How to Count Bots in Longitudinal Datasets of IP Addresses

Leon Böck*, Dave Levin §, Ramakrishna Padmanabhan #, Christian Doerr &, Max Mühlhäuser*

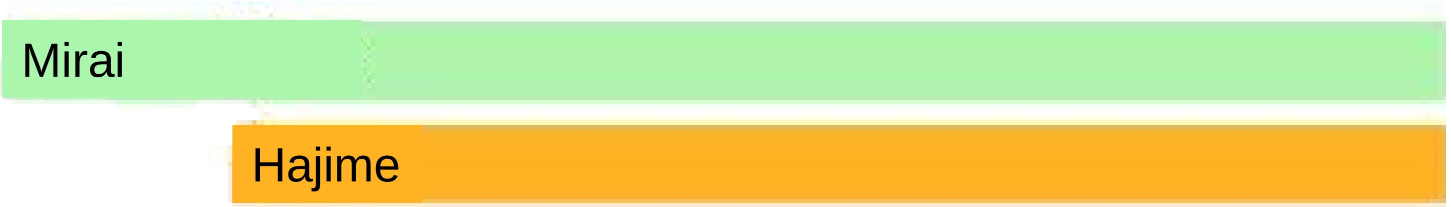
Technische Universität Darmstadt *

University of Maryland, College Park §

CAIDA, UCSD #

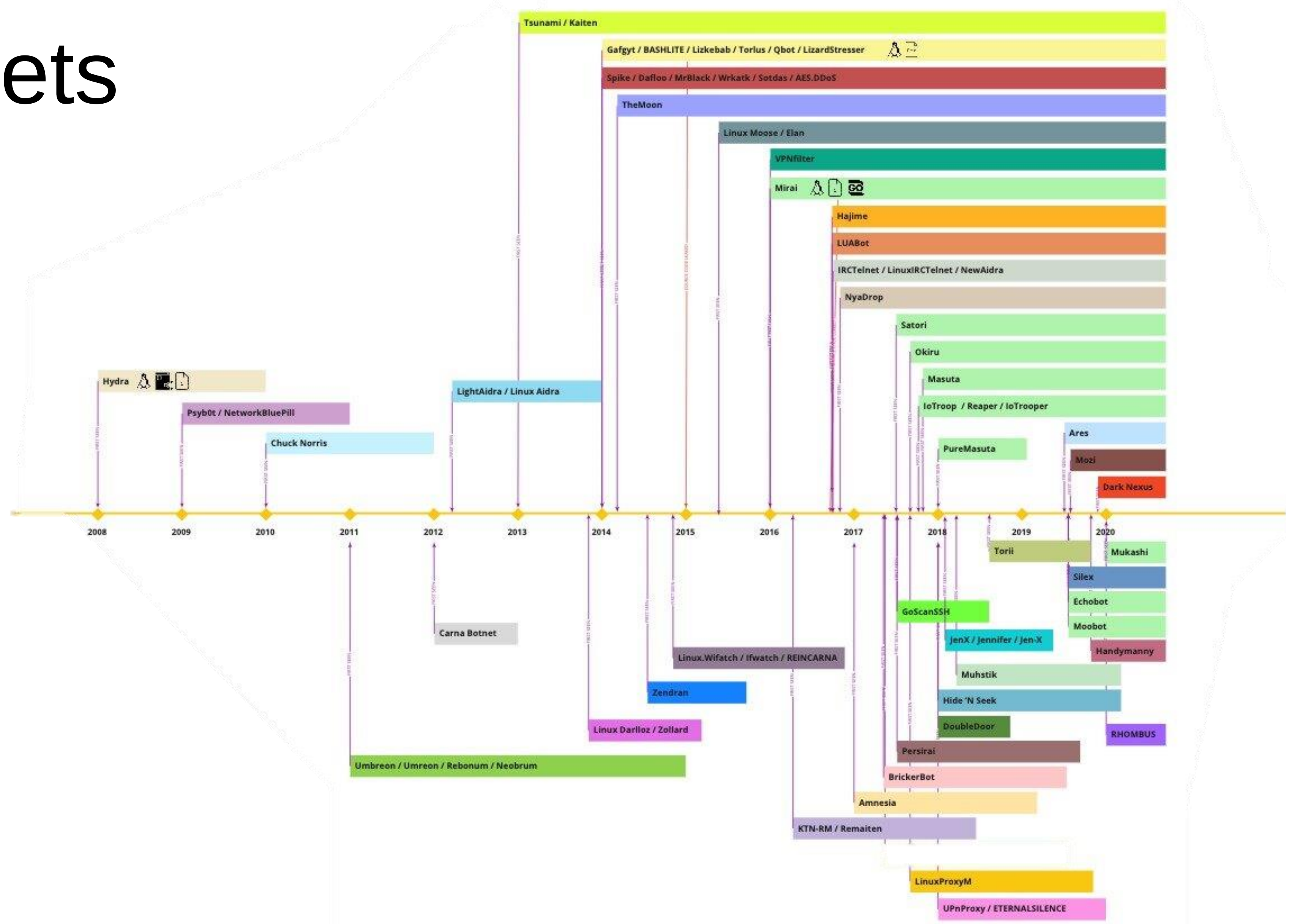
Hasso Plattner Institute, University of Potsdam &

IoT Botnets

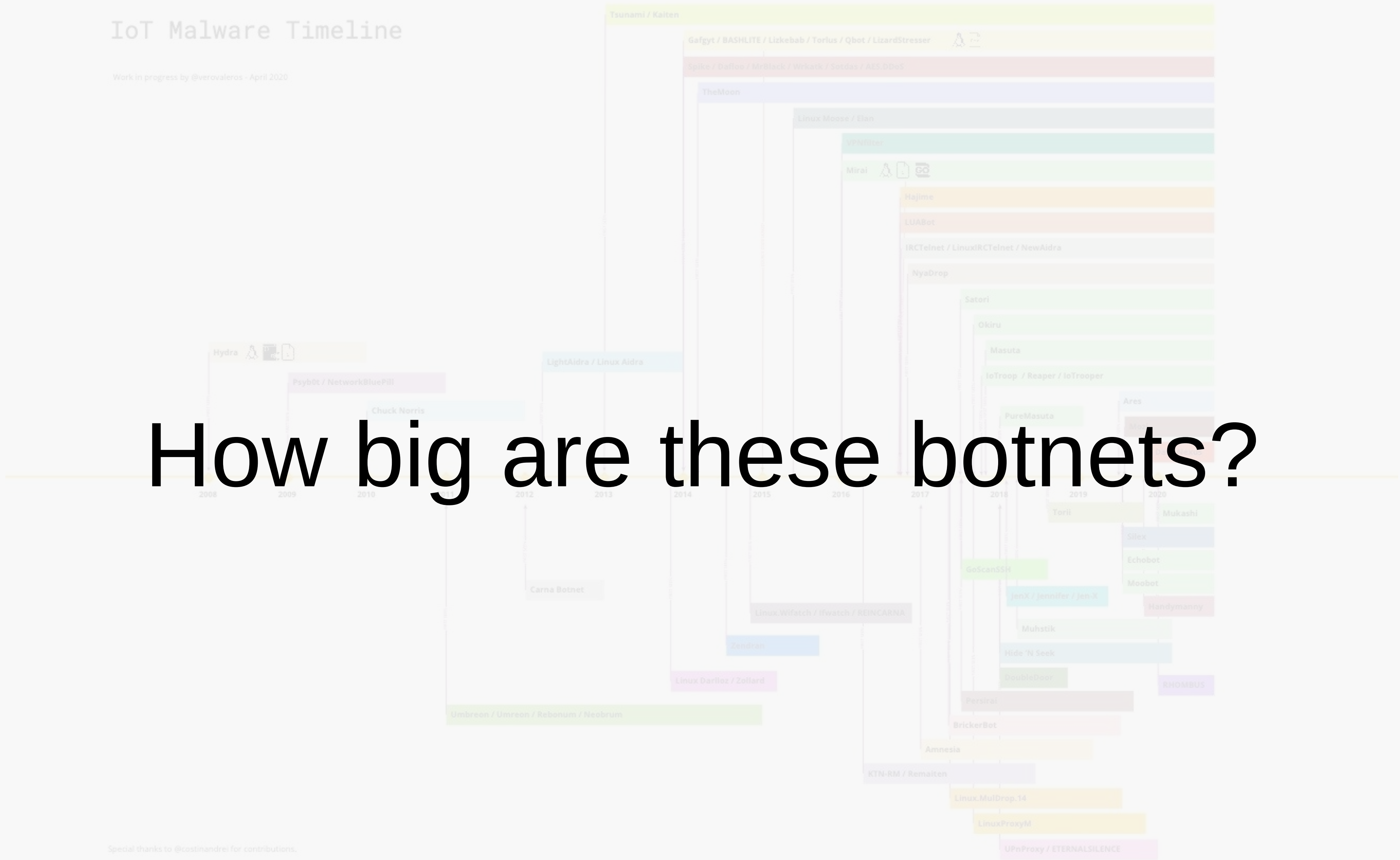


Work by Veronica Valeros @verovaleros
https://www.stratosphereips.org/a-study-of-iot-malware

IoT Botnets



Work by Veronica Valeros @verovaleros
https://www.stratosphereips.org/a-study-of-iot-malware



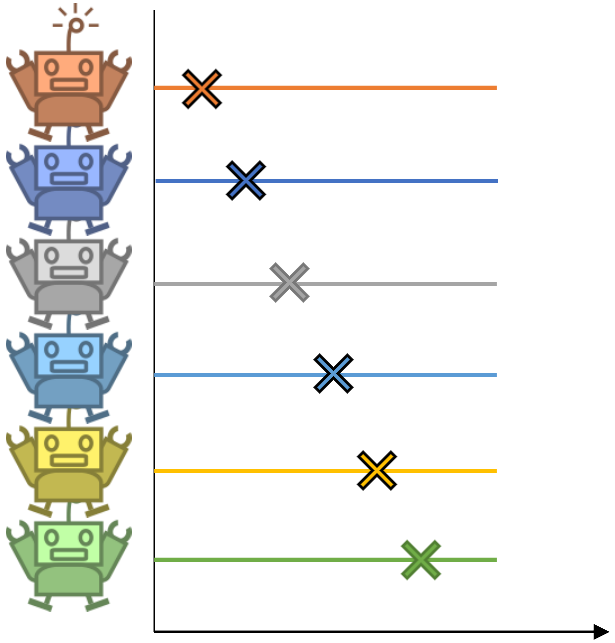
How big are these botnets?

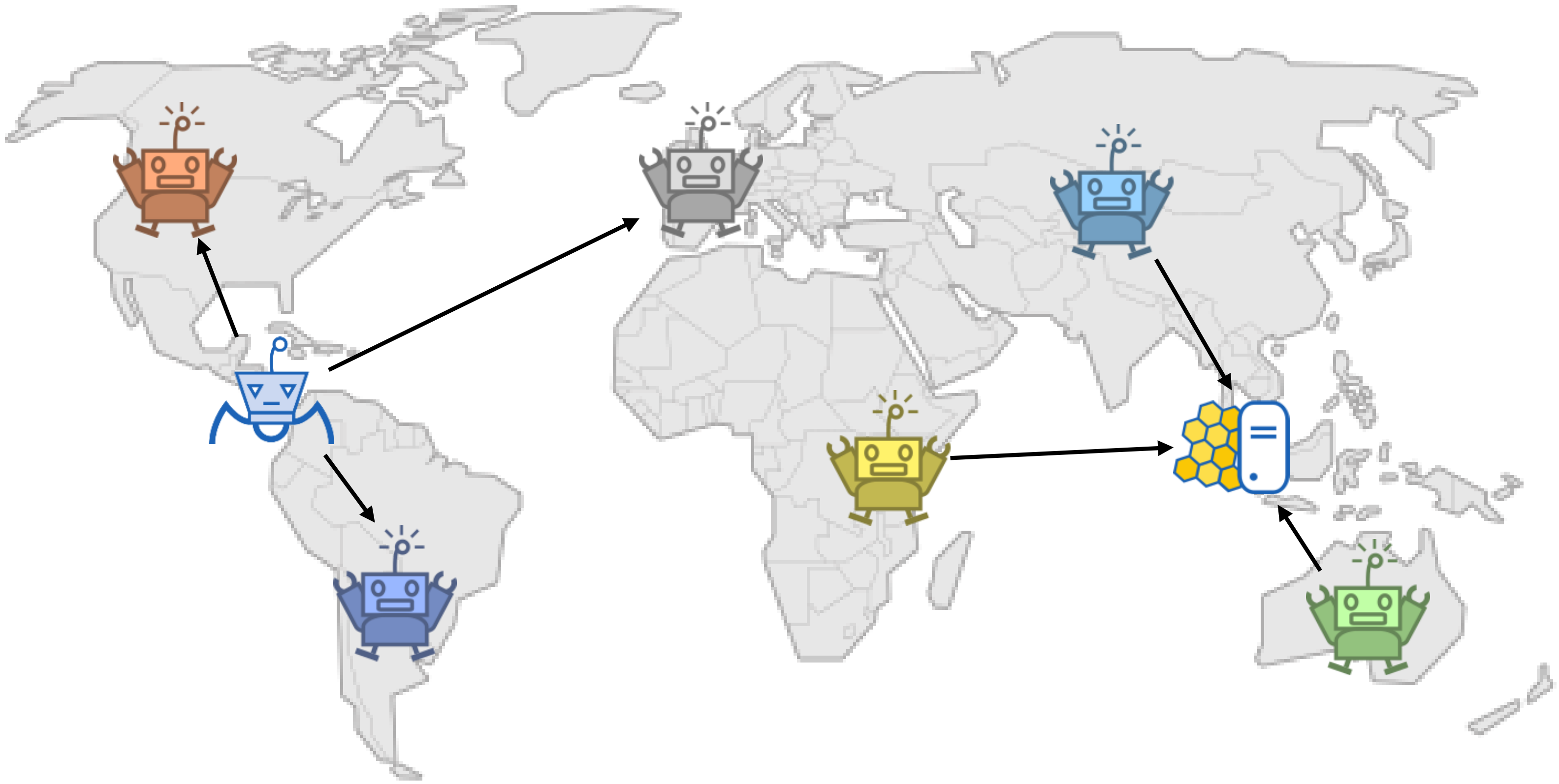
Veronica Valeros @verovaleros
https://www.stratosphereips.org/a-study-of-iot-malware

Bot

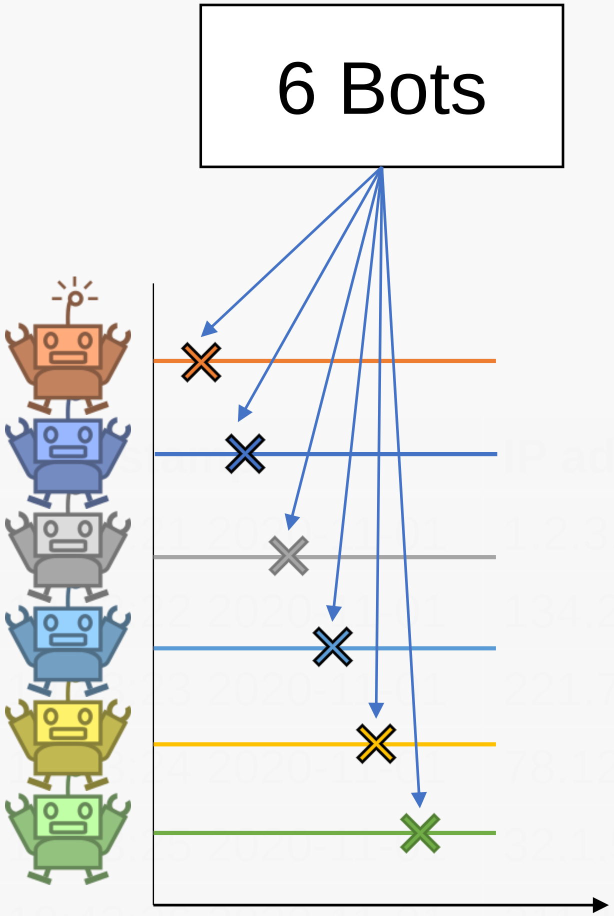
Crawler

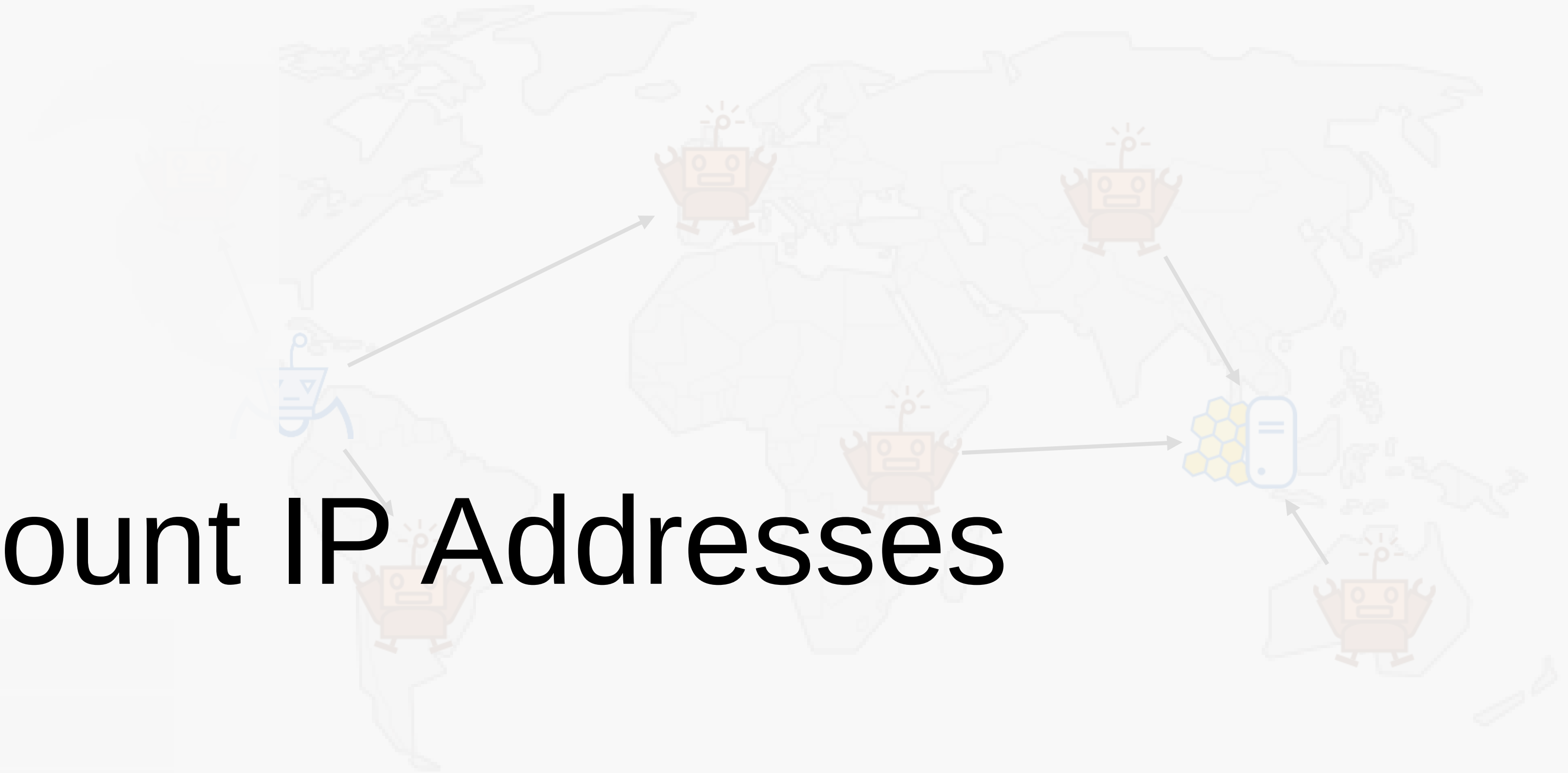
Honeypot





How big are these botnets?

