Unmasking the Shadows: Understanding and Detecting Residential IP Proxies

Elisa Chiapponi, Martynas Bouzis, Ruediger Zachmann, Olivier Thonnard
Amadeus IT Group

Aviation ISAC Q2 AvTech – 2024
Lufthansa Training & Conference Center, Seeheim, Germany

amadeus

amadeus
Global Security Operations

Funded by
the European Union
NextGenerationEU

Supported by:
Federal Ministry
for Economic Affairs
and Climate Action

on the basis of a decision
by the German Bundestag

Who am I

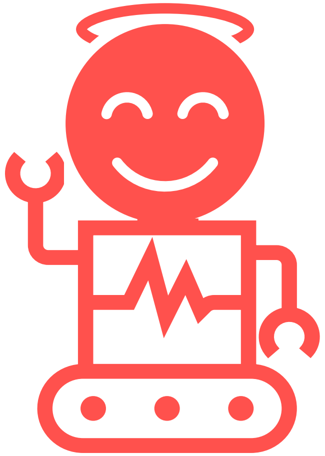
- Security Researcher in the Global Security Operations of Amadeus
  - Protection of web domains linked to the travel industry
  - Expertise in Network and Application Security

- Work based on:
  - Ph.D. research and collaborations
  - RESCUE – Resilient Cloud for Europe, IPCEI – Germany

