SPRAWOZDANIE 3

Pomiary, analiza i modelowanie systemów internetowych

WERONIKA SIERACKA
JOANNA MIELNICZUK
PIOTR OBERSKI
KAMIL POCZĄTEK
MACIEJ WOJTKO

Cel projektu:

Celem projektu jest zbadanie określonego zbioru danych z wykorzystaniem „Orange” do zautomatyzowanego odkrywania zależności oraz schematów.

Wstęp:

Poddano analizie zbiór danych zawierający informacje o ruchu w sieci, w tym – o ruchu w Darknecie.

Darknet jest to „ciemna strona internetu”, którą tworzą sklepy e-commerce, anonimowe fora dyskusyjne, strony internetowe oraz blogi. Kluczem dostępu jest odpowiednie ustawienie przeglądarki lub określone oprogramowanie. Baza zawiera 141530 rekordów oraz 86 encji. Możemy je podzielić na 2 rodzaje danych jakościowych oraz 84 rodzajów danych ilościowych.

Dane jakościowe:

- **Label1** - stosowane szyfrowanie
- **Label2** – kategoria ruchu

Dane ilościowe:

- **Flow ID** - połączenie Src IP, Dst IP, Src Port, Dst Port, Protocol
- **Src IP** - IP źródła
- **Src Port** - Port źródła
- **Dst IP** - IP docelowe
- **Dst Port** - Port docelowy
- **Protocol** - numer protokołu
- **Timestamp** - czas
- **Flow Duration** - liczba pakietów
- **Total FWD Packets** - liczba wysłanych pakietów
- **Total BWD Packets** - liczba otrzymanych (zwróconych) pakietów
- **Total Length of Fwd Packet** - całkowity rozmiar wszystkich wysłanych pakietów
- **Total Length of Bwd Packet** - całkowity rozmiar wszystkich otrzymanych (zwróconych) pakietów
- **Fwd Packet Length Min** - minimalny rozmiar wysłanych pakietów
- **Fwd Packet Length Max** - maksymalny rozmiar wysłanych pakietów
- **Fwd Packet Length Mean** - średnia wartość wysłanych pakietów
- **Fwd Packet Length Std** - odchylenie standardowe wysyłanych pakietów
- **Bwd Packet Length Max** – maksymalny rozmiar otrzymanych (zwróconych) pakietów
- **Bwd Packet Length Min** – minimalny rozmiar otrzymanych (zwróconych) pakietów
- **Bwd Packet Length Mean** - średnia wartość otrzymanych (zwróconych) pakietów
- **Bwd Packet Length Std** - odchylenie standardowe otrzymanych (zwróconych) pakietów

- **Flow Bytes/s** – liczba przepływu bytów na sekundę
- **Flow Packets/s** – liczba przepływu pakietów na sekundę
- **Flow IAT Mean** - średni czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Flow IAT Std** – odchylenie standardowe pomiędzy dwoma pakietami wysłanymi w przepływie
- **Flow IAT Max** – maksymalny czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Flow IAT Min** – minimalny czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Fwd IAT Total** - całkowity czas między dwoma wysłanymi pakietami
- **Fwd IAT Mean** - średni czas pomiędzy dwoma wysłanymi pakietami
- **Fwd IAT Std** – czas odchylenia standardowego pomiędzy dwoma wysłanymi pakietami
- **Fwd IAT Max** – maksymalny czas pomiędzy dwoma wysłanymi pakietami
- **Fwd IAT Min** – minimalny czas pomiędzy dwoma wysłanymi pakietami
- **Bwd IAT Total** – całkowity czas pomiędzy dwoma otrzymanymi (zwróconymi) pakietami
- **Bwd IAT Mean** – średni czas pomiędzy dwoma otrzymanymi (zwróconymi) pakietami
- **Bwd IAT Std** – czas odchylenia standardowego pomiędzy dwoma otrzymanymi (zwróconymi) pakietami
- **Bwd IAT Max** – maksymalny czas pomiędzy dwoma otrzymanymi (zwróconymi) pakietami
- **Bwd IAT Min** – minimalny czas pomiędzy dwoma otrzymanymi (zwróconymi) pakietami
- **Fwd PSH Flags** - ile razy flaga PSH została ustawiona w wysyłanych pakietach (0 dla UDP)
- **Bwd PSH Flags** - ile razy flaga PSH została ustawiona w otrzymanych (zwróconych) pakietach (0 dla UDP)
- **Fwd URG Flags** - liczba przypadków ustawienia flagi URG w wysyłanych pakietach
- **Bwd URG Flags** - liczba przypadków ustawienia flagi URG w otrzymanych (zwróconych) pakietach
- **Fwd Header Length** - całkowita liczba bajtów wykorzystanych na nagłówki (wysyłanie)
- **Bwd Header Length** - całkowita liczba bajtów wykorzystanych na nagłówki (otrzymywanie)
- **Fwd Packets/s** - liczba pakietów wysyłanych na sekundę
- **Bwd Packets/s** - liczba pakietów otrzymanych (zwróconych) na sekundę
- **Packet Length Min** - minimalna długość pakietu
- **Packet Length Max** - maksymalna długość pakietu
- **Packet Length Mean** – średnia długość pakietu
- **Packet Length Std** – odchylenie standardowe długości pakietu
- **Packet Length Variance** - wariancja długości pakietu
- **FIN Flag Count** - liczba pakietów z FIN
- **SYN Flag Count** - liczba pakietów z SYN

