A Labeled Data Set
For Flow-based Intrusion Detection

Anna Sperotto, Ramin Sadre,
Frank van Vliet, Aiko Pras

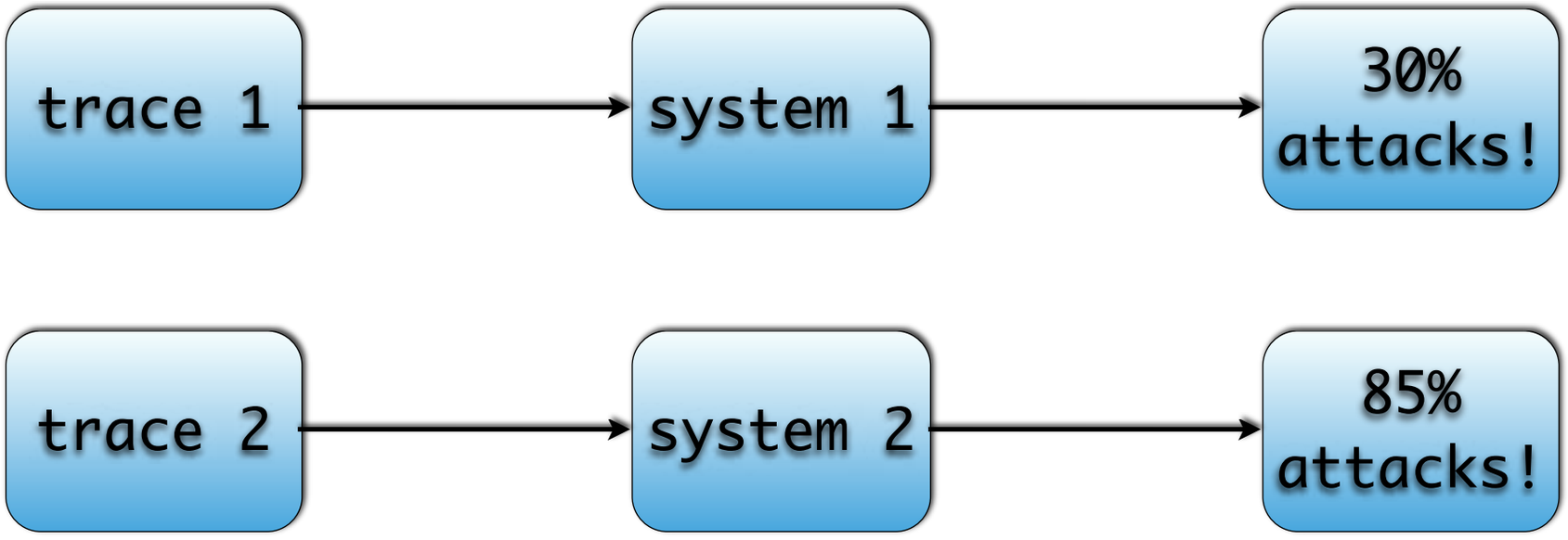
Design and Analysis of Communication Systems
University of Twente, The Netherlands

NMRG Workshop on Netflow/IPFIX Usage in Network Management
Maastricht - July 30, 2010

Contents

- Operational experience in trace collections
  - Experimental Setup
- Data processing and labeling
- The labeled data set

Introduction



- Systems are evaluated on proprietary traces
- No shared ground truth
- Results cannot be directly compared!

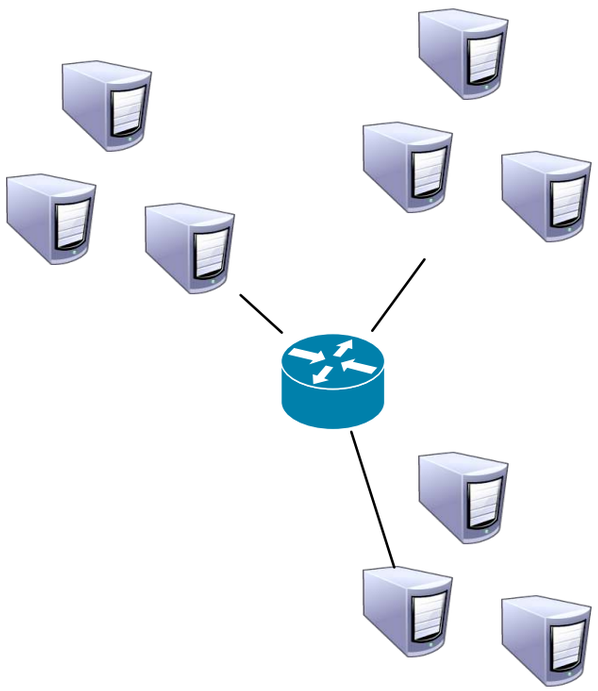
Data set requirements

We want the data set to be:

- *realistic data*
- *complete* and *correct* in labeling
- achievable in an acceptable *labeling time*
- sufficient *trace size*

