Politecnico di Torino

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 $\rightarrow$ cybercrime

Cybercrime

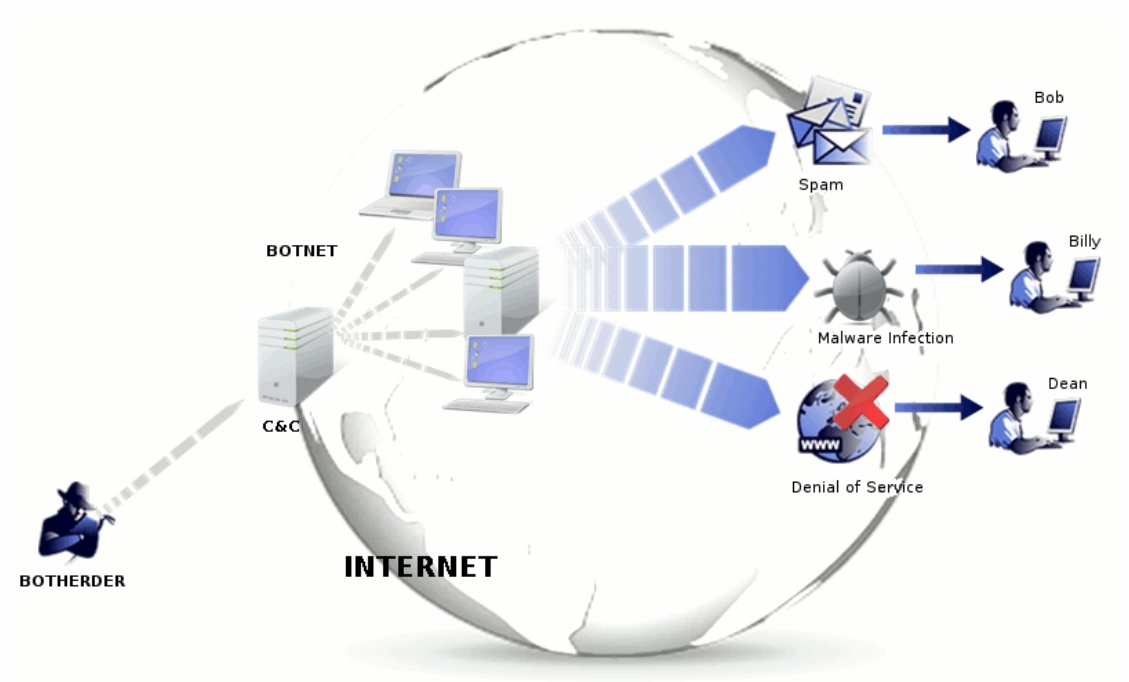
1

[1] http://evilfingers.blogspot.com/2009/10/current-business-outlook-caused-by.html

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 $\rightarrow$ big network of compromised computers.

