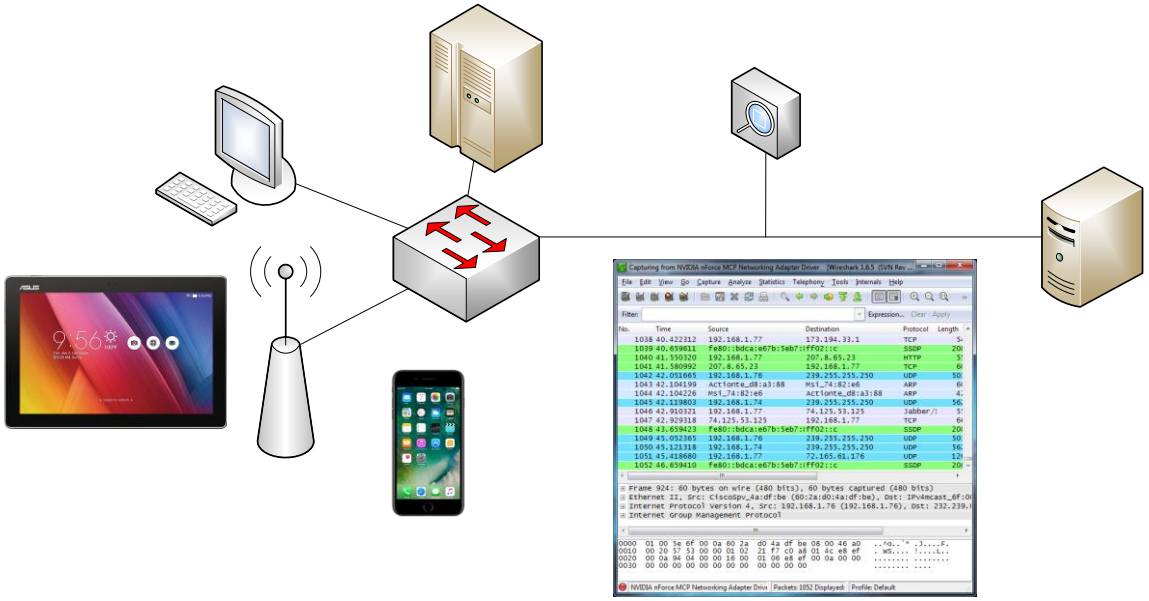
On Reliability of JA3 Hashes for Fingerprinting Mobile Applications

Petr Matoušek, Ivana Burgetová, Ondřej Ryšavý, Malombe Victor

Brno University of Technology, Faculty of Information Technology
Božetěchova 1/2, 612 66 Brno
Czech Republic

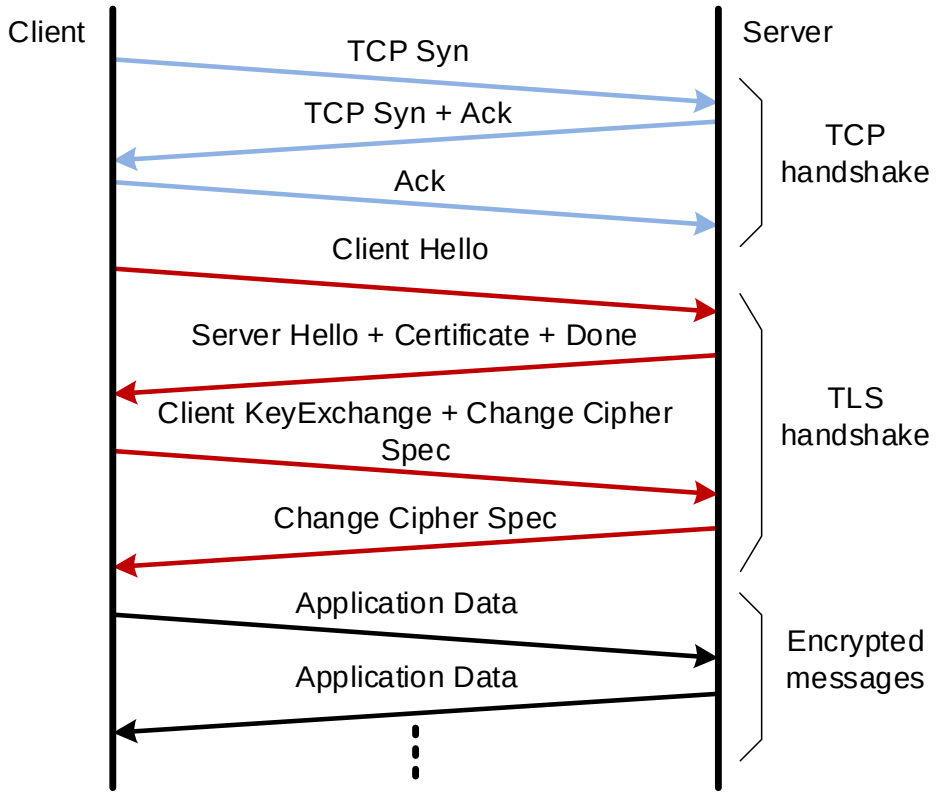
Detection of Mobile Devices in Network Traffic