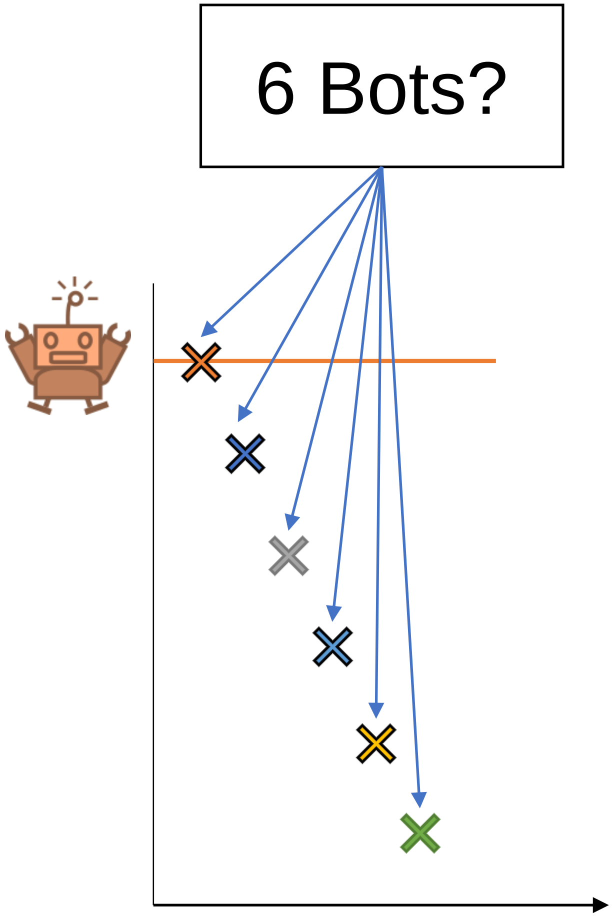

6 Bots



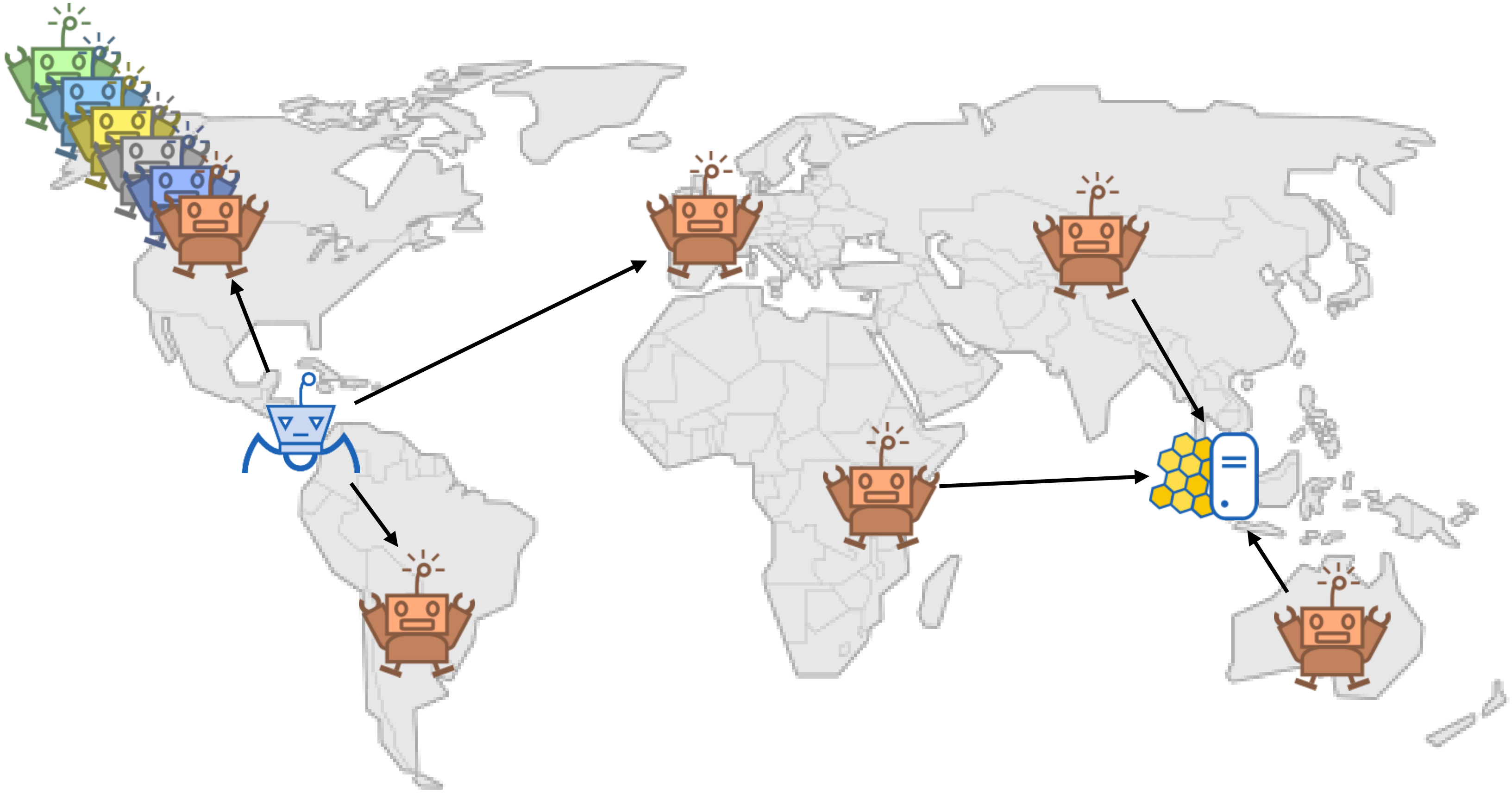
Count IP Addresses

[illegible]stamp	IP address
[illegible]21 2020-11-01	1.2.3.4
[illegible]22 2020-11-01	134.21.31.45
[illegible]23 2020-11-01	221.76.87.91
[illegible]24 2020-11-01	78.12.14.65
[illegible]25 2020-11-01	32.1.5.122
19:43:26 2020-11-01	211.1.23.65