- **RST Flag Count** – liczba pakietów z RST
- **PSH Flag Count** - liczba pakietów z PSH
- **ACK Flag Count** - liczba pakietów z ACK
- **URG Flag Count** - liczba pakietów z URG
- **CWE Flag Count** - liczba pakietów z CWE
- **ECE Flag Count** - liczba pakietów z ECE
- **Down/Up Ratio** - współczynnik pobierania i wysyłania
- **Average Packet Size** - średnia wielkość pakietu
- **Fwd Segment Size Avg** - średni rozmiar (wysyłanie)
- **Bwd Segment Size Avg** - średni rozmiar (odbieranie)
- **Fwd Bytes/Bulk Avg** - średnia liczba wysyłanych bajtów
- **Fwd Packet/Bulk Avg**- średnia liczba odbieranych(zwracanych) pakietów
- **Fwd Bulk Rate Avg** – średnia wysyłana stawka hurtowa
- **Bwd Bytes/Bulk Avg**- średnia liczba odbieranych(zwracanych) bajtów
- **Bwd Packet/Bulk Avg** - średnia liczba odbieranych(zwracanych) pakietów
- **Bwd Bulk Rate Avg** – średnia odbieranych(zwracanych) stawka hurtowa
- **Subflow Fwd Packets** - średnia liczba wysyłanych pakietów w podprzepływie
- **Subflow Fwd Bytes** - średnia liczba wysyłanych bajtów w podprzepływie
- **Subflow Bwd Packets** - średnia liczba odbieranych(zwracanych) pakietów w podprzepływie
- **Subflow Bwd Bytes** - średnia liczba odbieranych(zwracanych) pakietów w podprzepływie
- **FWD Init Win Bytes** - całkowita liczba bajtów wysłanych w początkowym oknie
- **Bwd Init Win Bytes** - całkowita liczba bajtów odbieranych(zwracanych) w początkowym oknie
- **Fwd Act Data Pkts** - liczba wysyłanych pakietów z co najmniej 1 bajtem ładunku danych TCP
- **Fwd Seg Size Min** - minimalny rozmiar segmentu wysyłanego
- **Active Mean** - średni czas, w którym przepływ był aktywny przed przejściem w stan bezczynności
- **Active Std** - odchylenie standardowe czasu, w którym przepływ był aktywny przed przejściem w stan bezczynności
- **Active Max** - maksymalny czas, przez który przepływ był aktywny przed przejściem w stan bezczynności
- **Active Min** – minimalny czas, przez który przepływ był aktywny przed przejściem w stan bezczynności
- **Idle Mean** - średni czas, w którym przepływ był bezczynny, zanim stał się aktywny
- **Idle Std** - odchylenie standardowe czasu, w którym przepływ był bezczynny przed aktywacją
- **Idle Max** - maksymalny czas bezczynności przepływu przed aktywacją
- **Idle Min** – minimalny czas bezczynności przepływu przed aktywacją

Częstotliwość zbierania danych:

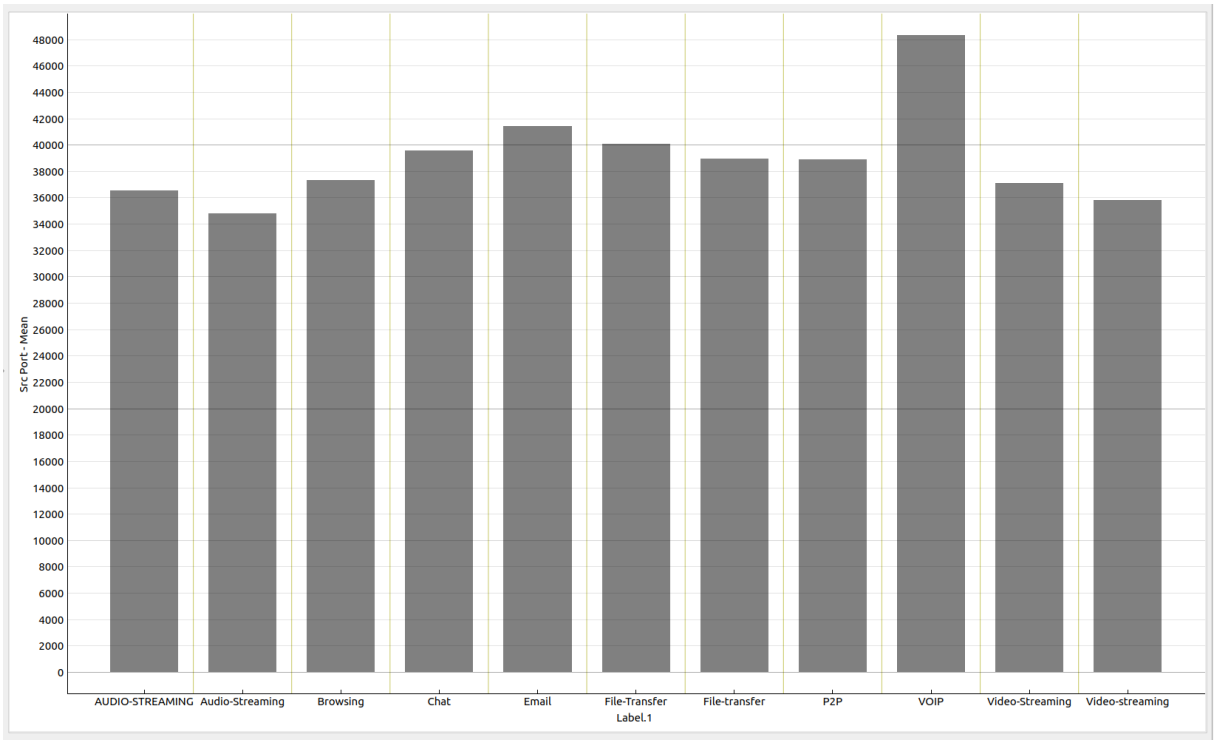
Baza danych zawiera informacje zbierane w sposób ciągły przez ponad rok. Pierwsze rekordy pochodzą z 4 stycznia 2015r., a ostatnie z 25 lutego 2016r.

