CS@Mines

Poster: Data Collection for ML Classification of Encrypted Messaging Applications

Jason Hussey, *Colorado School of Mines (presenting)*
Ethan Taylor, *Colorado School of Mines*
Kerri Stone, *iCR, inc.*
Tracy Camp , *Colorado School of Mines*

https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-outage-instagram-whatsapp-signal-down-b1932505.html

"Signal is regularly used by journalists and investigators to protect sources identity"

Users in 2020:

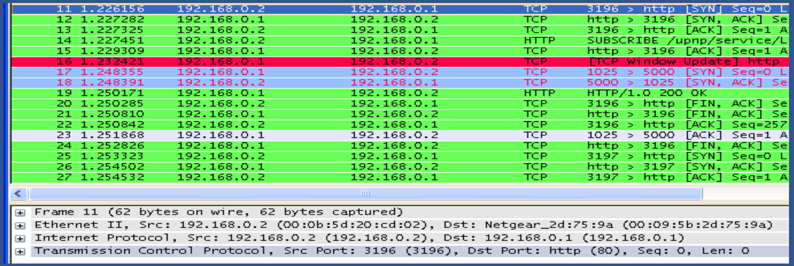
WhatsApp, 2 billion
Telegram, 400 million
Signal, 20 million

https://www.businessofapps.com/data/signal-statistics/
https://www.businessofapps.com/data/telegram-statistics/
https://www.businessofapps.com/data/whatsapp-statistics/

Signal: The Pros and Cons of a Truly Private Chat App

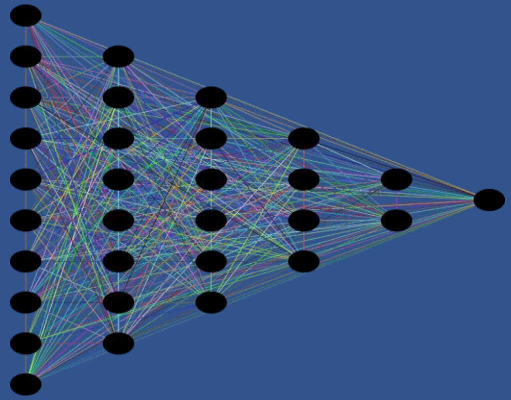
Signal, the encrypted messaging app, is seeing record numbers of downloads amid the pandemic and nationwide protests. It might make sense for you, too.

https://www.wsj.com/articles/signal-the-pros-and-cons-of-a-truly-private-chat-app-11592127002



11	1.226156	192.168.0.2	192.168.0.1	TCP	3196 > http [SYN] Seq=0 L
12	1.227282	192.168.0.1	192.168.0.2	TCP	http > 3196 [SYN, ACK] Se
13	1.227325	192.168.0.2	192.168.0.1	TCP	3196 > http [ACK] Seq=1 A
14	1.227451	192.168.0.2	192.168.0.1	HTTP	SUBSCRIBE /upnp/service/L
15	1.229309	192.168.0.1	192.168.0.2	TCP	http > 3196 [ACK] Seq=1 A
16	1.232421	192.168.0.1	192.168.0.2	TCP	[TCP Window Update] http
17	1.248355	192.168.0.1	192.168.0.2	TCP	1025 > 5000 [SYN] Seq=0 L
18	1.248391	192.168.0.2	192.168.0.1	TCP	5000 > 1025 [SYN, ACK] Se
19	1.250171	192.168.0.1	192.168.0.2	HTTP	HTTP/1.0 200 OK
20	1.250285	192.168.0.2	192.168.0.1	TCP	3196 > http [FIN, ACK] Se
21	1.250810	192.168.0.1	192.168.0.2	TCP	http > 3196 [FIN, ACK] Se
22	1.250842	192.168.0.2	192.168.0.1	TCP	3196 > http [ACK] Seq=257
23	1.251868	192.168.0.1	192.168.0.2	TCP	1025 > 5000 [ACK] Seq=1 A
24	1.252826	192.168.0.1	192.168.0.2	TCP	http > 3196 [FIN, ACK] Se
25	1.253323	192.168.0.2	192.168.0.1	TCP	3197 > http [SYN] Seq=0 L
26	1.254502	192.168.0.1	192.168.0.2	TCP	http > 3197 [SYN, ACK] Se
27	1.254532	192.168.0.2	192.168.0.1	TCP	3197 > http [ACK] Seq=1 A

Frame 11 (62 bytes on wire, 62 bytes captured)
Ethernet II, Src: 192.168.0.2 (00:0b:5d:20:cd:02), Dst: Netgear_2d:75:9a (00:09:5b:2d:75:9a)
Internet Protocol, Src: 192.168.0.2 (192.168.0.2), Dst: 192.168.0.1 (192.168.0.1)
Transmission Control Protocol, Src Port: 3196 (3196), Dst Port: http (80), Seq: 0, Len: 0



Research Summary

- **Network traffic classification** is used to identify the nature of traffic on a network.
- Entities capable of monitoring net-work traffic use classification for all manner of reasons, including **identification of mobile applications being used on the network**.
- It is possible that the usage of encrypted messaging applications by users on these networks can be detected, **betraying elements of their privacy.**
- We describe a system that:
  - leverages campus network resources to generate real-world data
  - alongside a more curated dataset captured from Android application traffic.
- We also explore the ability of machine learning (ML) models to accurately classify traffic from these encrypted messaging applications.

