PDS Projekt – Analýza statistické podobnosti síťových služeb

Vojtěch Fiala

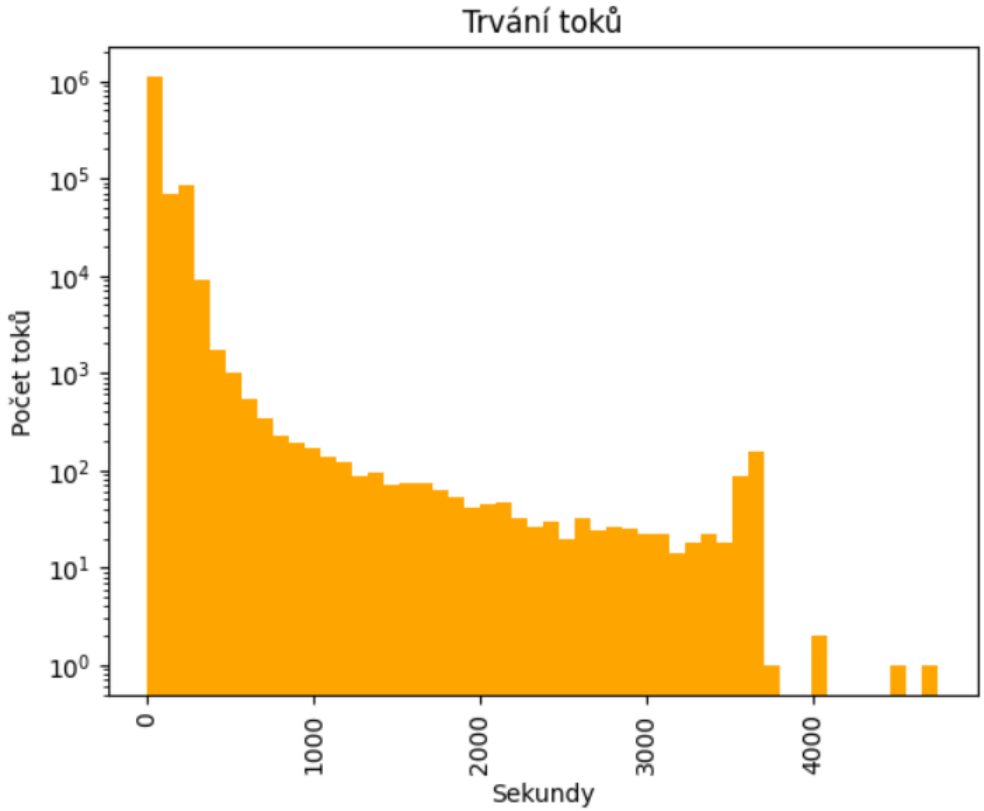
April 22, 2024

Statistická analýza - TLS_SNI

- 1 287 422 řádků
- Každý rádek obsahuje informace o jednom HTTPS toku
- Nejčastěji navštěvovaná adresa je www.google.com – adresy jsou primárně ve formátu URL (bez protokolu).
- Záznam o 19 910 různých adresách
- Průměrně zaznamenáno 64.66 toků na jednu adresu, vysoká směrodatná odchylka
- 83.78% všech unikátních adres serveru se vyskytlo pouze 1-4x. V kontextu všech toků se ale jedná o pouhých 1.75%
- Toky se vytvářely s pár adresami opakovaně, s novými adresami výjimečně
- Nejčastější Top-Level doména je .com – 80.81% všech toků. Vyskytují se i adresy, které TLD nemají (např. 8.8.4.4)
- Nejčastější doména 2. řádu je .google – Google služby tvoří více než 50% všech toků
- Časté toky k DNS službám – DNS over HTTPS

Statistická analýza - TIME

- Dataset pochází z 23. 11. 2023 v době cca mezi 14:40 a 16:10
- Dataset obsahuje toky nasbírané za 1h 27m
- Průměrná doba trvání toku byla 40.66 vteřiny
- 99% toků bylo kratší než 290 vteřin.



Obrázek: Rozložení délky trvání toku

Statistická analýza - TIME

- Množství vznikajících i končících toků se napříč časem mírně snižuje
- Nejvíce času otevřen tok se službou mail.google.com – 21 dní, 9 hodin, 38 minut, 14 sekund
- Průměrná doba přenosu paketu je 0.9 vteřiny. Nejvyšší doba přenosu paketu vychází 16 vteřin (doba toku / počet paketů).

Statistická analýza - PACKETS/BYTES

- Jako upload se počítaly pakety/byty ve směru klient → server
- Jako download pak naopak server → klient
- Průměrný počet paketů na tok je 193.14. Průměrný upload je 67.7 paketu a průměrný download 125.43
- 64.95% paketů bylo ve směru downloadu
- Průměrný počet bytů poslaných ve směru upload je 36 151, průměrný download 140 818
- 79.57% přenesených bytů bylo ve směru downloadu
- Zjištěna korelace mezi počtem paketů a bytů v toku
- 1 paket měl v průměru 373.56 bytů. Nejvíce zaznamenáno 1500 bytů na paket – běžný limit MTU

Statistická analýza - PACKETS/BYTES

- Zdaleka nejaktivnější tok se službou www.googleapis.com – 24 896 311 paketů
- Velmi aktivní toky i na další Google služby, cloudové úschovny souborů a solocoo.tv – přenos televizního vysílání
- www.google.com je z hlediska počtu přenesených paketů velmi vysoko, ale počtu bytů už tolik ne. Naopak www.youtube.com se s přenesenými pakety na první příčky neřadí, ale z pohledu přenesených bytů ano
- www.googleapis.com je nejaktivnější i z hlediska přenesených bytů

Statistická analýza – TCP_FLAGS

- Nejčastější "vlajky"jsou 27 – ACK, PSH, SYN, FIN
- 3.74% toků neukončeno ani FIN, ani RST
- 24.36% ukončeno RST, 14.32% nejdříve FIN, pak i RST
- 71.90% ukončeno FIN
- Přes RST nejčastěji ukončovány DNS služby a www.google.com (a další google služby)
- V tocích končících RST přeneseno 30.2% paketů a 33% bytů

Statistická analýza – PPI

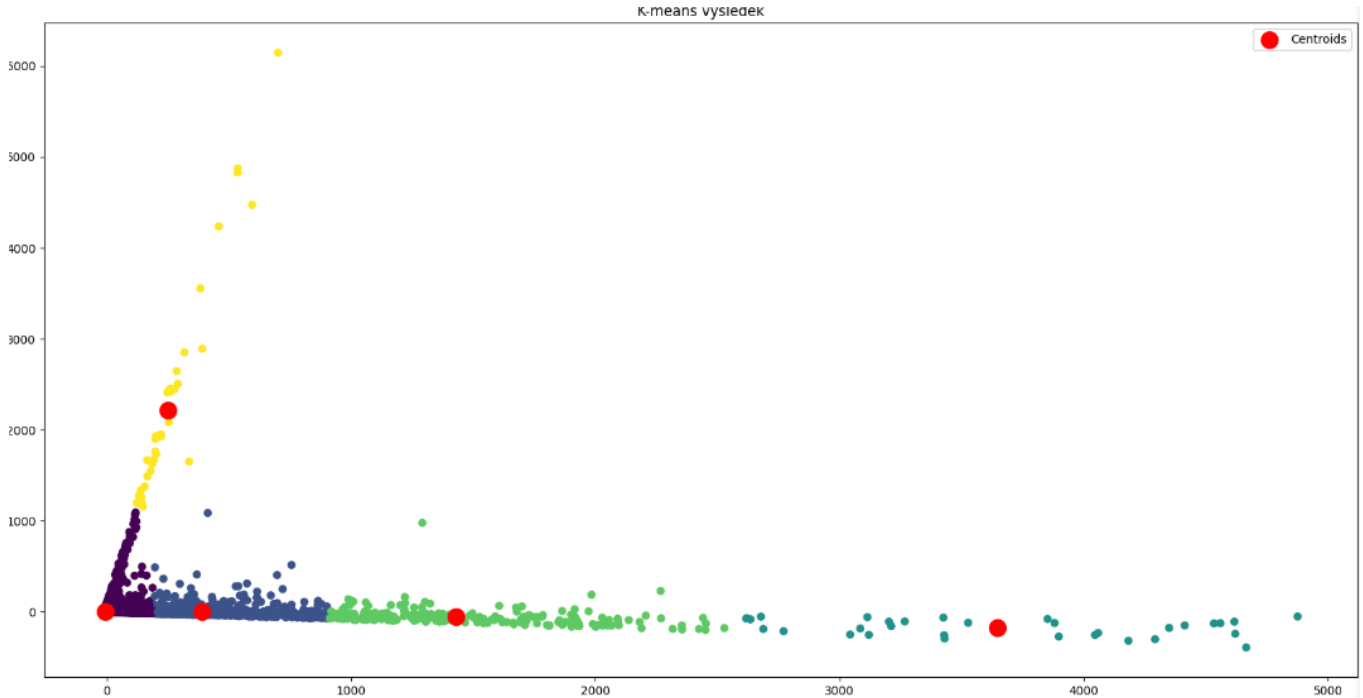
- První paket v každém toku byl téměř vždy SYN
- 4.23% nekorektně zaznamenaný 3-Way Handshake (nesprávně pořadí, RST po SYN...)
- 2x zaznamenaný první paket ACK (avšak validní komunikace – server reagoval)
- 44.99% paketů z prvních 100 paketů toku mělo payload velikost 0 – Signalizační pakety. Z toho 58.44% při uploadu.

Shluková analýza – Předzpracování

- 6 datasetů lišící se atributy. Z toho 1 speciální zaměřený pouze na službu www.youtube.com
- Random sampling na 1/10 původní velikosti, pro DBSCAN a Hierarch. shlukování až na 1/100 z důvodu efektivity
- Odstranění krajních hodnot přes Z-score
- V některých případech odfiltrovány toky na adresu, která se vyskytla méně než 5x
- Redukce dimenzionality na 2 atributy přes PCA

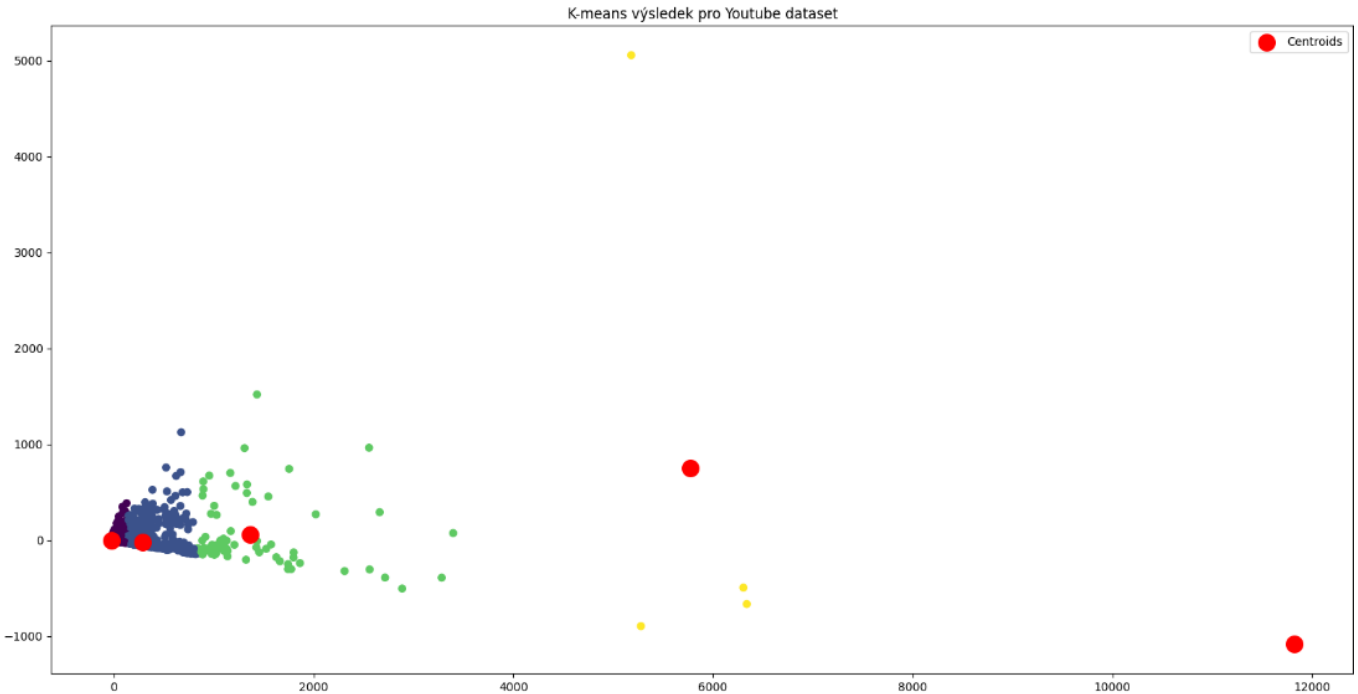
Shluková analýza – K-Means