Kategoryzacja danych:

W naszej bazie danych możemy skategoryzować ruch na:

- **Audio-Stream** – aplikacje audio, które wymagają ciągłego i stabilnego strumienia danych, np. Vimeo i Youtube
- **Browsing** - pod tą etykietą mamy ruch HTTP i HTTPS generowany przez użytkowników podczas przeglądania informacji w Firefox i Chrome
- **Chat** - identyfikuje aplikacje do obsługi wiadomości błyskawicznych, np. ICQ, AIM, Skype, Facebook i Hangouts
- **Email** - próbki ruchu wygenerowane przy użyciu klienta Thunderbird oraz kont „Alice" i „Bob". Klienci zostali skonfigurowani do dostarczania poczty przez SMTP/S i odbierania jej przy użyciu protokołu POP3/SSL w jednym kliencie i IMAP/SSL w drugim
- **P2P** - służy do identyfikacji protokołów udostępniania plików, takich jak Bittorrent
- **Transfer** - aplikacje ruchu, których głównym celem jest wysyłanie lub odbieranie plików i dokumentów. Dla naszego zestawu danych przechwycono transfery plików Skype, sesje ruchu FTP przez SSH (SFTP) i FTP przez SSL (FTPS)
- **Video-Stream** - aplikacje wideo, które wymagają ciągłego i stabilnego strumienia danych. Przechwycono ruch z YouTube (wersje HTML5 i flash) oraz usług Vimeo przy użyciu Chrome i Firefox
- **VoIP** - grupuje cały ruch generowany przez aplikacje głosowe. W ramach tej etykiety przechwycono rozmowy głosowe za pomocą Facebooka, Hangouts i Skype

Na poniższym rysunku znajduje się histogram zawierający ilość danych z każdej kategorii. Został on stworzony przy użyciu oprogramowania o nazwie „Orange”.



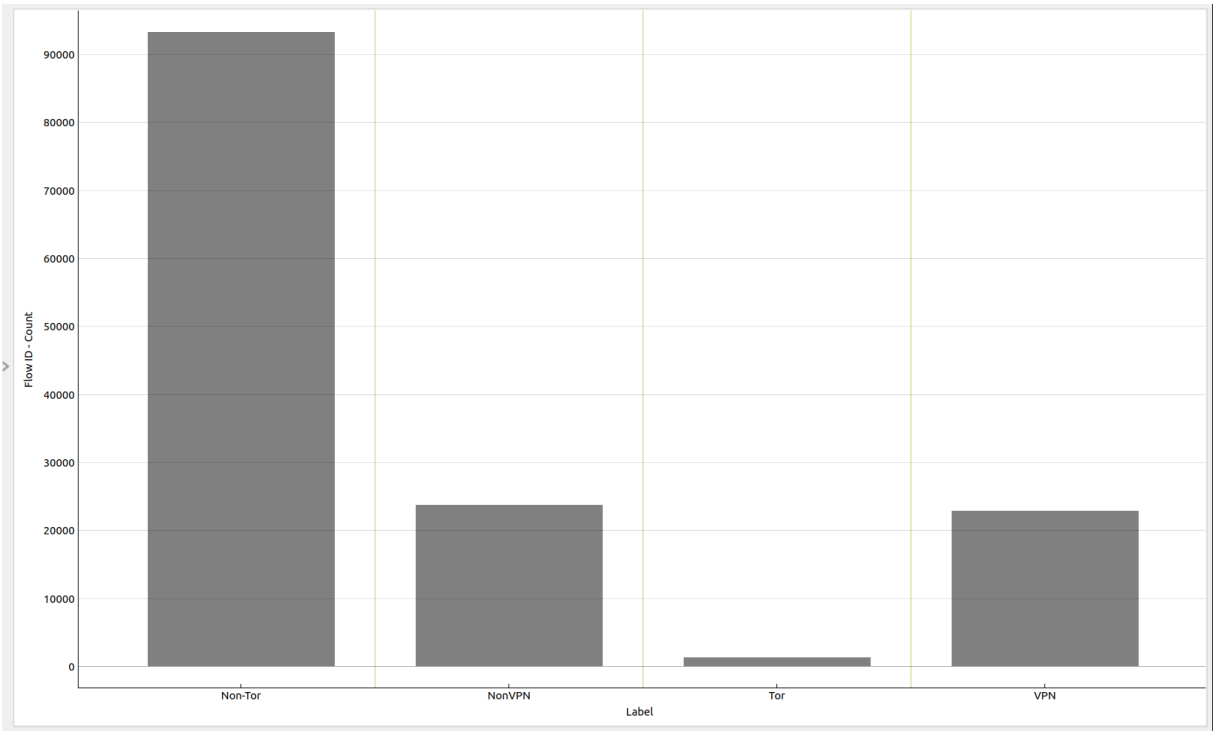
Jak możemy zauważyć największą ilość danych zebrano o VoIP, a najmniejszą o Audio-Streaming.