The requirements will determine the collection setup

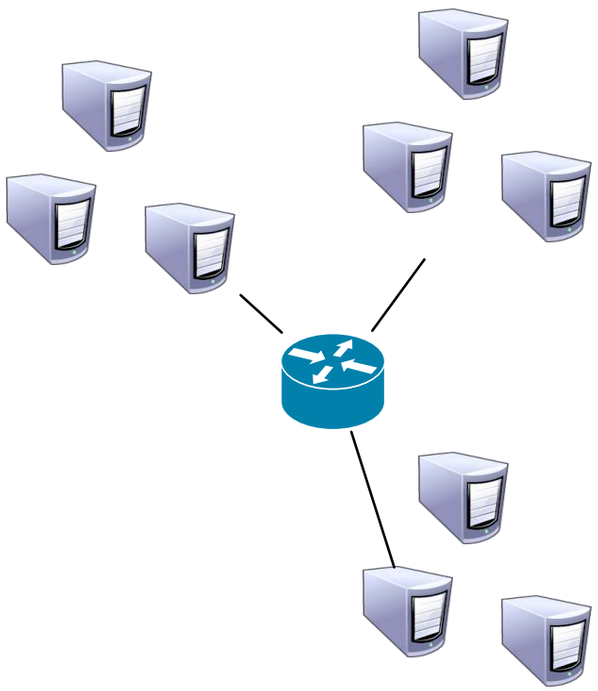
Measurement scale



NETWORK

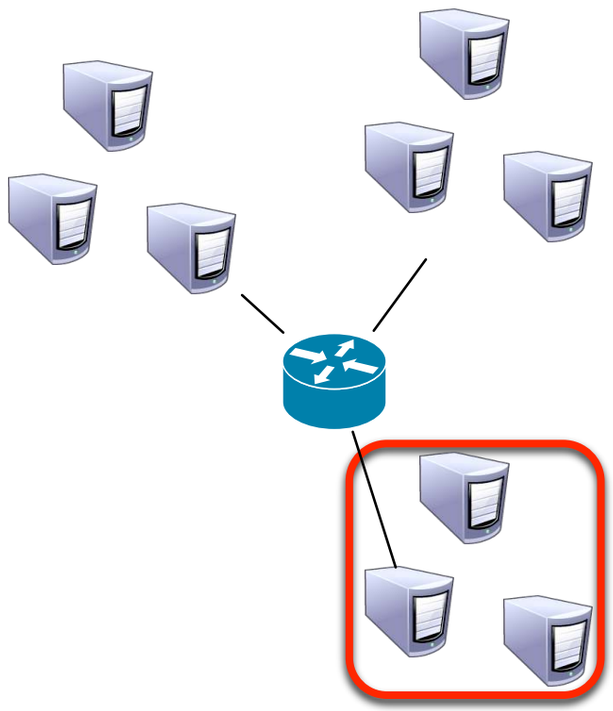
- realistic
- not complete
- it does not scale

Measurement scale



NETWORK

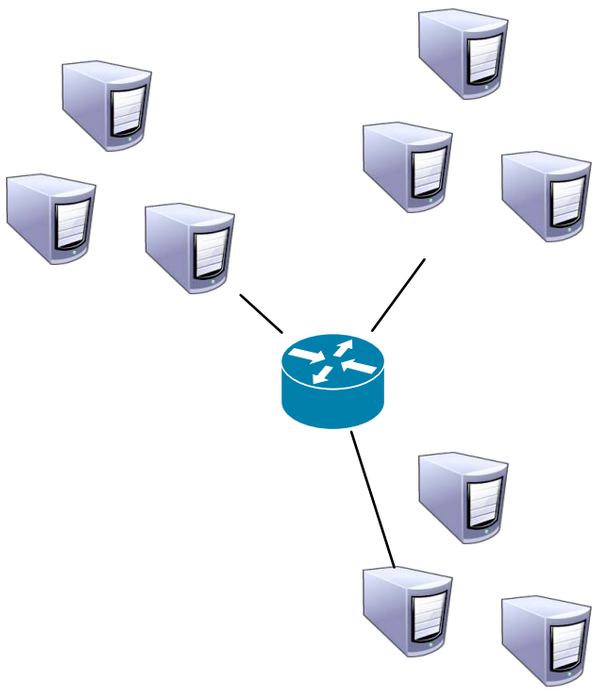
- realistic
- not complete
- it does not scale



