Spamming Botnets: Signatures and Characteristics

Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, Ivan Osipkov

Presented by Hongyu Gao
Mar. 04 2009
Motivation

Introduction

Design of AutoRE

Experimental Results

Spamming Botnet Characteristics

My Comments
Motivation

✧ Botnets have been widely used for sending spam emails at a large scale.
  ✧ Detecting and blacklisting individual bots is difficult.
  ✧ Little effort has been devoted to understanding the aggregate behaviors of botnets.
Introduction

- **Botnet**
  - A group of compromised host computers (bots)
  - Controlled by a small number of commander hosts (bot masters)
Introduction, cont’d

- High level idea
  - Use email dataset from a large email service provider (MSN Hotmail)
  - Focus on URLs embedded in email content
  - Derive signatures for spam based on URLs
  - Detect spam using signatures
AutoRe: Signature Based Botnet Identification

- A completely automatic tool
- Take as input a group of emails
- Produce a set of spam URL signatures and a list of botnet host IP addresses

Three modules:
- URL preprocessor
- Group selector
- RegEx generator
AutoRe: Signature Based Botnet Identification, Cont’d
URL Pre-processing

- Extract URL string, source server IP address and email sending time
- Partition URLs into groups based on their Web domains

<table>
<thead>
<tr>
<th>Time</th>
<th>URLs</th>
<th>Source ASes</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://www.lympos.com/n/?167&amp;brokenacclaim">http://www.lympos.com/n/?167&amp;brokenacclaim</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://www.lympos.com/n/?167&amp;acceptoraudience">http://www.lympos.com/n/?167&amp;acceptoraudience</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://shgeep.info/tota/indexx.html?ikjija.cvqxjby,hvx">http://shgeep.info/tota/indexx.html?ikjija.cvqxjby,hvx</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://shgeep.info/tota/indexx.html?ivvx_ceh.cvqxjby,hvx">http://shgeep.info/tota/indexx.html?ivvx_ceh.cvqxjby,hvx</a></td>
</tr>
</tbody>
</table>

Figure 2: Examples of polymorphic URLs.
URL Group Selection

- Assume the bursty property of botnet email traffic
- Construct n time window
- $S_i(k)$ is defined as the total number of IP addresses that sent at least one URL in group i in window k
- URL groups with sharp spikes are higher ranked
Signature Tree Construction

- The root node is set to the domain name
- Start with the most bursty and distributed substring
- Incrementally expand the signature tree
- Until no eligible substring remains
- The path from root to leaf defines a keyword-based signature
Signature Tree Construction, Cont’d

Figure 5: Example input URLs and the keyword-based signature tree constructed by AutoRE.
Regular Expression Generation

✧ The detailing process
  ✧ Given the keyword-based signatures, apply a set of predefined rules to generate regular expressions for the substring between keywords.

✧ The generalization process
  ✧ Takes the generated regular expressions and further groups them.
Regular Expression Generation, Cont’d

Figure 6: Generalization: Merging domain-specific regular expressions into domain-agnostic regular expressions.
Evaluation

- Emails were sampled from Nov. 2007, Jun. 2007 and Jul. 2007 (sampling rate 1:25000)

<table>
<thead>
<tr>
<th>Month</th>
<th>Nov 2006</th>
<th></th>
<th>June 2007</th>
<th></th>
<th>July 2007</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CU</td>
<td>RE</td>
<td>CU</td>
<td>RE</td>
<td>CU</td>
<td>RE</td>
<td></td>
</tr>
<tr>
<td>Num. of spam campaigns</td>
<td>1,229</td>
<td>519</td>
<td>1835</td>
<td>591</td>
<td>2826</td>
<td>721</td>
<td>7,721</td>
</tr>
<tr>
<td>Num. of ASes</td>
<td>3,176</td>
<td>1,398</td>
<td>4,495</td>
<td>1,906</td>
<td>4,141</td>
<td>1,841</td>
<td>5,916</td>
</tr>
<tr>
<td>Num. of botnet IPs</td>
<td>88,243</td>
<td>23,316</td>
<td>113,794</td>
<td>19,798</td>
<td>85,036</td>
<td>29,463</td>
<td>340,050</td>
</tr>
<tr>
<td>Num. of spam emails</td>
<td>118,613</td>
<td>26,897</td>
<td>208,048</td>
<td>26,637</td>
<td>159,494</td>
<td>40,777</td>
<td>580,466</td>
</tr>
<tr>
<td>Total botnet IPs</td>
<td>100,293</td>
<td></td>
<td>131,234</td>
<td></td>
<td>113,294</td>
<td></td>
<td>340,050</td>
</tr>
</tbody>
</table>

Table 1: Some statistics pertaining to the botnets identified by AutoRE.
Evaluation, Cont’d

- Low false positive rate
Evaluation, Cont’d

- Domain-agnostic generation improves the detection rate without affecting false positive rate.
For most spam campaigns, 90% of the destination Web pages are at least 75% similar.
Evaluation, Cont’d

- Pages from different campaigns are different
Spamming Botnet Characteristics

- Botnet IP Addresses are distributed and dynamic
For each campaign, the emails are sent almost simultaneously.
It is uncommon for different spam campaigns to overlap
My comments

- If the URLs are presented in image, this tool will be likely to miss them.
- This tool focuses on “bursty” and “distributed” characteristics of spamming botnets. However, if a botnet is not sending spam in a “bursty” or “distributed” way, e.g. when the botnet is small or it keeps sending spam in a long period of time, it is likely to evade the detection.
The authors assume at first the “bursty” and “distributed” nature of spamming botnets. Based on the assumption, they design a tool to detect botnets that behave in a “bursty” and “distributed” way. At last they use the detection result to prove that spamming botnets are “bursty” and “distributed”.

The assumption can not be confirmed in this way.