Improving Protection against Internet Attacks through Contextual Feature Pairing

Georgiana Ingrid Stoleru

Supervised by: Assoc. Prof. PhD Dragoș Teodor Gavriluț

Faculty of Computer Science, "Alexandru Ioan Cuza" University

July, 2018

Table of Contents

Table of Contents

Overview

Definition

A **security attack** represents an attempt to gain **unauthorized access** to information resources or services, or to cause **damage** to information systems.
(*Big Data Security Management*, Zaiyong Tang and Youqin Pan)

Overview

Steps at which detection occurs

- **Downloading**
- **Writing** on disk
- **Reading** from disk
- **Execution**

Table of Contents

Status Quo

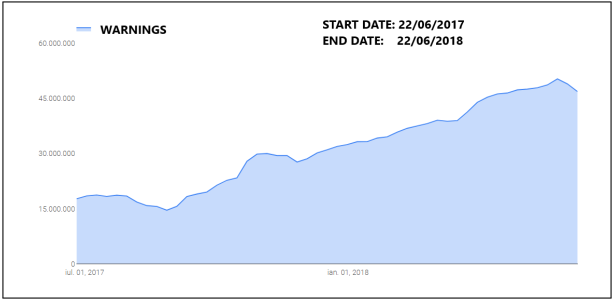
- Increasing number of malicious URLs:

Phishing URLs	Malicious URLs
> 45000 per week	> 15000 per week

Google Safe Browsing Analysis
22/06/2017 -> 22/06/2018

- Short life span of a malicious URL:
  - Average phishing web site: **54** hours (AntiPhishing Working Group, June 2018)

Statistical Indicators



Weekly number of displayed warnings (Google Safe Browsing)

Related work: Blacklists

Standard detection technologies:

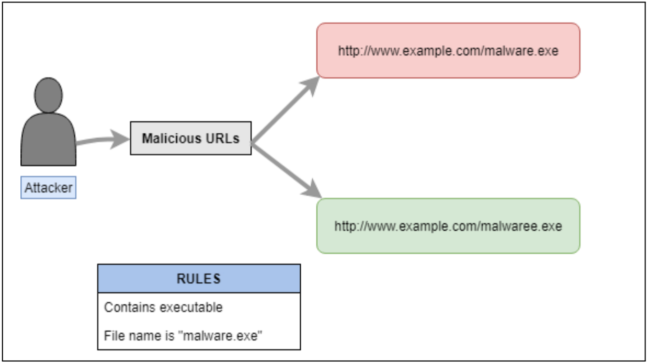
- **Blacklists**

Malicious URLs
http://www.comprealm.net/wordpress/1w0jkheYE8/
http://www.icb.cl/ZxavoDe/
http://www.chungcusamsoraprimier.com/DW8dXe/
http://www.service-pc.com.ro/7o9opMY/
http://www.minami.com.tw/P4UDGp/

URLs hosting Emotet samples

Related work: Subroutines

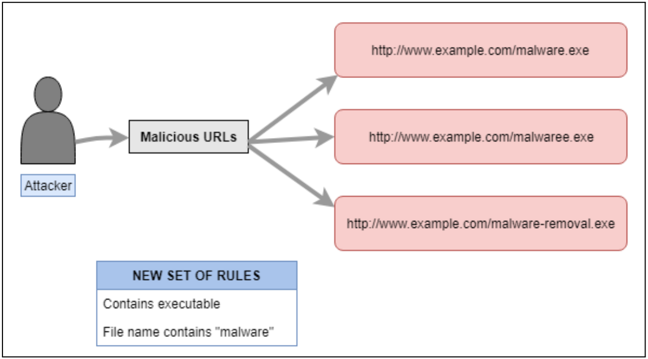
- Subroutines



Malicious URL Attack and Subroutine Defense flow

Related work: Subroutines

- Subroutines



Malicious URL Attack and Subroutine Defense flow

Table of Contents

Subroutines description

Description of the approach

- Detection technique consisting of various **sets of rules**

Remarks

- An approach to the identification of **similar URLs**
- A step towards **feature extraction**

