Cyber Security Datasets for Machine Learning

Hayretdin Bahsi

Centre for Digital Forensics and Cyber Security
Tallinn University of Technology

Cyber Security and Machine Learning

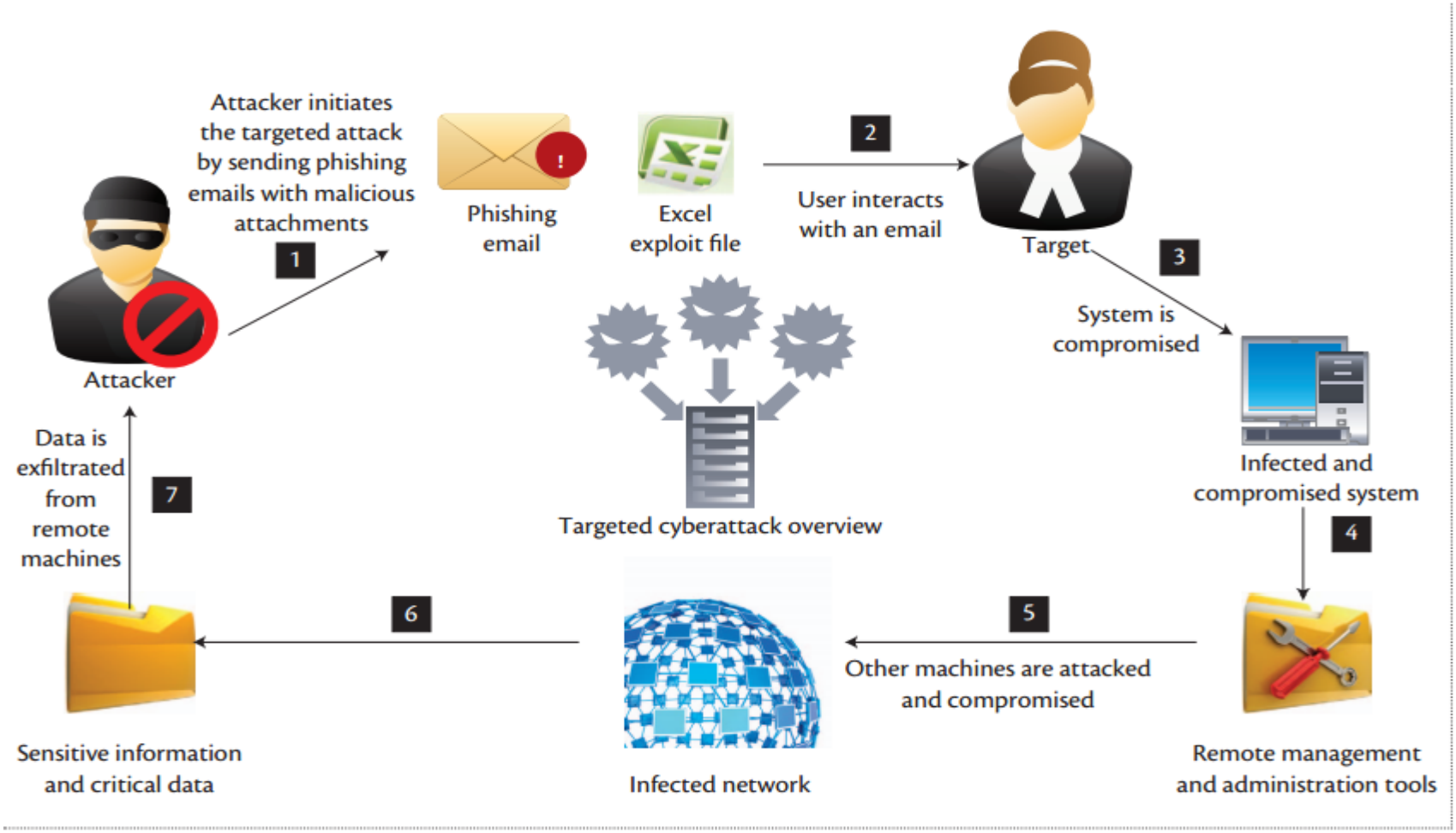
- Main cyber security areas that benefit from machine learning
  - Intrusion detection
    - Detection of web attacks
    - Botnet detection
    - Detection of SCADA attacks
  - Malware scanning
  - Phishing detection
  - Cyber threat intelligence

