Towards botnet detection: What botnet characteristics can be detected in real-life network environments by using flow data?

Research Center Cyber Defence
Universität der Bundeswehr München

Christian Dietz
christian.dietz[at]unibw.de

NMRG Workshop (Prag): 24.05.2015

UNIVERSITY OF TWENTE.

DACS
Design and Analysis of Communication Systems

Introduction

"Cyber crime costs global economy $445 billion a year!

> **…** About 40 million people in the United States, roughly 15 percent of the population, has had personal information stolen by hackers, it said, while high-profile breaches affected 54 million people in Turkey, 16 million in Germany and more than 20 million in China.**"** **…**
>
> ---
> Source: Reuters, London, Mon Jun 9, 2014

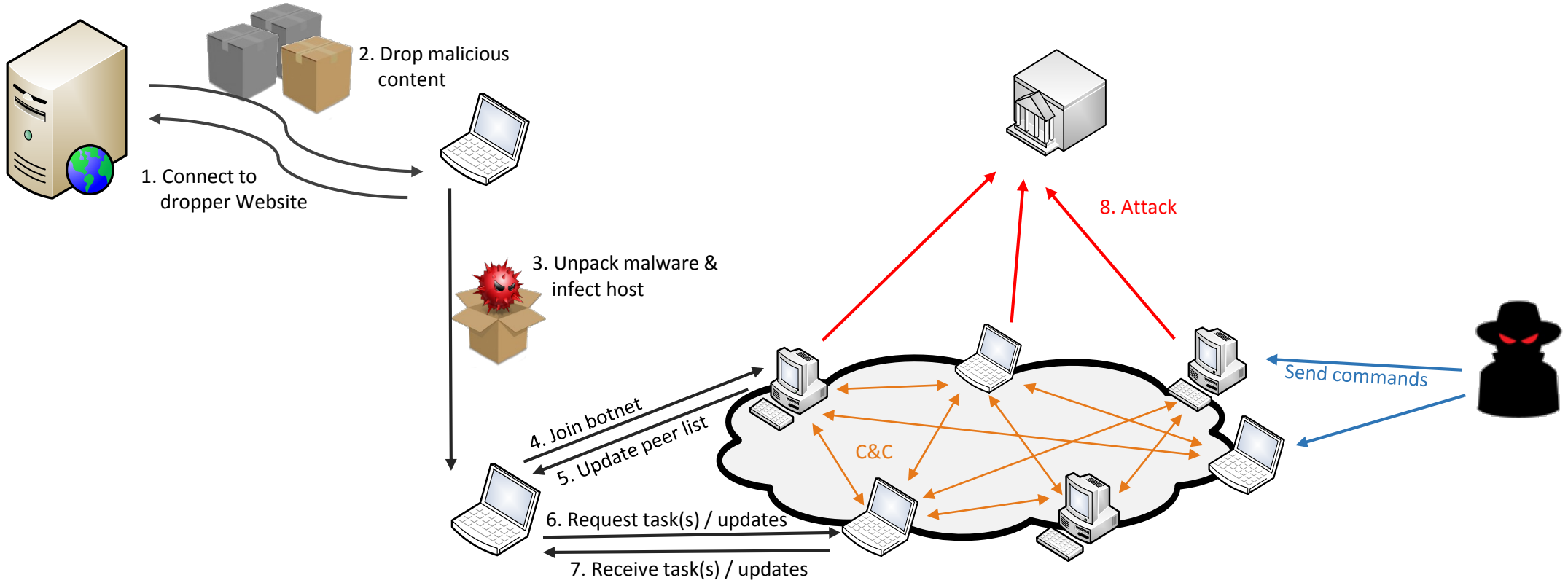
- Structure
  - Approach
  - Early results
  - Outlook

Introduction

Botnets:

- Provide infrastructure for various cyber criminal activities e.g. SPAM, DDoS, financial fraud, data theft, extortion
- A botnet is network of malware infected hosts under the command of a botmaster.
- Command and control infrastructure (C&C): IRC, HTTP, P2P

Botnet life-cycle



Challenges:


Malware updates


P2P


Fast-Flux


Encryption


Big Data


Anonymization (TOR, …)

Icon source: *https://www.iconfinder.com/*

Requirements

- Adaptability
- Accuracy
- Efficiency
- Scalability
- Privacy-preserving
- Universal applicability
- Traceability
- …

State of the art

Knowledge based:

- Rules
- Signatures
- Filters
- Experts

Drawbacks:

- Only detect known threats
- Amount rules necessary

Anomaly based:

- Machine learning
- Training/Validation data
- Automatically adapted to changed conditions

Drawbacks:

- False positives
- Availability of data
- Traceability of the detection process

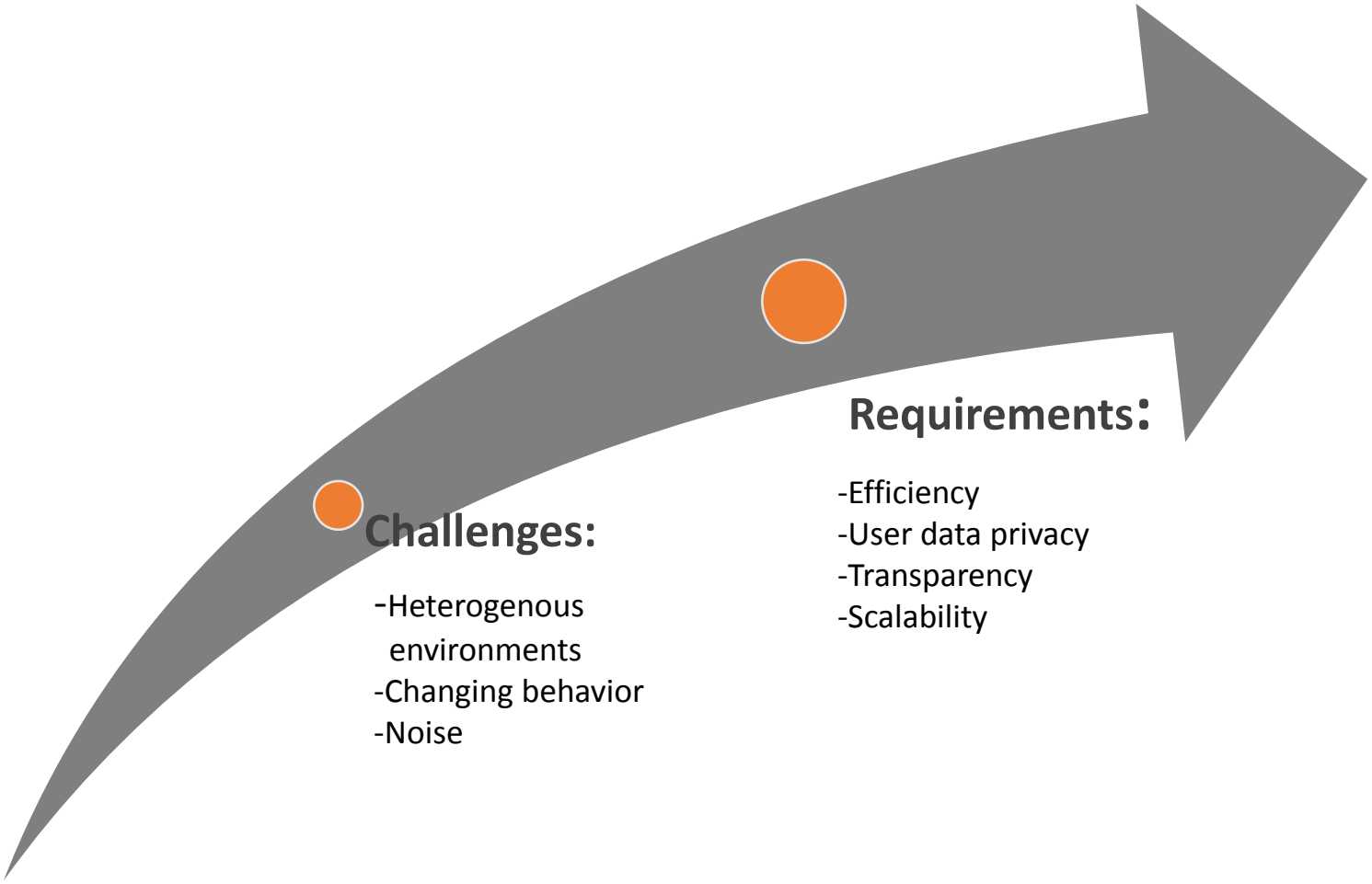
Approach

- What is limiting the success of existing detection approaches?
  - Privacy concerns?
  - Amount of data?
  - Availability of data?
  - Heterogenous behavior and environments!

    -> Standardized formats do not guarantee standardized behavior/noise description!

