



SPRAWOZDANIE 2

Pomiary, analiza i modelowanie systemów internetowych



WERONIKA SIERACKA
JOANNA MIELNICZUK
PIOTR OBERSKI
MACIEJ WOJTKO
KAMIL POCZĄTEK

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 wysłanych 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 bajtów 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 wysłanego
- **Active Mean** - średni czas, w którym przepływ był aktywny przed przejściem w stan beczynności
- **Active Std** - odchylenie standardowe czasu, w którym przepływ był aktywny przed przejściem w stan beczynności
- **Active Max** - maksymalny czas, przez który przepływ był aktywny przed przejściem w stan beczynności
- **Active Min** – minimalny czas, przez który przepływ był aktywny przed przejściem w stan beczynności
- **Idle Mean** - średni czas, w którym przepływ był beczynny, zanim stał się aktywny
- **Idle Std** - odchylenie standardowe czasu, w którym przepływ był beczynny przed aktywacją
- **Idle Max** - maksymalny czas beczynności przepływu przed aktywacją
- **Idle Min** – minimalny czas beczynności przepływu przed aktywacją

Częstotliwość zbierania danych:

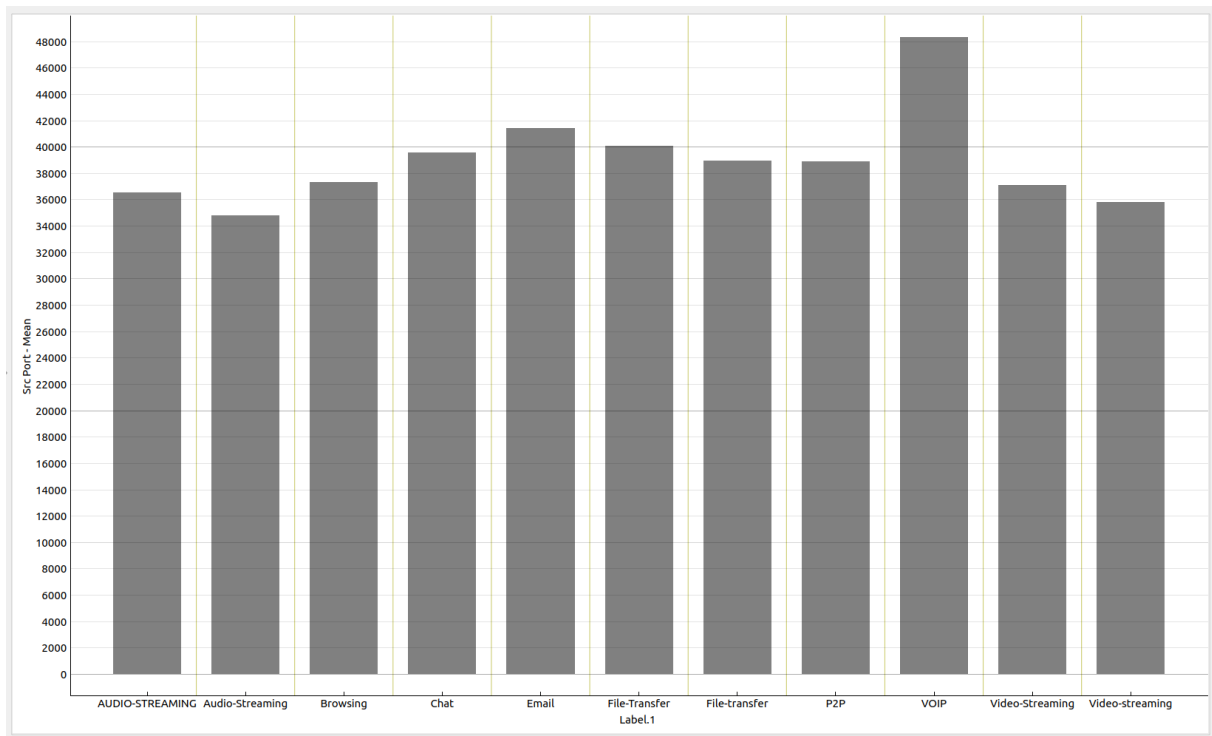
Baza danych zawiera informacje zbierane w sposób ciągły przez ponad rok. Pierwsze rekordy pochodzą z 4 stycznia 2015r., a ostatnie z 25 lutego 2016r.

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”.