Funded by
the European Union
NextGenerationEU

Supported by:
Federal Ministry
for Economic Affairs
and Climate Action

on the basis of a decision
by the German Bundestag



Sophisticated bots

Bots

Random user interactions generation

Mimic of human interaction to reach their goals

CAPTCHA solving

Examples of sophisticated bots attacks

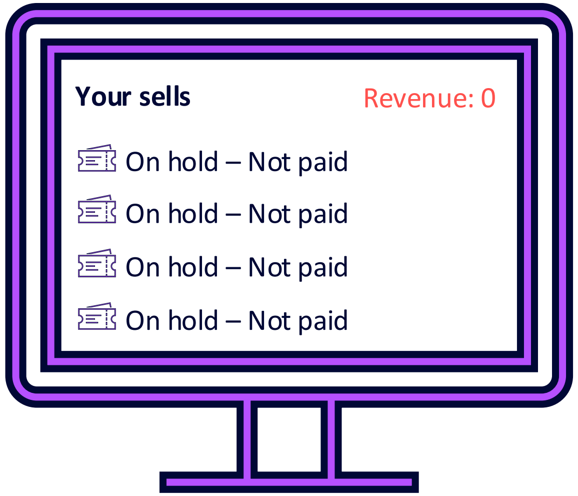
Web scraping

Content Reselling

Illicit Aggregators

Requests
Purchases
Revenues

Slow connections



Your sells Revenue: 0

On hold – Not paid

On hold – Not paid

On hold – Not paid

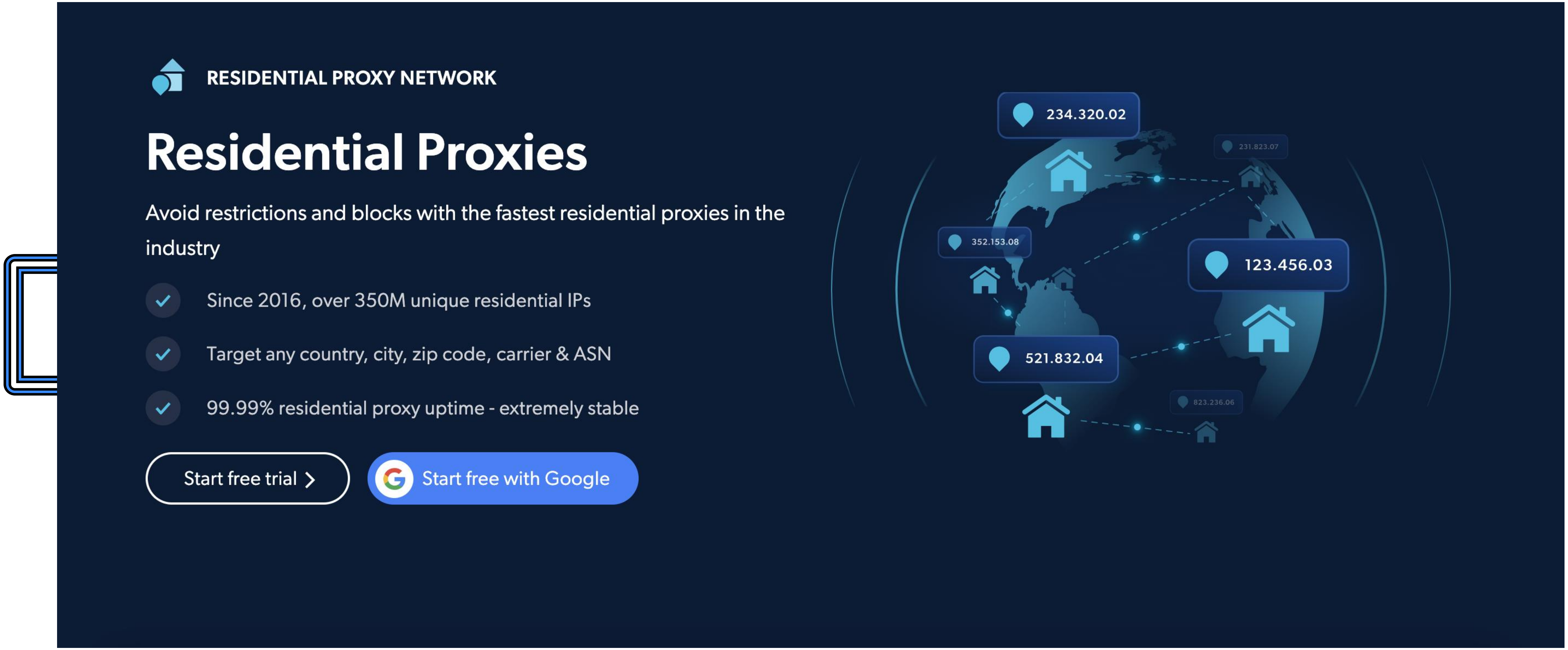
On hold – Not paid

Denial Of Inventory

Artificial Price Increase-Decrease

“Application Layer DDoS”

Arms race



RESIDENTIAL PROXY NETWORK

Residential Proxies

Avoid restrictions and blocks with the fastest residential proxies in the industry

- ✓ Since 2016, over 350M unique residential IPs
- ✓ Target any country, city, zip code, carrier & ASN
- ✓ 99.99% residential proxy uptime - extremely stable

Residential IP Proxies (RESIPs)

- Large networks of **residential devices** (smartphones, laptops, tablets,…)
- Devices **owned** by genuine users who **share** their usage
- No application layer information about being proxied
  - **Indistinguishable** from the requests sent directly by the residential devices at this layer
  - **High probability of false positives** for the traditional server-side bot detection techniques
- Advanced bot traffic **heavily rely** on RESIPs

Advantages for the attacker

Tens of millions of residential IPs

No private distributed infrastructure

Solved!

Automated services

Good reputation IPs

No direct traceability

Recruitment process

Free services (e.g. VPN)

Bandwidth payment

Mobile SDKs included by app developers

Infected devices (IoT)

External references:
- M. Frappier et al., Illegitimate residential proxy services: the case of 911.re and its IOCs, 2022.
- X. Mi et al., “Your Phone is My Proxy: Detecting and Understanding Mobile Proxy Networks,” in NDSS 2021.
- A. Vastel. “Ever wonder how proxy providers & BaaS providers obtain residential proxies?”, 2022.

Legittimate but…

Shady Device Recruitment

Credential and stuffing attacks

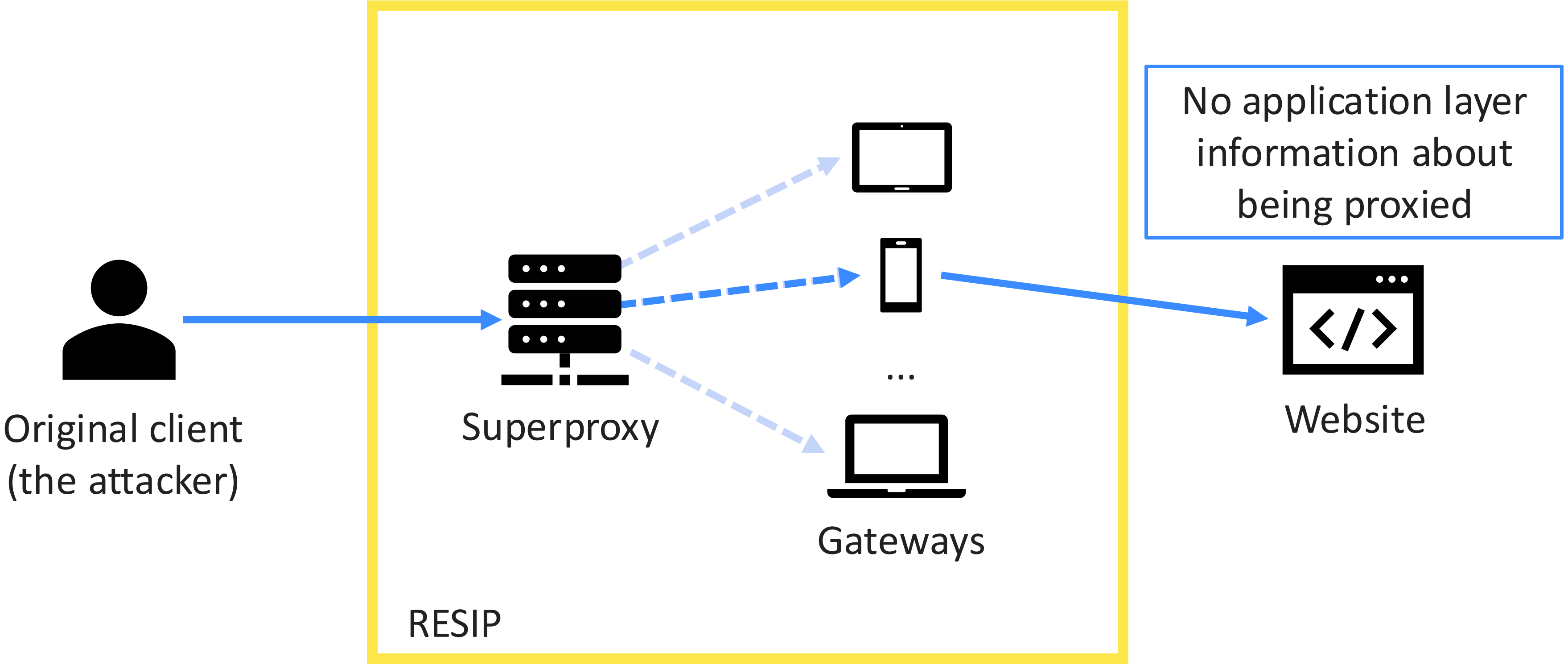
Social Media Spam

Fast Flux Proxies

Cryptojacking

External references:
- M. Frappier et al., Illegitimate residential proxy services: the case of 911.re and its IOCs, 2022.
- B. Krebs, The Rise of “Bulletproof” Residential Networks, 2019.
- X. Mi et al. Resident Evil: Understanding Residential IP Proxy as a Dark Service, IEEE S&P 2019.
- M. Yang et al., An Extensive Study of Residential Proxies in China. ACM SIGSAC CCS 2022.

RESIP infrastructure



Both direct and RESIP connections are indistinguishable at the application layer **but** are there differences at the **transport layer?**

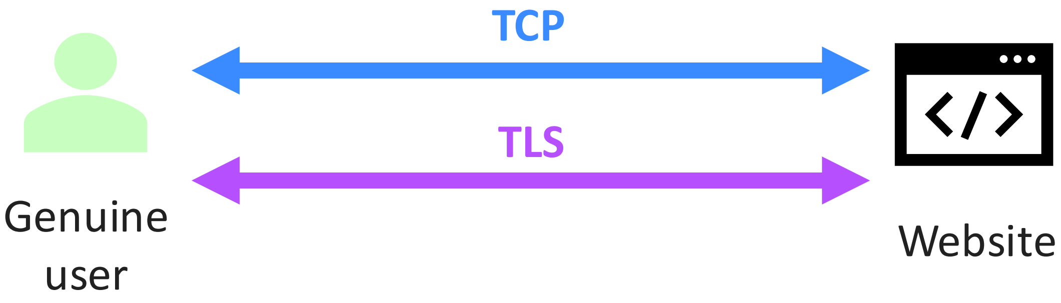
Can we characterize the **IP addresses** used by RESIP and leverage the collected insights for detection?

Outline

1. RESIP Detection based on Round Trip Times

2. RESIP IP addresses analysis

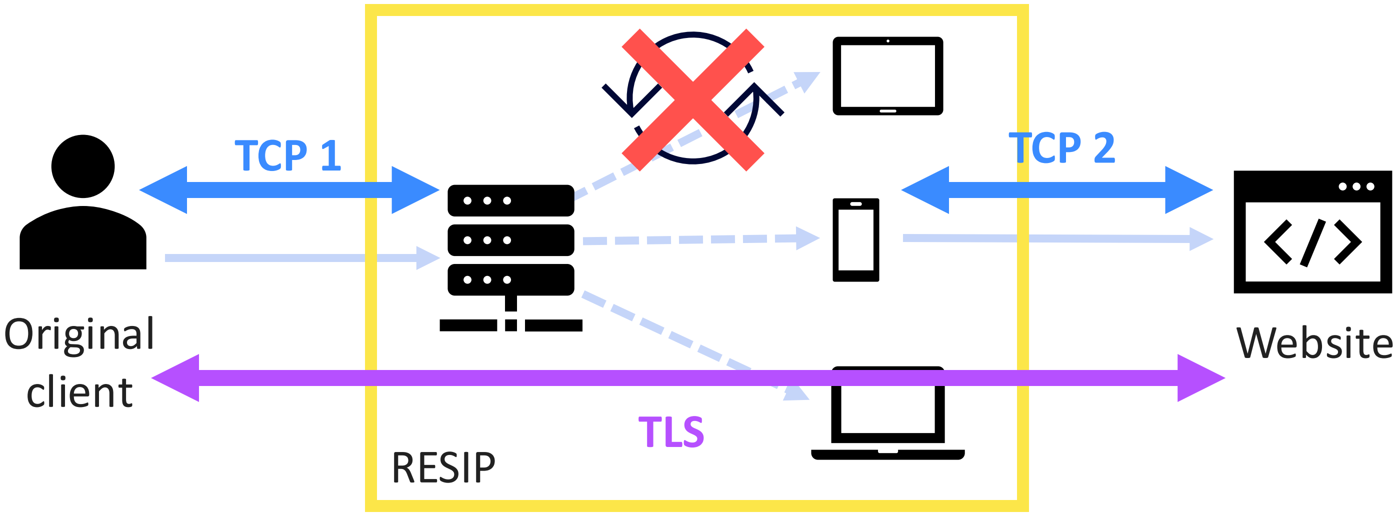
Direct connections



TCP: Transmission Control Protocol

TLS: Transport Layer Security

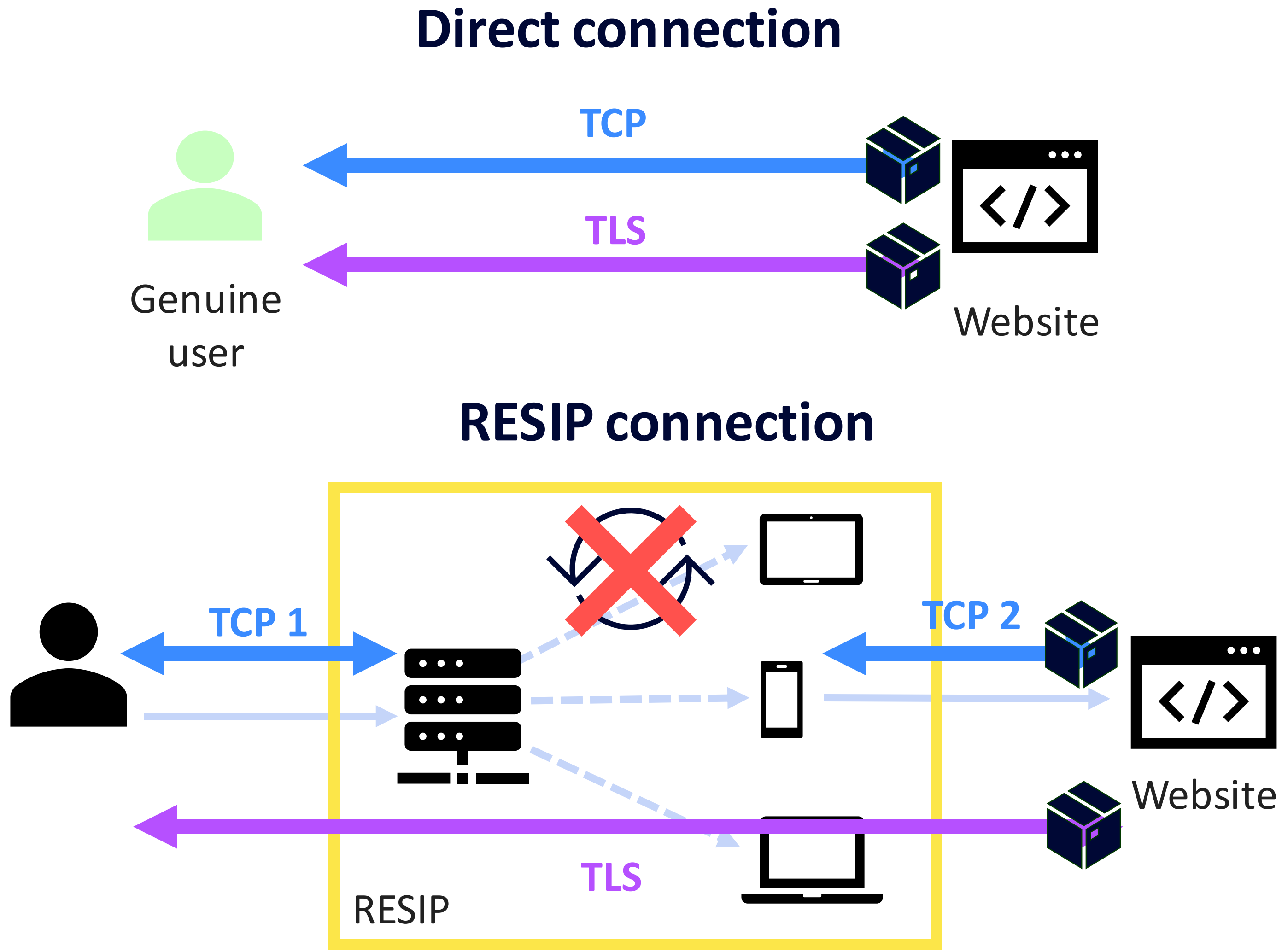
RESIP connection



TCP: Transmission Control Protocol

TLS: Transport Layer Security

How can we check it at the server side?



RTT_{TLS} ~ RTT_{TCP} for direct connections

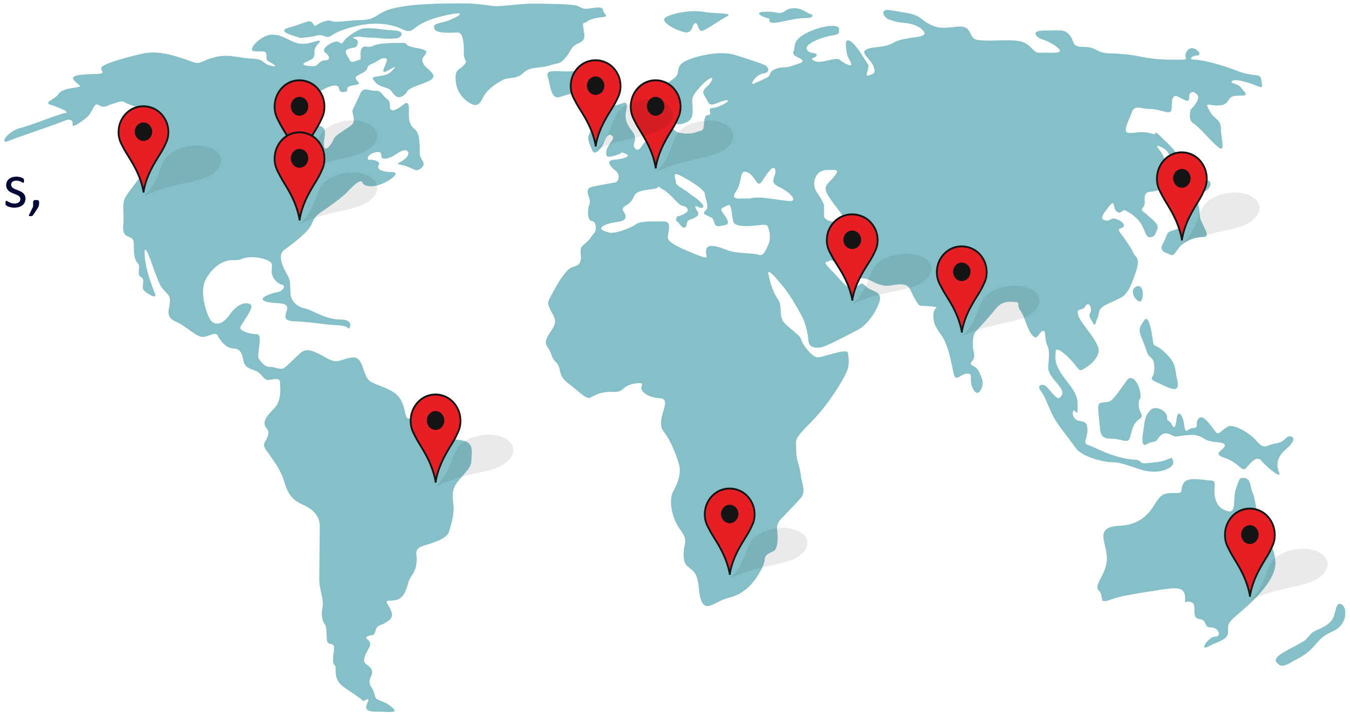
RTT_{TLS} >> RTT_{TCP} for RESIP connections