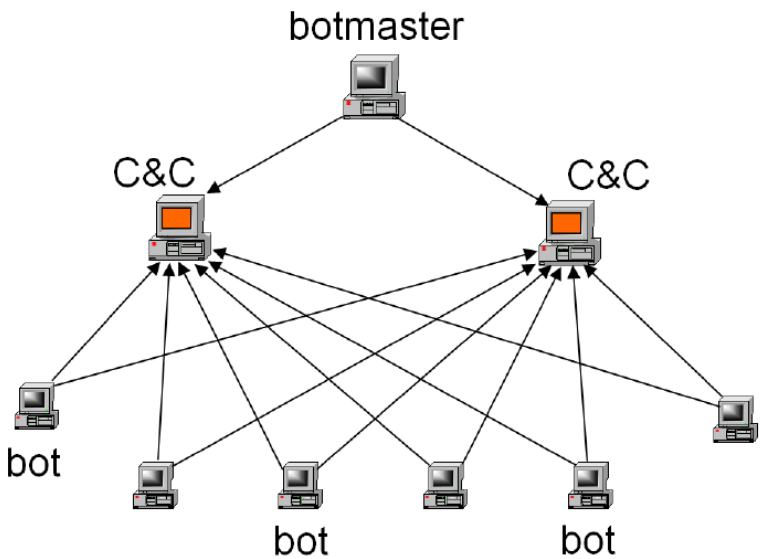
Botnet I



2

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

Botnet II



3

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

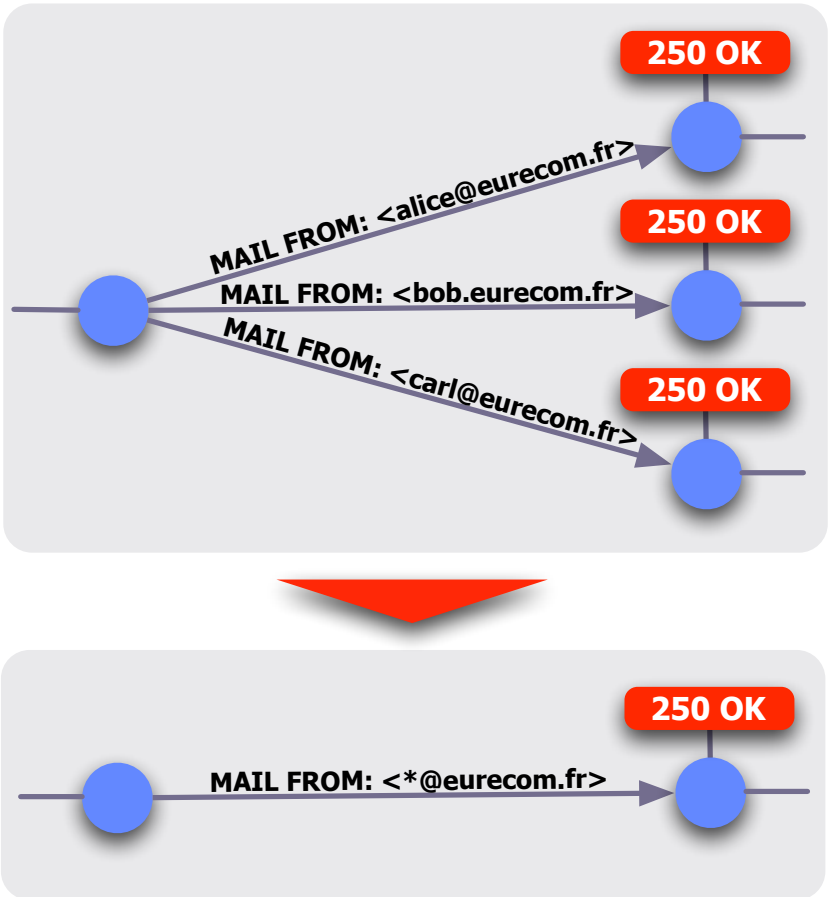
Analysis

- The malware analysis can be:
  - Static $\rightarrow$ manual and error prone (Reverse Engineering approach).
  - Dynamic $\rightarrow$ 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.

SMTP Finite state machine



Mozzie

4

[4] http://www.fusedfilm.com/2010/07/interview-willie-garson-talks-white-collar-season-2-and-first-season-blu-ray-release/

