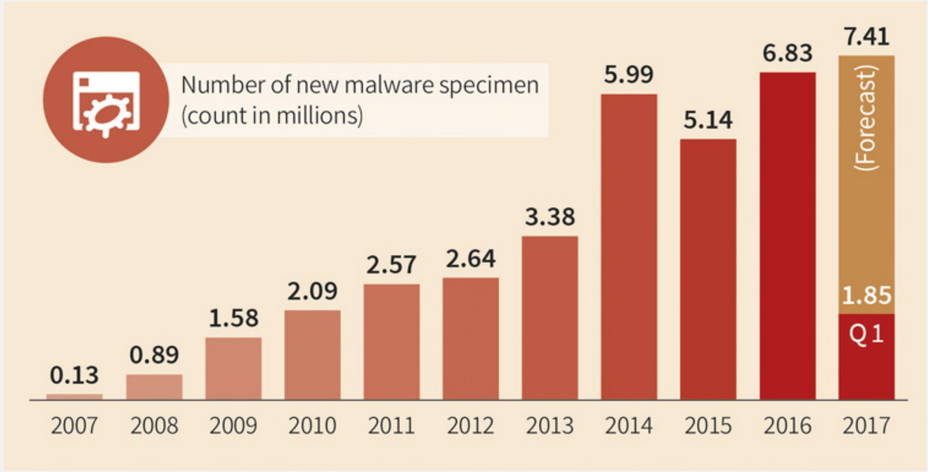
SAPIENZA
UNIVERSITÀ DI ROMA

Sandbox, evasive malware and text mining against Phishing

Federico Palmaro

Dealing with Malware today

The proliferation of malware in recent years has followed an exponential growth.



Source: G-Data Security Blog

Static and Dynamic analysis are two methodologies for the study of malware, each with their own strengths and weaknesses.

State of the art

Sandbox

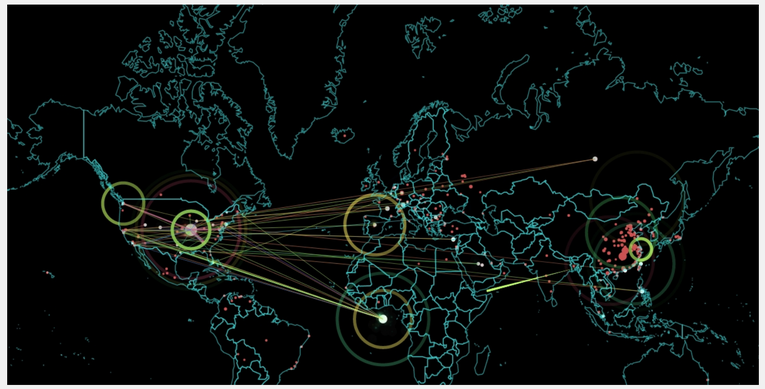
- A sandbox is a controlled environment in which analysts can observe the behavior of a malware sample while it is running.

Malware evasion

- Many samples contain sections designed specifically to fingerprint details of the surrounding software environment, comparing them with the known values of dynamic analysis tools: such techniques are called evasive techniques.

