Real-time Pattern Detection in IP Flow Data using Apache Spark

International Symposium on Integrated Network Management (IM 2019)

May 9, 2019

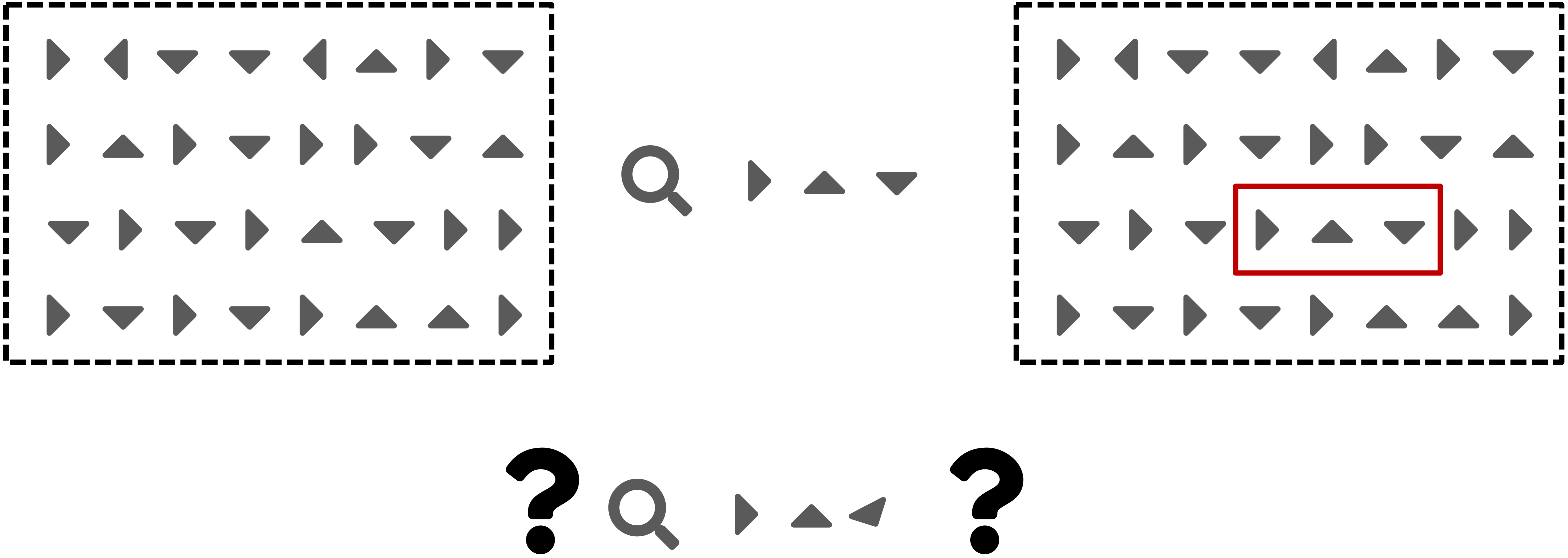
Milan Cermak, Martin Lastovicka, **Tomas Jirsik**

Institute of Computer Science, Masaryk University, Brno

CSIRT-MU

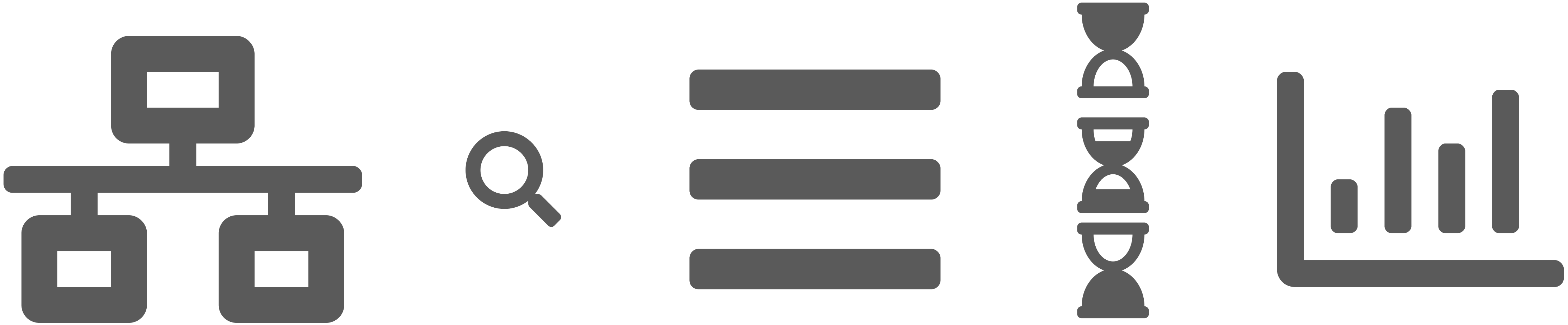
Attack Detection in Network Flow Records

