VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

IDENTIFIKÁCE MOBILNEJ KOMUNIKÁCIE POMOCOU TLS ODTLAČKOV

SEMESTRÁLNÍ PROJEKT
TERM PROJECT

AUTOR PRÁCE **ANDREJ ZAUJEC**
AUTHOR

VEDOUCÍ PRÁCE **Ing. PETR MATOUŠEK, Ph.D. M.A.**
SUPERVISOR

BRNO 2021

Obsah

Kapitola 1

Úvod

TLS protokol slúži na zašifrovanie komunikácie medzi dvoma bodmi v sieti. V rámci *TCP/IP* balíku protokol leží medzi aplikačnou a transportnou vrstvou. Tento protokol sa snaží predísť odpočúvaniu, manipulácií a falšovaniu správ. Pred ustanovením šifrovaného kanálu si musia klienti medzi sebou spraviť *TLS* handshake(podanie ruky). Medzi najzaujímavejšie správy, ktoré sú posielané patria *ClientHello* a *ServerHello* a to z dôvodu obsahu týchto správ. V týchto správach sú informácie ohľadom toho aké šifry preferuje odosielateľ, aké podporuje skupiny, akú verziu *TLS* protokolu používa a mnoho ďalšieho. Po uskutočnení handshaku sa ďalšia komunikácia javí už len v zašifrovanej podobe. V dnešnej dobe je TLS protokol majoritne využívaný na bezpečné prenášanie *HTTP* protokolu. Servery, ktoré podporujú *HTTP* protokol šifrovaný skrz *TLS* počúvajú na rezervovanom porte 443.

Hlavným cieľom tohoto projektu je zozbierať TLS odtlačky mobilných aplikácia a na základe mnou zvolenej metódy a atribútov z odtlačov odhadnúť, ktorá mobilná aplikácia práve komunikoval skrz daný šifrovaný kanál. Taktiež súčasťou projetu je popis získania a spracovanie odtlačkov, vyhodnotenie a diskusia ohľadom mnou zvolenej metódy.

Kapitola 2

Analýza problému a popis použitej metódy

Medzi najznámejšie metódy vytvárania odtlačok *TLS* spojenia patrí metóda s názvom *JA3*. Táto metóda vytvára message digest(prehlad správy) pomocou hašovacieho algoritmu *MD5*. Samotná správa pred zahašovaním obsahuje verziu *TLS*, podporované šifry, rozšírenia, podporované skupiny, formát Eliptických kriviek. Rozlišujeme medzi *JA3* a *JA3S*, kedy v prípade *JA3* sú získané atribúty z *ClientHello* správy a v prípade *JA3S* sú tieto atribúty získané zo *ServerHello*. Kombináciou *JA3* a *JA3S* vieme veľmi slušne rozlišovať aké aplikácie komunikujú i keď stále je tu možnosť zlepšenia. Veľmi často sa pre lepšenie rozlíšteľnosti jednotlivých komunikácií ešte využívaný atribút *SNI(Server Name Indication)*[1]. Tento atribút pochádza zo *ClientHello* správa a obsahuje indikáciu *hostnamu* serveru s ktorým sa chystá klient komunikovať.

Ja som si zvolil v tomto projekte ako spôsob vytvárania odtlačok poslednú spomínanú možnosť a to je identifikácia pomocou trojice *JA3*, *JA3S*, *SNI*. Samotnú implementáciu vytvárania JA3 som prevzal z oficiálneho repozitáru[1]. Následne som upravil daný skript, aby dokázal získať informáciu *SNI*.

[1] https://github.com/salesforce/ja3

Kapitola 3

Získanie a spracovanie datasetu

Sieťová komunikácia bola získaná z fyzického telefónu iPhone X, ktorý mal naištalovaný softvér *iOS* vo verzií 14.0.1. Pomocou nástroja rvictl[1], ktorým disponuje operačný systém *MacOS* bol vytvorený virtuálny interface na ktorom bolo možné odpočúvať komunikáciu telefónu. Na nahrávanie samotnej komunikácie a ukladanie do súborov formátu *pcap* som použil nástroj Wireshark[2]. Na samotné spracovanie *pcap* súborov som si upravil oficiálny *JA3* skript. Odpočúvanie jednotlivých aplikácií prebiehla nasledovne. Každá aplikácia bežala priemerne tri minúty, kde po spustení časovača som spustil danú aplikácia a následne som klikal na všetky možné prvky, ktoré sa dali aktivovať aby som vynútil čo najviac komunikácie. Počas tohoto klikania som aplikáciu aspoň päťkrát vypol a znovu spustil.

Zoznam sledovaných aplikácií je v tabulke 3.1. Táto tabulka obsahuje celkový počet odtlačkov, ktorý bol zaznamenané počas sledovanie behu aplikácia a následne obsahuje, koľko z týchto odtlačkov bolo považovaných ako za jedinečné pre danú aplikáciu. Toto rozhodnutie jedinečnosti, že daný odtlačok patrí danej aplikáciu je založená na tom, že daný odtlačok obsahoval ako podreťazec kľúčové slovo vo svojom *SNI* atribúte.