Threat Intelligence

Target Domains

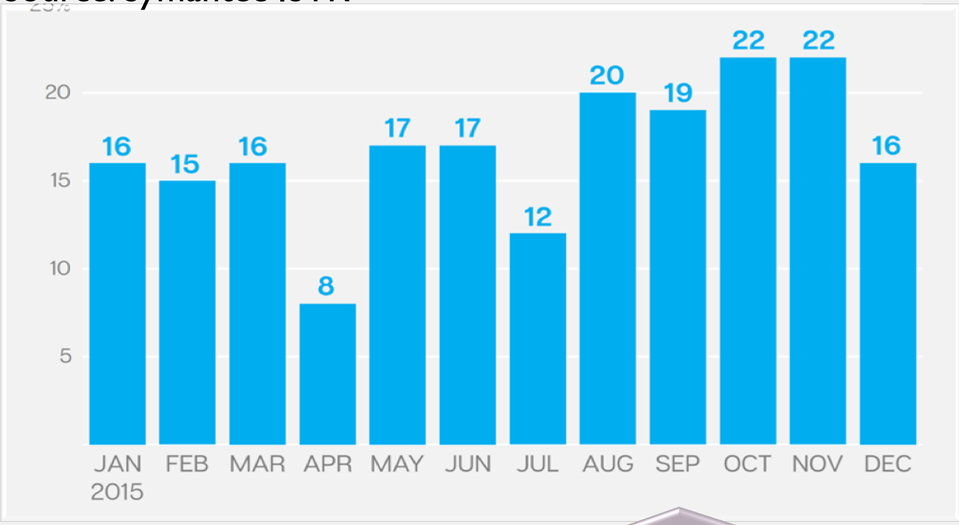


Language of Attackers

Target Emails

How relevant is the problem of evasion?

Source: Symantec ISTR



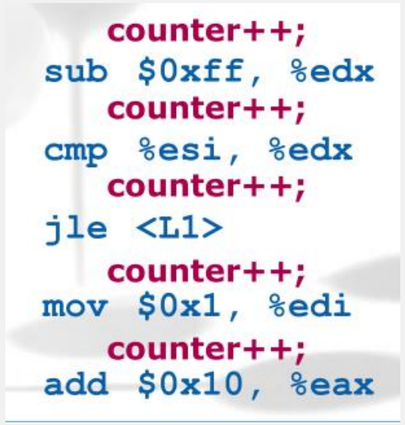
Approximately 16 percent of malware is routinely able to detect and identify the presence of a virtual machine environment, peaking at around 22 percent in Q4

Furtim malware, which in 2016 stunned the malware analysis community by showing more than 400 evasive techniques that could elude static and behavioral detection techniques used at the time.

Dynamic binary instrumentation

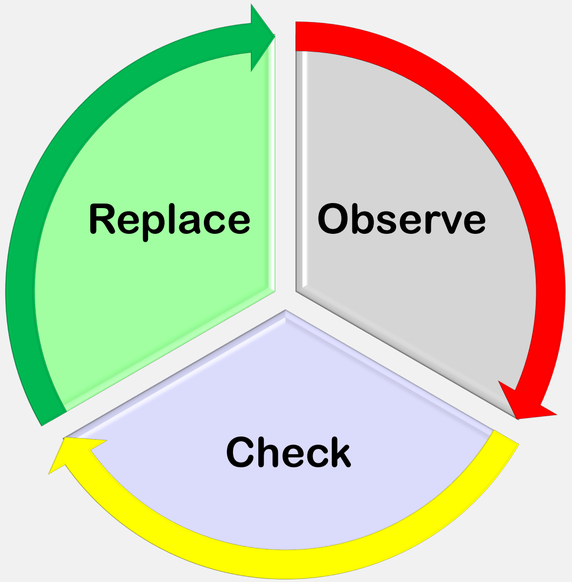
- A program analysis technique used to understand the behavior of a binary application at run time through the injection of instrumentation code. Such code executes as part of the normal instruction stream after being injected in the program.

We can insert code during execution in a transparent way!



Observe-check-replace paradigm

The process to provide fake responses to malware is based on three steps: **Observe, Check and Replace**



