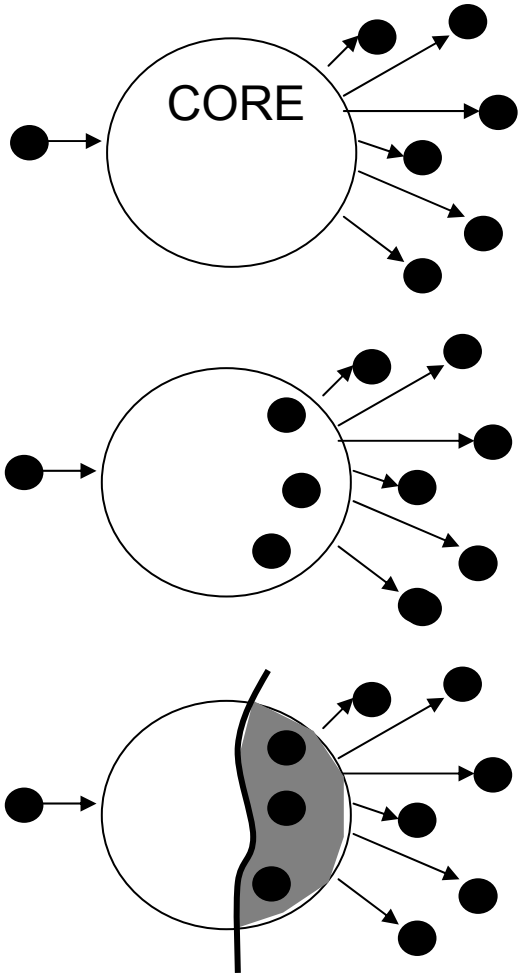
A Large-Scale Study of Link Spam Detection by Graph Algorithms

Hiroo Saito — University of Tokyo. JST, ERATO
Masashi Toyoda — University of Tokyo
Masaru Kitsuregawa — University of Tokyo
Kazuyuki Aihara — 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

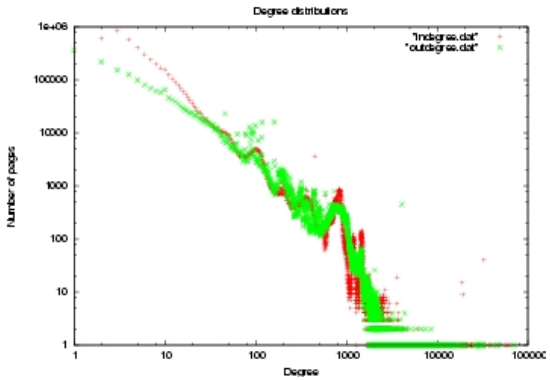
- Japanese Web archive crawled in May 2004
  - 96 million pages, 4.5 billion links
  - 60% pages in Japanese, 40% in other languages
- Site graph
  - Top of site: URL linked from 3 or more servers
  - A site is a set of URLs below the top URL
  - 5.9 million sites, 283 million links

