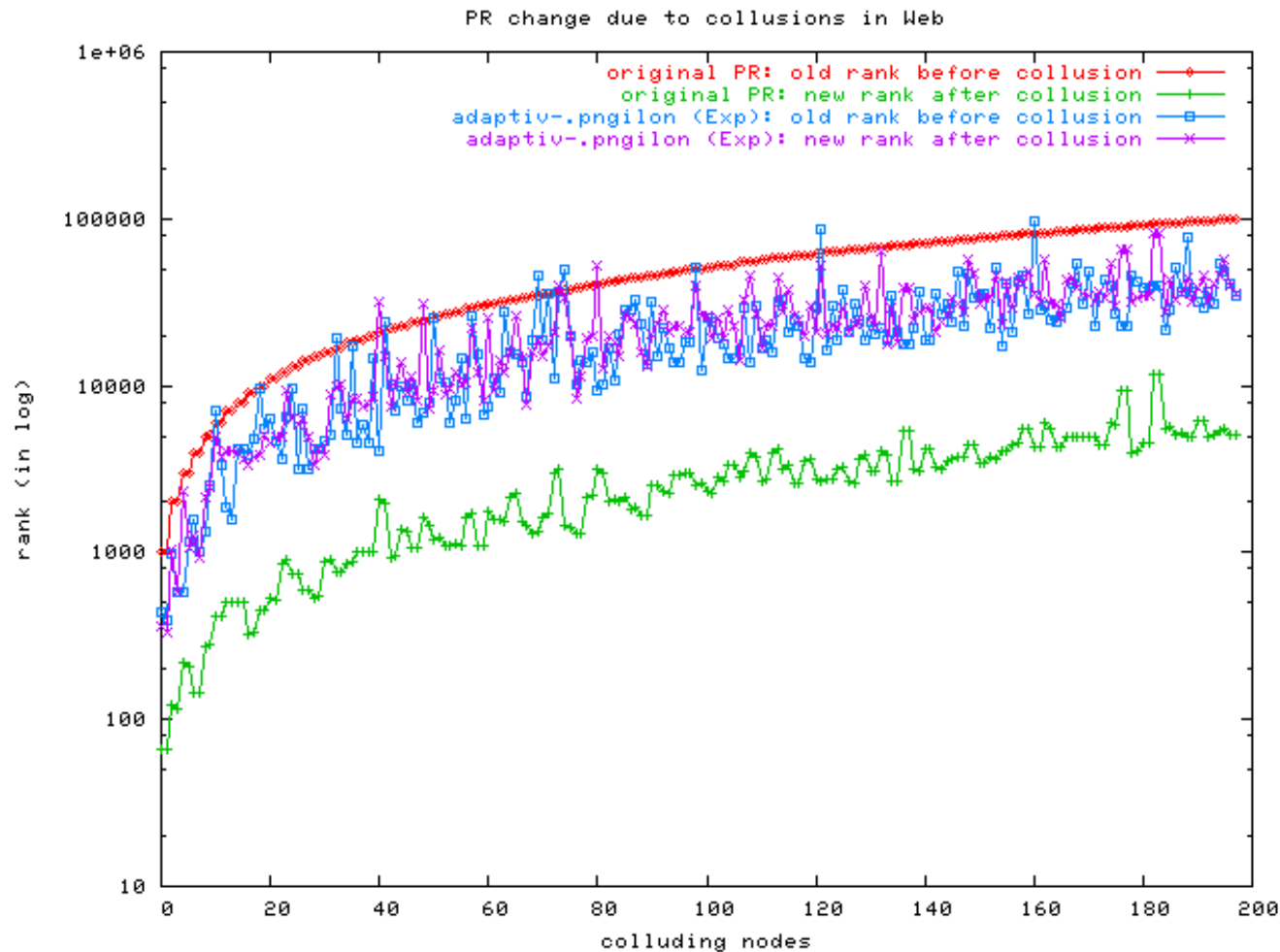
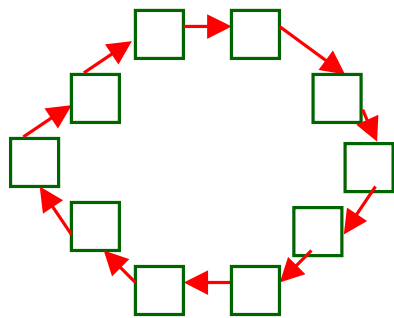


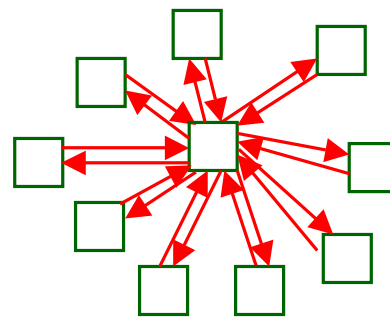
Figure 7:  $W$  – new PR rank after *Collusion200*.