- Traditionally, mobile device detection was based on specific fatures obtained from network protocols: DHCP options, HTTP User-agent, cookies...
- Due to massive encryption, we cannot obtained these features without decryption – available only on end devices or web proxy.
  - In 2019, 99,18% of the mobile traffic is encrypted.
  - Non encrypted protocols: DNS (0,31 %), HTTP (0,32%), ICMP/IGMP (0,07%)
- *Who is the sender? What kind of data are transmitted?*



TLS communication

- During TLS handshake, TLS parameters are negotiated.



Hypothesis:

- TLS handshake parameters depend on packages and methods that were used to build the application. These parameters are relatively stable and related to the application.

=> TLS handshake can be used for application fingerprinting [1,2].

[1] Husák, M., Čermák, M., Jirsík, T., Čeleda, P.: Https traffic analysis and client identification using passive SSL/TLS, fingerprinting. EURASIP Journal on Information Security (2016)

[2] Anderson, B., Paul, S., McGrew, D.: Deciphering malware's use of TLS (without decryption). Journal of Computer Virology and Hacking Techniques pp. 195-211 (2018)

How to Create a TLS Fingerprint using JA3 algorithm [3]

1. Extract selected fields from Client Hello: version, cipher suites, extensions, supported groups, EC formats.
2. Concatenate data in decimal format into one string.
3. Compute MD5 hash of the string => JA3 fingerprint of a client.

Version, Cipher Suites, Extensions, Supported Groups, EC format

0x00000303 - 49195,49196,52393,49199,49200,52392,158,159,49161,49162,49171,49172,51,57,156,157,47,53 - 65281,0,23,35,13,16,11,10 - 0x00000017,0x00000018,0x00000019 - 0

↓ Hex to Decimal Format

771, 49195-49196-52393-49199-49200-52392-158-159-49161-49162-49171-49172-51-57-156-157-47-53, 65281-0-23-35-13-16-11-10, 23-24-25, 0

↓ 32-bit MD5 hash

n8bvbvyZuTPF4tj89PaJVQ

- Similarly, we can compute JA3S fingerprint extracted from Server Hello data exchange [3] => it describes the server application.

[3] For original JA3 algorithm, see https://github.com/salesforce/ja3 or https://ja3er.com/

Research questions related to the JA3 fingerprints

- How reliable are JA3 fingerprints for mobile apps identification?
- Are JA3 fingerprints of mobile apps unique?
- How stable are JA3 fingerprints of mobile apps wrt. OS version?

FIT

Observations

1. Random values in TLS Cipher Suites, Extensions and Supported Group [4-6]
   - The large set of unique JA3 hash values for a given app.

   ⇒ Ignore these values from the fingerprint.

   ⇒ Reduction of JA3 hash values: # Opera hashes reduced from 155x to 4x.