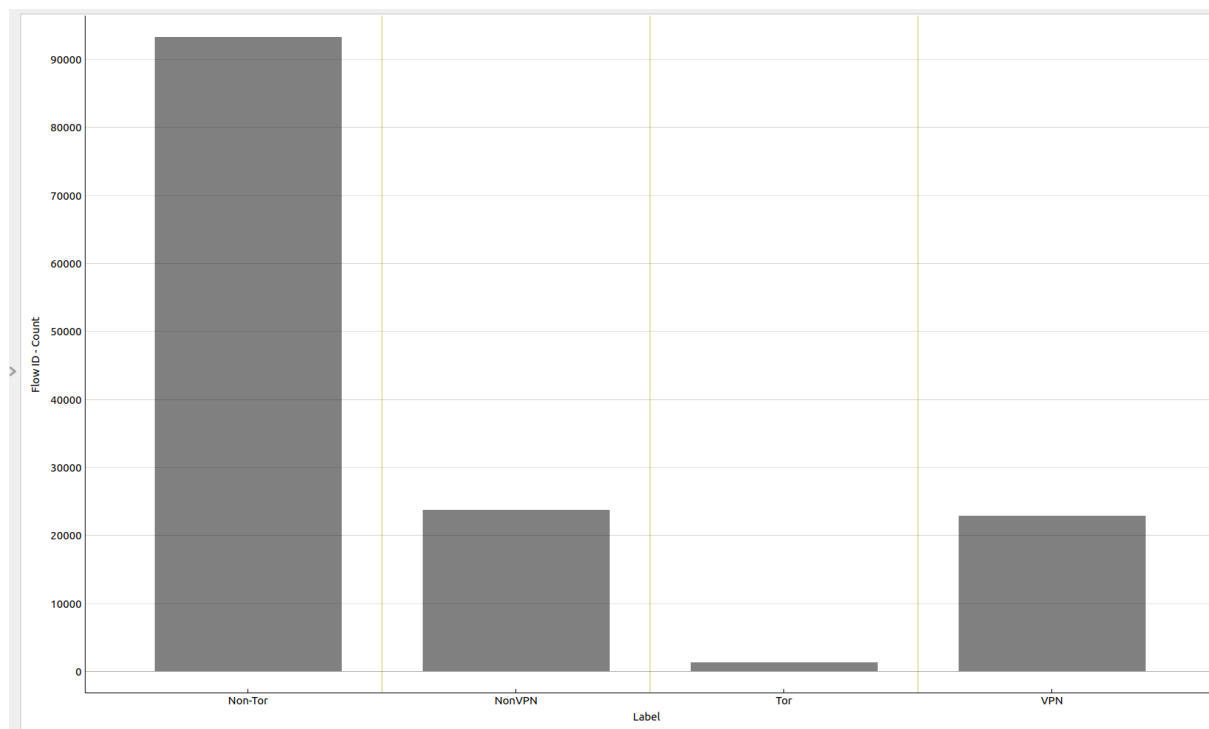
Jak możemy zauważyć największą ilość danych zebrano o VoIP, a najmniejszą o Audio-Streaming.

Dodatkowo – przy etykietowaniu danych wkraść 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”.

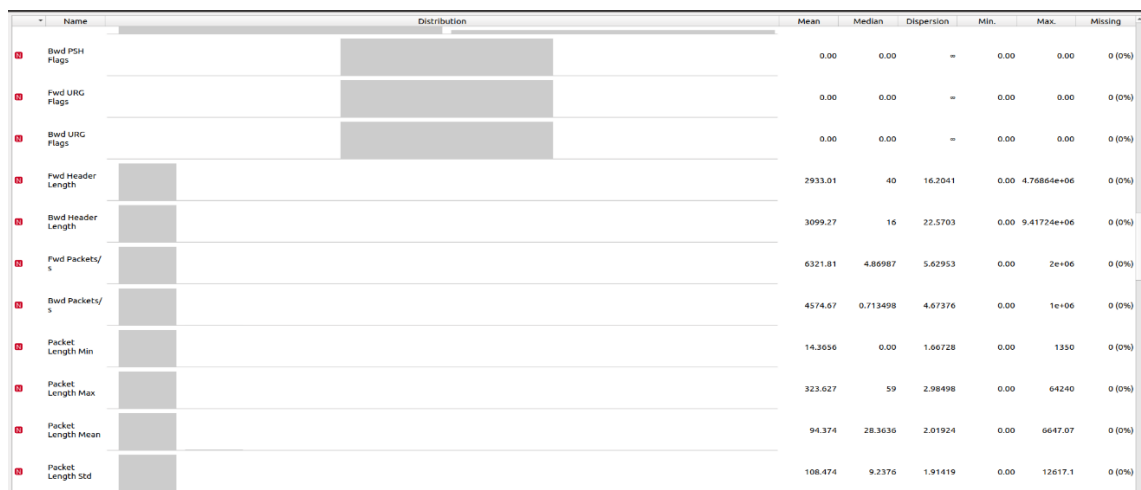


Podstawowa statystyka danych:

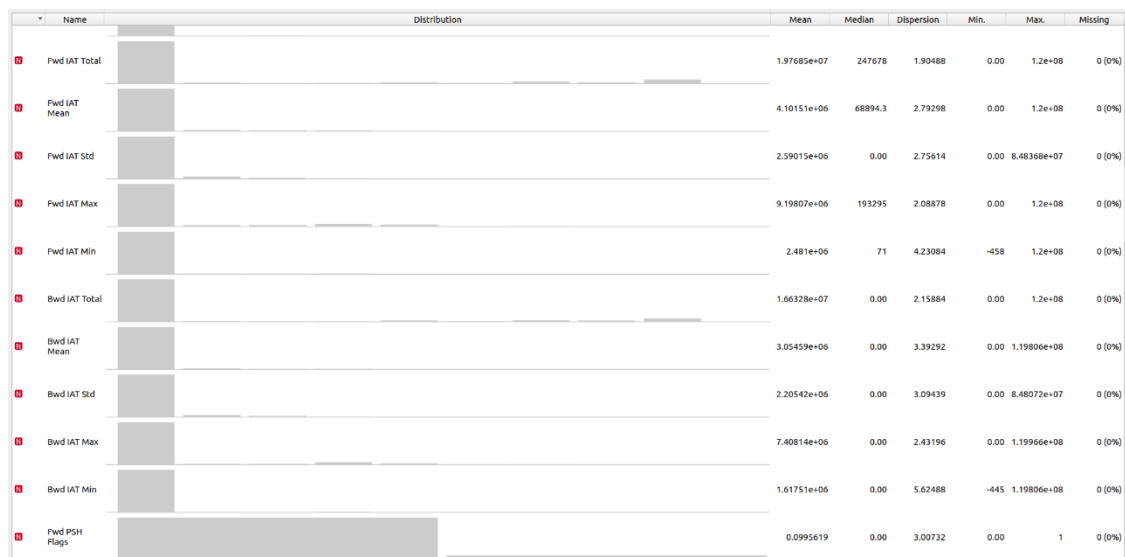
Wykorzystano oprogramowanie „Orange” do wyliczenia dla każdej encji średniej, mediany, dyspersja, minimum, maksimum oraz ilości brakujących wartości w procentach.

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%)

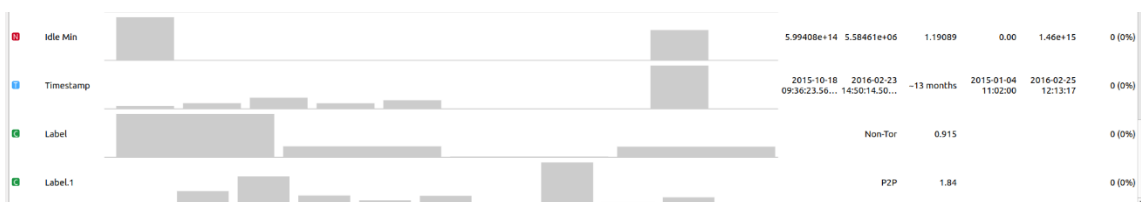
Rysunek 1.



Rysunek 2.



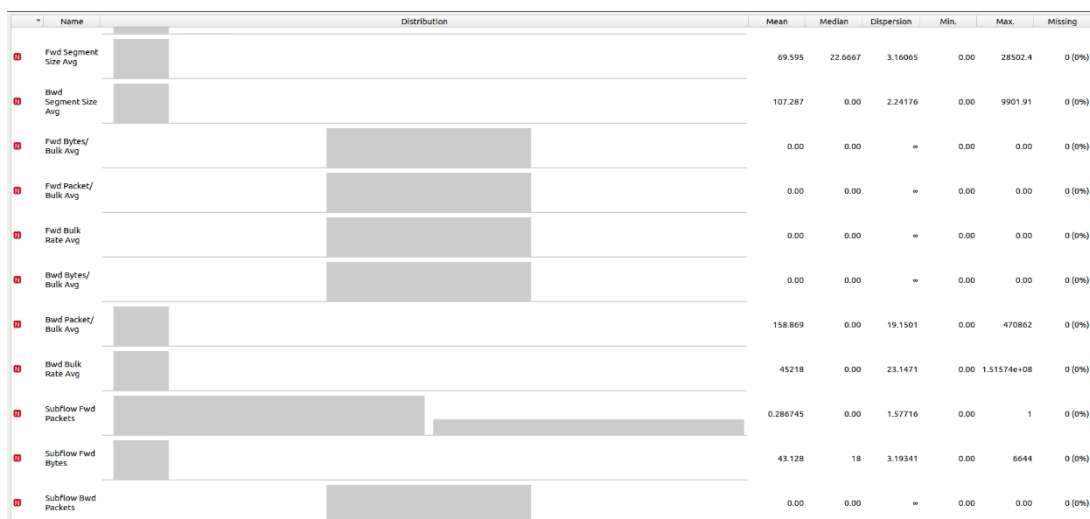
Rysunek 3.



Rysunek 4.



Rysunek 5.



Rysunek 6.



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 wysłanego
- **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-C4yhjsVkVc>
- 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 Pack-ets/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”:

