VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA SÍŤOVÉ KOMUNIKACE PRO ÚČELY PROFILOVÁNÍ MOBILNÍCH APLIKACÍ

NETWORK COMMUNICATION ANALYSIS FOR MOBILE APPLICATION PROFILING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **MATEJ MELUŠ**
AUTHOR

VEDOUCÍ PRÁCE **Ing. IVANA BURGETOVÁ, Ph.D.**
SUPERVISOR

BRNO 2020

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav informačních systémů (UIFS) Akademický rok 2019/2020

Zadání bakalářské práce

Student: **Meluš Matej**
Program: Informační technologie
Název: **Analýza síťové komunikace pro účely profilování mobilních aplikací**
Network Communication Analysis for Mobile Application Profiling
Kategorie: Data mining

Zadání:

1. Seznamte se s nástroji Android studio a Wireshark.
2. Po dohodě s vedoucí vyberte mobilní aplikace, kterými se budete dále zabývat (min. 2 aplikace)
3. Pomocí výše uvedených nástrojů vytvořte datovou sadu, která bude obsahovat síťovou komunikaci vybraných mobilních aplikací.
4. Vytvořenou datovou sadu vhodným způsobem předzpracujte pro další použití.
5. Po dohodě s vedoucí vyberte vhodnou metodu pro vytvoření profilu mobilních aplikací na základě jejich síťové komunikace.
6. Zvolenou metodu implementujte a otestujte na vytvořené datové sadě.
7. Zhodnoťte dosažené výsledky.

Literatura:

- Matoušek, Petr: Síťové aplikace a jejich architektura. Brno: VUTIUM, 2014, 396 s., ISBN 978-80-214-3766-1.
- Ajaeiya, G., Elhajj, I.H., Chehab, A. et al.: Mobile Apps Identification Based on Network Flows, Knowl Inf Syst (2018) 55: 771, https://doi.org/10.1007/s10115-017-1111-8
- Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Third Edition. Morgan Kaufmann Publishers, 2012, 703 p., ISBN 978-0-12-381479-1

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Burgetová Ivana, Ing., Ph.D.**
Vedoucí ústavu: Kolář Dušan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2019
Datum odevzdání: 14. května 2020
Datum schválení: 24. října 2019

Abstrakt

Cieľom práce bolo vytvoriť profily mobilných aplikácií na základe ich sieťovej komunikácie, ktorá je obsiahnutá vo vytvorených dátových sadách. Za týmto účelom bol vytvorený nástroj, ktorý dokáže tieto profily z dátových sád extrahovať. Profily obsahujú JA3 odtlačky, IP adresy a doménové mená, ktoré sú na základe počtu výskytov v dátových sadách asociované s konkrétnou aplikáciou. Neupravené profily popisujú dôležitosť jednotlivých identifikátorov, ktoré aplikácia pri sieťovej komunikácií využíva. Pri upravení profilov je možné ich využiť pri identifikácii aplikácie v sieťovej prevádzke.

Abstract

Goal of this thesis was to create profiles of mobile applications based on their network communication, which is stored in created data sets. For this purpose, a tool that can extract profiles from data sets was created. Profiles contain JA3 fingerprints, IP addresses and hostnames, that are associated with specific application based on number of occurrences in data sets. Unmodified profiles describe importance of individual identifiers used by application in network communication. Modified profiles can be used to identify application in network traffic.

Kľúčové slová

profilovanie, sieťová komunikácia, analýza sieťovej komunikácie, JA3 odtlačok, IP adresa, doménové meno, mobilné aplikácie, dolovanie z dát, identifikácia mobilných aplikácií, Python

Keywords

profiling, network communication, network communication analysis, JA3 fingerprint, IP address, hostname, mobile applications, data mining, mobile application identification, Python

Citácia

MELUŠ, Matej. *Analýza síťové komunikace pro účely profilování mobilních aplikací*. Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Ivana Burgetová, Ph.D.

Analýza síťové komunikace pro účely profilování mobilních aplikací

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pani Ing. Ivany Burgetovej Ph.D. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal.

.....
Matej Meluš
27. mája 2020

Poďakovanie

Rád by som poďakoval vedúcej práce, pani Ing. Ivane Burgetovej Ph.D, za odbornú pomoc, rady a ochotu pri riešení práce. Taktiež by som rád poďakoval svojej rodine za podporu.

Obsah

Kapitola 1

Úvod

Na prvý pohľad sieťová komunikácia môže pôsobiť zmätočne a zložito, no všetka komunikácia na sieti podlieha dohodnutým pravidlám. Vďaka týmto pravidlám je možné počítače spájať a vymieňať dáta medzi nimi. Pochopením týchto pravidiel je možné porozumieť tomu, čo za dáta sú v sieti odosielané a prijímané. Na základe toho je možné odvodiť, aká komunikácia sa na sieti odohráva.

Dolovanie z dát (*data mining)* je proces získavania netriviálnych informácií z dátových sád, ktoré môžu byť užitočné. Na získavanie takýchto netriviálnych informácií sú používané viaceré metódy, medzi ktoré patrí aj štatistika či strojové učenie. Využitie informácií, ktoré sú získané týmto spôsobom sú rôzne a závisia od toho o aké dáta sa jedná. V rámci tejto práce sa jedná o získavanie informácií zo sieťovej komunikácie, z čoho sa predpokladá využitie týchto informácií v príbuznej oblasti.

Cieľom tejto práce je získanie informácií, ktoré budú popisovať mobilnú aplikáciu na základe sieťovej komunikácie. Ide o spojenie znalostí počítačových sietí a dolovanie z dát. Získané informácie predstavujú profil mobilnej aplikácie. Pod pojmom profil mobilnej aplikácie je v tomto prípade možné rozumieť združenie informácií, ktoré popisujú chovanie konkrétnej aplikácie na úrovni siete. Tým, že je väčšina komunikácie mobilnej aplikácie šifrovaná, nie je na profilovanie vhodná. Preto sa pri profilovaní berú do úvahy hlavne informácie, ktoré predchádzajú tejto šifrovanej komunikácii. Súčasťou práce je taktiež vytvorenie dát, z ktorých sú informácie pre profily získavané. Tieto dáta predstavujú záznamy sieťovej komunikácie, ktorá obsahuje komunikáciu mobilnej aplikácie. V rámci tejto práce bola taktiež testovaná využiteľnosť profilov, teda efektivita konkrétnych postupov získavania informácií z dátových sád.

Práca je členená do ôsmych kapitol. Prvú kapitolu tvorí úvod, v ktorom je načrtnuté v akom odvetví informatiky sa práca nachádza a jej rozdelenie. V druhej kapitole je popísaná teória počítačovej siete, ktorá je v rámci práce potrebná a tretia kapitola sa venuje použitým technológiám. Štvrtá a piata kapitola obsahuje návrh implementácie a samotnú implementáciu riešenia práce na úrovni kódu a programov. V šiestej kapitole sú popísané dátové sady, ktoré boli v rámci práce vytvorené. Siedma kapitola obsahuje samotné profily, ktoré boli vytvorené spolu s overením ich použiteľnosti. V ôsmej kapitole sú zhrnuté výsledky práce.

Kapitola 2

Profilovanie sieťovej komunikácie

V tejto kapitole sú popísané teoretické informácie, na ktorých je postavený návrh ostatných častí práce. Sieťová komunikácia je rozsiahla téma a preto sú v rámci tejto kapitoly popísané hlavne tie časti, ktoré priamo súvisia s riešením práce. Kapitolu je možné rozdeliť na dve hlavné časti. Informácie použité pri profilovaní a profilovanie samotné.

2.1 Teória

Cieľom tejto podkapitoly je popísať tie časti sieťovej komunikácie, aby boli pokryté všetky informácie, ktoré boli použité pri návrhu riešenia. Podkapitola obsahuje všeobecný popis architektúry, ktorý má popísať základný princíp sieťovej komunikácie. Ďalej sú osobitne popísané časti sieťovej komunikácie, ktorým je v rámci tejto práce venovaná väčšia pozornosť.

TCP/IP model

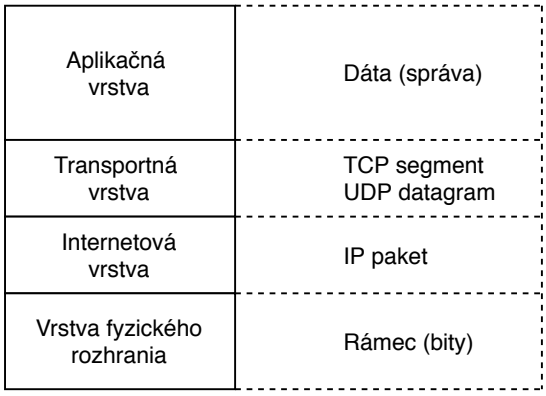
Ako sa píše v knihe [11, 14], architektúra počítačových sietí bola od počiatku navrhovaná tak, aby boli odlíšené tri časti komunikačného systému. Prenos signálov, spoľahlivý prenos a aplikačnú vrstvu, ktorá poskytuje služby. Ako štandard internetu v dnešnej dobe je používaný TCP/IP, jeden z dvoch hlavných modelov. Tento model je založený na TCP[1] a IP[2] protokoloch.

Jednotlivé vrstvy TCP/IP modelu sú zoradené od najvyššej vrstvy po najnižšiu následovne:

4. Aplikačná vrstva
3. Transportná vrstva
2. Internetová vrstva
1. Vrstva fyzického rozhrania

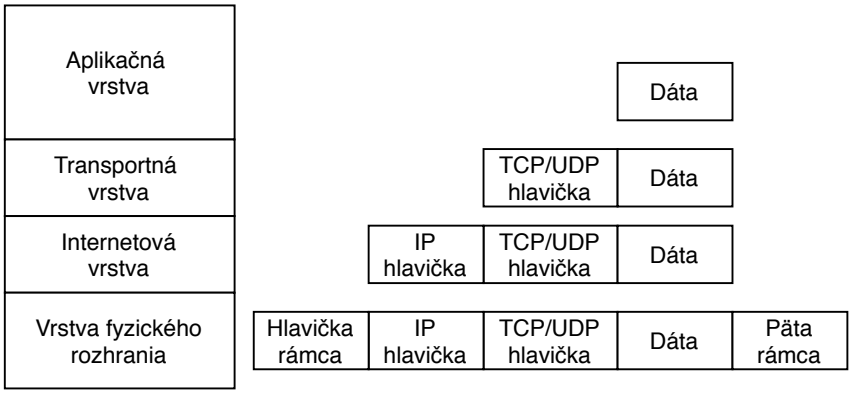
Ako je možné vidieť na obrázku 2.1, každá vrstva modelu má základnú dátovú jednotku, teda PDU[3]. Na aplikačnej vrstve sa jedná o *správu (Data)*, na transportnej vrstve sú to *TCP/UDP*[4] *pakety*, na internetovej vrstve ide o *IP datagram* a na vrstve fyzického rozhrania

[1] Transmission Control Protocol
[2] Internet Protocol
[3] Process Data Unit
[4] User Datagram Protocol



Obr. 2.1: TCP/IP model spolu s dátovými jednotkami jednotlivých vrstiev (názvy jednotiek boli získané z [15])

ide o *ethernetový rámec*. Ako je popísané v [11, 21], pri komunikácií po sieti dochádza k zapuzdreniu a rozbaleniu dát.



Obr. 2.2: Model zapuzdrenia dátových jednotiek na jednotlivých vrstvách TCP/IP modelu

Ako je popísané v [11, 21], pri odosielaní dát sú dáta z aplikačnej vrstvy postupne zapuzdrené do PDU nižších vrstiev. Tento proces je zobrazený na obrázku 2.2. Dáta sú zapuzdrené postupne od aplikačnej vrstvy až po vrstvu fyzického rozhrania. Na každej vrstve je pridaná hlavička, ktorá obsahuje informácie dôležité pre danú vrstvu. Táto hlavička je následne pridaná na začiatok PDU, ktorá bola vytvorená predošlou vrstvou. Bližšie sú hlavičky a ich časti dôležité v rámci tejto práce popísané v nasledujúcich paragrafoch. Takto zapuzdrené dáta sú odoslané. Pri prijatí dát na danom zariadení sú postupným rozbaľovaním získané potrebné informácie podľa podobného princípu, aký bol použitý pri zapuzdrení.

Pri tvorení PDU existujú pravidlá, podľa ktorých sú informácie na každej vrstve uložené. Tieto pravidlá sú aplikovateľné v rámci danej sieťovej architektúry. Ako sa píše v článku [17], sada takýchto pravidiel sa nazýva protokol. Každý protokol popisuje informácie na jednej konkrétnej vrstve a zabezpečuje dorozumenie určitej časti sieťovej komunikácie.

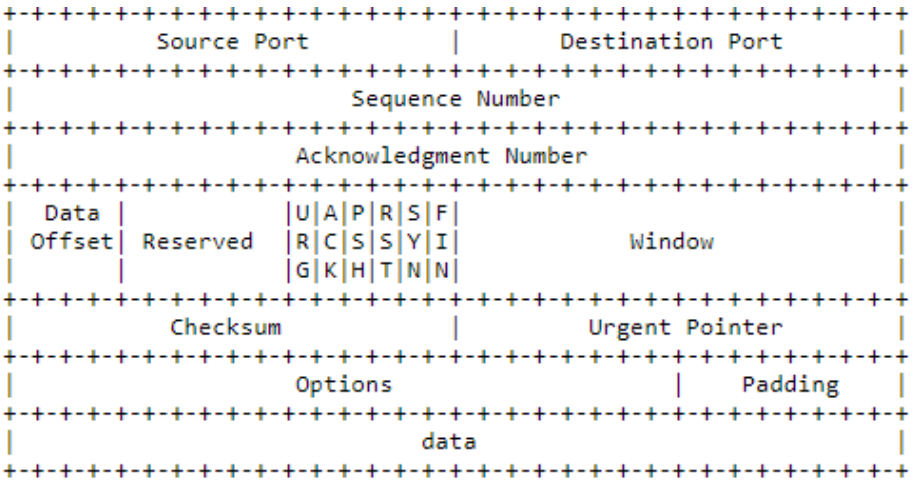
Aplikačná vrstva

Tak ako píše autor v [11, 26–29], aplikačná vrstva zabezpečuje spracovanie dát na najvyššej úrovni a je tvorená aplikáciami a procesmi. Protokoly tejto vrstvy je možné rozdeliť na tie, ktoré vykonávajú služby na základe používateľa (napr. protokol FTP - File Transfer Protocol) a na tie, ktoré zabezpečujú sieťové funkcie (napr. DNS - Domain Name System).

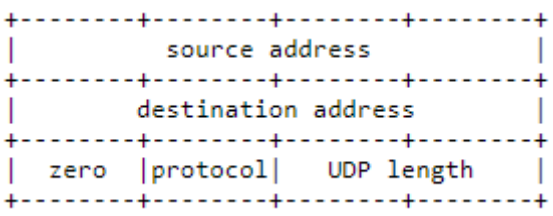
Adresovanie[5] na aplikačnej vrstve je rozdielne podľa aplikácií. Rozdielne adresovanie je možné poukázať na porovnaní elektronickej pošty, kde je formát adresy *user@host* a systému WWW, kde je formát URL[6].

Transportná vrstva

Podľa [11, 24–29] transportná vrstva zabezpečuje spojenie medzi procesmi. Týmto spojením potom transportné protokoly posielajú aplikačné dáta, ktoré sú rozdelené do paketov. Základné protokoly transportnej vrstvy sú TCP a UDP. Okrem segmentácie aplikačných dát a posielanie týchto segmentov (paketov), ktoré plnia oba protokoly, transportná vrstva sa stará aj o iné úlohy, ktoré sú špecifické pre daný protokol.



Obr. 2.3: Formát TCP hlavičky (Obrázok prevzatý z *RFC 793*)



Obr. 2.4: Formát UDP hlavičky (Obrázok prevzatý z *RFC 768*)

Ako je vidieť pri porovnaní obrázkov 2.3 a 2.4, hlavička protokolov TCP a UDP sa líšia v počte informácií. Informácie, ktoré sú v hlavičke TCP navyše oproti hlavičke UDP zaisťujú okrem iného aj spoľahlivý prenos dát. TCP vďaka týmto informáciám umožňuje kontrolovať

[5] Identifikácia cieľa na základe jednoznačnej informácie
[6] Uniform Resource Locator

poradie poslaných paketov, potvrdenie o prijatí a teda detekcia straty časti aplikačných dát. TCP protokol sa využíva pri aplikáciach, kde je dôraz na doručenie všetkých paketov.

Ako sa píše v [11, 67–68], vytvorené TCP spojenie je perzistentné a jeho vytvorenie je zložitejšie ako pri UDP. Na vytvorenie takéhoto spojenia sa používa mechanizmus trojfázového podania ruky (ang. *TCP 3–way handshake*). Tento mechanizmus spočíva vo výmene paketov s nastavenými príznakmi a poradovými číslami. Príznaky, ktoré sú v paketoch počas vytvárania spojenia nastavené, sú v poradí SYN, SYN+ACK, ACK. Po ustanovení cez toto spojenie prebieha komunikácia, teda výmena dát. Ukončenie spojenia je pomocou výmeny štyroch paketov, s príznakmi FIN a ACK spolu s poradovými číslami.

Protokol UDP nevytvára spoľahlivé spojenie, čím je zvýšená rýchlosť prenosu paketov, no nezabezpečuje spoľahlivé doručenie. Doručenie paketov v správnom poradí a detekciu straty paketov teda musí zabezpečiť zdroj. Protokol UDP sa používa prevažne na pri aplikáciach, kde je dôraz na rýchlosť doručenia paketov.

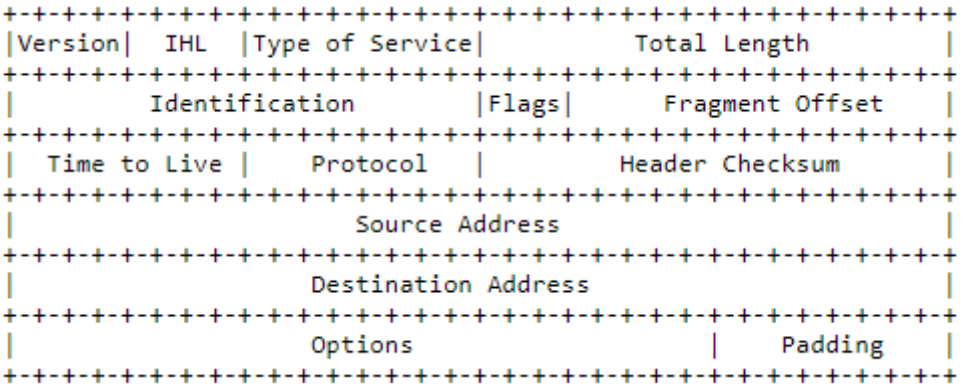
V hlavičkách oboch spomínaných protokolov je obsiahnuté číslo zdrojového a cieľového portu. Tieto porty identifikujú služby, ktoré spolu komunikujú. Zdrojový port označuje službu, ktorá dáta posiela a cieľový port označuje službu, ktorá dáta prijíma. Čísla portov sa delia na tri kategórie:

- Rezervované (0–1023)
- Registrované (1024–49151)
- Dynamické (49152–65535)

Rezervované porty slúžia na štandardné služby, ako napríklad DNS(53) či HTTPS[7](443).

Internetová vrstva

Ako sa píše v [11, 22–28], internetová vrstva sa stará o vytváranie datagramov, smerovanie a adresovanie. Ako hlavný protokol je možné považovať prenosový IP protokol. Medzi ostatné protokoly, ktoré vrstva používa, patri protokoly ICMP (Internet Control Message Protocol), IGMP (Internet Group Management Protocol), ARP (Address Resolution Protocol) a RARP (Reverse ARP).



Obr. 2.5: Formát IPv4 hlavičky (Obrázok prevzatý z *RFC 791*)

[7]Hypertext Transfer Protocol Secure

Na internetovej vrstve pri TCP/IP modely slúži k identifikácií sieťového rozhrania IP adresa. Táto IP adresa je jednoznačná v rámci celého internetu. Pri IP protokole verzii 4 sa jedná o 32 bitovú adresu, v prípade verzie 6 sa jedná o 128 bitovú adresu. Formát IP adresy pri verzii 4 je `x.x.x.x`, kde `x` nadobúda hodnoty z rozsahu 0 až 255. Kontrola IP adries je v réžií zariadení, ktoré operujú na internetovej vrstve. IP adresy môžu byť nastavené staticky, kedy sa IP adresa nemení a dynamicky, kedy je IP adresa pridelená len na určitú dobu, po uplynutí ktorej môže byť priradená inému zariadeniu. Každé sieťové rozhranie má okrem fyzickej adresy priradenú minimálne jednu IP adresu. Na prevod IP adresy na fyzickú a opačne slúžia protokoly ARP a RARP. Zdrojová a cieľová IP adresa sa nachádza v IP hlavičke, ako je možné vidieť v obrázku 2.5.

Vrstva fyzického rozhrania

Ako autor popisuje v [11, 22–27], vrstva fyzického rozhrania okrem iného zabezpečuje zapuzdrenie IP datagramov do rámcov a obsahuje funkcie, vďaka ktorým je možný prístup k fyzickému médiu. Na identifikáciu sieťového rozhrania počítača (karta NIC[8]) sa na tejto vrstve používa fyzická adresa. Pri technológiách založených na Ethernete ide o 48 bitovú adresu, kde prvá polovica bitov adresy definuje výrobcu karty a druhá polovica je zvolená výrobcom.

DNS

V rámci tohoto paragrafu sú popísané tie časti systému DNS, ktoré určitým spôsobom súvisia s návrhom, implementáciou a vyhodnotením práce. Zdroj popísanej teórie je [11, 107–134].

Pre používateľov je vhodnejšia identifikácia sieťových zariadení pomocou doménových mien a na úrovni sieťových zariadení samotných je výhodnejšia identifikácia pomocou IP adries. Základnou úlohou DNS služby je práve mapovanie doménových mien na IP adresy. DNS služba obsahuje všetky doménové mená spolu s IP adresami. Kvôli veľkému objemu dát sú tieto informácie rozdelené na viacere počítače, ktoré sú nazývané DNS servery. Samotný proces vyhľadávania v DNS systéme sa nazýva rezolúcia.

Informácie zo systému DNS sa k používateľským programom dostanú pomocou klientského programu *Resolver*. Resolver dotazuje DNS server, následne interpretuje odpoveď a získané dáta posúva požívateľskému programu, ktorý dáta žiadal. Pokiaľ má resolver vlastnú cache pamäť a hľadaná informácia sa v nej nachádza, DNS server nedotazuje. V prípade, že sa hľadaná informácia v tejto pamäti nenachádza, resolver dotazuje najbližší (lokálny) DNS server. Na základe typu dotazu sa resolveru vráti odpoveď. V prípade iteratívneho dotazu, resolver dostane v odpovedi hľadanú informáciu alebo adresy serverov, ktoré sú najbližšie k vyhľadávanej adrese. V prípade rekurzívneho dotazu sa resolveru vracia odpoveď o hľadanej informácií.

TLS

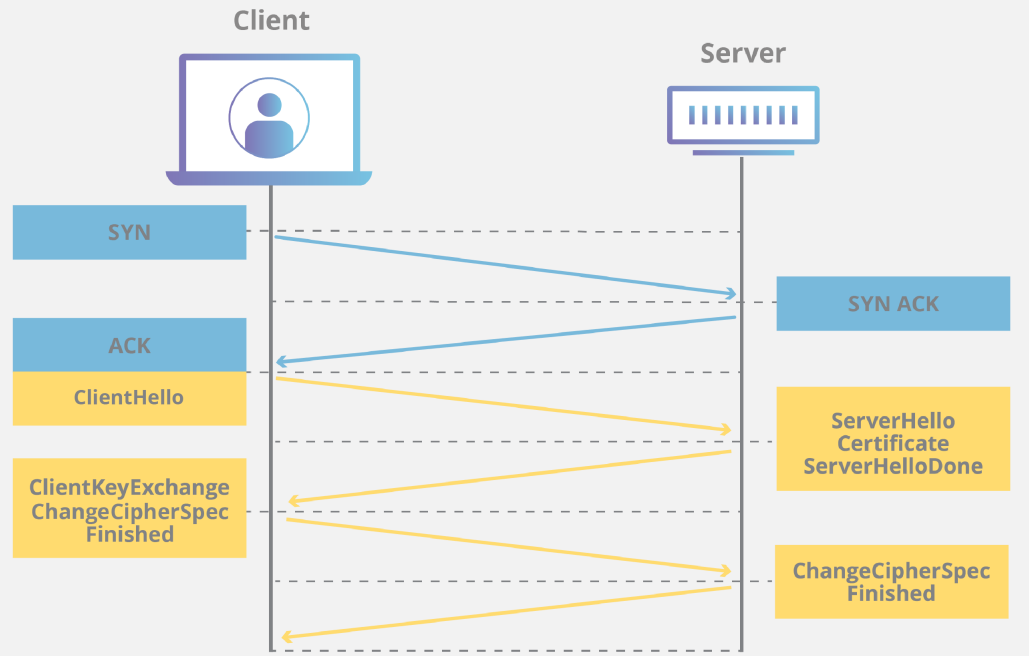
TLS (Transport Layer Security) je podľa dokumentu [13] protokol, vďaka ktorému je možné pre klient/server aplikácie komunikovať bez toho, aby boli správy odpočúvané či upravované. Primárny cieľ TLS je poskytnúť bezpečný kanál pre komunikantov. Takýto kanál by mal poskytovať autentifikáciu ako serveru tak klienta, utajenosť dát (dáta sú viditeľné len

[8] Network Interface Card

pre koncové uzly komunikácia) a neporušenosť dát (zmeny správ útočníkom by mali byť zaznamenané).