- Elbow metoda pro určení počtu clusterů – 5 clusterů
- Kvalita vyhodnocena přes Davies Boulding skóre (0 dobré, 1 a dál špatné)
- Nejlepší výsledek 0.456
- Většina (98.95%) hodnot v jednom clusteru. V dalších clusterech primárně přenos souborů (solocoo.tv)



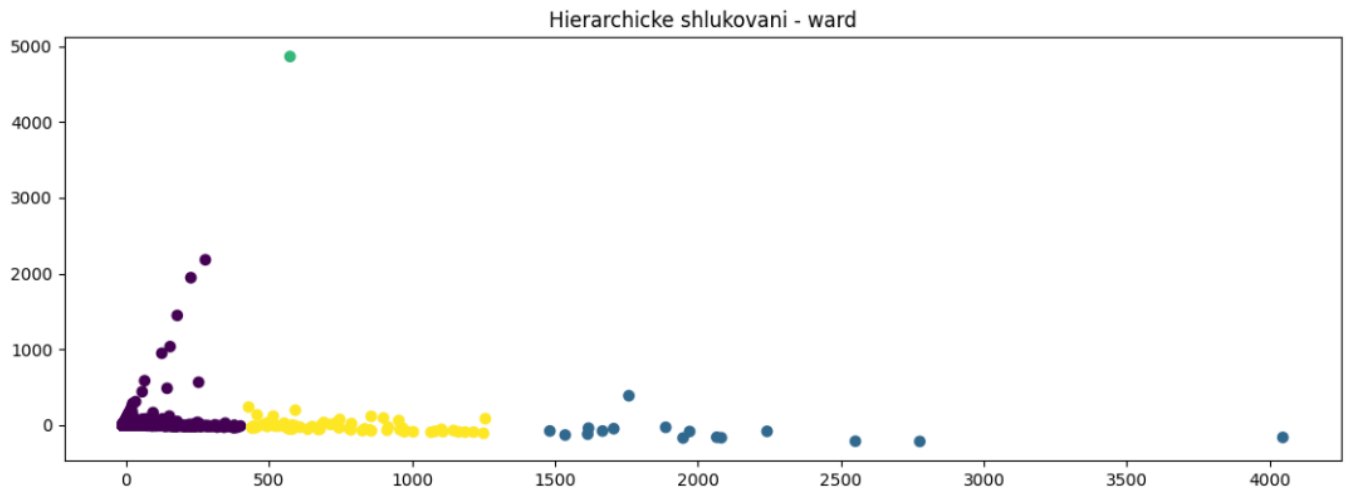
Obrázek: Výsledek K-means shlukování nad obecným datasetem 0

- V Youtube datasetu lepší rozdělení na clustery.
- Clustery se liší hlavně průměrnou délkou trvání toku a počtu přenesených bytů/paketů



