A Study of Link Farm Distribution and Evolution using a Time Series of Web Snapshots

Young-joo Chung, Masashi Toyoda, Masaru Kitsuregawa

Institute of Industrial Science
The University of Tokyo
Japan

OUTLINE

Motivation

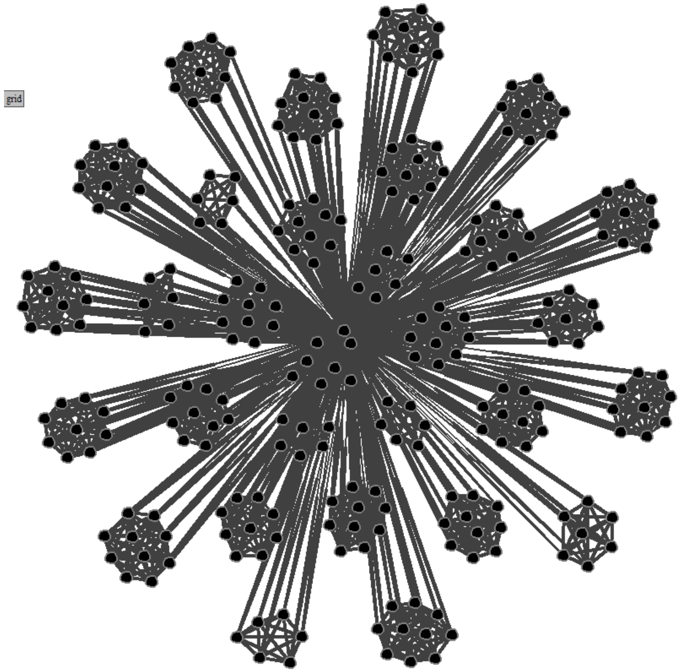
Approach

Experiment

Summary and Future Work

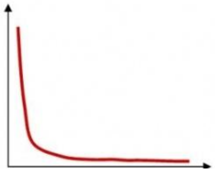
Link Farm

- Spammers create densely connected link structures to boost rank score of a target spam pages [Gyöngyi et al. VLDB 2005]

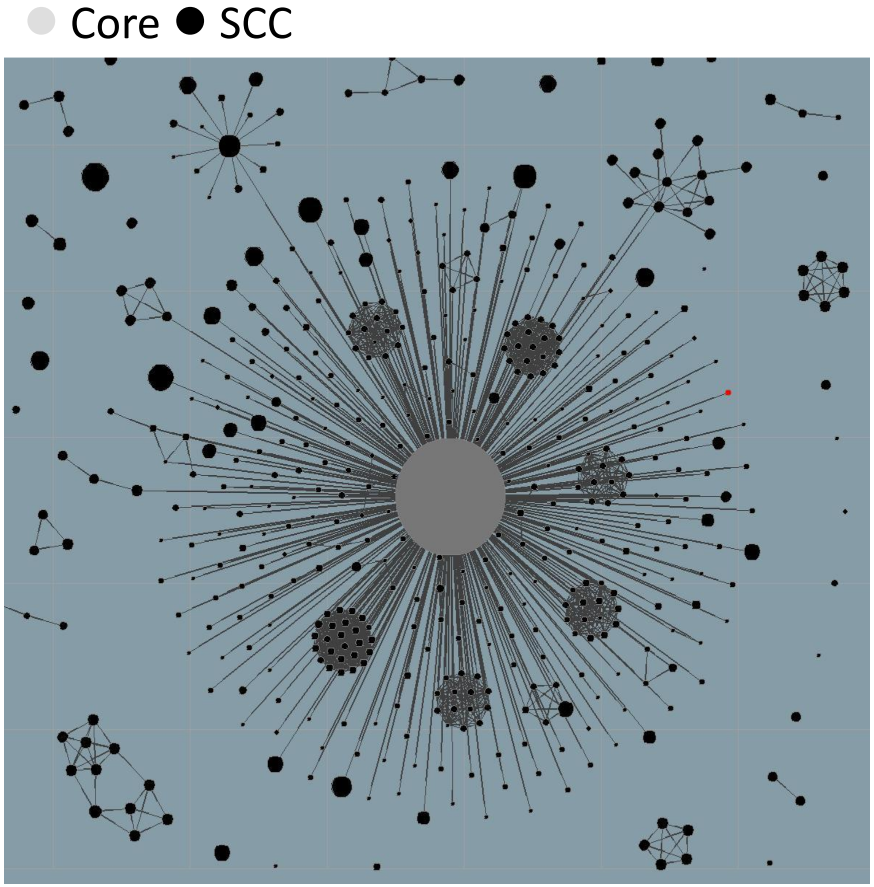


Strongly Connected Component and Link Farm

- SCC Decomposition of the Web graph [Broder et al. 2000]



  - Size distribution of SCCs follows the power-law.
  - The largest SCC (Core) is about 30% of all nodes
- Most large SCCs around the core are the link farm [Saito et al. AIRWEB 2007]



Distribution and Evolution of Link Farm

- Link farms in the core of the Web
  - To extract link farms in the core, apply recursive SCC decomposition with node filtering
  - Observe the size distribution of obtained SCCs
- Evolution of link farms in time series of Web snapshots
  - Find out the corresponding link farms from Web snapshots