Domains

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

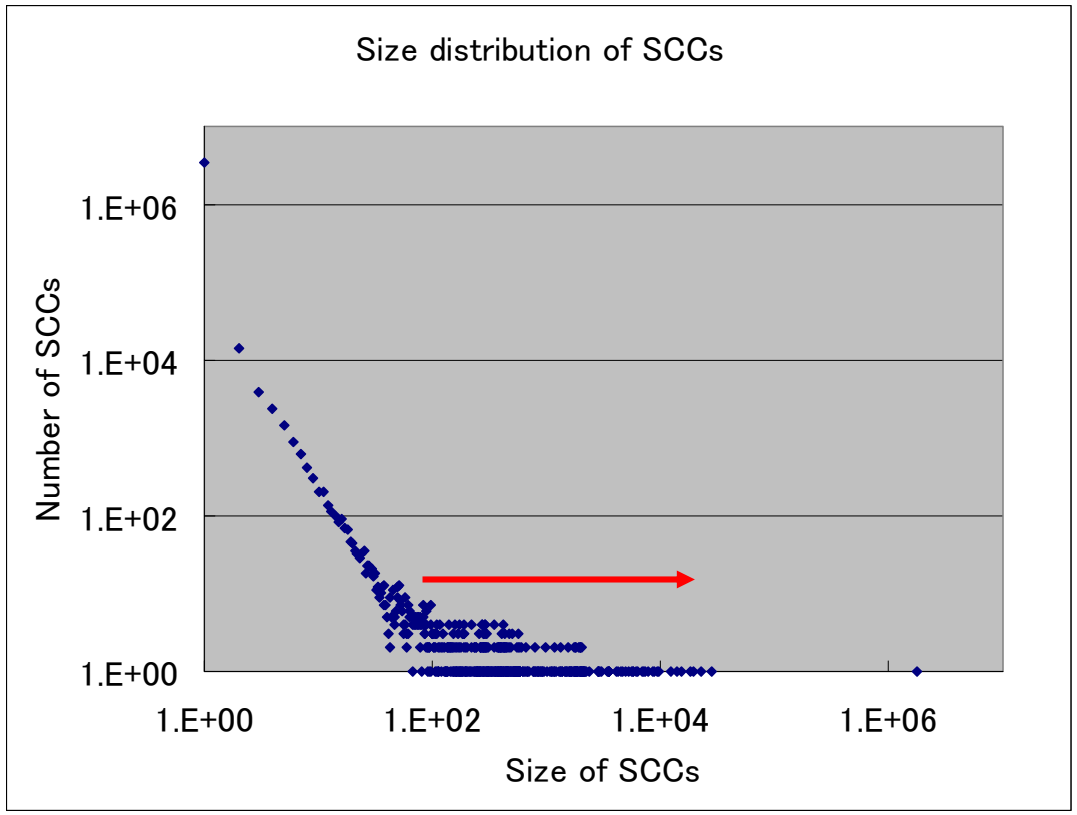
Degree

max. of indegree	61,006
avr. of indegree	48
max. of outdegree	70,294
avr. of outdegree	48