IP Addresses change



6 Bots?



CARDCount

Considering Address Reassignment Durations when Counting

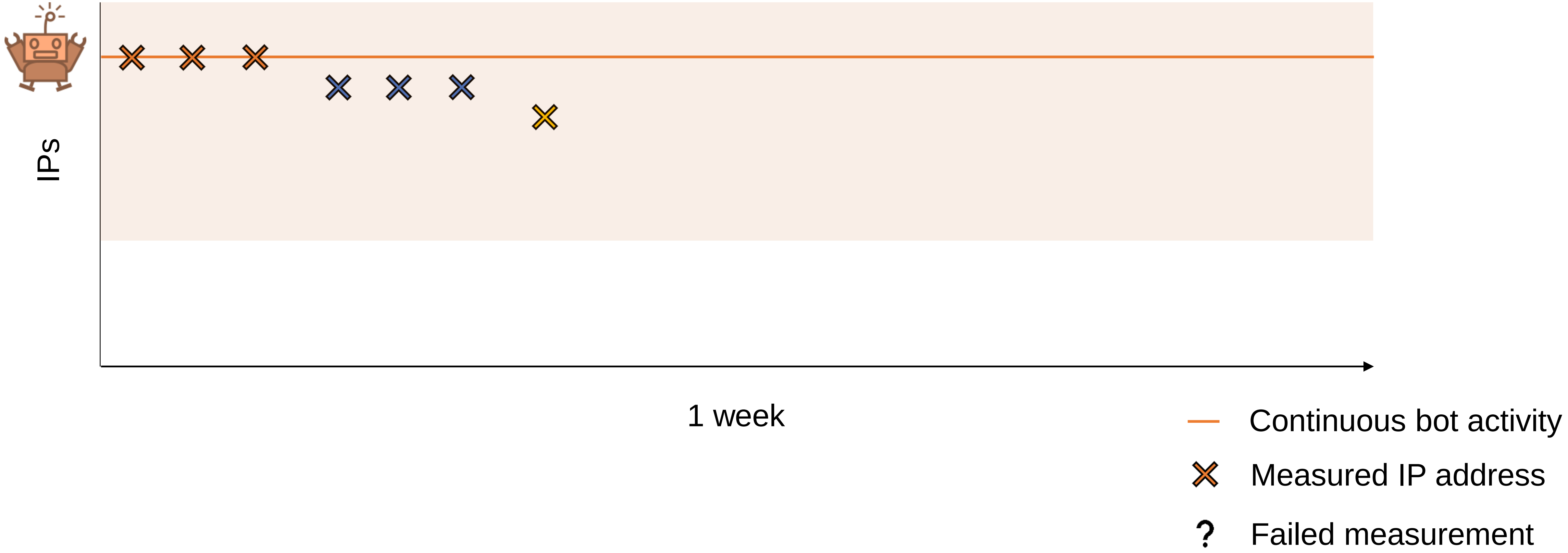
Accurate botnet size estimation

Provides confidence intervals

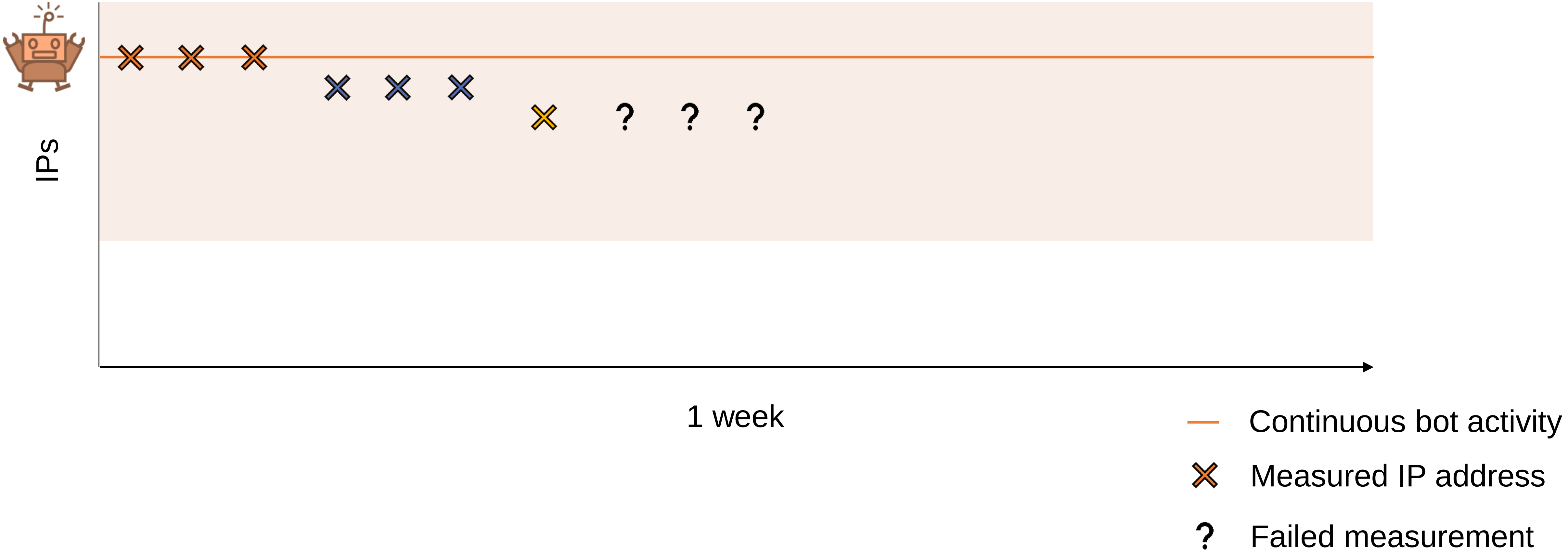
Accurate for long measurement durations

Resilient to incomplete data

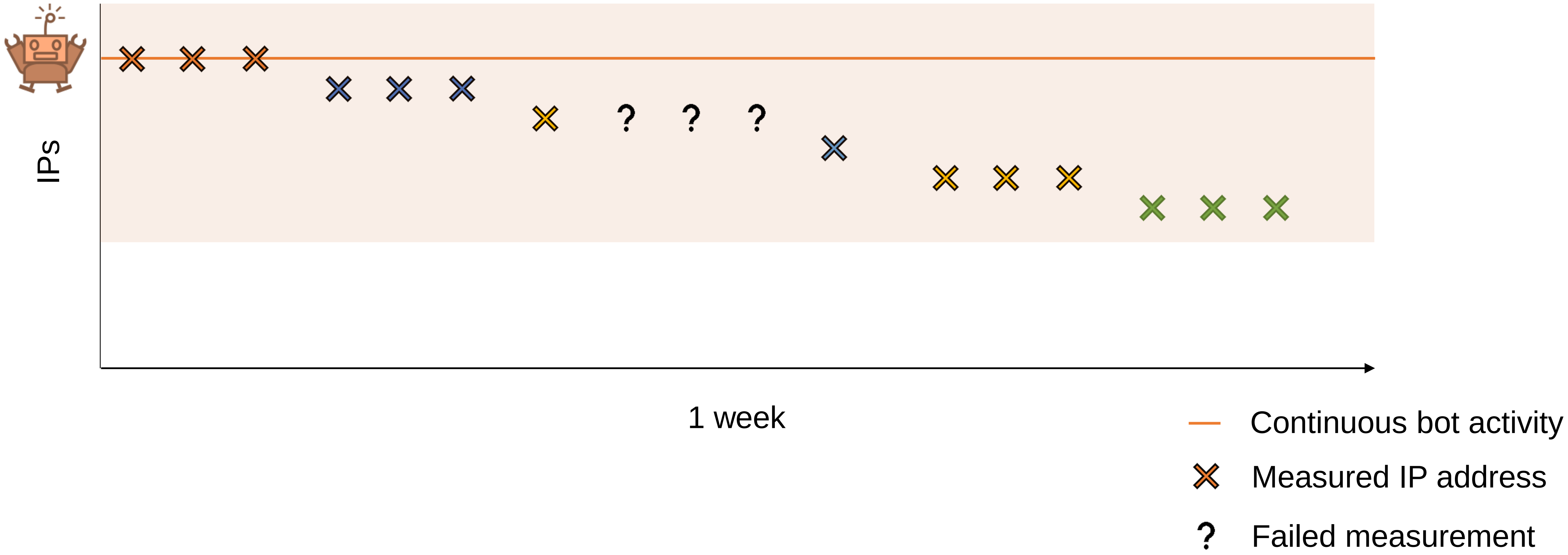
State of the Art



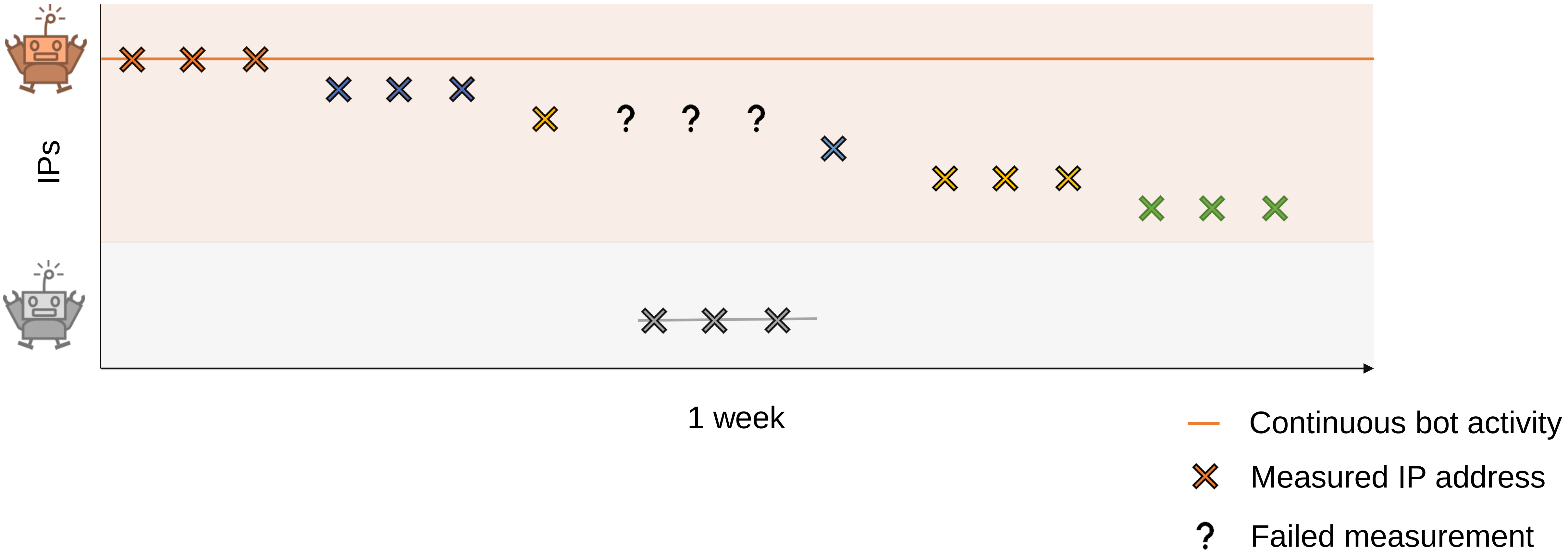
State of the Art



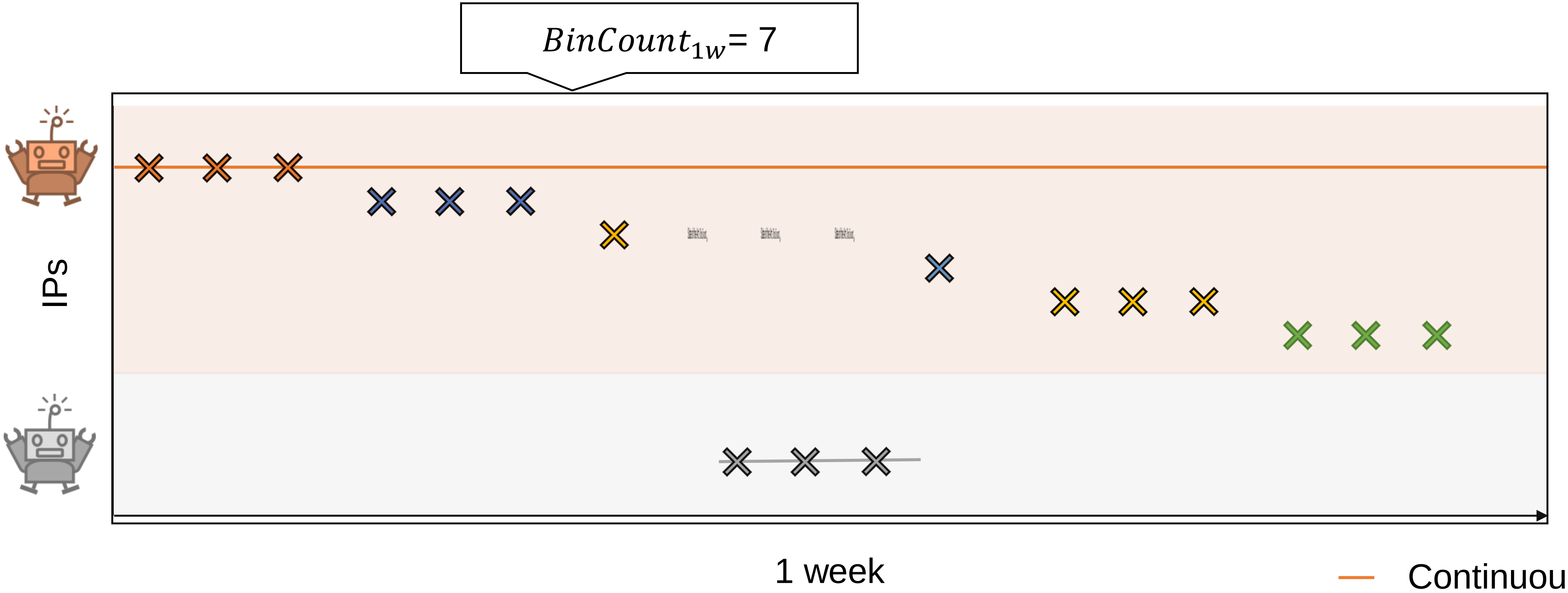
State of the Art



State of the Art



State of the Art: $BinCount_{\omega}$

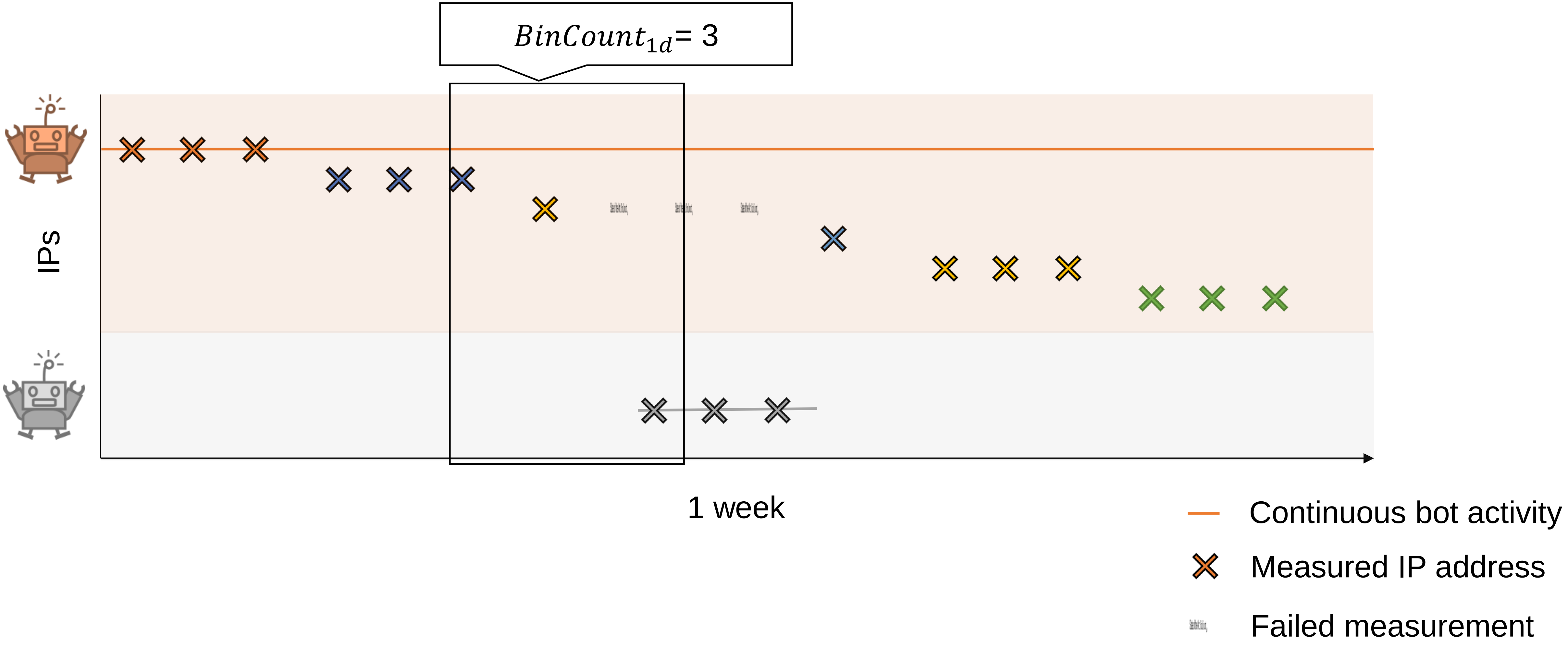


— Continuous bot activity

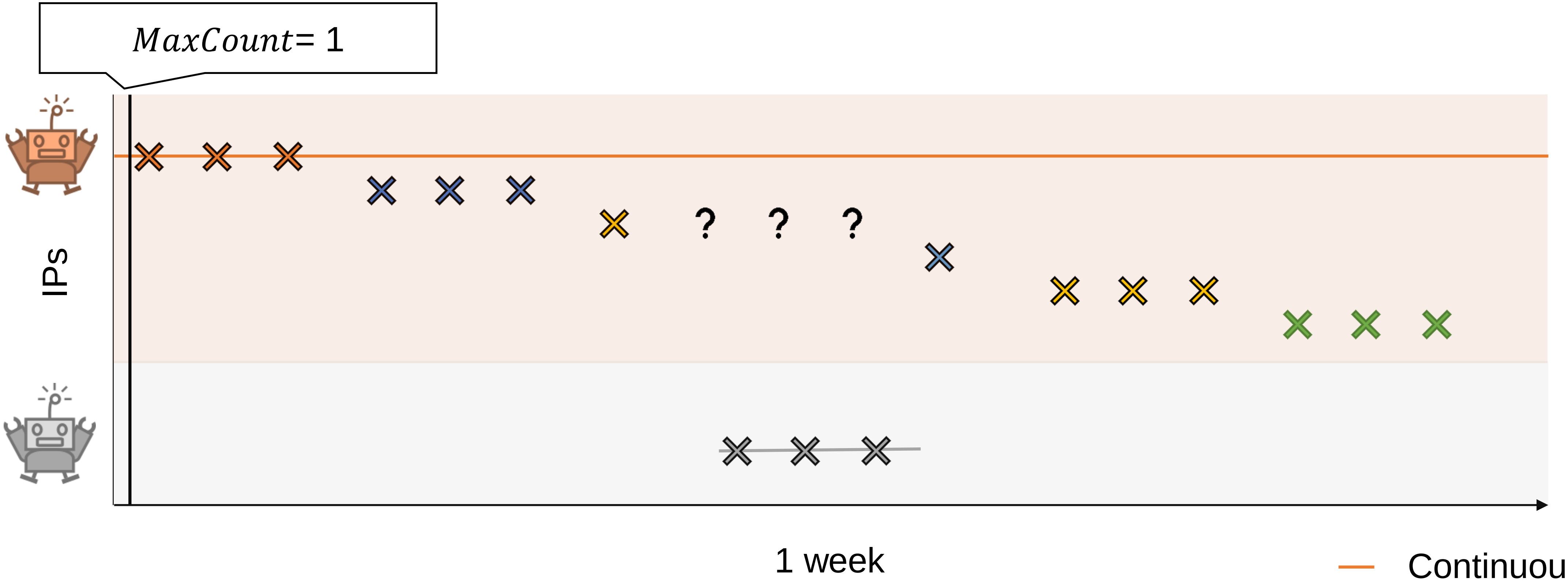
✕ Measured IP address

Failed measurement

State of the Art: $BinCount_{\omega}$



State of the Art: $MaxCount$



— Continuous bot activity

✕ Measured IP address

? Failed measurement

State of the Art: $MaxCount$

$MaxCount$= 1
New IPs
1 week

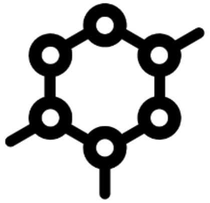
$$MaxCount_{AS}$$

— Continuous bot activity

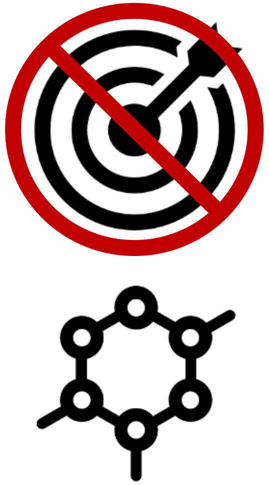
✕ Measured IP address

? Failed measurement

Comparison

BinCount	MaxCount	CARDCount
		

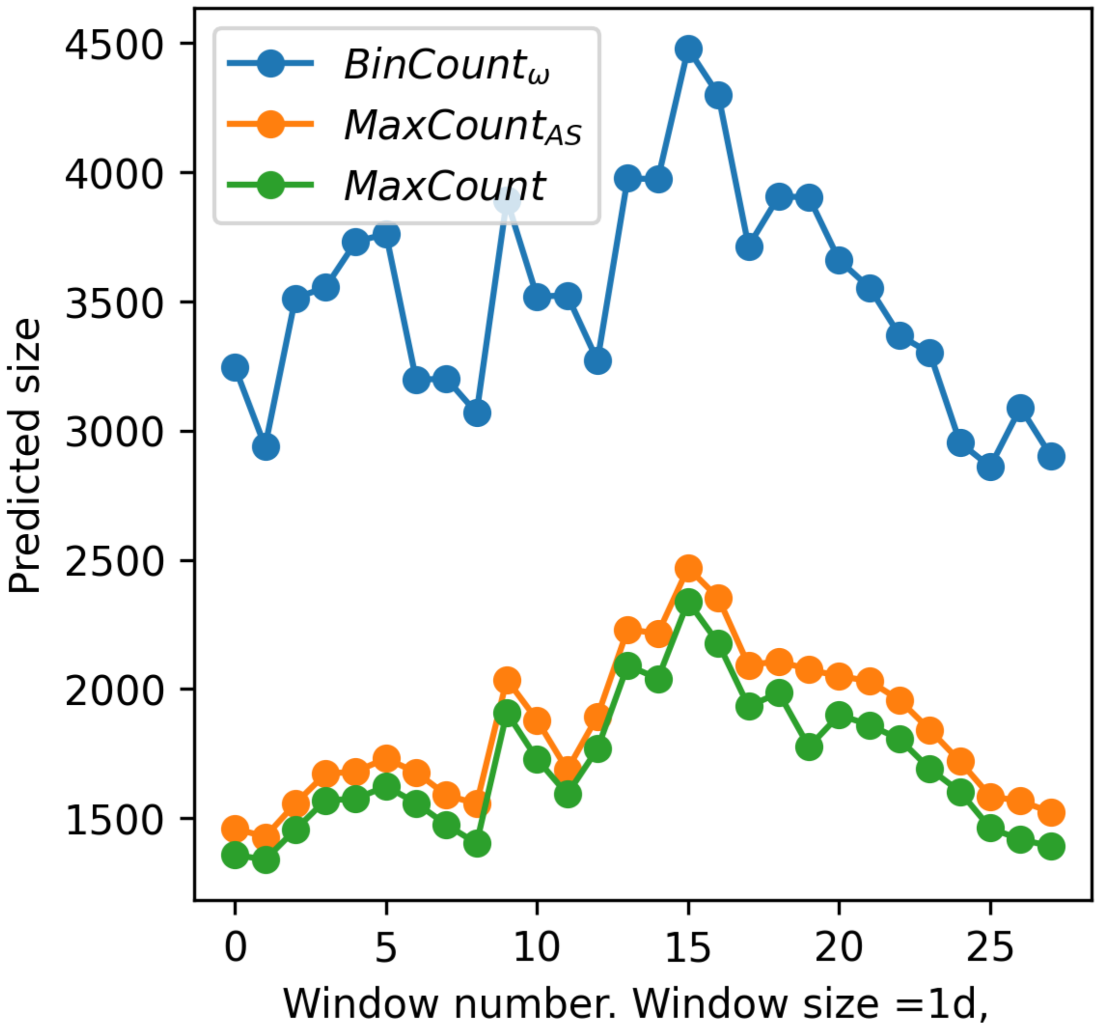
Comparison

BinCount	MaxCount	CARDCount
		

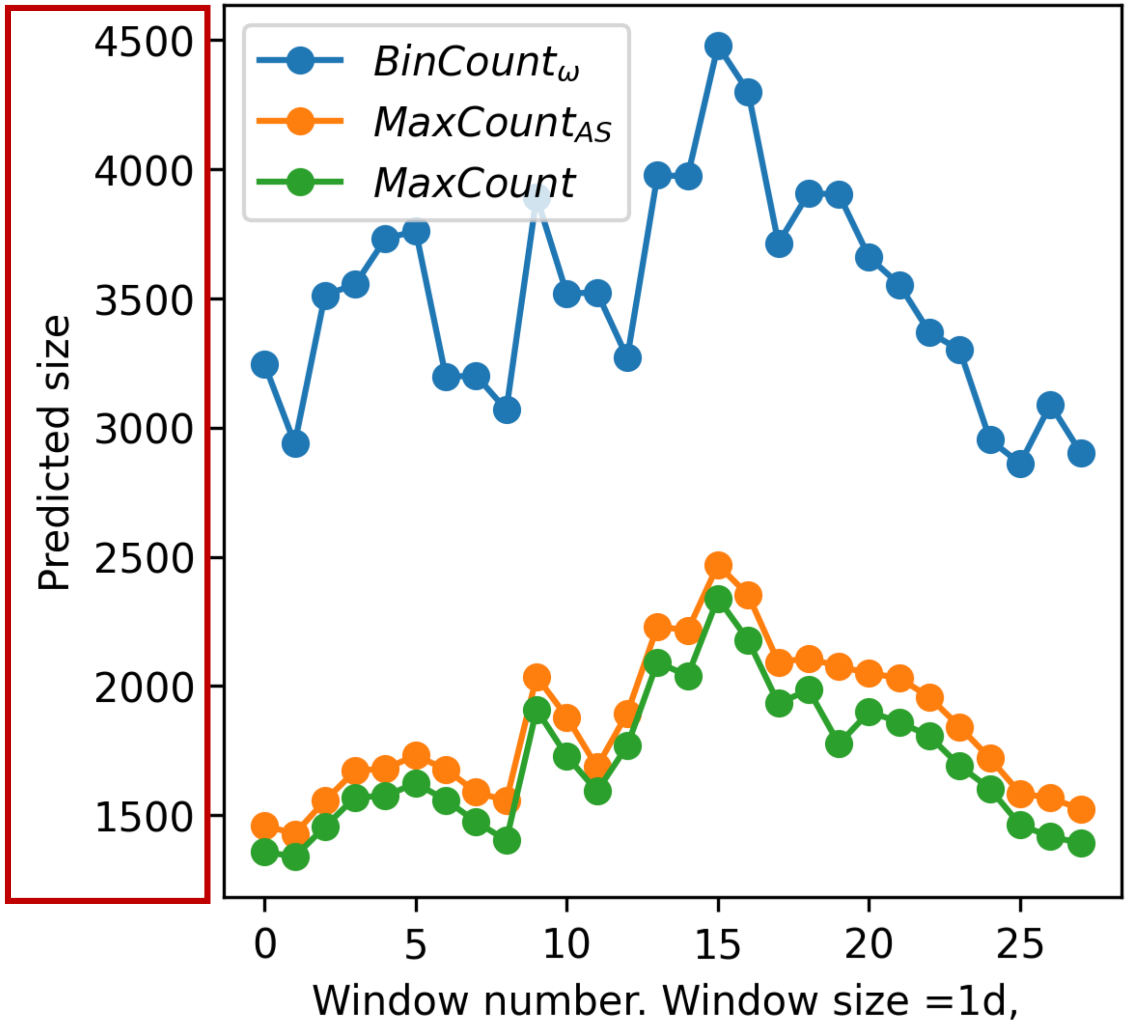
Comparison

BinCount	MaxCount	CARDCount
🚫🎯	🎯	🎯
⬡	🚫🕓	
	🚫⬡	🕓
		⬡

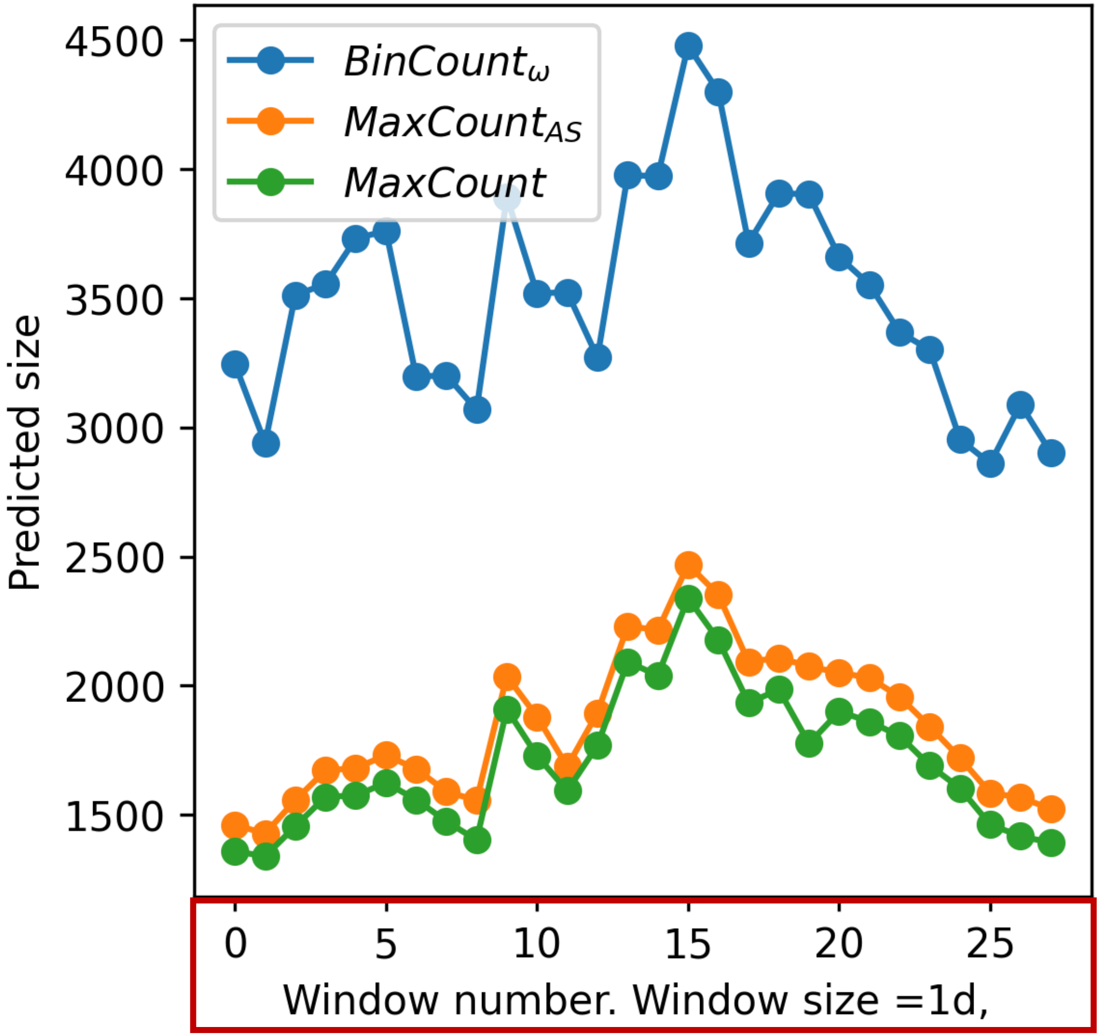
Mirai botnet size



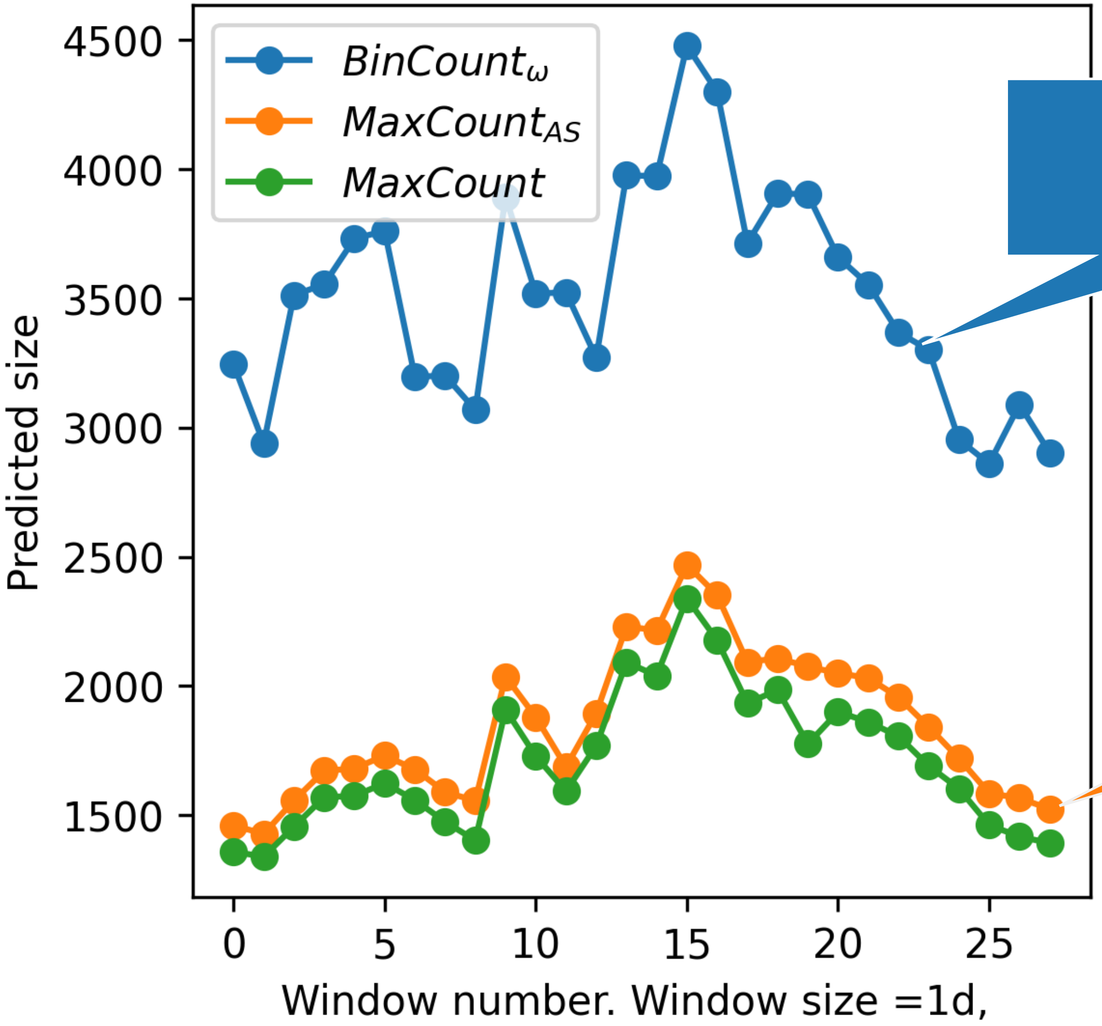
Mirai botnet size



Mirai botnet size



Mirai botnet size

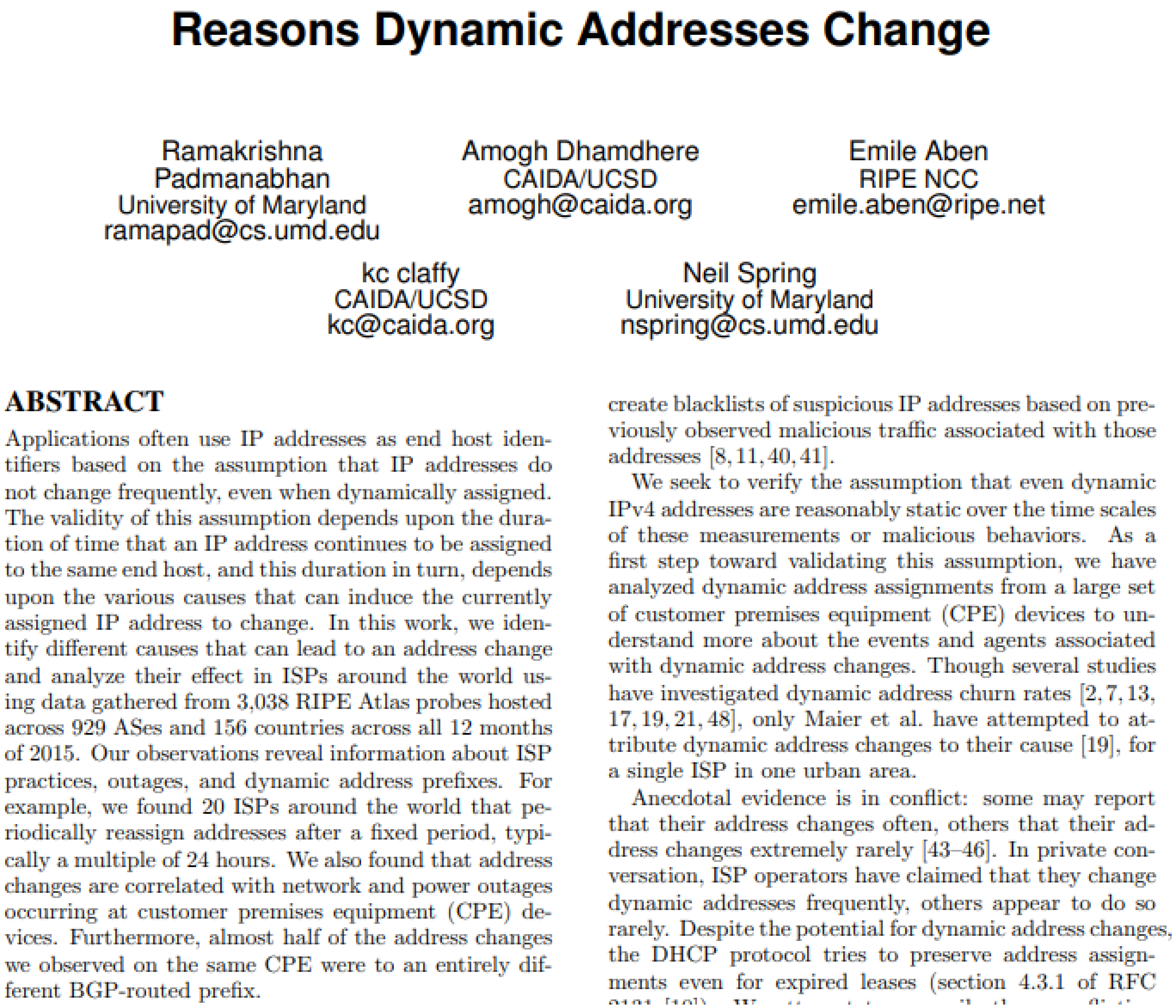


Overestimation

Lower bound

IP Addresses change

IP Addresses change predictably



Reasons Dynamic Addresses Change

Ramakrishna Padmanabhan
University of Maryland
ramapad@cs.umd.edu

Amogh Dhamdhere
CAIDA/UCSD
amogh@caida.org

Emile Aben
RIPE NCC
emile.aben@ripe.net

kc claffy
CAIDA/UCSD
kc@caida.org

Neil Spring
University of Maryland
nspring@cs.umd.edu

ABSTRACT

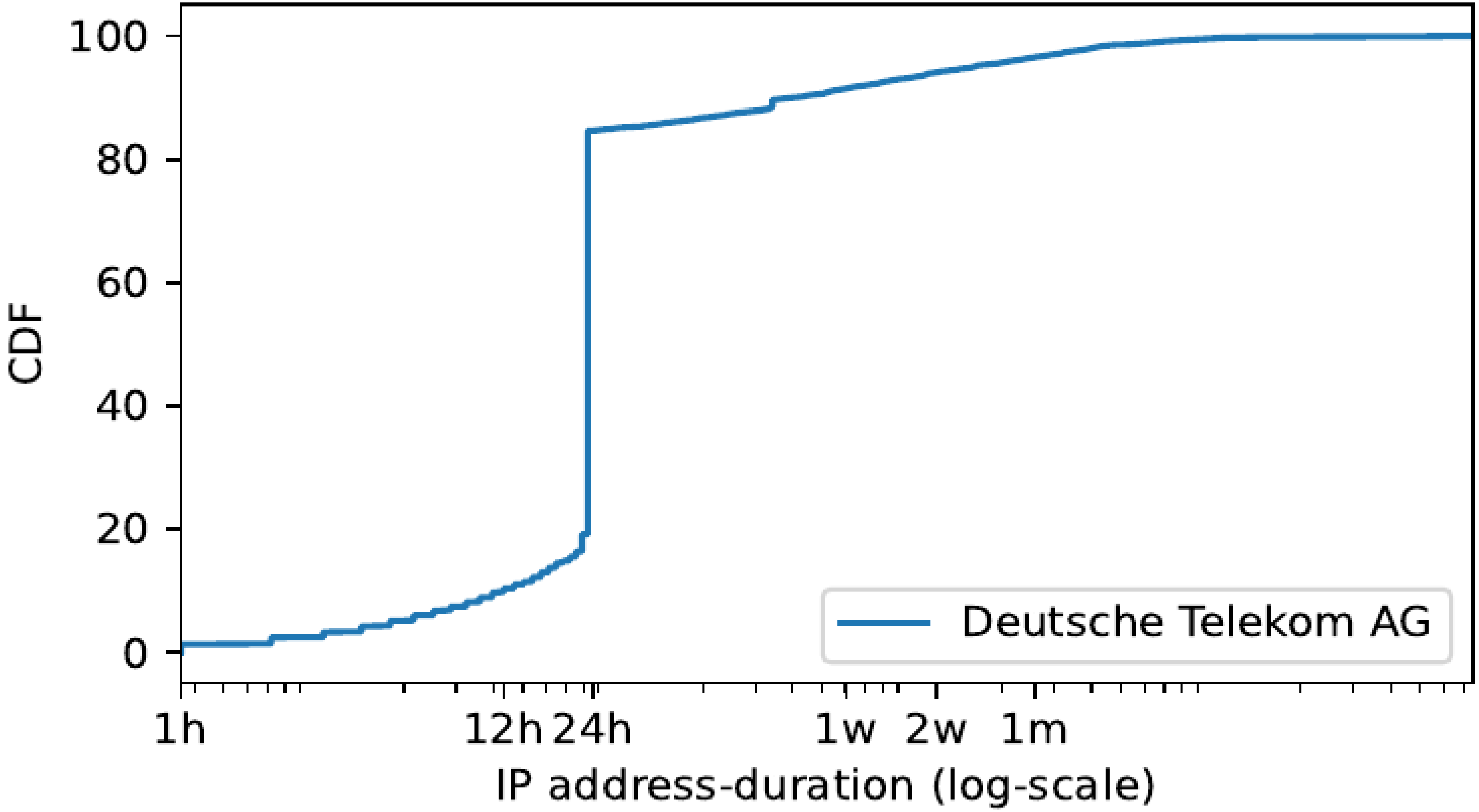
