COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

NeighborWatcher: A Content-Agnostic Comment Spam Inference System

Jialong Zhang and Guofei Gu

Secu**r**e **C**ommunication and **C**omput**e**r **S**y**s**tems Lab

Department of Computer Science & Engineering

Texas A&M University

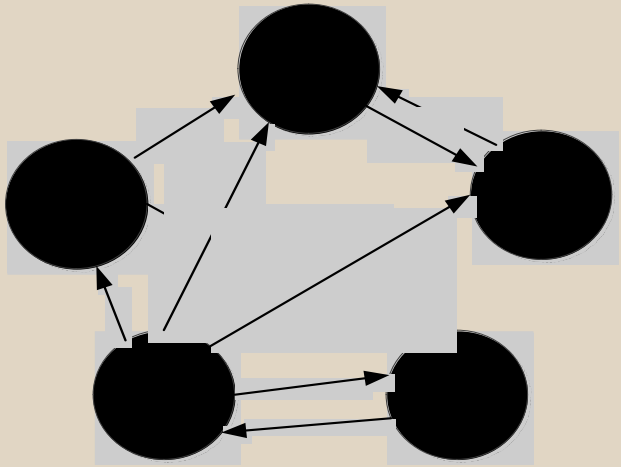
Spamdexing

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Buying **Viagra**, Lowest Price **Viagra** - Online Pill Store, Big Discounts
sagescholars.berkeley.edu/about-us/message-from-the-chancellor
cheap **viagra** caverta online **free shipping viagra** post ejaculation extra strength **viagra** from cadana quel est le mieux **viagra cialis** levitra apcalis over the counter ...

Uk **Viagra** Sales, Free **Viagra** Sample - Canadian Pharmacy ...
sagescholars.berkeley.edu/david
viagra cialis pack purchase fda approved **viagra** no prescription where to purchase **viagra** at a store. **viagra** para mulheres. **viagra** cheap and **free shipping** ...

Viagra Uk Cheap Purchase Buy, Buy Cheap **Viagra** Online - Pill ...
blogs.library.ucla.edu/special
Jul 12, 2011 – uk bristol **cialis viagra** ... compare the cost of **cialis** levitra **viagra** in new zealand **viagra** uk ... cheap **viagra** pay by mastercard **free shipping** ...



Comment Spamming

Comment Spamming

- **Comment spamming:** refers to the behavior of *automatically* and *massively* posting random/specific comments to benign third-party websites that allow user generated content
  - Blog
  - Forum
  - GuestBook

- **Benefits**
  - Improve search rank
  - Not to be easily blocked
  - Scalability
  - Low cost