TCP: Transmission Control Protocol

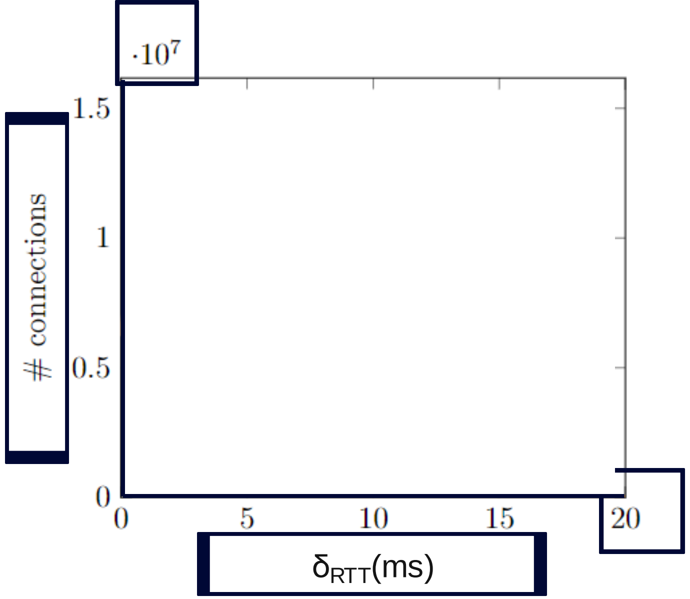
TLS: Transport Layer Security

From theory to practice

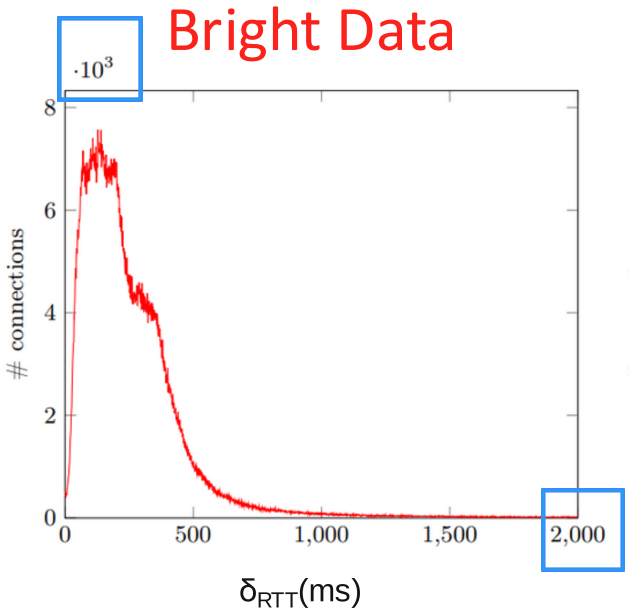
- **2** client/server machines in **11** locations all over the world
- **4** RESIP providers (BrightData, Oxylabs, Proxyrack, Smartproxy)
- **4 months** experiment
- **92M+** connections