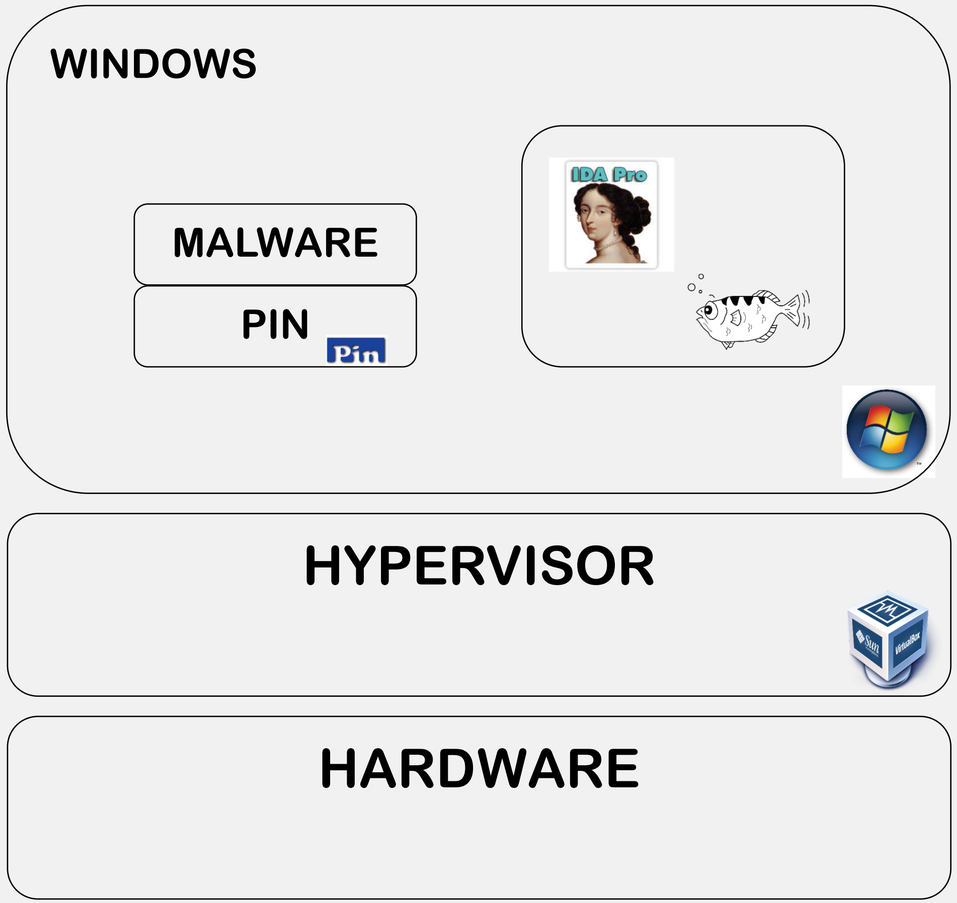
Observe: monitor all activities (e.g., calls) of running process

Check: analyze input parameters and return values of previously selected actions

Replace: fix "wrong" values with other considered valid *(as a real environment would yield)*

IsDebuggerPresent is a library call that returns true if a debugger is present in a system

Implementation overview



Using this infrastructure, we have managed to find and fight over 100 different types of evasive techniques!

What is phishing

- Phishing e-mails are used by malicious actors with the aim of obtaining sensitive information from a victim, deceiving or blackmailing them. An inattentive or uninformed user may often fail to recognise if an e-mail is sent by an authentic sender or is a scam.

Countermeasures in the wild

Blacklists

Lists of detected phishing URLs, IP addresses or keywords. Provide the result with low overhead, but do not provide protection against zero-hour phishing attacks.