Comment Spamming

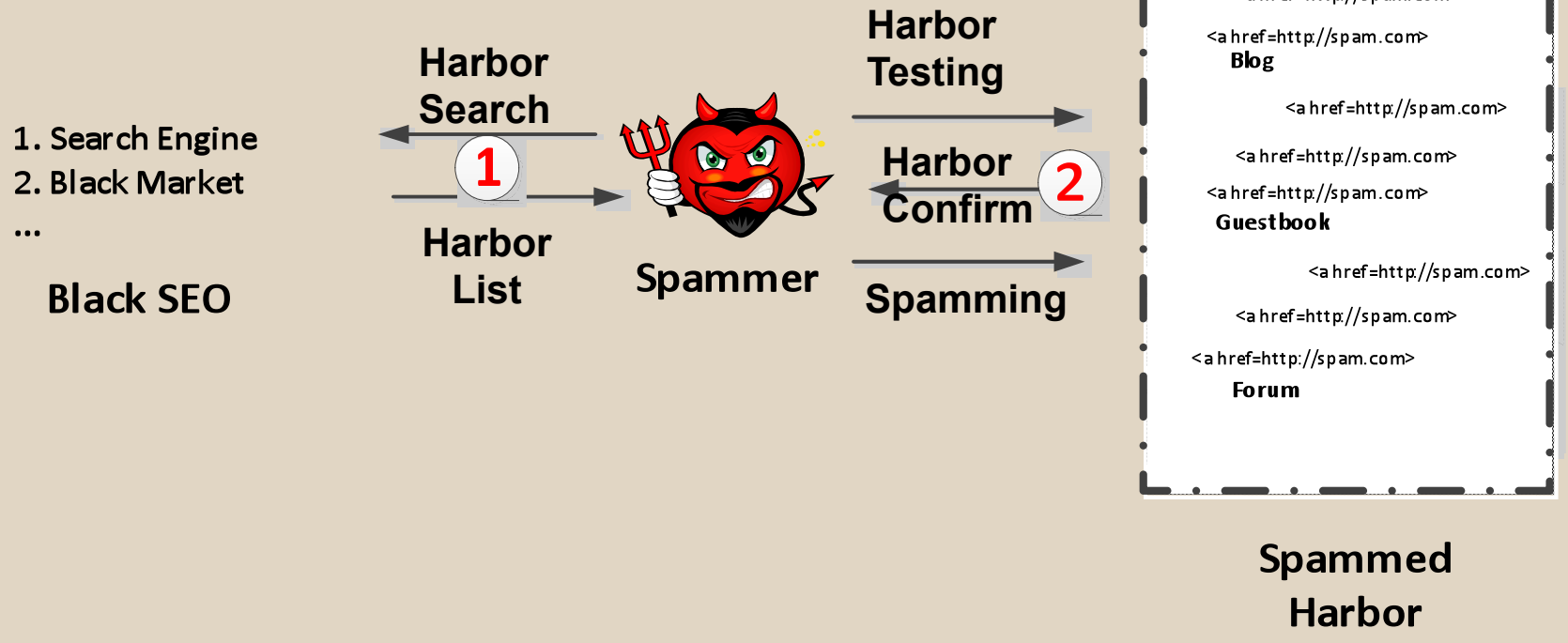
- **Current research**
  - Content-based detection (Y. Shin et al. [*INFOCOM'12*])
    - Easy to evade
  - Context-based detection (Y. Niu et al. [*NDSS'07*])
    - Low coverage
  - Honey blogs
    - Passive
    - Low coverage

Motivation

- **Comment spamming**
  - What are the properties of those benign websites (Harbors), which spammers usually spam on ?
  - What's the infrastructure of those harbors?
  - Can we exploit the infrastructure of harbors to help detect spam?

- **Assumption**
  - Each spammer has relative stable spamming harbors
  - Spammers intend to massively and automatically post the spam URL on their spamming harbors

Threat Model



Data Collection

- **Seeds**
  - 10,000 spam links from previous work (J. Zhang et al.[*RAID'12*])
- **Methods**
  - Search spam links in Google
    - Security websites that report search links as spam
    - Benign websites that link to search links
    - Spam harbors

Data Collection

Harbor

Spam

Data Collection

- **Seeds**
  - 10,000 spam links from previous work (J. Zhang et al.[*RAID'12*])
- **Methods**
  - Search spam links in Google
    - Security websites that report search links as spam
    - Benign websites that link to search links
    - Spam harbors
  - Extract search results with embedded hyperlinks tags in their content (e.g., [URL]…[/URL])

	Blog	Forum	GuestBook	Other	Total
# of search results	27,846	29,860	31,926	500,717	590,349
# of harbors (domain)	4,807	2,515	3,878	27,713	38,913
# of active harbors	4,685	2,185	3,419	25,642	35,931
# of postings	532,413	640,073	1,469,251	6,497,263	9,139,000

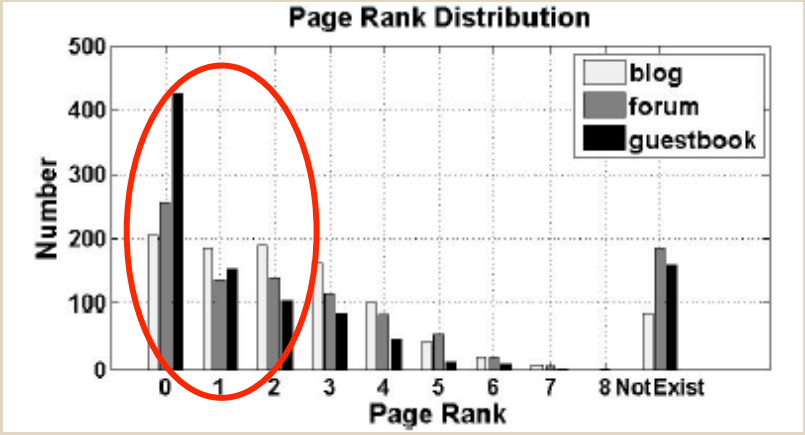
Comment Spamming Study

What are the properties of those benign websites (Harbors), which spammers usually spam on