Obrázek: Výsledek K-means shlukování nad Youtube datasetem

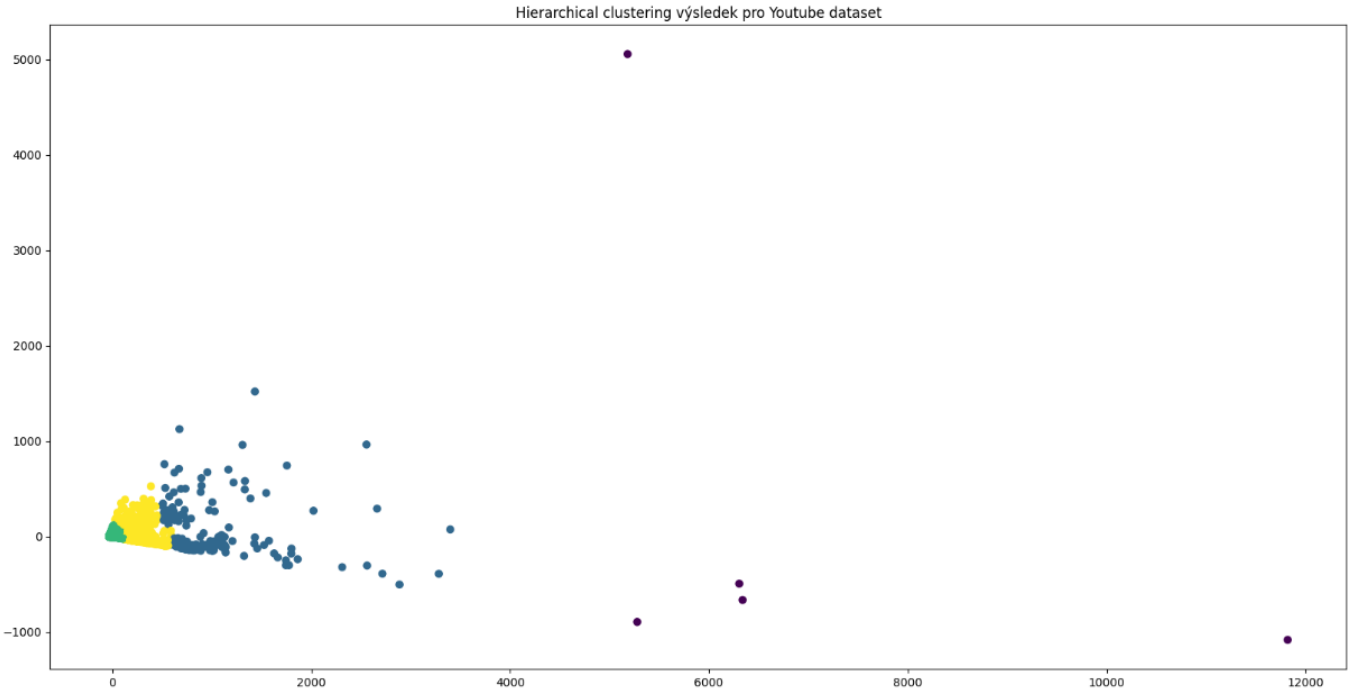
- 4 metody, většina má problém overfittingem, pouze Ward vypadá kvalitně (všechny hodnoty nejsou v pár clusterech).
- 4 clustery
- Vyhodnocení opět přes DB Skóre – nejlepší s metodou Ward je 0.35 – nižší (lepší) oproti K-means
- Ačkoliv lepší DB skóre, 99.37% hodnot v jednom clusteru. V ostatních opět primárně přenos souborů (solocoo...)



Obrázek: Výsledek hierarchického shlukování nad obecným datasetem 2

Hierarchické shlukování – Youtube

- Podobné K-means



Obrázek: Výsledek hierarchického shlukování nad obecným datasetem 2

Shluková analýza – DBSCAN

- Vyhodnocení podle Silhuette skóre. 1 – dobré, -1 – špatné. Nejlepší výsledek 0.643
- V jednom clusteru oproti ostatním pouze 94.22% všech záznamů
- Vytváří mnoho clusterů – pro Youtube dataset více než 40
- Statistiky pro průměry Youtube datasetu zpracovaného DBSCANem se liší od předchozích – jsou nižší

Discussion