TLS sa skladá z dvoch hlavných častí. Prvá časť je tzv. *handshake protocol*, ktorý rieši autentifikáciu komunikantov, šifrovacie módy, parametre a kľúče. Druhá časť je tzv. *record protocol*, ktorý používa dohodnuté parametre a rozdeľuje komunikáciu na samostatné záznamy, ktoré sú individuálne chránené.

Handshake protocol prebieha po vytvorení TCP spojenia. Ako popisuje článok [2], počas TLS podania rúk (tzv. *TLS handshake*) je medzi klientom a serverom vykonaných viacero vecí. Medzi ktoré patrí definovanie používanej verzie TLS, definovanie používaných množín šifrovacích algoritmov, verifikovanie identity serveru a generovanie kľúčov, ktoré budú použité na šifrovanie. Tento proces je zobrazený na obrázku 2.6.



Obr. 2.6: Výmena paketov tvoriacich TCP spojenie, nasledujúca výmenou paketov popisujúcich *TLS handshake* (Obrázok prevzatý z článku [2])

2.2 Profilovanie

Podľa autora v článku [6], profilovanie je možné popísať ako technológiu, vďaka ktorej je možné získať znalosti z dát tým, že poukazuje na možné korelácie medzi dátami, ktoré sú analyzované. V článku autor spomína profil, ktorý popisuje jednu osobu. Takýto profil popisuje rôzne zvyky konkrétnej osoby, na základe nahraných dát tejto osoby.

Z toho je možné vyvodiť ekvivalent, ktorý sa v rámci tejto práce používa. V rámci tejto práce profilovanie teda znamená zoskupenie informácií o správaní aplikácie na sieti a následná interpretácia týchto informácií, ktoré profil poskytuje.

Metódy profilovania

V práci [19] autor okrem iného popisuje rôzne metódy klasifikácie v sieťovej komunikácií. Princíp niektorých týchto metód je použitý aj pri návrhoch v rámci tejto práce. Konkrétne sa jedná o klasifikáciu na základe portov, klasifikácia na základe obsahu dát, ktoré aplikácia posiela a algoritmy strojového učenia. V rámci tejto práce je profilovanie inšpirované prvými dvoma metódami.

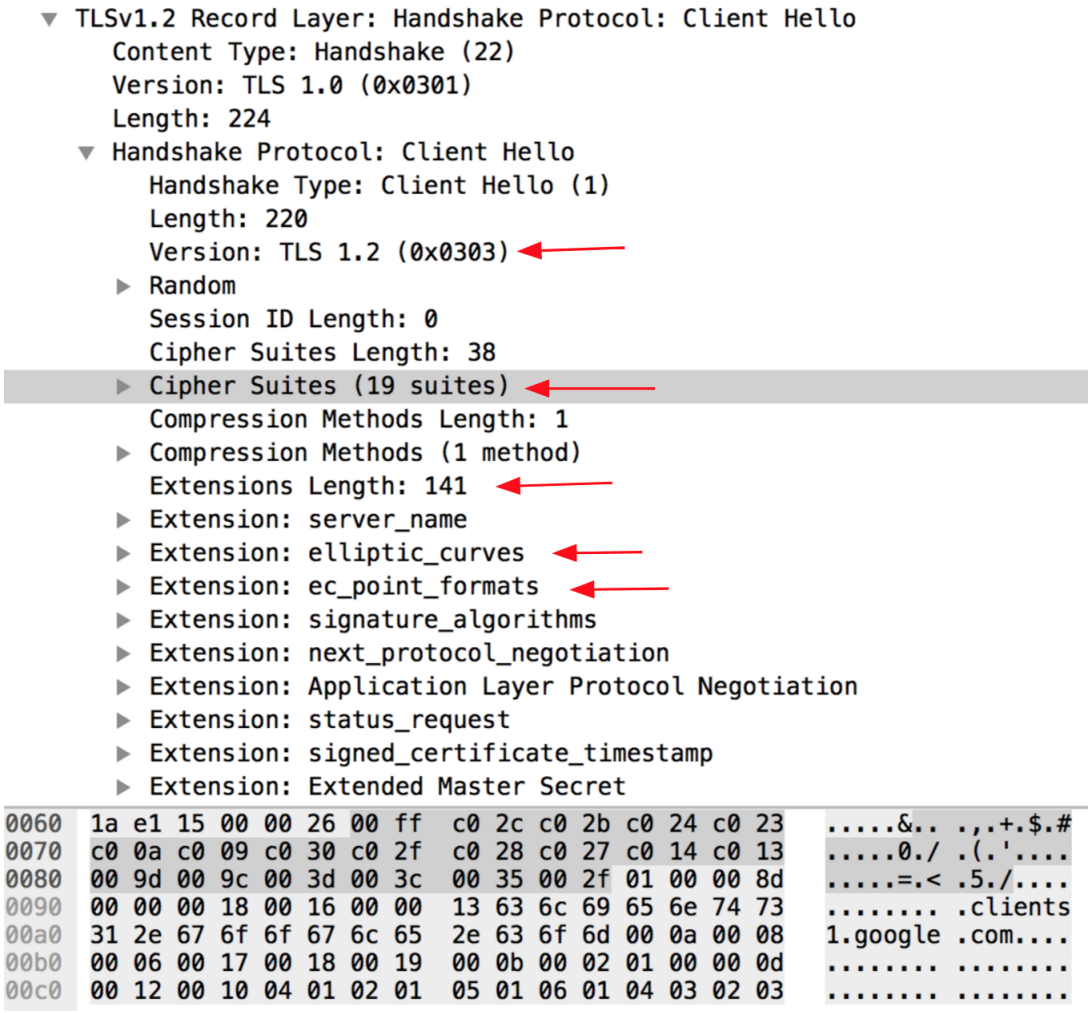
Klasifikovanie podľa portov je založené na znalosti o tom, ktoré porty sú používané ktorými aplikáciami. Táto metóda je rýchla, no ak aplikácie používajú dynamické porty, znižuje sa jej využitie.

Metóda Klasifikovania na základe obsahu dát ktoré aplikácia posiela, je taktiež známa aj ako DPI (Deep Packet Inspection). Princíp tejto metódy spočíva v analyzovaní dát, ktoré aplikácia posiela a vytvorení charakteristickej signatúry tejto aplikácie. Na základe tejto signatúry je aplikácia identifikovaná v sieťovej prevádzke.

JA3

JA3 je metóda, ako popisujú autori v blogu [1], ktorú je možné využiť na vytvorenie odtlačku klienta na základe TLS paketu. JA3 odtlačok môže predstavovať časť profilu, keďže ide o jednu z vlastností aplikácie. Na vytvorenie odtlačku metóda JA3 využíva práve informácie z TLS paketov, ktoré klient a server posiela po TCP trojfázovom podaní rúk. Klient posiela *Client Hello* paket, ktorý obsahuje informácie odvodené z použitých metód a balíčkov použitých pri tvorbe klientskej aplikácie. Tým, že sú tieto informácie viditeľné, je možné časť z nich použiť za cieľom vytvorenia JA3 odtlačku. Server odpovedá paketom *Hello Server*, ktorý obsahuje vyhovujúcu konfiguráciu.

Z dát, ktoré sú zobrazené na obrázku 2.7, sú na vytvorenie JA3 odtlačku použité položky *Version, Cipher Suited*, zoznam rozšírení (*extensions*), *elliptic_curves a ec_point_formats*. Z decimálnych hodnôt týchto položiek je vytvorený reťazec, v ktorom sú tieto hodnoty oddelené čiarkou („,“). Ak položka obsahuje viaceré hodnoty, ako napríklad zoznam rozšírení, tieto hodnoty sú rozdelené pomlčkou („–“). Príklad takéhoto reťazca je:
`771,49195-49196-52393-49199-49200-52392-49171-49172-156-157-47-53,65281-0-23-35-13-5-16-11-10,29-23-24,0`
Takýto reťazec je následne transformovaný MD5 algoritmom, čím vzniká JA3 odtlačok. Výsledný odtlačok má tvar `6f5e62edfa5933b1332ddf8b9fb3ef9d`



Obr. 2.7: Príklad *Client Hello* paket, zobrazený vo *Wireshark* (Obrázok prevzatý z [1])

Kapitola 3

Použité nástroje

V tejto kapitole sú popísané hlavné nástroje použité v rámci práce. Konkrétne sú popísané nástroje použité pri tvorbe dátových sád, použité aplikácie a zvolený programovací jazyk.

3.1 Nástroje použité na tvorbu dátových sád

Táto podkapitola popisuje samotné nástroje, ktoré sú použité pri tvorbe dátových sád. Konkrétne nástroje na spustenie Android zariadení a zaznamenávanie sieťovej komunikácie. Ich využitie, rozdiely, výhody a nevýhody použitia sú popísané v kapitole 5.1.

Android Studio

Android Studio[1] je oficiálne vývojové prostredie pre Android aplikácie, ktoré je založené na *IntelliJ IDEA*[2]. Android Studio disponuje emulátorom Android zariadení. Týmto spôsobom je možné jednoducho využiť viaceré Android verzie.

Verzia Android Studia použitá v rámci práce je *3.5.1*. Pri tvorbe dátových sád s použitím Android Studia emulátoru boli použité zariadenia s Android verziami *7.1.1*, *8.1* a *9.0*.

Oracle VM VirtualBox

Oracle VM VirtualBox[3] je multi–platformná virtualizačná aplikácia. Použitím tejto aplikácie je možné vytvoriť virtuálne identifikovateľné Android zariadenie. V rámci práce bola používaná verzia *6.0.12r133076*. Verzie Android, ktoré boli použité pri vytvorení virtuálnych zariadení boli *7.1*, *8.1* a *9.0*.

V rámci práce je touto aplikáciou vytvorené aj virtuálne zariadenie, ktoré je použité na zachytávanie sieťovej komunikácie.

Wireshark

Wireshark[4] je analyzátor sieťových protokolov. V rámci práce je použitý na odchytávanie sieťovej komunikácie Android zariadení. Použitím tohoto programu je možné záznam uložiť do formátu *.pcap*. Ako je popísané na stránke [10], *.pcap* súbory sú vytvárané sieťovými analyzátormi a obsahujú pakety zaznamenané v sieťovej komunikácií.

[1] https://developer.android.com/studio/intro
[2] https://www.jetbrains.com/idea/
[3] https://www.virtualbox.org/wiki/VirtualBox
[4] https://www.wireshark.org/

V rámci práce bola použitá verzia *3.0.5* pre Windows a verzia *2.6.10* pre Ubuntu. Verzia pre Ubuntu je použitá v kombinácií s virtuálnymi zariadeniami vytvorených použitím aplikácie *VirtualBox*.

3.2 Vybrané mobilné aplikácie

Všetky mobilné aplikácie boli stiahnuté na Android zariadenia z *Google Play*[5]. Mobilné aplikácie boli vybrané na základe osobných preferencií v kombinácií s poradím aplikácií v rebríčku na *Google Play*. Dôraz pri výbere aplikácií bol daný na existenciu sieťovej komunikácie.

Cestovné poriadky CP[6] je oficiálna slovenská aplikácia pre vyhľadávanie v cestovných poriadkoch vlakovej, autobusovej a mestskej hromadnej dopravy. Túto aplikáciu som vybral na základe jej aktívneho používania. V rámci práce bola použitá verzia *1.5.0*.

Můj vlak[7] je mobilná aplikácia pre vlaky Českých dráh. Táto aplikácia bola vybraná na základe podobnej funkcionality s aplikáciou *Cestovné poriadky CP*. V rámci práce bola používaná verzia *1.15.1*.

Ako sa píše v článku [18], *Reddit* je masívna kolekcia fór a v čase písania článku bola 18. najpopulárnejšou stránkou celosvetovo. Na základe toho spolu s preferenciami, bola vybraná mobilná aplikácia *Reddit*[8]. V rámci práce bola použitá verzia *2020.6.0*.

Aplikácie *Seznam.cz*[9] je prehliadač od Seznamu. Táto aplikácia bola vybraná výhradne na základe odporúčaní na *Google Play* spolu s predpokladom, že spôsob akým aplikácia funguje sa líši od ostatných použitých aplikácií.

3.3 Vybraný programovací jazyk

Celá programová časť práce je implementovaná v jazyku *Python*, verzia *3.8.0*. Ako sa píše v článku [9], *Python* je jeden z najpopulárnejších programovacích jazykov v roku 2020 a medzi jeho hlavné využitia je možné zaradiť aj analýzu dát. Tento jazyk bol vybraný na základe jeho použiteľnosti a osobných preferencií.

[5] https://play.google.com/store?hl=en
[6] https://play.google.com/store/apps/details?id=cz.chaps.cpsk&hl=sk
[7] https://play.google.com/store/apps/details?id=cz.cd.mujvlak.an&hl=sk
[8] https://play.google.com/store/apps/details?id=com.reddit.frontpage&hl=sk
[9] https://play.google.com/store/apps/details?id=cz.seznam.sbrowser&hl=sk

Kapitola 4

Návrh

Táto kapitola obsahuje návrhy riešení a implementácií praktických časti. Praktické časti predstavujú možné spôsoby tvorby dátových sád, ich reprezentácia a spracovanie, vytvorenie a reprezentácia profilu a návrh rôznych pokusov. Každý pokus je vytvorený s určitým cieľom. Ako jeden z hlavných cieľov je možné považovať použiteľnosť rôznych metód, konkrétne pri tvorbe dátových sád. Na to nadväzujú pokusy, ktoré porovnávajú výsledky tvorby dát. Medzi hlavné ciele pokusov patrí porovnanie výsledkov vyhľadávania profilov v dátových sád.

4.1 Návrh tvorby dátových sád

Táto podkapitola popisuje návrh na tvorbu dátových sád. Ide o navrhnutie najvhodnejšieho spôsobu vytvorenia dátových sád. Sú tu taktiež opísané dôvody či rozhodnutia, ktoré daný proces ovplyvnili.

Tým, že nie sú k dispozícií žiadne informácie o sieťovej komunikácii vybraných aplikácií, s ktorými by bolo možné výsledky kontrolovať, vytvorené dátové sady sú jediným zdrojom informácií. Preto je kladený dôraz na tvorbu dátových sád. Dátová sada obsahujúca chybné či zavádzajúce informácie môže znehodnotiť výsledný profil. Cieľom je teda vytvoriť dátové sady, vďaka ktorým je možné čo najpresnejšie definovať profil aplikácie. Preto bol pri tvorbe návrhu tvorenia dátových sád braný dôraz na určité faktory, ktoré by dátová sada mala spĺňať.

Ako sa píše v [4], dáta sú kvalitné ak spĺňajú požiadavky použitia. V tomto prípade sú to požadované dátové sady, z ktorých je možné čo najpresnejšie vytvoriť profil aplikácie. Rôzne faktory ovplyvňujú kvalitu dát, s ktorými sa pracuje. Jedným z týchto faktorov je *konzistencia* a *presnosť*. Cieľom je teda tvorba dátových sád, ktoré sú presné a zároveň konzistentné.

Presnosť

Presnosť dátovej sady je možné hodnotiť rôzne. Ako hlavný faktor presnosti dátovej sady bolo to, či dátová sada obsahuje kompletnú sieťovú komunikáciu aplikácie. To znamená, že ak bolo zaznamenávanie komunikácie ukončené pred tým, ako ukončila komunikáciu aplikácia, časť tejto komunikácie chýba. To isté platí aj v prípade, ak bolo zaznamenávanie zapnuté až po tom, čo aplikácia iniciovala komunikáciu.
Preto je pri tvorbe dátovej sady dôležité dbať na precíznosť, aby k takýmto situáciám ne-

došlo.

Príliš rozsiahly záznam je taktiež možné brať za menej presný. To z toho dôvodu, že nie je možné na isto povedať, ktorá komunikácia zo záznamu môže byť asociovaná s aplikáciu. To za predpokladu, že je isté, že aplikácia v čase tvorby záznamu komunikovala po sieti. Komunikáciu ktorá nesúvisí s aplikáciu je možné nazvať šumom.

Na odstránenie relatívne veľkej časti šumu je možné použiť systém, pri ktorom by sieť na ktorej komunikácia prebieha bola do určitej miery izolovaná. To je možné docieliť použitím virtuálnych strojov zapojených do samostatnej siete.

Konzistencia

Ako je v článku [5] popísaná premena dát na informáciu a znalosť, v tejto práci ide o proces získania určitých znalostí z dátových sád. Rôznorodosť obsahu dátových sád môže spôsobiť nesprávnu interpretáciu dát a tým vytvorenie nesprávneho záveru. Ako príklad je možné uviesť dátovú sadu, ktorá obsahuje komunikáciu inej aplikácie. Výsledný profil by bol tým pádom ovplyvnený týmito dátami a výsledný profil by nezodpovedal realite.
Dôraz na konzistenciu obsahu dátových sád má za cieľ dosiahnuť čo najlepšiu možnú interpretáciu dát. Ako prvý krok v zaistení konzistencie obsahu dát je dodržiavať to, že jedna dátová sada obsahuje jedno zapnutie a vypnutie práve jednej aplikácie.

Rovnorodosť obsahu dát je možné docieliť aj tým, že pri tvorení záznamu sú prevedené tie isté kroky. Jedna množina dátových sád tým pádom bude obsahovať rôzne záznamy tej istej situácie, čím je možné presnejšie danú situáciu popísať. Vďaka tomu je možné presnejšie popísať konkrétne spôsoby spustenia aplikácie. Ako príklad je možné uviesť počiatočné spustenie aplikácie. Z dátových sád, ktoré obsahujú len počiatočné spustenia aplikácie je možné tento proces popísať lepšie, ako by boli v dátových sadách zamiešané aj záznamy obsahujúce napríklad iné aplikácie. Tým, že je možné získať relatívne presný popis takejto konkrétnej situácie, pri spájaní týchto menší častí je možné vytvoriť prehľadný popis viacerých situácií.

4.2 Návrh profilu

Ako popisuje autor v článku [6], jeden z viacerých druhov profilov je osobný profil. Takýto profil opisuje práve jednu osobu a môže byť vytvorený výhradne zo získaných dát reprezentujúcu túto osobu. Rovnaký cieľ je aj pri tvorení profilu v rámci tejto práce. Pri tvorbe profilu je snaha reprezentovať konkrétnu aplikáciu výhradne na základe jej sieťovej komunikácie, ktorá bola nahraná.

Je vhodné poukázať na to, že nie sú dostupné informácie o tom, aké dáta má aplikácia k dispozícií, aké informácie v rámci sieťovej komunikácie potrebuje či ako má vyzerať ideálna sieťová komunikácia tejto aplikácie. Preto návrh profilu a práce s ním bol vytvorený na základe teoretických znalostí sietí. Z týchto znalostí boli následné vyslovené určité predpoklady a tieto predpoklady sú v iných častiach práce následne overované. Ako jeden z hlavných predpokladov je existencia dát, ktoré sú v rámci sieťovej komunikácie aplikácie vždy prítomné. Na základe tohoto je možné považovať vyšší počet dátových sád pre aplikáciu za výhodu.

Prvá a hlavná myšlienka pri tvorbe návrhu profilu bolo nájsť informácie obsiahnuté v sieťovej komunikácií, pomocou ktorých je možné konkrétnu aplikáciu z komunikácie identifikovať. Druhá myšlienka bola zameraná na jedinečnosť týchto informácií. Teda nie len

aplikáciu identifikovať, ale taktiež nezmýliť si ju s inou aplikáciu. Keďže je väčšina sieťovej komunikácie v súčastnosti zakódovaná, v rámci práce bol zvolený prístup zamerať sa na informácie ktoré sú dostupné, prípadne informácie ktoré je z dostupných dát možné vytvoriť.

Základný tvar profilu

V tejto kapitole sú popísané vybrané atribúty ktoré budú tvoriť profil aplikácie a dôvody, kvôli ktorým boli zvolené. Medzi hlavné atribúty patrí JA3 odtlačok, hostname a IP adresa. Popis týchto atribútov je v podkapitolách 2.1 a 2.2.

JA3 odtlačok bol zvolený na základe článku [1], kde autor popisuje ako by kombinácia odtlačku klienta a serveru mala zvýšiť presnosť identifikácie zakódovanej komunikácie medzi daným klientom a serverom. Z toho bolo implikované využitie časti tejto metódy, konkrétne JA3 odtlačok klienta. V článku autor popisuje, že aj napriek možnej neznalosti IP adries či domén serverov je možné vďaka tejto metóde s relatívnou presnosťou identifikovať komunikáciu na základe TLS výmeny medzi klientom a serverom. Na základe týchto vyjadrení bolo zvolené pri tvorbe profilu kontrolovať JA3 odtlačky obsiahnuté v poskytnutých dátových sadách.

Použitie IP adresy ako jeden z atribútov tvoriaci profil bolo zvolené na základe jej významu. Ako je popísané v [11, 28], na základe IP adresy je možné vytvárať spojenia a posielať dáta. Ak teda aplikácia komunikuje po sieti, je možné predpokladať, že má nejakým spôsobom zadefinované, aké adresy má použiť na vytvorenie spojenia. Ako je v kapitole 2.1 popísané, IP adresy a hostname spolu súvisia. Na získanie IP adresy je vo väčšine prípadov použitý systém DNS, ktorý je popísaný v kapitole 2.1, paragraf DNS. Využitie oboch týchto informácií vychádza z toho, že aj keď spolu tieto atribúty úzko súvisia, oba popisujú komunikáciu aplikácie z iného pohľadu. Hostname je možné považovať za informáciu, ktorú má aplikácia uloženú. Využíva ju v prípade, že nemá k dispozícií IP adresu, ktorej táto hostanme zodpovedá. IP adresa je naopak použitá pri aktívnej komunikácií aplikácie.

Použitie portov ako jeden z atribútov definujúci aplikáciu na základe jej komunikácie nebolo primárne zvolené. Ako je popísané v [11, 28–29], porty sa rozdeľujú podľa rozsahu, do ktorého spadajú. Využiteľnosť portov priradených štandardným službách je pri profile minimálna, keďže sa nejedná o port špecifický pre aplikáciu. Tým, že dynamické porty sú určené na krátkodobé využitie a na rôznych systémoch môžu identifikovať rôzne služby, ich použiteľnosť v profile je nízka. Jediné porty, ktoré by teoreticky boli použiteľné, sú registrované porty. To za predpokladu, že aplikácia používa nejaké rezervované porty. Pravdepodobnosť využitia portov ako identifikátor profilu je nízka, no v rámci práce by mala byť overená.

Modifikácie profilu

Cieľom modifikácie profilov je zistiť použiteľnosť jednotlivých atribútov profilu. Pod použiteľnosťou je možné rozumieť, aký dopad bude mať atribút na celkový výsledok pri vyhľadávaní profilu.

Prvá modifikácia profilu predstavuje vytvorenie špecifického či jedinečného profilu. Takýto profil by mal obsahovať len informácie, ktoré nezdieľa so žiadnym iným profilom. Preto, že je možné vytvoriť viaceré profily pre jednu aplikáciu (rôzne spôsoby spustenia zariadenia, rôzne verzie operačného systému), vyfiltrované informácie by mali byť len voči profilom iných aplikácií, nie iných profilov tej istej aplikácie. Cieľom tejto modifikácie je získať informácie, ktoré sú jedinečné pre komunikáciu aplikácie pri danom nastavení.