SUBNETWORK

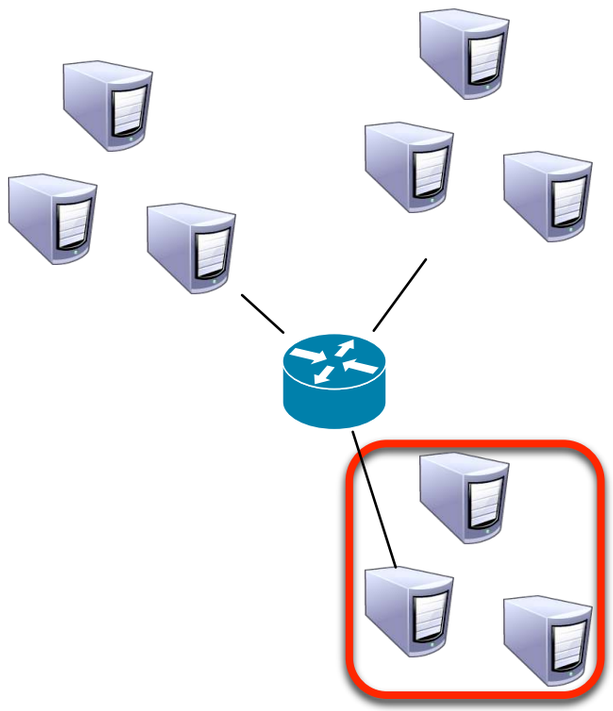
- realistic
- not complete

Measurement scale



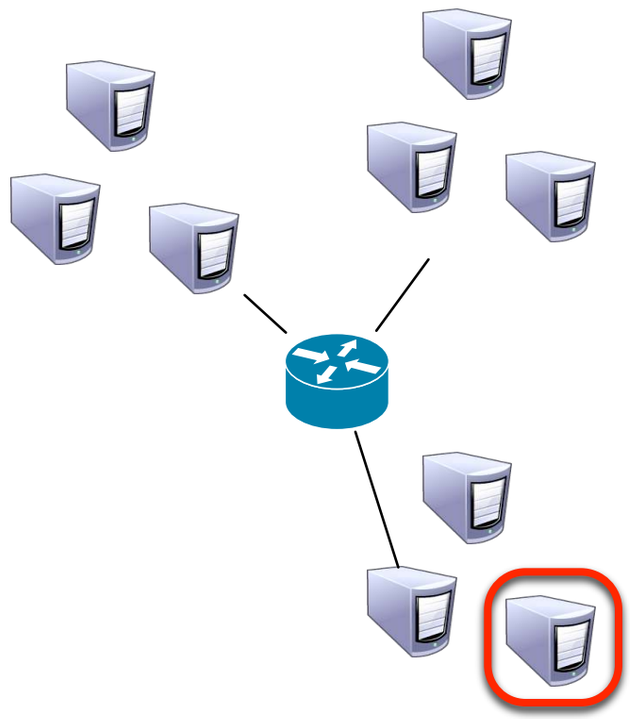
NETWORK

- realistic
- not complete
- it does not scale



SUBNETWORK

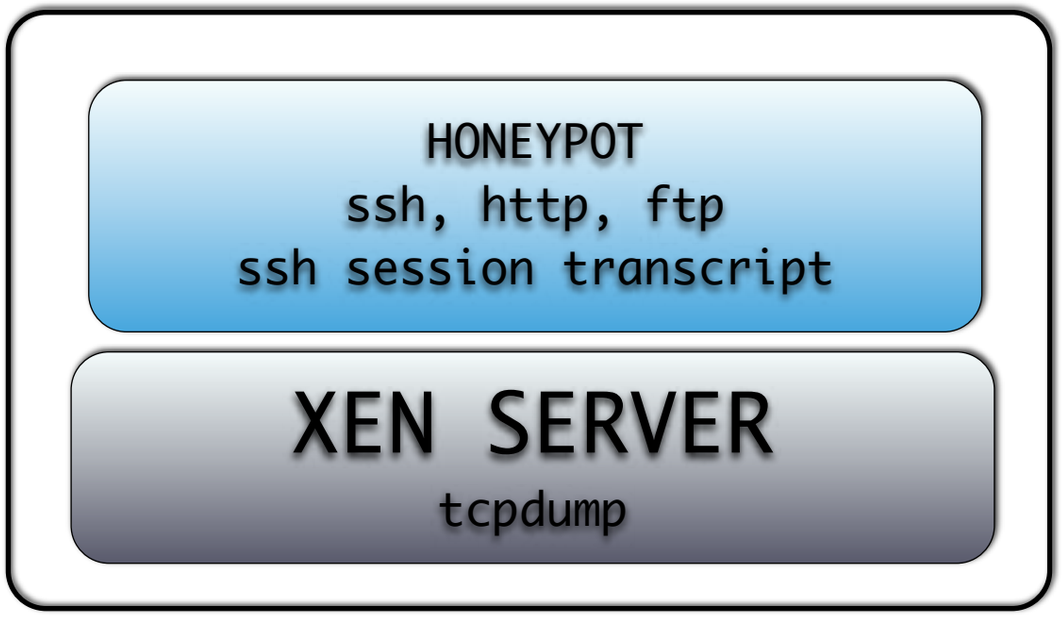
- realistic
- not complete



SINGLE HOST

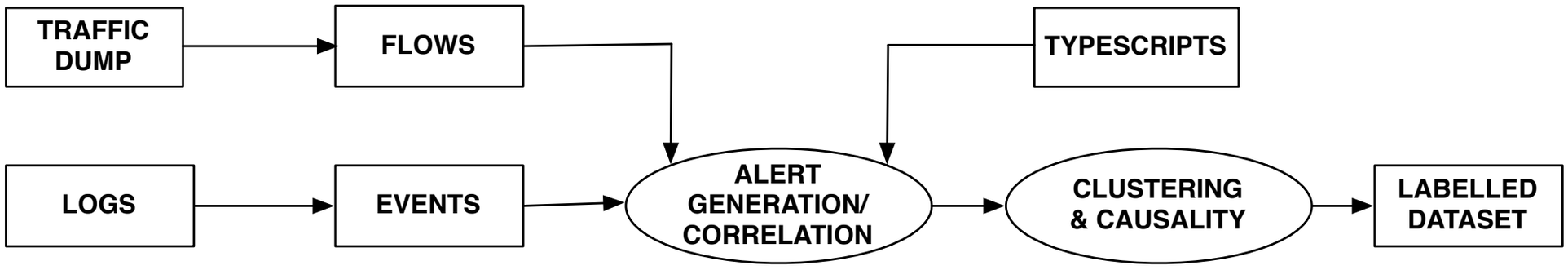
- realistic
- *enhanced logging* (honeypot)