OUTLINE

Motivation

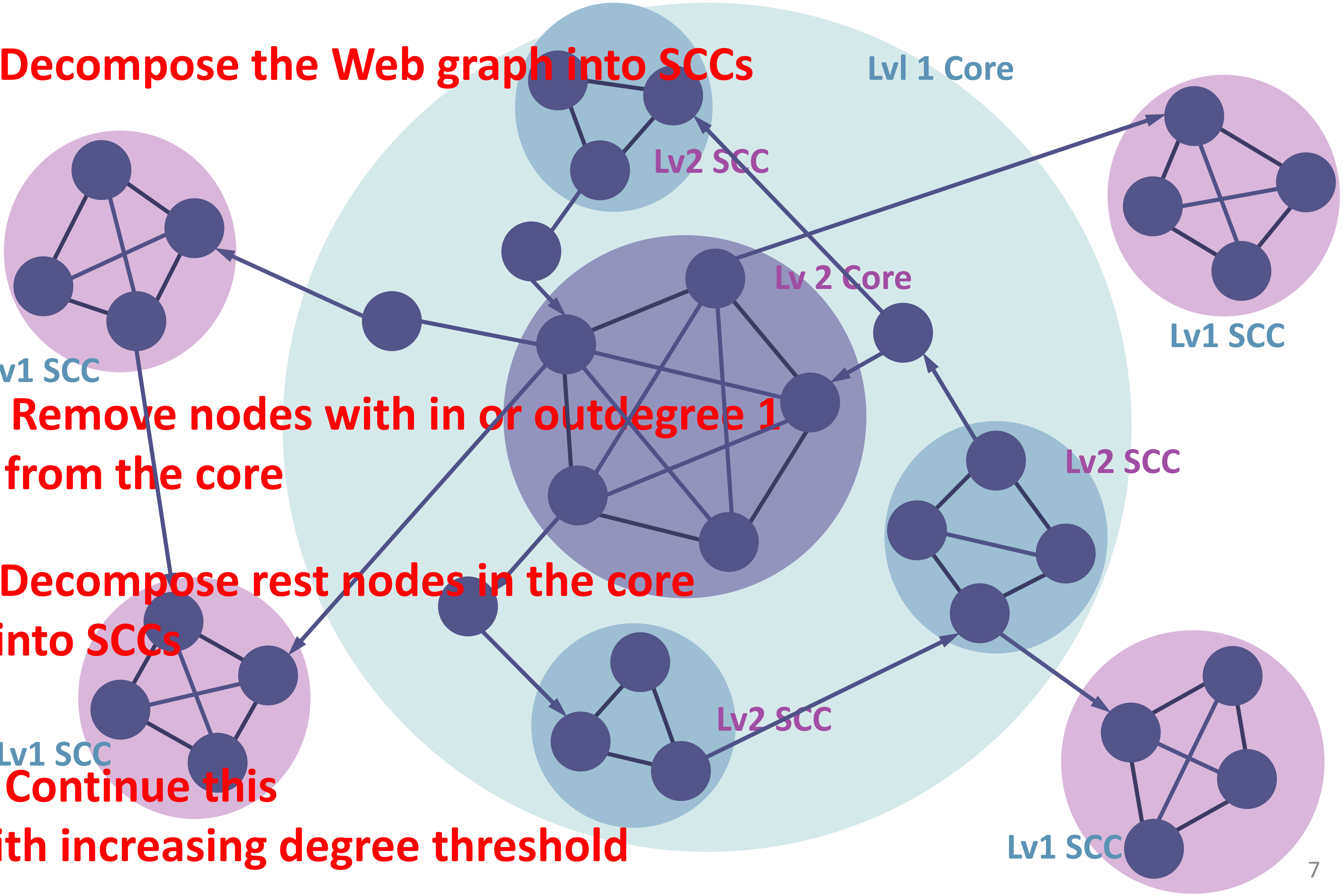
Approach

- Link farm extraction method
- Link farm evolution metrics

Experiment

Summary and Future Work

Recursive SCC Decomposition with Node Filtering



1. Decompose the Web graph into SCCs

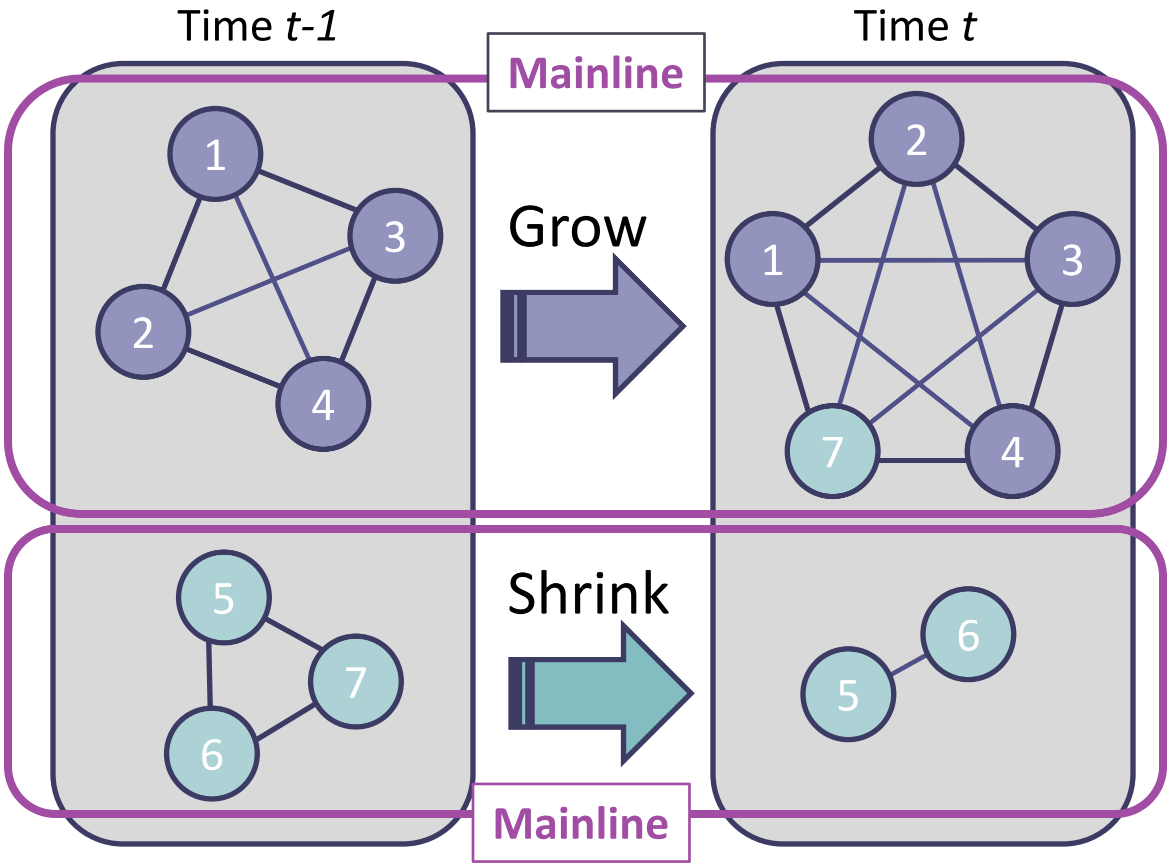
2. Remove nodes with in or outdegree 1
from the core

