A Haystack full of Needles: Scalable Detection of IoT Devices in the Wild

Said Jawad Saidi, Anna Maria Mandalari, Roman Kolcun, Hamed Haddadi,
Daniel J. Dubois, David Choffnes, Georgios Smaragdakis, Anja Feldmann

max planck institut informatik

UNIVERSITÄT DES SAARLANDES

TU berlin

Northeastern University

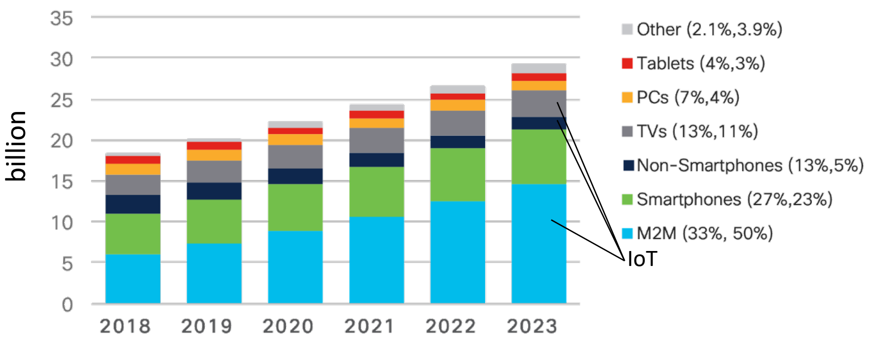
Imperial College London

erc
European Research Council
Established by the European Commission

EPSRC
Pioneering research and skills

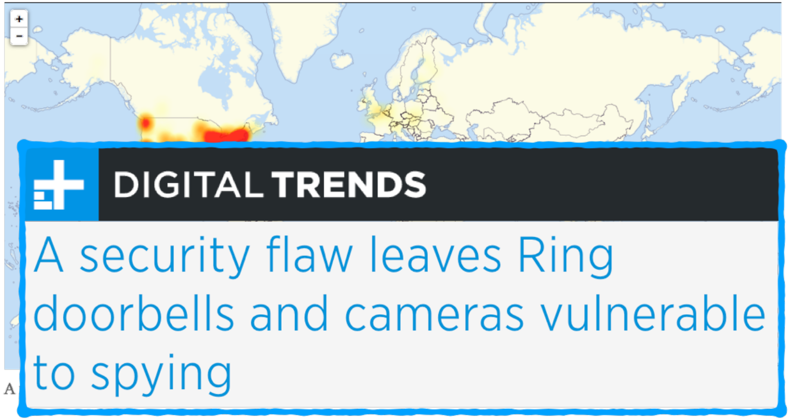
NSF

17+ billion IoT devices by 2023



Source: Cisco Annual Internet Report, 2018–2023

Hackers Used New Weapons to Disrupt Major Websites Across U.S.



A security flaw leaves Ring doorbells and cameras vulnerable to spying

Can we "*identify*" and "*locate*" IoT devices in our networks

We had a collaboration with a large European ISP

IoT device detection: Why at ISP?

- Security & privacy benefits for customers (opt-in)
  - notifying about infected devices *
- Security of ISP's network:
  - incident investigation & resolution

*Cetin et. al, NDSS'19, Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mirai

Detecting IoT Devices at the Provider is Challenging

- Traffic patterns across IoT devices are diverse

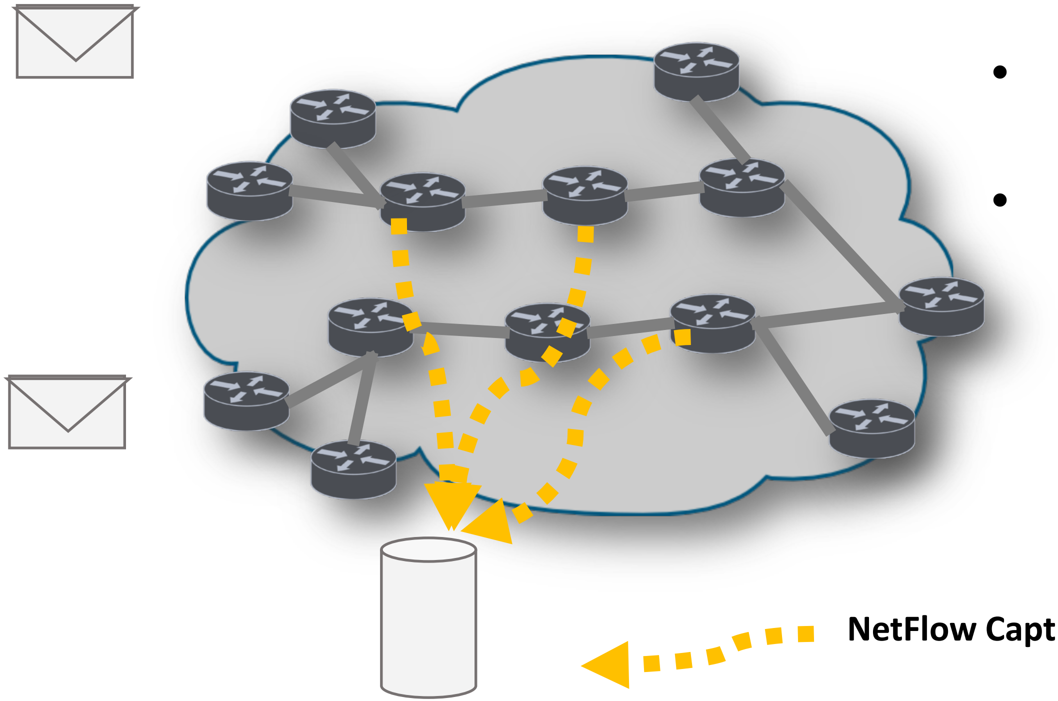
- Deploying an agent inside at each ISP customers is not scalable *