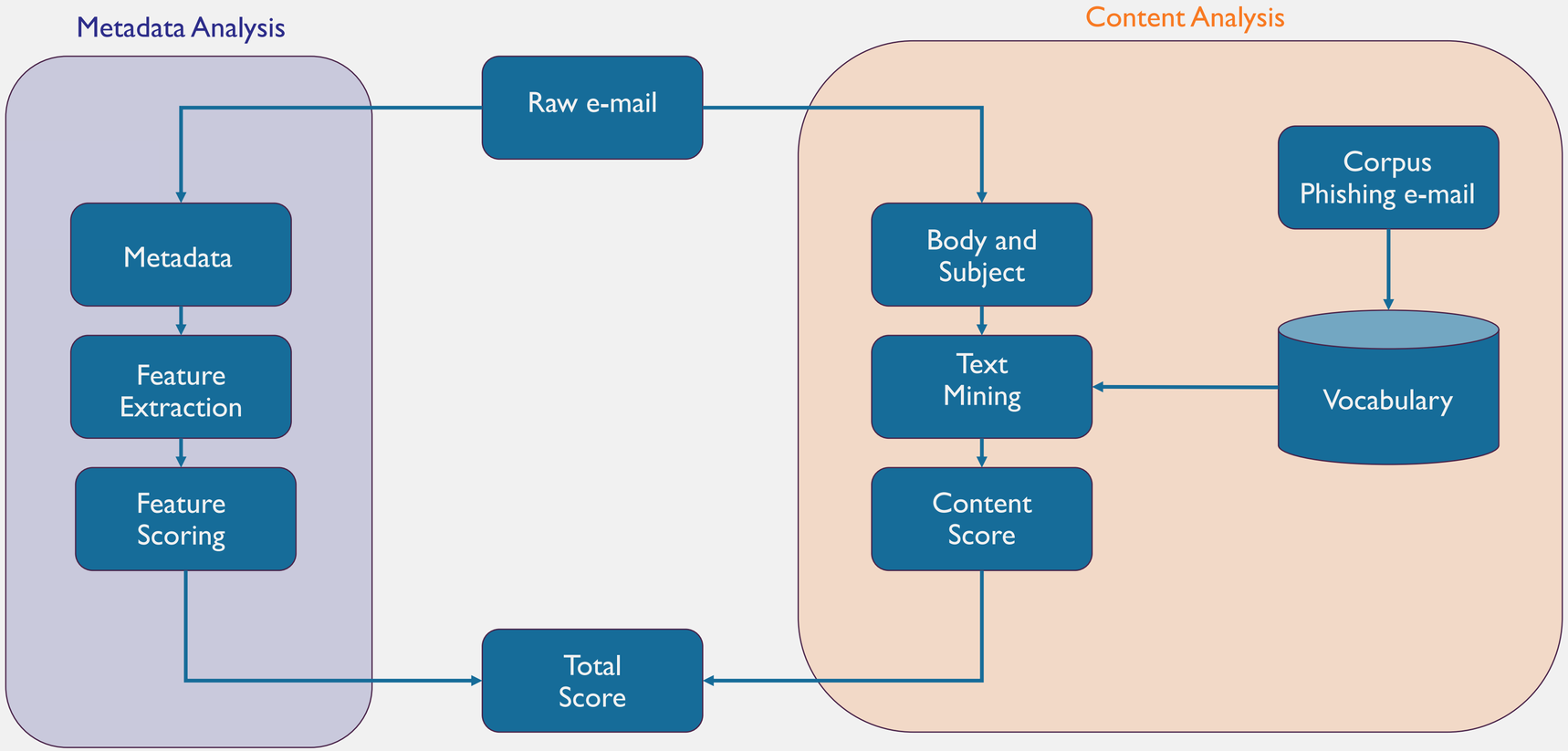
Heuristics

Characteristics that are found to exist in phishing attacks in reality. If a set of general heuristic tests are identified, it can be possible to detect zero-hour phishing attacks.

Data Mining

Document classification or clustering problem, where models are constructed by taking advantage of Machine Learning and clustering algorithms.

Proposed solution



Hands-on

```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  3675
NO PUBLIC DOMAIN
CONTENT-TYPE: text/plain
ON A TOTAL OF 0 URLs 0 ARE UNSAFE
ON A TOTAL OF 227 PUNTACTION MARKS 1 ARE !
ON A TOTAL OF 227 PUNTACTION MARKS 38 ARE $
ON A TOTAL OF 347 WORDS CONTENT 80 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 5 WORDS SUBJECT 3 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 347 WORDS 30 ARE MISSPELLED
NO ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 3
--------------------------------------------------------------------------------
```

⟸ Legitimate e-mail

Phishing e-mail ⟹

```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  306
NO PUBLIC DOMAIN
CONTENT-TYPE: text/plain
ON A TOTAL OF 0 URLs 0 ARE UNSAFE
ON A TOTAL OF 85 PUNTACTION MARKS 0 ARE !
ON A TOTAL OF 85 PUNTACTION MARKS 1 ARE $
ON A TOTAL OF 619 WORDS CONTENT 175 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 6 WORDS SUBJECT 3 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 619 WORDS 10 ARE MISSPELLED
ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 5
--------------------------------------------------------------------------------
```

Results

- REALISTIC DATASET
- USER EDUCATION
- 99,2% ACCURACY

Contact

Federico Palmaro

+39 331 3830430

Federico-1

Fede.nik94@gmail.com

linkedin/federico-palmaro-634092116/

SAPIENZA
UNIVERSITÀ DI ROMA