## Experiment 2: Collusion22

- Model various colluding subgraphs.
- Methodology:
  - 3 colluding groups:



G1: 10-node ring



G2: 10-node star topology



G3: 2-node ring

□ node  
→ referential link

# Experiment result of *Collusion22* (I)

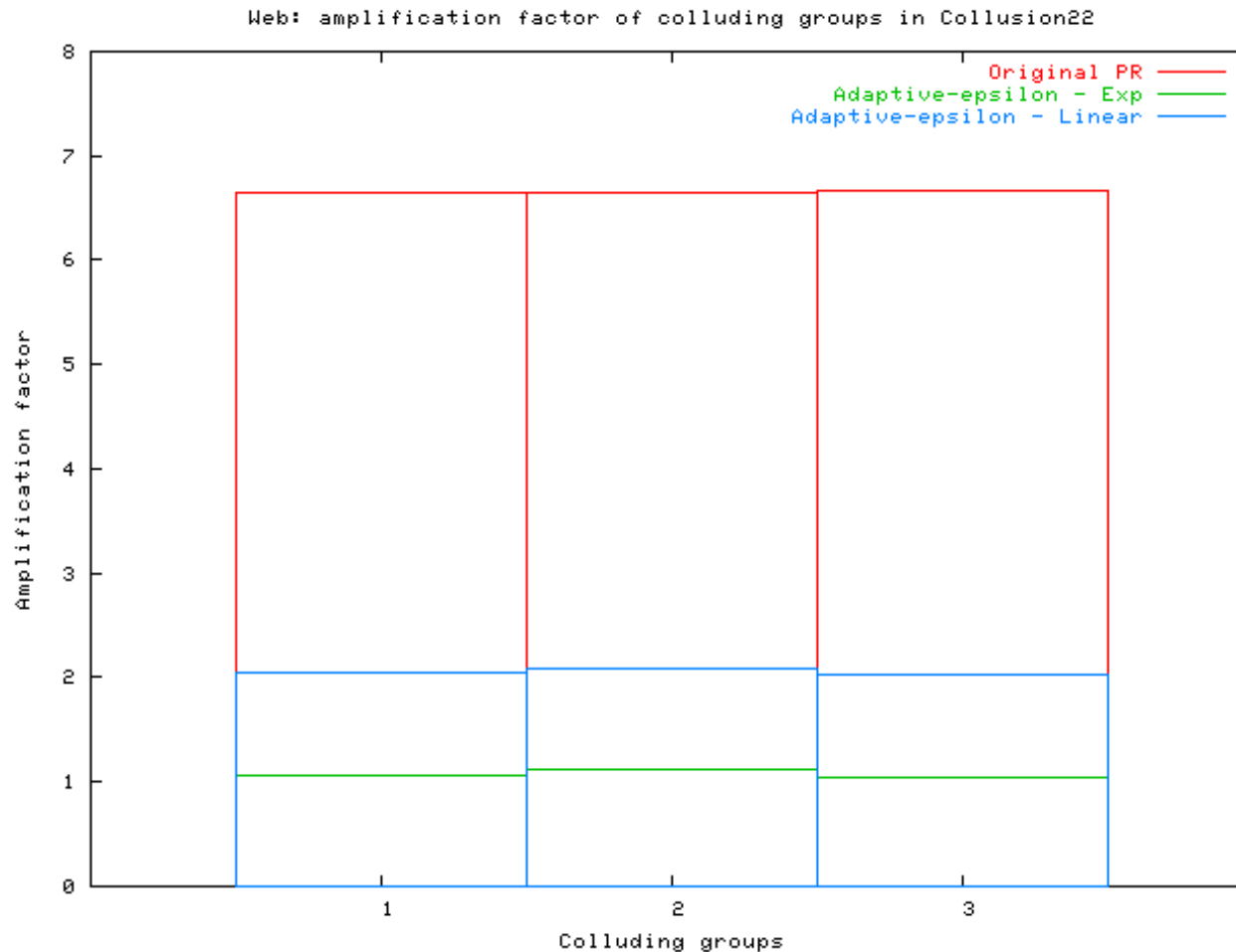


Figure 8: Amplification factors of the 3 colluding groups in *Collusion22*.