- **Quality of harbors**
  - Intuition: the higher quality spam harbors have, the more effective comment spamming is
  - PageRank
  - Lifetime
  - Google Indexing Interval

Quality of Harbors

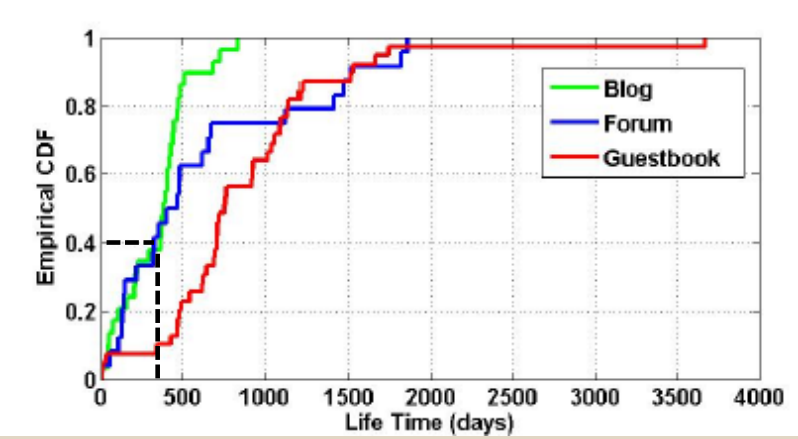
- **PageRank**
  - Randomly choose 1,000 spam harbors in each category



Spammers choose harbors regardless of their reputation to compensate the relatively poor quality of individuals

Quality of Harbors

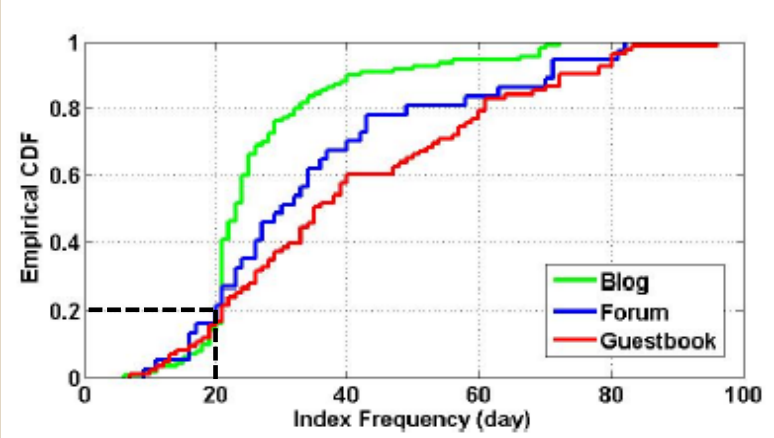
- **Life Time**
  - Definition: the time interval between the posting time of the first spam and the recent spam.
  - Randomly choose 100 harbor in each category



Spammers tend to explore some stable harbors which they can keep spamming on

Quality of Harbors

- **Google Indexing Interval:**
  - Definition: the time difference between two consecutive Google crawling time (Google cache) of same spam harbor
  - Randomly choose 100 harbor in each category



There exists a long time lag between spamming time and indexing time

What's the infrastructure of spam harbors?