Applications often use IP addresses as end host identifiers based on the assumption that IP addresses do not change frequently, even when dynamically assigned. The validity of this assumption depends upon the duration of time that an IP address continues to be assigned to the same end host, and this duration in turn, depends upon the various causes that can induce the currently assigned IP address to change. In this work, we identify different causes that can lead to an address change and analyze their effect in ISPs around the world using data gathered from 3,038 RIPE Atlas probes hosted across 929 ASes and 156 countries across all 12 months of 2015. Our observations reveal information about ISP practices, outages, and dynamic address prefixes. For example, we found 20 ISPs around the world that periodically reassign addresses after a fixed period, typically a multiple of 24 hours. We also found that address changes are correlated with network and power outages occurring at customer premises equipment (CPE) devices. Furthermore, almost half of the address changes we observed on the same CPE were to an entirely different BGP-routed prefix.

create blacklists of suspicious IP addresses based on previously observed malicious traffic associated with those addresses [8, 11, 40, 41].

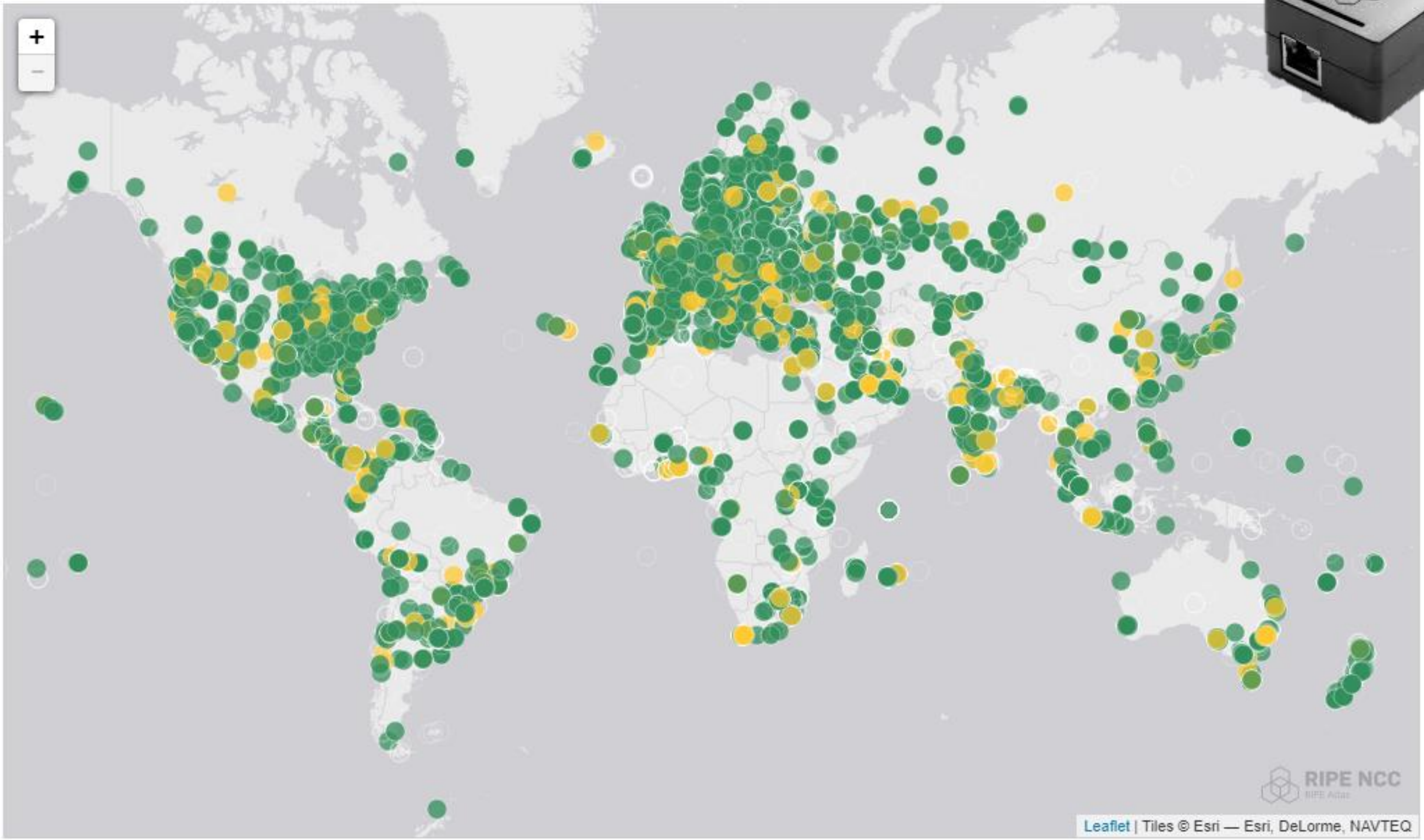
We seek to verify the assumption that even dynamic IPv4 addresses are reasonably static over the time scales of these measurements or malicious behaviors. As a first step toward validating this assumption, we have analyzed dynamic address assignments from a large set of customer premises equipment (CPE) devices to understand more about the events and agents associated with dynamic address changes. Though several studies have investigated dynamic address churn rates [2, 7, 13, 17, 19, 21, 48], only Maier et al. have attempted to attribute dynamic address changes to their cause [19], for a single ISP in one urban area.