Description

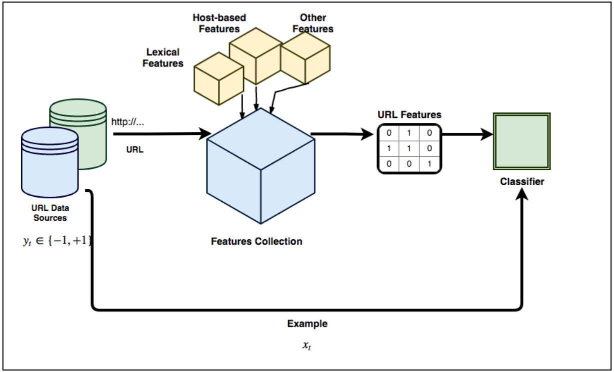
Advantages

- The **certainty** that the URL will be detected
- The **effectiveness** in a suite of attacks (E.g.: `http://www.example.com/image.png`)

Disadvantages

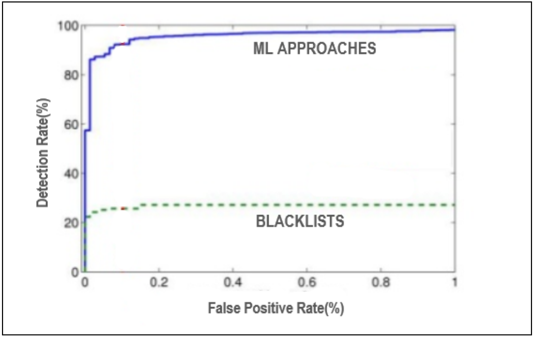
- The ease of **evading detection**
- The **rules** are **manually specified**

Machine Learning Classification System



URL Classification System

Remarks



Related Work

These approaches often involve a high rate of False Positives.

Solution

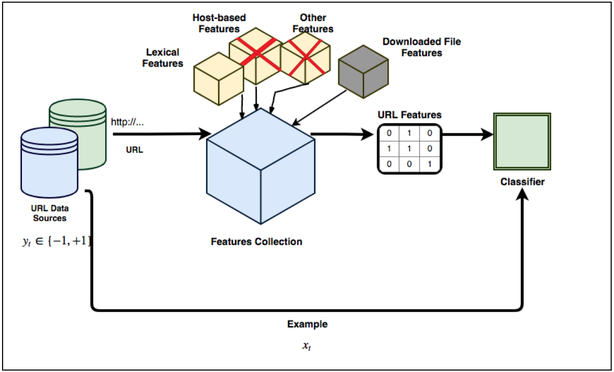

Host-based
Features
Other
Features
Lexical
Features
Downloaded File
Features
http://...
URL
URL Features
0 1 0
1 1 0
0 0 1
Classifier
URL Data
Sources
$y_t \in \{-1, +1\}$
Features Collection
Example
x_t

Table of Contents

URL Features

1^{st} Remark

We extracted the **lexical features** of the URLs (**148** features).

2^{nd} Remark

We **discretized** the continuous features.

Advantages of discretization

- **Memory** space
- **Resiliency** to change

URL Features Classification

Domain Features

- **Domain** is **IP** address
- TLD is common
- Domain is **randomly** generated

Directory Features

- Subdirectory tokens
- Existence of **small words**
- Existence of **random words**

URL Features Classification

Content Features

- File content is an **executable** or a document
- File content has a **known extension**
- File name is **randomly** generated

Argument Features

- URL contains **parameters**
- Parameters can indicate log-in information

Example

Malicious URL Example

Consider the **malicious** URL:

`http://cdn.discordapp.com/attachments/402490727474528267/407242837365751809/d.exe`

Category	Feature	Description of the feature
Content related	URL-IS-EXEC	The content is an executable file
Content related	FILENAME-IS-ALPHANUMERIC	The downloaded file name contains alphanumeric symbols
Content related	KNOWN-EXTENSION	The extension of the downloaded file belongs to a predefined list
URL related	HTTP-PROTOCOL	The protocol used is "http"
Domain related	KNOWN-TLD	The tld is common
Directory related	PREV-DIGIT	The last but one split contains only digits
Directory related	PREV-SHORT	The last but one split has a small length

A part of the extracted features for the URL

1^{st} Remark

We **downloaded** the files corresponding to the URLs in the database.

2^{nd} Remark

We extracted the **features** corresponding to the **downloaded files** (**8413** features).