Setup



- daily used services with enhanced logging
- direct connection to the Internet
- attack exposure
- complete tcpdump of the traffic (offline flow creation)

Data set creation



Preprocessing

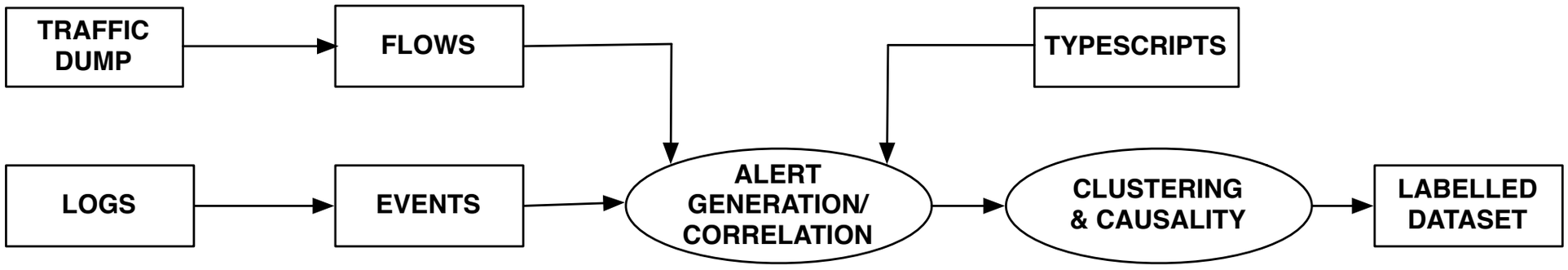
- packets ⇨ flows

$$F = (I_{src}, I_{dst}, P_{src}, P_{dst}, Pckts, Octs, T_{start}, T_{end}, Flags, Prot)$$

- logs ⇨ log events

$$L = (T, I_{src}, P_{src}, I_{dst}, P_{dst}, Descr, Auto, Succ, Corr)$$

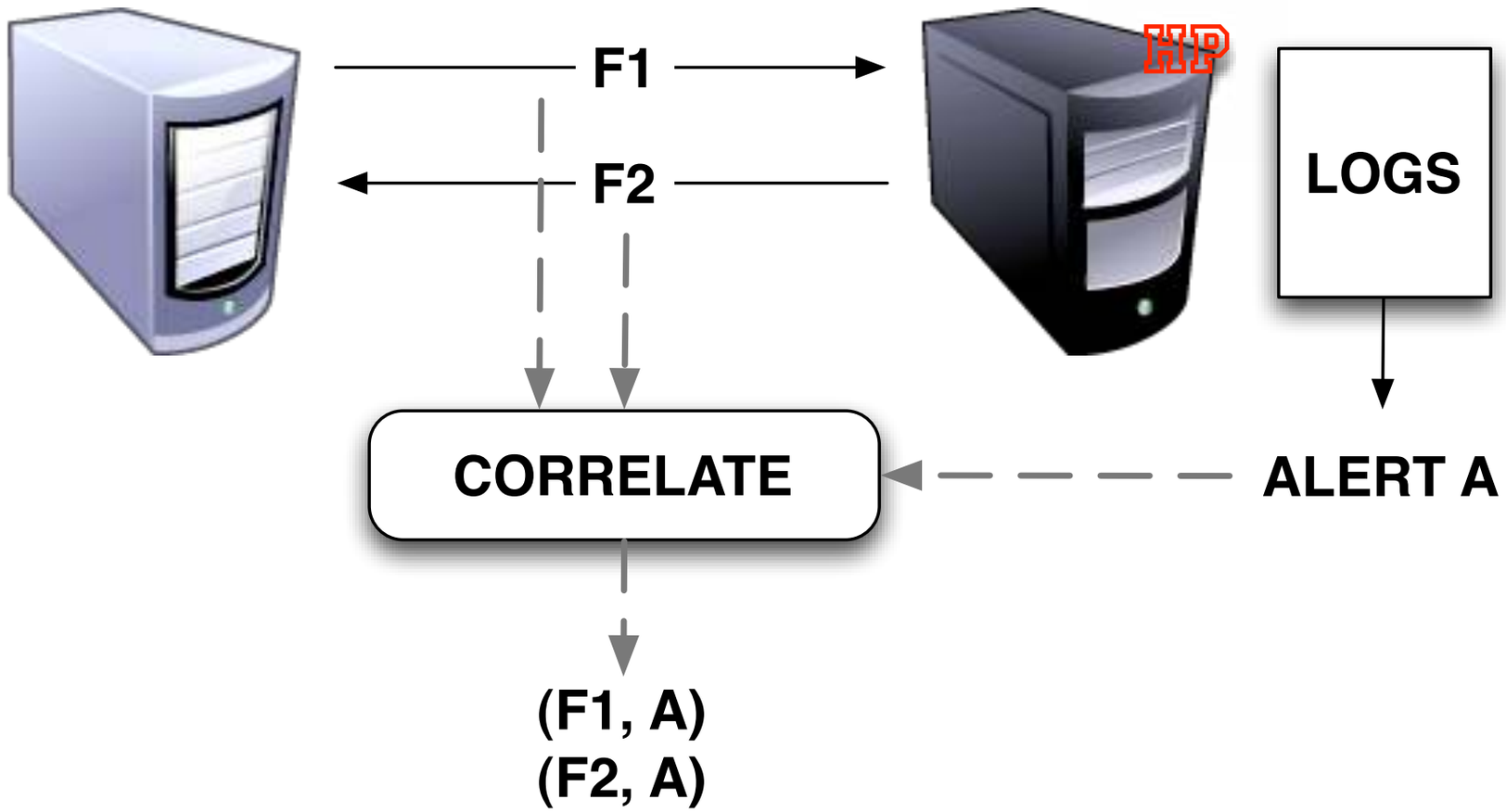
Data set creation



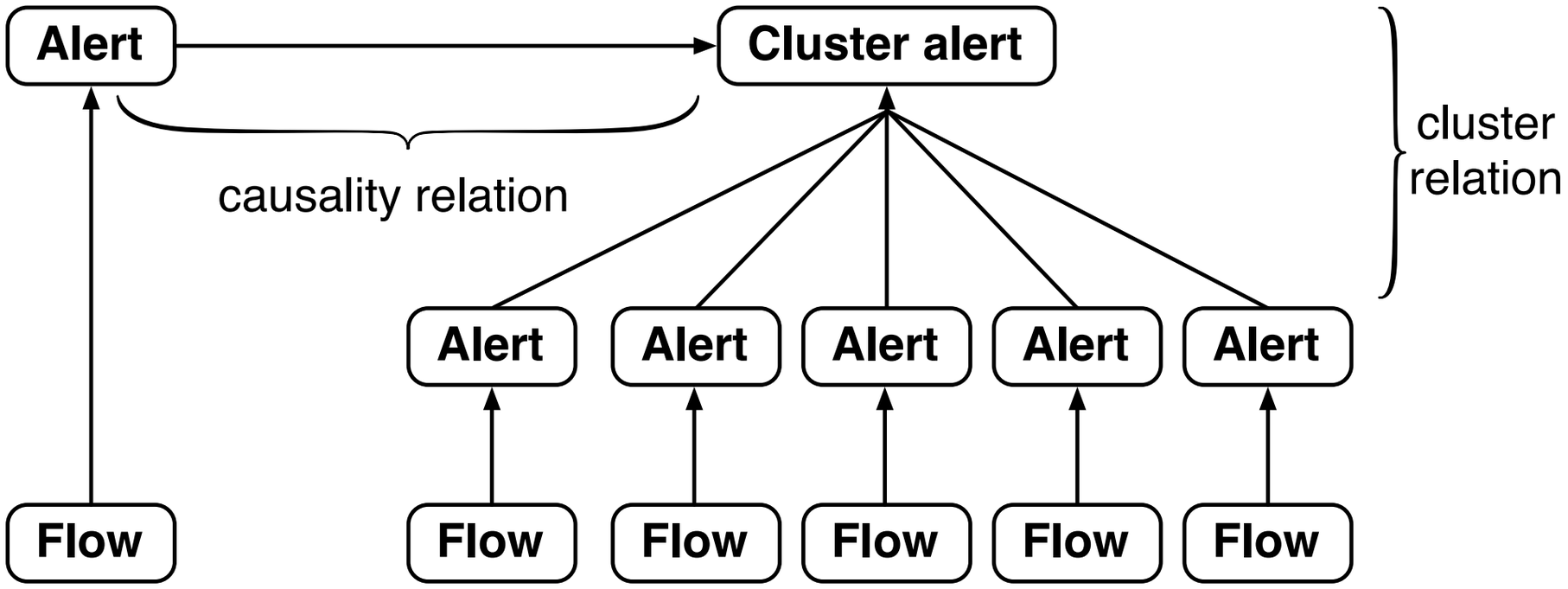
- The correlation process will results in *alerts*

$$A = (T, Descr, Auto, Succ, Serv, Type)$$

Correlation procedure



Cluster and Causality



- Hierarchic view of the alerts to enrich the data set with extra information on the traffic
- Group simple alerts into cluster alerts
  - high level view of malicious activities