Mozzie - Definition

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

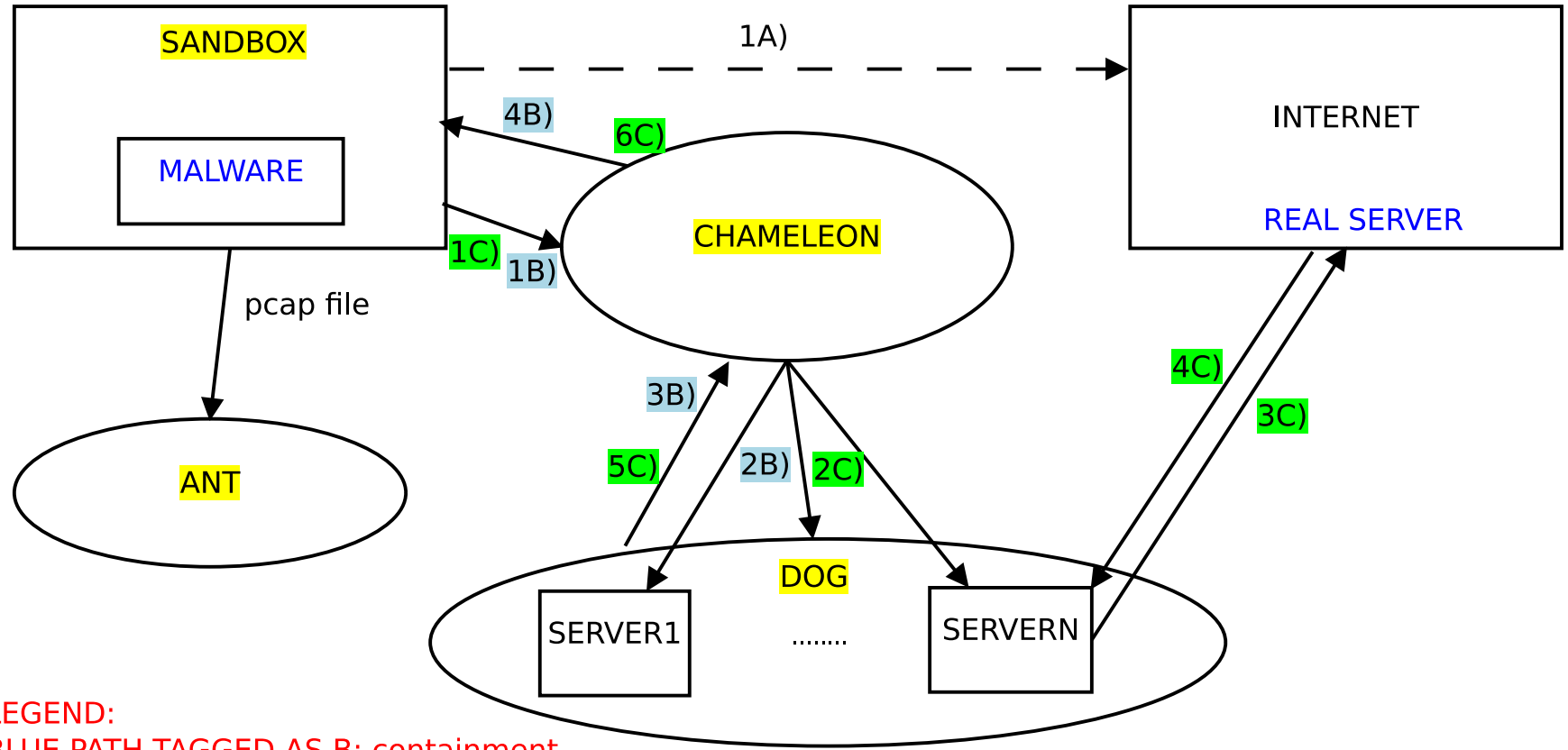
Mozzie is composed of:

- 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.

Mozzie - Dictionary

- In the dictionary there are two kinds of keys:
  - IP based (Protocol, Destionation IP, Destination Port) $\rightarrow$ detailed.
  - Port based (Protocol, None, Destination Port) $\rightarrow$ generic.
- The values of the dictionary are the finite state machines (FSM).
- Port based $\neq$ 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).

