Extending Black Domain Name List by Using Co-occurrence Relation between DNS Queries

Kazumichi Sato, Keisuke Ishibashi, Tsuyoshi Toyono, and Nobuhisa Miyake
Outline

• Motivation
  – Method for detecting botnets by using blacklist
  – Coverage of blacklist

• Proposed Method
  – Extend blacklist by using co-occurrence relation
  – Problems of naively using co-occurrence relation
  – Eliminating popular domain names and heavy user effect

• Experimental Results

• Conclusion
Motivation

• Botnet threats increasing
  – Launching DDoS attacks
  – Sending spam e-mail
  – Stealing personal information
  – Infecting other hosts

• Finding infected hosts and stopping malicious activities is necessary
Black Domain Name List

• Match black domain names with DNS queries to detect infected hosts
  – Bot sends DNS query to resolve domain name of C&C server
  – Black domain name list created by capturing and analyzing bots

• Block connections from infected hosts to C&C servers to stop malicious activities

1. Resolving black domain names

2. Connecting C&C server and receiving commands

3. Launching attacks

- DDoS attack
- Spam e-mail
- Infecting host

DNS server

C&C server

evil.com
(black domain name)

Infected hosts

Victim hosts
Problem of Black Domain Name List

- Blacklist does not cover all black domain names
  - Numerous new bots are observed every day, thus we cannot capture all bots
  - Some bots resolves many different black domain name, thus it is hard to maintain blacklist (e.g., Conficker worm)
Objective of Study

- Extending blacklist
  - Find unknown black domain names
    - Stop malicious activities by blocking connections from infected hosts to C&C servers
  - Using extended blacklist, find unknown infected hosts
    - Alert infected hosts to removing bot
Assumption

• One bot resolves several black domain names
  – For redundancy of C&C servers

Assumption

If two domain names are resolved by the same host frequently and one is black, the other is also black.
Approach Overview

1. Find hosts that resolve known black domain names
2. Find other domain names resolved by infected hosts
3. Add to blacklist
4. Find hosts that resolve unknown black domain names

Known black
- evil.com
- unknown.evil.com

Unknown black
- unknown.evil.com

Infected Hosts
- evil.com
- unknown.evil.com

Unknown Infected Hosts
- blacklist
Naive Scoring Method

• Our assumption
  – If two domain names are resolved by the same host frequently and one is black, the other is also black.

Focus on Co-occurrence relation

• Scoring method by using co-occurrence relation

$$C(d_1, d_2) = \frac{\text{# hosts that resolve } d_1 \text{ and } d_2}{\text{# hosts that resolve } d_1 \text{ or } d_2}$$

Co-occurrence rate

Total co-occurrence rate with black domain names

$$S(d) = \sum_{d_m \in \text{blacklist}} C(d_m, d)$$

If score is high, assume $d$ is black
Problem of Popular Domain Name

Many non-infected hosts are detected as infected

Popular domain name co-occurs with black frequently
Weight of Number of Non-infected Hosts

• Focus on number of non-infected hosts that resolve a domain name
  – Popular domain names are resolved by both infected and non-infected hosts
  – Black domain names are resolved by only infected hosts
• Define weight of number of non-infected hosts

\[
W(d) = \frac{\# \text{ infected that resolve } d}{\# \text{ infected that resolve } d + \# \text{ non-infected that resolve } d}
\]

If \(d\) is popular, \(W(d)\) is relatively small

If \(d\) is popular, value is relatively large
Problem of Infected Heavy User

Many domain names co-occur with black

Known blacklist

evil.com

Many non-infected hosts detected as infected

Non-infected Hosts

Infected Heavy User
Weight of Number of Queries

• Focus on number of domain names resolved by infected hosts
  – Add weight of number of queries to naive co-occurrence rate