challenges that everyone has to deal with



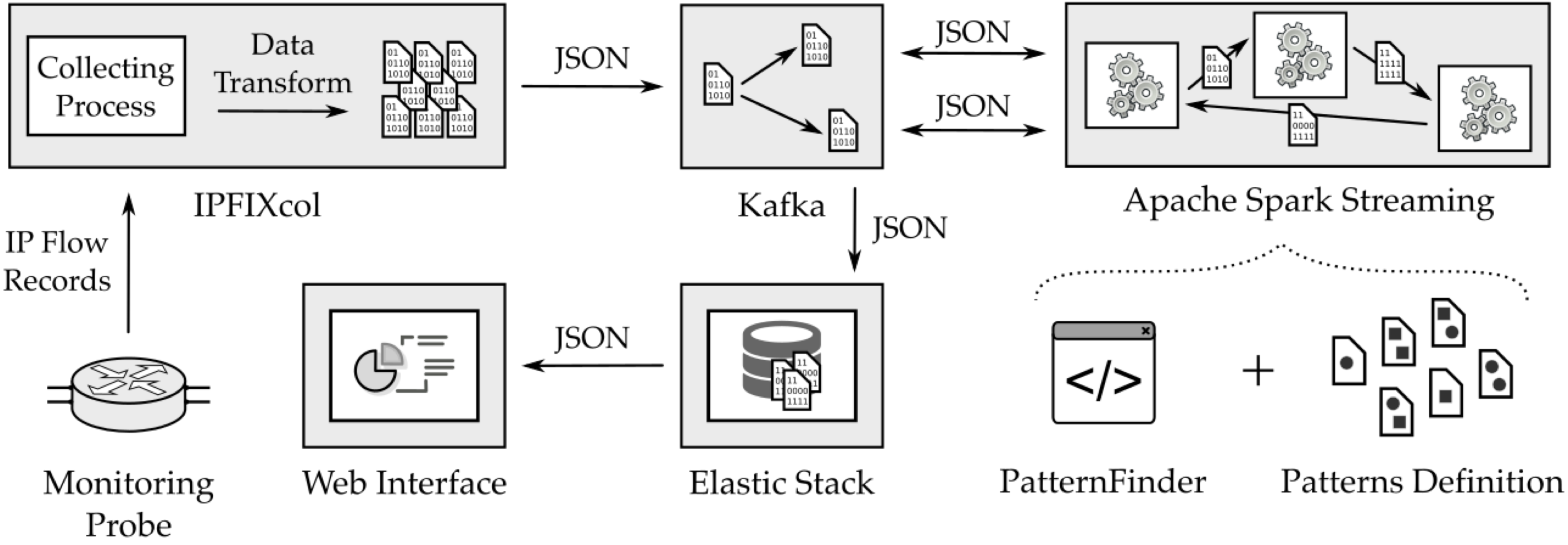
Attack Detection in Network Flow Records

challenges that everyone has to deal with II.