Anecdotal evidence is in conflict: some may report that their address changes often, others that their address changes extremely rarely [43–46]. In private conversation, ISP operators have claimed that they change dynamic addresses frequently, others appear to do so rarely. Despite the potential for dynamic address changes, the DHCP protocol tries to preserve address assignments even for expired leases (section 4.3.1 of RFC

IP Reassignments

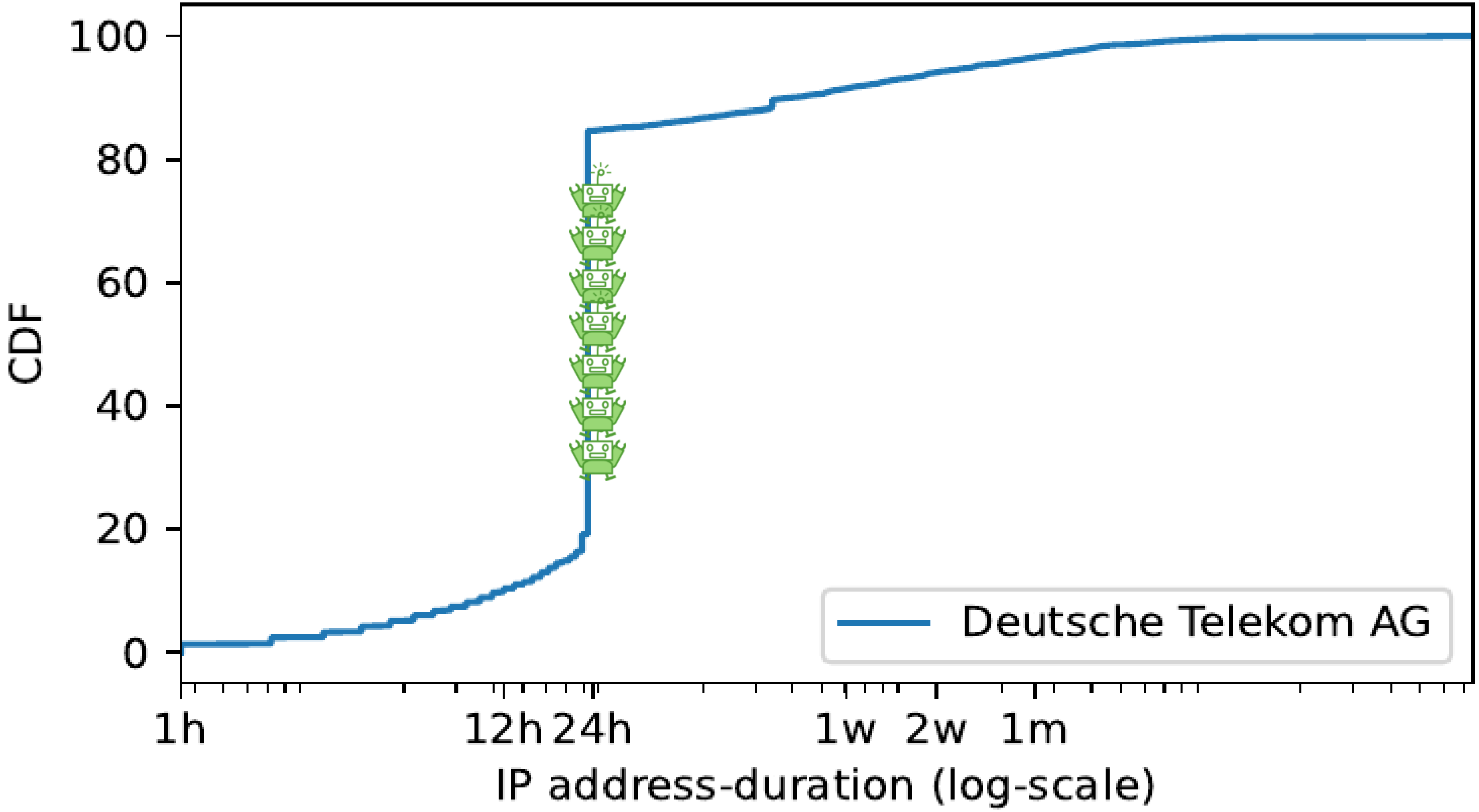


RIPE Atlas

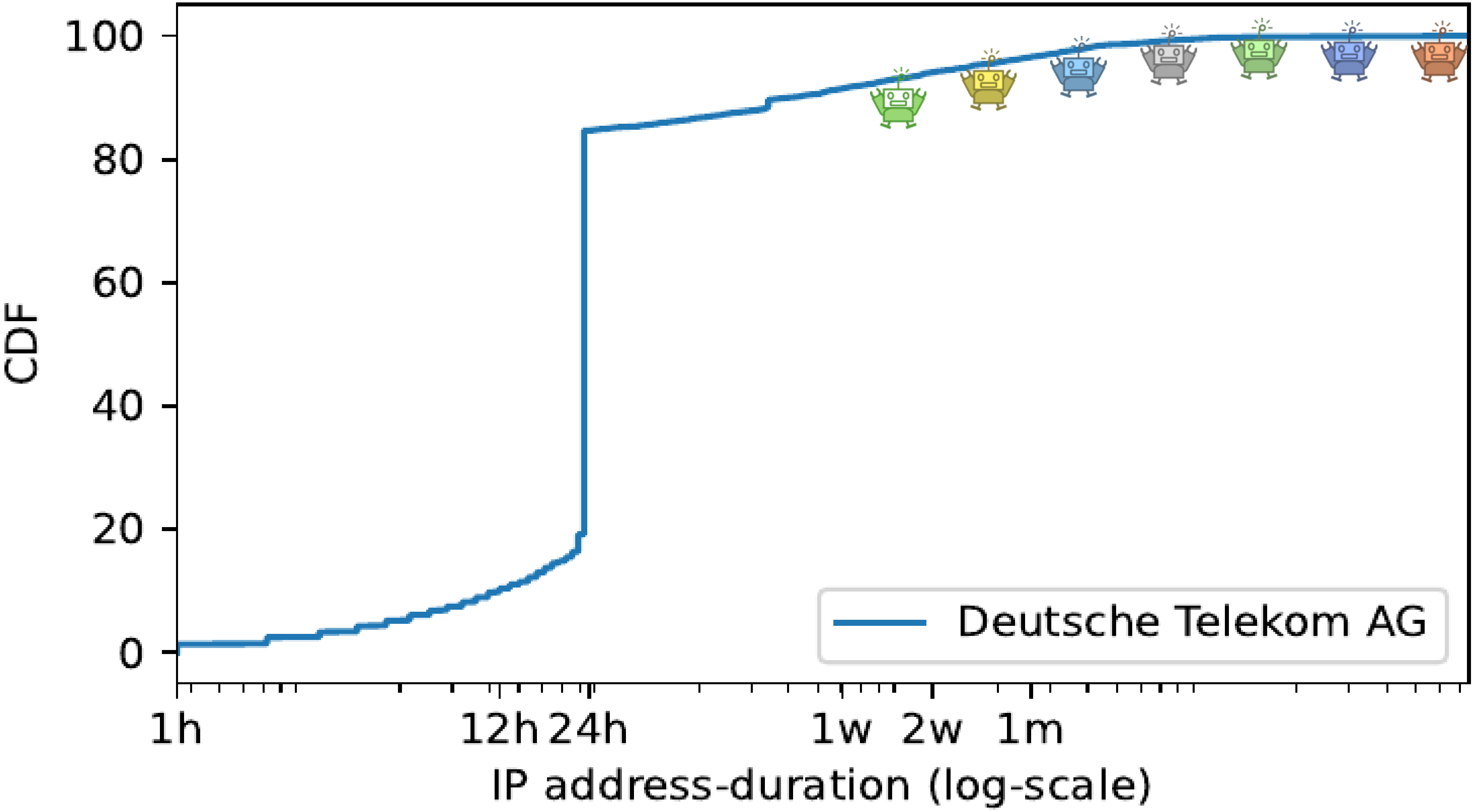

+
−
RIPE NCC
RIPE Atlas
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

RIPE NCC

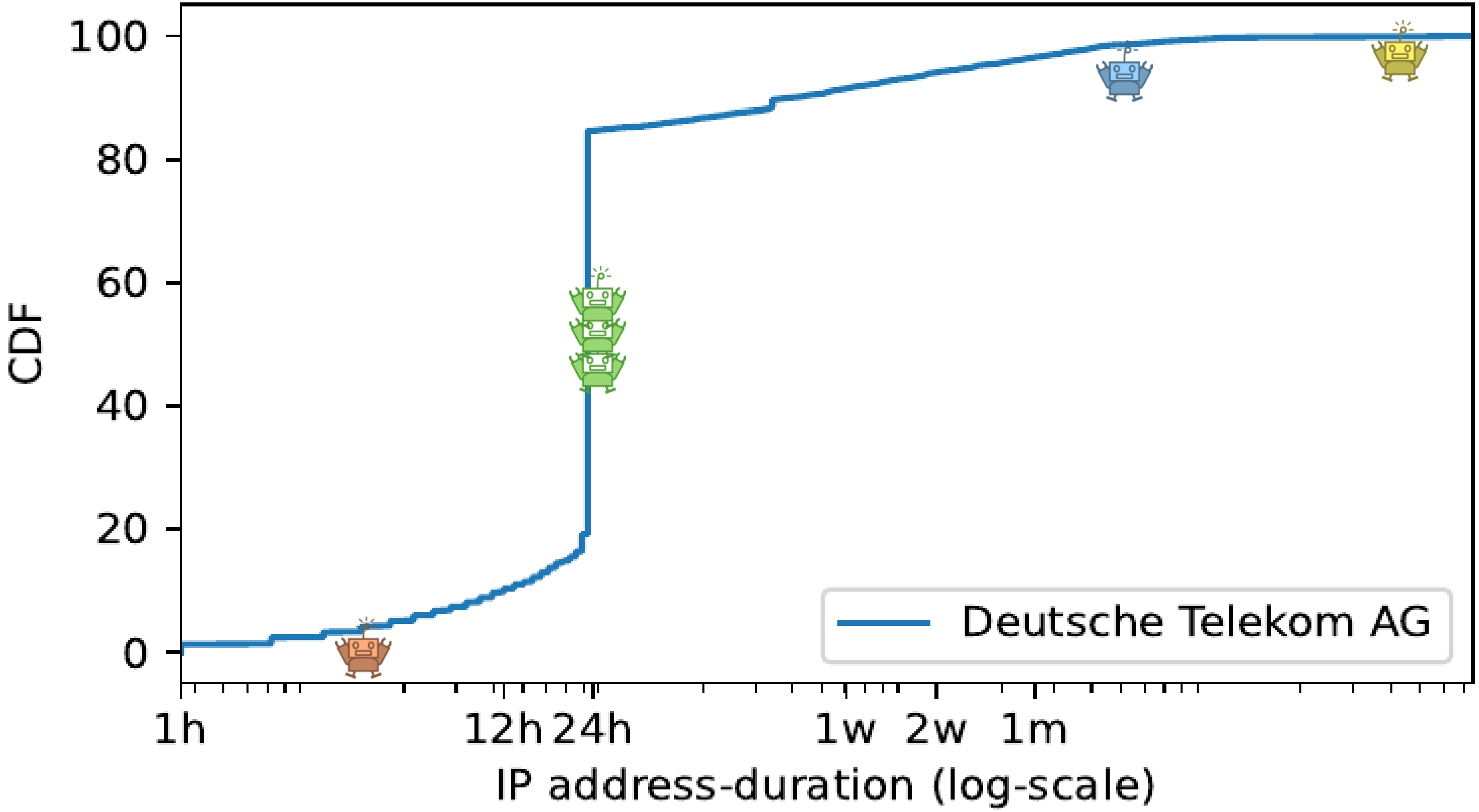
Example – 7 IP addresses over 1 week



Example – 7 IP addresses over 1 week



Example – 7 IP addresses over 1 week



CARDCount

Unique IP addresses

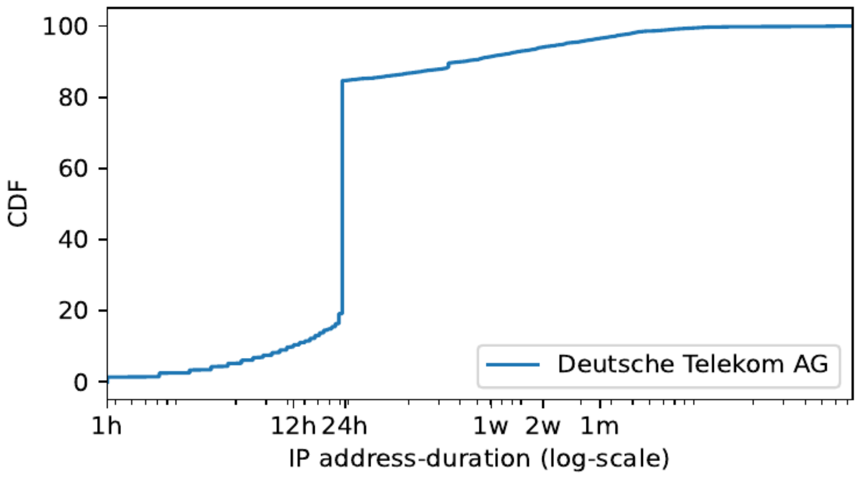
Measurement duration

$$CARDCount(A, D, T) = A * \frac{1}{n} * \sum_{i=1}^{n} \frac{d}{d + T}$$

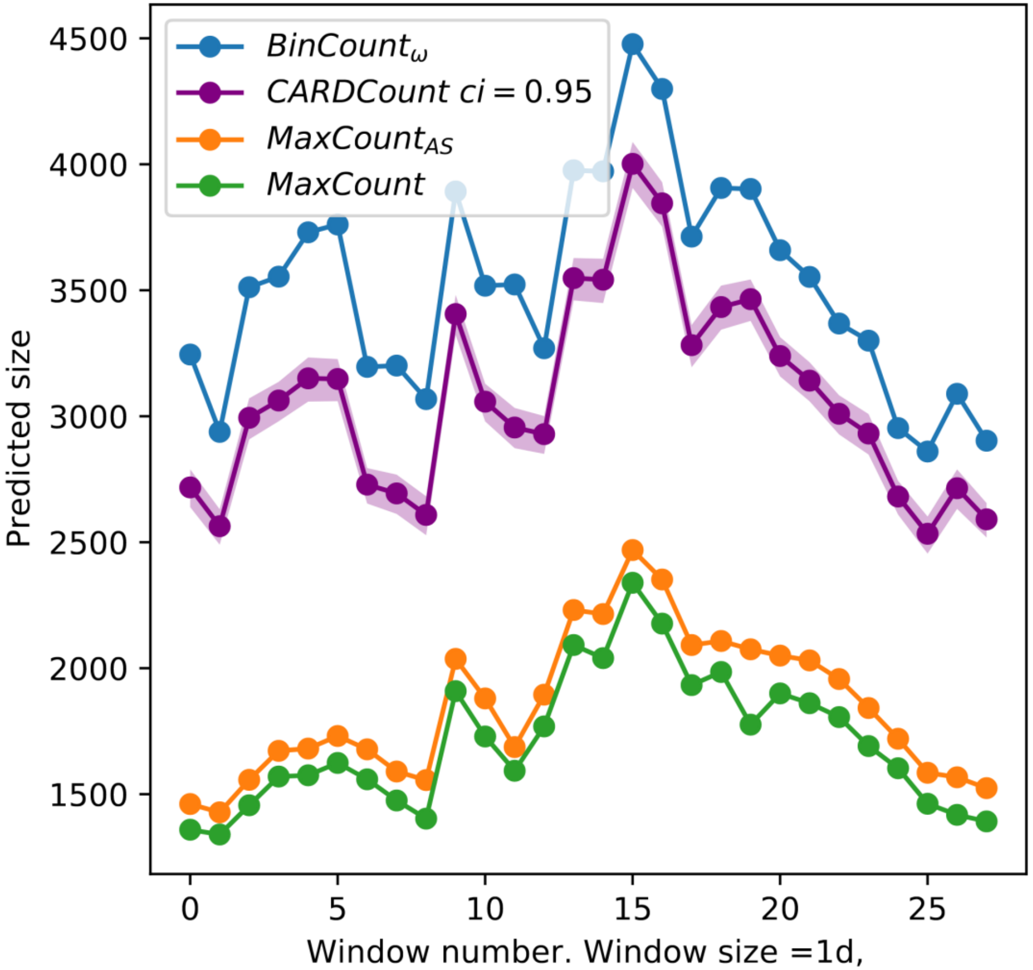
IP duration distribution

Example: 7 IPs over 7 days

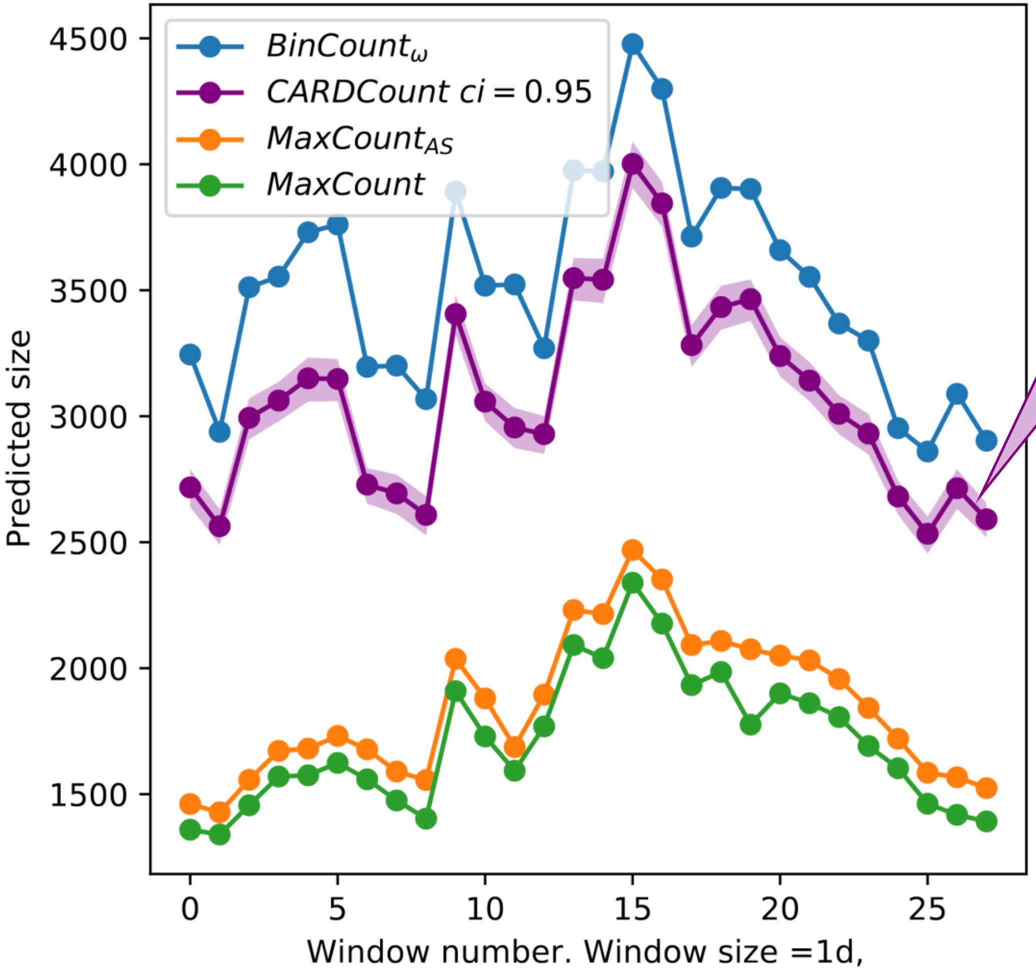
CARDCount(7, D_{DTAG}, 7d) = 1.59



Mirai Botnet Size



Mirai Botnet Size

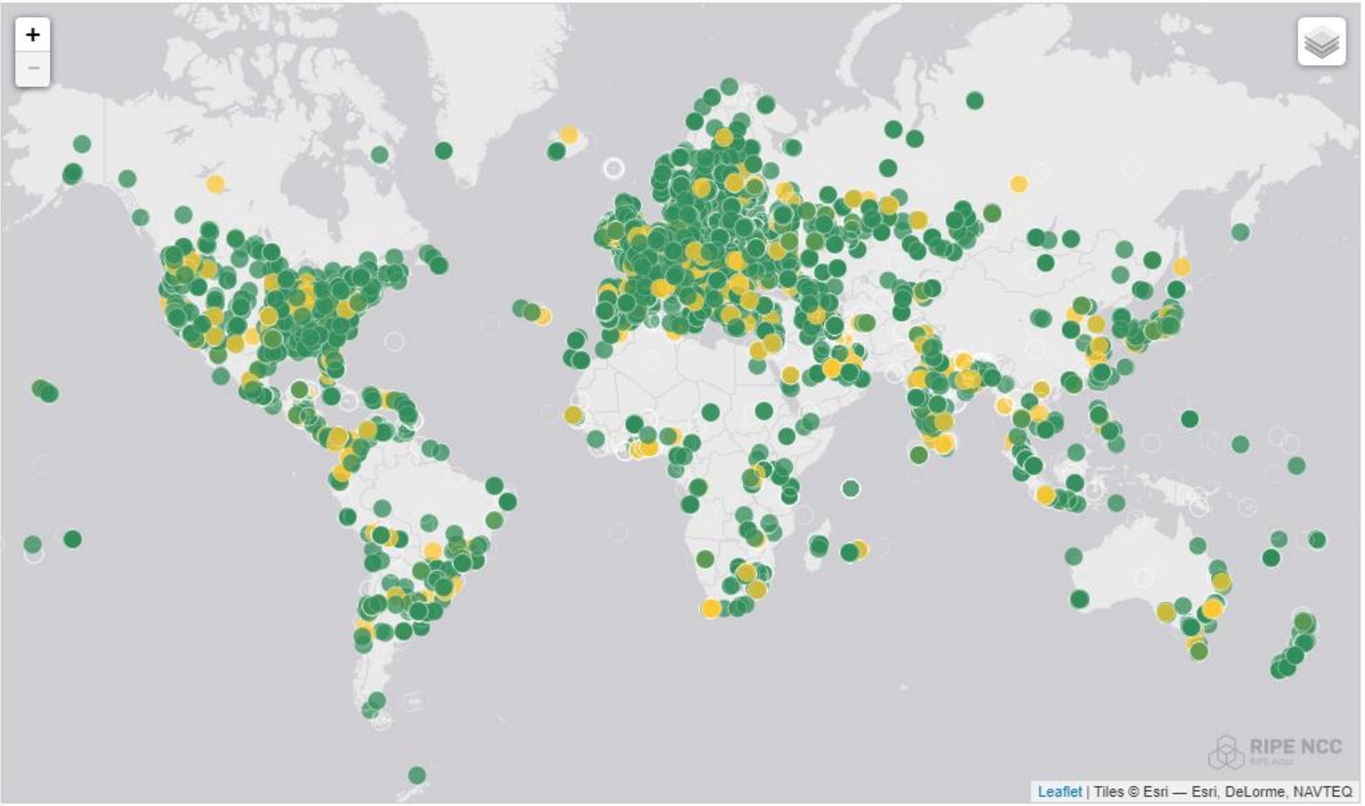


Confidence Intervals

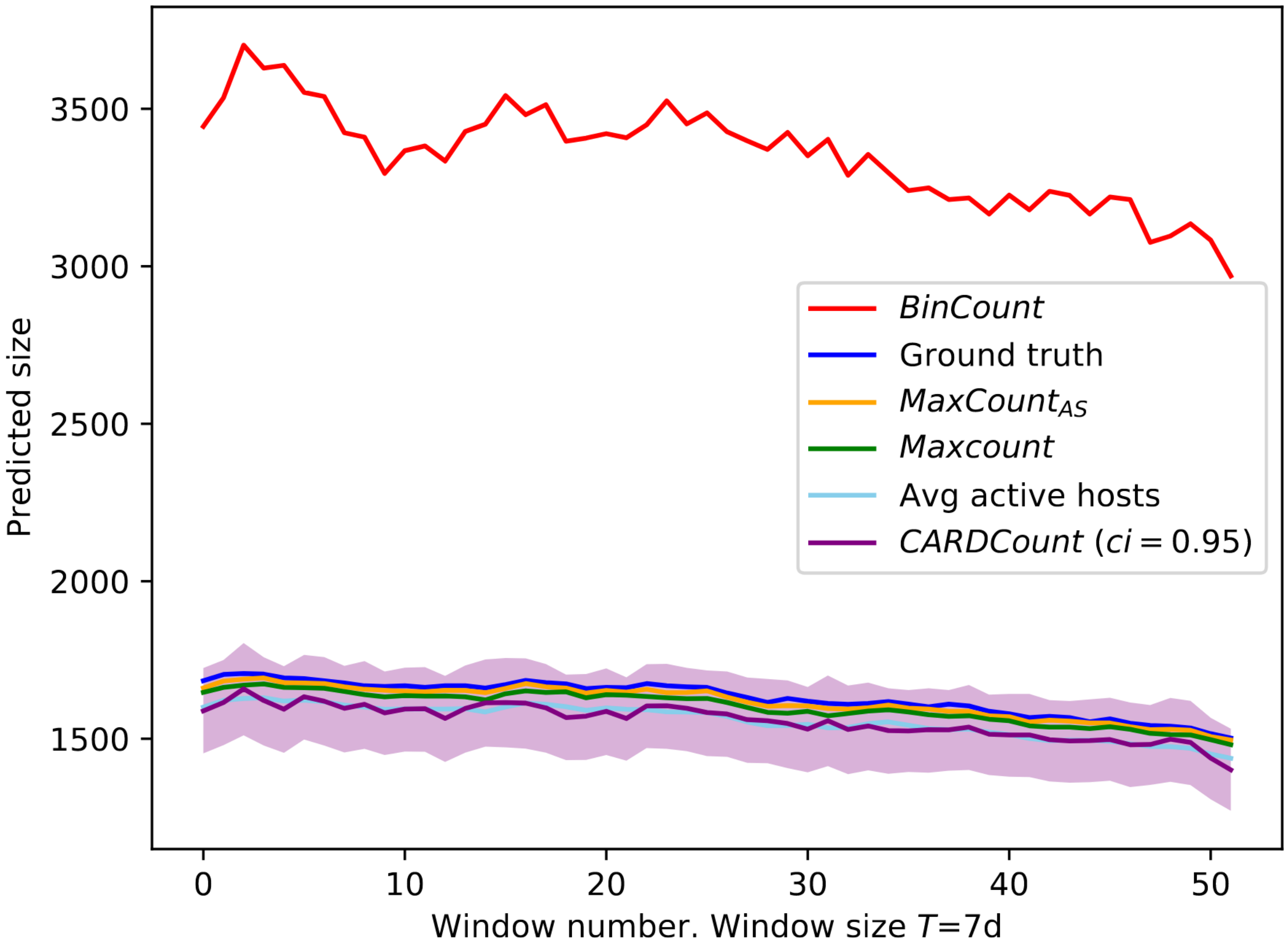
How accurate is CARDCount

Ground truth evaluation on RIPE Atlas

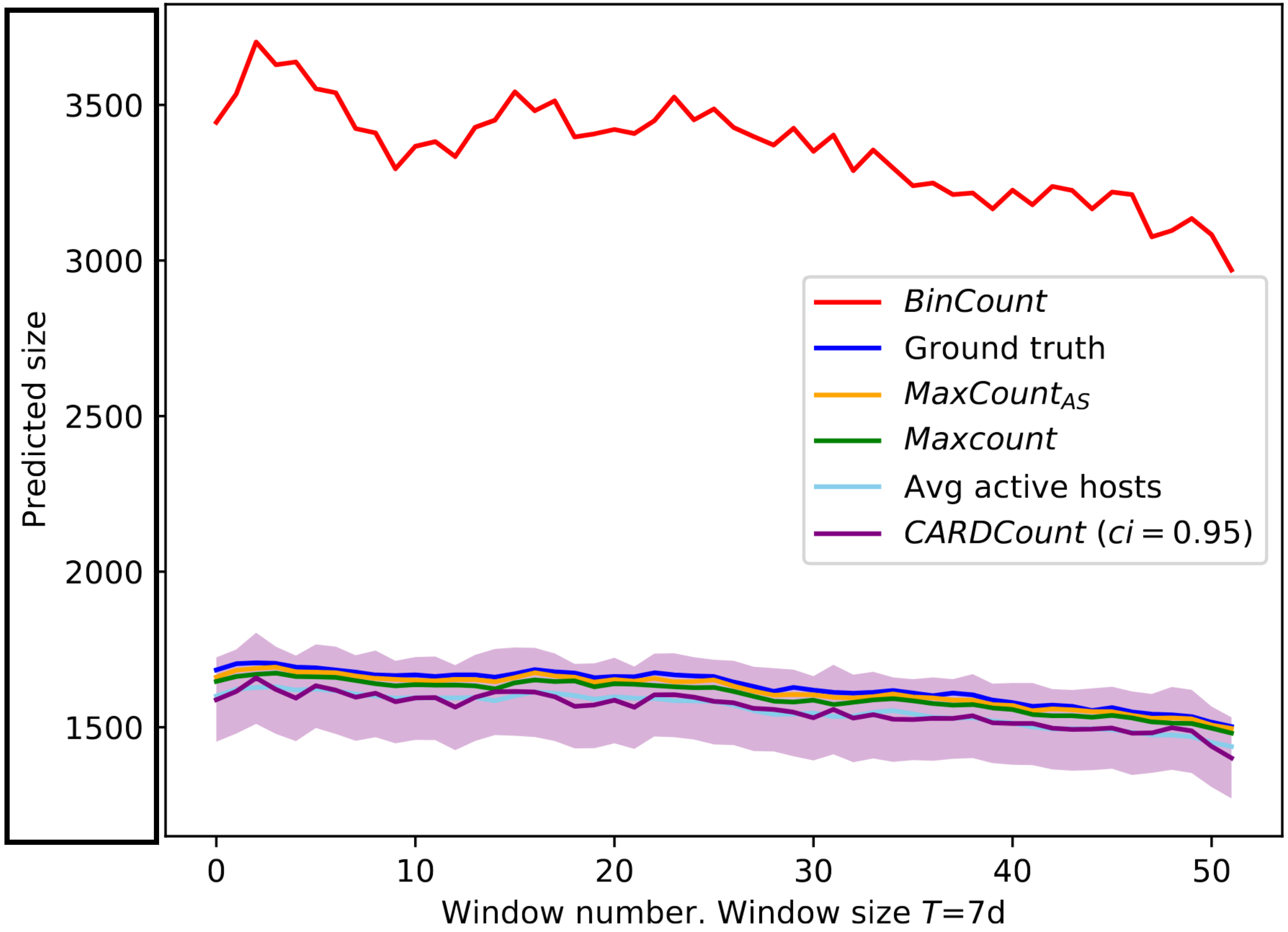
~1800 Devices
39 ASes
Precise IP duration
Ground truth activity