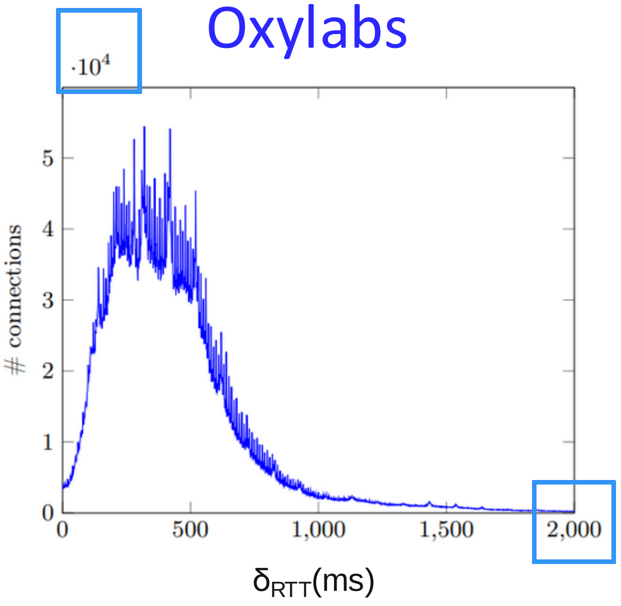


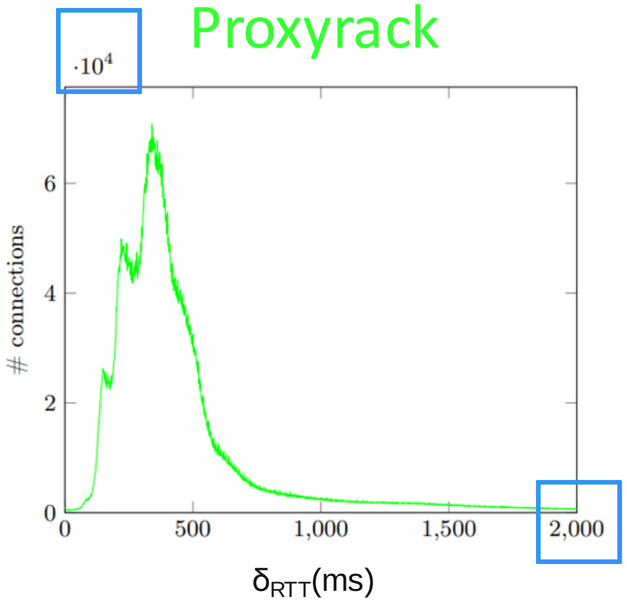
Direct Connections

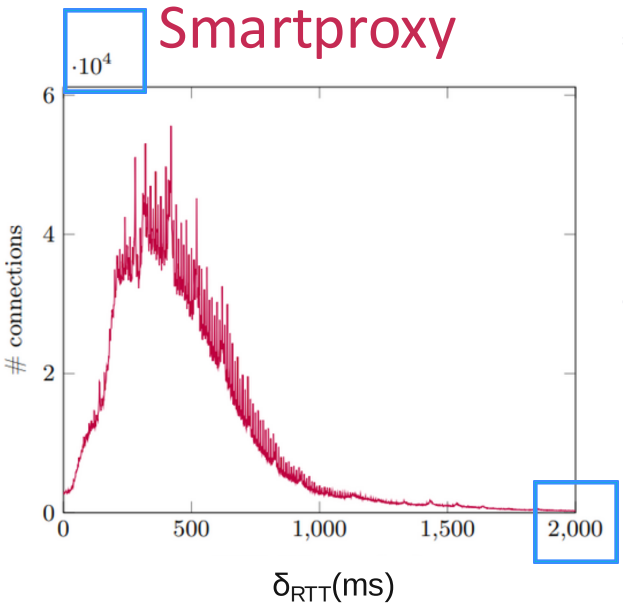


$$\delta_{RTT} = RTT_{TLS} – RTT_{TCP}$$


Bright Data


Oxylabs


Proxyrack


Smartproxy

RESIP Connections

amadeus

RTT Detection

- $\delta_{RTT} > 50ms$ → RESIP Connection

- Possible impacts on the detection technique:
  - Packet speed
  - TLS version

  } No impact

  - Client processing time → Browsers and hotspot increase the difference but below threshold for direct connections
  - Network delays
  - Geographic location of parties

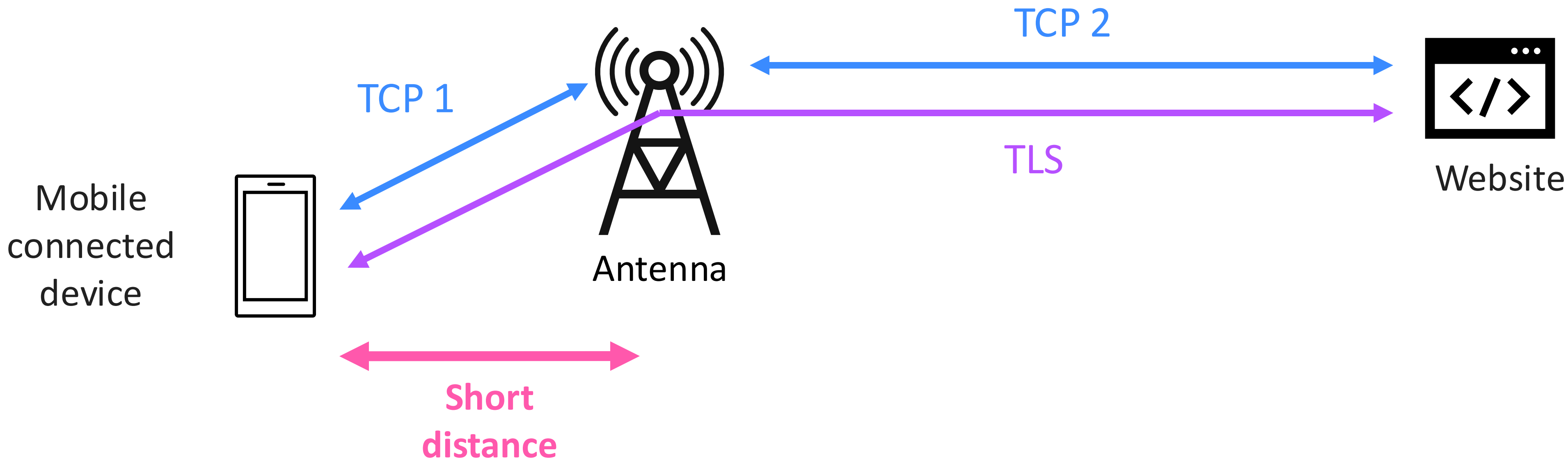
  } Small increase in false negatives

Implementation in real-world

- **Different** from current anti-bot techniques (JS insertion, parameters clustering)
- Amadeus convinced an anti-bot **third party** company to implement the technique
- Analysts currently **using the feature** to detect RESIP campaigns in combination with other parameters