Dodatkowo – przy etykietowaniu danych wkradł się błąd dotyczący wielkości liter w kategoriach Video-Stream i Audio-Stream.

Ponadto, dane można skategoryzować na podstawie zastosowanego szyfrowania:

- **Tor** – z języka angielskiego „The Onion Router”, wirtualna sieć zapobiegająca analizie ruchu sieciowego, zapewniająca użytkownikom prawie anonimowy dostęp do zasobów Internetu.
- **VPN** – z języka angielskiego „Virtual Private Network”, jest to tunel, przez który płynie ruch w ramach sieci prywatnej za pośrednictwem sieci publicznej (takiej jak Internet), można w nim kompresować lub szyfrować dane dla lepszej jakości czy bezpieczeństwa.

Na poniższym rysunku przedstawiono histogram prezentujący klasyfikację danych na podstawie wyżej wymienionego kryterium, wykonany w programie „Orange”.



Podstawowa statystyka danych:

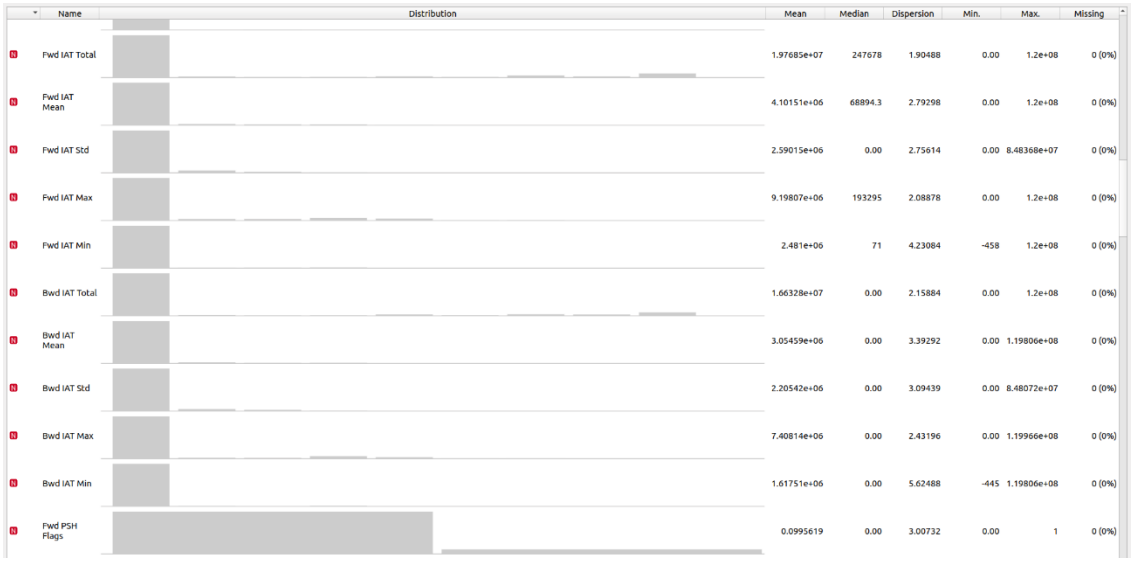
Wykorzystano oprogramowanie „Orange” do wyliczenia dla każdej encji średniej, mediany, dyspersja, minimum, maksimum oraz ilości brakujących wartości w procentach.

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
FWD Init Win Bytes		5308.19	913	1.86421	0.00	65535	0 (0%)
Bwd Init Win Bytes		1766.76	0.00	4.28126	0.00	65535	0 (0%)
Fwd Act Data Pkts		96.8816	1	16.3376	0.00	113325	0 (0%)
Fwd Seg Size Min		15.8082	20	0.449811	0.00	44	0 (0%)
Active Mean		0.00	0.00	∞	0.00	0.00	0 (0%)
Active Std		0.00	0.00	∞	0.00	0.00	0 (0%)
Active Max		0.00	0.00	∞	0.00	0.00	0 (0%)
Active Min		0.00	0.00	∞	0.00	0.00	0 (0%)
Idle Mean		7.02803e+14	7.28125e+14	1.00436	0.00	1.46e+15	0 (0%)
Idle Std		5.52614e+13	0.00	3.49278	0.00	1.03e+15	0 (0%)
Idle Max		7.30588e+14	1.42773e+15	0.992101	0.00	1.46e+15	0 (0%)

Rysunek 1.