Intrusion Kill Chain

Reconnaissance	• Research, identification and selection of targets
Weaponization	• Coupling a remote access trojan with an exploit into a deliverable payload
Delivery	• Transmission of the weapon to the targeted environment
Exploitation	• Exploitation triggers intruders' code after the weapon is delivered to victim host
Installation	• Installation of a remote access trojan or backdoor
Commond and Control	• Establish a command and control channel to the attacker
Actions on Objectives	• Data exfiltration or moving inside the network

Hutchins, Eric M., Michael J. Cloppert, and Rohan M. Amin. "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains." *Leading Issues in Information Warfare & Security Research* 1 (2011): 80.

Highly Targeted Attack Scenario*



* Sood, Aditya K., and Richard J. Enbody. "Targeted cyberattacks: a superset of advanced persistent threats." *IEEE security & privacy* 1 (2013): 54-61.

What is Intrusion Detection System?

- Monitors network traffic or host data
- Looks for potential intrusions
  - Detects signatures
  - Detects anomaly
- Not a protection method
- Detection system
- Intrusion Prevention System

IDS Types


By Location
Network-based IDS
Host-based IDS
By Detection Method
Signature-based IDS
Anomaly-based IDS

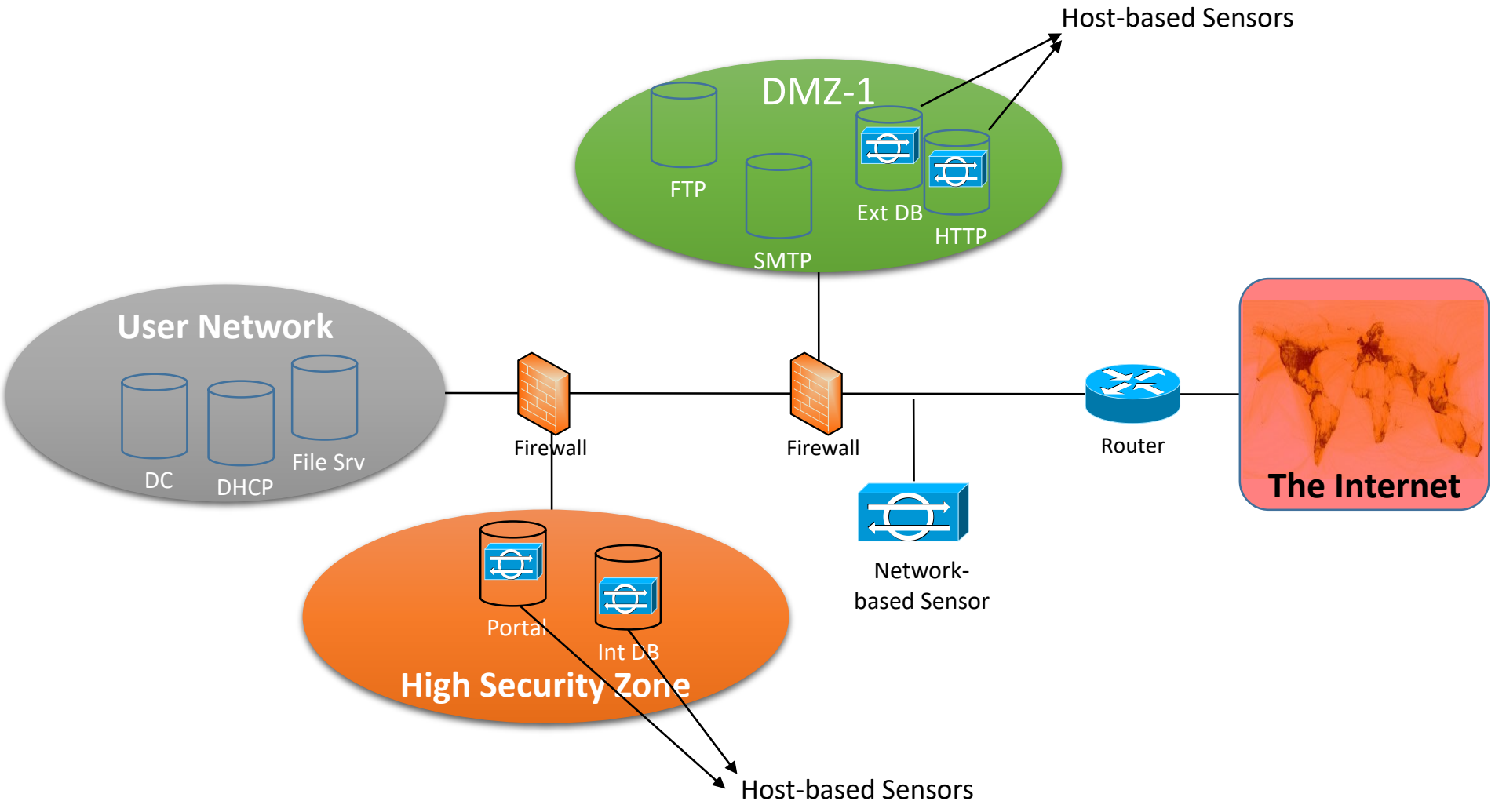
Network based IDS

- Monitors traffic like IDS
- Examines network packets
- Monitors packets with sensors
- Sensor sniffs traffic between network segments
- Promiscuous mode:
  - Sensor does not interfere with traffic
  - The sniffing NIC attached to a hub or a switch with mirroring feature can be used