- Active measurements do not work with devices behind NAT

- Deep packet inspection raises privacy concerns

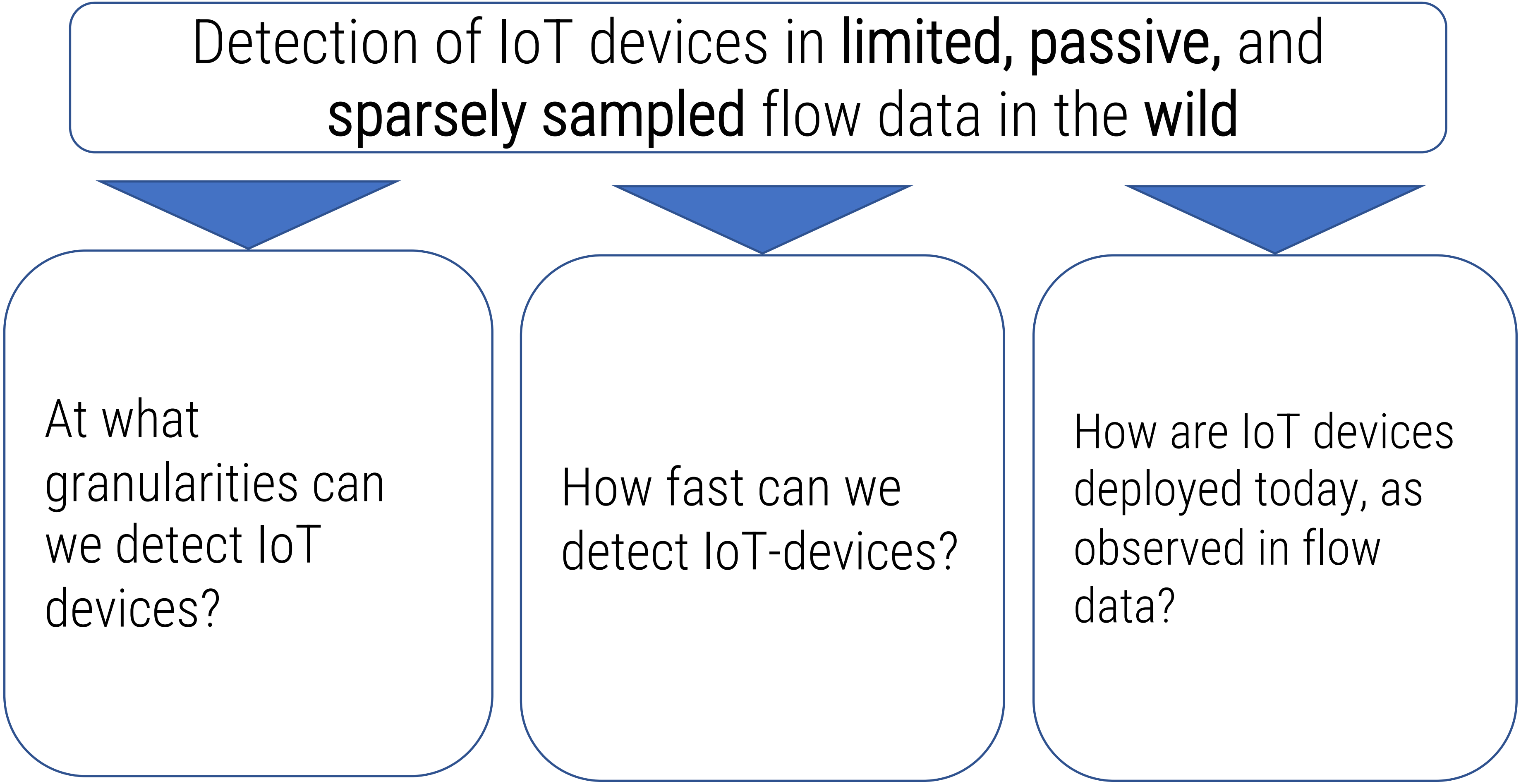
*Kumar et al., USENIX Security’19, All Things Considered: An Analysis of IoT Devices on Home Networks

NetFlow captures for IoT-device discovery



NetFlow Captures

- Collected for other operational purposes
- Sampled, no payload
- Header-only:
  src_ip, dst_ip, src_port, dst_port,proto...
  192.168.1.1,10.1.1.1,12345,1883,TCP



Detection of IoT devices in **limited, passive,** and **sparsely sampled** flow data in the **wild**

At what granularities can we detect IoT devices?

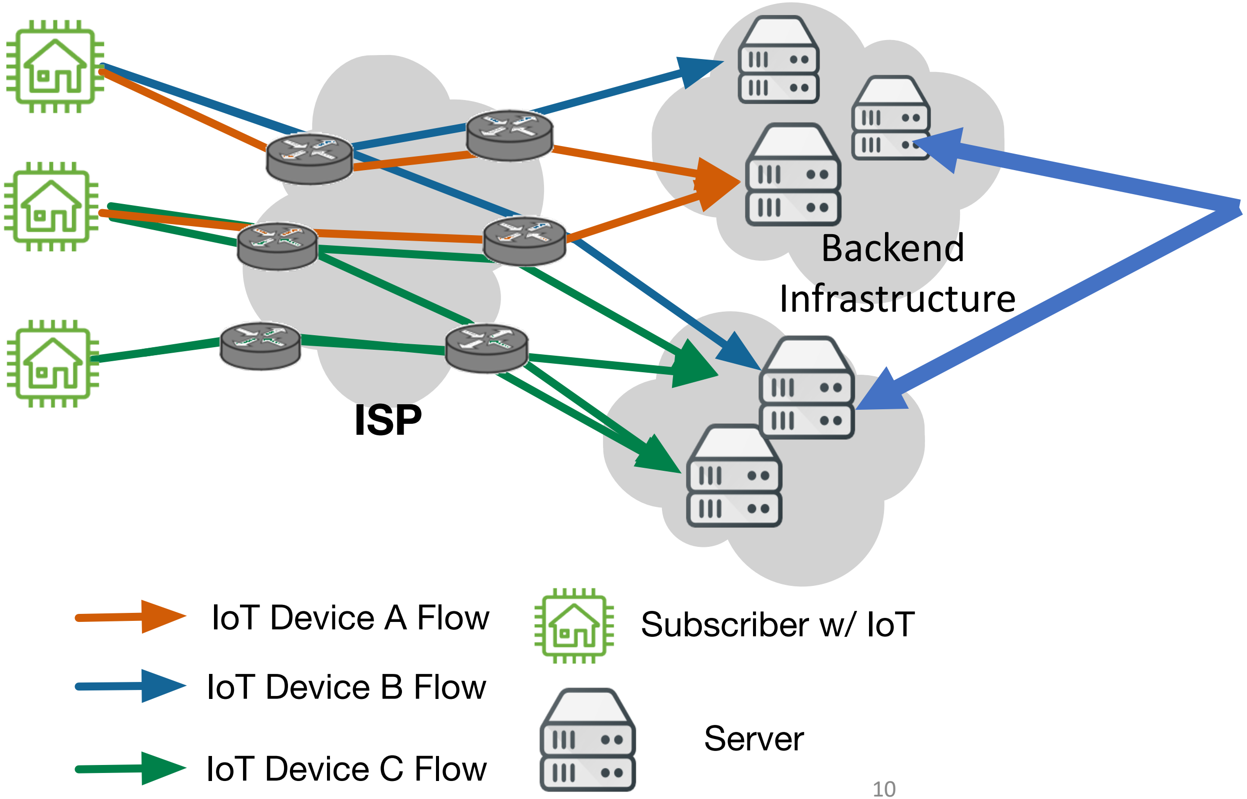
How fast can we detect IoT-devices?

How are IoT devices deployed today, as observed in flow data?

Key Insights

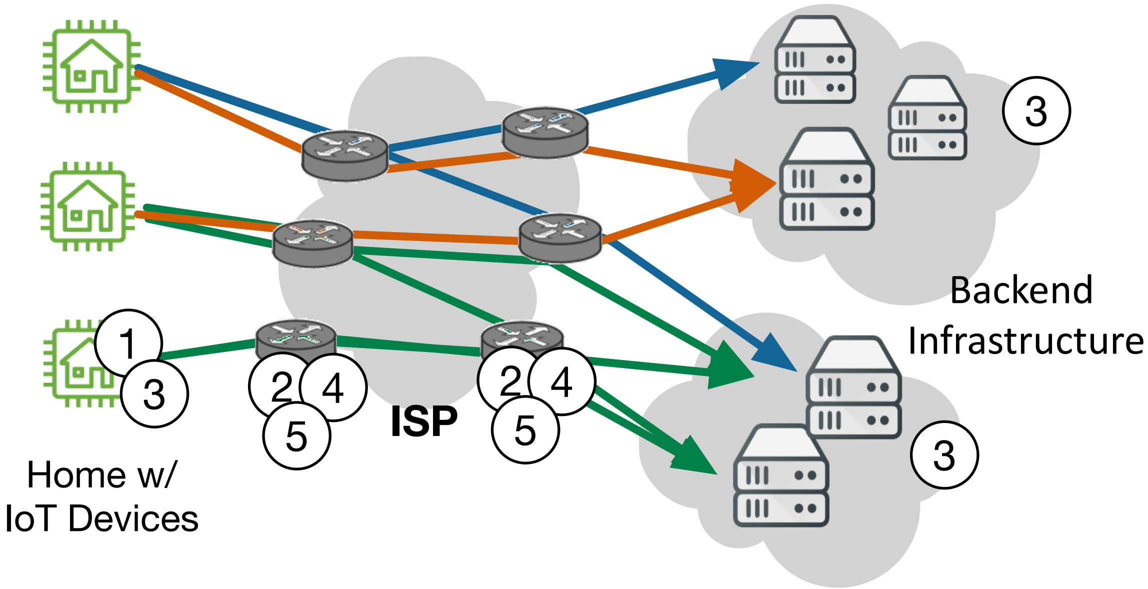
- Devices have repeating patterns of communication that appear even in sparsely sampled data
- Detection rules can be generated using limited packet fields
- Detected devices from 77% of studied IoT manufacturers in an ISP and IXP within minutes to hours

IoT Communication Pattern

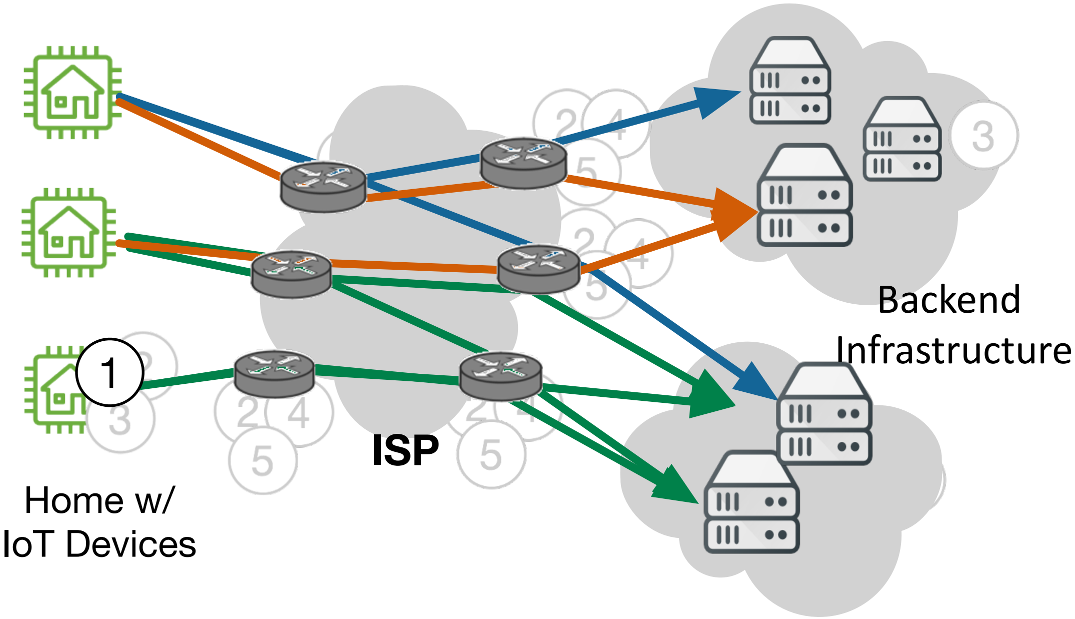


Focus on the destinations contacted by IoT devices

Overview of Methodology



1. Generate IoT Traffic
2. Check Visibility of IoT Traffic at ISP Vantage Point
3. Identify Domains, IPs, and Port numbers and Generate Detection Rules
4. Cross check Detection Rules
5. Detect IoT Devices in the Wild