    -> Exchange needs standardized formats and protocols plus normalized/standardized behavior descriptors!

Approach



Challenges:

-Heterogenous environments
-Changing behavior
-Noise

Requirements:

-Efficiency
-User data privacy
-Transparency
-Scalability

Hybrid multi-level detection approach:

- Filters
- Machine learning
- Behavior + Signatures
- Pseudonyms

Approach

- Research questions:

How does botnet query behavior of look like in real life network environments?

How can the observations enhance/complent current detection approaches?

State of the art

- Definition flow as used in this research:

  Set of records observed over a certain period of time sharing connection information plus a set of common properties derived from the data contained in the records captured at a network observation point.

- Examples:
  - Sinkhole trace:

    [1] "[2012-09-09 07:01:28.64365] bootstrap request from 81.214.XXX.XXX:1064"

    [2] "[2012-09-09 07:01:29.17498] bootstrap request from 81.214.XXX.XXX:1067"

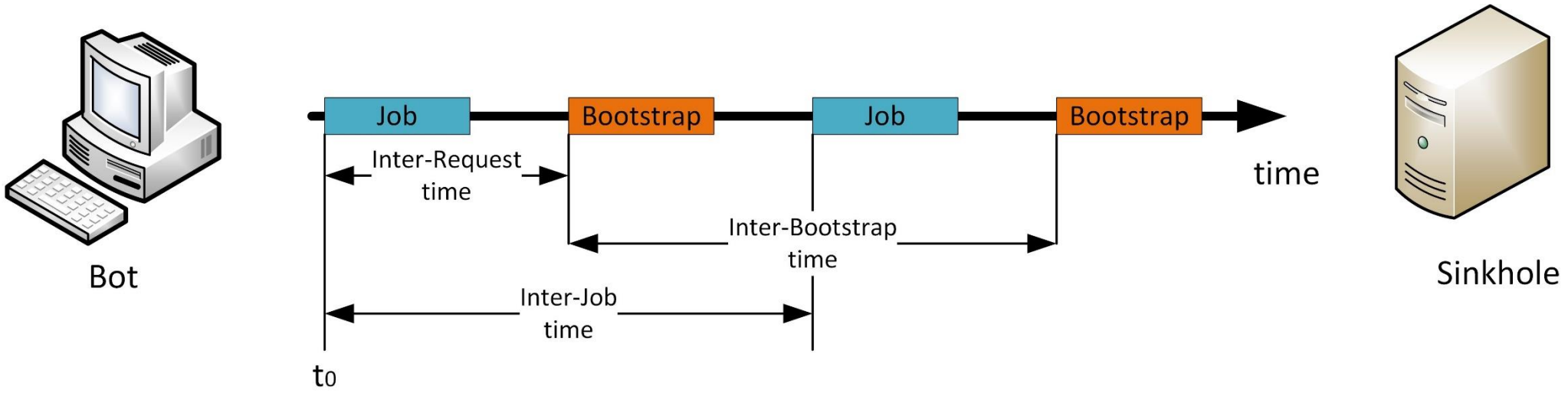
    [3] "[2012-09-09 07:01:29.37554] job request from 89.29.XXX.XXX:3265 - 0c274f674d8347509234a088d359df49, v126 \"relqq26\", os info: 5.1.2600, platform 2)"

  - Netflow:

    [1] 2012-09-09 07:01:28.64365, 2012-09-09 07:01:48.44789, 1.1.1.1, 8.8.8.8, 1234, 80, 10, 984, 0, 0, .A...., UDP, ...

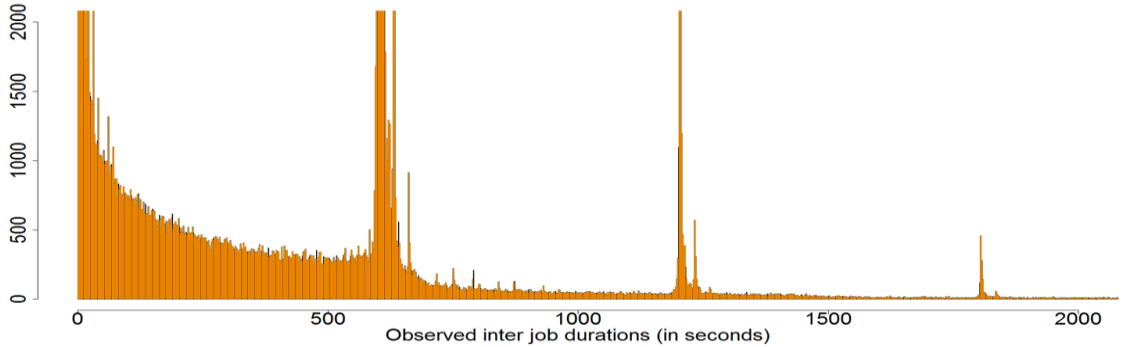
Early results

- What we measured:



Idea: From traffic to behavior descriptor

- Exchange of normalized (temporal) behavior descriptor
- Allows:
  - Efficient search for similar behavior in big data sets
  - If binarized, efficient algorithms could be used (K-nearest neighbor, Bloomfilter…)



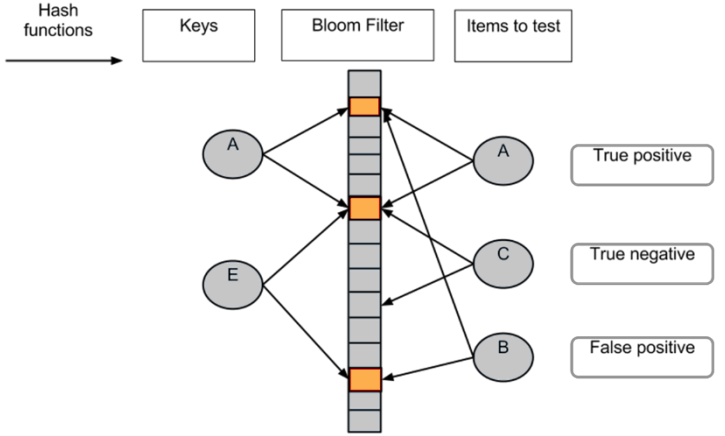
Noisy binarized (temporal) behavior signature

Normalized binarized (temporal) behavior signature

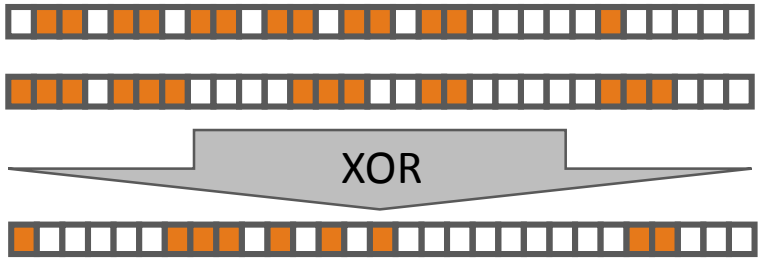
Outlook - Concept

Behavior signature database

Filter:



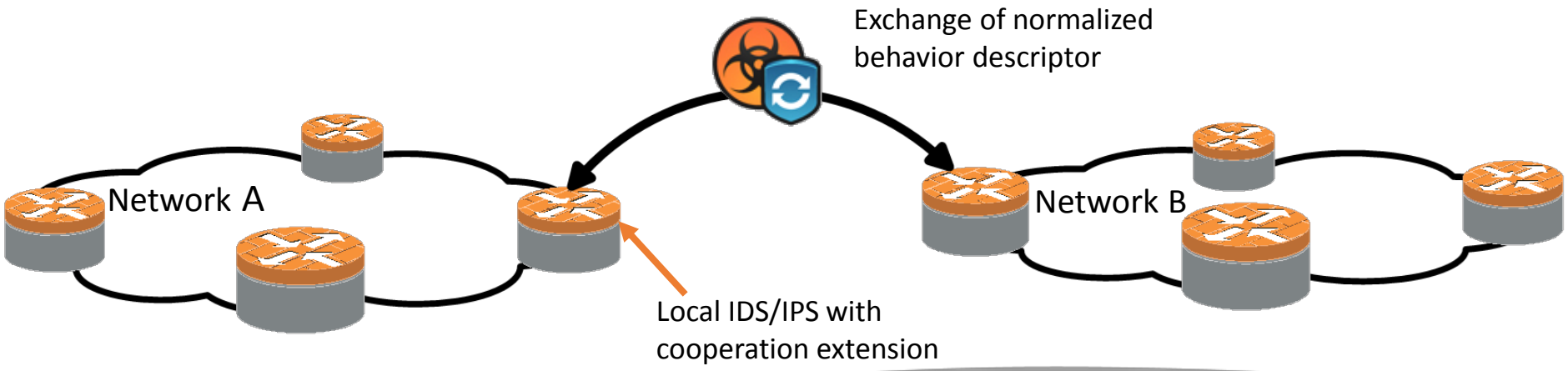
Matching (e.g. distance-based):