  - The sniffing NIC does not need an IP address
- Sensor sends intrusion information to a central console
- Central console is used to manage sensors and to view alarms

Host based IDS

- Sensors are installed on the hosts
- Can monitor only the host installed on
- Monitors OS logs
  - File system modifications
  - File accesses
  - Program running numbers
- Monitors application logs
  - Syslog
  - Database logs
  - Web server logs
- Monitors network packets coming to host

Possible Network Topology



KDD Cup 1999 Dataset

- MIT Lincoln Labs under DARPA and Air Force Research Laboratory Sponsorhip
- KDD Cup 1999
- Benchmark data for intrusion detection systems
- Network- and host-based data
- http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

KDD Cup 1999 Dataset (2)

- Main attack categories:
  - Denial-of-Service (DoS), user-to-root (U2R), Remote to Local Attack (R2L) and Probing Attack
- 21 specific attack categories:
  - Buffer_overflow, guess_passwd, portsweep,etc.
  - Each specific attack is mapped to a main attack category
- Test and train data are not from the same probability distribution
- New attack types in test data (which are not included in the training data)
- 42 attributes, 39 → continous, 3 → binary (0 or 1)

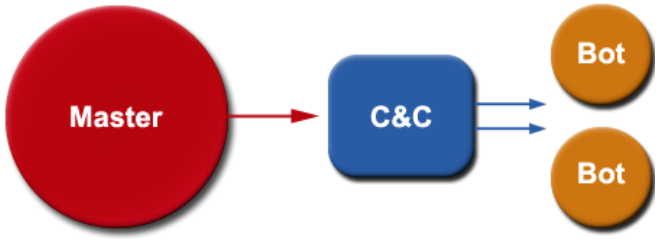
Botnets

- Botnet = Robot Network
- Network consists of bots
  - Semi-autonomous agent
  - Under control of attacker
  - Applies the commands received from remote controller
- Bot=Malware + Remote Control + Communication Channel

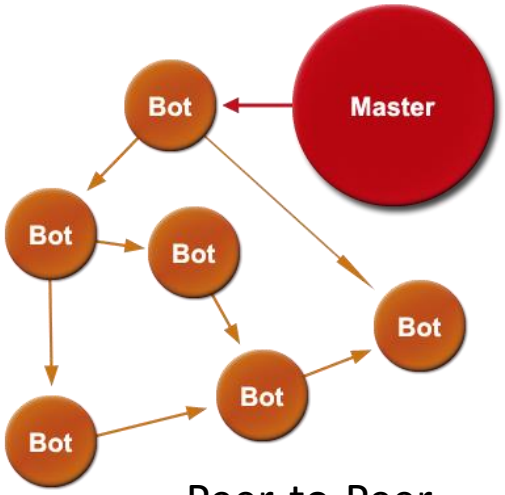
Botnets (2)