Meno aplikácie	**Počet zachytených odtlačkov celkovo**	**Počet odtlačkov danej aplikácie**
TikToK	183	37
Instagram	29	5
Messenger	7	3
Gmail	222	25
KalorickeTabulky	109	9
Binance	132	21
Blockfolio	79	9
Netflix	339	68
Medium	202	8

Tabuľka 3.1: Zvolené aplikácie v datasete a ich jednotlivé počty odtlačkov

Kontrola týchto kľúčových slov 3.1 je automatizovaná pričom výber týchto kľúčových slov prebehol ručne po nahliadnutý na získané odtlačky. Podstatou tohoto ručného výberu bolo dobre zohľadniť typické slová, ktoré sa často vyskytujú v *hostnamoch* serverov na ktoré sa aplikácie primárne dotazujú. Vhodne zvolené kľúčové slová následne dokážu pekne

[1] https://developer.apple.com/documentation/network/recording$_{a}$ $_{p}$$a$$c$$k$$e$$t$$_{t}$$r$$a$$c$$e$

[2] https://www.wireshark.org/

odlíšiť jednotlivé *SNI* atribúty a poukázať či sa nejedná o komunikáciu, ktorá sa vyskytuje veľmi často,ale netýka sa aplikácie priamo a spôsobuje šum v dátach. Príkladom takejto komunikácie može byť dotazovanie sa na reklamný server, volania operačného systému ohľadom synchronizácie s cloudom alebo získanie obsahu z dobre známych CDN(Content Delivery Network), ktoré používa veľmi veľa aplikácií na rýchle doručenie obsahu namiesto získavania daného obsahu z vlastnej siete.

Všetky odtlačky, ktoré vyhovovali spomínaným kritériam boli uložené do databázy odtlačkov, ktorú reprezentuje súbor vo formáte *csv*. Pričom hlavička tohoto súboru je ja3,ja3s,sni,app_name.

```
keywords = {
    "instagram": ["instagram","graph.facebook","cdninstagram",],
    "tiktok": ["tiktokcdn","tiktokv" , "tiktok",],
    "messenger": ["web.facebook", "fbcdn"],
    "gmail": [
        "googleusercontent",
        "googleapis",
        "mail.google",
        "inbox.google",
        "www.google",
    ],
    "blockfolio": ["api.blockfolio","blockfolio","cointelegraph",
    "coindesk",
    "cryptobriefing",
    ],
    "netflix": ["netflix","nflxso","nflxvideo",],
    "binance": ["binance", "bnbstatic", "hanqiweb", "shyqxxy", "riskified"],
    "twitter": ["twitter", "twimg",],
    "medium": [ "medium",],
    "kaloricketabulky": [ "kaloricketabulky",],
}
```

Výpis 3.1: Zvolené kľučové slová pre jednotlivé aplikácie

Kapitola 4

Testovanie a vyhodnotenie dátovej sady

Vyhodnotenie získanej databázy prebehlo pomocou testovacej sade odtlačkov, ktorá bola tentoraz vytvorená striedaním rôzneho poradia aplikácií, ktoré boli v spomínané v tabulke 3.1. Tentokrát avšak zachytávanie komunikácie prebiehalo asi 5 minút a počas týchto minút boli spomínané aplikácie v náhodnom poradí púštané a vypínané a následne aj náhodne aktivované rôzne prvky v aplikáciach. Táto testovacia sada obsahuje 305 odtlačkov, z ktorých po odfiltrovaní a anotovaní pomocou kľúčových slov zostalo 105. Správne anotovanie jednotlivých odtlačkov a priradenie k aplikácií bolo založené znovu na obsiahnutí kľúčového slova ako podreťazca v atribúte *SNI*.

Klasifikácia do jednotlivých aplikácií je založená na jednoduchom porovnaní celého odtlačku a teda ak daný odtlačok súhlasí s odtlačkom v databáze tak mu je následne pridelené meno aplikácie z databázy. Výsledky tohoto porovnania je možné vidieť na matici zámen 5.1. Kde jednotlívé označenia 0 až 10 odpovedajú tomuto zoznamu tried aplikácií: Instragram, TiktTok, Medium, Gmail, Binance, Netflix, Blockfolio, KalorickeTabulky, Twitter, Messenger, Unknown. Posledná spomínaná trieda Unkwown slúži ako kôš pre odtlačky, ktoré sme nevedeli klasikovať pomocou našej databázy. Výsledné metriky testovania vyšli nasledovne.