• Weighted co-occurrence rate

\[
C'(d_1, d_2) = \frac{\sum_{h \in \text{hosts that resolve } d_1 \text{ and } d_2} 1/\# \text{ domain names resolved by } h}{\# \text{ hosts that resolve } d_1 \text{ or } d_2}
\]

If host is heavy user, value is small and C’ increases little
Proposed Scoring Method

Eliminate influence of infected heavy user

Even if $d$ is resolved by infected heavy user, $C'$ increases little

$$S'_w (d) = \sum_{d_m \in \text{blacklist}} C'(d_m, d) \times W(d)$$

Eliminate influence of popular domain name

If $d$ is popular, $W(d)$ is small
Evaluation

• Validated our assumption by cross validation
  
  ![Diagram showing Blacklist connected to Validation List and Learning list]

  Validation List (Assume unknown black)

  Learning list (Assume known black)

• Validated high-scored domain names
  
  – Applied proposed method to all domain names in known blacklist and classified 100 high-scored domain names as black, legitimate, or unclear

• Validated effectiveness of extended blacklist
  
  – Found hosts that resolved domain names in extended blacklist
Dataset

• DNS traffic data
  – Captured during one hour in Feb. 2009

• Blacklist
  – Created by using honeypot during same period
  – Blacklist has about 270 domain names
Cross Validation Results

Proposed Scoring Method

Naive Scoring Method
Classification Results

- Domain names for top 100 scores consisted of
  - 39% black, 4% legitimate, and 56% unclear
- Domain names for top 20 scores
  - 70% black
  - No legitimate domain names included

<table>
<thead>
<tr>
<th>Score</th>
<th>Domain Name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.571</td>
<td>spy.nerashti.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.571</td>
<td>bla.bihsecurity.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.571</td>
<td>aaaaaaaaaaaaaaaa.locop.net</td>
<td>Black</td>
</tr>
<tr>
<td>0.500</td>
<td>icq-msg.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.319</td>
<td>mail.tiktikz.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.300</td>
<td>x.zwned.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.300</td>
<td>evolutiontmz.sytes.net</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.300</td>
<td>dcom.anxau.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.292</td>
<td>usa.lookin.at</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.292</td>
<td>rewt.buyacaddi.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.250</td>
<td>unkn0wn</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.250</td>
<td>google-analitucs.com/loader/</td>
<td>Black</td>
</tr>
<tr>
<td>0.222</td>
<td>netspace.err0r.info</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.203</td>
<td>win32.kernelupdate.info</td>
<td>Black</td>
</tr>
<tr>
<td>0.203</td>
<td>free.systemupdates.biz</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.200</td>
<td>zjjdte.3322.org</td>
<td>Black</td>
</tr>
<tr>
<td>0.200</td>
<td>ykln.3322.org</td>
<td>Unclear</td>
</tr>
<tr>
<td>0.200</td>
<td>dr27.mcboo.com</td>
<td>Black</td>
</tr>
<tr>
<td>0.189</td>
<td>china.alwaysproxy.info</td>
<td>Black</td>
</tr>
<tr>
<td>0.167</td>
<td>home.najd.us</td>
<td>Black</td>
</tr>
</tbody>
</table>
Details of Unclear Domain Names

• Some unclear domain names are suspicious
  – Domain name whose subdomain differs from known black domain name
    • ykln.3322.org (zjjdtc.3322.org is known black)

  – Domain name with format “<black>.<legitimate>”
    • www.h7smcnrwlsdn34fgv.info.<legitimate>

  – Domain name for DNSBL lookups
    • <IP address>.zen.spamhaus.org
Finding Unknown Infected Hosts

• Rate of increase of number of unknown infected hosts is only 3%
  – Insufficient rate
  – Need to improve proposed method
Conclusion

• Proposed scoring method for finding unknown black domain names

• Found unknown black domain names and extended blacklist
  – Stop malicious activities by using extended blacklist more effectively

• Cannot find unknown infected hosts sufficiently
  – Improve method for finding unknown infected hosts as future work
Thank You