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*
“Signal is regularly used by journalists and investigators to protect sources identity”

Users in 2020:
WhatsApp, 2 billion
Telegram, 400 million
Signal, 20 million

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.
Research Summary

• **Network traffic classification** is used to identify the nature of traffic on a network.

• Entities capable of monitoring network 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.
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
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.
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