NeighborWatcher: A Content-Agnostic Comment Spam Inference System

Jialong Zhang and Guofei Gu
Secure Communication and Computer Systems Lab
Department of Computer Science & Engineering
Texas A&M University
Spamdexing

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
• **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

1. Search Engine
2. Black Market
... 
Black SEO

Harbor Search

Harbor List

Spammer

Harbor Testing

Harbor Confirm

Spamming

Spammed Harbor

<a href=http://spam.com>Blog</a>
<a href=http://spam.com>Guestbook</a>
<a href=http://spam.com>Forum</a>
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

Mar 23, 2012 – a href="https://digilander.libero.it/redprofile/allgrannysex/index.html" pomnhubd (a href="" >aligrannysex (a href="%20" >..."
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])

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

![Empirical CDF of Life Time](image)

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
• 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
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
• System Design

System Architecture
NeighborWatcher

- Building Spamming Infrastructure
### Spamming Inference

<table>
<thead>
<tr>
<th>Spamming structure</th>
<th>Real posting structure</th>
<th>Modified Cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 0</td>
<td>1: 1</td>
<td></td>
</tr>
<tr>
<td>2: 0.25</td>
<td>2: 1</td>
<td></td>
</tr>
<tr>
<td>3: 0.35</td>
<td>3: 0</td>
<td></td>
</tr>
</tbody>
</table>

Diagram:

1. URL: url.com
2. Graph with nodes 1, 2, 3 connected by edges with weights 10, 50, 30, respectively.

Output: 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

<table>
<thead>
<tr>
<th>Sim. Threshold</th>
<th>Hit Rate</th>
<th>Hit Count</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>57.29%</td>
<td>432</td>
<td>322</td>
</tr>
<tr>
<td>0.4</td>
<td>75.93%</td>
<td>426</td>
<td>135</td>
</tr>
<tr>
<td><strong>0.5</strong></td>
<td><strong>97.14%</strong></td>
<td><strong>408</strong></td>
<td><strong>12</strong></td>
</tr>
<tr>
<td>0.6</td>
<td>97.8%</td>
<td>360</td>
<td>8</td>
</tr>
</tbody>
</table>
Evaluation

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

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

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

![Daily IPs and Daily Emails](chart.png)
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.
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