- DDoS attacks
- Click fraud
- Sending spam
- Distribute malware
- Phishing attacks

Botnets (3)



Client-Server Model



Peer-to-Peer

- Internet Relay Chat (IRC)-based
- HTTP-based
- DNS-based
- Peer to peer (P2P)

Botnet Detection

- Network flow information
- Data from French chapter of the honeynet project
  - Storm (non-P2P botnet)
  - Waledac botnet (P2P botnet)
- Non-malicious traffic from
  - Ericsson Research in Hungary
  - Lawrence Berkeley National Lab

Botnet Detection (2)

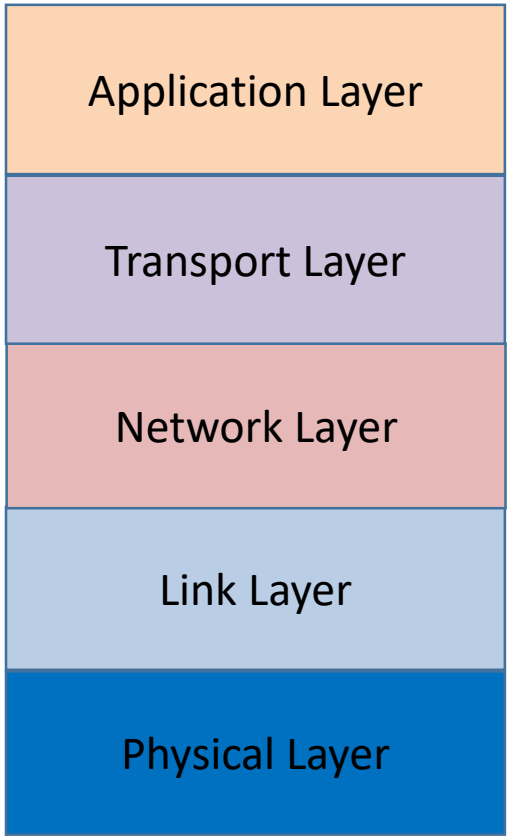
- Features
  - Flow source/destination IP addresses
  - Source/destination ports
  - Protocol type
  - Average packet length per flow
  - Total number of packets per flow etc.
- Labels
  - Botnet C&C,
  - normal traffic
    - P2P – Skype, bittorent, etc.
    - Non-P2P – http, ftp, etc
- https://www.uvic.ca/engineering/ece/isot/datasets/

An Integrated Botnet Dataset

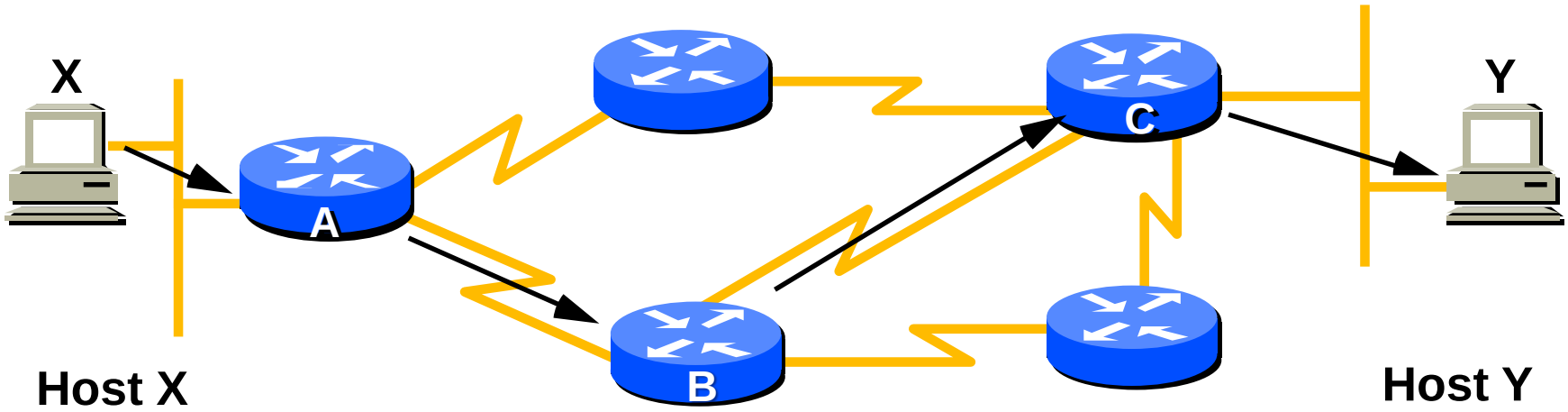
- Integration of different datasets into one set
- Representation of different bot types
- Data set content
  - 7 bot types in training data/ 16 bot types in test data
  - 43.92% of training data is malicious
  - 44.97% of test data is malicious
- http://www.unb.ca/cic/datasets/botnet.html

