**Feature Extraction**

In this task, we aim to extract feature for each sequentially collected packet. Given the existing data, the attributes of a data package are dynamic. Being dynamic means that some features of some data packets may not exist for some other packets. For example, TCP packets do not have the features of security socket layer, or the TLS packets intended for handshake layer have certificate attributes while those for record layer have not. Nevertheless, we need to turn any packet into a numeric vector so that it can be trained in machine learning algorithm. This document illustrates how a given packet, p, is to be transferred to a numeric vector.

In our task, the data packets are classified as either TCP packets or SSL/TLS packets. Moreover, SSL/TLS packets have more cryptographic features.

**Shared Features (TCP features):**

|  |  |  |
| --- | --- | --- |
| Feature | Type | Description |
| Come/Leave | boolean (0/1) | From the view of webserver, whether the packet is leaving from or coming to the server. |
| Protocol | enum: TCP, SSL2, SSL3, TLS1, TLS2, TLS3 | *Transform into a one-hot vector* |
| Length | integer |  |
| Interval | float (unit: ms) | The time interval from previous packet |
| Flag | Enum: NS, CWR, ECE, URG, ACK, PSH, RST, SYN, FIN | See: https://en.wikipedia.org/wiki/Transmission\_Control\_Protocol |
| Windows Size | integer |  |

**Dynamic Features (SSL/TLS features):**

The following features are optional for some packets.

*Dimension is zero if does not contain the feature.*

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | | | Type |
| Handshake Protocol | ClientHello | Length | integer |
| Cipher suites  *Enumerate all possible cipher suites* | enum:  TLS\_ECDHE\_ECDSA\_WITH\_AES\_256\_GCM\_SHA384  … |
| Cipher suites length | integer |
| Compression method  *Enumerate all possible compression method* | enum: null, … |
| Supported group length | Integer |
| Supported groups | Enum: x25519, secp256r1, … |
| Encrypt then mac length | Integer |
| Extended master secret | Integer |
| Signature hash algorithm | Enum: |
| ServerHello | Length | integer |
| Extended master secret (cant find this in wireshark) | Integer |
| Renegotiation info length | integer |
| Certificate | Average certificate length | integer |
| Maximum certificate length | integer |
| Minimum certificate length | integer |
| Number of Certificates | integer |
| Signature algorithm | Enum |
| Server Hello Done | Length | integer |
| Client Key Exchange | length | integer |
| Pubkey length | integer |
| Encrypted Handshake Message | length | integer |
| Change Cipher Spec | Change Cipher Spec | length | integer |
| Application Data Protocol | Application Data Protocol | length | integer |

*Each row corresponds to a packet*

For Discussion:

1. We need to complement some features (red ones) with more data.
2. What features should be used for classical machine learning algorithm? How do we aggregate the data?