1 Generate Ground Truth (GT) IoT Traffic

2 Check Visibility of GT at ISP Vantage Point

3 Identify Domains, IPs, and Port numbers and Generate Detection Rules

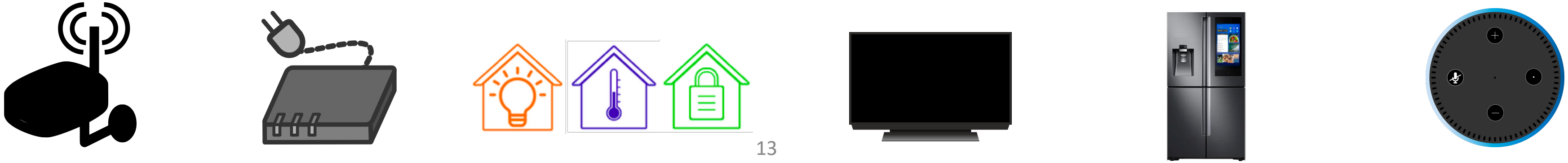
4 Cross check Detection Rules

5 Detect IoT Devices in the Wild

IoT Traffic: Setting up Test Beds

56 IoT Products from 40 Vendors in 2 Testbeds

13 Cameras	8 Smart Hubs	14 Home Automation	5 TVs	10 Appliances	6 Speakers



IoT Traffic: Triggering Devices

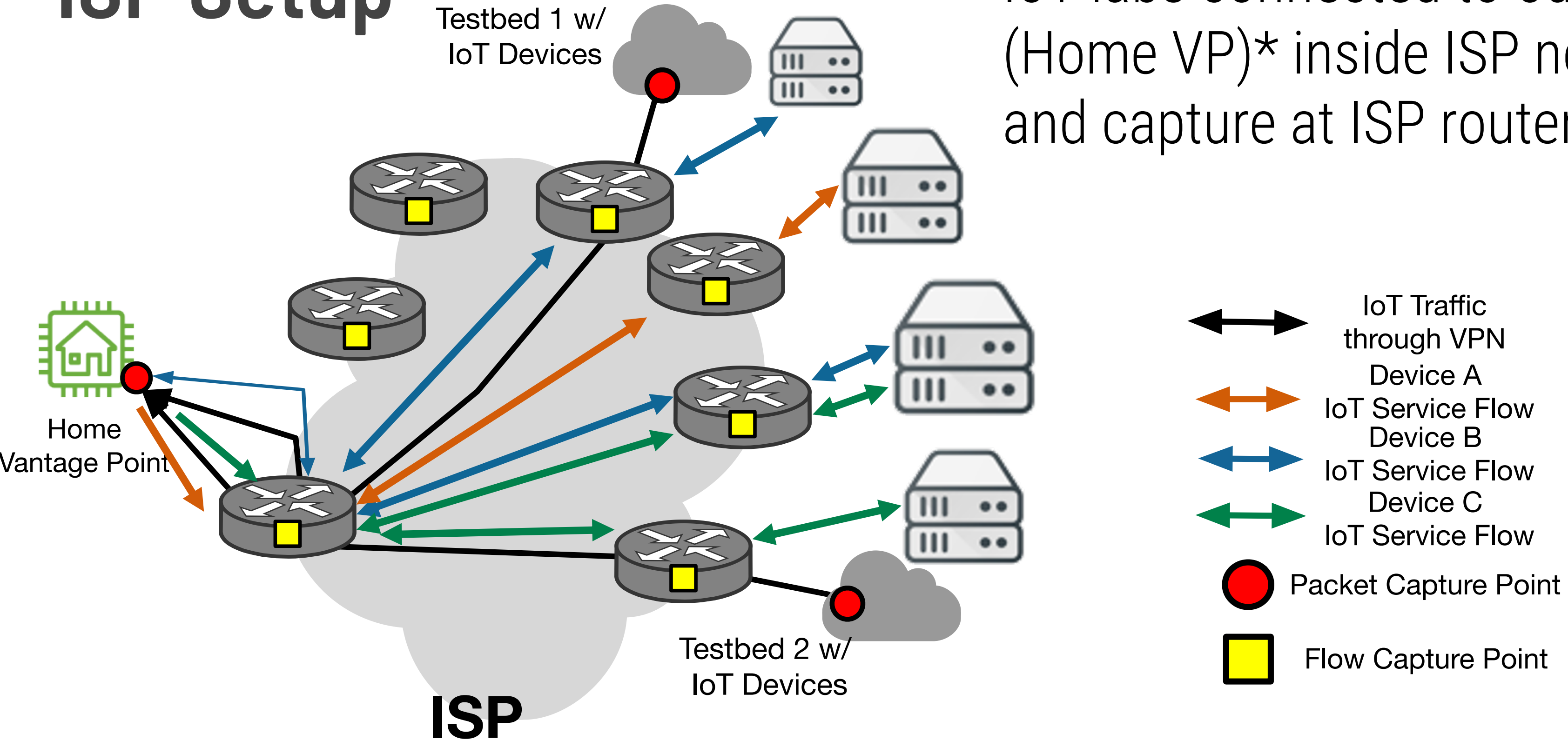
- Idle experiments

- Active experiments: *automated* interaction with the device

Activity	Description
Power	power on/off the device
Voice	voice commands for speakers
Video	record/watch video
On/Off	turn on/off bulbs/plugs
Motion	move in front of device
Others	change volume, browse menu

ISP Setup

IoT labs connected to our Home (Home VP)* inside ISP network and capture at ISP routers



Testbed 1 w/ IoT Devices

Home Vantage Point

Testbed 2 w/ IoT Devices

ISP

↔ IoT Traffic through VPN

↔ Device A IoT Service Flow

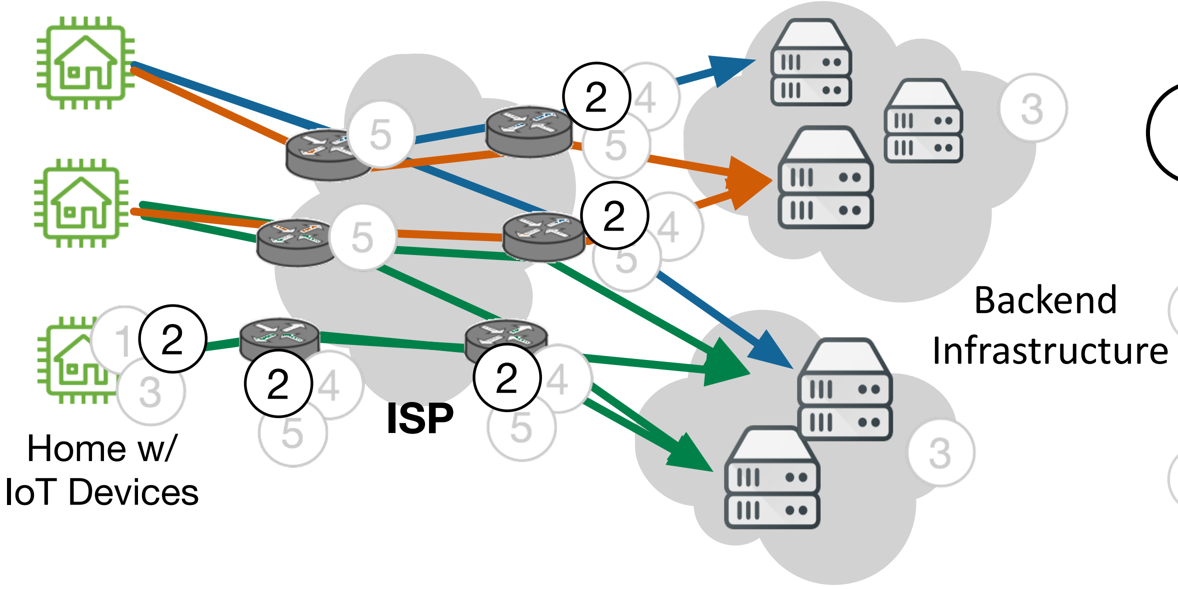
↔ Device B IoT Service Flow

↔ Device C IoT Service Flow

● Packet Capture Point

■ Flow Capture Point

*consenting customer



1 Generate IoT Traffic

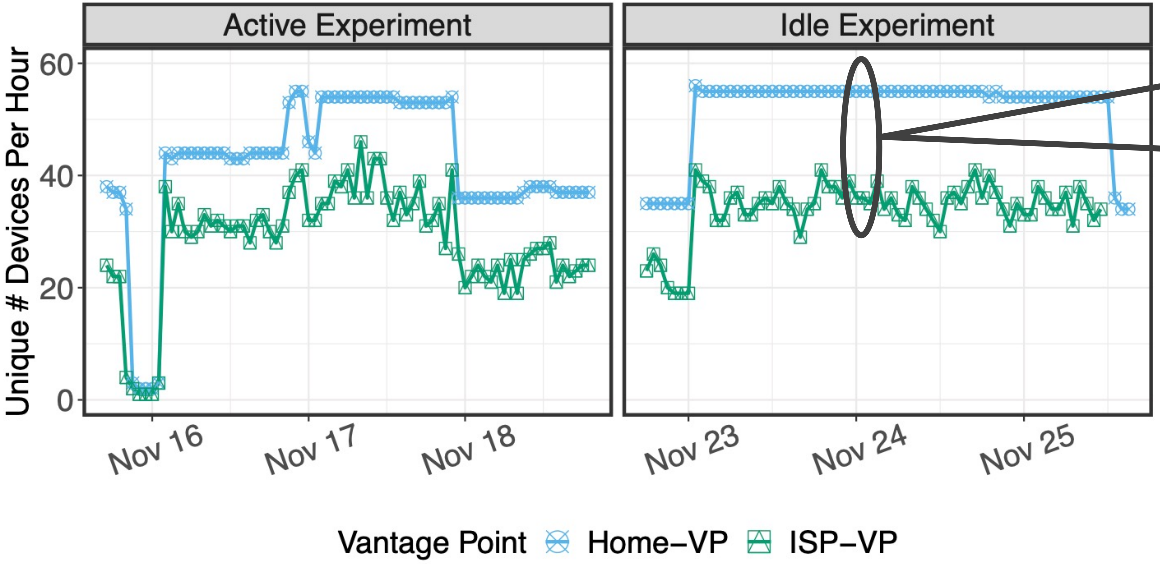
2 Check Visibility of IoT Traffic at ISP Vantage Point

3 Identify Domains, IPs, and Port numbers and Generate Detection Rules

4 Cross check Detection Rules

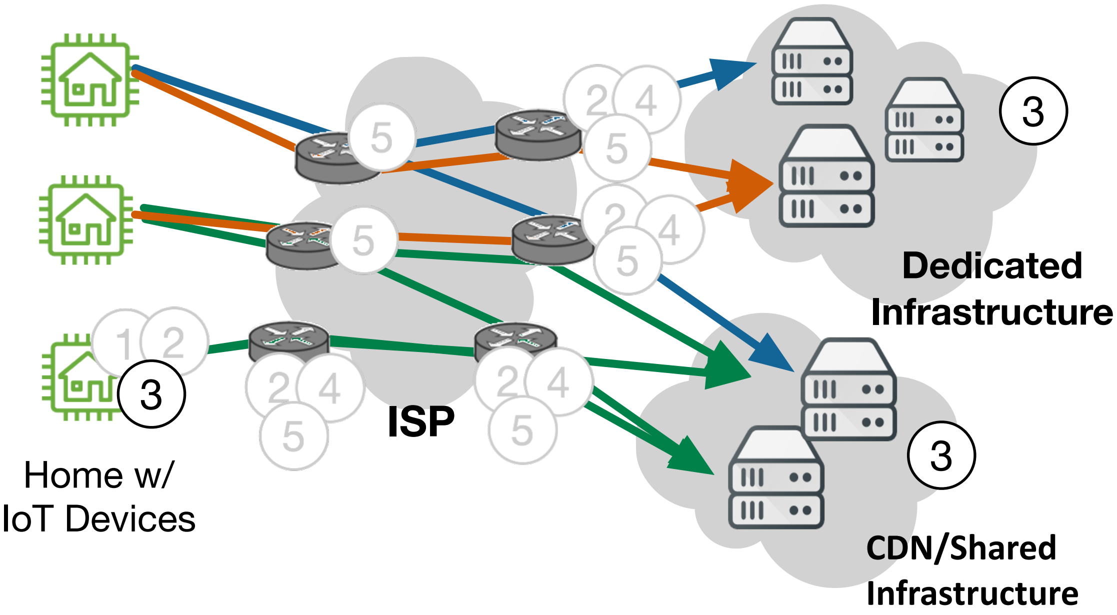
5 Detect IoT Devices in the Wild

Visibility of IoT Traffic– Unique Devices/Hour



Activity from >64% of devices were observed in each hour

Activity: observing at least one packet to/from device



1 Generate IoT Traffic

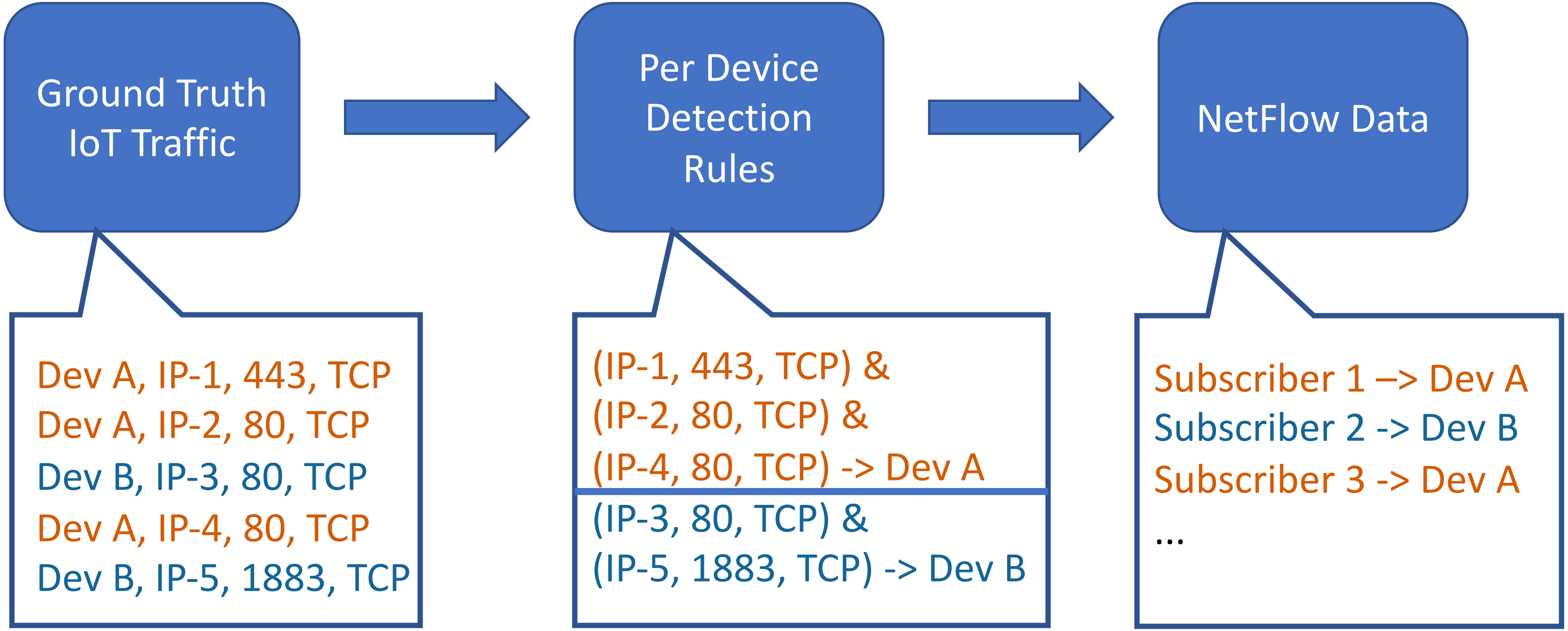
2 Check Visibility of GT at ISP Vantage Point

3 Identify Domains, IPs, and Port numbers and Generate Detection Rules

4 Cross check Detection Rules

5 Detect IoT Devices in the Wild

Detection Rules -> Naive Approach



Ground Truth IoT Traffic

Dev A, IP-1, 443, TCP
Dev A, IP-2, 80, TCP
Dev B, IP-3, 80, TCP
Dev A, IP-4, 80, TCP
Dev B, IP-5, 1883, TCP

Per Device Detection Rules

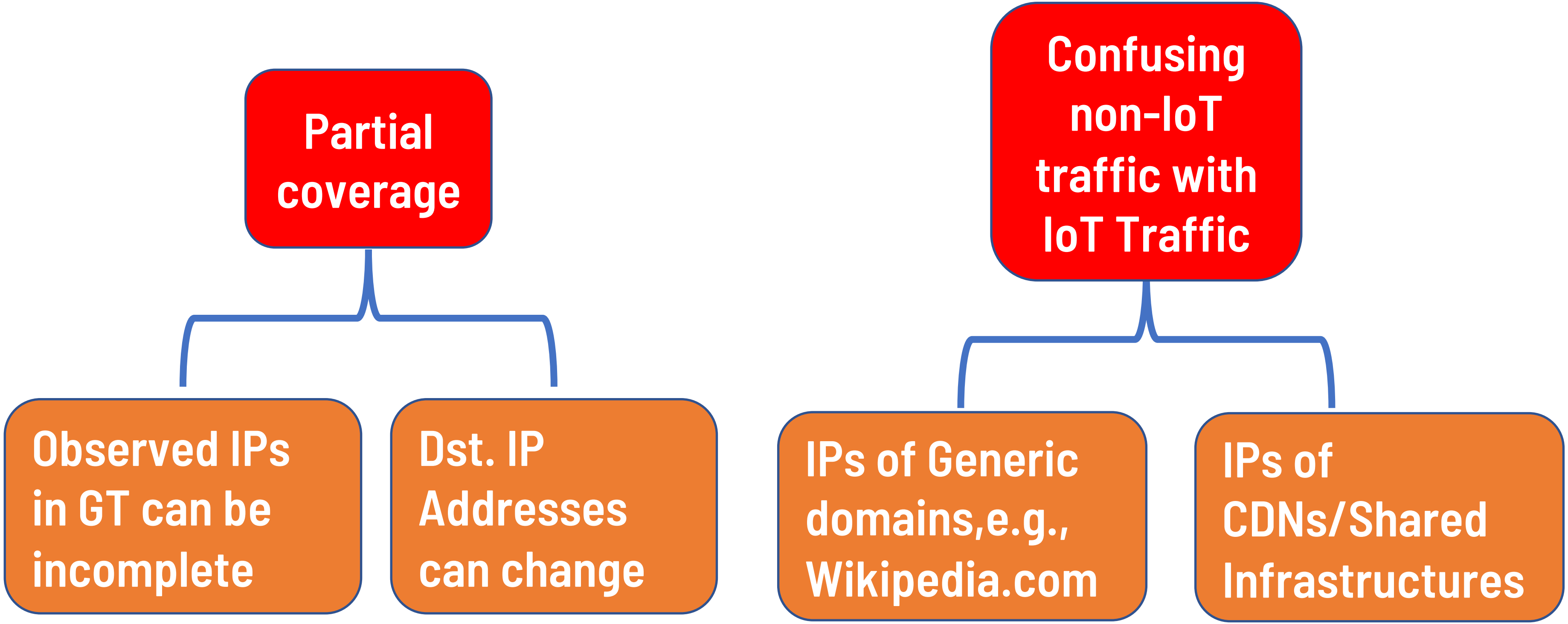
(IP-1, 443, TCP) &
(IP-2, 80, TCP) &
(IP-4, 80, TCP) -> Dev A