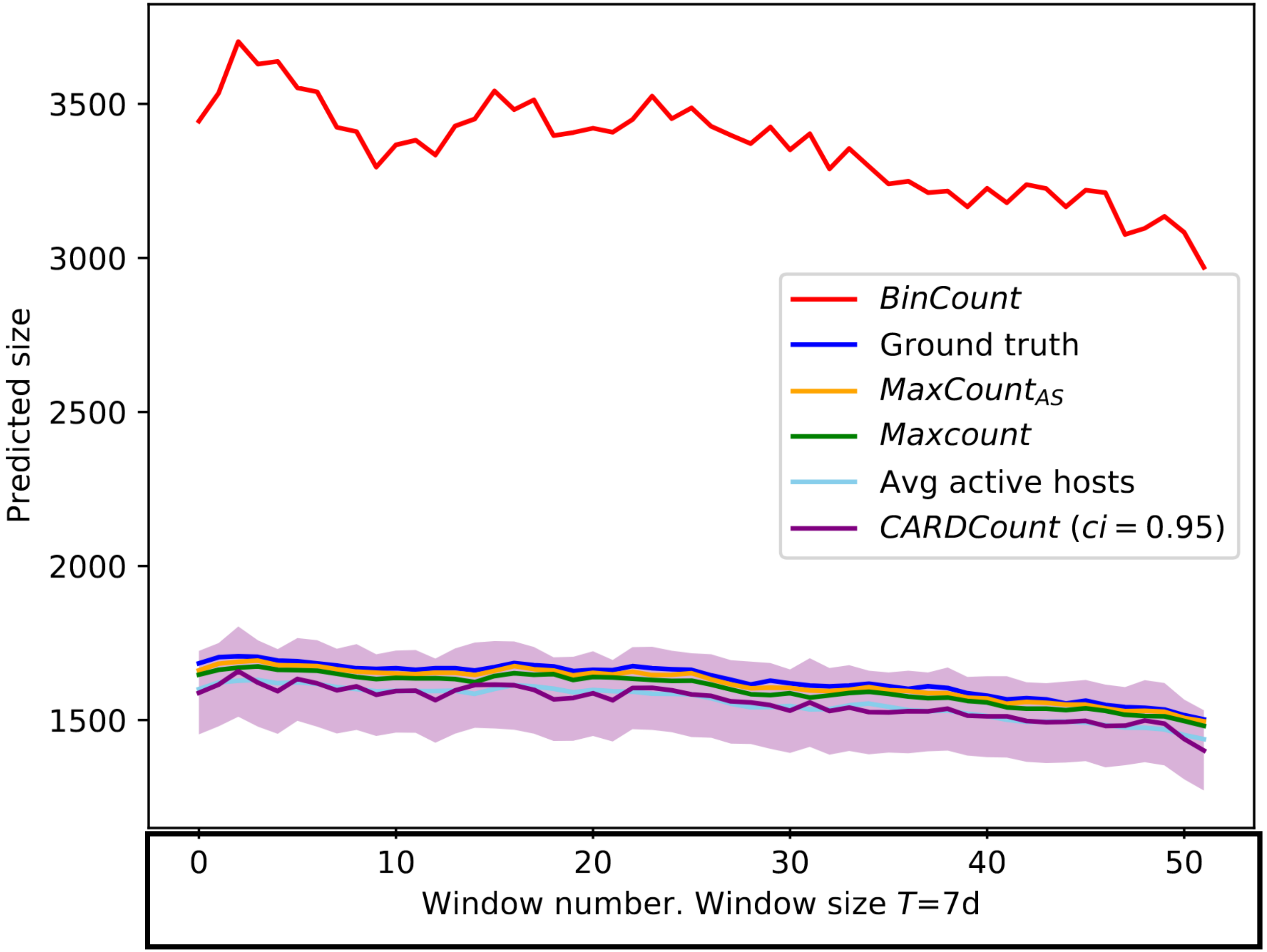
Ground Truth – RIPE Atlas



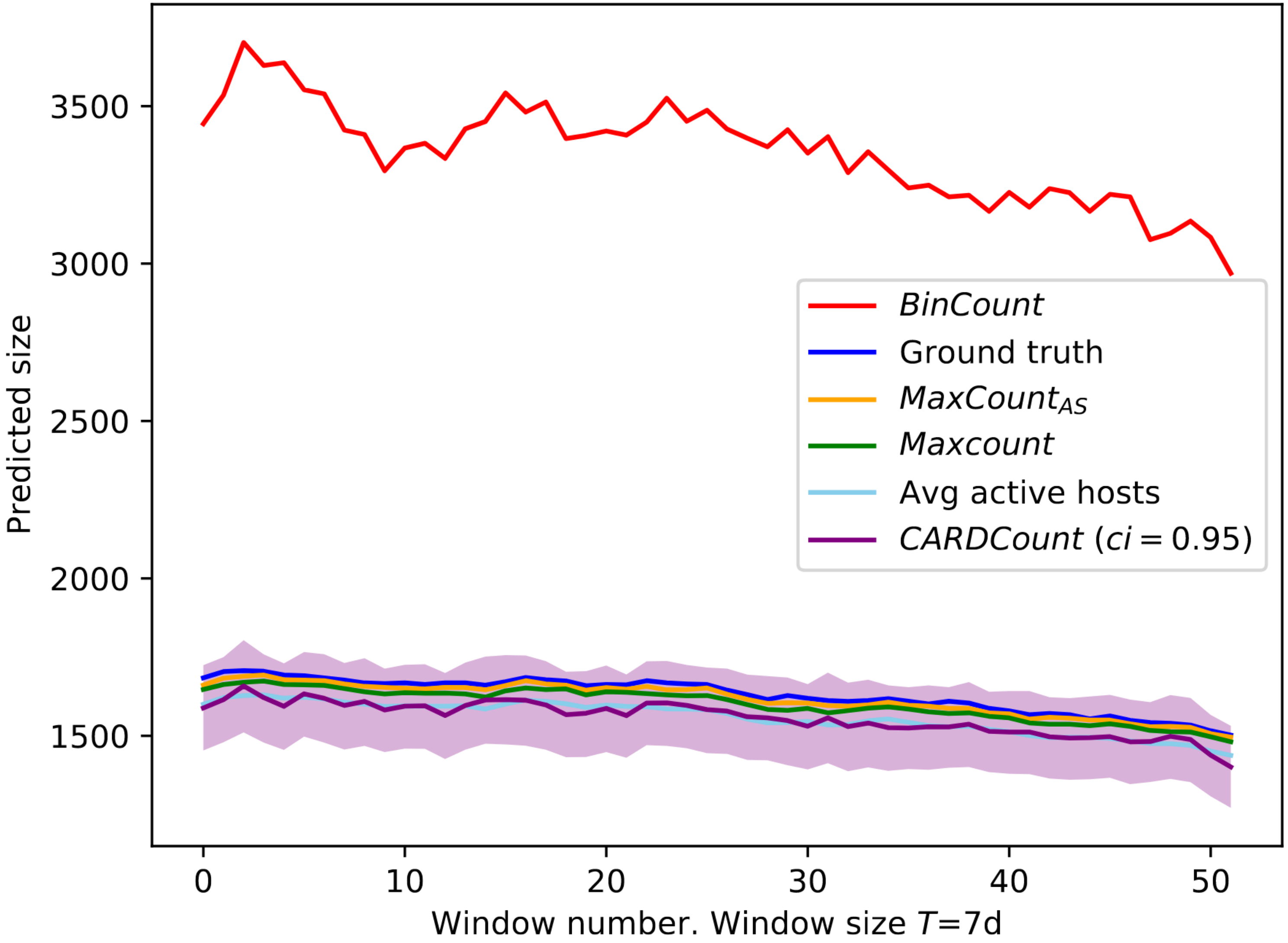
Ground Truth – RIPE Atlas



Ground Truth – RIPE Atlas



Ground Truth – RIPE Atlas



Confounding Factors

Short IP address durations

Bot churn

Capturing partial bot activity

Accuracy of the address duration distributions

Shared IP addresses

Confounding Factors

Short IP address durations

Bot churn

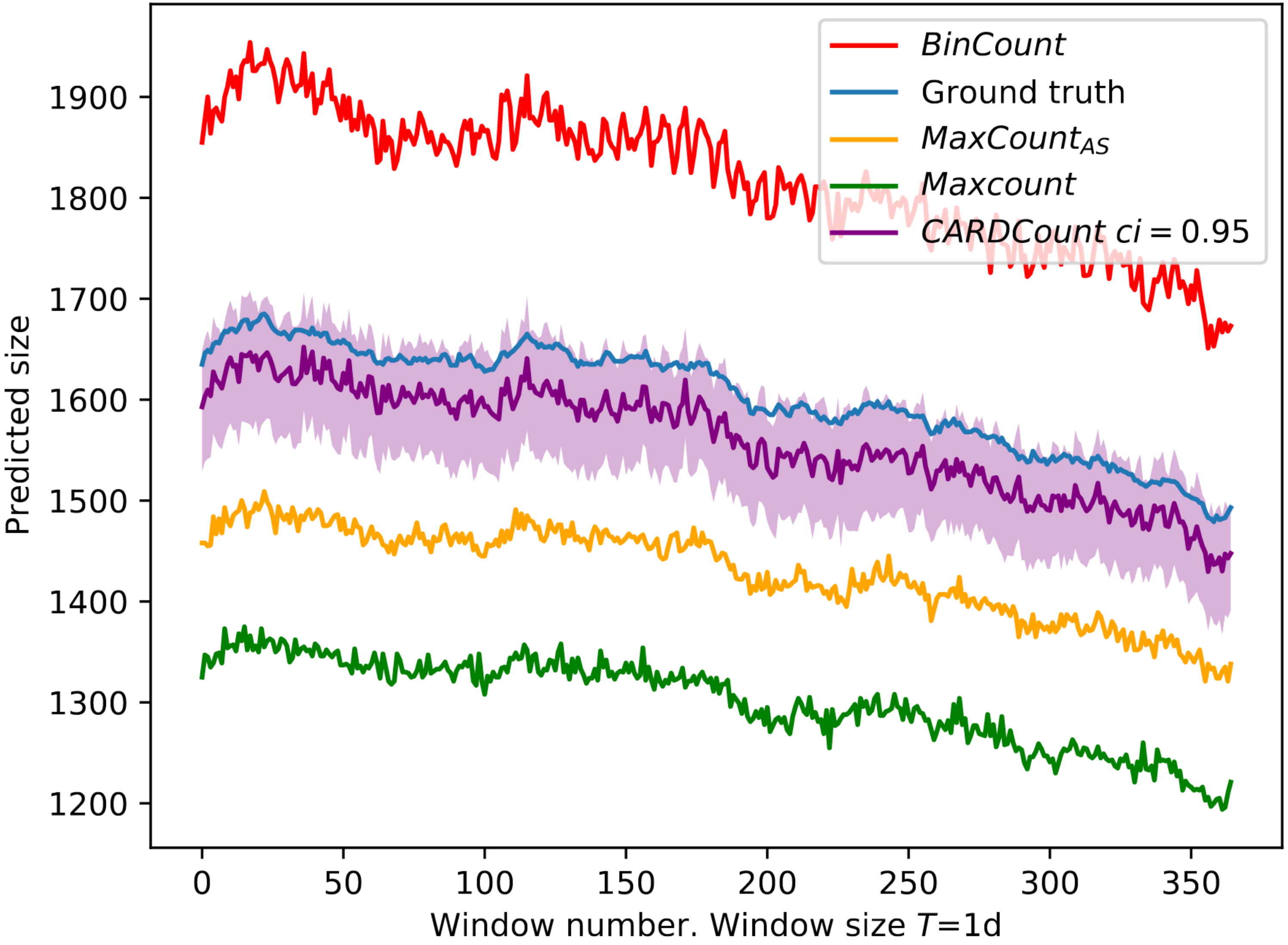
Capturing partial bot activity

Incomplete data

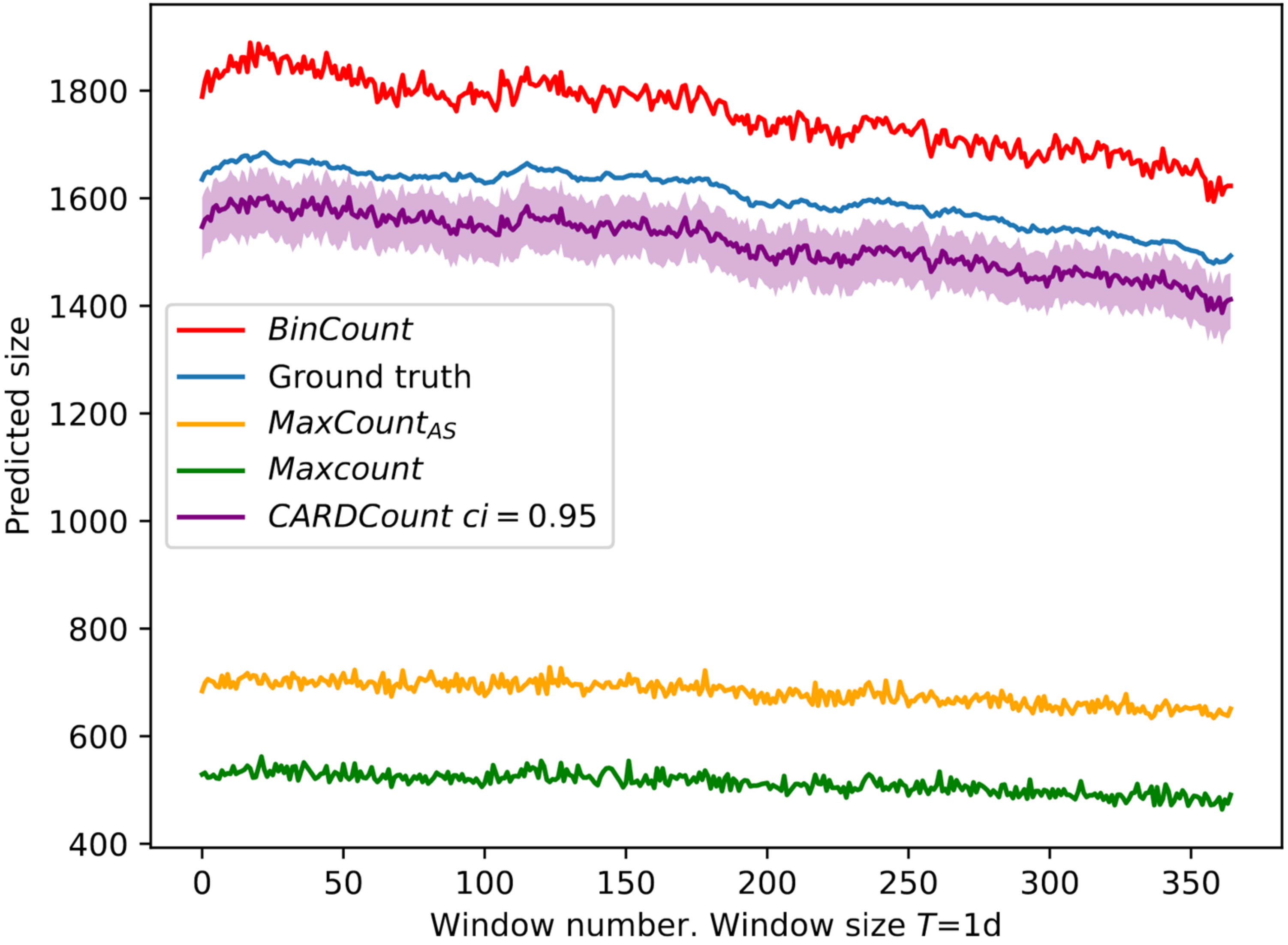
Accuracy of the address duration distributions

Shared IP addresses

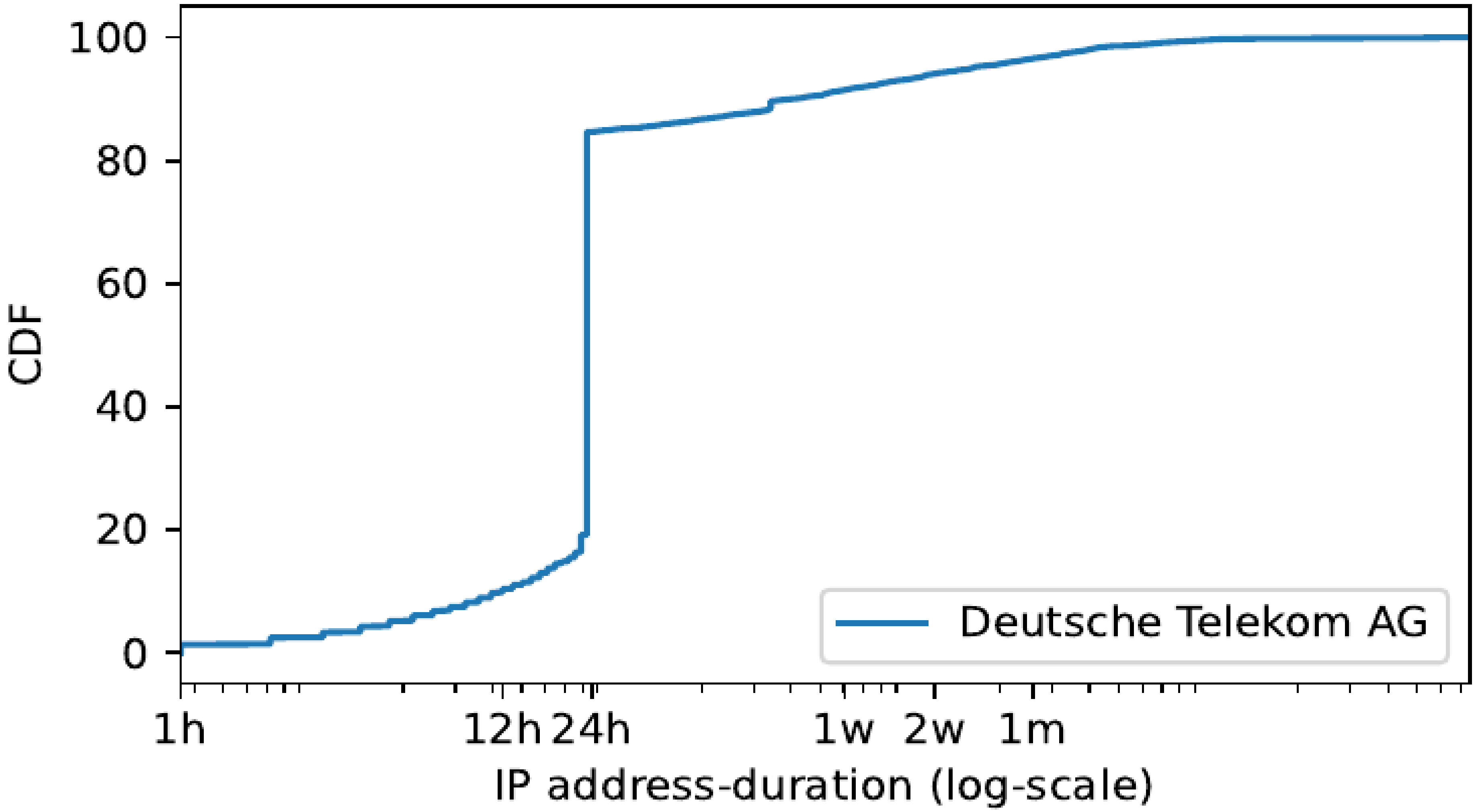
Dealing with incomplete data (80%)



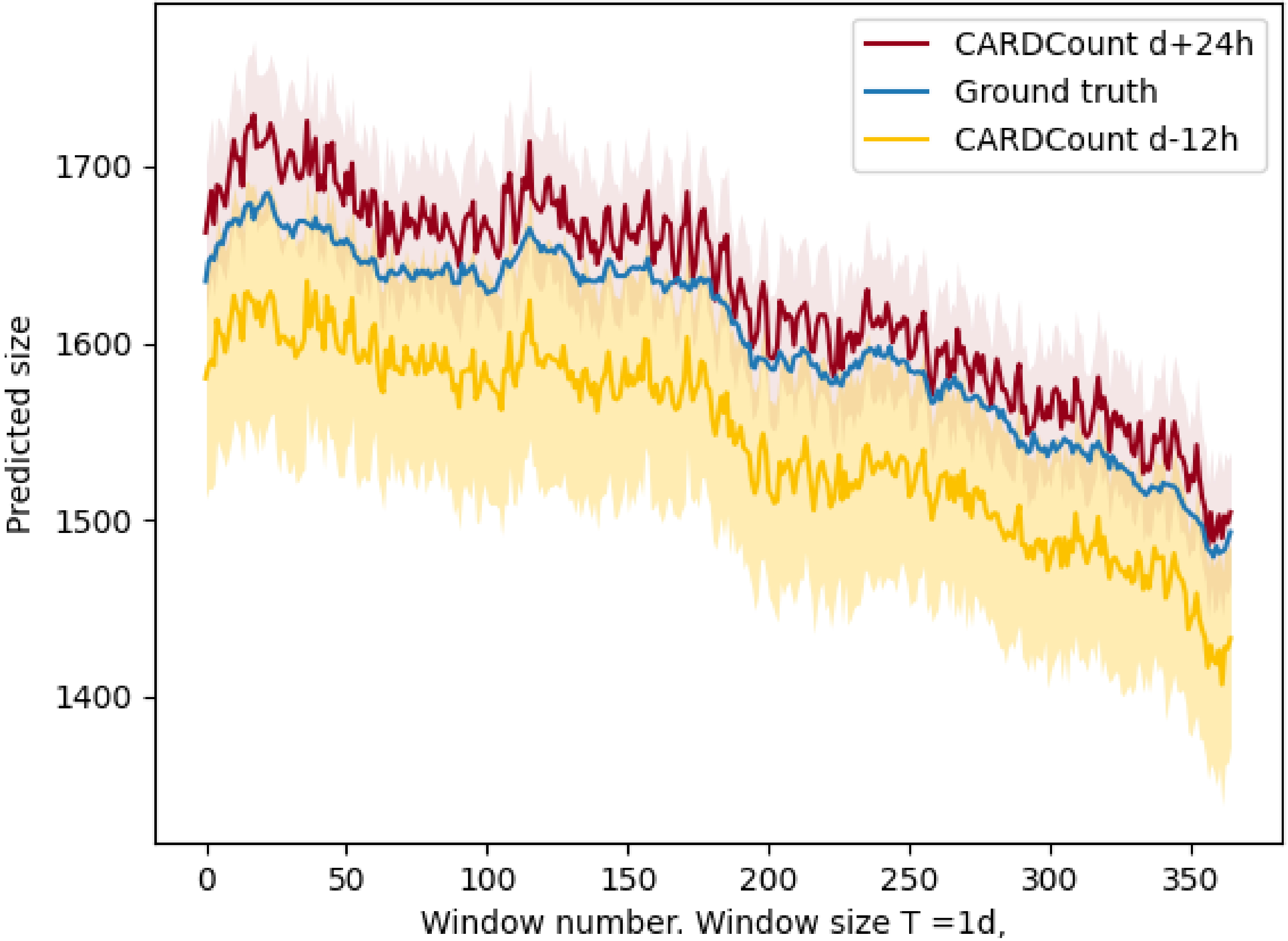
Dealing with incomplete data (30%)