CS@Mines

Methodology – Data Collection

WiFi Data Collection

- Partner with the ITS office to collect anonymous WiFi packet headers
- Leverage ntop’s n2disk utility
  - Zero copy drivers
- Extract just the IP and TCP/UDP headers and pre-process with tshark
- Multiprocess the tshark output into mongodb

Android Application Collection

- Rooted Android phones (Samsung and Xiaomi)
- X-compiled strace attached to Signal messaging app process
- netstat polling for verification
- tcpdump on a Ubuntu station serving as AP
- Filter the PCAP file to only those flows identified by socket calls in trace

```
23903 getsockopt(84, SOL_SOCKET, SO_DOMAIN, [10], [4]) = 0
23903 socket(AF_UNIX, SOCK_STREAM|SOCK_CLOEXEC, 0) = 93
23903 connect(93, {sa_family=AF_UNIX, sun_path="/dev/socket/fwmarkd"}, 110) = 0
23903 sendmsg(93, {msg_name=NULL, msg_namelen=0, msg_iov=[{iov_base="\1\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0", iov_len=16}, {iov_base=NULL, iov_len=0}], msg_iovlen=2,
msg_control=[{cmsg_len=20, cmsg_level=SOL_SOCKET, cmsg_type=SCM_RIGHTS, cmsg_data=[84]}], msg_controllen=24, msg_flags=0}, 0) = 16
23903 recvfrom(93,  <unfinished ...>
23904 socket(AF_INET6, SOCK_STREAM, IPPROTO_IP <unfinished ...>
23903 <... recvfrom resumed> "\0\0\0\0", 4, 0, NULL, NULL) = 4
23904 <... socket resumed> )          = 98
23903 connect(84, {sa_family=AF_INET6, sin6_port=htons(443), inet_pton(AF_INET6, "::ffff:76.223.92.165", &sin6_addr), sin6_flowinfo=htonl(0), sin6_scope_id=0}, 28) = -1
EINPROGRESS (Operation now in progress)
23903 socket(AF_UNIX, SOCK_STREAM|SOCK_CLOEXEC, 0) = 93
23903 connect(93, {sa_family=AF_UNIX, sun_path="/dev/socket/fwmarkd"}, 110) = 0
23903 sendmsg(93, {msg_name=NULL, msg_namelen=0, msg_iov=[{iov_base="\6\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0", iov_len=16},
```

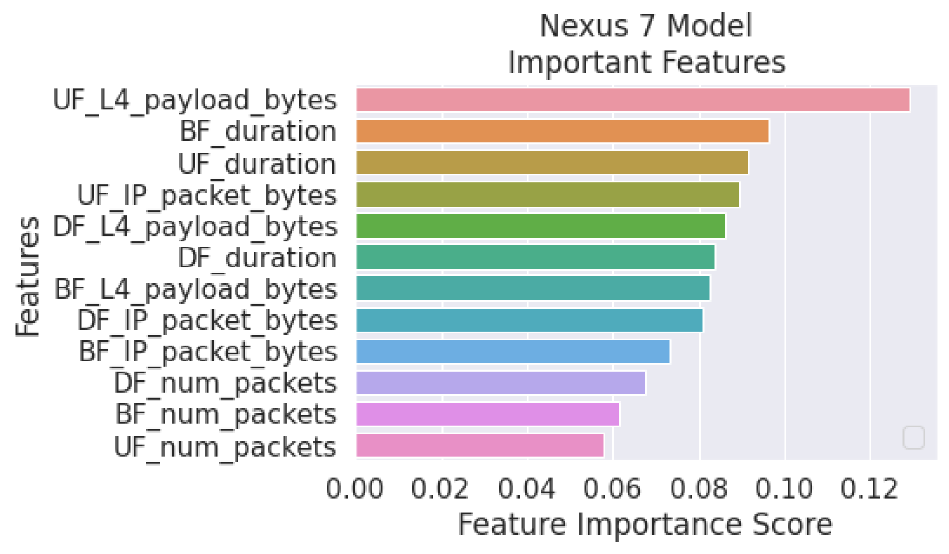
Methodology – Data Analysis

- Traffic object we examine is the bi-directional flow
  - Uniquely identified by the 5-tuple of source IP, source port, destination IP, destination port, and which protocol (TCP or UDP)
  - These are not features, just unique identifiers
- Direction, timing, and size are preserved as a ‘feature’
- Many other statistical features can then be created to describe these flows
  - E.g., total bytes sent, momentum of the conversation, in addition to the mean, max, min, variance, etc.

CS@Mines

ML applications

- **Some initial proof-of-concept multi-class classification**
- **Off the shelf classifiers; in our experiments Random Forests worked very well.**
- **Trained a classifier on MIRAGE data's Nexus 7 flows to classify apps from a different phone's flows**
- **In this particular case, the upstream L4 payload was of high importance.**
  - **This intuitively suggests that the client side behavior is an important discriminator**



Future Work

- Describe the system and considerations in greater detail to assist researchers
  - Emphasizing the partnership opportunities with host institutions
  - Allow other researchers to similarly extend the MIRAGE dataset
- ML applications
  - Extending the MIRAGE dataset with our own custom applications in the same format
  - Applying classifiers to 'real world' WiFi dataset from Mines
  - Expanding the 'positive class' from just a single application to the genre of Encrypted Messaging Applications