Stream4Flow: Real Time Analysis

distributed data stream processing framework

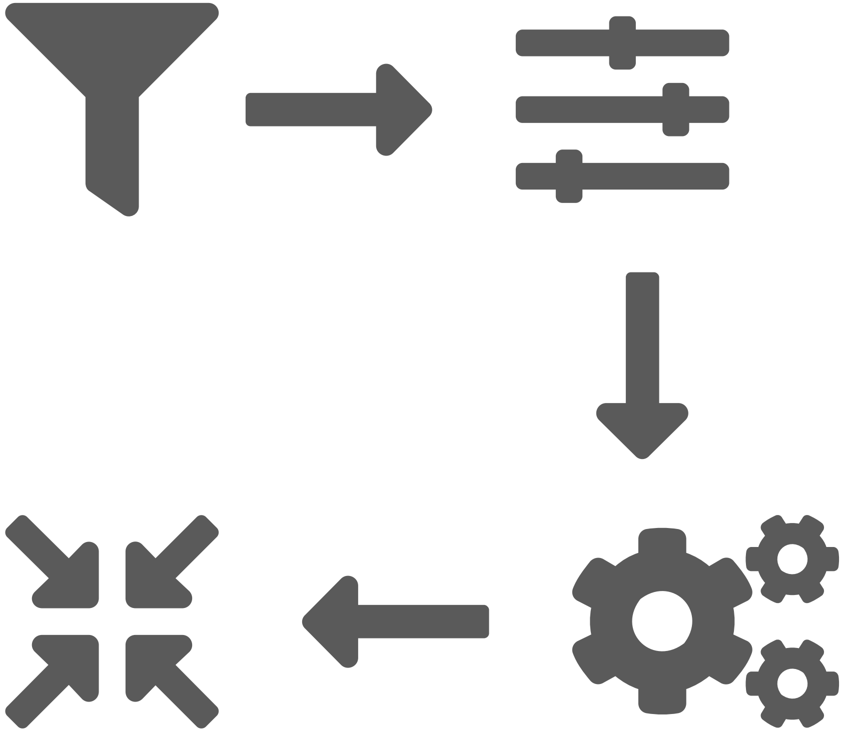


IM 2019: Real-time Pattern Detection in IP Flow Data using Apache Spark

Milan Cermak et al., Institute of Computer Science, Masaryk University, Brno

PatternFinder

taking advantage of similarity search



```
distance_function:
       biflow_quadratic_form

patterns:
     - name: anomaly
       request: [23, 8983, 9098]
       response: [24, 1125, 9101]

distribution:
     anomaly:
       intervals: [0, 3, 5, 6, 7, 11]
       weights: [3, 2, 1, 1, 2, 3]
```

Pattern Definition

discovery of general attack patterns

Dataset

- Only network traffic of interest
- Include attack variations
- Creation
  - Real-world dataset 👎
  - Artificial dataset 👍

Pattern

- Easy to determine from dataset
- Statistical aggregations of attack characteristics

SSH Authentication Attack Use-case

from theory to real-world

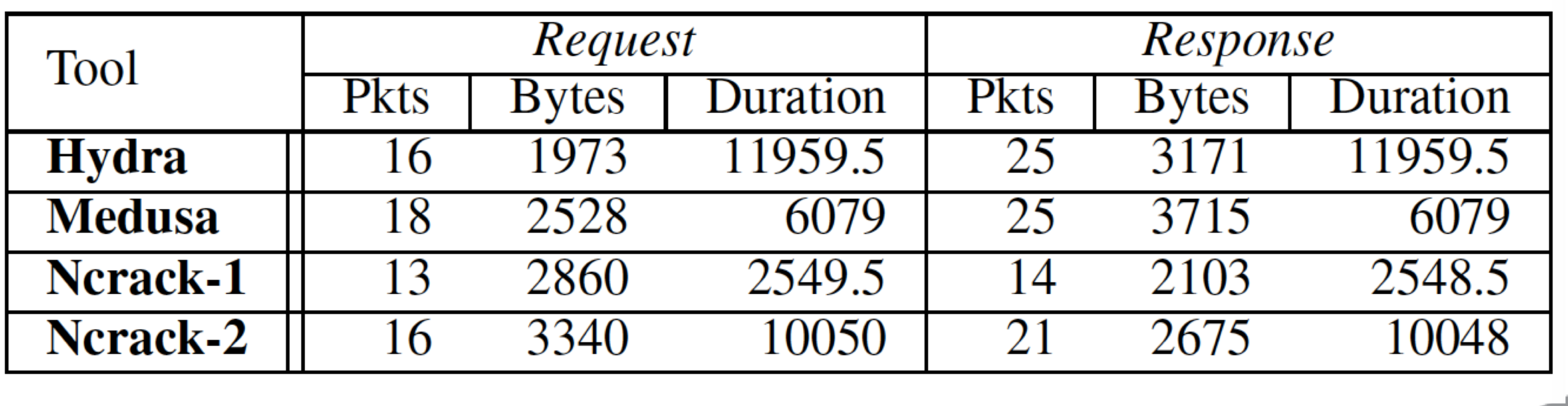
Pattern Definition

Hydra, Medusa, or Ncrack?