- **Infrastructures of harbors**

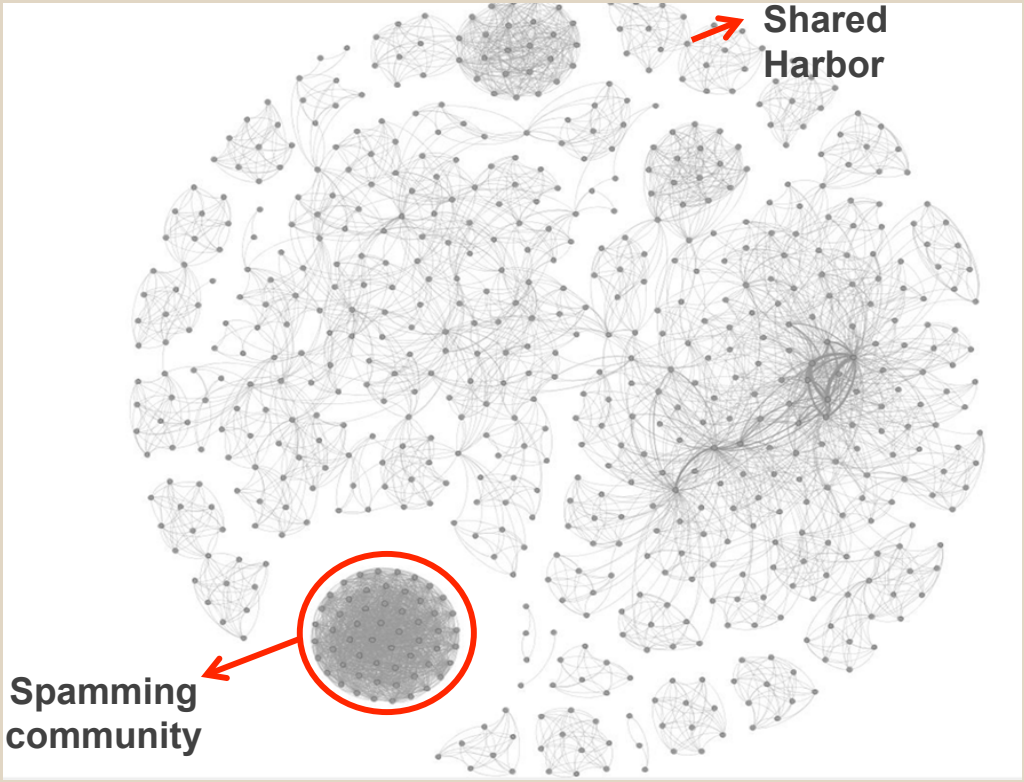
–Intuition: spammers build artificial relationships among spam harbors

–Relation graph

–How spammers use their infrastructure to distribute spam

Infrastructure of Harbors

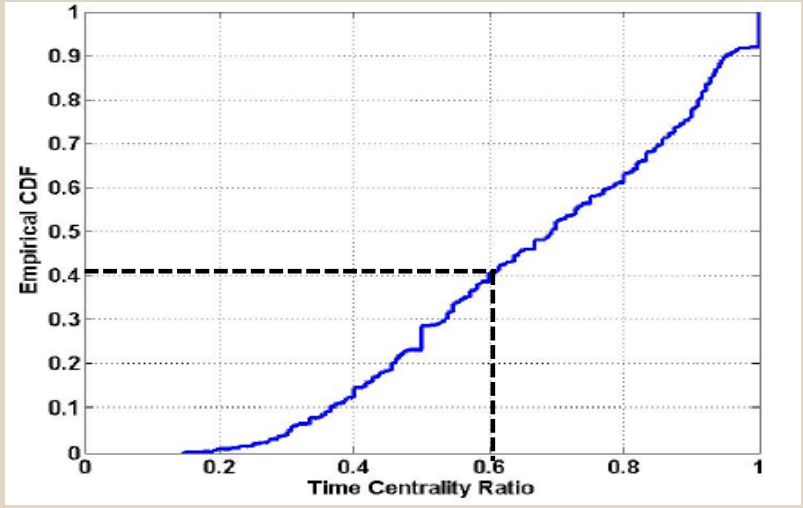
- **Relation Graph**



- Spammers build artificial relationships among those harbors
- Although different spammers may have different strategies to find their harbors, there exist some intersections among them

Infrastructure of Harbors

- **Distributing Spam**
  - Whether spam messages are distributed at the same time
  - Time Centrality Ratio: the maximal number of harbors that post the spam in the same month over the total number of harbors that post this spam



- Spammers tend to utilize their spam infrastructure in similar time

NeighborWatcher

Can we exploit the infrastructure of harbors to help detect comment spam?