Druhá modifikácia profilu predstavuje rozšírenie „rozsahu“ profilu aplikácie. Ide o opak toho, čo v prvej modifikácií. Teda spojenie viacerých profilov reprezentujúcich jednu aplikáciu za rôznych nastavení. Cieľom tejto modifikácie je získať informácie, podľa ktorých je možné aplikáciu identifikovať bez ohľadu na to, aké sú nastavenia pri spustení aplikácie.

Tretia modifikácia predstavuje kombináciu predošlých dvoch modifikácií. Cieľom tejto modifikácie je získať informácie v profile, ktoré sú jedinečné pre konkrétnu aplikáciu bez ohľadu na to, aká verzia operačného systému je použitá a ako je zariadenie spustené.

4.3 Návrh pokusov

Ako je popísané v úvode tejto kapitoly, pokusy robené v rámci tejto práce majú rôzne ciele. Okrem iného je možné za pokusy považovať aj modifikovanie profilov preto, že tieto modifikácie dokážu potvrdiť alebo vyvrátiť rôzne predpoklady.

Veľkú časť práce tvoria pokusy s vyhľadávaním profilov v dátových sadách. Cieľom týchto pokusov spočíva v porovnaní výsledkov pre rôzne profily, nastavenia a modifikácie profilov. Na základe týchto výsledkov by malo byť následne možné vyhodnotiť použiteľnosť jednotlivých modifikácií profilov a nastavení. Pod nastaveniami sú myslené rôzne hodnoty atribútov a obmedzení.

Vyhľadávanie profilov

Samotné vyhľadávanie profilu v dátovej sade je proces, ktorého výsledkom je číslo, ktoré nadobúda hodnotu z rozsahu <0, 100>. Toto číslo prezentuje istotu, s ktorou je na základe získaných informácií možné povedať, že sa daný profil v dátovej sade našiel. Iná formulácia môže znieť, s akou istotou je možné profil identifikovať v dátovej sade.

Ako je popísané v návrhu tvorby profilu, každý profil bude obsahovať určité množstvo položiek. Tieto položky sa delia na JA3 odtlačky, IP adresy a hostname. Pre proces vyhľadávania má každá z týchto položiek priradenú váhu podľa toho, pod aký atribút patrí. Váha týchto položiek označuje dôležitosť danej položky v profile. Váha položky je nezáporné celé číslo. Každá položka má taktiež výskyt, ktorý značí v koľkých dátových sadách sa táto položka vyskytla. Výskyt položky je kladné celé číslo.

Princípom vyhľadávania profilu v dátovej sade predstavuje zistenie, ako je daný profil v dátovej sade pokrytý. Na získanie tohoto pokrytia je potrebné vypočítať koeficient profilu. Význam aj výpočet tohoto koeficientu je popísaný v následujúcej podkapitole.

Výpočet koeficientu profilu

Koeficient predstavuje číslo, špecifické pre konkrétny profil. Toto číslo je základná a najmenšia jednotka, ktorá je použitá na výpočet hodnôt položiek profilu. Hodnota položky reprezentuje akú časť profilu daná položka tvorí. Jej výpočet je popísaný rovnicou 4.1.

$$item\_value = item.occurrence * attribute\_weight \tag{4.1}$$

kde

- *item.occurrence* – predstavuje počet výskytov danej položky profilu
- *attribute_weight* – predstavuje váhu atribútu v ktorom sa *item* nachádza (IP, hostname, JA3)

Výpočet koeficientu záleží taktiež od relevantných položiek profilu. Relevantné položky profil predstavujú položky, ktorých výskyt je vyšší, ako nastavená minimálna hranica pre daný profil. Minimálna hranica výskytu je percento, ktoré je nastavené podľa potreby. Predstavuje to minimálny výskyt, ktorý položka musí mať, aby bola považovaná za relevantnú. Cieľom tejto hranice je možnosť skúmať aký vplyv má zmena počtu takýchto položiek v procese vyhľadávania.

Maximálne pokrytie profilu (max_cover) je hodnota, ktorej výpočet je popísaný v rovnici 4.2. Táto hodnota predstavuje súčet hodnôt všetkých relevantných položiek daného profilu.

$$max\_cover = \sum_{ja3}^{JA3s} ja3\_value + \sum_{ip}^{IPs} ip\_value + \sum_{hn}^{Hostnames} hn\_value \tag{4.2}$$

kde

- $JA3s, IPs, Hostname$ – sú množiny atribútov profilu – relevantné JA3 odtlačky, IP adresy, hostname
- $ja3, ip, hn$ – sú jednotlivé položky konkrétnych atribútov
- $ja3\_value, ip\_value, hn\_vaue$ – hodnota konkrétnej položky vypočítaná pomocou rovnice 4.1.

Koeficient profilu predstavuje prevrátenú hodnotu maximálneho pokrytia profilu, teda $\frac{1}{max\_cover}$.

Princíp vyhľadávania profilu v dátovej sade spočíva v zistení pokrytia vyhľadávaného profilu v použitej dátovej sade. To predstavuje vyhľadanie všetkých relevantných položiek profilu, ktoré sa nachádzajú v použitej dátovej sade a vypočítanie súčtu ich hodnôt. Tento súčet reprezentuje, na koľko percent je hľadaný profil pokrytý v dátovej sade.

Overenie použiteľnosti portov

Ako je popísané v 4.2, predpokladaná použiteľnosť portov pre identifikáciu je nízka, no je potrebné to overiť. Použiteľnosť portov je možné vyhodnotiť podľa ich výskytu na prieč dátovými sadami. Ak daná aplikácia používa konkrétny port často a tento port je špecifický pre profil danej aplikácie, je vhodné brať porty ako relevantný faktor. To je overiteľné porovnaním výskytu jednotlivých portov na prieč všetkými dátovými sadami pre každú aplikáciu.

Kapitola 5

Implementácia

Táto kapitola popisuje praktické prevedenie prevedenie a implementáciu návrhu. Je tu popísané ako boli dátové sady vytvorené, aké nástroje boli použité, popis skriptov ktoré implementujú rôzne časti návrhu.

5.1 Dátové sady

Všetky dátové sady boli vytvorené manuálne. Vďaka tomu bolo možné dodržiavať všetko, čo bolo v návrhu popísané. Dve metódy, ktoré boli pri tvorbe použité boli využitie Android Studia a virtualizácie. Pri oboch metódach boli použité tie isté postupy. Rozdiel v týchto postupoch spočíva v tom, či pred zapnutím aplikácie došlo k vymazaniu dát, ktoré aplikácia používa. Tieto dva postupy mali reprezentovať situáciu, kedy je aplikácia zapnutá prvý krát a situáciu, kedy bola už aplikácia spustená a má k dispozícií určité dáta.

Obe metódy používajú rovnaký postup vytvorenia záznamu, ktorý vyzerá takto:

1. Spustenie zaznamenávania sieťovej komunikácie
2. Vymazanie všetkých dát, ktoré má aplikácia potenciálne uložené
3. Spustenie aplikácie, dôraz na úplné spustenie a načítanie aplikácie
4. Vypnutie aplikácie
5. Ukončenie zaznamenávania a uloženia záznamu do *pcap* súboru

Pri tvorbe dátových sád, ktoré predstavujú zapnutie aplikácie spolu s dátami aplikácie je tento postup použitý bez kroku 2.

Pre jednu kombináciu android zariadenia (odhliadnuc od spôsobu spustenia) a aplikácie existuje jedna množina dátových sád. Táto množina je z polovice tvorená dátovými sadami, ktoré obsahujú spustenie aplikácie bez vymazania dát aplikácie a druhá polovica sú dátové sady, kde boli pred spustením aplikácie všetky jej dáta vymazané.

Hlavným rozdielom týchto metód je spôsob spustenia android zariadenia. Tento rozdiel je v následujúcich podkapitolách bližšie popísaný.

Použitie Android Studia

Pri tejto metóde bolo Android zariadenie spustené použitím android emulátoru, ktorým Android Studio disponuje. Samotné Android Studio je spustené na stroji s operačným

systémom Windows 10. Na tom istom stroji je spustený program *Wireshark(verzia=3.0.5)*, ktorý zaznamenáva sieťovú komunikáciu na sieti, do ktorej je pripojený.

Pri tvorbe dátových sád boli použité tri operačné systémy. *Android 9.0*, *Android 8.1* a *Android 7.1*. Na všetkých spomenutých operačných systémoch boli vytvorené dátové sady pre všetky aplikácie rovnakým postupom, ako je popísané v úvode tejto podkapitoly.

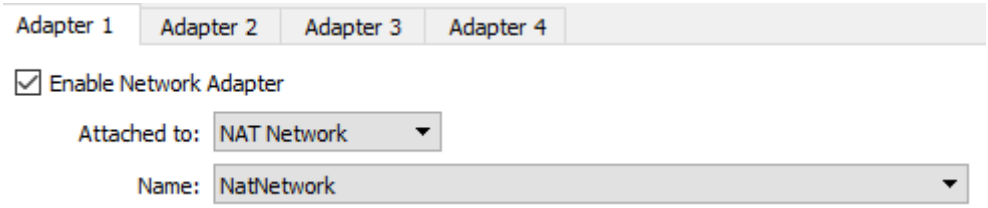
Ako je popísané v podkapitole 4.1, vytvorené dátové sady by mali čo najpresnejšie popisovať sieťovú komunikáciu aplikácie. Pri použití tejto metódy z pohľadu siete je Android zariadenie to isté, ako zariadenie na ktorom je Android Studio spustené (v tomto prípade *Windows*). Preto nie je na prvý pohľad možné presne rozlíšiť, ktorá komunikácia patrí aplikácií a ktorá *Windows* zariadeniu.

Je vhodné predpokladať, že pri tvorbe dátovej sady bude prítomná aj komunikácia iných zariadení, ktoré sa v sieti nachádzajú. Odstránením takejto komunikácie z dátovej sady sa zvýši presnosť vytvorenej sady. To je docielené filtráciou, pri ktorej ostanú v dátovej sade len tie pakety, v ktorých sa zdrojová alebo cieľová IP adresa zhoduje s IP adresou Android zariadenia.

Použitie virtuálneho zariadenia

Ako je popísané v kapitole 4.1, cieľom použitia tejto metódy je presnejší obsah dátových sád. Pri tejto metóde je použitá dvojica virtuálnych zariadení. Prvé zariadenie z dvojice slúži na spustenie programu *Wireshark*, ktorý dokáže zaznamenávať sieťovú komunikáciu všetkých použitých virtuálnych zariadení. Operačný systém tohoto zariadenia je *Ubuntu*. Druhým zariadením je virtuálne mobilné zariadenie. V rámci práce boli použité tri virtuálne mobilné zariadenia, podobne ako pri použití Android Studia. Konkrétne operačné systémy *Android 9.0*, *Android 8.1* a *Android 7.1*. Všetky virtuálne zariadenia sú vytvorené použitím technológie *Oracle VirtualBox*[1].

Primárny dôvod používania tejto metódy je vytvorenie dátových sád, bez zbytočného šumu. Odstránenie šumu je v tomto prípade docielené použitím siete, na ktorej prebieha komunikácie výhradne virtuálnych zariadení. Keďže jedno zariadenie slúži výhradne na zaznamenávanie tejto komunikácia, je potrebné aby tieto zariadenia boli pripojené do jednej siete. To je docielené nastavením konfigurácie siete každého použitého virtuálneho zariadenia tak, ako je zobrazené na obrázku 5.1.



Obr. 5.1: Nastavenie siete pri virtuálnych zariadeniach

Vytvorenie virtuálneho mobilného zariadenia prebehlo podľa postupu, ktorý je popísaný v [16]. Po vytvorení virtuálneho Android zariadenia bolo pred spustením potrebné upraviť nastavenie zariadenia. V kategórií **Display** položku **Graphics Controller** na hodnotu „VBoxSVGA“. Toto nastavenie umožnilo spustenie užívateľského rozhrania Android zariadenia a jeho potreba záleží od zariadenia, na ktorom je *VirtualBox* spustený.

[1] https://www.virtualbox.org/

Po vytvorení všetkých potrebných zariadení je postup vytvorenia dátových sád rovnaký, ako pri použití Android Studia. Taktiež platí to, že polovica dátových sád pre každú kombináciu Android zariadenia a aplikácie je vytvorená s vymazaním dát aplikácie a druhá polovica bez mazania týchto dát.

Výhoda tejto metódy oproti použitiu Android Studia spočíva v tom, že Android zariadenie je samostatné a preto je možné ho podľa IP adresy v sieti identifikovať. Tým je možné odfiltrovať tú časť komunikácie, ktorá nesúvisí s Android zariadením.

5.2 Vytvorenie profilu

Vytvorenie profilu je implementované ako funkcia, ktorú je potenciálne možné použiť ako skript. Deklarácia tejto funkcie je zobrazená vo výpise 5.1

```
def create_profile(app_name: str, operation_sys: str,
                   dataset_folder_name: str, my_ip: str,
                   save_oneaddrcom: bool = False, max_datasets: int = None,
                   one_dataset_only: bool = False,
                   specific_dataset_name: str ='Temp',
                   print_dataset_name: bool =True)
```

Výpis 5.1: Deklarácia funkcie na vytvorenie profilu

Parametre `app_name, operation_sys` slúžia na uloženie informácie, akú aplikáciu profil reprezentuje a aký operačný systém zariadenie používa. Parametre `dataset_folder_name, my_ip` sú najdôležitejšie v rámci tejto funkcie. Poskytujú informáciu u tom, kde sa nachádza vybraná množina dátových sád, z ktorých má byť profil vytvorený. Informácia o IP adrese zariadení slúži na rozlíšenie, ktoré pakety sú odchádzajúce a ktoré prichádzajúce. Pri získavaní informácií z dátových sád táto IP adresa taktiež predstavuje zariadenie na ktorom je Android spustený, či Android zariadenie samotné.

Ostatné parametre sú voliteľné. Ich použitie je v špecifických prípadoch. Konkrétne parametre `max_datasets, one_dataset_only, specific_dataset_name` sú použité v prípade, ak je potrebné vytvoriť profil z konkrétneho počtu dátových sád. Pri ponechaní východzích hodnôt je profil tvorený zo všetkých dátových sád, ktoré sa v nastavenej zložke nachádzajú. Parameter `save_oneaddrcom` je použitý vtedy, ak je vyžadované ukladanie a kategorizovanie paketov do konkrétnych komunikácií v rámci dátovej sady. V prípade použitia tohoto parametru sú všetky vytvorené komunikácie uložené. Parameter `print_dataset_name` má čisto informačnú funkciu, aby bolo pri spracovaní dátových sád vidieť, ktorá dátová sada je práve spracovávaná.

Hlavné kroky tejto funkcie je možné zoradiť.

1. Vytvorenie inštancií triedy `Dataset`
2. Vytvorenie inštancií, ktoré reprezentujú dátové sady na úrovni kódu
3. Použitie týchto inštancií na vytvorenie profilu

Tieto kroky sú detailnejšie popísané v následujúcich paragrafoch.

Spracovanie dátových sád

Spracovaním dátových sád v tomto prípade znamená vytvorenie kódovej reprezentácie jednotlivých *pcap* súborov. Každý *pcap* súbor je na úrovni kódu reprezentovaný inštanciou triedy `Dataset`. Každý objekt tejto triedy má atribúty, ktoré obsahujú extrahované informácie z dátovej sady, ktorú daná inštancia reprezentuje. Tieto atribúty sú zobrazené vo výpise 5.2.

```
class Dataset:
        self.name
        self.dns_addresses
        self.ip_addresses
        self.one_address_communications
        self.tls_fingerprints
```

Výpis 5.2: Atribúty triedy `Dataset`

Atribút `self.name` reprezentuje meno *pcap* súboru, ktorý inštancia reprezentuje. Atribút `self.dns_addresses` obsahuje IP adresy, ktoré boli získané výhradne z DNS dotazov. Každá z týchto adries tvorí dvojicu s hostname, ktorú DNS dotaz obsahoval. Atribút `ip_addresses` obsahuje všetky IP adresy, ktoré dátová sada obsahuje. Každá adresa je reprezentovaná v atribúte práve jeden krát. Atribút `tls_fingerprints` je zoznam inštancií, ktoré reprezentujú všetky JA3 odtlačky vyskytnujúce sa v dátovej sade. Inštancie obsahujúce JA3 odtlačok sú bližšie popísané nižšie v tejto podkapitole. Atribút `one_address_communications` je voliteľný a obsahuje zoznam inštancií triedy `OneAddressCommunication`, ktorej vzhľad je popísaný vo výpise 5.3.

```
class OneAddressCommunication:
        self.ip_src
        self.ip_dst
        self.srcport
        self.dstport
        self.all_packets
        self.cipher_suites
        self.ja3_fingerprint
```

Výpis 5.3: Atribúty triedy `OneAddressCommunication`

Keďže jedna inštancia triedy `OneAddressCommunication` predstavuje jedno spojenie, ktoré dátová sada obsahuje, atribúty `self.ip_src, self.ip_dst, self.srcport` a `self.dstport` definujú dané spojenie. Atribút `all_packets` obsahuje všetky pakety, ktoré je na základe IP adries a portov zaradiť do tejto komunikácie. Parametre `cipher_suites` a `ja3_fingerprints` sú atribúty, ktoré sú naplnené iba vtedy, ak sa v tejto komunikácií nachádza aj *TLS Client Hello* paket. Z tohoto paketu sú v tom prípade extrahované a uložené číselné sady spolu s JA3 odtlačkom, ktorý je z tohoto paketu vytvorený.

Na sprístupnenie paketov obsiahnutých v *pcap* súboroch v rámci kódu, je použitý Python modul s názvom *pyshark*(*verzia=0.4.2.9*)[2]. Použitím tohoto modulu je možné získať iterátor daného *pcap* súboru, ktorý umožňuje prístup k paketom.

[2] http://kiminewt.github.io/pyshark/

Na získanie JA3 odtlačkov z dátovej sady nebolo potrebné implementovať vlastné riešenie, keďže autori tejto metódy poskytujú Python modul v ktorom je táto metóda implementovaná. Konkrétne sa jedná o modul s názvom *ja3(verzia=1.0.0)*[3]. Metóda `ja3.process_pcap()`, ktorá patrí do tohoto modulu, vracia zoznam inštancií. Každá táto inštancia obsahuje informácie získané z *TLS Client Hello* paketov ktoré dátová sada obsahuje. Tieto inštancie okrem samotného JA3 odtlačku obsahujú aj reťazec, z ktorého bol odtlačok vytvorený.

Na použitie spomínanej metódy `ja3.process_pcap()` je potrebné získať iterátor súboru *pcap*. Tento iterátor je získaný použitím modulu s názvom *dpkt(verzia=1.9.2)*[4], konkrétne triedy `dpkt.pcap.Reader`.

V článku [1] autori píšu, ktoré dáta z *TLS* paketu obsahujúci *Client hello* protokol používajú na tvorbu reťazca, z ktorého je následne vytvorený JA3 odtlačok. Jedno zo spomínaných dát je z rozšírenia *elliptic_curves*. Pri aplikované metódy bolo na miesto spomínaného rozšírenia použité rozšírenie *supported_groups*. V dokumente [3] sa ale píše o premenovaní rozšírenia *elliptic_curves* na rozšírenie *supported_groups*, čím sa vysvetľuje zistená nezrovnalosť.

Na vytvorenie inštancie triedy `Dataset` je potrebný iterátor, ktorý umožňuje prístup k paketom a zoznam inštancií, ktoré obsahujú informácie o JA3 odtlačku. Postupným priechodom sú z paketov získané informácie, ktoré sú uložené v atribútoch objektu. Pri každej iterácií je paket skontrolovaný či disponuje požadovanými informáciami. Ak disponuje, tieto informácie sú extrahované a uložené ako informácia inštancie.

Prvá informácia, ktorá je z paketu extrahovaná je IP adresa, ktorá predstavuje druhého komunikanta s ktorým zariadenie komunikuje. Získanie tejto adresy je na základe toho, akú IP adresu má zariadenie. IP adresy sú získané priamo z IP vrstvy paketu a sú uložené v atribúte `self.ip_addresses` inštancie triedy `Dataset`.

Ďalej sa kontroluje, či je paket DNS odpoveď. V pozitívnom prípade je z tohoto paketu získaná IP adresa, ktorá odpovedá hostname, ktorá bola obsiahnutá v DNS dotaze. Tieto dve informácie sú následne spojené do dvojice, kde sa hostname stáva kľúčom a IP adresa, ktorá mu zodpovedá, sa stáva hodnotou. Táto dvojica je uložená do atribútu `self.dns_addresses` inštancie triedy `Dataset`.

V prípade, že je zvolený voliteľný parameter, je prevedené zaradenie paketu do komunikácie. Takáto komunikácia reprezentuje jedno spojenie definované zdrojovou IP adresou, zdrojovým portom, cieľovou adresou a cieľovým portom. Množina takýchto komunikácií reprezentuje množinu vytvorených spojení, ktoré sú v dátovej sade obsiahnuté. Tento atribút je voliteľný z rôznych dôvodov. Jedným z dôvodov je vysoká pamäťová náročnosť pri skladovaní.

Vytvorenie profilu

Profil je vytvorený z množiny inštancií triedy `Dataset`, ktoré sú získané v predošlom kroku. Výstupom tohoto kroku je inštancia triedy `AppProfile`. Vzhľad tejto triedy je zobrazený vo výpise 5.4. Táto trieda slúži na reprezentáciu aplikácie na základe jej sieťovej komunikácie, ktorá je obsiahnutá vo vytvorených dátových sadách.

```
class AppProfile:
        self.app_name
```

[3] https://github.com/salesforce/ja3
[4] https://dpkt.readthedocs.io/en/latest/

```
self.operation_system
self.number_of_datasets
self.specified
self.ip_addresses
self.hostnames
self.ja3digest_fps
self._min_occurrence_percentage
self._weight_ja3fp
self._weight_ip
self._weight_hn
self.ip_addresses_broadly
```

Výpis 5.4: Atribúty triedy `AppProfile`

Atribúty `self.app_name`, `self.operation_system`, `self.number_of_datasets` a `self.specified` majú informačnú funkciu. Vypovedajú o tom, akú aplikáciu profil reprezentuje, aký operačný systém bol použitý pri tvorbe dátových sád, koľko dátových sád bolo pri tvorbe profilu použitých a či je profil základný alebo bol nejako špecifikovaný. Skupina atribútov `self.ip_addresses, self.hostnames, self.ja3digest_fps` obsahuje všetky IP adresy, hostname a JA3 odtlačky, ktoré sa našli vo všetkých použitých dátových sadách. Každá jedna položka obsiahnutá v jednom z týchto troch atribútov obsahuje taktiež informáciu, v koľkých dátových sadách sa vyskytla. Množina atribútov, ktoré začínajú podčiarkovníkíkom (ang. underscore) reprezentujú nastavenie profilu. Atribúty `self._weight_ja3fp, self._weight_ip, self._weight_hn` slúžia ako nastavenie váh jednotlivých atribútov profilu. Od týchto váh záleží vyhľadávanie profilu. Na upravenie jednotlivých váh profilov sú implementované funkcie `update_profiles_properties()` a `update_weight_properties()`. Atribút `self._min_occurrence_percentage` je informácia, aká percentuálna hranica má byť pre profil nastavená. Toto nastavenie definuje množinu relevantných položiek profilu. Je možné nastaviť hodnotu z rozsahu <0, 1>. Tieto nastavenia sú dôležité hlavne pri vyhľadávaní profilu. Od ostatných atribútov sa líšia v tom, že sú meniteľné podľa potreby aj po vytvorení profilu. Pri zmene sú určité obmedzenia. Ako popisuje autor článku [12], atribúty začínajúce práve jedným podčiarkovníkom sú považované za chránené a nemali by byť menené priamo. Použitím Python dekorátoru `@Property` je možné upravovať tieto atribúty s tým, že nastavená hodnota vždy podlieha obmedzeniam, ktoré sú určené. `self.ip_addresses_broadly` je sekundárny atribút, ktorý pôvodne slúžil na zistenie celkového výskytu IP adries. Ak bola v dátovej sade táto adresa použitá viac krát s rôznymi portami, bola započítaná viac krát. To zavádzalo pri analýze profilu, keďže nebolo možné povedať, či ten výskyt zodpovedá jednej či viacerým dátovým sadám. Aktuálne je tento atribút použitý na vytvorenie prehľadu výhradne tých IP adresách, ktoré boli získané z DNS odpovedí.