Extracted features

- Behaviour in virtual environments
- File format from the geometrical point of view
- Packed/ Obfuscated file

Feature	Value	Description of the feature
IS-MSIL	1(True)	The sample is a MSIL file
SET-STARTUP	1(True)	The program adds itself to startup
IS-PACKED	0(False)	The program is packed with a known packer
RANDOM-WORDS	2	There have been identified two random words in the file
NUMBER-CLASSES	4	The program contains four classes
NUMBER-RESOURCES	1	The program contains one resource
NUMBER-ICONS	0	There are 0 icons in the resources section

A part of the extracted features for the file

Table of Contents

Formal Model: Algorithm

Algorithm

We have used the **OSC** algorithm, a derived version of the **Perceptron**, because it is adjusted for a low number of **False Positives**.

Advantages of OSC

- Verdict provided in **linear time**
- Low number of false positives
- Less resource demanding

Feature Selection: F2-Score

F2-Score

$$\mathbf{F2} = 5 \times \frac{precision \times recall}{4 \times precision + recall} \tag{1}$$

Remark

F2-Score is a **Uni-variate** feature selection method.

Conditional Mutual Information Maximization Criterion

- It does not select a feature similar to already pickes ones.
- Naive Bayes Classifier together with CMIM criterion provide the same error rates as AdaBoost or SVMs.

Table of Contents

Dataset

Data Selection

1 million samples (Bitdefender Cyber Threat Intelligence Lab)

Data Filtering - 1^{st} Step

- Not **executable** content
- Clusters consisting of highly **similar** URLs
- A lower life span than the average for a malicious URL
- **98163** malicious samples and **234574** benign ones

Dataset

Data Filtering - 2^{nd} Step

- **Inconsistencies** removal
- Sequences of features related to benign samples
- **Duplicated** data
- **11107** malicious samples and **31247** benign ones

Training

OSC-U

- Number of features: **148**
- Source of features: URLs
- Number of epochs: **2000**
- Dataset: **11107** malicious samples and **31247** benign samples

Se	Tn	Tp	Acc
44.002%	31,247	4,887	85.31%

148 features — 2000 epochs

Training

Remark

In order to be used in practice, the model should provide a high detection rate.

Solution

- Increase the precision of the model by adding **features** extracted from **files**.
- Apply **feature selection** on the set of file features.

Training

OSC-UF

Number of features from URLs: **148**

Number of features from files: **256**

Feature Selection Algorithm: **F2-Score**

Number of epochs: **2000**

Se	Tn	Tp	Acc
72.57%	31,247	8,060	92.8%

404 features — 2000 epochs

Training

Remark

We notice a considerable **evolution**, but still with a detection rate lower than 75%.
F2-Score **scores** each of the features **individually**.

Solution

Choose a feature selection algorithm which selects only features which carry additional information about the class to predict. (**Conditional Mutual Information Maximization criterion**)

Training

OSC-UFF

Number of features from URLs: **148**

Number of features from files: **256**

Feature Selection Algorithm: **CMIM criterion**

Number of epochs: **2000**

Se	Tn	Tp	Acc
94.34%	31,247	10,478	98.5%

404 features — 2000 epochs

Improving the model

- Making features **linearly separable** by **mapping**
- New space with $m(m+1)/2$ features
- **Logical conjunction** between initial features

Training

OSC-CM

Number of features from URLs: **148**

Number of features from files: **81,810**

Feature Selection Algorithm: **CMIM criterion**

Number of epochs: **2000**

Se	Tn	Tp	Acc
95.26%	31,247	10,580	98.75%

81,810 features — 2000 epochs

Training

OSC-CM1

Number of features from URLs: **148**

Number of features from files: **81,810**

Feature Selection Algorithm: **CMIM criterion**

Number of epochs: **10000**

Se	Tn	Tp	Acc
96.60%	31,247	10,729	99.10%

81,810 features — 10,000 epochs

Real world detection data

Statistical indicators

- Test environment: **Bitdefender's technologies**
- Period: **1 month**
- Subset from the classified data: **15,273** malicious samples and **34,727** clean samples

FP + TN	FP	FP rate	FN + TP	TP	TP rate
34,727	52	0.0015%	15,273	12,140	79.49%

Real world detection data

Table of Contents

Future directions

- Extend the approach for different **protocols**
- Add further categories of **features** (e.g.:host-based)
- Port the algorithms on the **GPU** of the clients
- Process data in the **cloud**
- Take into account the **reputation** of a sample

Thank you!

Q&A