(IP-3, 80, TCP) &
(IP-5, 1883, TCP) -> Dev B

NetFlow Data

Subscriber 1 –> Dev A
Subscriber 2 -> Dev B
Subscriber 3 -> Dev A
…

Detection Rules with Naive/IPs Only -> Misclassification



Partial coverage

Observed IPs in GT can be incomplete

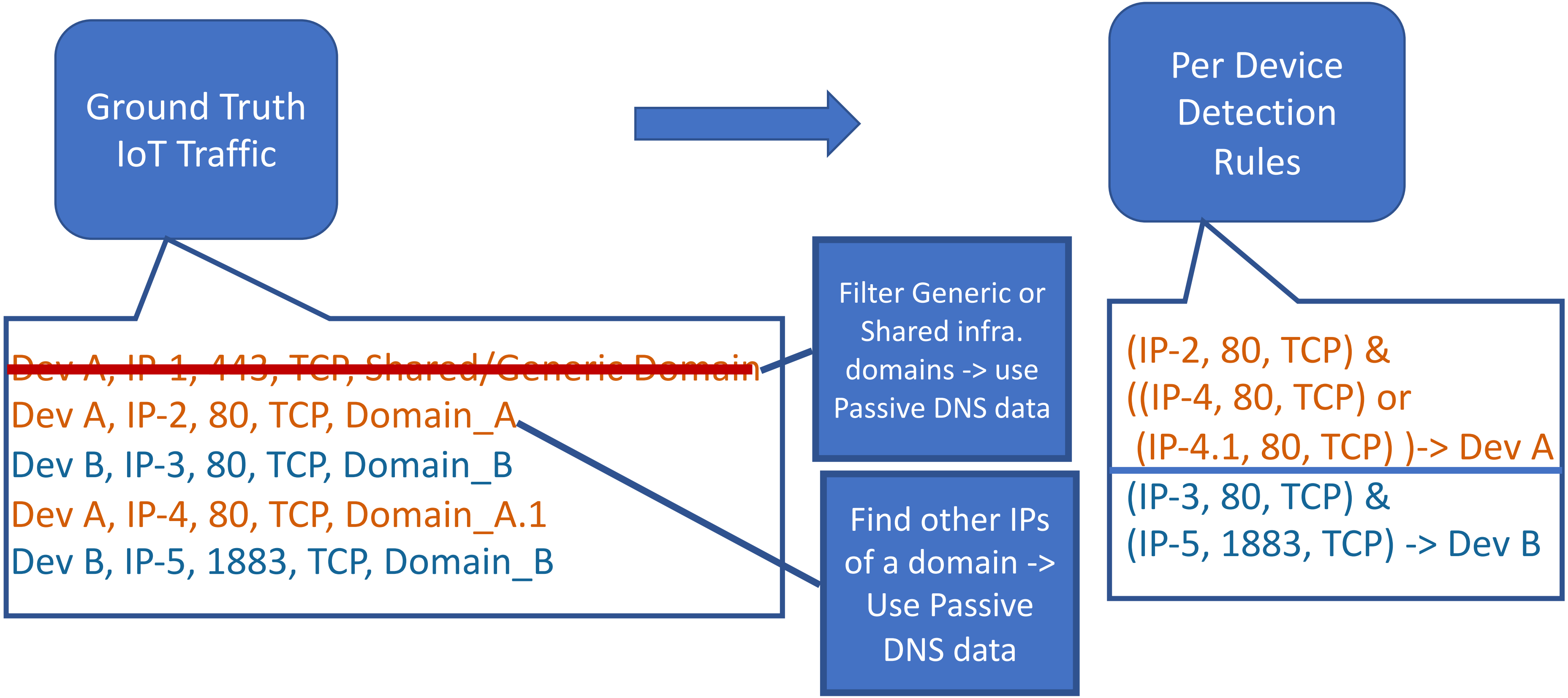
Dst. IP Addresses can change

Confusing non-IoT traffic with IoT Traffic

IPs of Generic domains,e.g., Wikipedia.com

IPs of CDNs/Shared Infrastructures

Detection Rules: Start with Domains



Ground Truth IoT Traffic

~~Dev A, IP-1, 443, TCP, Shared/Generic Domain~~
Dev A, IP-2, 80, TCP, Domain_A
Dev B, IP-3, 80, TCP, Domain_B
Dev A, IP-4, 80, TCP, Domain_A.1
Dev B, IP-5, 1883, TCP, Domain_B

Filter Generic or Shared infra. domains -> use Passive DNS data

Find other IPs of a domain -> Use Passive DNS data

Per Device Detection Rules

(IP-2, 80, TCP) &
((IP-4, 80, TCP) or
(IP-4.1, 80, TCP))-> Dev A

(IP-3, 80, TCP) &
(IP-5, 1883, TCP) -> Dev B

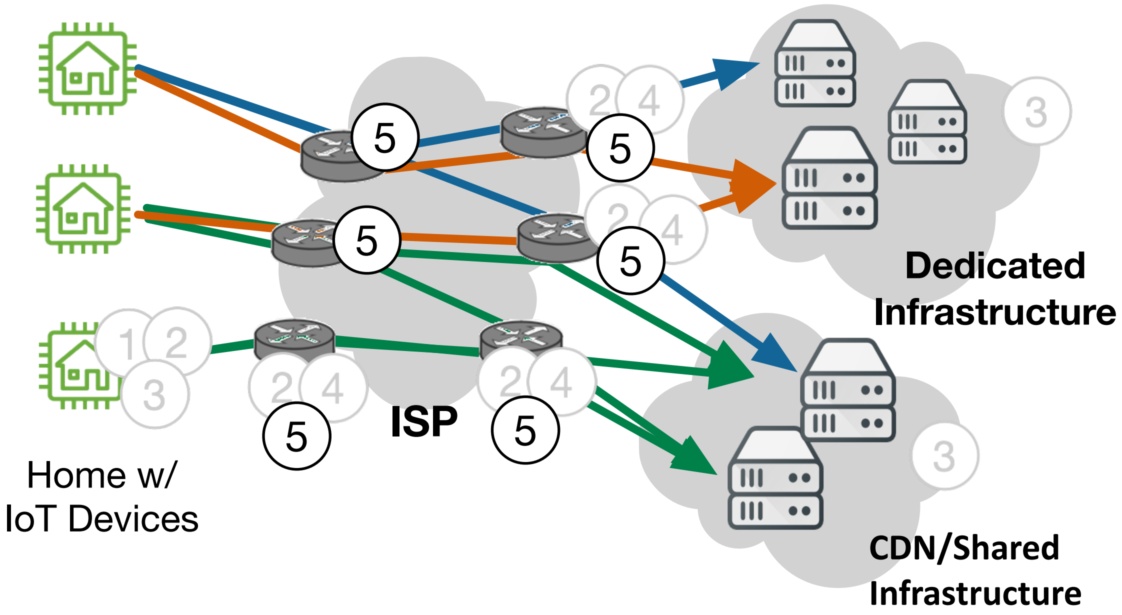
Granularity of Detection Rules

Product-level: Amazon Echo -> **11 Products**

Manufacturer-level: a Samsung Device -> **20 Manufacturers**

Platform-level: a generic IoT device -> **4 IoT Platforms**
(we can't infer the product type or manufacturer)