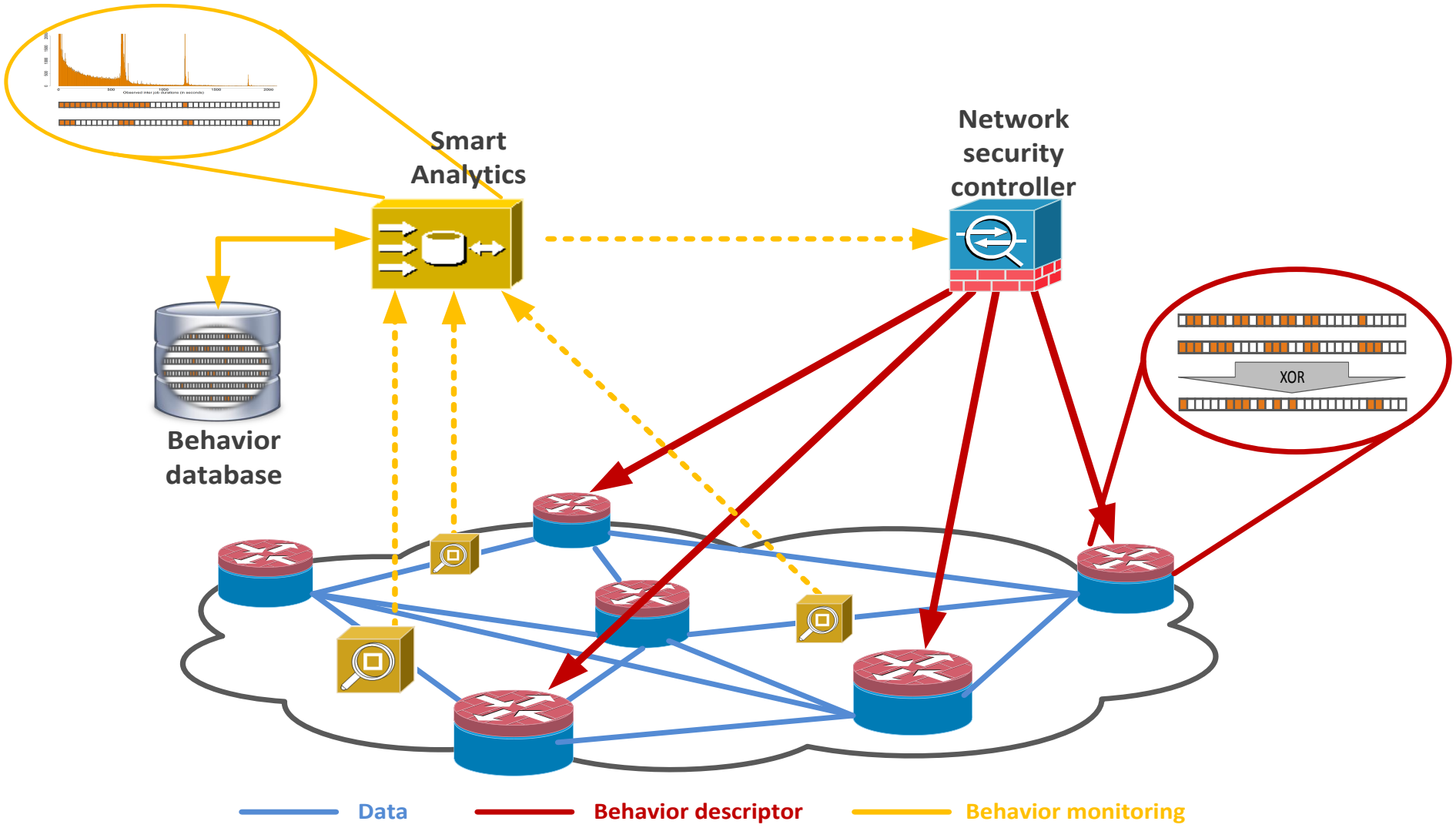
Hamming distance = 10

Use case: (1) Cooperative network operator

- Exchange of normalized behavior descriptor
- Benefit:
  - Efficient search for similar behavior in big data sets of heterogenous network environments
  - If binarized, efficient algorithms could be used (K-nearest neighbor, Bloomfilter…)



Use case: (2) Single (SDN) network operator



- Data?
- Use cases?
- Ideas?
- …?