3. Decompose rest nodes in the core
into SCCs

4. Continue this
with increasing degree threshold

Evolution of Link Farm

Find out the corresponding SCCs in time series of Web snapshots



Corresponding SCC

a SCC in the previous time that shares the most hosts with the SCC in Time t

Mainline

A pair of SCC and its corresponding SCC.
If multiple corresponding SCCs exist,
choose the largest one

OUTLINE

Motivation and Goal

Approach

Experiment

- **Datasets**
- The result of Japanese dataset
- The result of WEBSPAM-UK dataset
- The result of link farm evolution

Summary and Future Work

Datasets

- Japanese Web archive
  (e-Society and Info-plosion project supported by MEXT*)
  - Crawled for 10 years from 1999, about 10 billion pages
  - Focusing on Japanese pages, but 40% pages written in other languages.
  - Host graphs from 2004 to 2006
    - Only hosts in 2006 snapshot are included

	2004	2005	2006
Host	2,978,223	3,702,029	4,017,250
Edge	67,956,304	83,072,645	82,077,459

- WEBSPAM-UK Dataset
  - Public dataset obtained by crawling hosts with .co.uk domain
  - Label data exist. (Normal, Spam, Undecided)

	2006	2007
Host	11,402	114,529
Edge	730,774	1,836,441
Labeled host	10,662	6,479
|Labeled| / |total|	93.5%	5.7%

*Ministry of Education, Culture, Sports, Science and Technology of Japan.

OUTLINE

Motivation and Goal

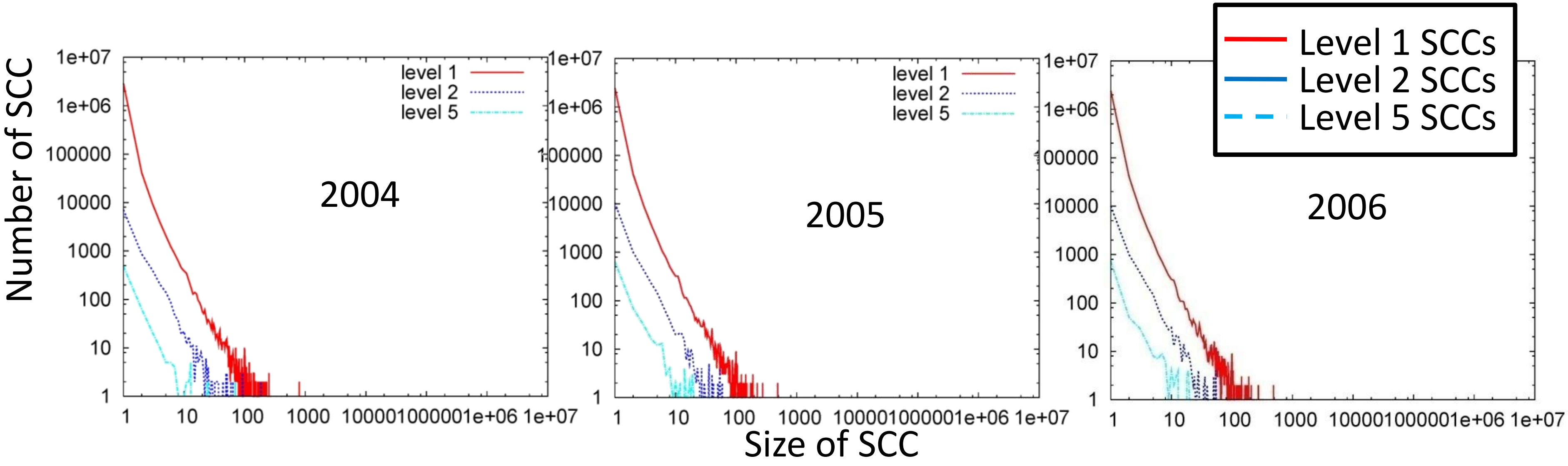
Approach

Experiment

- Datasets
- **The result of Japanese dataset**
- The result of WEBSPAM-UK dataset
- The result of link farm evolution