$$Precision = 76.50\%$$

$$Recall = 66.87\%$$

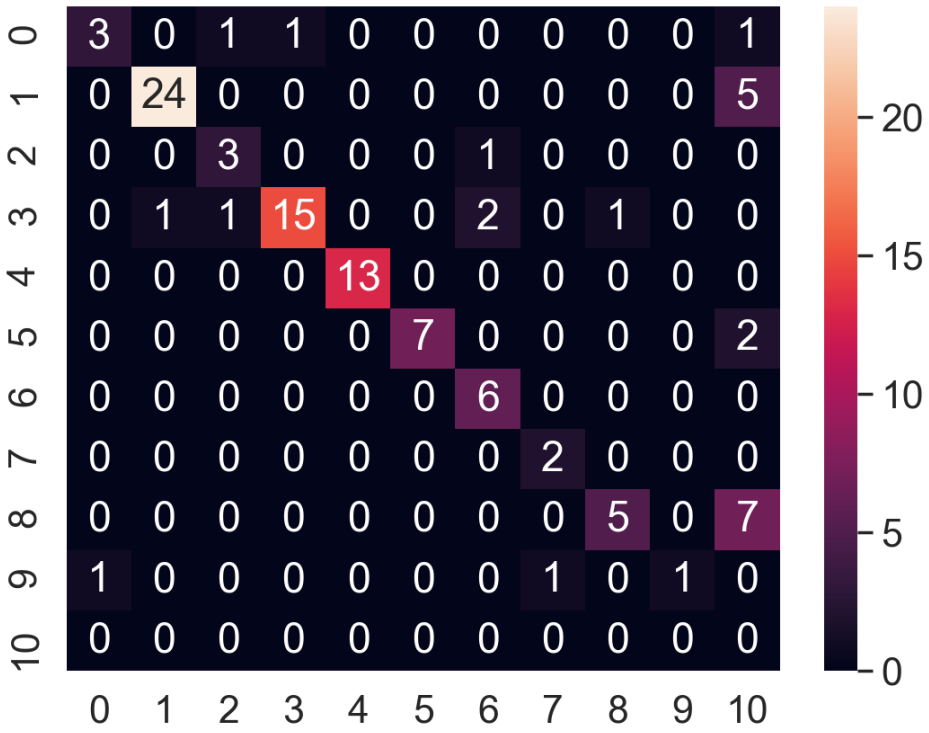
$$Accuracy = 75.96\%$$

Kapitola 5

Záver

Je zrejmé, že použitie trojice *JA3,JA3S,SNI* na určovanie odtlačkov mobilných aplikáci má dobrú úspešnosť a aj to je dôvodom prečo je tak rozšírené a používané. Zlepšenie presnosti tohoto spôsobu určovania odtlačkov v mojom prípade vidím v zväčšaní databázy odtlačkov. Ak sa pozrieme na maticu zámen tak je badateľné, že Twitter bol najmenej úspešne identifikovateľný pričom ostatné aplikácie boli na tom porovnateľne lepšie. Dôvodom slabej presnosti odtlačkov Twitteru bol prefix, ktorý majú zahrnutý v *SNI* atribúte, napríklad *api-43-0-0.twitter.com*. Po prezretí testovacej sady som objavil na každom ich odtlačku vždy iný *SNI* čo značí, že tieto prefixy sú často obmienané a teda v našom prípade ak sme ich neznamenali počas trénovacej fázy všetky tak potom nám dané odtlačky už v testovaní nebudú súhlasiť práve na obmieňanom *SNI* atribúte i keď bude *JA3* a *JA3S* súhlasiť.

Metriky Precision a Recall sa javia trochu nižšie, ale po prezretí matice zámen, je zrejmé, že práve spomínaný Twitter ťahá tieto metriky najviac dole z hladiska priemer. Tento problém by mohol byž vyriešený ako už bolo spomenuté zväčšením databázy odtlačkov. Ostatné odtlačky z pohľadu matice zámen obstáli veľmi dobre. Ďalším problémom tejto metódy na ktorý som ja narazil môže byť určenie odtlačku aplikáciam, ktoré majú spoločného majiteľa a to v prípade Instagramu a Messengeru kedže obe vlastní Facebook. Uprostred komunikáciá oboch aplikácií je časté badať využívanie rovnakých odtlačkov pričom celkový kontext naznačuje niečo iné. Preto ďalším zlepšením tejto metódy by mohlo byť aj pozeranie sa na odtlačky, ktoré predchádzali a použiť túto informáciu pri rozhodovaní ohľadom správnej aplikácie.



Obr. 5.1: Matica zámen

Literatúra

[1] Matoušek, P., Burgetová, I., Ryšavý, O. a Victor, M. On Reliability of JA3 Hashes for Fingerprinting Mobile Applications. In: *Digital Forensics and Cyber Crime. ICDF2C 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering.* Springer International Publishing, 2021, sv. 351, s. 1–22. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. DOI: 10.1007/978-3-030-68734-2_1. ISBN 978-3-030-68733-5. Dostupné z: https://www.fit.vut.cz/research/publication/12307.