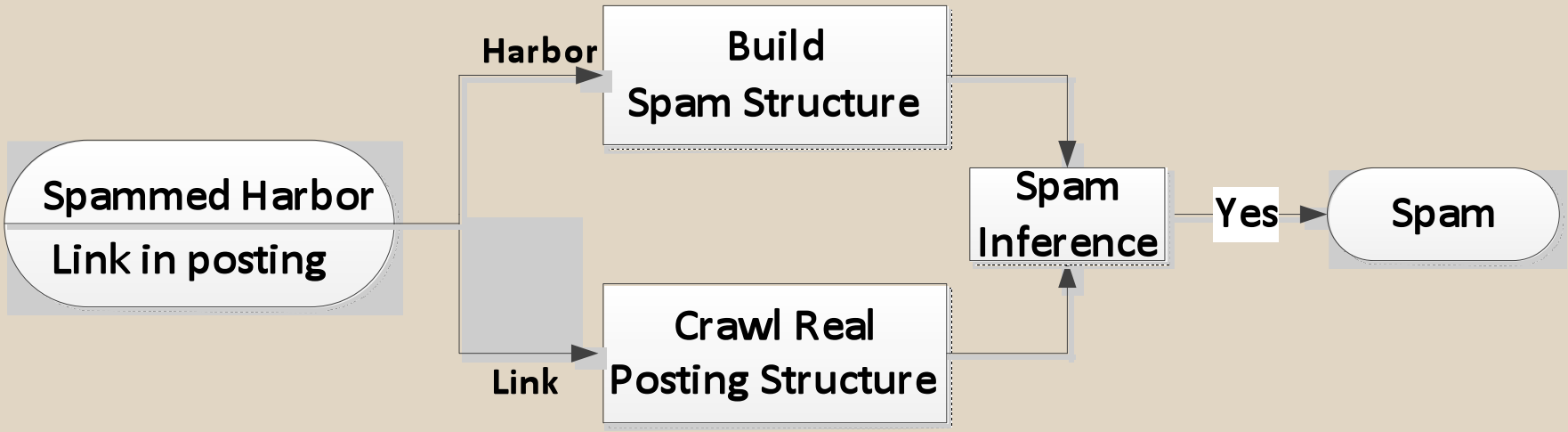
- **Detection of comment spam**

–Intuition: if a link is posted on a set of harbors that have a close relationship at a similar time, it has a high possibility to be spam