Mozzie - Overview



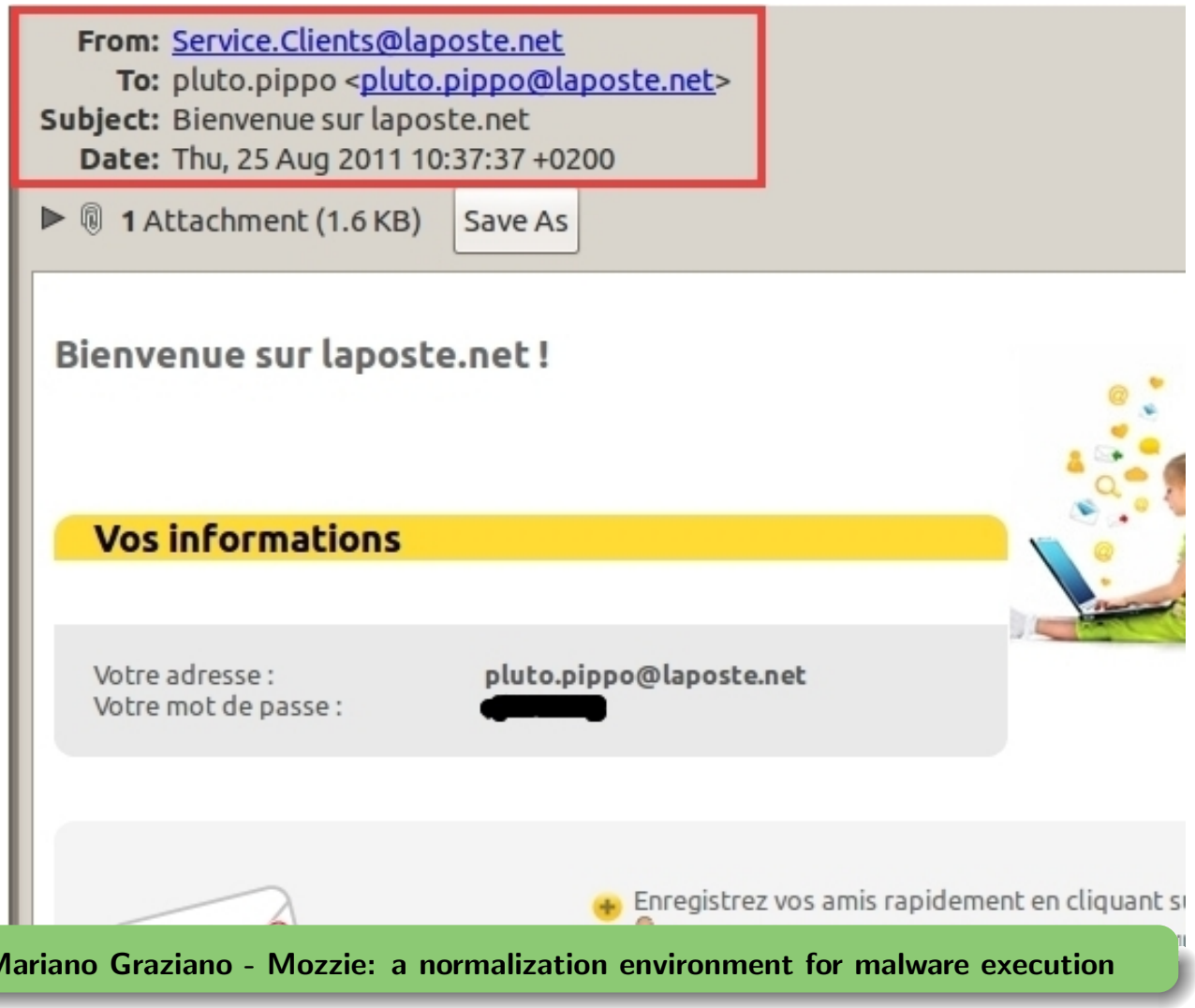
LEGEND:
BLUE PATH TAGGED AS B: containment
GREEN PATH TAGGED AS C: contact real server

Mozzie - Example

```
[2011-08-29 14:52:32,553] INFO    Chameleon              It's a TCP packet
[2011-08-29 14:52:32,554] INFO    Chameleon              Hash: 127.0.0.1:9000 <--> 10.0.0.91:56097
[2011-08-29 14:52:32,554] INFO    Chameleon              Adapting the packet...
[2011-08-29 14:52:32,554] INFO    Chameleon              Redirecting to the malware 10.0.0.91:56097
* 1 FETCH (UID 1 BODY[HEADER] {254}
content-type: multipart/related; type="text/html"; boundary="-"
Date: Thu, 25 Aug 2011 10:37:37 +0200
From: <Service.Clients@laposte.net>
To: pluto.pippo <pluto.pippo@laposte.net>
X-Unsent: 1
Mime-Version: 1.0
Subject: Bienvenue sur laposte.net

)
A00016 OK Completed
```

Mozzie - Example



Mozzie - Results

- Experiments have been gradual.
- It has been tested with different protocols (HTTP, IRC, IMAP, DNS) $\rightarrow$ 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.