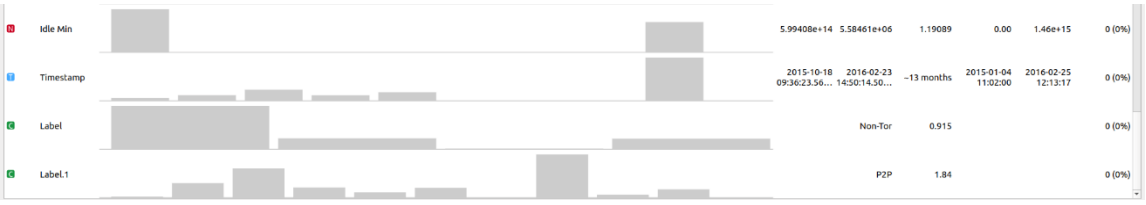
Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Bwd PSH Flags		0.00	0.00	∞	0.00	0.00	0 (0%)
Fwd URG Flags		0.00	0.00	∞	0.00	0.00	0 (0%)
Bwd URG Flags		0.00	0.00	∞	0.00	0.00	0 (0%)
Fwd Header Length		2933.01	40	16.2041	0.00	4.76864e+06	0 (0%)
Bwd Header Length		3099.27	16	22.5703	0.00	9.41724e+06	0 (0%)
Fwd Packets/s		6321.81	4.86987	5.62953	0.00	2e+06	0 (0%)
Bwd Packets/s		4574.67	0.713498	4.67376	0.00	1e+06	0 (0%)
Packet Length Min		14.3656	0.00	1.66728	0.00	1350	0 (0%)
Packet Length Max		323.627	59	2.98498	0.00	64240	0 (0%)
Packet Length Mean		94.374	28.3636	2.01924	0.00	6647.07	0 (0%)
Packet Length Std		108.474	9.2376	1.91419	0.00	12617.1	0 (0%)

Rysunek 2.



Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Fwd IAT Total		1.97685e+07	247678	1.90488	0.00	1.2e+08	0 (0%)
Fwd IAT Mean		4.10151e+06	68894.3	2.79298	0.00	1.2e+08	0 (0%)
Fwd IAT Std		2.59015e+06	0.00	2.75614	0.00	8.48368e+07	0 (0%)
Fwd IAT Max		9.19807e+06	193295	2.08878	0.00	1.2e+08	0 (0%)
Fwd IAT Min		2.481e+06	71	4.23084	-458	1.2e+08	0 (0%)
Bwd IAT Total		1.66328e+07	0.00	2.15884	0.00	1.2e+08	0 (0%)
Bwd IAT Mean		3.05459e+06	0.00	3.39292	0.00	1.19806e+08	0 (0%)
Bwd IAT Std		2.20542e+06	0.00	3.09439	0.00	8.48072e+07	0 (0%)
Bwd IAT Max		7.40814e+06	0.00	2.43196	0.00	1.19966e+08	0 (0%)
Bwd IAT Min		1.61751e+06	0.00	5.62488	-445	1.19806e+08	0 (0%)
Fwd PSH Flags		0.0995619	0.00	3.00732	0.00	1	0 (0%)

Rysunek 3.

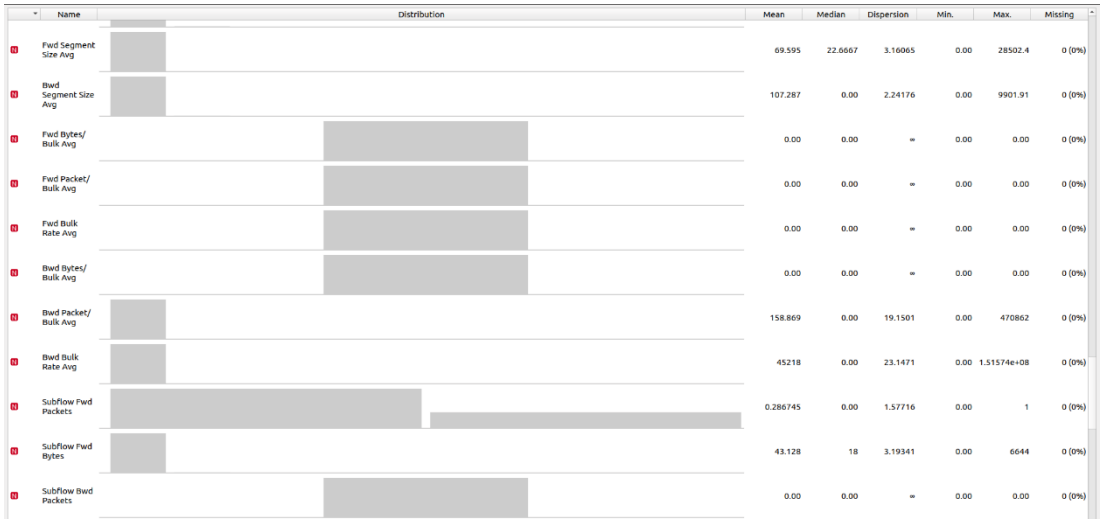


Idle Min		5.99408e+14	5.58461e+06	1.19089	0.00	1.46e+15	0 (0%)
Timestamp		2015-10-18 09:36:23.56...	2016-02-23 14:50:14.50...	~13 months	2015-01-04 11:02:00	2016-02-25 12:13:17	0 (0%)
Label			Non-Tor	0.915			0 (0%)
Label.1			P2P	1.84			0 (0%)

Rysunek 4.