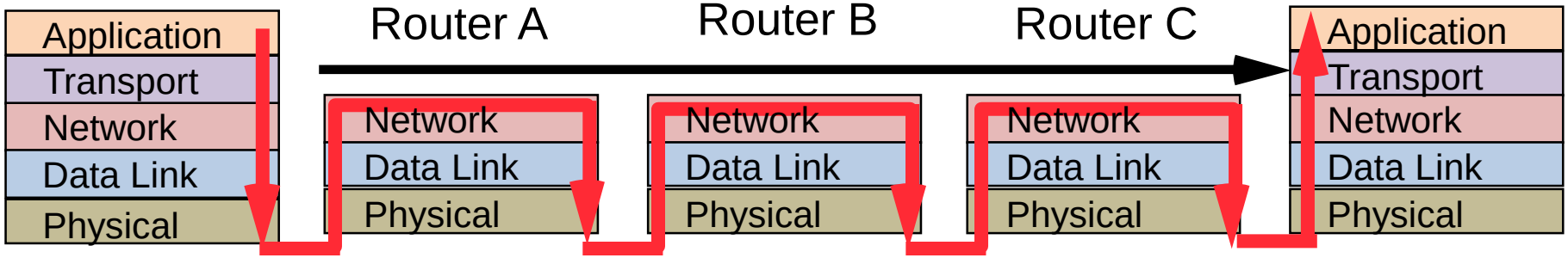
TCP/IP

- Application:
  - Network applications FTP, SMTP, HTTP
  - Interface to users
- Transport:
  - Data transmission between ports
  - TCP, UDP
- Network:
  - Routing of datagrams from source to destination
  - IP, routing protocols
- Link:
  - Data transfer between neighboring network elements
  - Ethernet, Wi-Fi, Bluetooth
- Physical: bits “on the wire”