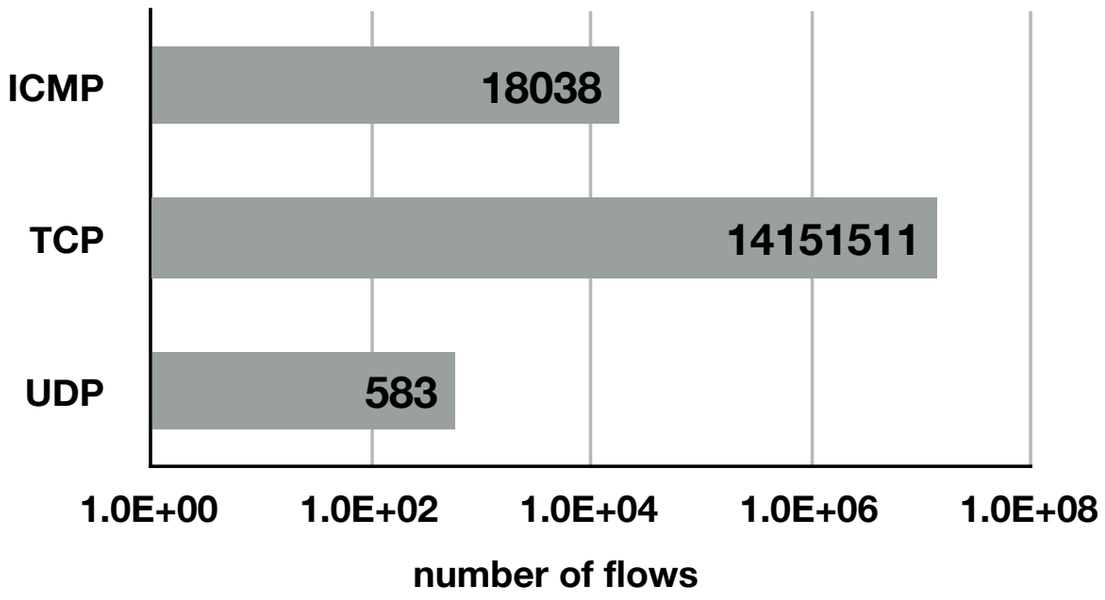
Implementation

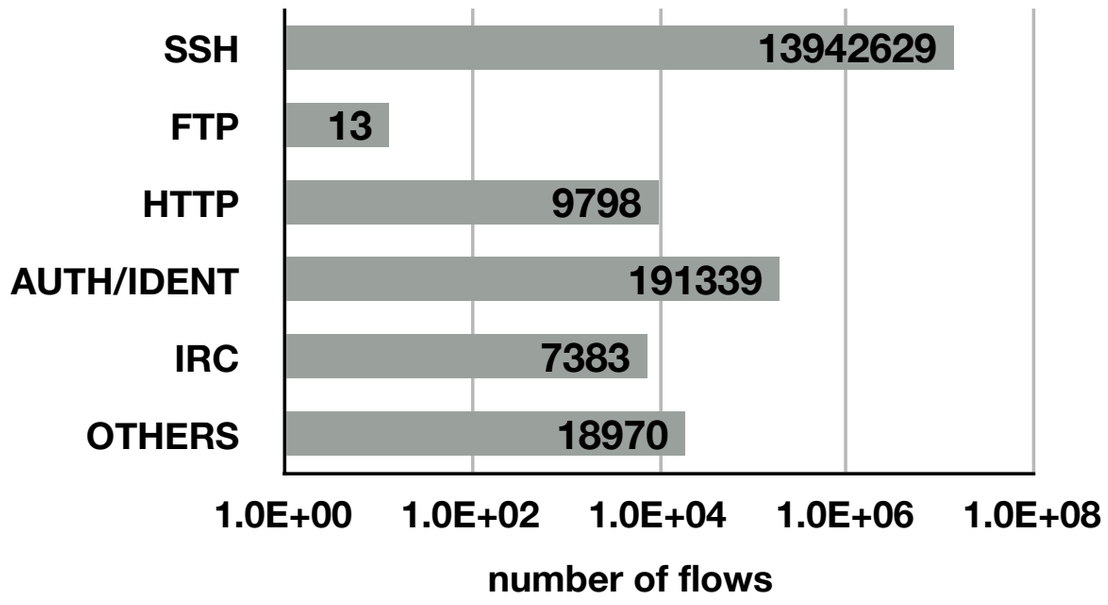
Packets to flows	AUTOMATIC	• softflowd
Logs to log events	SEMI-AUTOMATIC MANUAL	• shell scripts • discriminate between manual/ automated attacks
Alert correlation	SEMI-AUTOMATIC	• correlation procedure • extensible for other attacks
Cluster and causality	MANUAL	• analysis of typescripts