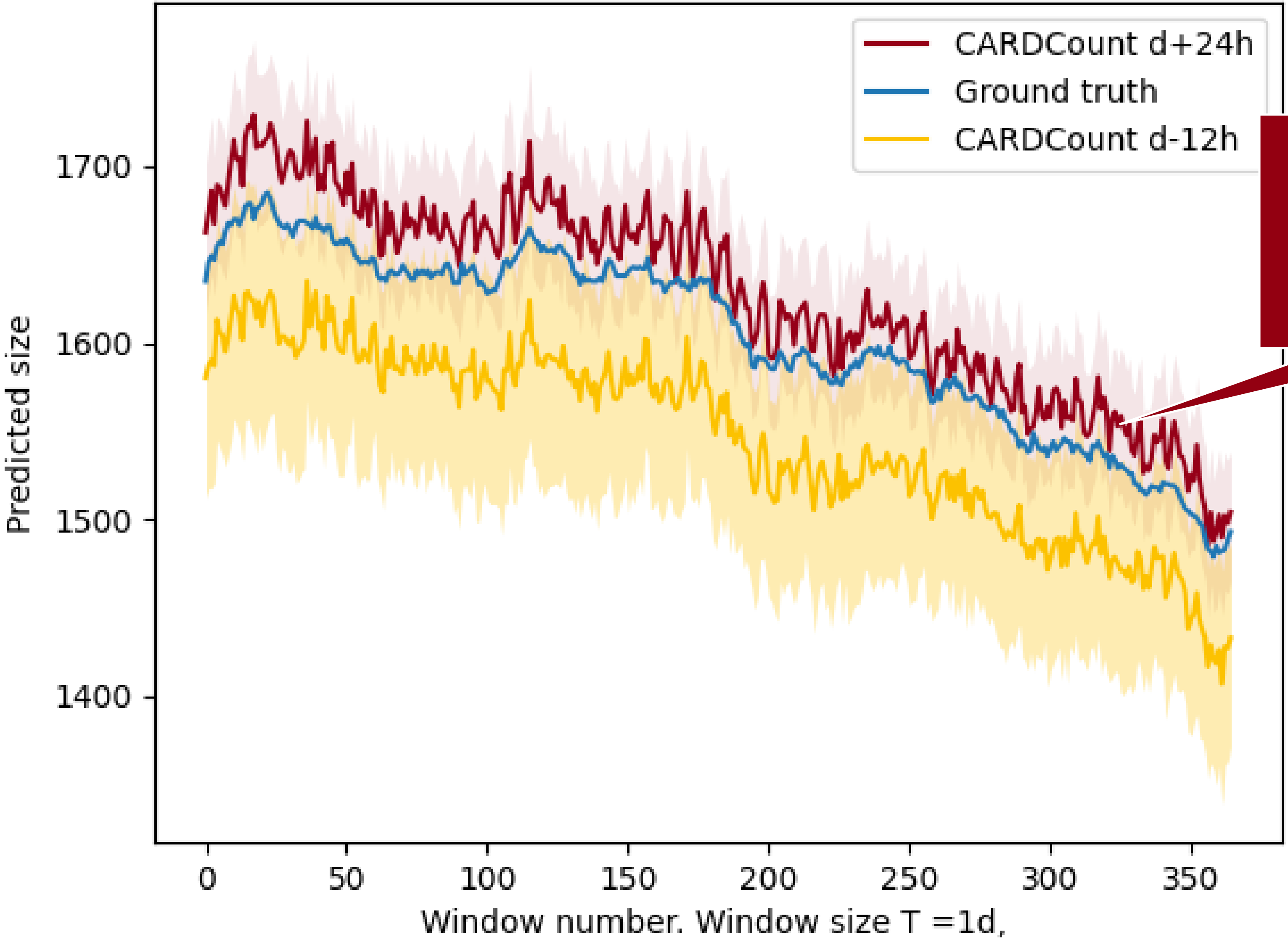
Accuracy of Address Duration Distributions



Accuracy of Address Duration Distributions

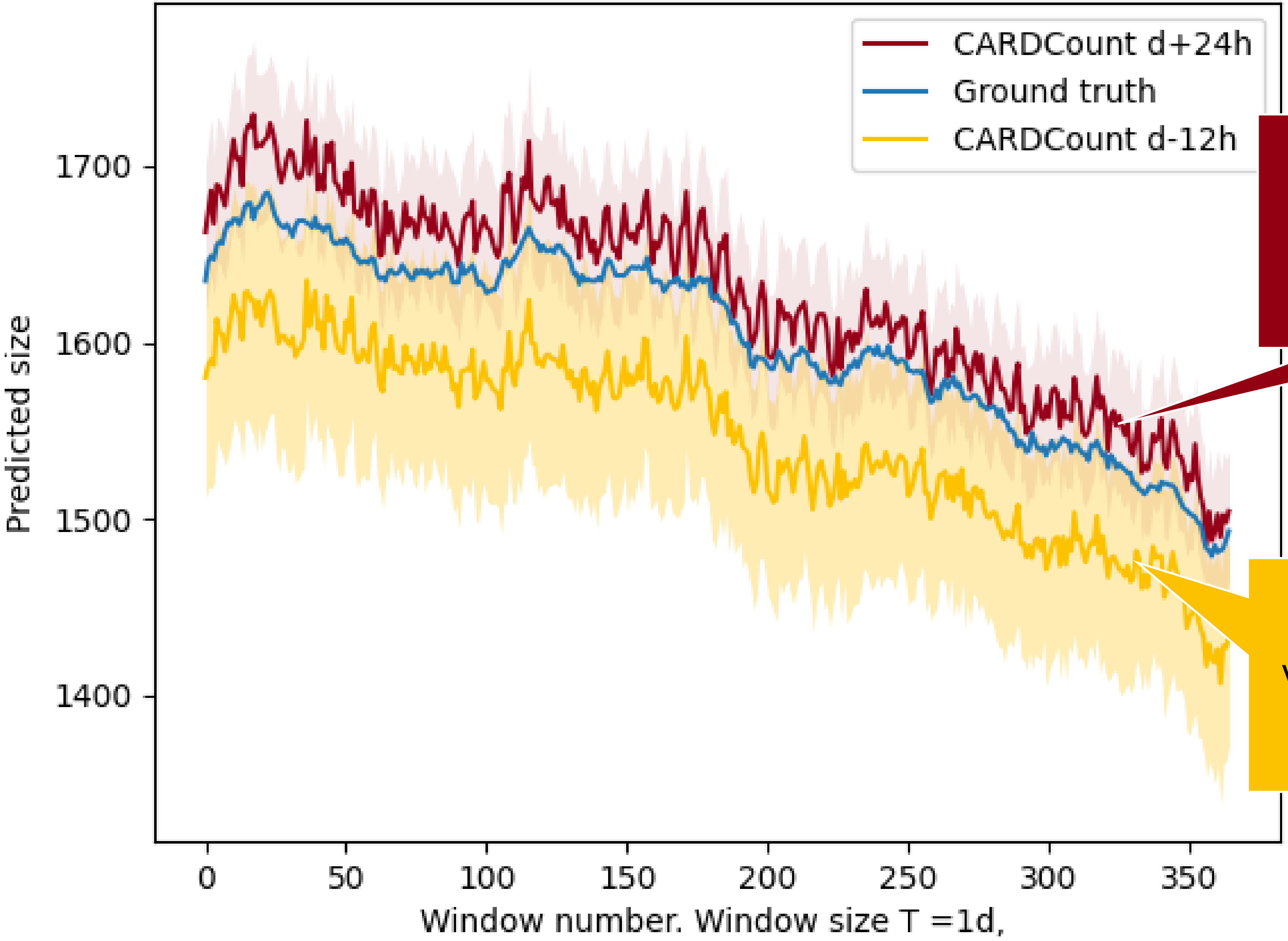


Accuracy of Address Duration Distributions



$\forall d \in D, d = d + 24h$

Accuracy of Address Duration Distributions



$\forall d \in D, d = d + 24h$

$\forall d \in D, d = d - 12h$

Confounding Factors

Short IP address durations

Bot churn

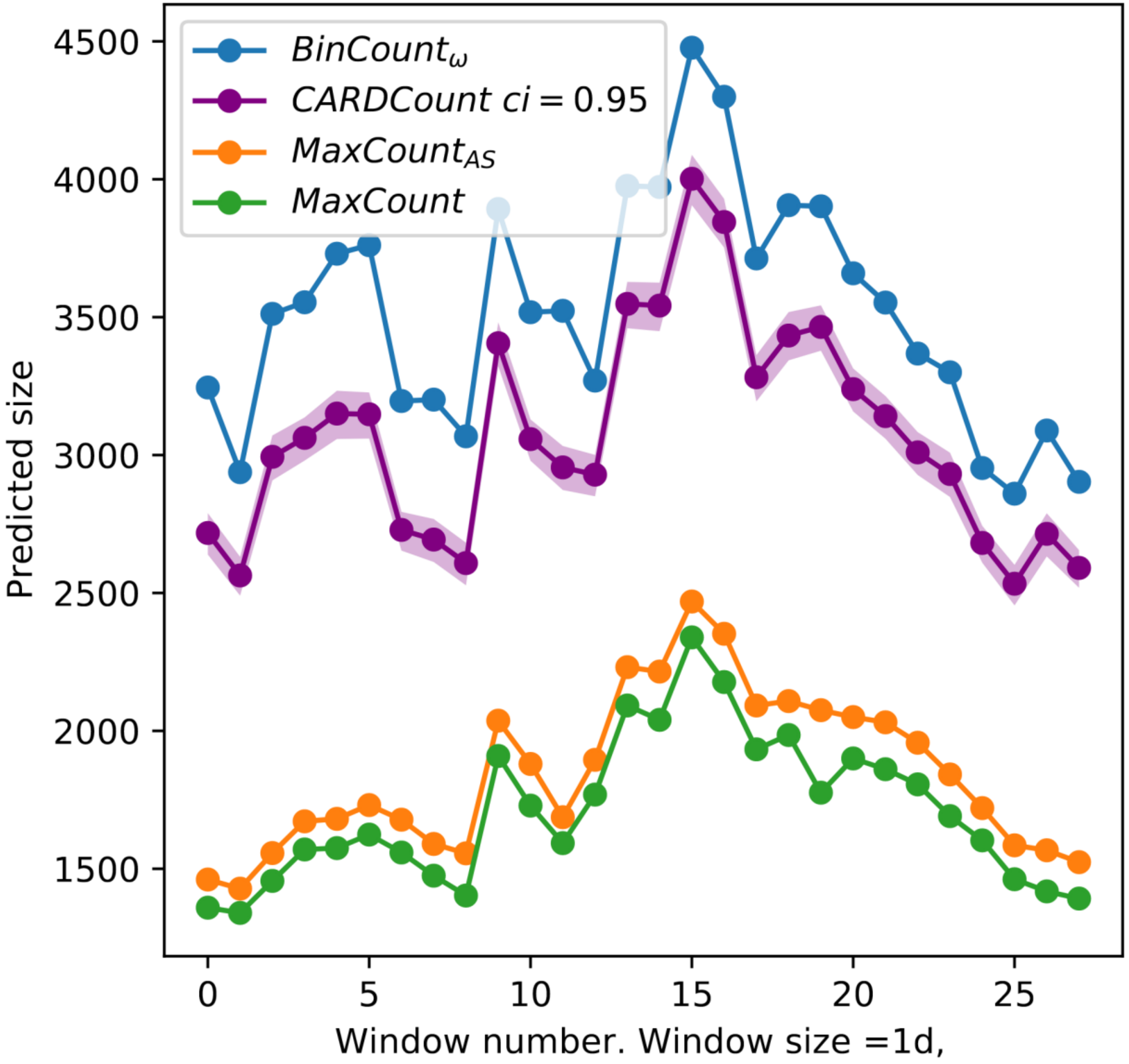
Capturing partial bot activity

Accuracy of the address duration distributions

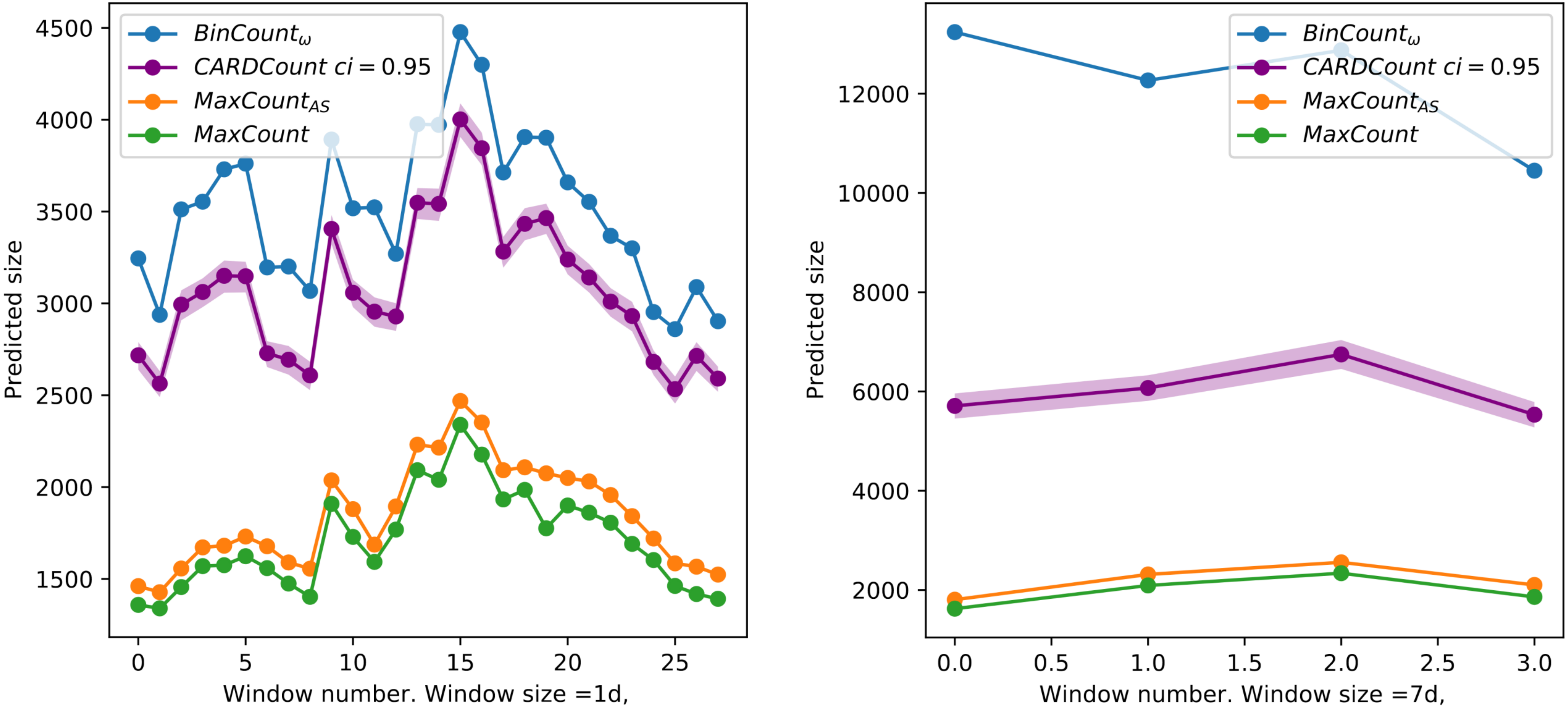
Shared IP addresses

CARDCount is most accurate under realistic conditions

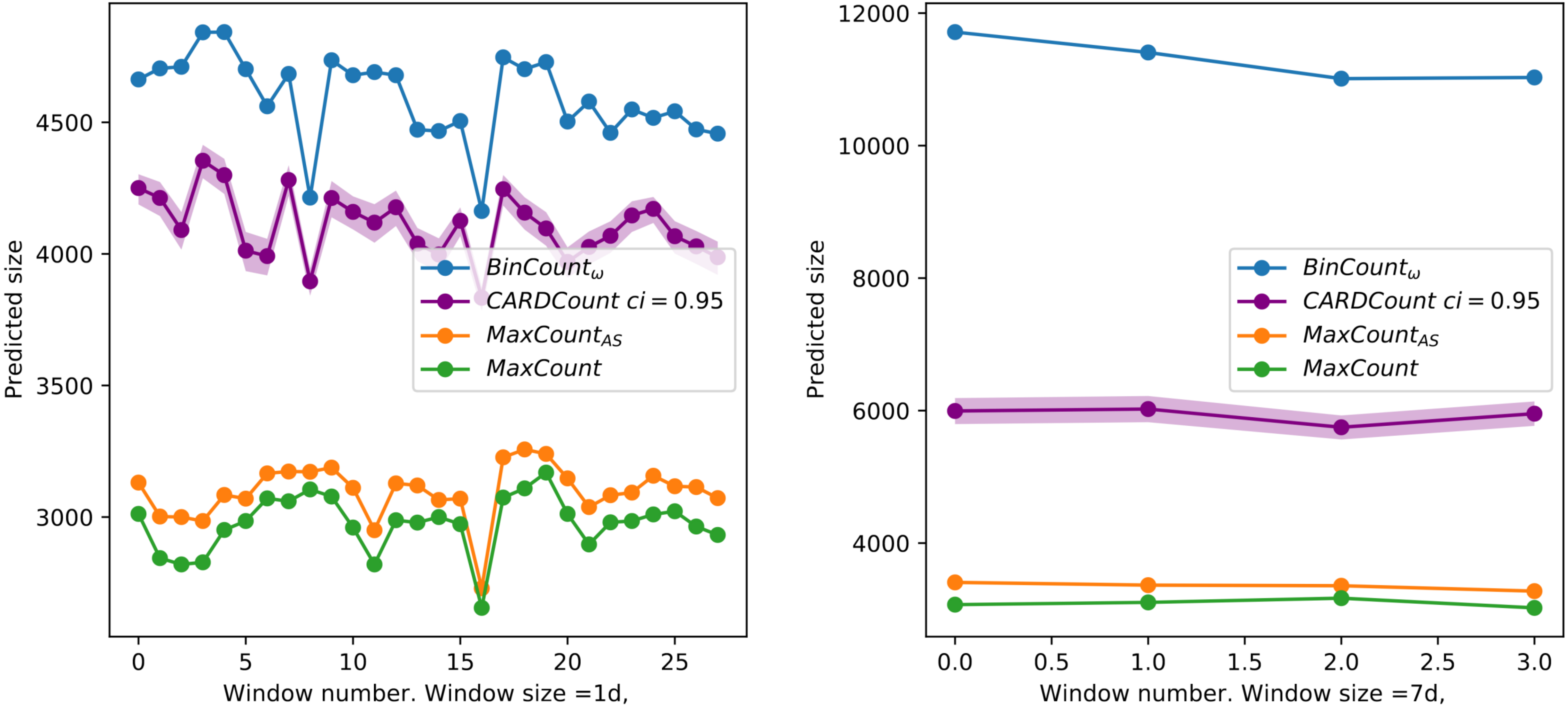
Mirai Botnet



Mirai Botnet



Hajime Botnet



Conclusion

CARDCount provides better size estimation

Relies on IP duration distributions

- Sign up for RIPE
- Convince ISPs to share distributions

Code: https://github.com/cardcount

Contact:

- boeck@tk.tu-darmstadt.de
- @lboeck@infosec.exchange
- @_LeonBock