V predošlom kroku boli získané inštancie triedy `Dataset`, ktoré reprezentujú dátové sady obsahujúce sieťovú komunikáciu vybranej aplikácie. Postupným prechádzaním týmito inštanciami sú získané všetky IP adresy, hostname aj JA3 odtlačky, ktoré sa v poskytnutých dátových sadách našli. Pri tejto iterácií je taktiež zistený výskyt každej jednej položky. Tieto dvojice, položka a jej výskyt naprieč dátovými sadami, je podla typu položky vložená do atribútu profilu. Výskyt položky atribútu v dátovej sade značí, že sa tam táto položka vyskytla minimálne jeden krát. Dátová sada môže obsahovať viacero spojení na jednu adresu

s rôznymi portami, no z pohľadu výskytu IP adresy v dátovej sade je započítaná len jeden krát.

Pri inicializácií inštancie triedy `AppProfile` sú všetky váhy atribútov nastavené na rovnakú hodnotu, konkrétne hodnotu 1. Minimálna hranica výskytu je pri inicializácií inštancie nastavená na 0, čo znamená, že každá položka ktorú profil obsahuje je braná ako relevantná.

Modifikácia profilov

Modifikácie profilov, ktoré sú popísané v kapitole 4.2, sú implementované v dvoch hlavných funkciách.

Na vytvorenie špecifického profilu je implementovaná funkcia s názvom `get_specific_profile_by_pseudo_disjunction()`. Táto funkcia vytvorí pseudo zjednotenie všetkých profilov, ktoré majú slúžiť ako filter pre vytvorenie špecifického profilu. Použitie výrazu pseudo je preto, lebo funkcia netvorí zjednotenie v pravom slova zmysle. Vytvorí zjednotenie všetkých položiek, ktoré sa vyskytujú aspoň v dvoch profiloch zo všetkých poskytnutých. Tým je vytvorený profil, ktorý obsahuje výhradne položky, ktoré nie sú pre žiaden profil jedinečné. Následne je vytvorený prienik profilu, ktorý má byť špecifický a novo vytvoreného profilu, ktorý reprezentuje pseudo zjednotenie. Všetky položky obsiahnuté v tomto prieniku sú zo špecifického profilu odstránené a profilu ostanú len položky, ktoré sú pre profil jedinečné.

Na vytvorenie profilu spojením viacerých profilov je implementovaná funkcia `merge_profiles()`. Výstupom tejto funkcie je nový profil, ktorý obsahuje všetky informácie z profilov použitých na jeho tvorbu, s aktuálnymi hodnotami. Počet dátových sád, z ktorých je nový profil vytvorený, je súčet všetkých dátových sád, z ktorých boli vytvorené použité profily. Tento istý princíp je aplikovaný na všetky položky. To znamená, že počet výskytov konkrétnej položky profilu predstavuje počet výskytov tejto položky naprieč všetkými dátovými sadami, ktoré boli použité na tvorbu použitých profilov. V následujúcich kapitolách je tento typ profilu označovaný tiež ako *super profil*.

Exportovanie, uloženie a načítanie profilu

Možnosť ukladania a načítania inštancií triedy `AppProfile` bola pridaná z toho dôvodu, aby sa predišlo nutnosti opakovaného tvorenia toho istého profilu bez zmien v dátových sadách. Po vytvorení inštancie profilu je táto inštancia automaticky uložená. Ukladanie je implementované metódou serializácie. Ako je písané na stránke [14], serializácia je proces, v ktorom je objekt transformovaný do formátu, ktorý je možné uložiť a neskôr deserializovať (teda získať originálny objekt). Na serializáciu a deserializáciu je použitý Python modul s názvom *pickle(verzia=4.0)*[5].

Po vytvorení inštancie profilu je vytvorený export, ktorý je možné uložiť a zobraziť ako textový súbor. Export obsahuje meno aplikácie ktorú profil reprezentuje, počet použitých dátových sád. V exporte sú reprezentované aj obsahy atribútov IP adries, hostname a JA3 odtlačkov spolu s výskytom jednotlivých položiek. Pred každou takouto tabuľkou je informačná tabuľka, ktorá zobrazuje počty položiek s rovnakým výskytom. Na vytvorenie exportu bol použitý Python modul s názvom *PrettyTable(verzia=0.7.2)*[6].

[5] https://docs.python.org/3/library/pickle.html
[6] https://github.com/jazzband/prettytable

5.3 Vyhľadanie profilu v dátovej sade

Princíp vyhľadávania profilu v dátovej sade aj maximálne pokrytie hľadaného profiluje sú popísané v kapitole 4.3. Tieto dve informácie sú kľúčové pri vyhľadávaní. V následujúcich podkapitolách je popísaná implementácia ich výpočtu a následne samotný proces vyhľadávania profilu.

Získanie koeficientu profilu

Akonáhle inštancia profilu existuje, má nastavené váhy atribútov a minimálny percentuálny výskyt. Ku koeficientu je pristupované ako k atribútu, ale v skutočnosti je to metóda triedy `AppProfile` s dekorátorom `@Property`. Vďaka tejto metóde implementácie, tento atribút vracia aktuálnu hodnotu koeficientu profilu.

Koeficient je priamo spojený s počtom relevantných položiek profilu. Prvý krok je teda zistenie počtu relevantných položiek. Tento počet je zistený na základe atribútu, ktorý určuje minimálny výskyt. Súčet všetkých položiek, ktorých výskyt je vyšší ako určená hranica, predstavuje počet relevantných položiek.

Druhý krok je získanie maximálneho možného pokrytia profilu. Toto maximum je vypočítané iterovaním cez všetky položky profilu. Pri každej položke je vypočítaná jej hodnota použitím rovnice 4.1. Hodnoty všetkých relevantných položiek sú následne sčítané a tým je vypočítané maximálne možné pokrytie profilu. Táto rovnica je zobrazená v 4.2.

Následné obrátenie tejto hodnoty vracia koeficient profilu. Všetky tieto kroky sú implementované v metóde profilu, ku ktorej je vďaka dekorátoru možné pristupovať ako k atribútu. Tým je docielené to, že koeficient je vždy aktuálny podľa toho, aké váhy a minimálna hranica sú nastavené.

Proces vyhľadávania profilu

Pri vyhľadávaní profilu v dátovej sade je potrebné zistiť, ktoré položky profilu sa v tejto dátovej sade nachádzajú. V rámci implementácie sa k dátovej sade, v ktorej je profil vyhľadávaný, pristupuje ako k inštancií triedy `AppProfile`. Táto inštancia je vytvorená z dátovej sady, v ktorej sa vyhľadáva.

Získanie pokrytia profilu aplikácie v dátovej sade predstavuje identifikáciu profilu. Pokrytie profile je vypočítané pri iterácií položkami profilu, ktorý reprezentuje dátovú sadu. Pri každej položke je skontrolované, či je položka obsiahnutá aj vo vyhľadávanom profile. V prípade, že áno, je vypočítaná hodnota tejto položky rovnicou 4.1. Súčet hodnôt všetkých položiek nadobúda hodnotu z rozsahu $< 0, 1 >$. Vynásobením tohoto súčtu číslom 100 je získané percentuálne pokrytie vyhľadávaného profilu vo zvolenej dátovej sade.

Získanie pokrytia profilu v dátovej sade (ktorá je v rámci funkcie reprezentovaná inštanciou triedy `AppProfile`) je implementované vo funkcií, ktorá je zobrazená vo výpise 5.5.

```
def calculate_profiles_probability_by_another_profile(
                                                app_profile,
                                                dataset_profile)
```

Výpis 5.5: Deklarácia funkcie `calculate_profiles_probability_by_another_profile`

Oba parametre v tejto funkcii sú inštancie triedy `AppProfile`, kde parameter `app_profile` predstavuje profil, ktorý je vyhľadávaný. Parameter `dataset_profile` predstavuje profil, ktorý reprezentuje dátovú sadu, v ktorej sa profil aplikácie vyhľadáva.

Implementované sú aj funkcie, ktoré slúžia na vytvorenie výstupu zobrazujúci výsledok vyhľadávania profilu. Všetky tieto funkcie používajú funkciu 5.5.
Funkcia `look_for_app_in_dataset()` vyhľadá konkrétny profil v konkrétnej dátovej sade. Táto funkcia umožňuje nastaviť minimálnu percentuálnu hranicu výskytu pre vyhľadávaný profil. Funkcia `export_profile_in_dataset_probability` vypočíta pravdepodobnosť výskytu profilov zo vstupného zoznamu profilov v dátových sadách vo vstupnom zozname dátových sád. Výstup tejto funkcie je textový súbor, ktorý obsahuje výsledky vyhľadávaní zobrazených do tabuľky.

Overenie použiteľnosti portov

Implementácia tohoto pokusu je takmer rovnaká ako vytvorenie profilu. Konkrétne časť, ktorá sa skladá z iterovania paketmi v dátových sadách a súčasnému zaznamenávaniu výskytu. Konkrétne je tento pokus implementovaný vo funkcií
`get_ports_occurrences_in_datasets_table()`. Výsledkom tejto funkcie sú tabuľky, ktoré zaznamenávajú výskyt jednotlivých zdrojových aj cieľových portov, ktoré sa vyskytli vo vybraných dátových sadách.

Kapitola 6

Vytvorené dátové sady

Táto kapitola sa venuje vytvoreným dátovým sadám, ich spracovaniu a využitiu. Pre účely tejto bakalárskej práce boli vytvorené dve skupiny dátových sád. Prvá skupina dátových sád bola vytvorená ako zdrojové dáta pre tvorbu profilu. Tieto dátové sady a ich spracovanie sú popísané v podkapitolách 6.1 až 6.4. Druhá skupina dátových sád bola vytvorená na použitie pri vyhľadávaní profilov v rôznych dátových sadách. Táto skupina je popísaná v podkapitole 6.3. Okrem iného táto kapitola popisuje aj rozdiely a výhody v použití rôznych metód tvorby dátových sád.

6.1 Dátové sady

Počet dátových sád pre danú aplikáciu závisí od cieľa, na čo budú dané dátové sady použité. Napríklad počet dátových sád na overenie informácie bude menší, ako počet dátových sád použitých na tvorbu profilu aplikácie.

Všetky dátové sady boli vytvárané dvomi spôsobmi. Polovica dátových sád bola vytvorená tak, že dáta, ktoré si aplikácia ukladá na zariadení (*cache*), boli ponechané a druhá polovica je tvorená sadami, kde je daná aplikácia spustená až po tom, čo boli tieto dáta vymazané. Použitie oboch spôsobov malo za cieľ simulovať obe situácie, v akých môže byť aplikácia reálne spustená. S použitím cache dát - napríklad zapnutie aplikácie prvý krát, s vymazaním cache dát - opätovné spustenie aplikácie.

Pri tvorbe dátových sád, ktoré sú následne použité na profilovanie aplikácie, som sa držal pravidla kde v jednej dátovej sade je spustená len jedna aplikácia jeden krát. Týmto som chcel vopred eliminovať duplikáty dát v rámci jednej sady, ktoré by mohla zneprehľadniť profil.

6.1.1 Použitie Android Studia

V tejto podkapitole sú popísané dátové sady obsahujúce sieťovú komunikáciu aplikácie, ktoré boli vytvorené použitím Android Studia a rôznych Android verzií. Implementácia tejto metódy je popísaná v kapitole 5.1.

V tabuľkách 6.1 až 6.3 je možné vidieť, že použitím Android Studia bolo vytvorených 60 dátových sád pre aplikácie *Seznam.cz, Reddit a Můj vlak* a 20 dátových sád pre aplikáciu *Cestovné poriadky CP*. Ako je z tabuliek vidieť, pre každú Android verziu bolo vytvorených 20 dátových sád pre jednu aplikáciu. Absencia časti dátových sád pre aplikáciu *CP* je spôsobená tým, že pri použití Android Studia a Android verziách 8.1 a 7.1 *Google Play* neumožňoval aplikácie stiahnuť.

Android 9.0 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	9761
Reddit	20	6607
Můj vlak	20	151
Cestovné poriadky CP	20	1839

Tabuľka 6.1: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 9.0, spusteného cez Android Studio

Android 8.1 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	6995
Reddit	20	6323
Můj vlak	20	122

Tabuľka 6.2: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 8.1, spusteného cez Android Studio

Android 7.1 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	7058
Reddit	20	3087
Můj vlak	20	124

Tabuľka 6.3: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 7.1, spusteného cez Android Studio

6.1.2 Použitie virtuálneho stroja

V tejto podkapitole sú popísané dátové sady, ktoré boli vytvorené použitím virtuálneho Android zariadenia. Tabuľky 6.4 až 6.6 obsahujú informácie o dátových sadách vytvorených pre jednotlivé Android verzie. Pri Android verzií 9.0 bolo vytvorených pre každú aplikáciu 100 dátových sád a pri ostatných verziách bolo vytvorených 20 dátových sád pre každú aplikáciu, čiže 140 dátových sád pre každú aplikáciu.

Pri tejto metóde tvorby dátových sád bolo možné vytvoriť aj dátové sady pre aplikáciu *Cestovné poriadky CP* na všetkých použitých Android verziách. To vďaka tomu, že namiesto znemožnenia stiahnutia aplikácie na verziách Android 7.1 a 8.1, *Google Play* zobrazil upozornenie, že aplikácia nemusí byť optimalizovaná s daným zariadením.

Android 9 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	100	2502
Reddit	100	859
Můj vlak	100	138
Cestovné poriadky CP	100	682

Tabuľka 6.4: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 9.0 spusteného na virtuálnom stroji

Android 8.1 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	4412
Reddit	20	808
Můj vlak	20	122
Cestovné poriadky CP	20	899

Tabuľka 6.5: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 8.1 spusteného na virtuálnom stroji

Android 7.1 aplikácia	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	3410
Reddit	20	703
Můj vlak	20	112
Cestovné poriadky CP	20	780

Tabuľka 6.6: Informácie o dátových sadách pre jednotlivé aplikácie, vytvorené na Android 7.1 spusteného na virtuálnom stroji

Internetový prehliadač	Počet dátových sád	Priemerný počet paketov v sade
Seznam.cz	20	1842
Reddit	20	964
Můj vlak	20	2883
Cestovné poriadky CP	20	1195

Tabuľka 6.7: Informácie o dátových sadách pre ekvivalenty aplikácií v internetovom prehliadači, spustený na virtuálnom stroji

Dátové sady popísané v tabuľke 6.7 boli vytvorené s cieľom poukázania na rozdielne chovanie aplikácie a chovanie prístupu na danú „doménu“ cez internetový prehliadač. Ekvivalenty aplikácií boli následovné

- *Cestovné poriadky CP* – `https://m.cp.hnonline.sk/vlakbusmhd/spojenie/`
- *Můj vlak* – `https://www.cd.cz/spojeni-a-jizdenka/`
- *Reddit* – `https://www.reddit.com`
- *Seznam.cz* – `https://www.seznam.cz`

Na tieto domény bolo pristupované cez internetový prehliadač mobilného Android zariadenia, konkrétne prehliadač *Google Chrome*[1](*verzia=80.0.3987.132*).

Z tabuľky 6.7 nie je možné vidieť žiadne rozdiely, ktoré by platili pre všetky aplikácie. Porovnanie komunikácie aplikácie a jej ekvivalentu cez internetový prehliadač je popísané v kapitole 7.

[1]Prehliadač bol jednou zo základných aplikácií Android zariadenia, t.j. nebol stiahnutý z *Google Play*

6.2 Vyhodnotenie vytvorených dátových sád

Ako je popísané v podkapitole 4.1, každá dátová sada obsahuje jedno zapnutie a vypnutie práve jednej aplikácie. Každá dátová sada by teda mala obsahovať podobný obnos informácií. Ako je ale možné vidieť v tabuľkách 6.1 až 6.6, dátové sady vytvorené pomocou Android Studia majú vyšší počet paketov. To je spôsobené práve použitím rozdielnych metód pri tvorbe týchto dátových sád. Dôvod tohoto rozdielu bol načrtnutý v podkapitole 5.1, v rámci popisu implementácie metódy použitia Android Studia.

Dátové sady, ktoré boli vytvorené použitím virtuálneho Android zariadenia obsahujú len sieťovú komunikáciu daného Android zariadenia (prípadne pár jednotiek paketov ktoré patria k zariadeniu tvoriace záznam). Ako je načrtnuté v kapitole 5.1, pri použití Android Studia na tvorbu dátových sád je predpoklad prítomnosť komunikácie, ktorá nepatrí Android zariadeniu. Tento predpoklad vychádza z faktu, že pri použití Android Studia sú dané Android zariadenie a zariadenie, na ktorom je spustené, z pohľadu siete totožné. Tým, že boli všetky dátové sady vytvorené použitím rovnakého postupu, je možné považovať tieto rozdiely v počtoch paketov ako potvrdenie predpokladu, ktorý bol popísaný.

Na základe toho je možné povedať, že všetky dátové sady obsahujú komunikáciu daného zariadenia. Rozdiel spočíva v tom, či sa jedná o výhradne komunikáciu Android zariadenia, alebo o komunikáciu zariadenia, ktorej časť patrí použitému Android zariadeniu. Odhliadnuc od tohoto rozdielu, všetky vytvorené dátové sady sú použiteľné na tvorbu profilov.

6.3 Dátové sady použité na testovanie

Táto skupina dátových sád predstavuje množinu dátových sád, ktoré sú použité pri vyhľadávaní profilov v reálnych záznamoch. Postup, akým sú tieto dátové sady tvorené je podobný tomu, ako je tvorená predošlá skupina dátových sád (6.1). Na rozdiel od predošlej, v tejto skupine sa nachádzajú aj dátové sady, ktoré obsahujú spustenie viacerých aplikácií, či naopak žiadnej aplikácie. Pri tvorbe dátovej sady v tejto skupine, ktorá má obsahovať práve jedno spustenie konkrétnej aplikácie, pred vytvorením takejto sady sú obe zariadenia vypnuté, aby sa uistilo ukončenie spojení, ktoré by mohli ovplyvniť vyhľadávanie.

Časť týchto dátových sád je vytvorené s tým, že pri testovaní bude známe, ktorá aplikácia bola počas tvorby sady spustená. Druhá časť sád je poskytnutá bez tejto znalosti a má slúžiť na vytvorenie predstavy, ako je možné využiť profily bez toho, aby bolo jasné či je identifikácia aplikácie správna.

V tabuľke 6.8, stĺpec „Spustená aplikácia v dátovej sade“ značí, ktorá aplikácia bola počas tvorby sady zapnutá. Vytvorenie takejto dátovej sady prebehlo tak, že pred vytvorením záznamu bolo zariadenie vypnuté, aby sa ukončili všetky spojenia, ktoré mohli ostať otvorené a pred zapnutím aplikácie sú vymazané jej cache dáta. Dátová sada „All applications“ obsahuje záznam sieťovej komunikácie, počas ktorej boli postupne spustené všetky použité aplikácie. Dátová sada „No applications“ naopak obsahuje záznam komunikácie, počas ktorej nebola spustená žiadna z použitých aplikácie. Pri dátových sadách označených ako „Unknown“ nie je známe či a ktorá aplikácia bola spustená počas tvorby záznamu.

Spustená aplikácia v dátovej sade	Počet paketov
No applications	2508
All applications	9127
Seznam.cz	4007
Reddit	743
Můj vlak	717
Cestovné poriadky CP	1275
Unknown 0	1236
Unknown 1	7461

Tabuľka 6.8: Dátové sady použité pre testovanie vyhľadávania profilov

Ak je v riadku konkrétne meno aplikácie, značí to, že žiadna iná aplikácia nebola počas tvorby záznamu zapnutá.
Pri dátových sadách, kde nie je známe ktoré aplikácie boli pri tvorbe záznamu zapnuté, je možné predpokladať štyri možnosti.

1. Žiadna aplikácia nebola zapnutá
2. Všetky aplikácie boli zapnuté
3. Viac aplikácií bolo zapnutých
4. Jedna aplikácia bola zapnutá

6.4 Extrahované dáta z dátových sád

V podkapitole 5.2 sú popísané informácie získavané z dátových sád ako aj postup ich získania. V tejto podkapitole sú popísané poznatky k týmto informáciám na základe už vytvorených dátových sád. Tieto poznatky boli získané manuálnou kontrolou dátových sád. Konkrétne sa jedná o položky ktoré by mali tvoriť jednotlivé profily, teda JA3 odtlačok, IP adresy a hostname.

IP adresy

Podobne ako je popísané v kapitole 5.2, IP adresy je možné rozdeliť do dvoch kategórií. IP adresy, ktoré sú priamo použité na smerovanie dát (t.j. Informácie uložené na internetovej vrstve paketu) a IP adresy, ktoré sú získané použitím DNS. Vytvorený profil obsahuje obe tieto množiny adries.

Množina IP adries, ktoré sú získané z DNS dotazov a odpovedí je prirodzene menšia až rovnaká. Je možné ju považovať za presnejšiu, ak sa predpokladá, že všetky DNS dotazy a odpovede v dátovej sade patria skúmanej aplikácií. Z tohoto predpokladu teda vyplýva, že všetky IP adresy získané týmto spôsobom taktiež patria k aplikácií. Nevýhoda použitia výhradne tejto množiny IP adries v profile spočíva v situácií, kedy aplikácia z nejakého dôvodu nepotrebuje zistiť IP adresu cez DNS (napr. je adresa uložená v rámci cache pamäte). Vtedy aj napriek tomu, že aplikácia s takou IP adresou komunikuje, v profile zaradená nie je a tým je profil nepresnejší.

Výhodou množiny IP adries, ktoré sú získané z internetovej vrstvy paketov spočíva v tom, že žiadna IP adresa s ktorou zariadenie komunikuje nie je vynechané. Tým je možné

odstrániť situáciu popísanú v predošlom paragrafe. Nevýhodou použitia tejto množiny spočíva práve v tom, že každá IP adresa s ktorou zariadenie komunikovalo je asociovaná s aplikáciou. Na čiastočné odstránenie tejto nevýhody slúži práva informácia o počte výskytov danej IP adresy. Vyšší počet výskytov predstavuje vyššiu istotu o tom, že daná IP adresa patrí ku komunikácií aplikácie a nejedná sa len o šum.

JA3 odtlačky a hostname

Funkcia použitá na získanie JA3 odtlačkov je implementovaná autormi tejto metódy. Vytvorené odtlačky tomu taktiež zodpovedajú. V podkapitole 2.2 je popísaný postup vytvorenia odtlačku JA3. Zoznam rozšírení je použitý ako časť reťazca, z ktorého je následne vytvorený odtlačok. Tento zoznam je tvorený typmi konkrétnych rozšírení, ktoré je možné vidieť na stránke [7]. Tým, že je použitý typ rozšírenia a nie jeho obsah, je tento zoznam použitých rozšírení všeobecnejší, čo môže znížiť použiteľnosť JA3 odtlačku pri identifikácií aplikácie.

Tým, že nie je k dispozícií zoznam hostname ktoré aplikácia používa, tento zoznam je vytvorený z DNS odpovedí, ktoré sa v dátovej sade nachádzajú. To za predpokladu, že každá DNS odpoveď patrí k aplikácií. Pri hostname, ktorých počet výskytov je vysoký je možné predpokladať ich spojenie s aplikáciou. Nevýhoda tejto metódy je, že hostname môže patriť k aplikácií, no IP adresu ktorá tejto hostname odpovedá je uložená v cache pamäti aplikácie a tým nie je potrebné ju zisťovať za každým. Takto je možné uvažovať prevažne pri hostname, ktorých počet výskytu sa pohybuje okolo hranice 50%. To z toho dôvodu, že polovica dátových sád pre každý profil je tvorená s vymazaním cache pamäte.

Predspracovanie dátových sád

Po dohode s vedúcou práce boli v rámci predspracovania dátových sád všetky dátové sady manuálne skontrolované, či sa v zázname nachádza celá sieťová komunikácia mobilnej aplikácie. Okrem manuálnej kontroly je predspracovanie dát súčasťou samotnej tvorby profilu.

Kapitola 7

Experimentálne výsledky

Túto kapitolu je možné rozdeliť na dve hlavné časti. Prvá časť sa venuje samotnému profilovaniu aplikácií na základe sieťovej komunikácie, rôznym nastaveniam pri profilovaní, popisu a porovnaniu jednotlivých profilov.

Druhá časť sa venuje experimentovaniu so získanými profilmi. Experimenty spočívajú vo vyhľadávaní profilov v rôznych dátových sadách. Pod pojmom vyhľadávať je možné rozumieť, s akou istotou je možné povedať, že bola daná aplikácia zapnutá počas tvorby dátovej sady, na základe profilu ktorý aplikáciu popisuje.

Použitie portov pri tvorbe profilu

Ako je popísaný predpoklad o využití portov v kapitole 4.2, cieľom tohoto testu bolo overiť, či je možné použiť porty ako identifikačný faktor profilu. V tabuľkách C.1 až C.4 sú zobrazené porty, ktorých výskyt bol vyšší ako 50%, čím je možné ich považovať za relevantné.