Mobile connections false positives

- Mobile TCP Terminating Proxies



- δ_{RTT}(= RTT_{TLS}–RTT_{TCP}) is smaller than RESIP one
  - Confirmation from semi-controlled and real-world data collections

Detection evasion

- **Downgrading** to HTTP
  - Downgrading **not allowed** + possible generalization
- **Breaking TLS** at the RESIP
  - Technically feasible **BUT**
    - Clients need to accept root certificate from the gateway
    - Gateways devices have access to the content
    - Increased workload for gateways
- **Delaying TCP** packets at the gateway
  - **Unfeasible** since RESIP do not control directly the gateways

Outline

1. RESIP Detection based on Round Trip Times

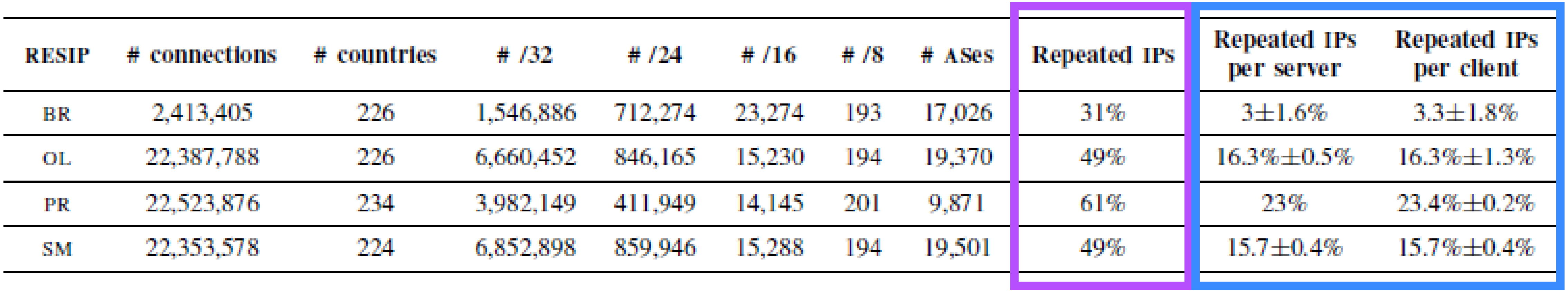
2. RESIP IP addresses analysis

RTT Dataset

- 4 RESIP providers
- 69M+ RESIP connections
- IP analyses:
  - Gateway assignation
  - Machines distribution
  - Amount of machines
  - External RESIP datasets comparison

Gateways Assignation

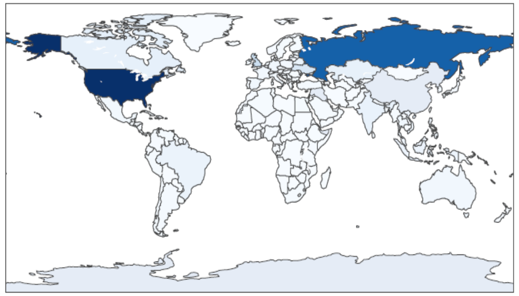
1. Minimization of gateway IP repetitions in a single client-server path but not on among all paths



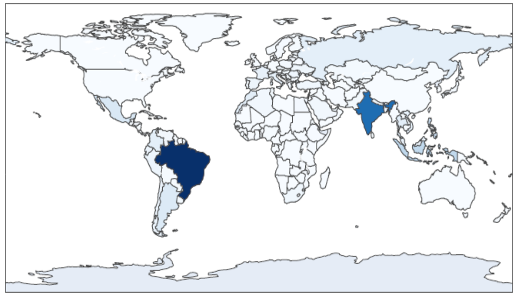
RESIP	# connections	# countries	# /32	# /24	# /16	# /8	# ASes	Repeated IPs	Repeated IPs per server	Repeated IPs per client
BR	2,413,405	226	1,546,886	712,274	23,274	193	17,026	31%	3±1.6%	3.3±1.8%
OL	22,387,788	226	6,660,452	846,165	15,230	194	19,370	49%	16.3%±0.5%	16.3%±1.3%
PR	22,523,876	234	3,982,149	411,949	14,145	201	9,871	61%	23%	23.4%±0.2%
SM	22,353,578	224	6,852,898	859,946	15,288	194	19,501	49%	15.7±0.4%	15.7%±0.4%

Machines distribution

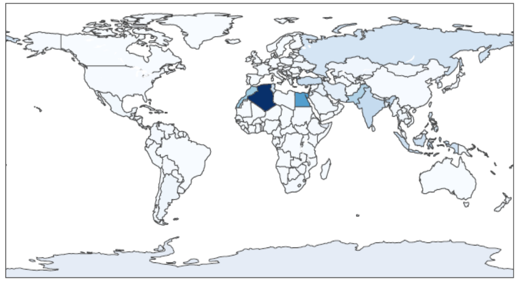
2. Similar gateways geographical distribution for two providers