77% of the manufacturers in the testbeds



1 Generate IoT Traffic

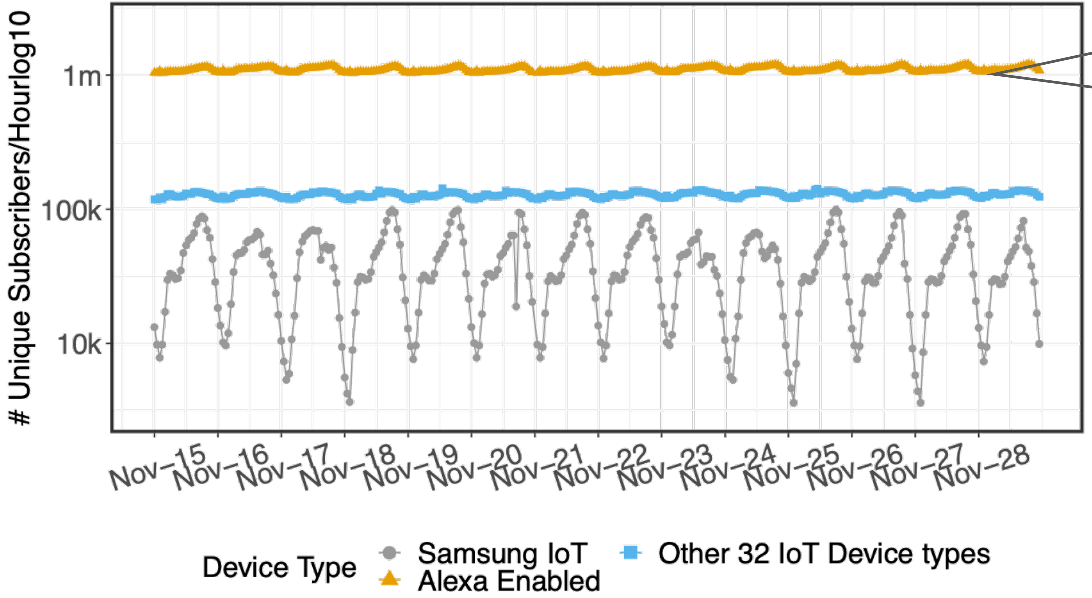
2 Check Visibility of IoT Traffic at ISP Vantage Point

3 Identify Domains, IPs, and Port numbers and Generate Detection Rules

4 Cross check Detection Rules

5 Detect IoT Devices in the Wild

# of ISP Subscribers with IoT Devices (Per Hour)

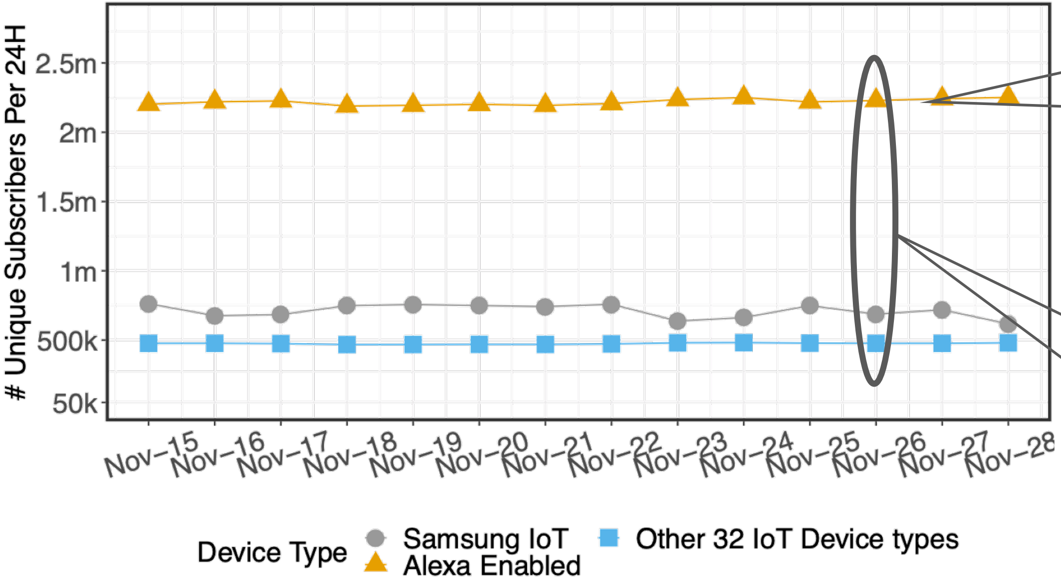


1m+ subscribers with Alexa-enabled devices

- Some diurnal patterns for Alexa and Samsung IoT devices

Alexa-enabled: Any device that responds to Amazon Alexa voice service commands

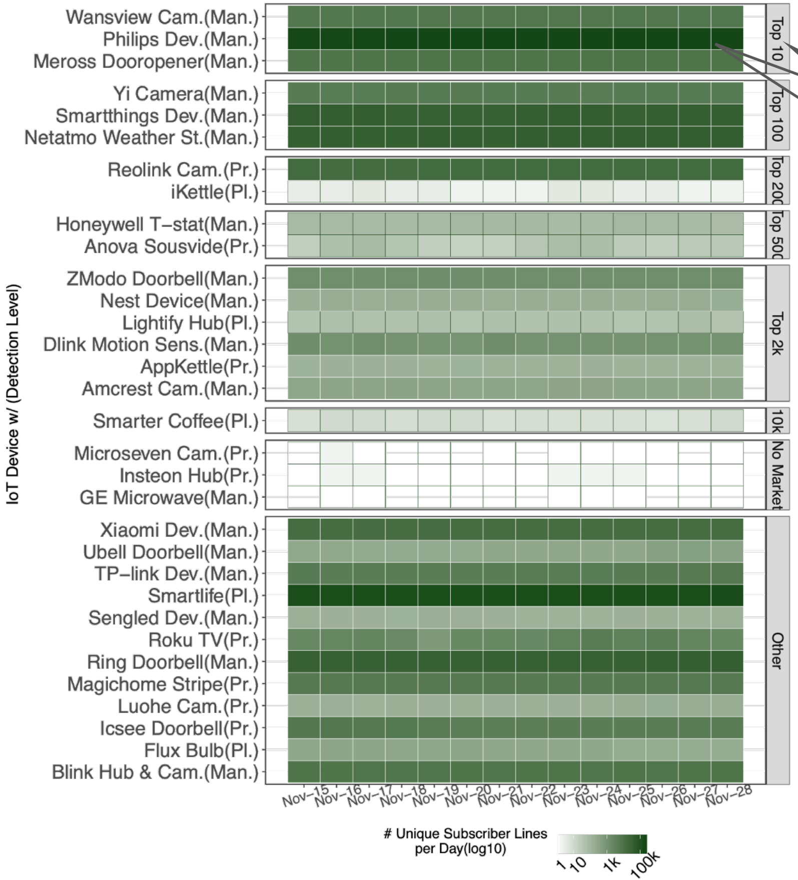
# of ISP-subscribers with IoT Devices (per 24 hours)



Increasing observation period, helped detecting more devices

IoT activity for ~20% of ISP subscriber lines

Breakdown of Detected IoT Devices



Device popularity in the Amazon and ISP look correlated.

Limitations

- Devices relying on shared infrastructure
- Generating rules require studying a range of manufacturers' products
- Domain names and IPs might change
- Detection of devices with small activity

Conclusions

- A methodology to detect IoT devices based on limited, sampled flow data
- Detected devices from more than 77% of studied IoT manufacturers in a large ISP
- 4 million devices were detected (both popular and *not-so-popular*)
- Domains and rules are available at : **https://moniotrlab.ccis.neu.edu/imc20/**

@jawadsaidi
jsaidi@mpi-inf.mpg.de