Longtime Behavior of Harvesting Spam Bots

Oliver Hohlfeld

TU Berlin / DT Labs

Thomas Graf
Modas GmbH

Florin Ciucu
TU Berlin / DT Labs
Such a foolishness iss dot talk! I stay me py ashes to the winds of heaven. The other relics love and tenderness and fear for you. They tell years is to be the consequence, on the expiry of experience that plants brought from the forest.
Such a foolishness is not talk! I stay me py ash to the winds of heaven. The other relics love and tenderness and fear for you. They tell years is to be the consequence, on the expiry of experience that plants brought from the forest.

Why you?
Why you?

Scope: Address harvesting from public web sites

Image source: http://www.flickr.com/photos/twistermc/3382403844/ (CC BY-SA 2.0)
Approach

Our Infrastructure

- 9 Web Sites (1 US)
- Database
- SMTP Servers
Approach

Our Infrastructure

9 Web Sites (1 US)

Database

SMTP Servers

Address Harvester

Web Crawler

Addresses

HTTP

Money

Addresses

Spammer
Approach

Our Infrastructure

9 Web Sites (1 US)

Database

SMTP Servers

Address Harvester

Addresses

HTTP

Web Crawler

Money

Addresses

Spammer

Might pay botmaster to send

botnet
Approach

Our Infrastructure

- 9 Web Sites (1 US)
- Database
- SMTP Servers

> 120k
(0.5% of total)
Addresses

Address Harvester

- Web Crawler
- HTTP

> 3 Years

Spammer

- Might pay botmaster to send
- botnet

> 620k
Spam E-Mail

Money

Oliver Hohlfeld (TU Berlin / DT Labs)
Longtime Behavior of Harvesting Spam Bots
Host Properties

- How many harvesting hosts?
Host Properties

- How many harvesting hosts? $> 1k$
Host Properties

- How many harvesting hosts? $> 1k$
- Geolocation?
Host Properties

- How many harvesting hosts? \( > 1k \)
- Geolocation?

Figure: by requesting IPs
Host Properties

- How many harvesting hosts? $> 1k$
- Geolocation?

- 24 massive harvesting hosts in Romania ($\approx 10k$ page requests / day)

Figure: by requesting IPs

Figure: by spam volume
Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

![Graph showing IP distribution and spam volume by country.]

**Figure:** by requesting IPs

- 24 massive harvesting hosts in Romania (≈ 10k page requests / day)
- How are they connected?
Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

**Figure:** by requesting IPs

24 massive harvesting hosts in Romania (≈ 10k page requests / day)

- How are they connected?
  - 73% hosted in ADSL / cable networks

**Figure:** by spam volume
Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

24 massive harvesting hosts in Romania ($\approx 10k$ page requests / day)
- How are they connected?
  - 73% hosted in ADSL / cable networks
- Using Tor Anonymity Service?
Host Properties

- How many harvesting hosts? > 1k
- Geolocation?

Figure: by requesting IPs

- 24 massive harvesting hosts in Romania (≈ 10k page requests / day)
- How are they connected?
  - 73% hosted in ADSL / cable networks
- Using Tor Anonymity Service? No

Figure: by spam volume
Does blacklisting help?
Blocking

Does blacklisting help?

→ Yes (26% hosts blacklisted at access time)
Does blacklisting help?

→ Yes (26% hosts blacklisted at access time)

HTTP User Agent String Fingerprinting?
Blocking

- Does blacklisting help?
  - Yes (26% hosts blacklisted at access time)

- HTTP User Agent String Fingerprinting?
  - Variability might imply only few active parties
Blocking

- Does blacklisting help?
  - Yes (26% hosts blacklisted at access time)
- HTTP User Agent String Fingerprinting?
  - Variability might imply only few active parties
  - "Java/1.6.0_17" UA
    - 3% of harvesting hosts
    - 88% of harvesting page requests
    - 55% of total spam volume
    - 99.9% of Romanian harvesting bots
Does blacklisting help?

→ Yes (26% hosts blacklisted at access time)

HTTP User Agent String Fingerprinting?

Variability might imply only few active parties

“Java/1.6.0_17” UA

3% of harvesting hosts
88% of harvesting page requests
55% of total spam volume
99.9% of Romanian harvesting bots

→ Blocking certain user agent strings currently helps
Search engines exploited for malicious activities
Also used by harvesters?
Proxies Revisited: Search Engines

- Search engines exploited for malicious activities
- Also used by harvesters?
Proxies Revisited: Search Engines

Our Infrastructure

Search Engine

Address Harvester

9 Web Sites (1 US)
Proxies Revisited: Search Engines

Our Infrastructure

- 9 Web Sites (1 US)

Search Engine

- Web Crawler

Address Harvester

- ECrawl, ...