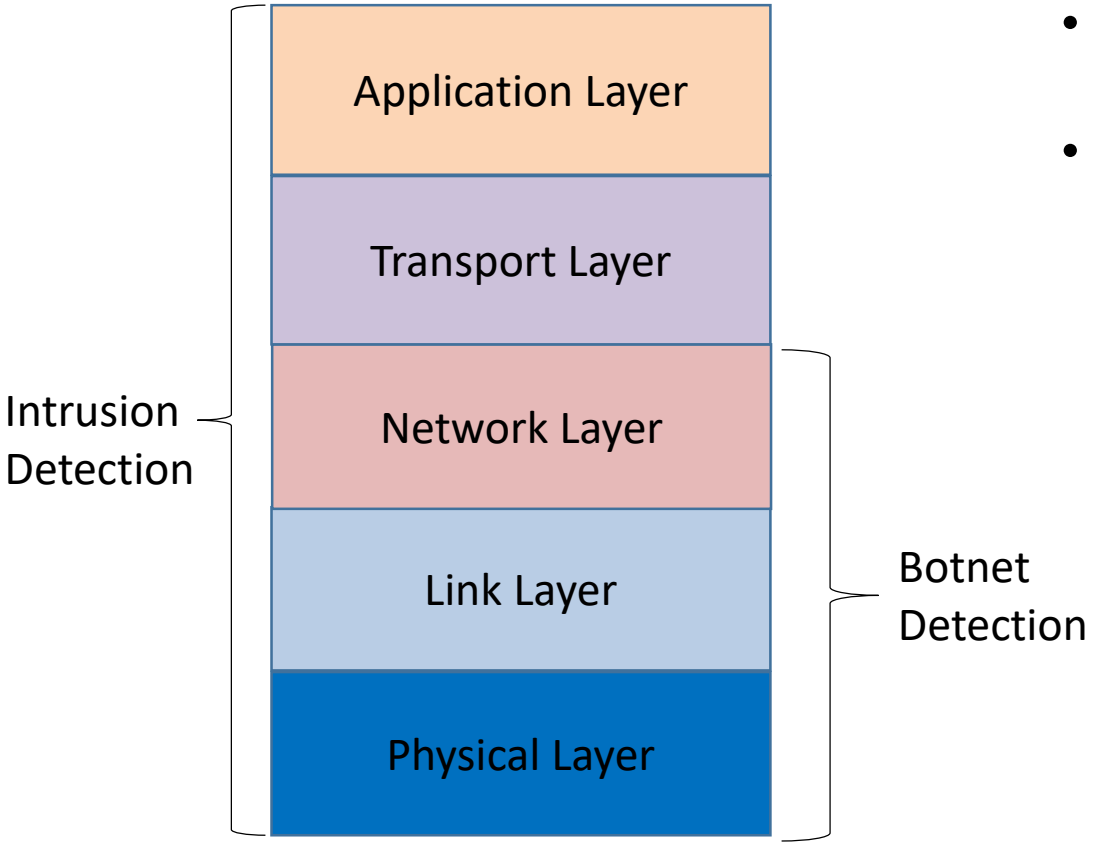


Wide Area Network Traffic





Intrusion Detection vs Botnet Detection

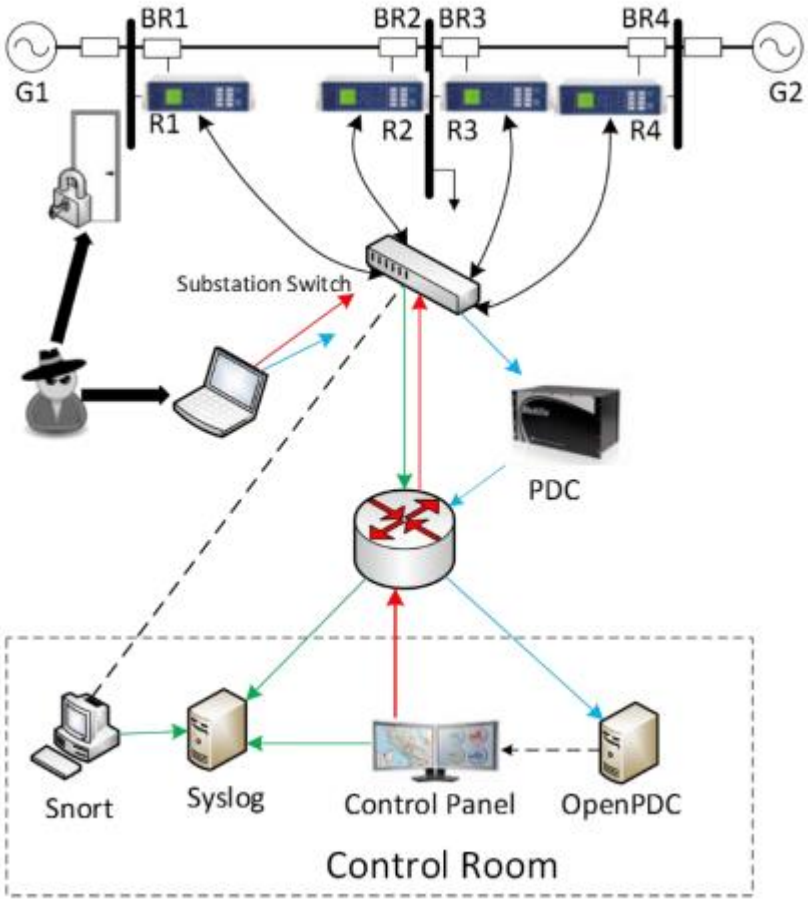


- Including all layers is better for detection
- However, it may not be optimal for large-scale implementations:
  - Requires extra computational power
  - Can not check the encrypted communication
  - Less privacy concern