Tieto tabuľky potvrdzujú, že žiadna z použitých aplikácií nepoužívala port, ktorý by bol pre ňu špecifický. Medzi porty, ktorých výskyt bol aspoň 50% patria porty 53, 80 a 433. Všetky tri porty majú priradenú určitú službu. Podľa [11, 28–29], port 53 adresuje službu DNS, port 443 identifikuje službu využívajúcu zabezpečeného prenosu a port 80 identifikuje prenos WWW stránok. V tabuľkách sa vyskytuje aj port 5353, ktorý je podľa [8] používaný pre Multicast DNS.

Z týchto zistení je možné potvrdiť, že porty v tomto prípade nie sú použiteľné ako identifikačný faktor profilu.

Internetový prehliadač a aplikácie

Pri porovnaní komunikácie aplikácie a internetového prehliadača na ekvivalent aplikácie neboli zistené žiadne rozdiely, ktoré by bolo možné brať ako smerodajné pri profilovaní aplikácie. Jediné zistenie, ktoré platí pri všetkých porovnaniach je to, že pri použití internetového prehliadača dátové sady obsahujú vždy len jeden relevantný JA3 odtlačok. Avšak tento odtlačok nie je špecifický pre prehliadač ako taký, pretože tento otlačok je nájdený aj v dátových sadách pre dve aplikácie.

Druhý rozdiel zistený z tohoto porovnania je to, že pre všetky aplikácie platí, že počet relevantných položiek hostanme a IP adries je menší, ako pri ich ekvivalente cez internetový prehliadač.

7.1 Vytvorené profily

V rámci bakalárskej práce bolo vytvorených 9 profilov pre aplikácie *Můj vlak*, *Reddit* a *Seznm.cz*. Pre aplikáciu *CP* bolo vytvorených 7 profilov. To zahŕňa aj modifikované profily. Všetky profily boli vytvorené z dátových sád popísaných v kapitole 6 a použitím postupov popísaných v kapitole 5.2. Každý profil je reprezentovaný inštanciou triedy `AppProfile` (výpis 5.4).

V následujúcich paragrafoch sú zobrazené vybrané profily, rozdelené podľa typu modifikácie. Tabuľky popisujúce jednotlivé profily sú upravené tak, aby zobrazovali tie položky profilu, ktoré je na základe ich počtu výskytov možné brať za relevantné. Položky s nižším počtom výskytov teda nie sú z dôvodu prehľadnosti v tejto časti práce zobrazené. Všetky vytvorené profily s ich všetkými položkami sú dostupné v priloženom pamäťovom médiu.

Všetky tabuľky, ktoré reprezentujú profil, obsahujú informácie o tom akú aplikáciu profil reprezentuje, koľko dátových sád bolo použitých na vytvorenie profilu, položky s najvyšším počtom výskytov a samotný počet výskytov všetkých položiek. Počet výskytu položiek značí, v koľkých dátových sadách sa daná položka vyskytla.

Základné profily

Základný profil reprezentuje sieťovú komunikáciu aplikácie pri konkrétnej kombinácií spustenia Android zariadenia a Android verzie. Na vytvorenie týchto profilov boli použité množiny dátových sád popísaných v 6.1. Pre aplikácie *Můj vlak*, *Reddit* a *Seznam.cz* bolo vytvorených 6 základných profilov a pre aplikáciu *Cestovné poriadky CP* bolo vytvorené 4 základné profily.

Tabuľky 7.1 až 7.4 reprezentujú základné profily všetkých použitých aplikácií spustených na virtuálnom Android zariadení. Tieto profily boli vybrané na základe najvyššieho počtu dátových sád.

Aplikácia	Cestovné poriadky CP		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
c60d01d600aacc2c04844595ce224279	100	216.58.201.66	79
3967ff2d2c9c4d144e7e30f24f4e9761	100	172.217.23.194	70
d8c87b9bfde38897979e41242626c2f3	100	172.217.23.202	56
66918128f1b9b03303d77c6f2eefd128	100	172.217.23.193	52
graph.facebook.com	96	216.58.201.65	51
e.crashlytics.com	95	216.58.201.67	51
googleads.g.doubleclick.net	50	172.217.23.206	50
pagead2.googlesyndication.com	46	172.217.23.195	50
pagead2.googleadservices.com	41	95.129.96.199	50
147.229.190.143	100	178.32.212.40	48
157.240.30.18	100	172.217.23.226	48
77.93.203.180	100	216.58.201.98	44
224.0.0.251	89	51.254.91.248	42

Tabuľka 7.1: Základný profil aplikácie *Cestovné poriadky CP*, na vytvorenie bolo použitých 100 dátových sád, ktoré boli vytvorené použitím virtuálneho stroja s operačným systémom Android 9

Aplikácia	Můj vlak		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d8c87b9bfde38897979e41242626c2f3	100	172.217.23.206	41
6f5e62edfa5933b1332ddf8b9fb3ef9d	100	172.217.23.238	37
e.crashlytics.com	95	224.0.0.22	28
settings.crashlytics.com	24	172.217.23.232	26
www.google-analytics.com	20	10.0.2.15	25
ipws2.timetable.cz	17	216.58.201.110	20
ssl.google-analytics.com	10	10.0.2.3	15
82.117.128.63	100	172.217.23.202	11
192.168.1.10	97	216.58.201.78	11
224.0.0.251	83	216.58.201.72	10
172.217.23.227	50		

Tabuľka 7.2: Základný profil aplikácie *Můj vlak*, na vytvorenie bolo použitých 100 dátových sád, ktoré boli vytvorené použitím virtuálneho stroja s operačným systémom Android 9

Pri prvom pohľade na tabuľky 7.1 a 7.2 je možné vidieť, určité spoločné prvky týchto profilov. Tieto profily obsahujú nízky počet hostname, ktoré aplikácia používala častejšie ako v polovici prípadov, o ktorých taktiež nie je možné prehlásiť, že sú špecifické pre danú aplikáciu.

Aplikácia	Reddit		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d8c87b9bfde38897979e41242626c2f3	100	register.appsflyer.com	44
6f5e62edfa5933b1332ddf8b9fb3ef9d	100	reports.crashlytics.com	41
3967ff2d2c9c4d144e7e30f24f4e9761	95	b.thumbs.redditmedia.com	38
gql.reddit.com	100	www.redditstatic.com	35
www.reddit.com	100	styles.redditmedia.com	31
oauth.reddit.com	100	192.168.1.10	100
strapi.reddit.com	100	151.101.193.140	98
e.reddit.com	100	151.101.129.140	92
t.appsflyer.com	87	151.101.1.140	88
alb.reddit.com	87	151.101.65.140	84
api.branch.io	67	224.0.0.251	77
gateway.reddit.com	66	52.222.147.6	64
cdn.branch.io	63	74.125.206.188	52
v.redd.it	56	172.217.23.227	50
external-preview.redd.it	55	13.227.219.145	37
e.crashlytics.com	50	172.217.23.202	34
preview.redd.it	50	216.58.201.78	31
api.appsflyer.com	46		

Tabuľka 7.3: Základný profil aplikácie *Reddit*, na vytvorenie bolo použitých 100 dátových sád, ktoré boli vytvorené použitím virtuálneho stroja s operačným systémom Android 9

Aplikácia	Seznam.cz		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
66918128f1b9b03303d77c6f2eefd128	100	77.75.76.55	76
d8c87b9bfde38897979e41242626c2f3	100	23.202.52.244	71
6f5e62edfa5933b1332ddf8b9fb3ef9d	100	77.75.77.37	66
3967ff2d2c9c4d144e7e30f24f4e9761	100	77.75.74.131	64
d32-a.sdn.cz	100	77.75.74.134	62
d39-a.sdn.cz	100	77.75.78.146	61
d32-a.sdn.szn.cz	100	77.75.76.72	60
d53-a.sdn.cz	100	77.75.75.108	58
d27-a.sdn.cz	100	77.75.75.107	57
ads.pubmatic.com	100	77.75.75.101	57
d15-a.sdn.cz	99	77.75.77.9	56
e.crashlytics.com	97	77.75.78.31	56
graph.facebook.com	96	77.75.74.132	54
d50-a.sdn.cz	95	77.75.77.43	54
data.flurry.com	94	77.75.74.133	54
d48-a.sdn.cz	92	77.75.77.212	54
d62-a.sdn.cz	92	77.75.74.137	53
d116-a.sdn.cz	89	77.75.79.16	52
v116-a.sdn.cz	84	77.75.75.104	52
login.seznam.cz	52	77.75.75.106	52
d41-a.sdn.szn.cz	50	77.75.75.102	52
d41-a.sdn.cz	50	87.248.118.22	52
147.229.190.143	100	77.75.74.136	52
157.240.30.18	100	87.248.118.23	51
185.66.190.10	100	77.75.79.133	51
224.0.0.251	87	172.217.23.195	50
77.75.78.55	77		

Tabuľka 7.4: Základný profil aplikácie *Seznam.cz*, na vytvorenie bolo použitých 100 dátových sád, ktoré boli vytvorené použitím virtuálneho stroja s operačným systémom Android 9

V porovnaní s predošlými základnými profilmi, pri profiloch zobrazených v tabuľkách 7.3 a 7.4 je možné vidieť vyšší počet relevantných položiek. Taktiež ale tieto profily obsahujú položky, ktoré sú zdielané naprieč viacerými až všetkými základnými profilmi.

Špecifické profily

Špecifické profily sú vytvorené zo základných profilov, ktoré boli popísané v predošlom paragrafe. Tento typ profilu obsahuje len položky, ktoré sú jedinečné pre profil v porovnaní s ostatnými vytvorenými profilmi. To znamená, že ak profil obsahuje určitú položku, žiaden iný profil v danej kategórií túto položku neobsahuje. Cieľom vytvorenia a použitia tohoto typu profilu bolo zníženie šance, že bude daný profil identifikovateľný v zázname, v ktorom by identifikovateľný nemal byť (tzv. *false positive*).

V tabuľkách 7.5 až 7.8, ktoré zobrazujú špecifické profily pre všetky aplikácie na virtuálnom Android zariadení so systémom Android 9.0, je na prvý pohľad vidieť rozdiel v počte JA3 odtlačkov. Bližšie je JA3 odtlačok rozoberaný v podkapitole 7.2.

Taktiež je možné vidieť rozdielnu zmenu v počte relevantných položiek pri vytvorení špecifických profilov oproti základným profilom. Zatiaľ čo špecifické profily *CP* a *Můj vlak* veľmi nízky počet relevantných položiek, pri špecifických profiloch *Reddit* a *Seznam.cz* sa tieto počty zmenili minimálne.

Aplikácia	Cestovné poriadky CP		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
pagead2.googlesyndication.com	46	gask.hit.gemius.pl	10
pagead2.googleadservices.com	41	77.93.203.180	100
tpc.googlesyndication.com	17	172.217.23.193	52
fonts.googleapis.com	16	95.129.96.199	50
fonts.gstatic.com	16	178.32.212.40	48
main.crws.cz	14	51.254.91.248	42
lh3.googleusercontent.com	14	151.80.66.32	12
www.googletagservices.com	13		

Tabuľka 7.5: Špecifický profil aplikácie *Cestovné poriadky CP*, vytvorený použitím vytvorených základných profilov (popísaných v 7.1 až 7.4) a maximálny počet výskytov je 100 dátových sád

Aplikácia	Můj vlak		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
ipws2.timetable.cz	17	82.117.128.63	100
ssl.google-analytics.com	10		

Tabuľka 7.6: Špecifický profil aplikácie *Můj vlak*, vytvorený použitím vytvorených základných profilov (popísaných v 7.1 až 7.4) a maximálny počet výskytov je 100 dátových sád

Aplikácia	Reddit		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
gql.reddit.com	100	preview.redd.it	50
www.reddit.com	100	api.appsflyer.com	46
oauth.reddit.com	100	register.appsflyer.com	44
strapi.reddit.com	100	reports.crashlytics.com	41
e.reddit.com	100	b.thumbs.redditmedia.com	38
t.appsflyer.com	87	www.redditstatic.com	35
alb.reddit.com	87	styles.redditmedia.com	31
api.branch.io	67	151.101.193.140	98
gateway.reddit.com	66	151.101.129.140	92
cdn.branch.io	63	151.101.1.140	88
v.redd.it	56	52.222.147.6	64
external-preview.redd.it	55	13.227.219.145	37

Tabuľka 7.7: Špecifický profil aplikácie *Reddit*, vytvorený použitím vytvorených základných profilov (popísaných v 7.1 až 7.4) a maximálny počet výskytov je 100 dátových sád

Aplikácia	Seznam.cz		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d32-a.sdn.cz	100	77.75.74.134	62
d39-a.sdn.cz	100	77.75.78.146	61
d32-a.sdn.szn.cz	100	77.75.76.72	60
d53-a.sdn.cz	100	77.75.75.108	58
d27-a.sdn.cz	100	77.75.75.107	57
ads.pubmatic.com	100	77.75.75.101	57
d15-a.sdn.cz	99	77.75.77.9	56
d50-a.sdn.cz	95	77.75.78.31	56
data.flurry.com	94	77.75.74.132	54
d48-a.sdn.cz	92	77.75.77.43	54
d62-a.sdn.cz	92	77.75.74.133	54
d116-a.sdn.cz	89	77.75.77.212	54
v116-a.sdn.cz	84	77.75.74.137	53
login.seznam.cz	52	77.75.79.16	52
d41-a.sdn.szn.cz	50	77.75.75.104	52
d41-a.sdn.cz	50	77.75.75.106	52
185.66.190.10	100	77.75.75.102	52
77.75.78.55	77	87.248.118.22	52
77.75.76.55	76	77.75.74.136	52
23.202.52.244	71	87.248.118.23	51
77.75.77.37	66	77.75.79.133	51
77.75.74.131	64		

Tabuľka 7.8: Špecifický profil aplikácie *Seznam.cz*, vytvorený použitím vytvorených základných profilov (popísaných v 7.1 až 7.4) a maximálny počet výskytov je 100 dátových sád

V tabuľkách 7.7 a 7.8 je možné vidieť, že počet položiek, ktorých výskyt je takmer 100%, je vyšší pri atribúte hostname. Položky IP adries naopak prevažujú v profiloch reprezentovaných tabuľkami 7.5 a 7.6

Super profily

Super profil predstavuje presný opak špecifického profilu. Je vytvorený zo všetkých dostupných profilov pre danú aplikáciu. Všetky profily z ktorých je super profil vytvorený reprezentujú komunikáciu danej aplikácie za určitej konfigurácie. Cieľom vytvorenia takéhoto profilu bolo obsiahnutie sieťovej komunikácie aplikácie pri rôznych spôsoboch spustenia Android zariadenia či Android verzie.

V tabuľkách 7.9 až 7.12 sú zobrazené super profily pre všetky použité aplikácie. Tieto profily boli vytvorené zo všetkých dostupných základných profilov predstavujúcich komunikáciu aplikácie.

Aplikácia	Cestovné poriadky CP		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
c60d01d600aacc2c04844595ce224279	120	172.217.23.194	109
d8c87b9bfde38897979e41242626c2f3	120	216.58.201.66	107
3967ff2d2c9c4d144e7e30f24f4e9761	120	172.217.23.226	96
66918128f1b9b03303d77c6f2eefd128	100	172.217.23.202	88
graph.facebook.com	154	172.217.23.206	80
e.crashlytics.com	153	95.129.96.199	80
googleads.g.doubleclick.net	78	216.58.201.65	78
pagead2.googlesyndication.com	75	216.58.201.98	71
pagead2.googleadservices.com	58	51.254.91.248	68
77.93.203.180	160	172.217.23.195	66
224.0.0.251	125	178.32.212.40	66
147.229.190.143	120	216.58.201.67	56
157.240.30.18	120	172.217.23.193	54

Tabuľka 7.9: Super profil pre aplikáciu *Cestovné poriadky CP*, vytvorený zo 160 dátových sád, ktoré boli vytvorené použitím všetkých kombinácií spustenia Android zariadenia a Android verzie, pri ktorých bolo možné dátové sady vytvoriť

Aplikácia	Můj vlak		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d8c87b9bfde38897979e41242626c2f3	120	172.217.23.238	76
6f5e62edfa5933b1332ddf8b9fb3ef9d	120	172.217.23.206	73
e.crashlytics.com	164	172.217.23.227	70
82.117.128.63	200	216.58.201.110	61
224.0.0.251	151	239.255.255.250	56
192.168.1.10	133	224.0.0.252	55

Tabuľka 7.10: Super profil pre aplikáciu *Můj vlak*, vytvorený z 200 dátových sád, ktoré boli vytvorené použitím všetkých kombinácií spustenia Android zariadenia a Android verzie

Tým, že tieto profily zlučujú profily rôznych konfigurácií pre jednu aplikáciu, je možné z nich odvodiť ktoré položky je možné asociovať s aplikáciou bez ohľadu na Android verziu či spôsob spustenia zariadenia.

Aplikácia	Reddit		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d8c87b9bfde38897979e41242626c2f3	120	gateway.reddit.com	127
6f5e62edfa5933b1332ddf8b9fb3ef9d	120	cdn.branch.io	126
3967ff2d2c9c4d144e7e30f24f4e9761	111	preview.redd.it	103
strapi.reddit.com	198	external-preview.redd.it	103
www.reddit.com	196	224.0.0.251	167
gql.reddit.com	196	192.168.1.10	140
oauth.reddit.com	194	151.101.193.140	136
e.reddit.com	194	151.101.129.140	125
t.appsflyer.com	160	151.101.1.140	122
alb.reddit.com	145	151.101.65.140	117
api.branch.io	136		

Tabuľka 7.11: Super profil pre aplikáciu *Reddit*, vytvorený z 200 dátových sád, ktoré boli vytvorené použitím všetkých kombinácií spustenia Android zariadenia a Android verzie

Aplikácia	Seznam.cz		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
6f5e62edfa5933b1332ddf8b9fb3ef9d	120	157.240.30.18	160
d8c87b9bfde38897979e41242626c2f3	120	77.75.78.55	144
3967ff2d2c9c4d144e7e30f24f4e9761	120	77.75.76.55	144
66918128f1b9b03303d77c6f2eefd128	101	77.75.77.9	132
d53-a.sdn.cz	199	185.66.190.10	129
d32-a.sdn.cz	194	77.75.78.31	126
ads.pubmatic.com	190	77.75.79.43	118
graph.facebook.com	187	77.75.77.37	118
d48-a.sdn.cz	186	77.75.74.134	118
e.crashlytics.com	177	77.75.74.132	116
data.flurry.com	177	77.75.74.131	116
d39-a.sdn.cz	171	87.248.118.22	116
d15-a.sdn.cz	170	77.75.74.136	114
d32-a.sdn.szn.cz	170	77.75.74.137	112
d50-a.sdn.cz	165	77.75.75.104	112
d62-a.sdn.cz	162	77.75.74.133	109
d116-a.sdn.cz	149	77.75.74.139	109
v116-a.sdn.cz	130	77.75.76.72	108
d27-a.sdn.cz	129	77.75.75.107	108
d41-a.sdn.cz	107	77.75.75.108	106
a.tribalfusion.com	105	77.75.75.106	105
d41-a.sdn.szn.cz	104	212.82.100.176	105
sync.mathtag.com	103	77.75.77.16	105
sync-tm.everesttech.net	103	46.228.164.11	104
pubmatic-match.dotomi.com	102	77.75.77.212	102
224.0.0.251	162	77.75.75.103	100
147.229.190.143	160		

Tabuľka 7.12: Super profil pre aplikáciu *Seznam.cz*, vytvorený z 200 dátových sád, ktoré boli vytvorené použitím všetkých kombinácií spustenia Android zariadenia a Android verzie

Pri super profiloch *Reddit*(7.11) a *Seznam.cz*(7.12) je možné vidieť, že aj na priek tomu, že žiadna položka nemá výskyt vo všetkých dátových sadách, najvyšší počet výskytov hostname presahuje najvyšší počet výskytov IP adresy približne o 15–20%.

Špecifické super profily

Tento typ profilu je vytvorený zo super profilov. Na vytvorenie týchto profilov boli využité super profily popísané v predošlom paragrafe.

Špecifický super profil je vytvorený odstránením položiek, ktoré sa vyskytujú aspoň v jednom inom dostupnom super profile. Výsledkom tohoto procesu je profil, ktorý obsahuje jedinečné prvky komunikácie aplikácie na rôznych Android verziách s rôznym spustením zariadenia. Špecifické super profily pre všetky použité aplikácie sú popísané v tabuľkách 7.13 až 7.16.

Využitím super profilu je dosiahnutý rozsiahlejší popis sieťovej komunikácie aplikácie. Následnou transformáciou na špecifický super profil je dosiahnutá jedinečnosť profilu. Táto kombinácia zvyšuje šancu na identifikáciu aplikácie v zázname a zároveň znižuje šancu na falošnú identifikáciu aplikácie.

Aplikácia	Cestovné poriadky CP		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
pagead2.googlesyndication.com	75	95.129.96.199	80
pagead2.googleadservices.com	58	51.254.91.248	68
77.93.203.180	160	178.32.212.40	66

Tabuľka 7.13: Špecifický super profil aplikácie *Cestovné poriadky CP*, vytvorený použitím vytvorených super profilov (popísaných v 7.9 až 7.12) a maximálny počet výskytov je 160 dátových sád

Aplikácia	Můj vlak		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
ipws2.timetable.cz	28	82.117.128.63	200
ssl.google-analytics.com	13	95.216.8.29	40

Tabuľka 7.14: Špecifický super profil aplikácie *Můj vlak*, vytvorený použitím vytvorených super profilov (popísaných v 7.9 až 7.12) a maximálny počet výskytov je 200 dátových sád

V tabuľkách 7.13 a 7.14 je možné vidieť, že oba tieto profily obsahujú len jednu IP adresu, ktorá sa vyskytla v pri každej komunikácií aplikácie.
Je taktiež možné vidieť, že ani jeden z týchto profilov neobsahuje hostname, ktorú by aplikácia použila pri tvorbe viac ako polovice dátových sád bez toho aby táto hostname bola použitá inými aplikáciami.
Z toho je možné vidieť, že komunikáciu aplikácií *CP* a *Můj vlak* je možné presnejšie identifikovať zameraním sa na IP adresy bez toho, aby boli tieto profily zameniteľné.

Aplikácia	Reddit		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
strapi.reddit.com	198	api.appsflyer.com	80
www.reddit.com	196	reports.crashlytics.com	79
gql.reddit.com	196	register.appsflyer.com	79
oauth.reddit.com	194	styles.redditmedia.com	75
e.reddit.com	194	www.redditstatic.com	73
t.appsflyer.com	160	b.thumbs.redditmedia.com	55
alb.reddit.com	145	151.101.193.140	136
api.branch.io	136	151.101.129.140	125
gateway.reddit.com	127	151.101.1.140	122
cdn.branch.io	126	52.222.147.6	77
preview.redd.it	103	151.101.113.140	60
external-preview.redd.it	103	99.86.237.244	60
v.redd.it	96		

Tabuľka 7.15: Špecifický super profil aplikácie *Reddit*, vytvorený použitím vytvorených super profilov (popísaných v 7.9 až 7.12) a maximálny počet výskytov je 200 dátových sád

Z profilov 7.15 a 7.16 je možné vidieť, že žiadna položka komunikácie sa nevyskytuje vo všetkých dátových sadách. Na priek tomu, oba profily obsahujú v pomere viac hostname položiek, ktorých výskyt presahuje 90%. Tieto položky je možné považovať za nevyhnutné pre danú aplikáciu
Pri profile 7.15 mali tri IP adresy výskyt v rozmedzí 50% až 75%. Profil je špecifický, z čoho je možné predpokladať oprávnenú asociáciu týchto adries s aplikáciou, pričom nespadajú do kategórie nevyhnutných položiek komunikácie aplikácie. Tým, že nevyhnutné položky tohoto profilu sú tvorené výhradne hostname atribútmi, je možné predpokladať presnejšiu identifikáciu tejto aplikácie pri zvýšení dôrazu na kontrolu hostname.