–NeighborWatcher

NeighborWatcher

- System Design

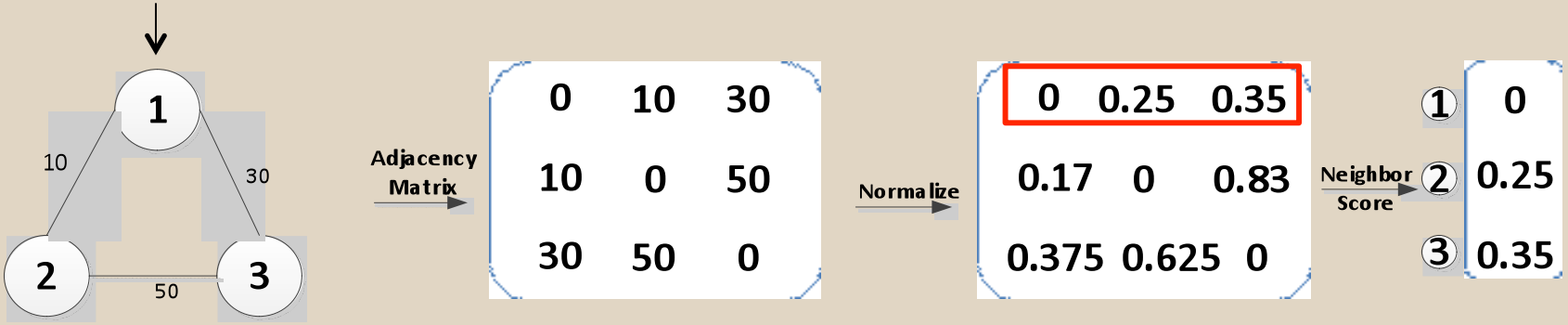


System Architecture

NeighborWatcher

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

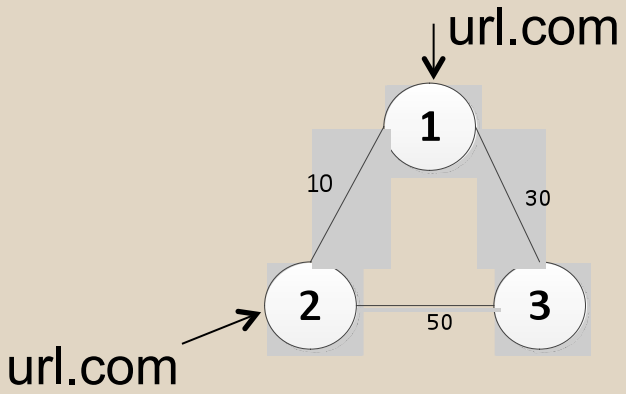
- Building Spamming Infrastructure



NeighborWatcher

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

- **Spamming Inference**

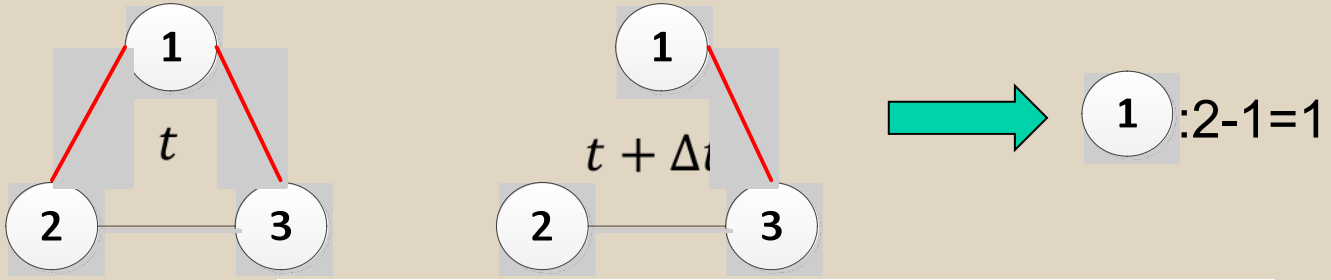


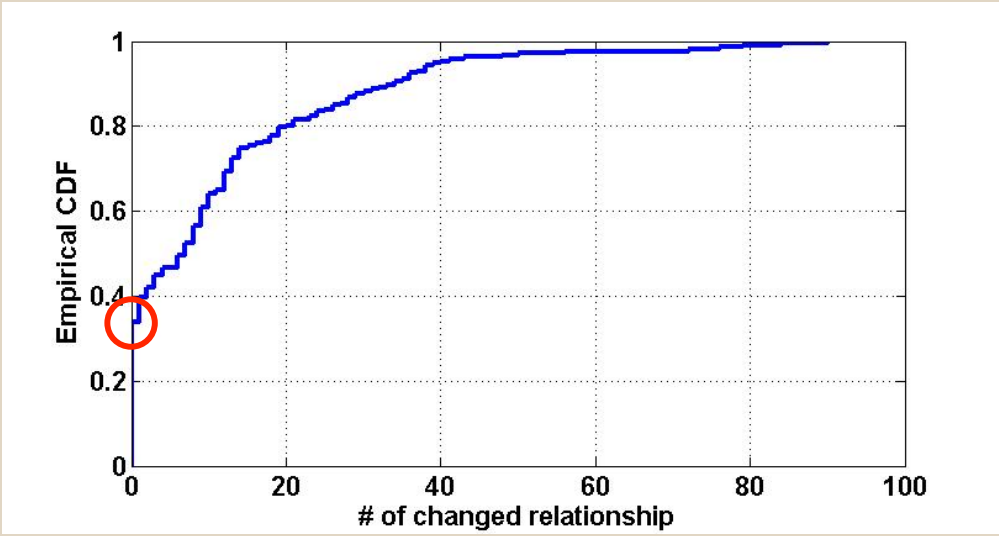
	Spamming structure	Real posting structure
1	0	1
2	0.25	1
3	0.35	0