List of Cipher Suites	List of Extensions	Supported Group	JA3 hash
49199-49200-49195-49196-52392-52393-49171-49161-49172-49162-156-157-47-53-49170-10	13172-0-5-10-11-13-65281-16-18	29-23-24-25	839868ad711dc55bde0d37a87f14740d
49199-49200-49195-49196-52392-52393-49171-49161-49172-49162-156-157-47-53-49170-10	13172-0-5-10-11-13-65281-16-18	29-23-24-25	839868ad711dc55bde0d37a87f14740d
56026-4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	60138-0-23-65281-10-11-35-16-5-13-18-51-45-43-27-19018-21	35466-29-23-24	ee972d7d47ec01a9cb9b04efb7346e32
60138-4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	39578-0-23-65281-10-11-35-16-5-13-18-51-45-43-27-56026-21	23130-29-23-24	cb4415a180704432d2e3f70f8dca5783
31354-4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	47802-0-23-65281-10-11-35-16-5-13-18-51-45-43-27-51914	43690-29-23-24	74a57a5f55ce2c9fa637b1f4567308b4
14906-4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	31354-0-23-65281-10-11-35-16-5-13-18-51-45-43-27-43690	56026-29-23-24	a10f93ffdc89d383db0f4437a0530569
49199-49200-49195-49196-52392-52393-49171-49161-49172-49162-156-157-47-53-49170-10	13172-0-5-10-11-13-16-18	29-23-24-25	5a291b49748c50adf1da70f8142d4cc4
49199-49200-49195-49196-52392-52393-49171-49161-49172-49162-156-157-47-53-49170-10	13172-0-5-10-11-13-16-18	29-23-24-25	5a291b49748c50adf1da70f8142d4cc4
4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	0-23-10-11-35-16-5-13-18-51-45-43-27	29-23-24	a839cfeed30d55439b09de5f1b47fa3a
4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	0-23-10-11-35-16-5-13-18-51-45-43-27	29-23-24	a839cfeed30d55439b09de5f1b47fa3a
4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	0-23-10-11-35-16-5-13-18-51-45-43-27	29-23-24	a839cfeed30d55439b09de5f1b47fa3a
4865-4866-4867-49195-49199-49196-49200-52393-52392-49171-49172-156-157-47-53-10	0-23-10-11-35-16-5-13-18-51-45-43-27	29-23-24	a839cfeed30d55439b09de5f1b47fa3a

[4] Benjamin, D.: Applying Generate Random Extensions And Sustain Extensibility (GREASE) to TLS Extensibility. IETF RFC 8701 (January 2020)
[5] Rescorla, E., Ray, M., Dispensa, S., Oskov, N.: Transport Layer Security (TLS) Renegotiation Indication Extension. IETF RFC 5746 (February 2010)
[6] Benjamin, D.:A Transport Layer Security (TLS) ClientHello Padding Extension. IETF RFC 7685 (October 2015)

Observations

2. Ads, tracking, web analytics create a “noise” during the training.
   - The same “noise” fingerprints are generated by various apps.

=> We need to exclude the “noise” from the training dataset.

=> We use blacklisting [7]

SrcIP	DstIP	Server Name Indication	JA3 Fingerprint
10.0.2.15	172.217.23.193	ci5.googleusercontent.com	d5dcde95b8fa38b5062a128f7eff0737
10.0.2.15	172.217.23.225	ci3.googleusercontent.com	d5dcde95b8fa38b5062a128f7eff0737
10.0.2.15	172.217.23.229	mail.google.com	81d2604dcc31ff39cdddb6079692b0b0
10.0.2.15	216.58.201.106	www.googleapis.com	193c522402283ed9e84b8bb38137829f
10.0.2.15	216.58.201.106	www.googleapis.com	3d9a16cdc1b2a98f6046af1c833054b8
10.0.2.15	216.58.201.74	android.googleapis.com	ca75d9d90e40897206fa2a08d9100df0
10.0.2.15	216.58.201.97	ci4.googleusercontent.com	d5dcde95b8fa38b5062a128f7eff0737

[7] See https://hosts-file.net/ad servers.txt, https://pgl.yoyo.org/adservers/, or https://gitlab.com/ookangzheng/dbl-oisd-nl

Observations

3. Ambiguity of JA3 fingerprints over applications.
   - Many JA3 hash values belong to multiple applications.

⇒ TLS fingerprint feature set need to be extended by server side features:
   - JA3S fingerprint,
   - Server Name Indication (SNI) field.

⇒ Does not work for apps communicating with multiple destinations, e.g., web browsers, etc.