Aplikácia	Seznam.cz		
JA3/Hostname/IP	Výskyt	JA3/Hostname/IP	Výskyt
d53-a.sdn.cz	199	185.66.190.10	129
d32-a.sdn.cz	194	77.75.78.31	126
ads.pubmatic.com	190	77.75.79.43	118
d48-a.sdn.cz	186	77.75.77.37	118
data.flurry.com	177	77.75.74.134	118
d39-a.sdn.cz	171	77.75.74.132	116
d15-a.sdn.cz	170	77.75.74.131	116
d32-a.sdn.szn.cz	170	77.75.74.136	114
d50-a.sdn.cz	165	77.75.74.137	112
d62-a.sdn.cz	162	77.75.75.104	112
d116-a.sdn.cz	149	77.75.74.133	109
v116-a.sdn.cz	130	77.75.74.139	109
d27-a.sdn.cz	129	77.75.76.72	108
d41-a.sdn.cz	107	77.75.75.107	108
a.tribalfusion.com	105	77.75.75.108	106
d41-a.sdn.szn.cz	104	77.75.75.106	105
sync.mathtag.com	103	212.82.100.176	105
sync-tm.everesttech.net	103	77.75.77.16	105
pubmatic-match.dotomi.com	102	46.228.164.11	104
77.75.78.55	144	77.75.77.212	102
77.75.76.55	144	77.75.75.103	100
77.75.77.9	132		

Tabuľka 7.16: Špecifický super profil aplikácie *Seznam.cz*, vytvorený použitím vytvorených super profilov (popísaných v 7.9 až 7.12) a maximálny počet výskytov je 200 dátových sád

Profil 7.16 má s profilom 7.15 spoločné to, že množinu nevyhnutných nutných položiek komunikácie tvoria výhradne hostname. Oproti ostatným profilom však obsahuje väčšie množstvo ako IP adries, tak hostname.

Tak ako pri profile 7.15, ani v tomto profile nie je IP adresa s výskytom vyšším ako 75%. Preto téza, ktorá bola vyslovená pri predošlom profile platí aj v tomto prípade.

Zhodnotenie

Výhoda základných profilov spočíva v tom, že z profilu je jasne vidieť správanie aplikácie za danej konfigurácie vďaka tomu, že pre každú konfiguráciu je vytvorený profil. To je ale dôvodom, že bez rôznych úprav, porovnaní či spájaní s inými profilmi je náročné zistiť, ktoré atribúty patri aplikácií samotnej, ktoré sú ovplyvnené inými faktormi a podobne. Tieto profily sú ale nevyhnutné pre všetky ďalšie typy profilov.

Špecifické profily prinášajú výhody zistenia, čo je špecifické pre daný profil. Vďaka tomu je vidieť položky, ktoré boli prítomné len pri spustení danej aplikácie a na základe výskytu takej položky je možné predpokladať, či aplikácií patrí alebo nie. Odstránené položky môžu patriť iným procesom zariadenia, ktoré komunikovali nezávisle od aplikácie, no taktiež mohlo ísť o položky ktoré má aplikácia spoločné s inými aplikáciami. Ako príklad je možné uviesť hostname *e.crashlytics.com*, ktorá sa nachádza takmer v každom zo základných profilov, no v prípade že sa táto hostname nájde v zázname, nie je možné povedať,

ktorý aplikácia bola spustená. Preto je použitie špecifických profilov výhodné v prípade potreby zníženia šancí na falošné identifikovanie aplikácie.

Vytvorenie super profilov slúži k získaniu predstavy, ktoré položky profilu aplikácie sú spoločné aj napriek rôznym verziám operačného systému či spôsobu spustenia zariadenia. To je možné vidieť na základe výskytu jednotlivých položiek a počtu dátových sád, ktoré boli použité. Ná základe výskytu je možné prehľadne vidieť, ktoré položky sú pre akú aplikáciu najdôležitejšie.

Špecifické super profily prinášajú lepší prehlad o tom, ktoré položky komunikácie sú asociované priamo s aplikáciou, alebo sú závislé od ďalších faktorov. Hlavnou výhodou oproti super profilu samotnému je faktor jedinečnosti. Pri niektorých profiloch to môže spôsobiť nízky počet relevantných položiek. No napriek tomu, je možné vidieť vo všetkých štyroch profiloch, že obsahujú položky ktoré je možné považovať za najdôležitejšie pri procese identifikácie.

Treba brať na vedomie fakt, že to, že ak určitá položka nie je obsiahnutá v profile neznamená, že k aplikácií nepatrí. Tieto položky je možné vidieť pri porovnaní tabuliek 7.13 až 7.16 s tabuľkami 7.9 až 7.12. Špecifický super profil teda zobrazuje atribúty komunikácie aplikácie, ktoré sú pre aplikáciu jedinečné v porovnaní s inými testovanými aplikáciami.

Taktiež je vhodné poukázať na to, že nešpecifické profily môžu obsahovať položku IP adresy s relatívne vysokým počtom výskytu, pričom nepatrí priamo žiadnej aplikácií. Jedná sa napríklad o IP adresy DSN serveru. Tým, že pri tvorbe dátových sád bolo zariadenie pripojené v rôznych sieťach, líšia sa aj dotazované DNS servery. Tým, že tieto adresy ale nie sú špecifické pre žiadnu aplikáciu, vytvorením špecifického profilu je možné tieto položky odstrániť.

Je možné tieto profily rozdeliť na dve časti podľa toho, čo je účelom tvorby profilu. Ak je cieľom tvorby profilu čo najlepšie popísať správanie aplikácie zo sieťovej komunikácie, najviac zachytia základné profily spolu so super profilmi. Ak je cieľom tvorby profilu identifikácia aplikácie, je lepšie použiť dáta ktoré sú minimálne duplikované, na čo sú zamerané špecifické profily. Využitie jednotlivých profilov pri identifikácií v rôznych záznamoch je popísané v kapitole 7.3.

7.2 Porovnanie JA3 odtlačkov

V tejto sekcií sú popísané rôzne porovnania JA3 odtlačkov naprieč profilmi. Účelom týchto porovnaní bolo overiť použiteľnosť odtlačku pri identifikácií aplikácií.

Pre zjednodušenosť textu, výraz „profil vytvorený pomocou Android Studia či virtuálneho stroja“ znamená, že daný profil bol vytvorený z dátových sád, ktoré boli vytvorené použitím Android Studia či virtuálneho stroja.

Na základe článku [1] bola vyslovená téza, v ktorej je použitý JA3 odtlačok ako primárny faktor pri identifikovaní a profilovaní aplikácie ako klienta na základe sieťovej komunikácie. V tejto podkapitole sú uvedené experimenty, ktoré slúžia na overenie použiteľnosti tejto tézy.

Každá tabuľka obsahuje JA3 odtlačok a číslo, ktoré značí výskyt daného JA3 odtlačku v profile. Číslo **X/Y** značí, že daný JA3 odtlačok sa pri tvorbe profilu vyskytol v **X** dátových sadách z **Y** celkových (použitých na tvorbu daného profilu).

JA3 odtlačok	Můj vlak	Seznam.cz	CP	Reddit
859ce948f69709d8d1263981cf4dd16d	0/100	1/100	0/100	0/100
f912c10788b92e747685beebaaef6cd1	0/100	0/100	1/100	0/100
c60d01d600aacc2c04844595ce224279	0/100	1/100	100/100	0/100
839868ad711dc55bde0d37a87f14740d	1/100	0/100	0/100	0/100
c9610a96cb052bf20d5e4e45d81abed8	0/100	1/100	1/100	0/100
3967ff2d2c9c4d144e7e30f24f4e9761	0/100	100/100	100/100	95/100
346ba115157830275e4b9249de1d2bba	4/100	3/100	3/100	4/100
6f5e62edfa5933b1332ddf8b9fb3ef9d	100/100	100/100	0/100	100/100
6ec2896feff5746955f700c0023f5804	4/100	1/100	3/100	1/100
33490b1d5377580b19f7f9b5849d7991	2/100	3/100	0/100	2/100
9c815150ea821166faecf80757d8826a	4/100	2/100	2/100	0/100
66918128f1b9b03303d77c6f2eefd128	1/100	100/100	100/100	1/100
33b38c8778b29bc991b5200b6141f35e	5/100	3/100	3/100	2/100
d8c87b9bfde38897979e41242626c2f3	100/100	100/100	100/100	100/100
8f41a697eff27e008f969cf7b5ba4117	0/100	13/100	0/100	0/100
6db5068dc42d8cfb4bd5ee3e169db25e	1/100	0/100	0/100	0/100
ee26b1f1aec16d6098768e2c67388ace	7/100	5/100	5/100	0/100
554719594ba90b02ae410c297c6e50ad	0/100	1/100	0/100	0/100
e9ec38c2b40ff3e300e9975dd7619902	0/100	0/100	0/100	4/100

Tabuľka 7.17: Porovnanie výskytu JA3 odtlačkov pre 4 profily na Android 9.0

Tabuľka 7.17 popisuje výskyt jednotlivých JA3 odtlačkov naprieč profilmi. Zvýraznené riadky upozorňujú na konkrétne JA3 odtlačky, ktoré by mohli byť brané ako špecifické pre daný profil aplikácie vďaka výskytu vo väčšine dátových sád, z ktorých bol profil tvorený. Okrem výnimiek, ktoré sú popísané v rámci následujúcich paragrafov, je vidieť, že žiaden JA3 odtlačok nie je špecifický len pre danú aplikáciu.

Pri tvorbe profilu je snaha zoskupiť informácie, ktoré danú aplikáciu budú nie len identifikovať ale taktiež odlišovať od ostatných profilov. Ak teda JA3 odtlačok, ktorý je asociovaný s aplikáciou, je asociovaný aj s inou aplikáciou, vierohodnosť identifikácie daného profilu

aplikácie klesá.

Tabuľky 7.18 a 7.19 ukazujú, že myšlienka popísaná pri opise tabuľky 7.17 platí aj pre profily vytvorené na iných Android verziách a nejedná sa len o anomáliu jednej verzie.

JA3 odtlačok	Můj vlak	Seznam.cz	CP	Reddit
33490b1d5377580b19f7f9b5849d7991	1/20	0/20	0/20	0/20
5353c0796e25725adfdb93f35f5a18f7	0/20	20/20	20/20	0/20
d8c87b9bfde38897979e41242626c2f3	20/20	20/20	20/20	20/20
6f5e62edfa5933b1332ddf8b9fb3ef9d	20/20	20/20	0/20	20/20
c60d01d600aacc2c04844595ce224279	0/20	0/20	20/20	0/20
ee26b1f1aec16d6098768e2c67388ace	0/20	2/20	1/20	0/20
346ba115157830275e4b9249de1d2bba	1/20	1/20	1/20	1/20
3967ff2d2c9c4d144e7e30f24f4e9761	1/20	20/20	20/20	16/20
6ec2896feff5746955f700c0023f5804	0/20	1/20	0/20	0/20
9256f440671627d0578d923befd27285	0/20	20/20	0/20	0/20
9c815150ea821166faecf80757d8826a	0/20	1/20	1/20	1/20
33b38c8778b29bc991b5200b6141f35e	2/20	1/20	1/20	0/20
554719594ba90b02ae410c297c6e50ad	1/20	0/20	0/20	0/20

Tabuľka 7.18: Porovnanie výskytu JA3 odtlačkov pre 4 profily na Android 8.1

JA3 odtlačok	Můj vlak	Seznam.cz	CP	Reddit
629b587f706aee60430ec3879c6edb66	1/20	2/20	0/20	2/20
ac9fa1e796361fe02ab026354107ad60	0/20	0/20	1/20	0/20
6db5068dc42d8cfb4bd5ee3e169db25e	0/20	1/20	0/20	0/20
8aea43da272f1b65778b22467058be4b	0/20	0/20	1/20	0/20
31010a807900df95eb46e153206784a8	0/20	0/20	20/20	0/20
9fc6ef6efc99b933c5e2d8fcf4f68955	20/20	20/20	20/20	20/20
f6a0bfafe2bf7d9c79ffb3f269b64b46	20/20	20/20	0/20	20/20
1ac3fb696535812eb572f9f9e6336897	0/20	1/20	0/20	0/20
89ba18b00454b90030aefdb560ea88fd	0/20	20/20	20/20	10/20
554719594ba90b02ae410c297c6e50ad	0/20	1/20	0/20	0/20
bc6c386f480ee97b9d9e52d472b772d8	0/20	1/20	0/20	0/20
66918128f1b9b03303d77c6f2eefd128	0/20	1/20	0/20	0/20
ee26b1f1aec16d6098768e2c67388ace	0/20	1/20	0/20	0/20
33b38c8778b29bc991b5200b6141f35e	0/20	1/20	1/20	0/20
6ec2896feff5746955f700c0023f5804	0/20	0/20	1/20	1/20
9c815150ea821166faecf80757d8826a	0/20	1/20	0/20	1/20
698e36219f3979420fa2581b21dac7ec	0/20	20/20	0/20	0/20
8498fe4268764dbf926a38283e9d3d8f	0/20	20/20	20/20	0/20

Tabuľka 7.19: Porovnanie výskytu JA3 odtlačkov pre 4 profily na Android 7.1

Pri porovnaní JA3 odtlačkov pre profily naprieč Android verziami, je možné vidieť pri porovnaní tabuľky 7.17 a 7.18, že JA3 nájdené vo viacerých profiloch sa zhodujú. Na druhú stranu, pri porovnaní týchto odtlačkov s odtlačkami v tabuľke 7.19, výskyt odtlačkov sa

zhoduje, no odtlačok samotný sa líši. Na základe tejto informácie je možné predpokladať, že JA3 odtlačok nie je špecifický pre daný operačný systém a jeho verziu.

JA3 odtlačok	AS Můj vlak	VM Můj vlak	AS Seznam	VM Seznam
bd1073cbb91ac524017faf1cee3bbd68	20/20	0/100	20/20	0/100
859ce948f69709d8d1263981cf4dd16d	0/20	0/100	0/20	1/100
ebf5e0e525258d7a8dcb54aa1564ecbd	0/20	0/100	20/20	0/100
839868ad711dc55bde0d37a87f14740d	0/20	1/100	0/20	0/100
e9ec38c2b40ff3e300e9975dd7619902	0/20	0/100	0/20	0/100
78c922015590b3a7822b159f62b05fc8	1/20	0/100	20/20	0/100
33b38c8778b29bc991b5200b6141f35e	0/20	5/100	0/20	3/100
6ec2896feff5746955f700c0023f5804	0/20	4/100	0/20	1/100
f912c10788b92e747685beebaaef6cd1	0/20	0/100	0/20	0/100
6db5068dc42d8cfb4bd5ee3e169db25e	0/20	1/100	0/20	0/100
3967ff2d2c9c4d144e7e30f24f4e9761	0/20	0/100	0/20	100/100
ee26b1f1aec16d6098768e2c67388ace	0/20	7/100	0/20	5/100
a0e9f5d64349fb13191bc781f81f42e1	0/20	0/100	1/20	0/100
6f5e62edfa5933b1332ddf8b9fb3ef9d	0/20	100/100	0/20	100/100
d8c87b9bfde38897979e41242626c2f3	0/20	100/100	0/20	100/100
554719594ba90b02ae410c297c6e50ad	0/20	0/100	0/20	1/100
33490b1d5377580b19f7f9b5849d7991	0/20	2/100	0/20	3/100
9c815150ea821166faecf80757d8826a	0/20	4/100	0/20	2/100
d311fcfe5b660d59dc616e20831c55a0	0/20	0/100	0/20	0/100
66918128f1b9b03303d77c6f2eefd128	0/20	1/100	0/20	100/100
9313b7e42f6ef7b8157a9a9dcf8d8751	1/20	0/100	0/20	0/100
4dedfe5d0b5f86a1517ccd32636433e5	0/20	0/100	19/20	0/100
346ba115157830275e4b9249de1d2bba	0/20	4/100	0/20	3/100
8f35687f7cd9ba7a693ccc31d712f6c0	0/20	0/100	0/20	0/100
3b5074b1b5d032e5620f69f9f700ff0e	0/20	0/100	0/20	0/100
a89ed63ceab8bcda5c99b08f0ef2f151	20/20	0/100	20/20	0/100
d29e16f2a6aaec2d3ac83b2076f81ca3	0/20	0/100	0/20	0/100
c9610a96cb052bf20d5e4e45d81abed8	0/20	0/100	0/20	1/100
28a2c9bd18a11de089ef85a160da29e4	1/20	0/100	0/20	0/100
69659b6dfbeaa53c063d2002cfecab13	1/20	0/100	1/20	0/100
8f41a697eff27e008f969cf7b5ba4117	0/20	0/100	0/20	13/100
c60d01d600aacc2c04844595ce224279	0/20	0/100	0/20	1/100
50d267ccce1b9cc18c13541d20ee4601	1/20	0/100	1/20	0/100

Tabuľka 7.20: Porovnanie výskytu JA3 odtlačkov profilov Můj vlak a Seznam.cz vytvorených pomocou Android Studia a pomocou virtuálneho stroja

Tabuľky 7.20 a 7.21 porovnávajú výskyt JA3 odtlačkov profilov vytvorených použitím Android Studia a virtuálneho stroja. Za týmto účelom je dané rozostavenie v tabuľke, kde sú vedľa seba profily pre jednu aplikáciu ale vytvorené rôznymi spôsobmi.

Farebné rozlíšenie v tabuľkách slúži na zvýraznenie JA3 odtlačkov, ktoré majú vysoký výskyt pre jeden až viac profilov. Žltá farba označuje JA3 odtlačky, ktoré majú vysoký výskyt v profiloch vytvorených pomocou Android Studia, sivá farba označuje JA3 odtlačky s vysokým výskytom, vytvorených pomocou virtuálneho stroja.

JA3 odtlačok	AS CP	VM CP	AS Reddit	VM Reddit
bd1073cbb91ac524017faf1cee3bbd68	0/20	0/100	20/20	0/100
859ce948f69709d8d1263981cf4dd16d	0/20	0/100	0/20	0/100
ebf5e0e525258d7a8dcb54aa1564ecbd	20/20	0/100	0/20	0/100
839868ad711dc55bde0d37a87f14740d	0/20	0/100	0/20	0/100
e9ec38c2b40ff3e300e9975dd7619902	0/20	0/100	0/20	4/100
78c922015590b3a7822b159f62b05fc8	20/20	0/100	18/20	0/100
33b38c8778b29bc991b5200b6141f35e	0/20	3/100	0/20	2/100
6ec2896feff5746955f700c0023f5804	0/20	3/100	0/20	1/100
f912c10788b92e747685beebaaef6cd1	1/20	1/100	0/20	0/100
6db5068dc42d8cfb4bd5ee3e169db25e	0/20	0/100	0/20	0/100
3967ff2d2c9c4d144e7e30f24f4e9761	0/20	100/100	0/20	95/100
ee26b1f1aec16d6098768e2c67388ace	0/20	5/100	0/20	0/100
a0e9f5d64349fb13191bc781f81f42e1	1/20	0/100	1/20	0/100
6f5e62edfa5933b1332ddf8b9fb3ef9d	0/20	0/100	0/20	100/100
d8c87b9bfde38897979e41242626c2f3	0/20	100/100	0/20	100/100
554719594ba90b02ae410c297c6e50ad	0/20	0/100	0/20	0/100
33490b1d5377580b19f7f9b5849d7991	0/20	0/100	0/20	2/100
9c815150ea821166faecf80757d8826a	0/20	2/100	0/20	0/100
d311fcfe5b660d59dc616e20831c55a0	20/20	0/100	0/20	0/100
66918128f1b9b03303d77c6f2eefd128	0/20	100/100	0/20	1/100
9313b7e42f6ef7b8157a9a9dcf8d8751	1/20	0/100	0/20	0/100
4dedfe5d0b5f86a1517ccd32636433e5	1/20	0/100	0/20	0/100
346ba115157830275e4b9249de1d2bba	0/20	3/100	0/20	4/100
8f35687f7cd9ba7a693ccc31d712f6c0	1/20	0/100	0/20	0/100
3b5074b1b5d032e5620f69f9f700ff0e	1/20	0/100	0/20	0/100
a89ed63ceab8bcda5c99b08f0ef2f151	20/20	0/100	20/20	0/100
d29e16f2a6aaec2d3ac83b2076f81ca3	1/20	0/100	1/20	0/100
c9610a96cb052bf20d5e4e45d81abed8	1/20	1/100	0/20	0/100
28a2c9bd18a11de089ef85a160da29e4	2/20	0/100	0/20	0/100
69659b6dfbeaa53c063d2002cfecab13	1/20	0/100	0/20	0/100
8f41a697eff27e008f969cf7b5ba4117	0/20	0/100	0/20	0/100
c60d01d600aacc2c04844595ce224279	0/20	100/100	0/20	0/100
50d267ccce1b9cc18c13541d20ee4601	1/20	0/100	0/20	0/100

Tabuľka 7.21: Porovnanie výskytu JA3 odtlačkov profilov CP a Reddit vytvorených pomocou Android Studia a pomocou virtuálneho stroja

Prvá vec, ktorú je možné všimnúť si v týchto tabuľkách je to, že ak má daný odtlačok výskyt v profile pre danú aplikáciu, pre profil tej istej aplikácie vytvorený druhým spôsobom je výskyt nulový. Môže nastať výnimka, kedy sa JA3 odtlačok vyskytuje v oboch profiloch pre danú aplikáciu, no výskyt je v tomto prípade minimálny a tým je odtlačok nepoužiteľný.

Dané farebné rozlíšenie taktiež umožňuje vidieť dve skupiny hlavných JA3 odtlačkov, kde je jedna skupina špecifická pre profily vytvorené pomocou Android Studia a druhá skupina špecifická pre profily vytvorené použitím virtuálneho stroja, pričom prienik týchto skupín je prázdna množina.

Ak by bol JA3 odtlačok špecifický pre konkrétnu aplikáciu, tento odtlačok by bol prítomný v oboch profiloch pre danú aplikáciu bez ohľadu na to, ako boli vytvorené dátové sady z ktorých bol profil vytvorený. Poznatok, ktorý je popísaný v tomto paragrafe túto tézu teda vyvracia.

	JA3 odtlačok	Můj vlak	Seznam.cz	CP	Reddit
	Android Studio				
1	a89ed63ceab8bcda5c99b08f0ef2f151	20/20	20/20	20/20	20/20
2	bd1073cbb91ac524017faf1cee3bbd68	20/20	20/20	0/20	20/20
3	ebf5e0e525258d7a8dcb54aa1564ecbd	0/20	20/20	20/20	0/20
4	78c922015590b3a7822b159f62b05fc8	1/20	20/20	20/20	18/20
	Virtuálny stroj				
1	d8c87b9bfde38897979e41242626c2f3	100/100	100/100	100/100	100/100
2	6f5e62edfa5933b1332ddf8b9fb3ef9d	100/100	100/100	0/100	100/100
3	66918128f1b9b03303d77c6f2eefd128	1/100	100/100	100/100	1/100
4	3967ff2d2c9c4d144e7e30f24f4e9761	0/100	100/100	100/100	95/100

Tabuľka 7.22: Extrakcia určitých odtlačkov z tabuliek 7.20 a 7.21

Tabuľka 7.22 obsahuje JA3 odtlačky, ktoré sú vďaka svojmu výskytu brané ako relevantné. Pri porovnaní riadkov s rovnakým číslom je vidieť podobnosť pomeru výskytu odtlačku naprieč profilmi. Všetky profily v tejto tabuľke, odhliadnuc od metódy použitej na tvorbu dátových sád, boli tvorené z dátových sád rovnakej Android verzie. Z toho je možné predpokladať, že JA3 odtlačok je ovplyvnený aj spôsobom, akým je zariadenie spustené[1] a nie je viazaný čisto na aplikáciu.

JA3 odtlačok	Můj vlak	Seznam.cz	CP	Reddit
ee26b1f1aec16d6098768e2c67388ace	1/20	1/20	2/20	1/20
6ec2896feff5746955f700c0023f5804	0/20	0/20	0/20	1/20
3967ff2d2c9c4d144e7e30f24f4e9761	0/20	0/20	0/20	1/20
33490b1d5377580b19f7f9b5849d7991	1/20	2/20	0/20	3/20
c9610a96cb052bf20d5e4e45d81abed8	0/20	0/20	0/20	1/20
346ba115157830275e4b9249de1d2bba	0/20	0/20	1/20	3/20
66918128f1b9b03303d77c6f2eefd128	20/20	20/20	20/20	20/20
e9ec38c2b40ff3e300e9975dd7619902	0/20	2/20	0/20	0/20
6db5068dc42d8cfb4bd5ee3e169db25e	0/20	0/20	0/20	1/20
d8c87b9bfde38897979e41242626c2f3	0/20	0/20	0/20	1/20
6f5e62edfa5933b1332ddf8b9fb3ef9d	0/20	0/20	0/20	1/20

Tabuľka 7.23: JA3 odtlačky pri profiloch použití internetového prehliadača

Na porovnanie JA3 odtlačkov profilu aplikácie a profilu jej variácie dostupnej cez internetový prehliadač slúži tabuľka 7.23. Profily boli vytvorené z dátových sád, ktoré obsahujú prístup na danú stránku namiesto aplikácie (napr. namiesto použitia aplikácie *Seznam.cz* bolo pristúpené na stránku `www.seznam.cz` cez internetový prehliadač).
Pri porovnaní s tabuľkou 7.17 je vidieť, že pri aplikáciach je nájdených viac JA3 odtlačkov

[1] Android na inteligentnom zariadení, Android spustený cez emulátor, Android spustený ako virtuálne zariadenie, atď.

a ich výskyt naprieč profilmi je rozdielny. V prípade internetového prehliadača je relevantný jediný JA3 odtlačok, ktorého výskyt je maximálny vo všetkých profiloch, čo značí, že nie je špecifický pre konkrétnu stránku.
Tento odtlačok je prítomný aj v profile aplikácií *CP* a *Seznam.cz*. Z toho vyplýva, že tento odtlačok nie je špecifický len pre prístup cez internetový prehliadač a preto je jeho použiteľnosť znížená.