→ Modified Cosine similarity

Evaluation

- **Stability of Spamming Structure**
  - Changed Relationship





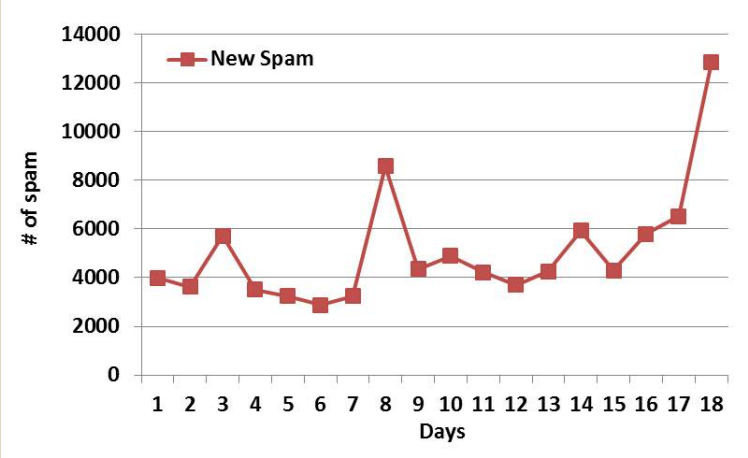
Evaluation

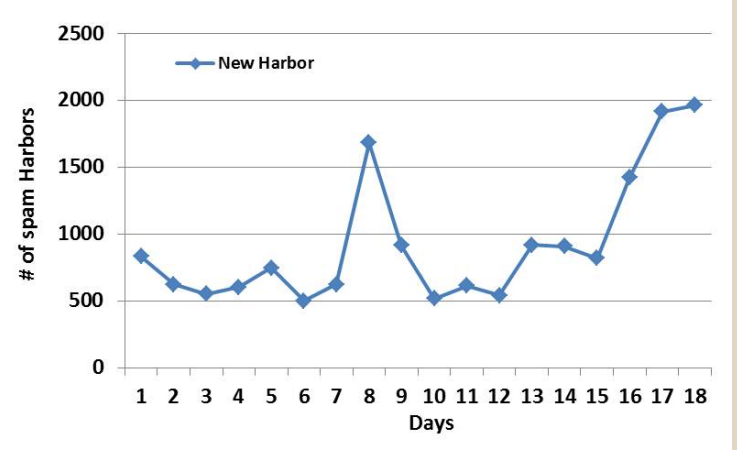
- **Effectiveness of Inference**
  - Ground Truth:
    - 500 verified spam
    - Top 20,000 domains from Alexa => 754 benign links
  - Hit Count: the number of correctly inferred spam
  - Hit Rate: the ratio of correctly inferred spam to the total number of inferred spam

Sim. Threshold	Hit Rate	Hit Count	False Positive
0.3	57.29%	432	322
0.4	75.93%	426	135
0.5	**97.14%**	**408**	**12**
0.6	97.8%	360	8

Evaluation

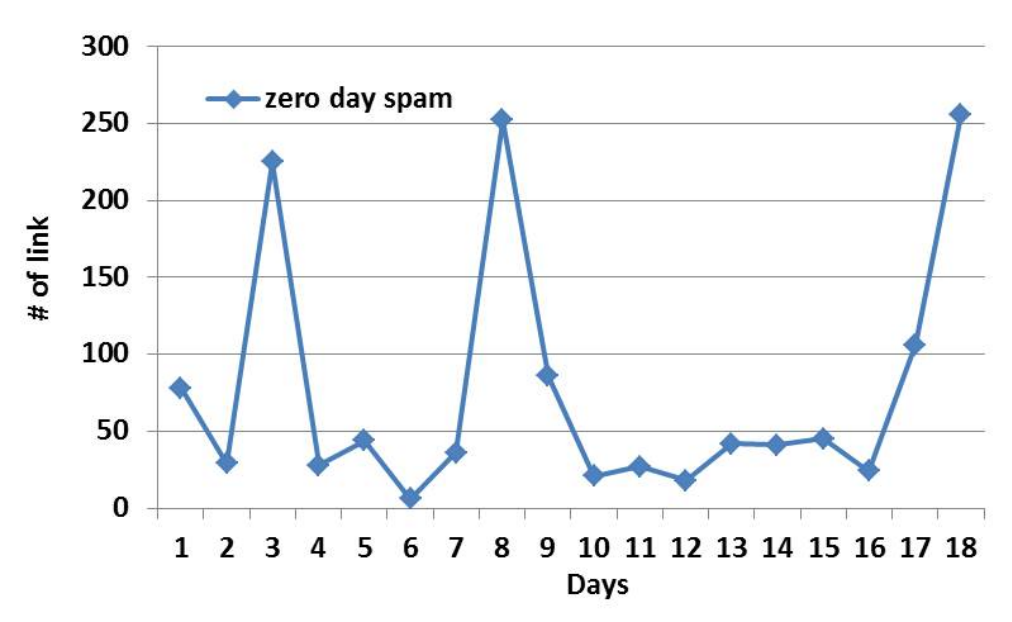
- **Constancy**
  - Whether our system can continue finding new spam over time?