- Example: Uniqueness of TLS features learnt from our dataset

Feature	Distinct items	Unique items	Uniqueness
JA3	30	21	70,00%
JA3+JA3S	122	114	93,44%
JA3+JA3S+SNI	154	153	99,35%

Distinct items = # of distinct JA3/JA3+JA3S/JA3+JA3S+SNI features in the dataset
Unique items = # of TLS features that are unique only to one mobile app in the dataset

Observations

4. Fingerprints stability over OS version is related to the used TLS libraries.

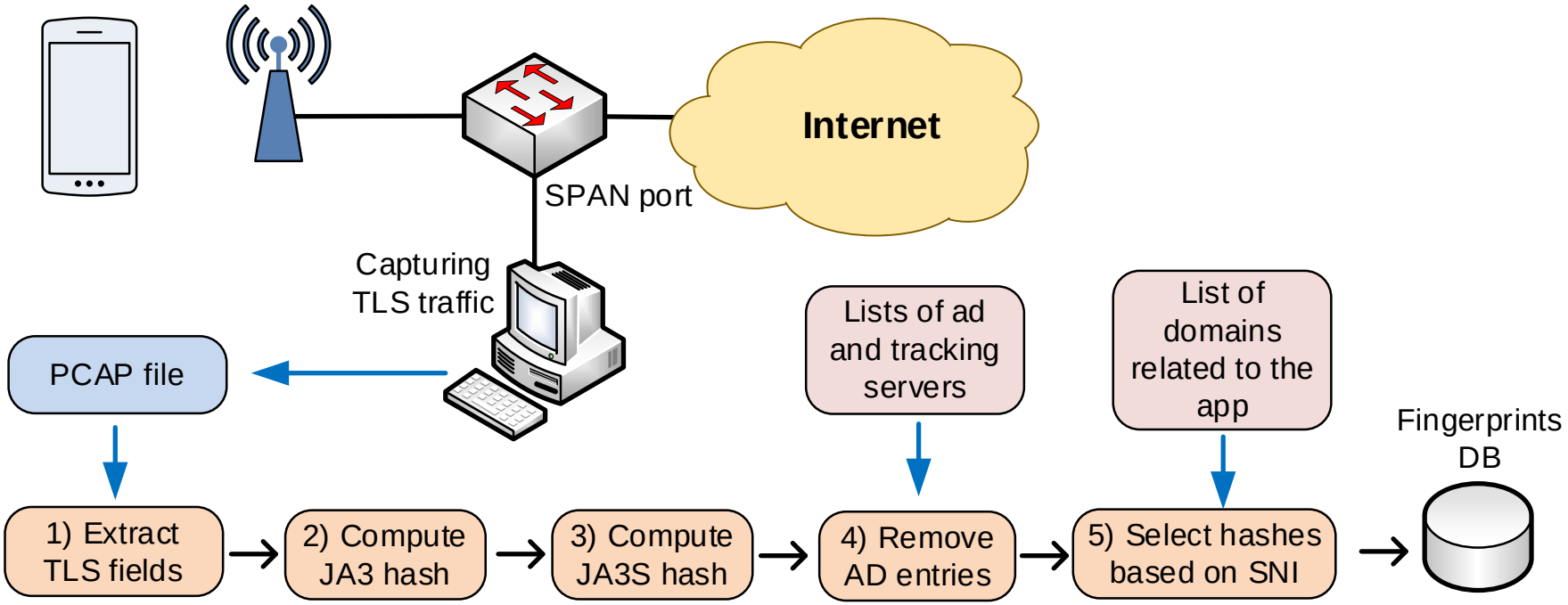
   $\Rightarrow$ TLS features may depend on the OS version.

Mobile App Feature	Android 7	Android 8.1		Android 9	
	present	added	missing	added	missing
CP JA3	1	1	1	0	0
CP JA3S	2	2	2	0	0
CP SNI	2	0	0	0	0
Mujvlak JA3	1	1	1	0	0
Mujvlak JA3S	1	0	0	0	0
Mujvlak SNI	1	0	0	0	0
Reddit JA3	1	2	1	0	0
Reddit JA3S	1	2	1	0	0
Reddit SNI	9	3	2	4	0
Seznam CZ JA3	3	3	3	2	1
Seznam CZ JA3S	11	0	0	2	0
Seznam CZ SNI	12	0	1	1	0

JA3 hash, JA3S hash, Server Name Indication (SNI) string for selected apps.

Learning Mobile App fingerprints

- Training dataset obtained using Android Virtual Studio Emulator or real devices.
- We observed about 30 different apps (datasets MA1-4, see Section 5.1).