Summary and Future Work

SCC Size Distribution and Decomposition in JP Dataset



Distributions of SCCs in the deep of the core follow Power law with similar exponent to level 1 SCCs

The fraction of the core size increases drastically from level 1 to level 2, and then keep similar value until level 10

# SCCs	1,888,550	9,055	612	127
Size of the largest SCC	749,166	520,554	301,120	195,926
|size of core| / |nodes|	25.15	93.60	99.51	99.85

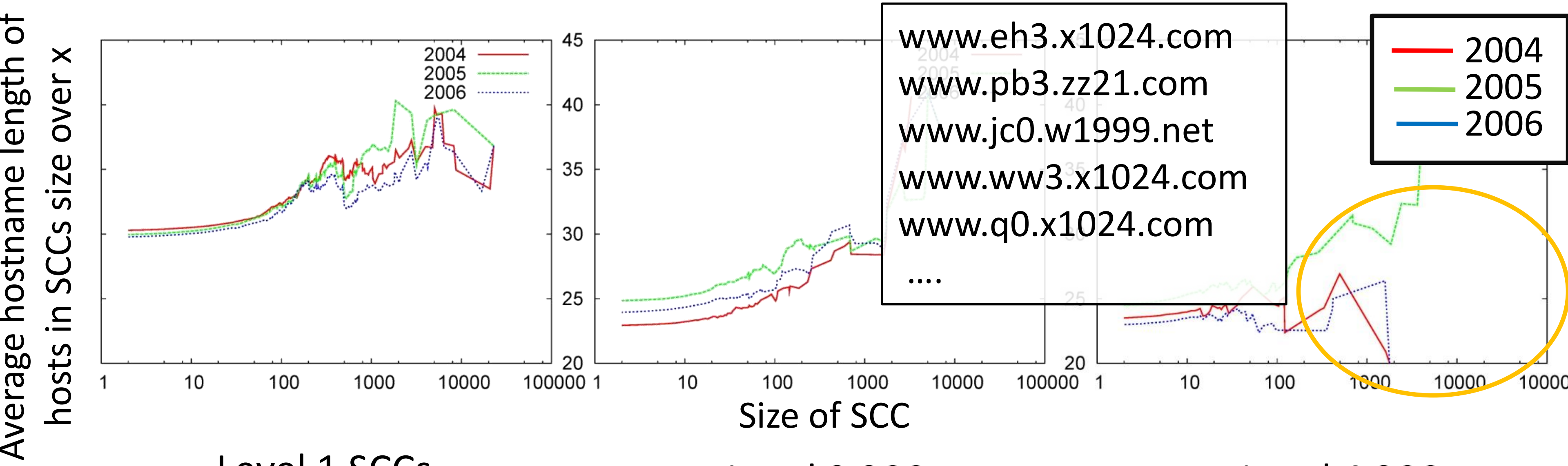
Spamicity by URL Properties

- Two metrics
  - Hostname length
    - Hosts with long URL are very likely to spam [Fetterly et al., WebDB 2004]
  - Spam keyword
    - URLs contain spam keywords are judged spam [Becchetti et al., AIRWEB 2006]
    - 114 Spam keywords are selected from SCCs(1000<) with frequency and by manual check
- If a SCC has many members whose URLs are long or contain spam keywords, that SCC is likely to be a link farm

Hostname in one SCC

```
www.cheap-motorcycle.co.uk
www.cheap-sports-tickets.co.uk
www.cheap-bank-loan.co.uk
www.cheap-taxi.co.uk
www.car-number-plate.net
www.cheap-cars.net
www.cheap-dvd-players.net
www.cheap-motor-car-insurance.co.uk
www.cheap-mortgage.net
www.cheap-loans-uk.net
www.cheap-motorbike-insurance.com
www.cheap-health-insurance.co.uk
www.cheap-insurance.co.uk
www.cheap-laptop-computers.co.uk
www.cheap-life-insurance.com
www.cheap-credit-cards.net
www.cheap-videos.com
www.medical-health-insurance.net
www.cheap-van.co.uk
www.cheap-gas-electricity.co.uk
www.cheap-car.net
www.cheap-medical-insurance.co.uk
```

Hostname Length of SCCs in JP Dataset



www.eh3.x1024.com
www.pb3.zz21.com
www.jc0.w1999.net
www.ww3.x1024.com
www.q0.x1024.com
….

