DE MONTFORT
UNIVERSITY
LEICESTER

Raw Network Traffic Data Preprocessing and Preparation for Automatic Analysis

Basil Alothman

PhD Student
Faculty of Technology
De Montfort University, Leicester, UK

Outline

I. Introduction

II. Overview of Steps

III. Steps in Detail

IV. Case Study using Real Network Traffic Data

V. Plotting Data Distribution

VI. Conclusion

Introduction

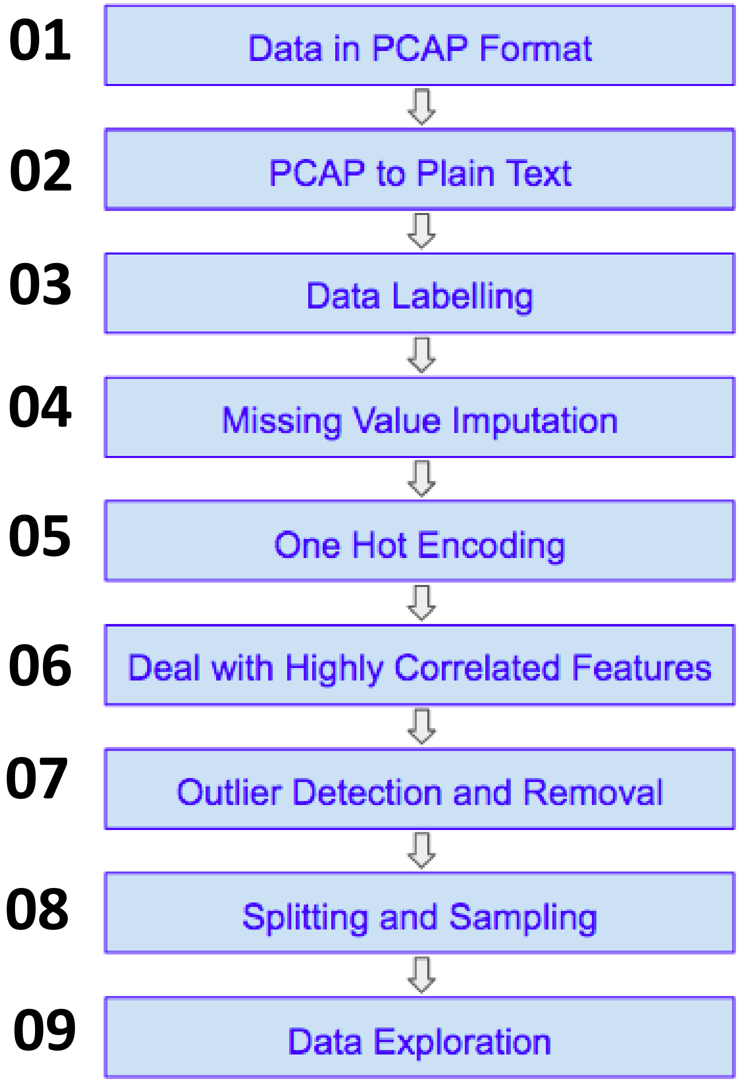
This presentation provides:

1. **Several steps that should be considered when carrying out network traffic data transformation from raw to a textual format.**

2. **Illustration of those steps in a case study using real, rather than simulated data.**

Overview of Steps

1. Some of these steps are essential for CSIRT teams in order to detect malicious network traffic with high accuracy.
2. Some can be optional.



01 Data in PCAP Format

02 PCAP to Plain Text

03 Data Labelling

04 Missing Value Imputation

05 One Hot Encoding

06 Deal with Highly Correlated Features

07 Outlier Detection and Removal

08 Splitting and Sampling

09 Data Exploration

Steps in Detail 1/3

1. **Raw Network Traffic Data can be obtained via capture tools such as WireShark (usually in PCAP format).**

2. **PCAP format can be transformed into CSV using tools such as FlowMeter (generates several useful features).**

3. **Resulting CSV data should be labelled (e.g. when generating training data).**

Steps in Detail 2/3

4. **Data should be checked for missing values, if any exist, these values should be imputed (several techniques).**

5. **Sometimes categorical features are represented numerically, this is not recommended and one-hot encoding can be used to represent such data correctly.**

6. **Some of the features in the generated data can be highly correlated, this case must be dealt with appropriately to achieve reliable results.**

Steps in Detail 3/3

7. Data should be checked for outliers, if any exist, they should be dealt with appropriately (depends on the purpose of analysis).

8. If the data contains multiple categories (e.g. Normal, DDoS, Worm … etc), it might be useful to have a separate dataset for each category (depends on the purpose of analysis).

9. Data exploration techniques can be used to inspect the data distribution.

Case Study 1/2

- Downloaded the ISCX Dataset (in PCAP format) http://www.unb.ca/cic/datasets/botnet.html

- Contains traffic data for multiple Botnets as well as Normal traffic.