The Dataset

dump file	24 GB
flows	14M
alerts	7.6M

- Flow breakdown

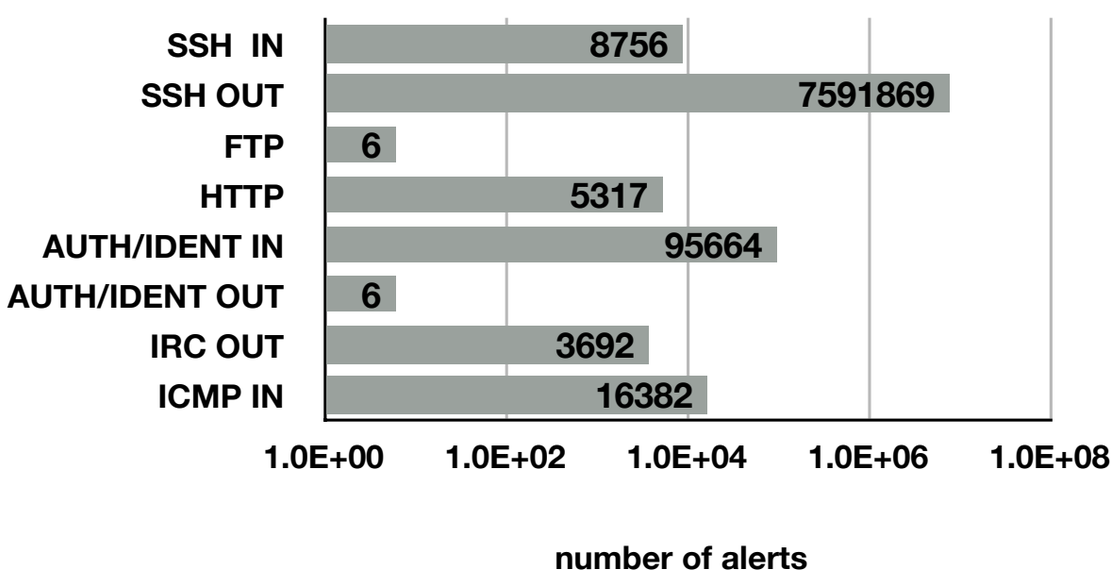


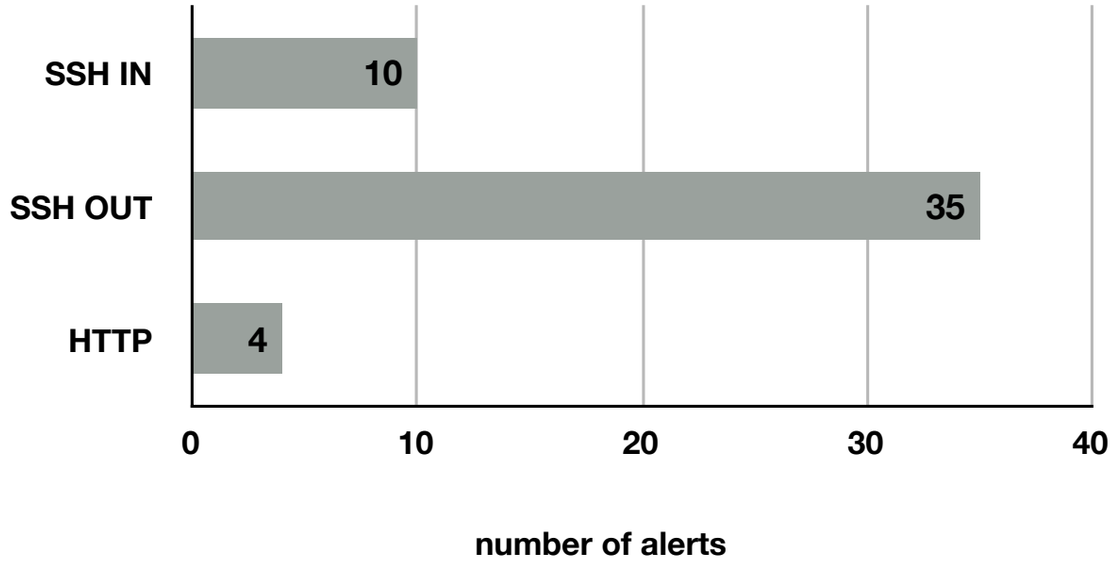


The Dataset

dump file	24 GB
flows	14M
alerts	7.6M

- Alert breakdown





The Dataset

- We labeled: 98,5% flows and 99,99% alerts
- Mainly malicious traffic:
  - ssh brute force attacks
  - automated http connections
- Small percentage of *side-effect traffic*
  - *auth/ident* on port 113
  - IRC traffic