Level 1 SCCs

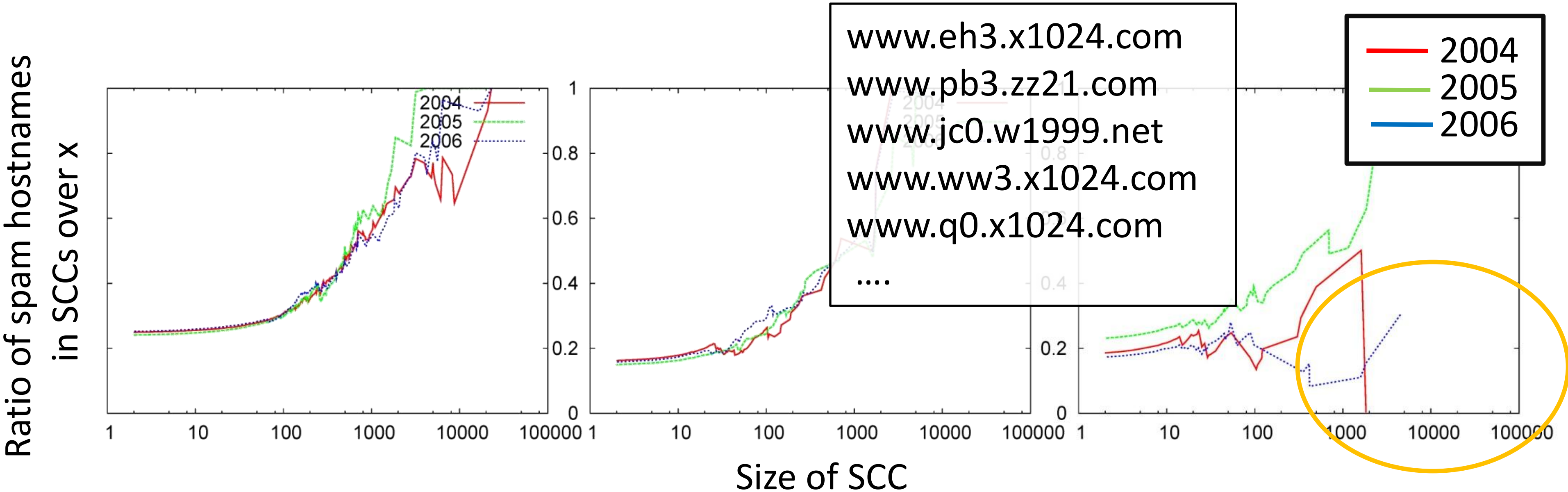
Level 2 SCCs

Level 4 SCCs

- As the size of SCC increases, the average hostname length also increases
- Large SCCs with short hostnames are manually checked, and we found that they are also spam.

Large SCCs have high spamicity!

Spam Keyword in Hostname in JP Dataset



www.eh3.x1024.com
www.pb3.zz21.com
www.jc0.w1999.net
www.ww3.x1024.com
www.q0.x1024.com
....

Level 1 SCCs

Level 2 SCCs

Level 4 SCCs

- As the size of SCC increases, the ratio of members containing spam keywords in their URL increases
- At the level 4, SCCs with low spamicity appeared.
- After manual check, we found out all hosts in such SCCs are spam without spam keyword in their URL

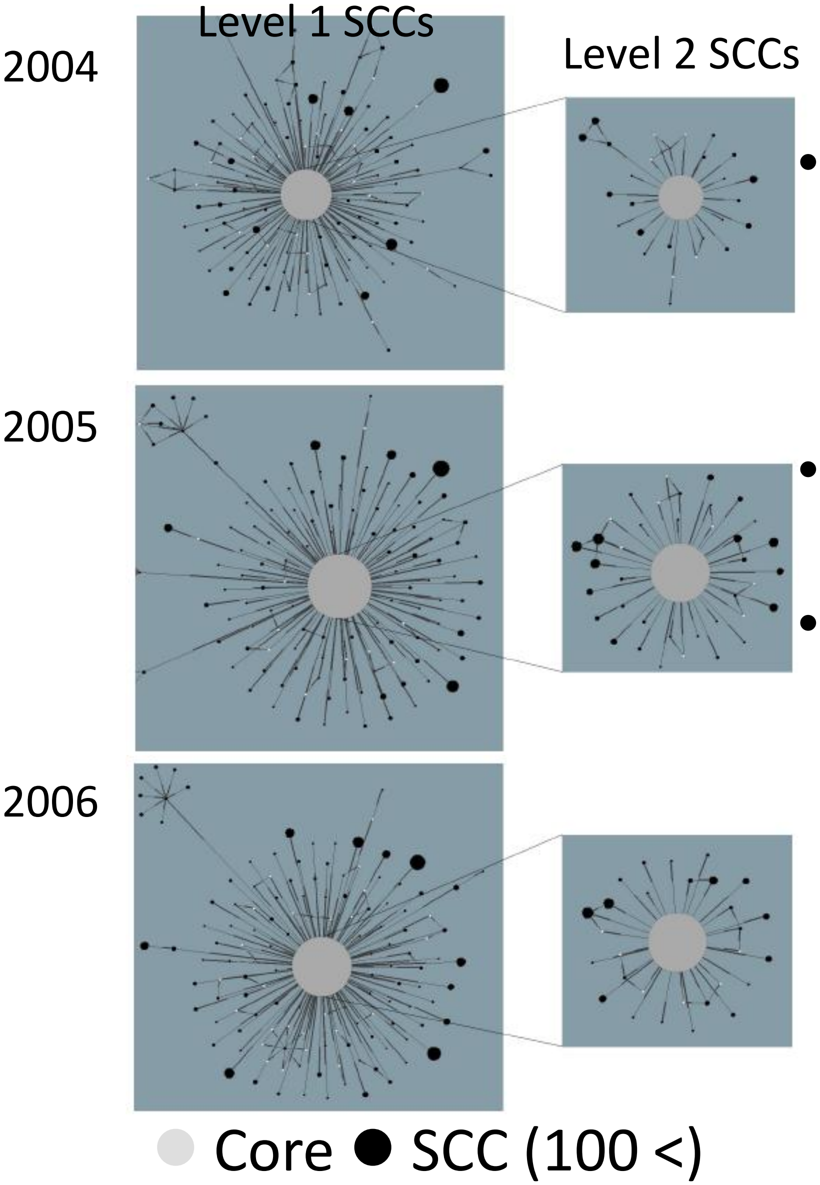
Large SCCs have high spamicity!