# Experiment result of *Collusion22* (II)

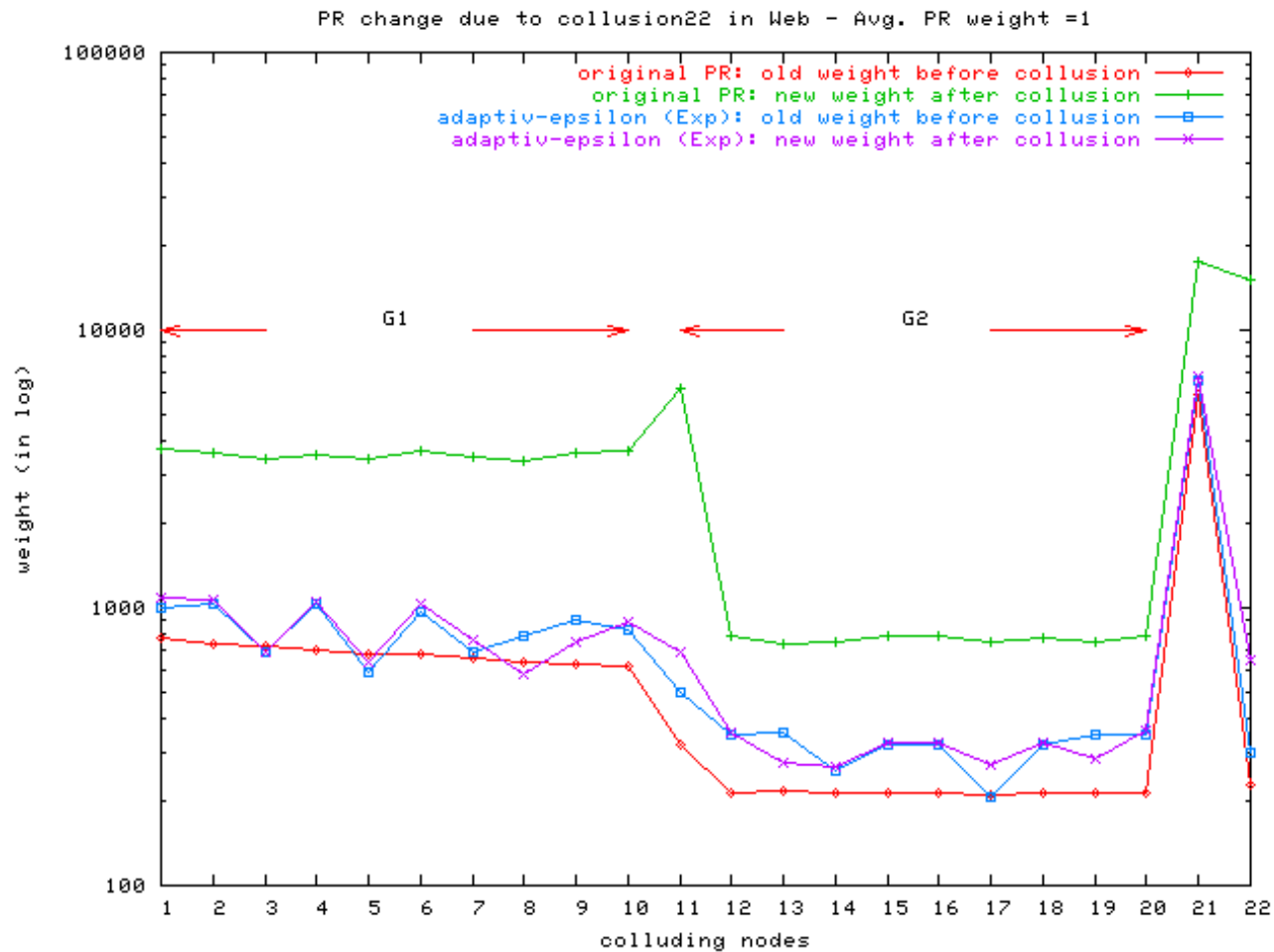


Figure 9:  $W$  – new PR weight after *Collusion22*.

# New top-25 URL list in $W$



Rank	Old list	New list
1	<del>http://www.yahwa.com/</del>	http://www.tucows.com/
2	<del>http://messenger.yahoo.com/</del>	http://www.yahoo.com/
3	http://www.tucows.com/	http://www.domaindirect.com/
4	http://www.domaindirect.com/	http://news.tucows.com/
5	http://news.tucows.com/	http://ispcentral.tucows.com/
6	http://ispcentral.tucows.com/	http://www.microsoft.com/
7	http://www.microsoft.com/	http://www.acme.com/software/thttpd
8	<del>http://www.microsoft.com/info/copyright.htm</del>	http://www.adobe.com/products/acrobat/readstep.html
9	http://www.adobe.com/products/acrobat/readstep.html	http://home.netscape.com/
10	http://home.netscape.com/	http://www.thecounter.com/
11	<del>http://www.dun.com/</del>	http://www.gendex.com/ged2html
12	<del>http://www.worldwidemart.com/scripts</del>	http://www.adobe.com/
13	http://www.acme.com/software/thttpd	http://www.worldwidemart.com/scripts
14	<del>http://search.internet.com/</del>	http://upload.tucows.com/contactus.html
15	http://upload.tucows.com/contactus.html	http://www.w3.org/
16	http://www.thecounter.com/	http://www.listbot.com/
17	http://www.listbot.com/	http://www.tucows.com/privacy.html
18	http://www.w3.org/	<del>http://www.worldwidemart.com/scripts/faq/wwwboard</del>
19	http://www.adobe.com/	<del>http://www.microsoft.com/windows/w/default.htm</del>
20	<del>http://www.tucows.com/search.html</del>	<del>http://www.mega.gov/</del>
21	http://www.tucows.com/privacy.html	<del>http://www.hadi.com/</del>
22	http://www.gendex.com/ged2html	<del>http://www.rsa.org/</del>
23	<del>http://chl.levels.ac.uk/mikee/personal.html</del>	http://search.internet.com/
24	<del>http://www.achbar.com/misc/privacy.html</del>	<del>http://www.mca.gov/</del>
25	<del>http://www.achbar.com/homepage.html</del>	http://chl.levels.ac.uk/mikee/personal.html

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

## **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.