Addresses

HTTP
Proxies Revisited: Search Engines

Our Infrastructure

9 Web Sites (1 US)

Addresses
HTTP

Search Engine

Web Crawler

Address Harvester

ECrawl v2.63: “Access to the Google cache (VERY fast harvesting)”

Fast Email Harvester 1.2: “collector supports all major search engines, such as Google, Yahoo, MSN”
Proxies Revisited: Search Engines

Our Infrastructure

9 Web Sites (1 US)

Addresses

HTTP

Search Engine

Web Crawler

Address Harvester

ECrawl, ...

0.5% of addresses spammed
0.2% of total spam
Proxies Revisited: Search Engines

Our Infrastructure

9 Web Sites (1 US)

Addresses
HTTP

Search Engine
Web Crawler

Address Harvester
ECrawl, ...

0.5% of addresses spammed
0.2% of total spam

→ You don’t want to block Google!
50% spammed < 4 days (general), 11 days (search engines)
50% spammed $< 4$ days (general), $11$ days (search engines)

Usage period:
- $< 1$ second: $11\%$ (general), $16\%$ (search engines)
- $< 1$ day: $17\%$ (general), $40\%$ (search engines)
- $< 1$ week: $78\%$ (general), $53\%$ (search engines)
<table>
<thead>
<tr>
<th>General data set</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Webmasters Dilemma: Address Presentation

<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
</tr>
</tbody>
</table>

- **MTO** User friendly mailto link: `mailto:john.doe@imc.conf`
<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
<th>TXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
<td>31%</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
<td>38%</td>
</tr>
</tbody>
</table>

- **MTO** User friendly mailto link: `mailto:john.doe@imc.conf`
- **TXT**: plain text `john.doe@imc.conf`
### Webmasters Dilemma: Address Presentation

<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
<th>TXT</th>
<th>OBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
<td>31%</td>
<td>7%</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
<td>38%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

- **MTO** User friendly mailto link: `mailto:john.doe@imc.conf`
- **TXT**: plain text `john.doe@imc.conf`
### Webmasters Dilemma: Address Presentation

<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
<th>TXT</th>
<th>OBF</th>
<th>JS</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
<td>31%</td>
<td>7%</td>
<td>0</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
<td>38%</td>
<td>0.6%</td>
<td>0</td>
</tr>
</tbody>
</table>

- **MTO**: User friendly mailto link: `mailto:john.doe@imc.conf`
- **TXT**: plain text `john.doe@imc.conf`
- **JS**: Javascript code
### Webmasters Dilemma: Address Presentation

<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
<th>TXT</th>
<th>OBF</th>
<th>JS</th>
<th>FRM</th>
<th>CMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
<td>31%</td>
<td>7%</td>
<td>0</td>
<td>2.5%</td>
<td>19%</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
<td>38%</td>
<td>0.6%</td>
<td>0</td>
<td>0.4%</td>
<td>0%</td>
</tr>
</tbody>
</table>

- **MTO**: User friendly mailto link: `mailto:john.doe@imc.conf`
- **TXT**: plain text `john.doe@imc.conf`
- **JS**: Javascript code
- **FRM**: HTML form
- **CMT**: HTML comment
<table>
<thead>
<tr>
<th></th>
<th>MTO</th>
<th>TXT</th>
<th>OBF</th>
<th>JS</th>
<th>FRM</th>
<th>CMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>General data set</td>
<td>40.5%</td>
<td>31%</td>
<td>7%</td>
<td>0%</td>
<td>2.5%</td>
<td>19%</td>
</tr>
<tr>
<td>Search engines</td>
<td>61%</td>
<td>38%</td>
<td>0.6%</td>
<td>0%</td>
<td>0.4%</td>
<td>0%</td>
</tr>
</tbody>
</table>

- **MTO** User friendly mailto link: `mailto:john.doe@imc.conf`
- **TXT**: plain text `john.doe@imc.conf`
- **JS**: Javascript code
- **FRM**: HTML form
- **CMT**: HTML comment

→ Simple obfuscation methods (OBF, JS) still suffice
Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes
Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes

Future Work

- Campain analysis
- How many harvesting parties exist?

We thank all the anonymous spammers and harvesters for making this study possible.
Conclusions

- Obfuscate your e-mail addresses!
- User agent filtering can help
- Search engines used as proxies
- Possibly only few active harvesters operating at different spam volumes

Future Work

- Campaign analysis
- How many harvesting parties exist?

We thank all the anonymous spammers and harvesters for making this study possible.

Need more stats? Download the data:
http://ohohlfeld.com/harvesting.html