7.3 Experimenty s vyhľadávaním profilov

V tejto podkapitole sú popísané experimenty, ktoré sa sústredia na identifikovanie profilu aplikácie v určitom zázname či dátovej sade. Myšlienka identifikácie spočíva v tom, s akou istotou je možné povedať, že daná aplikácia bola spustená v čase, keď sa záznam, v ktorom sa vyhľadáva, zhotovil. Cieľom týchto pokusov je zistiť efektívnosť a využiteľnosť typov profilov a ich jednotlivých atribútov. V podkapitole 4.3 sú popísané jednotlivé časti pokusov, ktoré sú manipulované. Konkrétne sa jedná o minimálnu hranicu počtu výskytov a váhy atribútov. V spomínanej kapitole je taktiež popísaný samotný proces vyhľadávania profilu v dátovej sade. Dátové sady, v ktorých sa profily vyhľadávajú, sú popísané v 6.3. V rámci tabuliek tejto podkapitoly, označenie *0 app* značí, že v dátová sada neobsahuje spustenie žiadnej aplikácie, *4 app* značí, že dátová sada obsahuje spustenie všetkých štyroch použitých aplikácií. Dátová sada s konkrétnym menom obsahuje spustenie aplikácie s rovnakým menom.

Rozdelenie pokusov

Primárne rozdelenie pokusov v rámci tejto podkapitoly je podľa toho, aký typ profilu bol použitý. Konkrétne *základný profil*, *špecifický profil*, *super profil* a *špecifický super profil*.

Sekundárne rozdelenie je pri každom type profilu. Tým je rozdelenie podľa nastavenia minimálnej hranice počtu výskytu položiek. Zmena minimálnej hranice počtu výskytov položiek ovplyvňuje počet položiek, ktoré sú v danom pokuse brané ako relevantné. Počet relevantných položiek pre profil v danom nastavení je zobrazený v zátvorke pri každom profile. Cieľom týchto zmien je zistenie, ako je ovplyvnená efektívnosť profilu znížením či zvýšením použitých položiek. Pri pokusoch boli použité tri hranice minimálneho počtu výskytov, konkrétne 50%, 75% a 90%.

Terciárne rozdelenie je podľa toho, aké boli nastavené váhy atribútov pri pokuse. Manipulovanie s váhami atribútov profilu má za cieľ zistiť, či je uprednostňovanie jedného atribútu výhodnejšie z pohľadu identifikácie aplikácie v zázname sieťovej komunikácie. Časť pokusov je zameraná aj na použiteľnosť JA3 odtlačkov práve znížením váhy tohoto atribútu. Informácie o nastavení váh atribútov pri danom pokuse sa nachádzajú v popise tabuliek.

Vo všetkých tabuľkách v rámci tejto podkapitoly, hodnoty reprezentujú istotu, s ktorou je možné na základe získaných informácií povedať, či bola aplikácia aktívna počas tvorby skúmanej dátovej sady.

7.3.1 Vyhľadávanie základných profilov

Základné profily, ako je popísané v 7.1, obsahujú položky ktoré sa vyskytli pri sieťovej komunikácií aplikácie pri jednej konkrétnej konfigurácií[2]. To znamená, že z jedného základného profilu nie je možné určiť, ktoré položky patria výhradne aplikácií a ktoré sú ovplyvnené inými faktormi konfigurácie.

[2]Konfiguráciou je myslená kombinácia spôsobu spustenia Android zariadenia a Android verzie

Fakt, že základné profily nie sú nijako filtrované od spoločných položiek taktiež môže ovplyvniť vyhľadávanie. To, či profil spoločné položky obsahuje alebo nie, nie je zo samotného profilu zistiteľné. Na to poukáže až porovnanie profilov medzi sebou, či vyhľadávanie profilu v dátových sadách, ktoré obsahujú komunikáciu inej aplikácie.

Identifikáciu základných profilov v dátových sadách taktiež ovplyvňuje aj počet relevantných položiek profilu. Pri nízkom počte relevantných položiek je šanca, že niektoré z týchto položiek sú spoločné pre viaceré profily. To môže spôsobiť, že profil bude čiastočne identifikovaný aj v dátovej sade, v ktorej aplikácia ktorú profil reprezentuje nebola spustená.

Na základe poznámok z predošlých paragrafov je v tejto podkapitole zobrazená len časť pokusov s vyhľadávaním základných profilov. Tie slúžia na potvrdenie predpokladov, ktoré boli popísané. V tabuľkách 7.24 až 7.29 sú zobrazené pokusy vyhľadávania základných profilov, pri ktorých boli váhy atribútov profilov nastavené rovnako, s výnimkou JA3 odtlačku. Pokusy s vyhľadávaním základných profilov s rôznymi váhami atribútov je možné vidieť v prílohe C.

Minimálny percentuálny výskyt: 50%

V tabuľkách 7.24 a 7.25 je možné všimnúť si rozdielny počet relevantných položiek pre každý profil. Dopad počtu relevantných položiek na vyhľadávane je možné vidieť pri porovnaní profilov *Můj vlak* a *Seznam.cz*. Istoty o výskyte či absencii profilu sú v prípade vyššieho počtu relevantných položiek presnejšie.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(20)	56.4	96.69	87.07	48.21	38.34	60.82
Můj vlak(7)	37.28	76.48	44.48	76.48	60.48	60.48
Reddit(26)	15.47	69.82	15.47	15.71	60.02	20.28
Seznam.cz(53)	13.66	75.27	21.11	16.13	16.13	74.05

Tabuľka 7.24: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(16)	49.87	95.52	82.53	47.59	34.24	55.84
Můj vlak(5)	31.29	65.41	41.88	65.41	41.88	41.88
Reddit(23)	7.11	64.84	7.11	7.11	53.42	7.11
Seznam.cz(49)	6.71	72.47	15.01	9.46	9.46	71.11

Tabuľka 7.25: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

Minimálny percentuálny výskyt: 75%

Pri porovnaní tabuliek 7.24 a 7.26 je vidieť, že zvýšenie minimálnej hranice počtu výskytov na 75% znížilo počet relevantných položiek pri profiloch *CP, Reddit* a *Seznam.cz* takmer o polovicu, pričom profil *Můj vlak* prišiel o jednu položku.

Pri tomto nastavení pre všetky profily platí, že pri identifikácií profilu v dátovej sade, v ktorej sa tento profil nachádza, sa istota nájdenia profilu zvýšila. Toto je dôsledok toho, že vyššou hranicou minimálneho výskytu každý profil obsahuje menej relevantných položiek.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(11)	46.18	100.0	92.54	45.7	45.7	73.65
Můj vlak(6)	31.83	83.13	48.35	83.13	65.74	65.74
Reddit(16)	18.04	69.36	18.04	18.37	63.59	24.67
Seznam.cz(25)	20.48	92.72	32.8	24.56	24.56	92.94

Tabuľka 7.26: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(7)	28.68	100.0	88.01	43.1	43.1	72.84
Můj vlak(4)	22.13	74.13	47.47	74.13	47.47	47.47
Reddit(13)	6.35	61.91	6.35	6.35	54.74	6.35
Seznam.cz(21)	9.45	91.25	24.27	14.36	14.36	91.51

Tabuľka 7.27: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

Minimálny percentuálny výskyt: 90%

Tabuľky 7.28 a 7.29 ukazujú, že pri nastavení minimálnej hranice výskytu na 90%, že po odstránení spoločných JA3 odtlačkov, profil *Reddit* obsahuje len položky, ktoré sú špecifické pre tento profil. To je vidieť z nulových hodnôt pri identifikácií profilu v iných dátových sadách. Z toho vyplýva, že aj základné profily môžu byť špecifické a použiteľné pri určitých nastaveniach.

Pri profile *CP* môže byť táto hranica považovaná za horšiu, pretože spôsobuje nízky počet relevantných položiek. Je možné predpokladať, že veľká časť týchto položiek je spoločná s profilom *Seznam.cz*, na základe výskytu profilu *CP* v dátovej sade *Seznam.cz*.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(9)	44.89	100.0	100.0	44.33	44.33	77.55
Můj vlak(5)	20.33	80.28	39.63	80.28	59.96	59.96
Reddit(11)	17.97	73.27	17.97	18.43	73.27	27.19
Seznam.cz(20)	20.36	100.0	35.27	25.29	25.29	95.32

Tabuľka 7.28: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(5)	20.37	100.0	100.0	39.71	39.71	79.63
Můj vlak(3)	0.0	66.78	32.53	66.78	32.53	32.53
Reddit(8)	0.0	63.29	0.0	0.0	63.29	0.0
Seznam.cz(16)	6.39	100.0	25.11	12.59	12.59	94.12

Tabuľka 7.29: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

Zhodnotenie

Tabuľky 7.24 až 7.29 slúžia ako podklady na potvrdenie predpokladov o použiteľnosti samostatných základných profilov. Ako je možné vidieť pri porovnaní základných profilov *Můj vlak* a *Seznam.cz*, počet relevantných položiek má veľký vplyv na identifikáciu profilu v dátovej sade, v ktorej by sa profil nachádzať nemal. Pri nízkom počte relevantných položiek je oveľa vyššia šanca na nesprávnu identifikáciu profilu a to z dôvodu, že nie je zaistené odfiltrovanie spoločných položiek. Kvôli tomuto nízkemu počtu, jedna nájdená položka v sade dokáže zvýšiť istotu identifikácie až o 20%. Potvrdenie tohoto predpokladu je možné vidieť pri porovnaní tabuliek 7.24 a 7.25 profilu *Můj vlak*, kde odstránenie JA3 odtlačkov pri vyhľadávaní (čo v tomto prípade tvorí rozdiel 2 položiek) spôsobí rozdiel takmer 19%.

Záleží taktiež od toho, koľko spoločných položiek profil obsahuje. Je možné povedať, že profily *CP* a *Reddit* vo väčšine prípadov obsahujú približne rovnaký počet relevantných položiek. Pomer spoločných položiek je zjavný na základe identifikácie týchto profilov v dátových sadách, ktoré by tieto profily nemali obsahovať. Zatiaľ čo *Reddit* identifikovaný relatívne presne, *CP* je identifikovaný s vysokou istotou aj v dátových sadách, ktoré obsahujú komunikáciu inej aplikácie.

Podkapitola 7.2 popisuje, že žiaden z profilov neobsahuje JA3 odtlačok, ktorý by bol špecifický pre danú aplikáciu a zároveň použiteľný na základe jeho výskytu. Z toho je možné sa na jeho použitie pri základných profiloch pozerať z dvoch pohľadov. Prvý pohľad je istota, ktorá popisuje istotu identifikácie profilu v dátovej sade, v ktorej sa profil nachádza. Druhý pohľad je istota, ktorá popisuje istotu identifikácie profilu v dátovej sade, v ktorej sa daný profil nenachádza. V tabuľkách 7.24 až 7.29 je možné si všimnúť, že pri nepoužití JA3 odtlačkov, vo všetkých prípadoch klesla istota identifikácie profilu v sade, v ktorej sa profil nenachádza[3]. Taktiež ale klesla aj v sade, v ktorej sa profil nachádza. Z toho by bolo možné usúdiť, že použitie JA3 odtlačku záleží na tom, ktorý pohľad identifikácie je dôležitejší.

Na základe získaných výsledkov je možné vyvodiť, že použiteľnosť základných profilov je závislá od počtu relevantných položiek profilu a toho, aká časť týchto položiek je zdieľaná inými profilmi.

[3]porovnávané s tabuľkou, kde je rovnaká minimálna hranica spolu s rovnakými váhami atribútov, okrem JA3 odtlačku

7.3.2 Vyhľadávanie špecifických profilov

V tejto časti sú popísané pokusy so špecifickými profilmi. Ako je popísané v podkapitole 5.2, špecifické profily neobsahujú žiadne položky, ktoré sú zdielané iným profilom. Použitie tohoto typu profilu odstraňuje problém falošnej identifikácie profilu v dátovej sade, v ktorej identifikovaný nemá byť.

Špecifické profily boli vytvorené výhradne zo základných profilov. V prípade, že profil aplikácie obsahuje väčšie množstvo týchto zdielaných položiek, po vytvorení špecifického profilu mu ostane nízke množstvo položiek. Nevýhoda nízkeho počtu relevantných položiek je popísaná v 7.3.1. Rozdiel v počte relevantných položiek špecifických profilov je možné vidieť v tabuľkách 7.30 až 7.35. Už pri tabuľke 7.30 je potvrdený predpoklad spoločných položiek na základe počtu relevantných profilových položiek v profiloch *CP* a *Můjvlak*.

Aplikácie *Reddit* a *Seznam.cz* majú v špecifických profiloch vyššie číslo relevantných položiek, čo svedčí o ich menších istotách identifikácie pri vyhľadávaní základných profilov.

V rámci tejto podkapitoly, výsledné tabuľky sú len dve pre každé nastavenie minimálnej hranice výskytu. To je z toho dôvodu, že pri špecifickom profile, žiaden profil neobsahuje špecifický JA3 odtlačok, ktorého výskyt by bol vyšší ako 33%. Dôsledkom toho pokusy s vyhľadávaním s a bez JA3 odtlačku dávajú identické výsledky.

Minimálny percentuálny výskyt: 50%

V tabuľkách 7.30 a 7.31 je možné vidieť, že zvýšenie váhy atribútu hostname taktiež zvýši istotu, s akou je možné identifikovať profil v danej dátovej sade.
Tým, že v profiloch *CP* a *Můj vlak* je veľmi nízky počet relevantných položiek, istota s ktorou je možné profil identifikovať môže nabrať len pár hodnôt. Dôsledkom čoho zmena jednej položky značí posun o niekoľko desiatok percent. Na druhú stranu, profily *Reddit* a *Seznam.cz* aj pri špecifickom profile obsahujú signifikantný počet relevantných položiek oproti predošlým dvom profilom. Z toho je možné povedať, že aj napriek nižšiemu číslu istoty má táto hodnota väčšiu váhu.
Nenulová hodnota v riadku s profilom *CP*, kde by profil nemal byť identifikovaný svedčí o tom, že profil obsahuje nejakú položku, ktorá bola prítomná pri tvorbe dátovej sady *Můj vlak*. Oproti predošlej podkapitole, nejedná sa o spoločné položky. Ak by sa o ne jednalo, spôsobilo by to nenulové hodnoty aj v profile *Můj vlak* vyhľadávaný v dátovej sade *CP*.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(3)	25.74	100.0	100.0	25.74	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(17)	0	75.09	0	0	60.23	0
specific Seznam.cz(43)	0	67.61	0	0	0	66.01

Tabuľka 7.30: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(3)	25.74	100.0	100.0	25.74	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(17)	0	85.77	0	0	68.8	0
specific Seznam.cz(43)	0	71.69	0	0	0	74.72

Tabuľka 7.31: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 2, IP adresa: 1

Minimálny percentuálny výskyt: 75%

Pri nastavení minimálnej hranice výskytu na 75% boli odstránené položky v profile *CP*, ktoré spôsobovali nenulové hodnoty v inej dátovej sade. Posunutie hranice na danú hodnotu spôsobilo signifikantný pokles počtu relevantných položiek pri profile *Seznam.cz*.
Toto nastavenie môže byť nevýhoda pre profily *CP* a *Můj vlak*, pretože počet ich relevantných položiek je 1, no pre ostatné profily to je možné brať ako výhodu, keďže oproti základným profilom, všetky tieto relevantné položky sú špecifické pre daný profil. Z toho je možné implikovať, že v prípade identifikácie profilu v dátovej sade, je možné to tvrdiť s väčšou istotou ako pri základných profiloch, keďže falošné identifikácie by mali byť odstránené vďaka tomu, že ide o špecifický profil.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(1)	0	100.0	100.0	0	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(10)	0	70.8	0	0	61.66	0
specific Seznam.cz(16)	0	88.45	0	0	0	88.79

Tabuľka 7.32: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(1)	0	100.0	100.0	0	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(10)	0	82.9	0	0	72.2	0
specific Seznam.cz(16)	0	87.39	0	0	0	90.52

Tabuľka 7.33: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 2, IP adresa: 1

Minimálny percentuálny výskyt: 90%

V tabuľkách 7.34 a 7.35 je možné vidieť extrémne istoty pri identifikácií profilov *CP* a *Můj vlak*. To je spôsobené tým, že pri daných nastaveniach je pre tieto profily len jedna položka relevantná a tým môžu nastať len dve situácie. Podľa toho, či sa táto položka našla, je aplikácia v dátovej sade identifikovaná.
Pri profiloch *Reddit* a *Seznam.cz* s danými nastaveniami je možné priložiť výsledným istotám vyššiu váhu kvôli väčšiemu počtu relevantných položiek, ktoré boli vyhľadávané. Zmena váhy atribútu hostname ovplyvnila pri danej konfigurácií iba tieto dva profily. Pre profil *Reddit* zmena váhy spôsobila zvýšenie istoty, z čoho je možné predpokladať vyššiu dôležitosť hostname položiek. Pri profile *Seznam.cz* zvýšenie váhy spôsobilo pravý opak, z čoho je možné predpokladať vyššiu dôležitosť IP adries pri tomto profile.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(1)	0	100.0	100.0	0	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(7)	0	72.46	0	0	72.46	0
specific Seznam.cz(12)	0	100.0	0	0	0	92.15

Tabuľka 7.34: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific CP(1)	0	100.0	100.0	0	0	0
specific Můj vlak(1)	0	100.0	0	100.0	0	0
specific Reddit(7)	0	84.03	0	0	84.03	0
specific Seznam.cz(12)	0	100.0	0	0	0	91.8

Tabuľka 7.35: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 2, IP adresa: 1

Zhodnotenie

V tabuľkách 7.30 až 7.35 je možné vidieť, že použitie špecifických profilov z väčšiny odstránilo falošnú identifikácie profilov, vďaka odstráneniu spoločných prvkov[4].

Výhoda a zároveň presnosť využitia tohoto typu profilu stúpa spolu s počtom relevantných položiek profilu. V prípade nízkeho počtu relevantných položiek môže byť vypočítaná istota nepresná, kvôli absencií určitých položiek. S vyšším počtom položiek toto riziko klesá.

Rôzne hranice minimálneho výskytu pri použití tohoto typu profilu korigujú primárne to, koľko položiek je potrebných na získanie istoty identifikácie s hodnotou 100%. Branie do úvahy položiek s nižším výskytom prináša riziko, že sa táto položka v danej skúmanej sade nevyskytuje, čo spôsobí nižšiu istotu identifikácie.

Pri použití špecifických profilov, zmena váhy atribútov vo väčšine pokusov zvýši istotu identifikácie. Vďaka jednému výsledku kde táto zmena mierne znížila výslednú hodnotu, je možné usúdiť, že to nie je pravidlo. To naznačuje dôležitosť hostname atribútov.

V porovnaní so základnými profilmi, špecifické profily prinášajú presnejšie výsledky pri identifikácií profilov v rôznych dátových sadách. Vyššia presnosť neznamená len vyššiu istotu identifikácie v dátovej sade, v ktorej sa profil nachádza. Tým, že žiaden profil neobsahuje spoločné položky, nedochádza k „falošným“ identifikáciám. Vďaka tomu je istota identifikácie ešte vierohodnejšia.

7.3.3 Vyhľadávanie super profilov

Super profily a základné profily sú veľmi podobné. Hlavný rozdiel spočíva v tom, že základné profily popisujú sieťovú komunikáciu aplikácie pri jednej konfigurácií a super profily by mali zastrešovať komunikácie aplikácie z viacero konfigurácií. Výsledky pri vyhľadávaní super profilov sú veľmi podobné tým, aké priniesli pokusy so základnými profilmi.

V tejto podkapitole sú uvedené len dve tabuľky, na ktorých sú poukázané určité zmeny oproti ostatným pokusom. Výsledky ostatných pokusov s vyhľadávaním super profilov je možné nájsť v prílohe C.

Rozdiel, na ktorý je vhodné upozorniť je pri minimálnej percentuálnej hranici výskytu. Treba dať do pozornosti, že minimálna percentuálna hranica pri super profiloch predsta-

[4]Spoločné prvky vytvorených profilov, ktoré sú k dispozícií. V prípade tejto práce ide o profily štyroch aplikácií

vuje inú hodnotu, ako pri základných profiloch. Pri hranici 50% boli počty relevantných položiek približne rovnaké ako pri základných profiloch. Pri hranici 75% sa rapídne zmenil počet relevantných položiek super profilu *Reddit*, ako je možné vidieť v tabuľke 7.36. Oproti základnému profilu ktorý mal 16 relevantných položiek, s rovnakou minimálnou percentuálnou hranicou výskytu má super profil *Reddit* len 7 relevantných položiek. Znížený počet relevantných položiek taktiež ovplyvňuje výsledky identifikácie.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(9)	40.69	100.0	100.0	43.46	53.52	76.51
sup Můj vlak(3)	29.32	100.0	61.17	100.0	61.17	61.17
sup Reddit(7)	12.8	100.0	12.8	12.8	87.74	12.8
sup Seznam.cz(15)	12.24	100.0	32.17	18.97	18.97	92.93

Tabuľka 7.36: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

Pri hranici 90%, počet relevantných položiek je pri super profiloch značne nižší. Ako je vidieť v tabuľke 7.37, počet relevantných položiek pri žiadnom profile neprekračuje hranicu 5. Ak je pri takomto nastavení v super profile položka braná ako relevantná, znamená to, že sa vyskytla takmer v každej dátovej sade, nezávisle od toho, ako bolo zariadenie spustené alebo aká verzia Android bola použitá. V tabuľke 7.37 je ale vidieť, že profily stále môžu obsahovať položky ktoré nie sú špecifické. Z toho je možné usúdiť, že nejaká položka môže byť pripisovaná aplikácií právom, ale to nemusí znamenať, že táto aplikácia je jediná ktorá ju používa

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(3)	0	100.0	100.0	32.76	32.76	65.74
sup Můj vlak(1)	0	100.0	0	100.0	0	0
sup Reddit(5)	0	100.0	0	0	100.0	0
sup Seznam.cz(5)	0	100.0	19.56	0	0	80.54

Tabuľka 7.37: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

Zmena váh atribútov profilu je závislá od toho, ktorých položiek obsahuje super profil viac. V prípade vyššieho počtu IP adries, zvýšením váhy IP adries sa zvýši aj istota identifikácie. Pre hostname to platí rovnako.

Na základe získaných výsledkov je možné povedať, že použiteľnosť super profilov pri vyhľadávaní je približne rovnaká, ako pri základných profiloch. Super profily sú výhodné tým, že poskytujú relatívne rozsiahly prehľad o tom, aké prvky komunikácie aplikácia používala s akou frekvenciou.

7.3.4 Vyhľadávanie špecifických super profilov

Tie položky, ktoré sú obsiahnuté v týchto profiloch predstavujú tie IP adresy a hostname, ktoré sú špecifické pre dané aplikácie naprieč rôznymi spôsobmi spustenia aplikácie. Dôsledkom odstránenia spoločných položiek super profilov, väčšina JA3 odtlačkov je v tomto procese odstránená. Dôsledkom toho táto podkapitola obsahuje najviac dve tabuľky pre každú minimálnu hranicu percentuálneho výskytu. Váha JA3 odtlačku teda nemá vplyv na dosiahnuté výsledky.