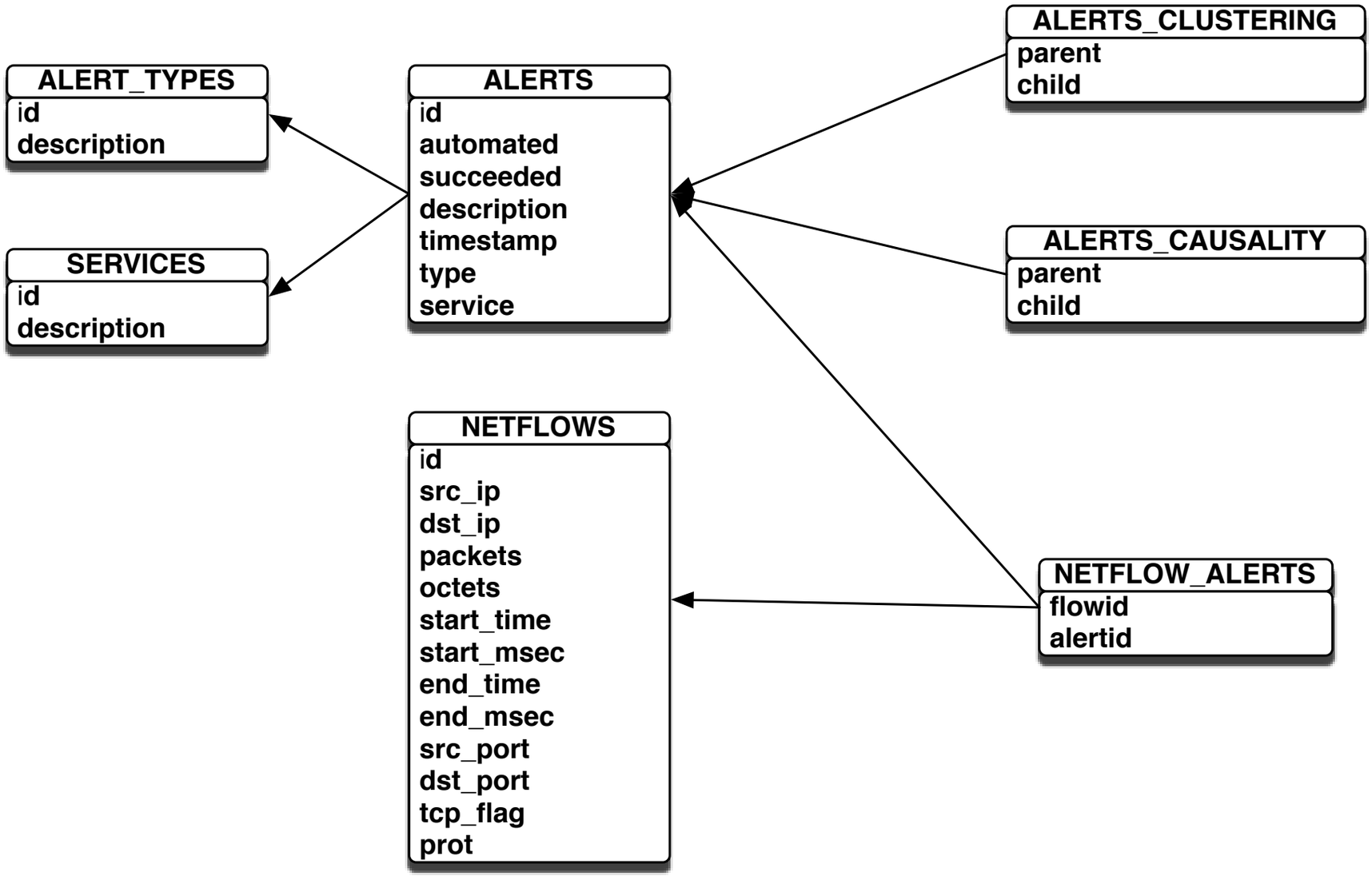
Conclusions

- We presented the first labeled data set for flow-based intrusion detection
  - http://traces.simpleweb.org/
  - Semi-automated correlation process
  - manual intervention is still needed
- Data set mainly constituted of malicious traffic
  - need to extend to benign traffic

Conclusions

- Reactions:
  - Since publication (October 2009) ~ 7 requests
  - We do not monitor the downloads at the webpage
  - In contact with Philipp Winter (Hagenberg University, AU): MSc Project "*Inductive Intrusion Detection in Flow-Based Network Data using One-Class Support Vector Machines*"

Implementation



Correlation procedure

Algorithm 1 Correlation procedure

1: **procedure** ProcessFlowsForService (s : service)
2: **for all** Incoming flows F_1 for the service s **do**
3: Retrieve matching response Flow F_2 such as
4: $F_2.I_{src} = F_1.I_{dst} \wedge F_2.I_{dst} = F_1.I_{src} \wedge F_2.P_{src} = F_1.P_{dst} \wedge F_2.P_{dst} = F_1.P_{src}$ $\wedge$
5: $F_1.T_{start} \leq F_2.T_{start} \leq F_1.T_{start} + \delta$
6: with smallest $F_2.T_{start} - F_1.T_{start}$;
7: Retrieve a matching log event L such as
8: $L.I_{src} = F_1.I_{src} \wedge L.I_{dst} = F_1.I_{dst} \wedge L.P_{src} = F_1.P_{dst} \wedge L.P_{dst} = F_1.P_{src} \wedge$
9: $F_1.T_{start} \leq L.T \leq F_1.T_{end} \wedge$ **not** $L.Corr$
10: with smallest $L.T - F_1.T_{start}$;
11: **if** L exists **then**
12: Create alert $A = (L.T, L.Descr, L.Auto, L.Succ, s, \texttt{CONN})$.
13: Correlate F_1 to A ;
14: **if** F_2 exists **then**
15: Correlate F_2 to A ; $L.Corr \leftarrow$ **true** ;
16: **end if**
17: **end if**
18: **end for**