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Packet Length Variance		54881	85.3333	14.4559	0.00	1.59191e+08	0 (0%)
FIN Flag Count		0.455691	0.00	1.22041	0.00	2	0 (0%)
SYN Flag Count		0.462284	0.00	1.82663	0.00	7	0 (0%)
RST Flag Count		0.0385713	0.00	8.88075	0.00	71	0 (0%)
PSH Flag Count		61.5232	0.00	13.1208	0.00	48025	0 (0%)
ACK Flag Count		268.252	2	19.9171	0.00	709023	0 (0%)
URG Flag Count		0.00	0.00	∞	0.00	0.00	0 (0%)
CWE Flag Count		0.00	0.00	∞	0.00	0.00	0 (0%)
ECE Flag Count		0.00	0.00	∞	0.00	0.00	0 (0%)
Down/Up Ratio		0.607214	1	20.5817	0.00	4266	0 (0%)
Average Packet Size		108.9	33	1.81181	0.00	6647.19	0 (0%)

Rysunek 5.

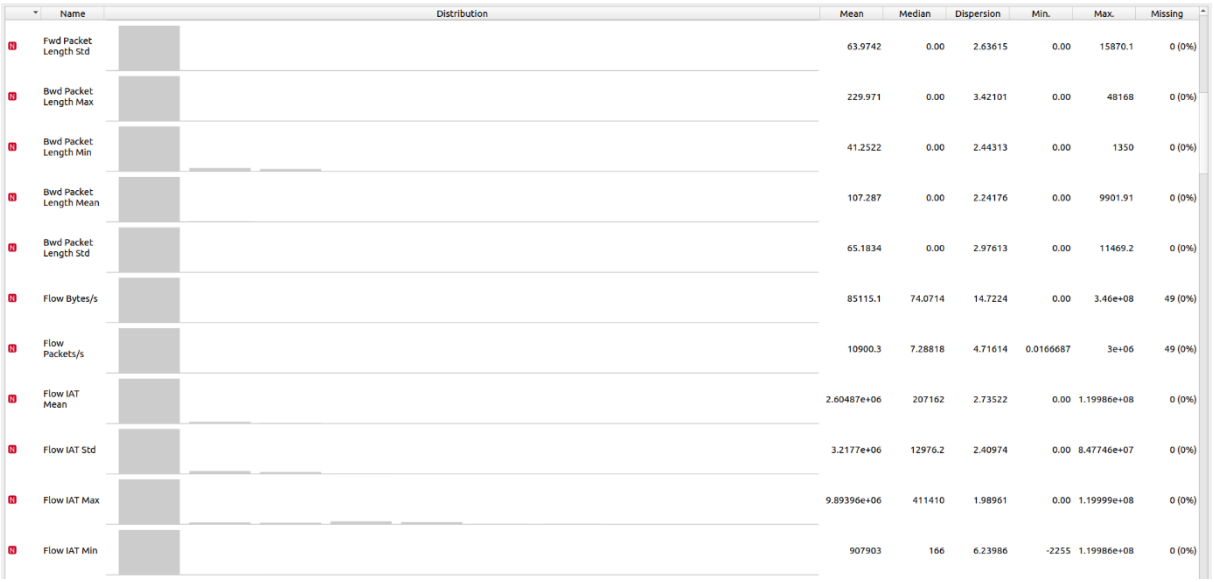


Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Fwd Segment Size Avg		69.595	22.6667	3.16065	0.00	28502.4	0 (0%)
Bwd Segment Size Avg		107.287	0.00	2.24176	0.00	9901.91	0 (0%)
Fwd Bytes/ Bulk Avg		0.00	0.00	∞	0.00	0.00	0 (0%)
Fwd Packet/ Bulk Avg		0.00	0.00	∞	0.00	0.00	0 (0%)
Fwd Bulk Rate Avg		0.00	0.00	∞	0.00	0.00	0 (0%)
Bwd Bytes/ Bulk Avg		0.00	0.00	∞	0.00	0.00	0 (0%)
Bwd Packet/ Bulk Avg		158.869	0.00	19.1501	0.00	470862	0 (0%)
Bwd Bulk Rate Avg		45218	0.00	23.1471	0.00	1.51574e+08	0 (0%)
Subflow Fwd Packets		0.286745	0.00	1.57716	0.00	1	0 (0%)
Subflow Fwd Bytes		43.128	18	3.19341	0.00	6644	0 (0%)
Subflow Bwd Packets		0.00	0.00	∞	0.00	0.00	0 (0%)

Rysunek 6.

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Src Port		38450.3	43528	0.497389	0.00	65534	0 (0%)
Dst Port		18124.6	5355	1.22497	0.00	65535	0 (0%)
Protocol		10.3504	6	0.524789	0.00	17	0 (0%)
Flow Duration		2.08128e+07	416282	1.83019	0.00	1.2e+08	0 (0%)
Total Fwd Packet		152.801	2	15.5648	1	238161	0 (0%)
Total Bwd packets		154.642	1	22.1072	0.00	470862	0 (0%)
Total Length of Fwd Packet		112621	44	28.8698	0.00	7.69307e+08	0 (0%)
Total Length of Bwd Packet		130453	0.00	35.0024	0.00	6.70429e+08	0 (0%)
Fwd Packet Length Max		208.92	34	3.1085	0.00	64240	0 (0%)
Fwd Packet Length Min		15.6171	0.00	2.005	0.00	1350	0 (0%)
Fwd Packet Length Mean		69.595	22.6667	3.16065	0.00	28502.4	0 (0%)

Rysunek 7.



Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Fwd Packet Length Std		63.9742	0.00	2.63615	0.00	15870.1	0 (0%)
Bwd Packet Length Max		229.971	0.00	3.42101	0.00	48168	0 (0%)
Bwd Packet Length Min		41.2522	0.00	2.44313	0.00	1350	0 (0%)
Bwd Packet Length Mean		107.287	0.00	2.24176	0.00	9901.91	0 (0%)
Bwd Packet Length Std		65.1834	0.00	2.97613	0.00	11469.2	0 (0%)
Flow Bytes/s		85115.1	74.0714	14.7224	0.00	3.46e+08	49 (0%)
Flow Packets/s		10900.3	7.28818	4.71614	0.0166687	3e+06	49 (0%)
Flow IAT Mean		2.60487e+06	207162	2.73522	0.00	1.19986e+08	0 (0%)
Flow IAT Std		3.2177e+06	12976.2	2.40974	0.00	8.47746e+07	0 (0%)
Flow IAT Max		9.89396e+06	411410	1.98961	0.00	1.19999e+08	0 (0%)
Flow IAT Min		907903	166	6.23986	-2255	1.19986e+08	0 (0%)

Rysunek 8.

Wybrane parametry:

- **Idle Max** - maksymalny czas bezczynności przepływu przed aktywacją
- **FWD Init Win Bytes** - całkowita liczba bajtów wysłanych w początkowym oknie
- **Idle Mean** - średni czas, w którym przepływ był bezczynny, zanim stał się aktywny
- **Idle Min** – minimalny czas bezczynności przepływu przed aktywacją
- **Fwd Seg Size Min** - minimalny rozmiar segmentu wysyłanego
- **Subflow Fwd Packets** - średnia liczba wysyłanych pakietów w podprzepływie
- **Flow Duration** - liczba pakietów
- **Flow IAT Max** – maksymalny czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Flow IAT Min** – minimalny czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Flow IAT Mean** - średni czas pomiędzy dwoma pakietami wysłanymi w przepływie
- **Fwd Packets/s** - liczba pakietów wysyłanych na sekundę
- **Flow Packets/s** – liczba przepływu pakietów na sekundę
- **Bwd Init Win Bytes** - całkowita liczba bajtów odbieranych(zwracanych) w początkowym oknie
- **Protocol** - numer protokołu
- **FIN Flag Count** - liczba pakietów z FIN
- **Bwd Packets/s** - liczba pakietów otrzymanych (zwróconych) na sekundę
- **Fwd IAT Max** – maksymalny czas pomiędzy dwoma wysłanymi pakietami

- **Bwd Packet Length Mean** - średnia wartość otrzymanych (zwróconych) pakietów
- **Bwd Packet Length Min** – minimalny rozmiar otrzymanych (zwróconych) pakietów
- **Fwd IAT Total** - całkowity czas między dwoma wysłanymi pakietami
- **Label1** - stosowane szyfrowanie
- **Label2** – kategoria ruchu

Kroki wybierania danych:

Do wybrania encji wykorzystano poniższe artykuły:

- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9514531&fbclid=IwAR1OYH1KRD24uK9x3YV2Wr36qvcR8UDa82wWoiegOz2VJuF-C4yIhjsVkVc
- https://arxiv.org/ftp/arxiv/papers/2102/2102.08411.pdf?fbclid=IwAR04gmmR6_rdJiu0vTspwZofh5U2ONQHnaC1UNisq3EDWyM0M3BWUYq47JI

1. Usunięcie wierszy z brakującymi informacjami.

Bwd Packet Length Std	Flow Bytes/s	Flow Packets/s	Flow IAT Mean	Flow IAT Std
0	-	3273.322	611	NaN
0	0	1904.762	1050	-
NaN	NaN	Infinity	0	0
0	536.9445	Infinity	186239	0
0	0	5494.505	Infinity	0
NaN	NaN	3.0371	658523	NaN

Rysunek 3. Przykładowe niekompletne dane.

2. Wykorzystanie poniższych algorytmów do wybrania encji:
    - Principal component analysis (PCA)
    - Decision Tree Classifier
    - XGBClassifier
    - Gradient Boosting Classifier (GB)
    - Random Forest Regressor (RFR)
    - CNN-LSTM

    Powyższe algorytmy służyły do skategoryzowania danych na podstawie zastosowanego szyfrowania.

Wnioski:

Rysunek 1:

- Encje Active Mean, Active Std, Active Max i Active Min są jednolite dla wszystkich krotek i równe zero, dlatego nie są użyteczne w kontekście analizy bazy danych.

Rysunek 2:

- Encje Bwd PSH Flags, Fwd URG Flags oraz Bwd URG Flags są jednolite dla wszystkich krotek i równe zero, dlatego nie są użyteczne w kontekście analizy bazy danych.

Rysunek 5:

- Encje URG Flag Count, CWE Flag Count oraz ECE Flag Count są jednolite dla wszystkich krotek i równe zero, dlatego nie są użyteczne w kontekście analizy bazy danych.

Rysunek 6:

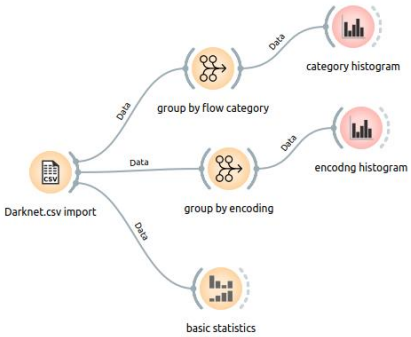
- Encje Fwd Bytes/Bulk Avg, Fwd Packet/Bulk Avg, Fwd Bulk Rate Avg oraz Bwd Bytes/Bulk Avg są jednolite dla wszystkich krotek i równe zero, dlatego nie są użyteczne w kontekście analizy bazy danych.