Minimálny percentuálny výskyt: 50%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific sup CP(2)	0	100.0	100.0	0	0	0
specific sup Můj vlak(1)	0	100.0	0	100.0	0	0
specific sup Reddit(15)	0	83.06	0	0	65.41	0
specific sup Seznam.cz(43)	0	66.98	0	0	0	73.4

Tabuľka 7.38: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific sup CP(2)	0	100.0	100.0	0	0	0
specific sup Můj vlak(1)	0	100.0	0	100.0	0	0
specific sup Reddit(15)	0	90.75	0	0	71.47	0
specific sup Seznam.cz(43)	0	70.97	0	0	0	80.11

Tabuľka 7.39: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 2, IP adresa: 1

V tabuľkách 7.38 a 7.39 je možné vidieť podobnosť výsledkov ako pri použití špecifických profilov, s tým rozdielom, že pri použití špecifických super profilov sú hodnoty istôt o niečo vyššie. Počet relevantných položiek má na tieto hodnoty vplyv, no špecifický super profil *Seznam.cz* má rovnaký počet relevantných položiek ako špecifický profil *Seznam.cz*. To naznačuje vyššiu efektívnosť špecifického super profilu aj v prípade, kedy výsledok nie je ovplyvnený počtom relevantných položiek.

Zvýšenie váhy hostname atribútu spôsobí taktiež zvýšenie istoty identifikácie profilu. To svedčí o tom, že pri tejto konfigurácií, hostname atribút má väčšiu hodnotu v rámci profilu ako atribút IP adries. Taktiež z toho vyplýva, že do relevantných položiek patria hostname aj IP adresy, no počet hostname položiek presahuje IP adresy.

Špecifické super profily *Reddit* a *Seznam.cz* pri danom počte relevantných položiek profilu sú v dátovej sade, v ktorej sa tieto aplikácie nachádzali, identifikované s vysokou istotou. Tým, že tieto profily neboli vôbec identifikované v dátových sadách, v ktorých sa aplikácie nenachádzali, sa zvyšuje vierohodnosť správnej identifikácie. Tým sa zvyšuje aj použiteľnosť týchto profilov.

Minimálny percentuálny výskyt: 75%

Tabuľka 7.40 reprezentuje výsledky pokusov pri danej minimálnej hranici výskytu a vypovedá o tom, že výsledok nezávisí od použitia JA3 odtlačku, ani od toho, či majú atribúty rôzne váhy.

Absencia JA3 odtlačku bola viditeľná už pri hranici 50%, čo pretrvá pri zvýšení tejto hranice. Žiaden dôsledok zmeny váhy atribútov svedčí o tom, že každý profil obsahuje len položky jedného typu atribútu.

Číslo relevantných položiek pri profiloch je relatívne nízke pri porovnaní s predošlými konfiguráciami, no treba mať na pamäti, že sú to položky ktoré boli asociované s aplikáciu pri viac ako 75% dátových sád, ktoré boli vytvorené rôznymi spôsobmi. Tento fakt, spolu s vysokým rozdielom pri identifikáciách naprieč dátovými sadami, predstavuje výpovednú hodnotu profilu.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific sup CP(1)	0	100.0	100.0	0	0	0
specific sup Můj vlak(1)	0	100.0	0	100.0	0	0
specific sup Reddit(6)	0	100.0	0	0	85.94	0
specific sup Seznam.cz(10)	0	100.0	0	0	0	89.57

Tabuľka 7.40: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1/2, IP adresa: 1

Minimálny percentuálny výskyt: 90%

Pri zvýšení minimálnej hranice výskytu na 90% sa identifikácia profilov správa rovnako, ako pri hranici 75% s tým rozdielom, že profily obsahujú menej relevantných položiek, čo spôsobí zvýšenie istoty identifikácie.

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
specific sup CP(1)	0	100.0	100.0	0	0	0
specific sup Můj vlak(1)	0	100.0	0	100.0	0	0
specific sup Reddit(5)	0	100.0	0	0	100.0	0
specific sup Seznam.cz(4)	0	100.0	0	0	0	75.81

Tabuľka 7.41: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1/2, IP adresa: 1

Zhodnotenie

Pri použití špecifických super profilov dochádza k presnejšej identifikácií profilov v dátových sadách. To je dôsledkom toho, že profily obsahujú len jedinečné položky, ktoré boli získané z dátových sád vytvorených rôznymi spôsobmi. Na základe toho je možné tieto položky považovať za smerodajné pri identifikácií profilov.

Na základe výsledkov z tabuliek 7.38 až 7.41 je možné vidieť, že tieto profily neobsahujú žiadne JA3 odtlačky, ktoré by boli špecifické pre aplikáciu odhliadnuc od spôsobu spustenia.

Pri práci so špecifickými super profilmi je taktiež vidieť, že prevažná väčšina položiek profilov je tvorená hostname atribútami. Z toho je možné usúdiť, že tento atribút má pri tejto konfigurácií väčší vplyv na identifikáciu profilu, čím sa zvyšuje aj jeho váha oproti IP adresám profilu.

Z porovnania pokusov v tejto podkategórií s pokusmi s inými typmi profilov je možné vidieť, že pri vhodnom nastavení minimálnej hranice výskytu je možné aplikáciu identifikovať najpresnejšie použitím špecifického super profilu. Pre profily *CP* a *Můj vlak* však využitie tohto typu profilu nemusí byť najvhodnejšie, keďže je aplikácia identifikovaná na základe jednej položky, aj keď sa táto položka vyskytuje v prevažnej väčšine dátových sád, z ktorých boli profily pre túto aplikáciu vytvorené.

Z predošlých pokusov s vyhľadávaním, je použitie špecifických super profilov efektívnejšie pri identifikácií profilu ako použitie super profilov. Špecifický super profil má naopak nižšie použitie pri vyobrazení toho, aké položky s akou dôležitosťou aplikácia používa, keďže sú odstránené všetky položky, ktoré nie sú špecifické pre aplikáciu profilu.

To je potvrdené aj identifikáciou aplikácií podľa super profilov a špecifických super profilov v dátových sadách, kde nie je známe ktoré aplikácie boli spustené. Výsledky týchto experimentov sú zobrazené v tabuľkách C.21 až C.28, ktoré sa nachádzajú v prílohách.

Kapitola 8

Záver

Cieľom tejto práce bola tvorba profilov vybraných aplikácií na základe analýzy sieťovej komunikácie týchto aplikácií, spolu s vytvorením dátových sád, ktoré takúto sieťovú komunikáciu obsahujú. V rámci práce sa pracovalo so štyrmi aplikáciami, pričom pre tri aplikácie bolo vytvorených 200 dátových sád a pre jednu aplikáciu 160 dátových sád obsahujúcich záznam sieťovej komunikácie danej aplikácie.

V rámci bakalárskej práce bol implementovaný nástroj, ktorý poskytuje extrakciu štyroch rôznych druhov profilov zo záznamov sieťovej komunikácie. Pri extrakcií dát zo záznamov sa zameriava na cieľové IP adresy, doménové mená a vybrané TLS pakety, z ktorých sú tvorené JA3 odtlačky. Tieto atribúty boli zvolené na základe ich využitia v sieťovej komunikácii. Je pravda, že tieto informácie nemusia byť vždy spoľahlivé, keďže ako IP adresa tak aj doménové meno nemusia byť úplne stabilné. Napriek tomu sa v rámci vykonaných experimentov darilo aplikácie na základe týchto informácií identifikovať. Okrem iného sú súčasťou tohoto nástroja aj funkcie, ktoré slúžia na rôzne modifikácie a porovnávanie profilov či vyhľadávanie profilov v iných záznamoch sieťovej komunikácie.

Použiteľnosť každého typu profilu bola overená identifikovaním aplikácie v rôznych záznamoch sieťovej prevádzky na základe každého profilu, ktorý ju popisoval. Experimentovanie s identifikáciou na základe profilu bolo pre každý typ profilu robené s rôznymi úpravami atribútov. Cieľom viacerých experimentov pre jeden typ profilu bolo získať čo najširší prehľad o tom, ako jednotlivé atribúty profilu ovplyvňujú efektivitu identifikácie. Na základe týchto experimentov bolo zistené, že časť profilov je vhodnejšia na popis komunikácie aplikácie bez ohľadu na to, či sú časti komunikácie jedinečné pre aplikáciu alebo nie. Druhá časť profilov má lepšie využitie pri identifikácii aplikácií v sieťovej prevádzke na základe prvkov jedinečných pre aplikáciu. Rôzne nastavenia atribútov ovplyvňujú identifikáciu aplikácie na základe toho, ako je daný profil stavaný. Najlepšie výsledky pri identifikácii boli dosiahnuté so špecifickými super profilmi, pri ktorých sa istota identifikácie aplikácie v dátovej sade pohybovala v rozmedzí 65% až 100%, pričom použitím týchto profilov žiadna aplikácia nebola identifikovaná v dátovej sade, v ktorej sa nenachádzala.

V prípade pokračovania v práci by bolo možné pri dolovaní z dát aplikovať metódy založené na strojovom učení či zamerať sa na hľadanie iných korelácií v získaných dátach. Prípadne sa zamerať na iné informácie alebo iné kombinácie informácií, ktoré by mali predpoklad na lepší popis správania aplikácie na úrovni siete.

Literatúra

[1] Althouse, J. *TLS Fingerprinting with JA3 and JA3S* [online], 15. januára 2019. Dostupné z: https://engineering.salesforce.com/tls-fingerprinting-with-ja3-and-ja3s-247362855967.

[2] Cloudflare. *What Happens in a TLS Handshake? | SSL Handshake* [online]. [cit. 2020-05-18]. Dostupné z: https://www.cloudflare.com/learning/ssl/what-happens-in-a-tls-handshake/.

[3] Gillmor, D. *Negotiated Finite Field Diffie-Hellman Ephemeral Parameters for Transport Layer Security (TLS)*. RFC 7919. August 2016.

[4] Han, J., Kamber, M. a Pei, J. *Data Mining: Concepts and Techniques*. 3. vyd. Morgan Kaufmann Publishers, 2011 [cit. 2020-04-14]. ISBN 978-0-12-381479-1. Dostupné z: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

[5] Hewitt, S. M. Data, Information, and Knowledge. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*. SAGE Publications. Apríl 2019, zv. 67, č. 4, s. 227–228, [cit. 2020-04-14]. DOI: 10.1369/0022155419836995. ISSN 0022-1554. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6437341/.

[6] Hildebrandt, M. Profiling: From data to knowledge. *Datenschutz und Datensicherheit - DuD*. September 2006, zv. 30, č. 9, s. 548–552, [cit. 2020-04-21]. DOI: 10.1007/s11623-006-0140-3. ISSN 1862-2607. Dostupné z: https://doi.org/10.1007/s11623-006-0140-3.

[7] IANA. *Transport Layer Security (TLS) Extensions* [online]. November 2005. 2020-04-17 [cit. 2020-04-28]. Dostupné z: https://www.iana.org/assignments/tls-extensiontype-values/tls-extensiontype-values.xhtml.

[8] IANA. *Service Name and Transport Protocol Port Number Registry* [online]. Apríl 2020 [cit. 2020-04-21]. Dostupné z: https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml?&page=92.

[9] Kamaruzzaman, M. *Top 10 In-Demand programming languages to learn in 2020* [online], 4. februára 2020 [cit. 2020-05-03]. Dostupné z: https://towardsdatascience.com/top-10-in-demand-programming-languages-to-learn-in-2020-4462eb7d8d3e.

[10] Keary, T. *PCAP: Packet Capture, what it is & what you need to know* [online], 9. decembra 2019 [cit. 2020-05-22]. Dostupné z: https://www.comparitech.com/net-admin/pcap-guide/.

[11] Matoušek, P. *Síťové aplikace a jejich architektura*. 1. vyd. Brno: VUTIUM, 2014. 396 s. ISBN 978-80-214-3766-1.

[12] Navone, E. C. *The @property Decorator in Python: Its Use Cases, Advantages, and Syntax* [online], 19. decembra 2019 [cit. 2020-04-17]. Dostupné z: https://www.freecodecamp.org/news/python-property-decorator/.

[13] Rescorla, E. *The Transport Layer Security (TLS) Protocol Version 1.3*. RFC 8446. August 2018.

[14] Sam, S. *Python object serialization* [online], 19. februára 2019 [cit. 2020-04-17]. Dostupné z: https://www.tutorialspoint.com/python-object-serialization.

[15] Socolofsky, T. a Kale, C. *A TCP/IP Tutorial*. RFC 1180. Január 1991.

[16] Summerson, C. *How to Install Android in VirtualBox* [online], 3. júla 2017 [cit. 2020-04-16]. Dostupné z: https://www.howtogeek.com/164570/how-to-install-android-in-virtualbox/.

[17] Techopedia. *Network Protocols* [online]. Apríl 2020 [cit. 2020-05-04]. Dostupné z: https://www.techopedia.com/definition/12938/network-protocols.

[18] Widman, J. *What is Reddit?* [online], 11. marca 2020 [cit. 2020-05-02]. Dostupné z: https://www.digitaltrends.com/web/what-is-reddit/.

[19] Yin, K. S. *Network Behavioral Analysis for Detection of Remote Access Trojans*. September 2019. Dostupné z: http://onlineresource.ucsy.edu.mm/bitstream/handle/123456789/2283/NetworkBehavioralAnalysis_KhinSweYin.pdf?sequence=1&isAllowed=y.

Príloha A

Obsah priloženého CD

- `/datasets` – vytvorené dátové sady
- `/exports` – všetky výpisy popisujúce profily, vyhľadávania v dátových sadách a pod.
- `/instances` – vytvorené a serializované profily
- `/scripts` – skripty spustiteľné z príkazovej riadky
- `/src` – zdrojové kódy implementácie
- `/unit_tests` – testy na overenie funkčnosti vybraných funkcií
- `/latex/` – zdrojové kódy pre vytvorenie technickej správy
- `manual.txt` – Detailnejší manuál popisujúci obsah priloženého média a spustenie programu
- `xmelus01-network-com-analysis-mob-app-profiling.pdf` – technická správa

Príloha B

Manuál

Podrobnejší manuál je dostupný v súbore `manual.txt`. Na spustenie programu je potrebný Python 3.8, TShark a prítomnosť konkrétnych Python balíčkov v systéme. Inštalácia balíčkov je možná príkazom

```
$pip install <názov balíčku>==<verzia>
```

prípadne

```
$pip3 install <názov balíčku>==<verzia>
```

Potrebné balíčky:

- `prettytable==0.7.2`
- `dpkt==1.9.2`
- `pyshark==0.4.2.9`
- `pyja3==1.0.0`

Implementovaný nástroj plní viaceré funkcie, pričom hlavné funkcie je možné spustiť z príkazovej riadky. Tieto skripty sa nachádzajú v zložke `scripts` a je potrebné ich z tejto zložky aj spúšťať. Každý skript obsahuje nápovedu ako skript správne spustiť. Zobrazenie nápovedy je možné príkazom

```
$python <názov skriptu> -h
```

prípadne

```
$python3 <názov skriptu> -h
```

Súbor `/src/main.py` obsahuje príklady, ako je možné pracovať s niektorými funkciami implementovaného nástroja. Je možné súbor upraviť a spustiť priamo zo zložky `/src` pomocou príkazu

```
$python main.py
```

prípadne

```
$python3 main.py
```

Príloha C

Výsledky experimentov

Výskyt použitých portov v dátových sadách

Počet dátových sád	160		
Vstupný port	Výskyt	Výstupný port	Výskyt
5353	105	53	160
		443	160
		5353	105

Tabuľka C.1: Výskyt portov pri aplikácií *CP*

Počet dátových sád	200		
Vstupný port	Výskyt	Výstupný port	Výskyt
		443	200
		53	176

Tabuľka C.2: Výskyt portov pri aplikácií *Můj vlak*

Počet dátových sád	200		
Vstupný port	Výskyt	Výstupný port	Výskyt
5353	108	53	200
		443	200
		5353	108

Tabuľka C.3: Výskyt portov pri aplikácií *Reddit*

Počet dátových sád	200		
Vstupný port	Výskyt	Výstupný port	Výskyt
5353	106	80	200
		53	200
		443	200
		5353	106

Tabuľka C.4: Výskyt portov pri aplikácií *Seznam.cz*

Vyhľadávanie profilov v dátových sadách

Základný profil, minimálna hranica výskytu 50%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(20)	51.57	97.13	88.82	49.83	38.48	63.31
Můj vlak(7)	32.36	79.58	51.81	79.58	65.69	65.69
Reddit(26)	10.18	80.14	11.76	11.92	67.24	14.93
Seznam.cz(53)	9.73	77.45	18.53	13.24	13.24	79.86

Tabuľka C.5: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 2, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(16)	44.78	96.3	85.58	49.78	35.14	59.93
Můj vlak(5)	25.58	71.73	52.5	71.73	52.5	52.5
Reddit(23)	4.43	78.1	6.17	6.17	63.86	6.17
Seznam.cz(49)	4.63	75.69	14.12	8.42	8.42	78.29

Tabuľka C.6: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 2, IP adresa: 1

Základný profil, minimálna hranica výskytu 75%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(11)	39.12	100.0	93.68	46.32	46.32	77.68
Můj vlak(6)	27.31	85.52	55.67	85.52	70.6	70.6
Reddit(16)	12.47	78.83	12.47	12.69	70.85	17.05
Seznam.cz(25)	12.76	90.93	25.5	17.85	17.85	93.19

Tabuľka C.7: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 2, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(7)	22.24	100.0	90.71	44.59	44.59	78.94
Můj vlak(4)	17.66	79.36	58.09	79.36	58.09	58.09
Reddit(13)	4.08	75.52	4.08	4.08	66.3	4.08
Seznam.cz(21)	5.47	89.87	19.7	11.15	11.15	92.39

Tabuľka C.8: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 2, IP adresa: 1

Základný profil, minimálna hranica výskytu 90%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(9)	36.97	100.0	100.0	45.29	45.29	81.52
Můjvlak(5)	17.04	83.48	49.4	83.48	66.44	66.44
Reddit(11)	12.3	81.7	12.3	12.62	81.7	18.61
Seznam.cz(20)	12.38	100.0	27.43	18.39	18.39	94.3

Tabuľka C.9: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 2, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
CP(5)	14.66	100.0	100.0	42.52	42.52	85.34
Můjvlak(3)	0.0	74.94	49.1	74.94	49.1	49.1
Reddit(8)	0.0	77.52	0.0	0.0	77.52	0.0
Seznam.cz(16)	3.53	100.0	20.71	10.39	10.39	93.5

Tabuľka C.10: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 2, IP adresa: 1

Super profil, minimálna hranica výskytu 50%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(16)	51.73	94.82	84.02	47.62	39.2	63.71
sup Můj vlak(6)	30.52	85.02	48.99	85.02	62.5	62.5
sup Reddit(21)	13.11	78.92	13.11	13.41	65.78	17.06
sup Seznam.cz(53)	9.47	73.7	16.96	12.0	12.27	77.15

Tabuľka C.11: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(12)	44.4	93.1	78.74	47.56	34.91	60.34
sup Můj vlak(4)	23.3	79.48	48.61	79.48	48.61	48.61
sup Reddit(18)	6.22	76.16	6.22	6.22	61.3	6.22
sup Seznam.cz(49)	4.93	71.84	12.94	7.63	7.63	75.54

Tabuľka C.12: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(16)	44.37	95.55	86.29	47.94	40.71	68.87
sup Můj vlak(6)	25.76	87.36	56.94	87.36	68.35	68.35
sup Reddit(21)	8.1	86.98	8.1	8.28	70.74	10.54
sup Seznam.cz(53)	6.51	76.07	15.24	9.99	10.18	82.46

Tabuľka C.13: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 2, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(12)	36.37	94.35	82.58	47.97	37.61	67.51
sup Můj vlak(4)	18.6	83.62	58.99	83.62	58.99	58.99
sup Reddit(18)	3.66	85.97	3.66	3.66	68.49	3.66
sup Seznam.cz(49)	3.31	74.93	12.45	6.96	6.96	81.63

Tabuľka C.14: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 2, IP adresa: 1

Super profil, minimálna hranica výskytu 75%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(9)	40.69	100.0	100.0	43.46	53.52	76.51
sup Můj vlak(3)	29.32	100.0	61.17	100.0	61.17	61.17
sup Reddit(7)	12.8	100.0	12.8	12.8	87.74	12.8
sup Seznam.cz(15)	12.24	100.0	32.17	18.97	18.97	92.93

Tabuľka C.15: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(6)	29.45	100.0	100.0	47.84	47.84	80.77
sup Můj vlak(3)	29.32	100.0	61.17	100.0	61.17	61.17
sup Reddit(7)	12.8	100.0	12.8	12.8	87.74	12.8
sup Seznam.cz(15)	12.24	100.0	32.17	18.97	18.97	92.93

Tabuľka C.16: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(9)	32.35	100.0	100.0	44.76	52.77	81.32
sup Můj vlak(3)	22.24	100.0	70.54	100.0	70.54	70.54
sup Reddit(7)	6.84	100.0	6.84	6.84	86.9	6.84
sup Seznam.cz(15)	6.74	100.0	25.32	14.15	14.15	92.21

Tabuľka C.17: Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 2, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(6)	21.51	100.0	100.0	48.38	48.38	85.95
sup Můj vlak(3)	22.24	100.0	70.54	100.0	70.54	70.54
sup Reddit(7)	6.84	100.0	6.84	6.84	86.9	6.84
sup Seznam.cz(15)	6.74	100.0	25.32	14.15	14.15	92.21

Tabuľka C.18: Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 2, IP adresa: 1

Super profil, minimálna hranica výskytu 90%

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(3)	0	100.0	100.0	32.76	32.76	65.74
sup Můj vlak(1)	0	100.0	0	100.0	0	0
sup Reddit(5)	0	100.0	0	0	100.0	0
sup Seznam.cz(5)	0	100.0	19.56	0	0	80.54

Tabuľka C.19: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	0 app	4 app	CP	Můj vlak	Reddit	Seznam.cz
sup CP(3)	0	100.0	100.0	39.53	39.53	79.33
sup Můj vlak(1)	0	100.0	0	100.0	0	0
sup Reddit(5)	0	100.0	0	0	100.0	0
sup Seznam.cz(5)	0	100.0	19.56	0	0	80.54

Tabuľka C.20: Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 2, IP adresa: 1

Neznáme dátové sady, minimálna hranica výskytu 50%

	Unknown 0	Unknown 1
sup CP (16)	73.6	69.28
sup Můj vlak (6)	45.5	59.01
sup Reddit (21)	17.72	85.47
sup Seznam.cz (53)	9.86	73.28

Tabuľka C.21: Vyhľadávanie super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	Unknown 0	Unknown 1
sup CP (12)	64.87	59.12
sup Můj vlak (4)	43.83	43.83
sup Reddit (18)	11.43	83.58
sup Seznam.cz (49)	5.34	71.4

Tabuľka C.22: Vyhľadávanie super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

	Unknown 0	Unknown 1
specific sup CP (2)	100.0	66.67
specific sup Můj vlak (1)	0	0
specific sup Reddit (15)	0	80.5
specific sup Seznam.cz (43)	0	75.38

Tabuľka C.23: Vyhľadávanie špecifických super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

Neznáme dátové sady, minimálna hranica výskytu 75%

	Unknown 0	Unknown 1
sup CP (9)	67.03	67.03
sup Můj vlak (3)	29.32	29.32
sup Reddit (7)	12.8	87.74
sup Seznam.cz (15)	13.27	74.64

Tabuľka C.24: Vyhľadávanie super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1, Hostname: 1, IP adresa: 1

	Unknown 0	Unknown 1
sup CP (6)	52.76	52.76
sup Můj vlak (3)	29.32	29.32
sup Reddit (7)	12.8	87.74
sup Seznam.cz (15)	13.27	74.64

Tabuľka C.25: Vyhľadávanie super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 0, Hostname: 1, IP adresa: 1

	Unknown 0	Unknown 1
specific sup CP (1)	100.0	100.0
specific sup Můj vlak (1)	0	0
specific sup Reddit (6)	0	85.94
specific sup Seznam.cz (10)	0	90.47

Tabuľka C.26: Vyhľadávanie špecifických super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

Neznáme dátové sady, minimálna hranica výskytu 90%

	Unknown 0	Unknown 1
sup CP (3)	67.24	67.24
sup Můj vlak (1)	0	0
sup Reddit (5)	0	100.0
sup Seznam.cz (5)	19.56	100.0

Tabuľka C.27: Vyhľadávanie super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1

	Unknown0	Unknown1
specific sup CP (1)	100.0	100.0
specific sup Můj vlak (1)	0	0
specific sup Reddit (5)	0	100.0
specific sup Seznam.cz (4)	0	100.0

Tabuľka C.28: Vyhľadávanie špecifických super profilov v neznámych dátových sadách. Nastavené váhy atribútov - JA3 odtlačok: 1/0, Hostname: 1, IP adresa: 1