Streaming Data Analysis Methods for TCP/IP Network Traffic

Daniel C. Ferreira, TU Wien
daniel.ferreira@tuwien.ac.at

TECHNISCHE UNIVERSITÄT WIEN

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...

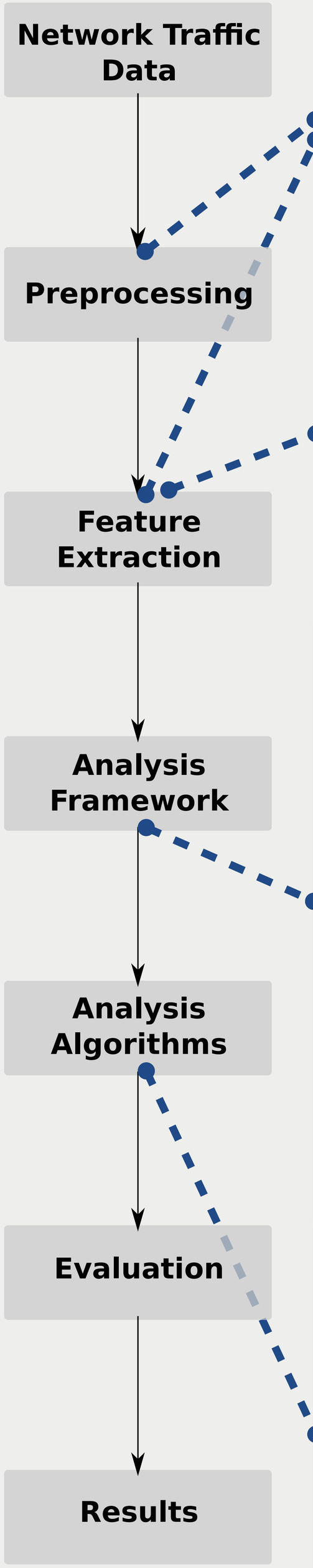
Collected Parameters

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
Conclusion: goal of the paper, focus of the paper, improvement claims, reproducibility

* blue means optional parameters

What Next?

- Publication with results about features:
  Daniel C. Ferreira, Félix Iglesias Vázquez, Gernot Vormayr, Maximilian Bachl, and Tanja Zseby. 2017. A Meta-Analysis Approach for Feature Selection in Network Traffic Research. In *Proceedings of the 2017 Reproducibility workshop, ACM SIGCOMM, Los Angeles, CA, USA, August 25, 2017*, 4 pages
- 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



Step-by-step Approach

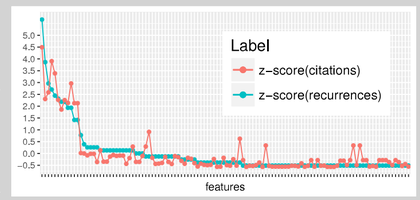
Meta-analysis of Features

What are the “best” features for traffic analysis?

Features (recurrences)	score[1]	Features (citations)	score[2]
octetTotalCount	5.8	octetTotalCount	4.6
packetTotalCount	3.9	ipTotalLength	3.9
flowDurationMilliseconds	3.1	destinationTransportPort	3.5
ipTotalLength	2.7	sourceTransportPort	3.0
destinationTransportPort	2.5	flowDurationMilliseconds	2.6

[1]: z-score(*recurrences*) [2]: z-score(*citations*)

(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

Problems:
- Kafka might not be fast enough (more testing needed)

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)

References

- *A Meta-Analysis Approach for Feature Selection in Network Traffic Research* (2017). P. 4. DOI: 10.1145/3097766.3097771
- TU Wien CN Group (2017b). *Network Traffic Analysis Database*. URL: https://www.cn.tuwien.ac.at/ns-dksp/ntadatabase.html