Power System Attack Datasets

- Power system dataset
- 37 type scenarios
  - Natural events (8)
  - No events (1)
  - Attack events (28)
- Attacker is assumed to be within industrial netwerk
- Simulated attacks
  - Remote command injection
  - Relay setting change
  - Data injection

Power System Attack Datasets (2)

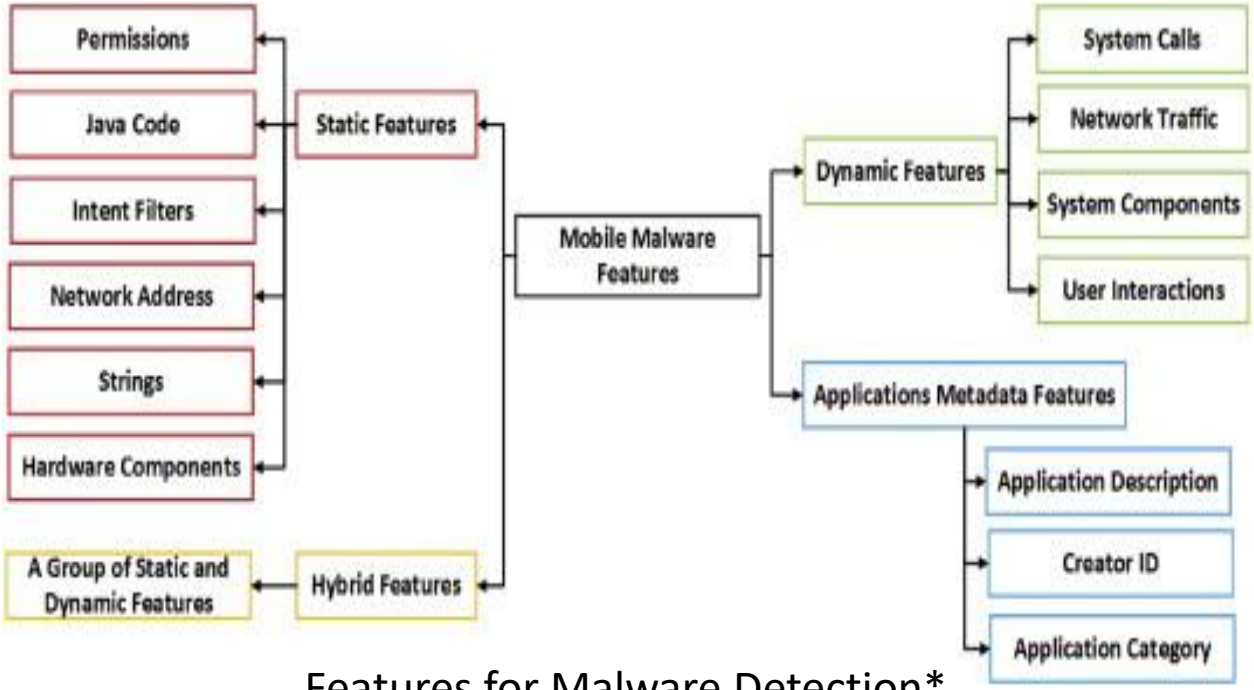


Power System Attack Datasets (3)

- Binary classification (attack vs other events)
- Three-class classification (attack, natural event, no event)
- Set of features (128)
  - 29 types of measurements per each phasor measurement unit (PMU) (19x4=116)
  - 12 columns for logs (control panel, snort and relay)
- https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets
- Additional dataset about gas pipeline system

Malware Detection

- Malware detection via static or dynamic features
- Many malware datasets



Features for Malware Detection*

* Feizollah, Ali, et al. "A review on feature selection in mobile malware detection." *Digital Investigation* 13 (2015): 22-37.

Links to Other Datasets

- http://www.unb.ca/cic/datasets/index.html
- http://www.azsecure-data.org/other-data.html

Overview

- Learning new attack types
- Lack of labeled data
- Performance considerations during training and testing
  - Application of feature selection
  - Consideration of system limits
- Attacks to machine learning algorithms
  - Attack-defence game