SCC decomposition

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

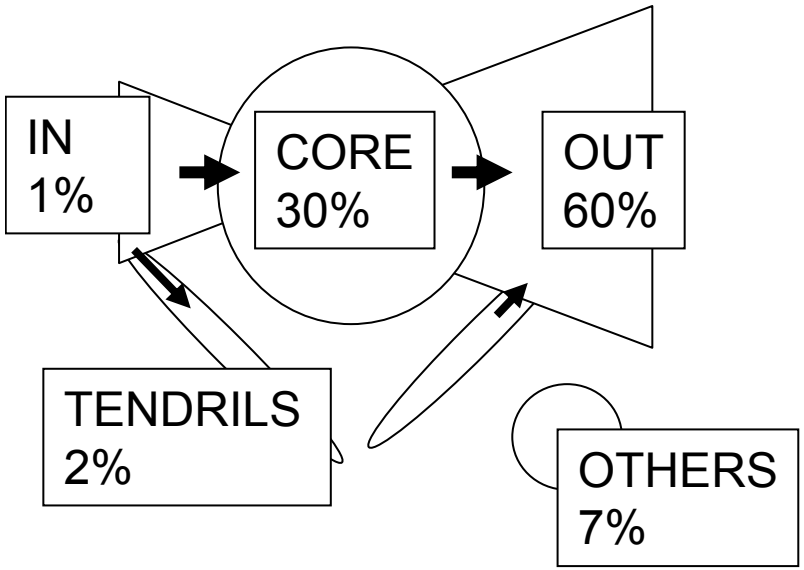


Sampling results

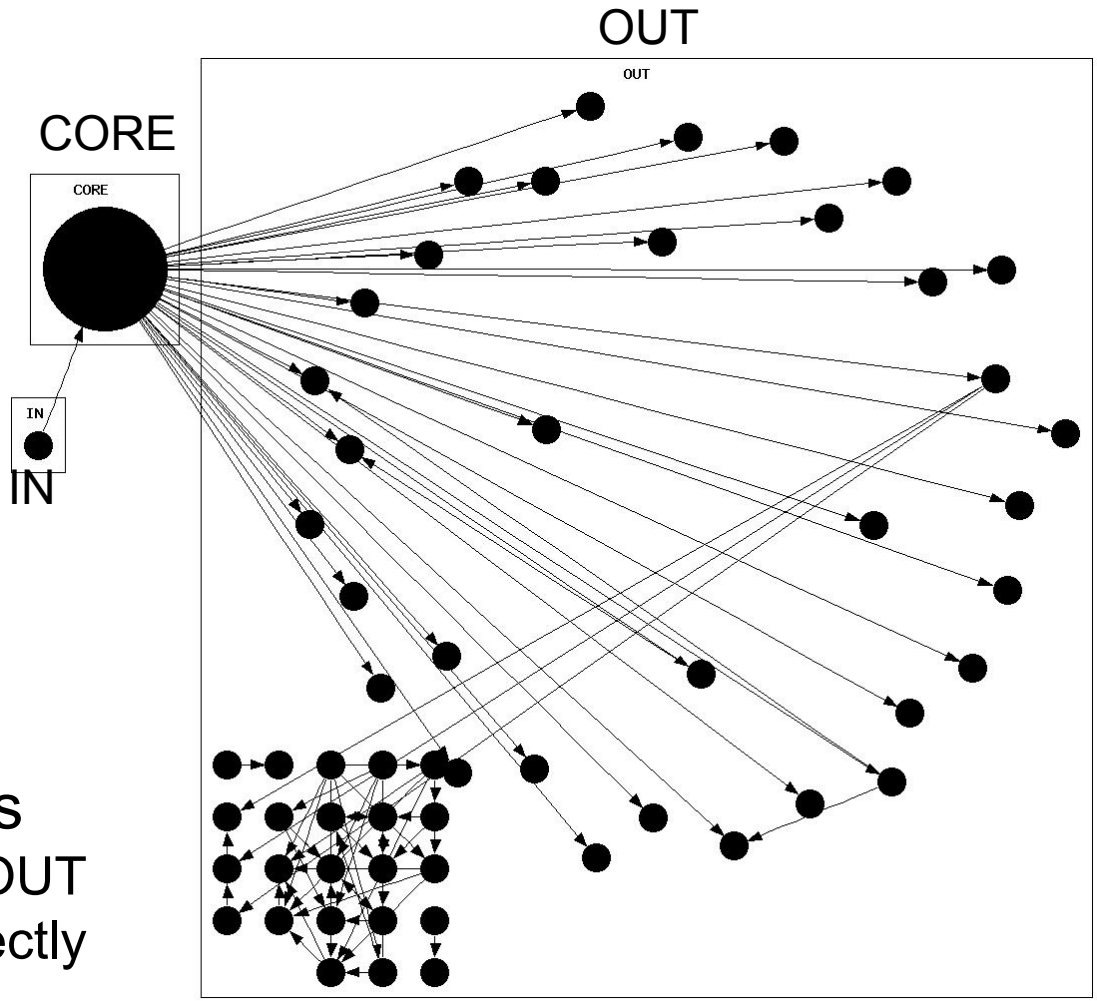
	spam	suspicious	non-spam
#sites	527	23	0
ratio (%)	95.8	4.2	0

Distribution of SCCs in the bow tie

- Bow-tie structure [Broder et al. 2000]



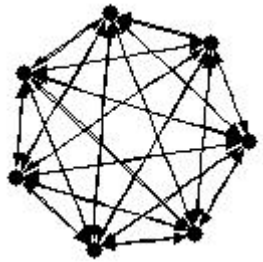
- Distribution of large SCCs
  - 450 / 552 (81%) SCCs in OUT
  - 385 / 450 (85%) SCCs directly connected to CORE
- CORE has many spam sites connecting to them



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(max. degree^4)
  - Large cliques are link farms (40 < n)
    - 26,931 maximal cliques, 8,346 sites (many duplicates)
    - 165 sample sites