Dataset Creation

- Virtual environment – attacker and server
- 3 tools, 5 different settings

Derived Patterns – median aggregation



Tool	*Request*			*Response*		
	Pkts	Bytes	Duration	Pkts	Bytes	Duration
Hydra	16	1973	11959.5	25	3171	11959.5
Medusa	18	2528	6079	25	3715	6079
Ncrack-1	13	2860	2549.5	14	2103	2548.5
Ncrack-2	16	3340	10050	21	2675	10048

Evaluation

comparison with others

Measurement

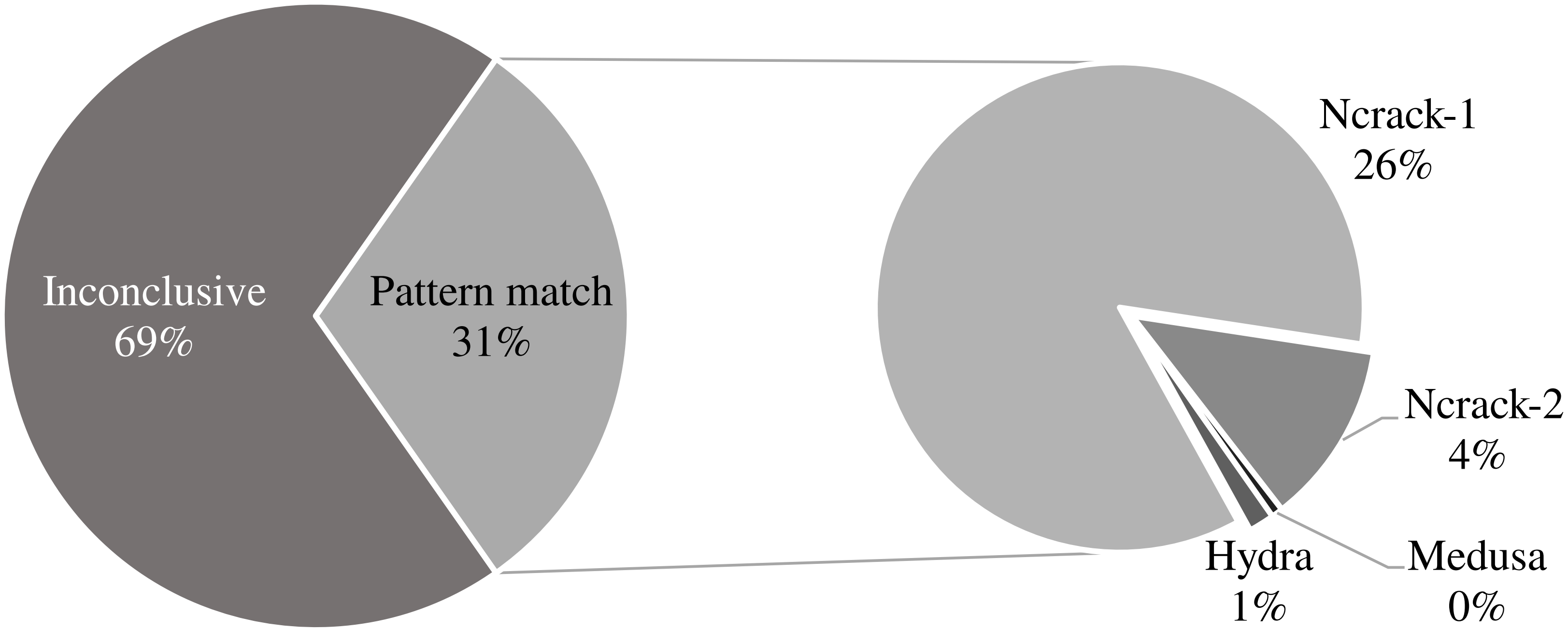
- one week period
- 478.98 M Flows, 5.54k Flows/second, 9.9k Flows/second in peak
- 21.91 TB data processed

Comparison

- Commercial solution Flowmon Anomaly Detection System
  - More than 30 login attempts in 5 min is an attack
- ADS **264** events from **75 IPs** vs PatternFinder **78** events from **42 IPs**
  - ADS overlapping events
- Accuracy 39%, precision 82%, recall 43%

Further Results

additional findings worth mentioning



Thank you for your attention

https://stream4flow.ics.muni.cz/

@csirtmu

Milan Cermak et al.

cermak@ics.muni.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

MUNI
ICS

CSIRT-MU