Spamicity of Large SCCs in JP Dataset

- We confirm a large SCC has a high spamicity
- Considering a SCC whose size is over 100 has a high spamicity, we found out 4.3%~7.2% hosts in the Web as a member of link farms, during 5 iterations.

		1	2	3	4	5
2004	# SCC	228	24	7	9	2
	# Host	182285	18650	9306	5032	242
2005	# SCC	167	32	18	13	7
	# Host	95347	38111	8236	15566	2789
2006	# SCC	180	26	21	6	8
	# Host	146015	26127	11092	9084	1499

Connectivity of Large SCCs in JP Dataset



- There was almost no connection between large SCCs in both the same level and the different level
- Link farms are isolated from each other
- For ranking algorithm based on the spam seed set, like Anti-TrustRank, comprehensive spam seed selection is needed for score propagation

OUTLINE

Motivation and Goal

Approach

Experiment

- Datasets
- The result of Japanese dataset
- **The result of WEBSPAM-UK dataset**
- The result of link farm evolution

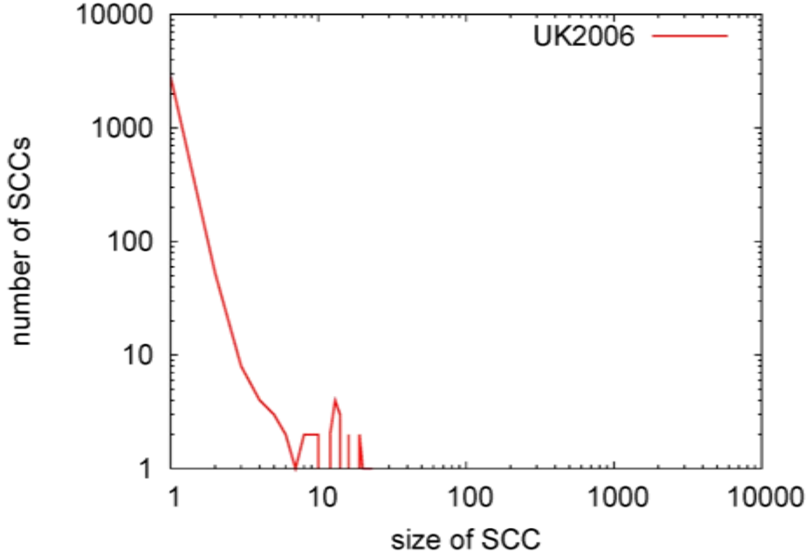
Summary and Future Work

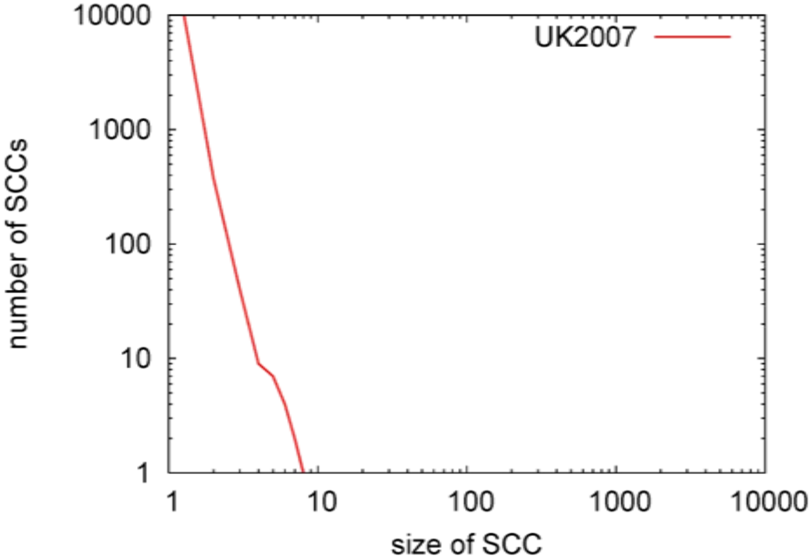
SCC Decomposition and Distribution in UK Dataset

Year	2006		2007	
Level	1	2	1	2
# of nodes	11,402	7,266	114,529	45,565
# of SCCs	2,935	574	54,822	969
Size of the core	7,945	6,683	59,160	44,564
|core| / |nodes| (%)	69.68	91.98	51.66	97.8
Size of 2nd largest SCC	73	6	8	3

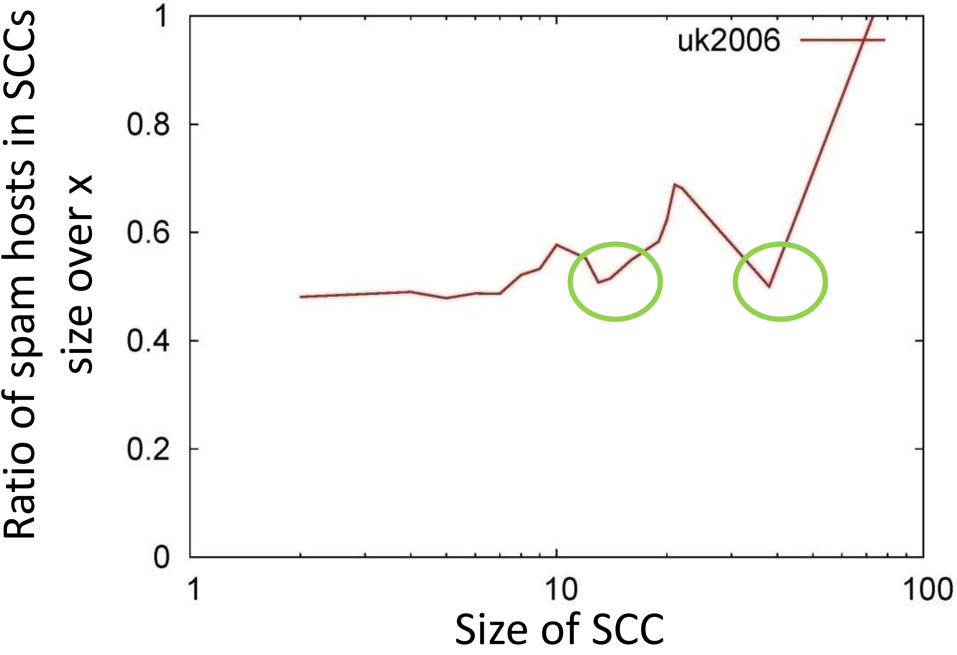
The fraction of the core was larger than that of JP dataset(25.1%)