- Used FlowMeter to transform it into CSV.

- Followed guidelines provided by ISCX team for labelling.

- Replaced missing values in each feature by Median of that feature.

Case Study 2/2

- Used one-hot encoding to represent source port, destination port and protocol fields.

- Removed highly correlated features (paper on FlowMeter has more details on these).

- Detected and removed Outliers.

- Split data into smaller datasets according to label (each Botnet has a separate dataset).

- Used Principal component analysis (PCA) and Partial Least Squares (PLS) to explore data.

Results of Partial Least Squares (PLS)

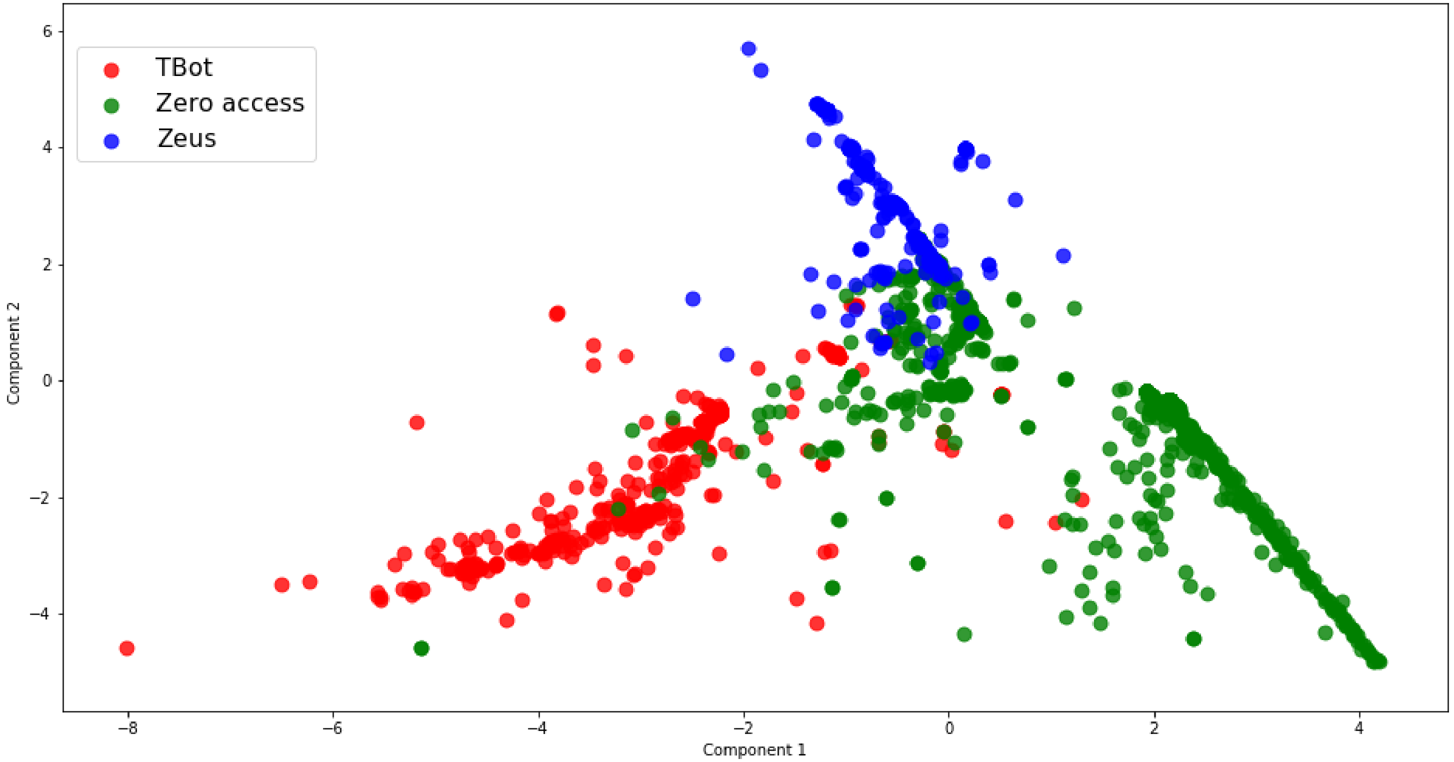


Figure : PLS Components for TBot, Zero access and Zeus data

- Plot shows clear separation of data from different Botnets.
- It shows data have different distributions.
- This is important for Data Mining and Machine Learning.

Conclusion

- This work presented several steps that should be considered when pre-processing raw network traffic data for data mining (e.g. making predictions).

- Some of these steps are essential, some others can be optional.

- Provided a case study using a freely available dataset.

- Results show the steps are indeed key to obtaining reliable results.

- More details in the paper.

Thank you..