Mozzie: a normalization environment for malware execution

Supervisors

prof. Antonio Lioy
ing. Andrea Atzeni

Candidate

Mariano Graziano

November 7, 2011
Mozzie - The beginning

- **Aim:**
  - Learn the protocols to control the network behaviour.

- **Why? (Motivations)**
  - The network context problem.
  - The network behaviour during the analysis is only logged and not controlled (Sandbox).
  - The problem of the repeatability of the analysis.
  - The problem of malware execution $\rightarrow$ Long-term containment system.
Introduction

- The term malware is generic (MALicious SoftWARE).
- Trojan horses, worms, virus, backdoor, scareware, rootkit.
- Virus detection is undecidable:
  - In general, detection of a virus is shown to be undecidable both by a-priori and runtime analysis, and without detection, cure is likely to be difficult or impossible. (Cohen, 1984).
- This field of computer security is based on approximations.
Introduction

- In 2011 Cohen’s theorem is still true but the landscape is completely different:
  - Internet has 1.97 billion users
  - In 2010 286 million unique variants of malware. (Symantec, 2010).
  - The average cost per incident of a data breach in the United States is $7.2 million.
  - Malware is used in targeted attacks (Operation Aurora, Stuxnet, Shady Rat).

- Malware is a real market, it has its trends → cybercrime
Cybercrime


Mariano Graziano - Mozzie: a normalization environment for malware execution
Internet

- The vector of these attacks is Internet.
- Internet is composed of a set of protocols:
  - A protocol is a set of rules for a communication.
  - Malware use Internet protocols and sometimes they have their modified versions.
- Malware do not damage the victim’s computer - No more vandalism.
- The victim’s computer is a valuable resource.
- Botnet → big network of compromised computers.
Botnet I

http://www.f-secure.com/en/web/labs_global/articles/about_botnets

Mariano Graziano - Mozzie: a normalization environment for malware execution
Botnet II

http://www.usenix.org/event/hotbots07/tech/full_papers/wang/wang_html/

3 Mariano Graziano - Mozzie: a normalization environment for malware execution
Analysis

- The malware analysis can be:
  - Static → manual and error prone (Reverse Engineering approach).
  - Dynamic → automated and faster (Based on the concept of sandbox).

- The second approach is very used to figure out the malware characteristics.

- During the dynamic analysis lack of attention in network behaviour.
Learning by doing

▶ **ScriptGen:**
  ▶ Developed by Corrado Leita (Researcher at Symantec).
  ▶ It is a set of protocol learning techniques.
  ▶ It aims at rebuilding portions of a protocol finite state machine through the observation of samples of network interaction between a client and a server implementing such protocol.
  ▶ No assumption is made on the protocol structure, and no a priori knowledge is assumed on the protocol semantics.

▶ **Finite state machine (FSM):**
  ▶ It is a tree.
  ▶ The vertices contain the server’s answer.
  ▶ The edges contain the client’s request.
Table 4. Anonymized distribution of Class A-subnets for attack cluster 17718

<table>
<thead>
<tr>
<th>Subnet of Origin</th>
<th>Nr of Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.x.x.x</td>
<td>451</td>
</tr>
<tr>
<td>B.x.x.x</td>
<td>193</td>
</tr>
<tr>
<td>C.x.x.x</td>
<td>168</td>
</tr>
<tr>
<td>D.x.x.x</td>
<td>160</td>
</tr>
<tr>
<td>E.x.x.x</td>
<td>159</td>
</tr>
<tr>
<td>F.x.x.x</td>
<td>123</td>
</tr>
<tr>
<td>G.x.x.x</td>
<td>113</td>
</tr>
<tr>
<td>H.x.x.x</td>
<td>100</td>
</tr>
<tr>
<td>I.x.x.x</td>
<td>91</td>
</tr>
<tr>
<td>J.x.x.x</td>
<td>90</td>
</tr>
<tr>
<td>Others</td>
<td>1602</td>
</tr>
</tbody>
</table>

The properties of the ScriptGen approach allow to perform a completely automated incremental learning of the activities as shown in [14]. ScriptGen-based honeypots are able to detect when a client request falls out of the current FSM knowledge (a 0-day attack or, more exactly, a yet unseen attack) by simply detecting the absence of a matching transition. In such case, the honeypot is thus unable to provide a valid answer to the attacker. We showed in [14] how the honeypot can react to this situation relying on a real host (an oracle) and acting as a proxy between the attacker and the real host. This allows the honeypot to continue the conversation with the attacker, and to collect a new sample of protocol interaction that can be used to automatically refine the protocol knowledge.
Mozzie

Mozzie - Definition

- Mozzie:
  - It is a normalization environment for malware execution.
  - It is based on ScriptGen.

Mozzie is composed of:

- Ant:
  - Responsible to create the dictionary containing the different finite state machines for each endpoint found in the pcap files.

- Chameleon:
  - Responsible to route the packets by changing the IP addresses and the ports from the source to the fake destination and vice-versa.

- Dog:
  - Responsible to follow the payload in the finite state machine and found, if possible, the correct answer, otherwise it tries to contact the real server.
In the dictionary there are two kinds of keys:
- IP based (Protocol, Destination IP, Destination Port) → detailed.
- Port based (Protocol, None, Destination Port) → generic.

The values of the dictionary are the finite state machines (FSM).

Port based ≠ Protocol based.

Incremental learning improves at every run the FSM.

Every server (Dog) will follow a FSM for the current endpoint (otherwise the generic FSM will be used).
Mariano Graziano - Mozzie: a normalization environment for malware execution
Mariano Graziano - Mozzie: a normalization environment for malware execution
Mariano Graziano - Mozzie: a normalization environment for malware execution
Mozzie - Results

- Experiments have been gradual.
- It has been tested with different protocols (HTTP, IRC, IMAP, DNS) → common C&C protocols and protocols used in illicit activities (DDoS, Spam).
- It has been tested with a real malware.
- The key factor is to have a good training set.
- Satisfactory results (it handles random nicknames, authentication).
- Repeatability is possible.
- Long-term containment system.
Mozzie - Pros & cons

Advantages:
✓ It controls the network context → analysis is repeatable.
✓ Containment.
✓ It can be integrated with every sandbox.

Disadvantages:
✗ Encrypted protocols not supported.
✗ Number of required network traces (Training set problem).
The current disadvantages will be handled in the future.
The end

Thank you for the attention.