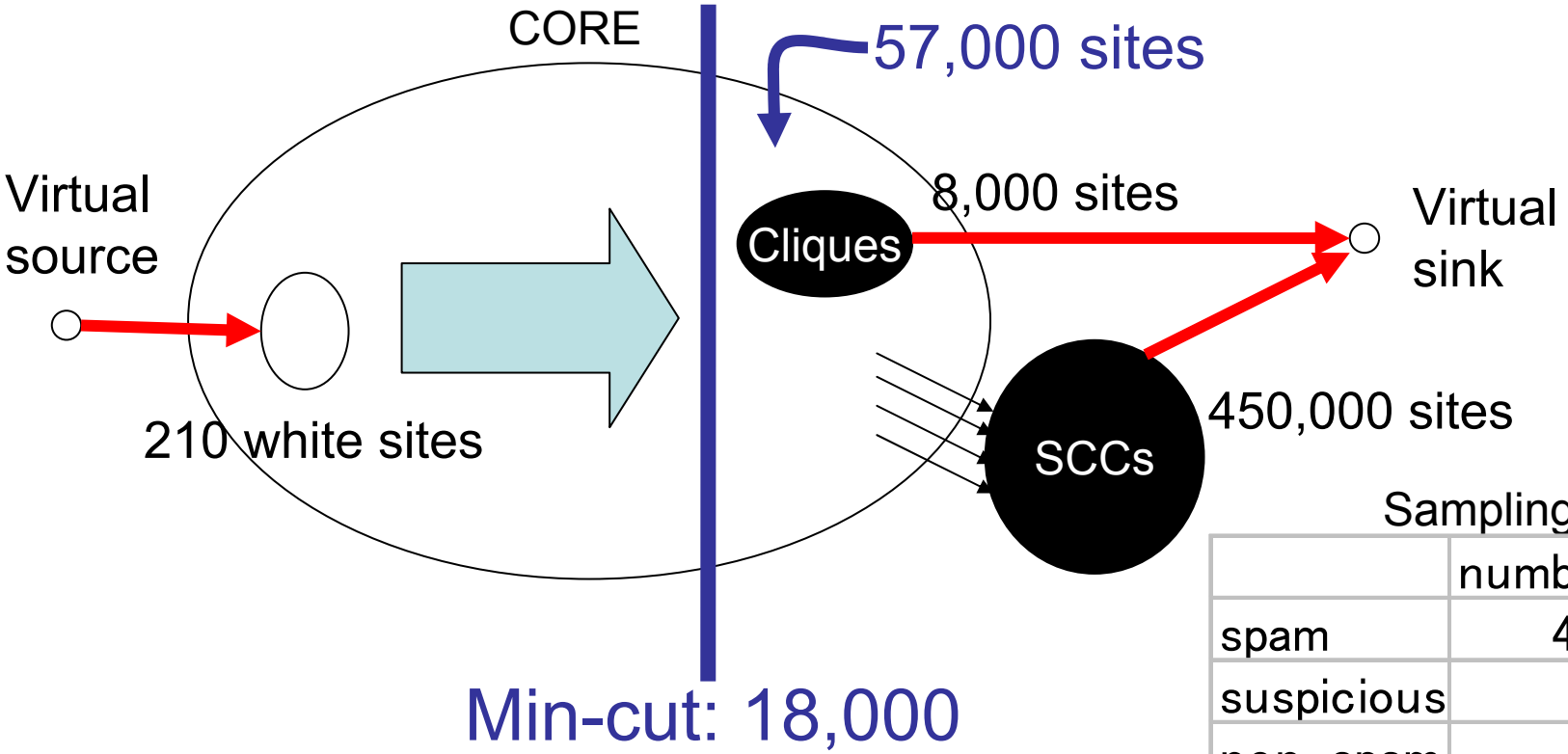
Sampling result

	spam	suspicious	non-spam
#sites	157	8	0
ratio (%)	95.2	4.8	0

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



Sampling result

	number	ratio (%)
spam	459	94.3
suspicious	27	5.5
non−spam	1	0.2

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