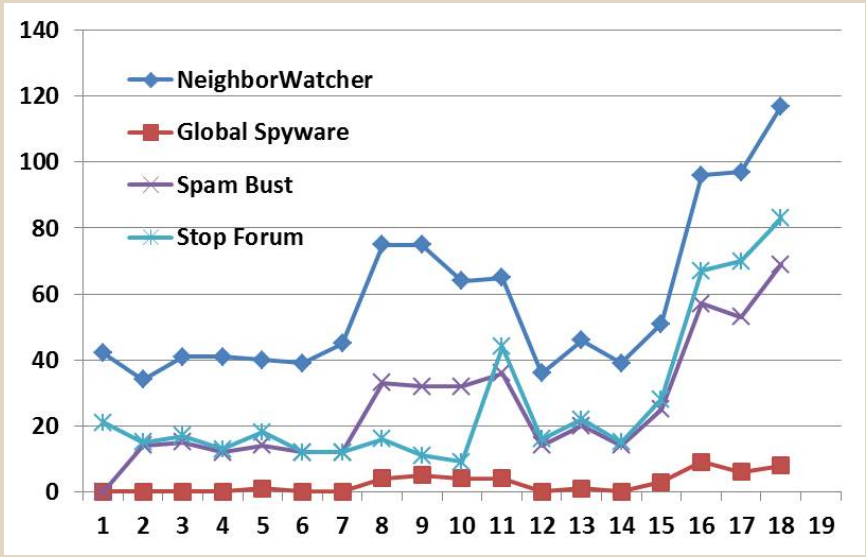
Applications

- **Early Warning**
  - Zero Day Spam: the spam that can not be searched out by Google at that time

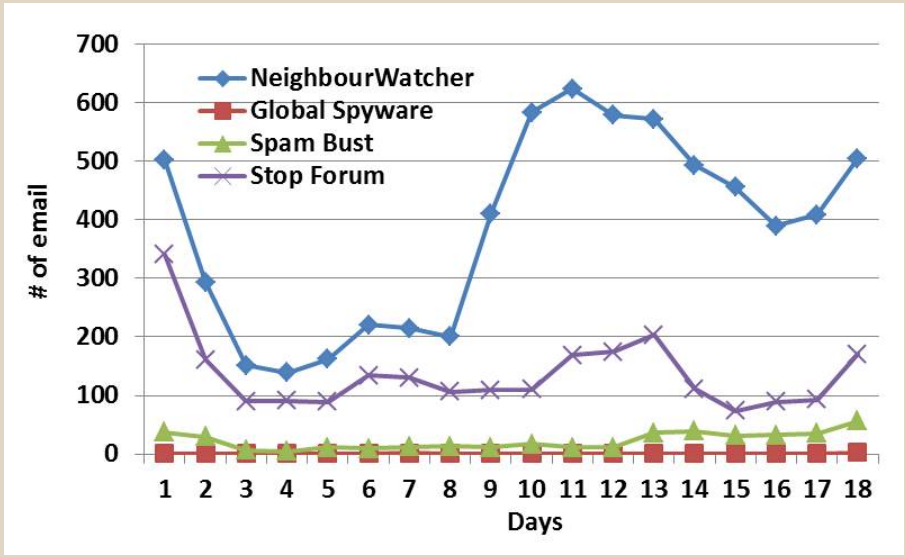


Applications

- **BlackLists**
  - Existing Blacklist
    - IPs
    - Emails
  - Whether our system can complement existing BlackLists



Daily IPs



Daily Emails

Summary

- We conduct a deep study on comment spam from a new perspective: spamming infrastructure
- We conclude that spammers prefer to keep utilizing their spam harbors for spamming
- We design a graph–based inference system to infer comment spam.

Q&A

COMPUTER SCIENCE & ENGINEERING
TEXAS A&M UNIVERSITY

Discussion

False negatives

- Only appeared on input harbor
- Crawl more harbors
- Need more time

False positives

- 5 are used for testing
- 7 are spammed on Alexa top 20,000 websites