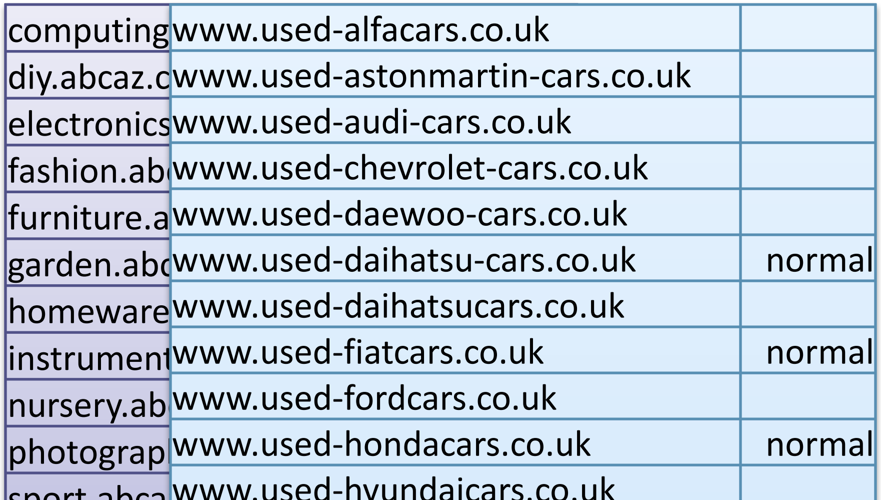
The sizes of SCC was much smaller than JP dataset





Spamicity of SCCs in UK Dataset





computing	www.used-alfacars.co.uk	
diy.abcaz.c	www.used-astonmartin-cars.co.uk	
electronics	www.used-audi-cars.co.uk	
fashion.ab	www.used-chevrolet-cars.co.uk	
furniture.a	www.used-daewoo-cars.co.uk	
garden.abc	www.used-daihatsu-cars.co.uk	normal
homeware	www.used-daihatsucars.co.uk	
instrument	www.used-fiatcars.co.uk	normal
nursery.ab	www.used-fordcars.co.uk	
photograp	www.used-hondacars.co.uk	normal
sport.abca	www.used-hyundaicars.co.uk	

- Large SCCs have high ratio of spam hosts
- 2 large SCCs have low spamicity
  - Shopping mall site with different hostnames for each category
  - Link farm with similar hostnames
- If we consider these 2 SCCs a link farm, total 282 host among 293 hosts were members of link farm(96.2%)

