Spam recognition by methods independent from text content

Virus Bulletin 2006 Montréal

Ralf Iffert
ISS C-Force

Mark Usher
ISS C-Force
Conventional spam filters are ineffective

- Circumvented by random text
- Outsmarted by spams without any text
- RBLs fooled by changing IPs with high frequency
Introduction of two spam detection methods independent from text analysis

- Structure Analysis
  - Analysing the HTML structure of the email

- Flow Analysis
  - Analysing the flow of incoming emails
Basic Idea

- Remove all content from HTML part
- Calculate a hash on the remaining HTML structure
- Add hash to a database that is used for spam analysis
Viagra $69.95 (10 tablets)
Valium $105.45 (30 tablets)
Cialis $99.95 (10 tablets)

And many other [http://kew76.obosome.com](http://kew76.obosome.com)

Viagra $69.95 (10 tablets)
Valium $105.45 (30 tablets)
Cialis $99.95 (10 tablets)

And many other [http://oyh42.obosome.com](http://oyh42.obosome.com)
Extract from source code of sample 1

\[
\begin{align*}
V &= i \leq e \leq q \leq j \\
&= 69,95 (1 \text{ $s6$} \text{ WU abIets})
\end{align*}
\]

Extract from source code of sample 2

\[
\begin{align*}
V &= \text{sig://4A3DDB22F7C25943 (structhash)} \\
&= j \leq q \leq j \\
&= 69 (10 \text{ OR s})
\end{align*}
\]
Structure Analysis – Workflows

**Manufacturer’s side**

- Master Spam Database
- Client Spam Database

**Client’s side**

- Spam Analysis

### Experimental Results

- **Voluminous inflow of spam**
- **Detection ratio: 45.1%**
  - Exclusively: 1.6%
  - Used for the local spam analysis
  - Updated by the Master Spam Database
- **False positive ratio: 0.010%**

---

Delivered

Blocked
Flow Analysis
Basic Idea

Identifying “similar“ emails arriving within a small time frame

→ Detection of whole spam threads
→ Similarity of emails is determined by similarity measures
Flow Analysis

<table>
<thead>
<tr>
<th>Type</th>
<th>SimilaritySignature</th>
<th>Sender</th>
<th>TimeStamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sig://90445BE9CB2FCCAE</td>
<td>sig://E98DBF65E5BAD2554</td>
<td>2006-10-12 14:23</td>
</tr>
<tr>
<td>2</td>
<td>sig://1141F80CD68C0F8</td>
<td>sig://E98DBF65E5BAD2554</td>
<td>2006-10-12 14:23</td>
</tr>
<tr>
<td>3</td>
<td>sig://062BAE7C5F13FC22</td>
<td>sig://E98DBF65E5BAD2554</td>
<td>2006-10-12 14:23</td>
</tr>
<tr>
<td>4</td>
<td>sig://165CCEEEA36B809</td>
<td>sig://E98DBF65E5BAD2554</td>
<td>2006-10-12 14:23</td>
</tr>
</tbody>
</table>

Further measures are conceivable…

Flow Database

- Administration of similarity measures
- Set of image URLs
- Columns of the Flow Database
- Type of similarity measure
- Set of attachments
- Similarity signature
- Time when entry was generated
- Sender of the email

Buy viagra online.
From € 9.85 per pill.
Click on the image to order online!!

International
Legal RX medications

CIALIS  
as low as $5.67
VALIUM  
as low as $3.8
VIAGRA  
as low as $3.5
XANAX  
as low as $2.85
SOMA  
as low as $0.88
ULTRAM  
as low as $1.83

Cheapest Cialis ever!!

From: cheap@viagra.com 23 PM

http://enter.your.credit.com/card/number.jpg
http://and.loose.info/all/your/money.gif
### Flow Analysis – Workflows

#### Usage of the Flow Database

- **Conditioned by the two parameters:**
  - Threshold for similar emails
  - Time-frame that is monitored
- **Only contains information of emails received within the time-frame**

#### Experimental Results

- **Detection ratio:** 72.7%
- **False positive ratio:** ~0%

#### Precondition

- About 50,000 spams per day or more!

#### Optimizations

- Delay email delivery according to time-frame
Computing time depends on:

- $P_i$ Averaged time to extract the similarity data of the $i^{th}$ similarity measure
- $M_i$ Averaged hash calculation time on the $i^{th}$ similarity data
- $R$ Time for one database request
- $N$ Number of similarity measures

→ Computing time per mail: of the $i^{th}$ similarity measure

$$\sum_{i=1}^{N} \left( P_i + M_i + R \right)$$
Memory requirements Flow Database

Required memory depends on
- $M$ Maximal number of entries in Flow Database
- $S$ Size in bytes of each Flow Database entry

→ Required memory: $M*S$

$M$ depends on
- $N$ Number of similarity measures
- $E$ Throughput in emails per minute
- $X$ Time-frame in minutes used for the Flow Analysis

→ $M = N*E*X$

$S$ depends on
- $C$ Number of bytes consumed by a Flow Database entry
- $O$ Memory overhead

→ $S = C+O$

→ Required memory: $N*E*X*(C+O)$

= 1.28 MB

Example
Future Work

Further approaches

- Automated detection of random text
- Usage of visual features
- Image signatures invariant against random variations
Many thanks for your attention!

Q & A

Ralf Iffert
ISS C-Force
riffert@iss.net