Rysunek 8:

- Procent brakujących wartości w encjach Flow Bytes/s oraz Flow Packets/s jest zbyt wysoki (ok. 49%), co powoduje, że dane nie mogą być uznane za wiarygodne.

Aby zbadać dokładniejszy rozkład pozostałych wartości należałoby usunąć wartości skrajne. Z histogramów można wywnioskować, że stanowią one bardzo mały procent danych.

Schemat z oprogramowania „Orange”:



Do sprawdzenia działania algorytmów opisanych w artykułach wykorzystano Google Colaboratory, notebook dostępny pod linkiem:

[Pomiary.ipynb - Colaboratory (google.com)]

Wykorzystano język Python oraz między innymi biblioteki:

- *pandas* - do wczytania i analizy danych
- *sklearn* - do stworzenia drzewa decyzyjnego oraz *random forest*
- *matplotlib* - do wygenerowania rysunku drzewa decyzyjnego

Wczytanie pliku csv do DataFrame z biblioteki *pandas:*

```
darknet_all_df = pd.read_csv("Darknet.csv", header=0)

darknet_all_df.head()
```

Wybór kolumn do analizy:

```
chosen_columns = [

    "Idle Max",

    "FWD Init Win Bytes",

    "Idle Mean",

    "Idle Min",

    "Fwd Seg Size Min",

    "Subflow Fwd Packets",

    "Flow Duration",

    "Flow IAT Max",

    "Flow IAT Min",

    "Flow IAT Mean",

    "Fwd Packets/s",

    "Flow Packets/s",

    "Bwd Init Win Bytes",

    "Protocol",

    "FIN Flag Count",

    "Bwd Packets/s",
```

```
    "Fwd IAT Max",
    "Bwd Packet Length Mean",
    "Bwd Packet Length Min",
    "Fwd IAT Total",
    "Label",
    "Label.1"
]
```

Zastąpienie *Label Non-VPN, Non-VPN, Non-Tor* etykietą *Non-Darknet*:

```
darknet_df['Label'].replace(to_replace=['Tor', 'VPN'], value='Darknet',
inplace=True)

darknet_df['Label'].replace(to_replace=['Non-Tor', 'Non-VPN',
'NonVPN'], value='Non-Darknet', inplace=True)
```

Funkcja sprawdzająca czy wiersze zawierają brakujące wartości i użycie jej na datasecie (wyeliminowano 49 wierszy).

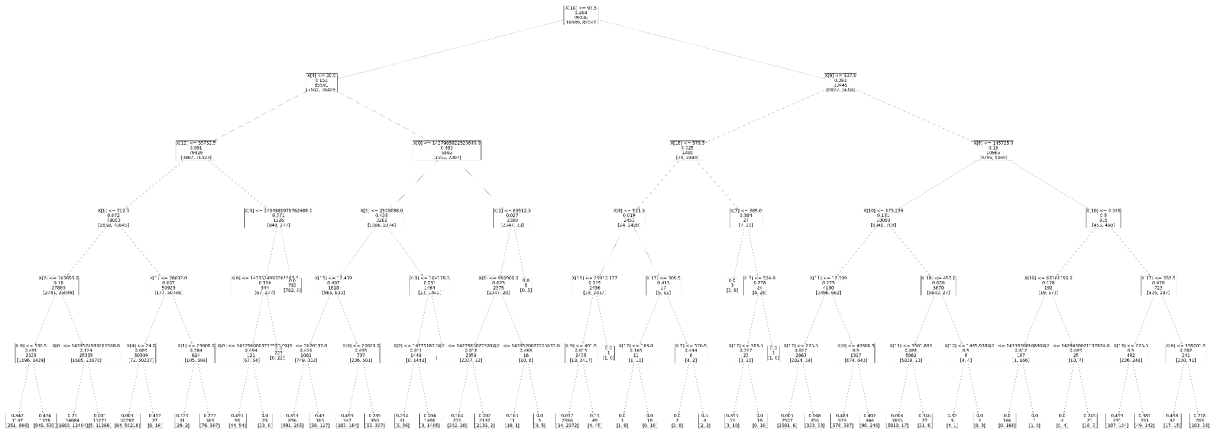
```
def hasMissingValues(row):
    for column, val in row.iteritems():
        if pd.isnull(val) or val == '' or (not isinstance(val, str) and
m.isinf(val)):
            return True
    return False

darknet_df = darknet_df[darknet_df.apply(lambda row: not
hasMissingValues(row), axis=1)]
```

Podział datasetu na części treningową i testową (proporcje 70:30):

```
train_X, test_X, train_Y, test_Y = train_test_split(X, Y,
test_size=0.3)

clf = DecisionTreeClassifier(splitter='best', max_depth = 6,
random_state=0)

clf.fit(train_X, train_Y)
```

```
plt.figure(figsize=(250, 100))

tree.plot_tree(clf, fontsize=50,label="none")
```

Powstałe drzewo decyzyjne ograniczono do 6 poziomów głębokości:



Pierwszy podział następuje dla “Bwd Packet Length Min” po wartości 97,5.

Od głównego wierzchołka odchodzą dwa "Fwd Seg Size Min" i "Flow IAT Mean" dzieląc odpowiednio po wartościach 30 i 537.