OUTLINE

Motivation and Goal

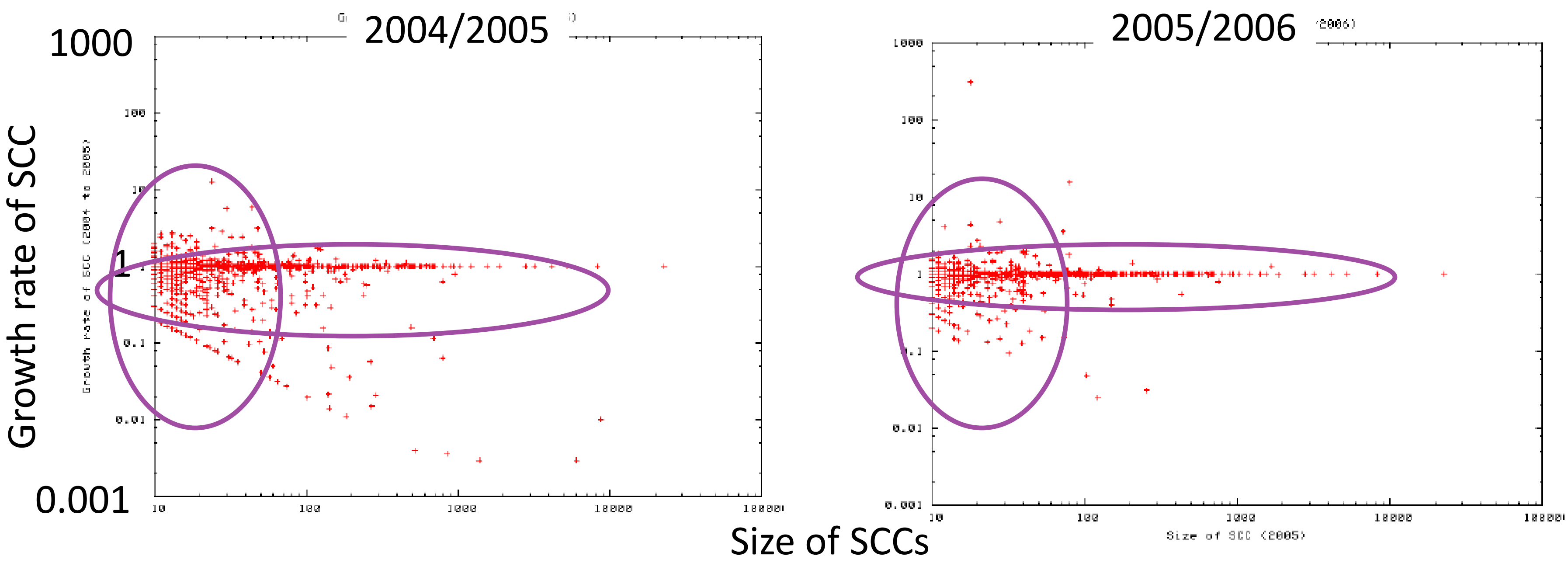
Approach

Experiment

- Datasets
- The result of Japanese dataset
- The result of WEBSPAM-UK dataset
- **The result of link farm evolution**

Summary and Future Work

Growth Rate of SCCs in JP Dataset



$$\text{Growth Rate} = \frac{|\text{SCC of the year}|}{|\text{SCC of previous year}|}$$

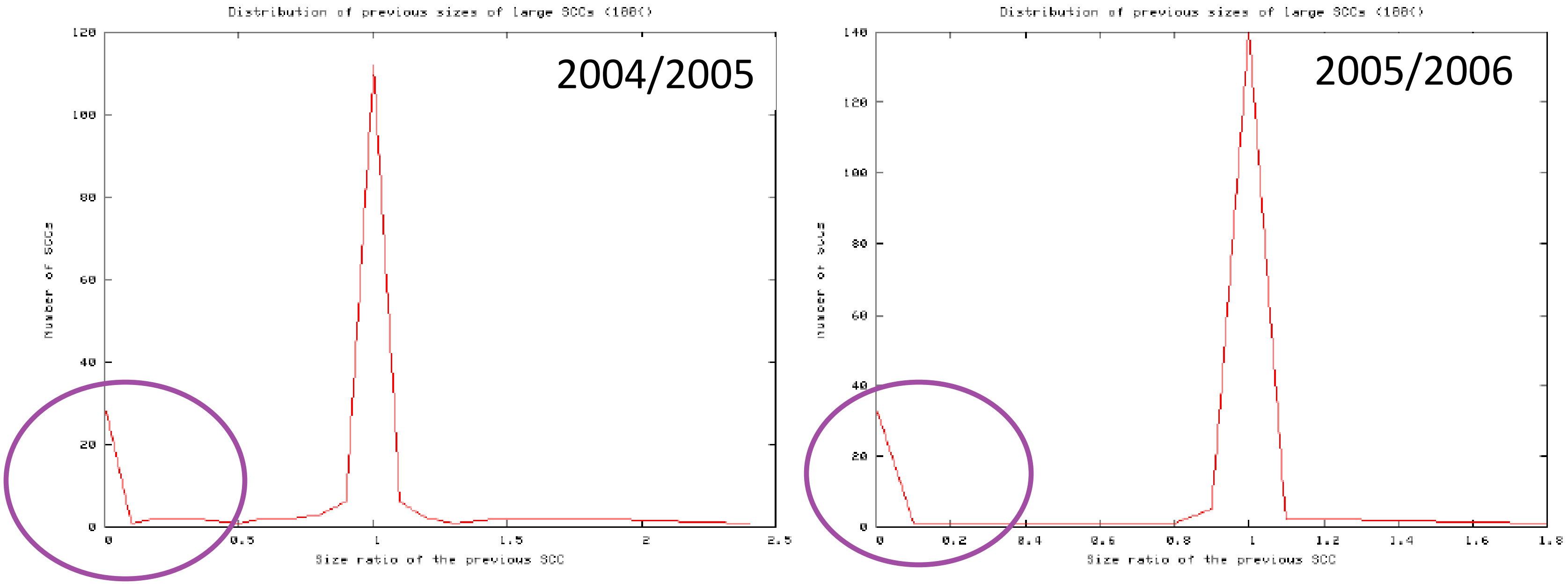
- Most SCCs did not changed in size

  This tendency gets stronger as the size of SCCs increases

- Small SCCs(size <100) follows Gibrat law, which means the growth rate is independent with its previous size

Previous size of Large SCCs in JP Dataset

Number of SCC



Size Ratio of SCC(100<) to the previous SCC

- Some large SCCs shrunk drastically during a year
- Spammers seem to either maintain their link farm or abandon, but do not bring them up
- To detect a newly appeared spam, it might not be helpful to tracking existing link farms

OUTLINE

Motivation and Goal

Approach

Experiment

Summary and Future Work

Summary

- Summary
  - Extracted SCCs in the core of the Web by recursive SCC decomposition
  - Evaluated the spamicity of large SCCs and confirmed that a large SCC has a high spamicity and isolated from each other
  - Observed the evolution of SCCs and found out large SCCs hardly grow
- Discussion
  - For the spam seed based ranking algorithm, comprehensive seed selection is needed
  - For the detection for new spam, tracking existing link farms is not helpful

Future Work

- Future Work
  - Observe the spam evolution with fine-grained time series of the Web snapshots
  - Observe the emergence and dissolution of link farms
- We are planning to distribute our host graph data to researchers.

Thank you for listening!