Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets

Network Traffic Measurement and Analysis Conference (TMA 2018)
June 28, 2018

Milan Cermak et al.
Institute of Computer Science, Masaryk University, Brno
Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets

Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno
TMA 2018: Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets
Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno
Research Problems
challenges that everyone has to deal with

- **Lack of research standards**
  missing rules for research data collection, analysis, sharing, and ethics of usage

- **Inaccessibility of appropriate datasets**
  real-world data cannot be reliable annotated and needs to be anonymized, artificial data are not sufficiently realistic and provides a limited set of network traffic

- **Inability to prove research results**
  it is complicated to prove properties of the proposed analytical method leading to limited acceptance of the results by industry

- **Missing verification of others researchers’ results**
  data and algorithms are kept in private which leads to the impossibility of research reproducibility
Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets
Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno
TMA 2018: Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets
Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno
The Basic Idea
what we realized during our research

- **Single event** full packet capture can be publicly shared
  units of network traffic with one type of network event contains only a minimum of personal data and can be publicly shared and easily annotated

- **Packet capture can be „simply“ manipulated**
  MAC and IP addresses can be changed to predefined values together with capture time and subsequently adapted to real-world data

- **Events can be mixed with each other** or with real-world data
  we usually have access to the real-world data, but we need an annotation or a ground truth
Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets

our goal is not to deal with all identified problems at this point, but to present a general solution in order to start a discussion of its usability
Semi-Labeled Datasets
we aim to cover all areas relevant to datasets usage

1. **Creation** of annotated units

2. **Use of semi-labeled datasets** composed of annotated units

3. **Sharing platform** for annotated units

4. Use of semi-labeled datasets for a research evaluation

**TMA 2018**: Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets
Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno
Challenges of Shared Datasets
usage and creation requirements to support applicability

- **Data anonymization**
  problems of application data and consistency in all packet layers

- **Traffic annotation**
  either inaccurate annotation of real-world datasets or accurate annotation of an artificial dataset but with insufficient authenticity

- **Capture parameters**
  network topology, capacity, utilization, and latency affects the dataset creation

- **Dataset recency**
  each fixed dataset becomes obsolete in time
Annotated Units

normalized and annotated packet traces containing a single event

**Creation** of full packet traces
- filter the desired traffic from an existing network
- capture a traffic form a prepared environment

Packet trace **normalization**
- change MAC and IP addresses to predefined values
- reset timestamp to zero epoch time

Units **annotation**
- store information about author, capture interface, network settings, and trace content

[Diagram of packet trace creation and normalization]

github.com/CSIRT-MU/trace-share
Annotated Units
besides benefits, there are still issues that need to be addressed

- No sensitive content of a traffic
- Accurate annotation
- Easily accessible data recency
- Uniformity of virtual environment
- Normalization problems
- Trace consistency preservation
Semi-Labeled Datasets

we aim to cover all areas relevant to datasets usage

1. **Creation** of annotated units

2. **Use of semi-labeled datasets** composed of annotated units

3. **Sharing platform** for annotated units

4. Use of semi-labeled datasets for a **research evaluation**
Combination of Annotated Units
how to create a semi-labeled dataset

1. Select annotated units based on your interest
2. Capture real-world network traffic within your environment
3. Compute characteristics of the real-world traffic capture
4. Modify annotated units to reflect characteristics of the real-world traffic
5. Merge annotated units and real-world traffic capture
Usage of Semi-Labeled Datasets

development of analytical methods using annotated units

Annotated Units → Semi-labeled Datasets → Real-world Traffic

Comprehension → Prototype → Finalised Method
Semi-Labeled Datasets
we aim to cover all areas relevant to datasets usage

1. **Creation** of annotated units

2. **Use of semi-labeled datasets** composed of annotated units

3. **Sharing platform** for annotated units

4. Use of semi-labeled datasets for a **research evaluation**
Sharing Platform Challenges

each of dataset sharing platforms suffers from common issues

- Data **anonymization**
  assisted anonymization of uploaded datasets should be one of the key features of a central dataset sharing platform

- Data **heterogeneity**
  sharing platform should have clearly defined types and format of datasets it collects

- Platform **sustainability**
  a necessity to have a founding and create the platform as an open community hub

- Initial **content**
  sharing platforms should contain a sufficient number of up-to-date datasets when launched
Data Sharing Platform

our plans with trace·share open platform

- **Community** hub
- **Storage and management** of annotated units
- **Assisted** uploading, normalization, annotation, and mixing of annotated units

- Inspired by OpenML platform (see [https://openml.org](https://openml.org))
- Prototype available at the end of the year (see [https://github.com/CSIRT-MU/traceshare](https://github.com/CSIRT-MU/traceshare))
Semi-Labeled Datasets
we aim to cover all areas relevant to datasets usage

1. **Creation** of annotated units

2. **Use of semi-labeled datasets** composed of annotated units

3. **Sharing platform** for annotated units

4. Use of semi-labeled datasets for a *research evaluation*
Challenges of Research Evaluation

an evaluation must give an objective metric of the method efficiency

▪ **Qualitative** aspect
  properties of a dataset itself whereas the network traffic capture must contain realistic, diverse data, that accurately reflect real-world traffic

▪ **Quantitative** aspect
  the process of evaluation giving an objective metric of the method efficiency, typically using confusion matrix with true positive, false positive, and false negative values
Evaluation Using Semi-Labeled Dataset
combination of qualitative and quantitative aspects

- **Ground truth** of the dataset based on inserted annotated units
- **Balanced** quantitative and qualitative aspects
- Unknown positives need to be **verified manually** and shared

![Evaluation Diagram]

- Annotated units
- Identified events
- Uncertainty
Semi-Labeled Datasets in a Nutshell

quick conclusion and a discussion of possible problems, solutions, (crazy) ideas, or anything else
Summary
what you should remember from this presentation

- No need to share the entire network traffic, share only selected events!
- **Combine events** between themselves and with real-world traffic
- **Share your differences** and provide annotated units to others
- **Prove** your research results!

- If you are interested in this topic contact me at cermak@ics.muni.cz
Prove your research by shared trace!

https://github.com/csirt-mu/trace-share

@csirtmu

Milan Cermak et al.
cermak@ics.muni.cz