Learning Results – Mobile App Fingerprints Database (MA1-4 datasets)

Mobile App	JA3 hashes	JA3S hashes	SNI strings	unique JA3+JA3S+SNI	Features
Accuweather	4	5	4	10	All
Boomplay Music	2	5	2	5	All
Cestovne Poriadky	2	4	2	4	All
Chrome	1			1	JA3
Discord	4	3	4	18	All
Duolingo	2	2	3	4	All
EquaBank CZ	2	3	2	4	All
Facebook	8	7		10	JA3+JA3S
Gmail	5	8	2	10	All
Google Calendar	1	1	1	1	All
KB klic	1	1	2	2	All
Mobilni banka	2	1	2	2	All
Muj vlak	2	1	1	2	All
Nextbike	2	6	5	11	All
Reddit	3	7		12	JA3+JA3S
Seznam	7	12		37	JA3+JA3S
Slack	3	4	3	6	All
TikTok	2	2	3	3	All
Tor	1			1	JA3
Viber	1	1	1	1	All
WhatsApp	5	5	4	7	All
Youtube	2	2	2	3	All

Detection using JA3 hash value only (MA4 dataset)

- Works well for C,D,K,L, other apps have the same JA3 hashes

		Real values															
Predicted values		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	X
	A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	14
	D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	12
	E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	K	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
	L	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	13
	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	X	5	5	0	0	1	6	1	4	9	1	1	0	13	6	1	1	139

Boomplay Music (A), EquaBank (B), Facebook (C), Gmail (D), Google Calendar (E), Chrome App (F), KB Klic (G), Mobilni Banka (H), NextBike (I), TikTok (J), WhatsApp (K), Youtube (L), Seznam CZ (M), Reddit (N), Muj vlak (O) and Cestovne Poriadky (P).

Letter X describes unknown traffic: column X (false positives), row X (false negatives)

Detection using JA3 + JA3S hash values only (MA4 dataset)

- Correctly identified A, B, C, D, F, H, I, K, L, M, N, O, P but high # of FP

Predicted values \ Real values	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	X
A	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	13
D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	18
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	4
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	10
M	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	2
N	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	2	0	0	1	4	1	3	3	1	2	0	3	0	0	0	131

Boomplay Music (A), EquaBank (B), Facebook (C), Gmail (D), Google Calendar (E), Chrome App (F), KB Klic (G), Mobilni Banka (H), NextBike (I), TikTok (J), WhatsApp (K), Youtube (L), Seznam CZ (M), Reddit (N), Muj vlak (O) and Cestovne Poriadky (P).

Letter X describes unknown traffic: column X (false positives), row X (false negatives)

Detection using JA3 hash + JA3S hash + SNI values (MA4 dataset)

- High accuracy by identification except of F

Predicted values \ Real values	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	X
A	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	8
D	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	1	0	0	0	6	0	1	2	1	1	1	1	0	0	0	170

Boomplay Music (A), EquaBank (B), Facebook (C), Gmail (D), Google Calendar (E), Chrome App (F), KB Klic (G), Mobilni Banka (H), NextBike (I), TikTok (J), WhatsApp (K), Youtube (L), Seznam CZ (M), Reddit (N), Muj vlak (O) and Cestovne Poriadky (P).

Letter X describes unknown traffic: column X (false positives), row X (false negatives)

Comparison of feature sets wrt. accuracy and precision

- JA3 is reliable only for specific apps and produces many false negatives.
- JA3+JA3S classification has the comparable accuracy but better recall.
- The best results are given by combination of JA3+JA3S+SNI.

	Total items	Accuracy	Precision	Recall
JA3	244	61,89%	23,53%	18,18%
JA3+JA3S	244	72,54%	49,46%	69,70%
JA3+JA3S+SNI	244	90,98%	86,67%	78,79%

Accuracy = (TP+TN) / (P+N) – the percentage of correctly classified samples
Precision = TP / (TP + FP) – exactness, the percentage of samples correctly labeled as true positives
Recall = TP / (TP + FN) = TP / P – completeness, the percentage of positive samples correctly identified

Summary

- The proposed methods enables automated creation of TLS fingerprint database based on JA3+JA3S+SNI feature set.
- The combination of these features produces reliable and unique fingerprints of the mobile app.
- Stability of fingerprints is related to the mobile apps version and the operating system.
- Detection can be easily implemented using IPFIX flow monitoring extended with TLS headers obtained from the network traffic.
- Proof-of-concept implementation is available in http://github.com/matousp/ja3-fingerprinting.

The research was supported by project "Integrated platform for analysis of digital data from security incidents" (Tarzan), 2017-2020, No. VI20172020062 granted by Ministry of Interior of the Czech Republic.