Global Approach

**Problems**
- Traffic analysis pipelines are complex
- What is the best data/preprocessing/analysis/... method?
- Most papers implement the full pipeline without justification for each component
- Is any of these parts “solved”?
- Where does a new researcher to the field start?
- Everyone seems to be doing different things, with no clear trends in the community

**Network Traffic Meta-analysis**
- What are people doing in Network Traffic Analysis?
- How are people doing it?
- What are the best practices? Are they being followed?
- Which approaches are lacking?

**Our Proposal**
We propose a formalized data structure that lets us address these questions (and more)
- A JSON file corresponds to each file
- Database is publicly available (71 papers) [TU Wien CN Group 2017b]
- Format documentation is public
- Contributions are welcome!
- Tools are coming...

**Collected Parameters**
Reference: title, authors, journal, open-access
Data: dataset name, availability, format, traffic type/protocol, captured/synthetic, year, length, anonymization
Preprocessing: feature selection technique + type (filter, wrapper,..., packet features + goal, flow features + key + goal + timeout + direction, flow aggregation features + key + goal + timeout, tools, normalization, transformations
Analysis Method: supervised/unsupervised/anomaly detection was used, tools, algorithm name + learning type + metric + source + parametrization provided
Evaluation: algorithm comparison was made, internal/external validation, dpi/port-based truth, real scenario, train/test split, methods names + types + metrics + sources

**Results**
- Publication with results about features:
- Format specification + database is completely public; tools will follow shortly
- Incentivize external researchers to contribute
- Continue improving/adjusting format
- Step-by-step approach: have a more in-depth look at each piece of the pipeline

References
- TU Wien CN Group (2017b). Network Traffic Analysis Database. URL: https://www.cn.tuwien.ac.at/re-disp/ntadatabase.html

Step-by-step Approach

**What are the “best” features for traffic analysis?**

<table>
<thead>
<tr>
<th>Features</th>
<th>Importance</th>
<th>Usage Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>TimeStamp</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Protocol</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Source</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Destination</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>Payload</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

(a) Most used and most cited features in our database.

(b) Frequency of use of features in publications in our database.

**Feature Learning**
**Need:**
- Represent traffic with numeric vectors

**Difficulties:**
- Choice of features is not obvious

**Current approach:**
- Learn which feature vectors to use
- Deep Learning has many successful uses of this idea, but not in network traffic

**Stream Processing**
**Need:**
- Stream capable framework
- Ability to run multiple algorithms for comparison

**Difficulties:**
- Avoid duplicate work
- Assert complete and easy reproducibility

**Current approach:**
- Modular structure, each module independent of others
- Each module takes streaming input, as it becomes available
- Each module is one Docker container
- Messaging done by Apache Kafka

**Stream Clustering**
**Need:**
- Clustering approach for a continuous stream of data

**Difficulties:**
- Stream is potentially infinite
- Input distribution changes throughout time (concept drift)

**Current approach:**
- Try existing state-of-the-art algorithms
- Identify deficiencies when applied to network traffic

**Problems:**
- A framework for testing is necessary (see Stream Processing, above)