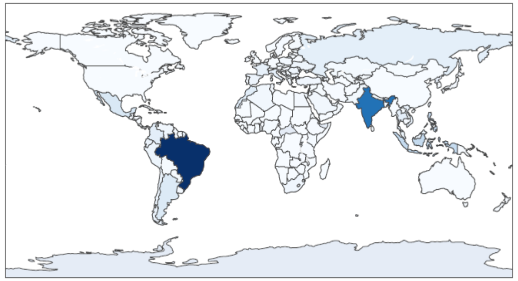
Bright Data



Oxylabs

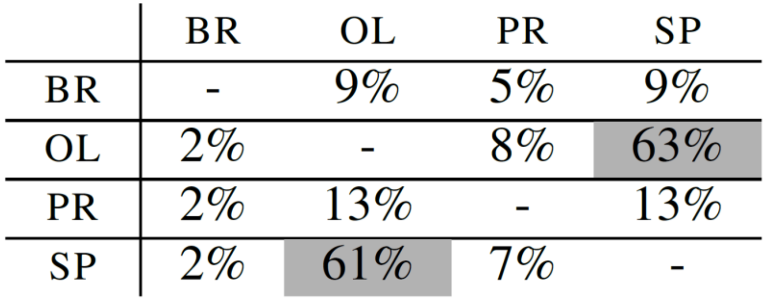


Proxyrack



Smartproxy

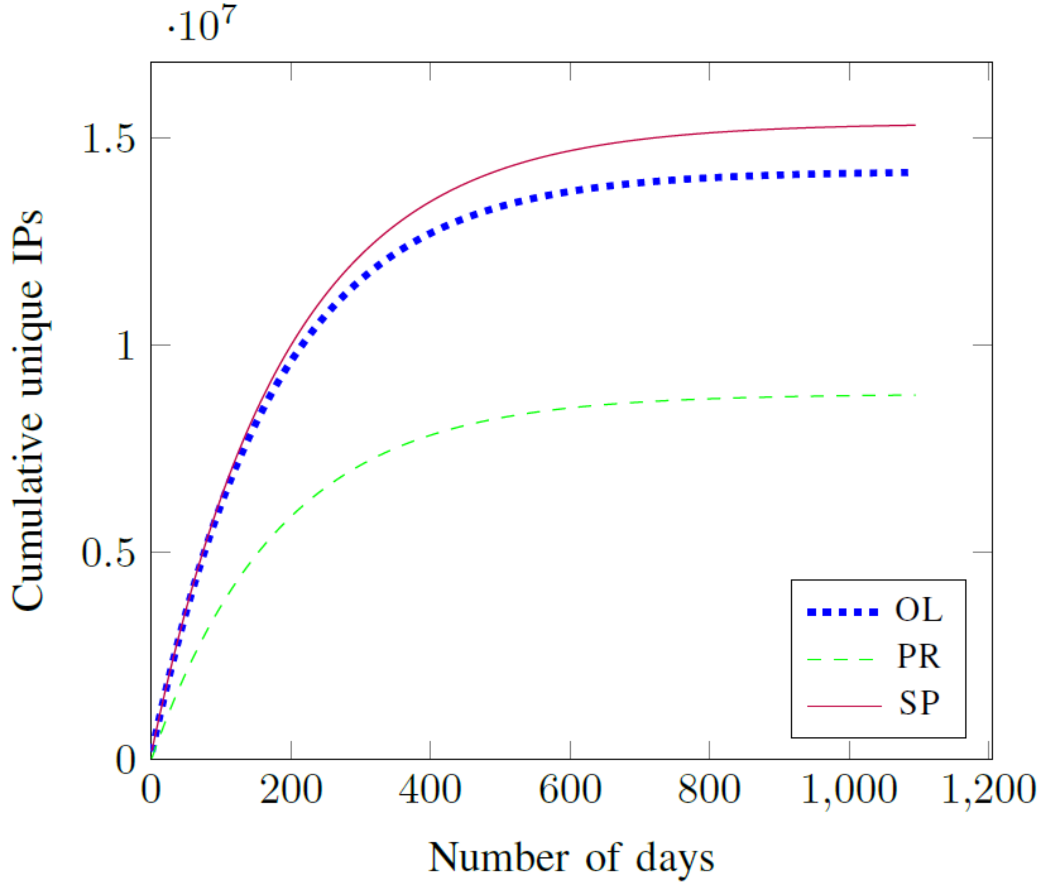
3. Shared IP Pool among providers



	BR	OL	PR	SP
BR	-	9%	5%	9%
OL	2%	-	8%	63%
PR	2%	13%	-	13%
SP	2%	61%	7%	-

Amount of machines

4. Advertised IP pool sizes do not match our observation and projections



External RESIP dataset comparison [1/2]

- External DS 1:
  - May 2017 - March 2018 (vs Jan 2022 – May 2022)
  - 6,419,987 RESIP IPs from 5 RESIP providers
- Sharing two RESIP providers with our study, BrightData and Proxyrack

DB	IP repetition	IP repetition BD	IP repetition PR
RTT DS	2.87 %	2.52 %	1.26 %
External DS 1	6.26 %	0.97 %	5.86 %

DB	/24 repetition	/24 repetition BD	/24 repetition PR
RTT DB	46.04 %	33.17 %	29.15 %
External DS 1	45.52 %	19.96 %	34.74 %

External DS: https://rpaas.site/

External RESIP dataset comparison [2/2]

- External DS 2:
  - April 2021 - October 2021 (vs Jan 2022 – May 2022)
  - 9,077,278 Chinese RESIP IPs from 6 RESIP providers

DB	IP repetition
RTT DS	5.22 %
External DS 2	8.04 %

DB	/24 repetition
RTT DB	54.33 %
External DS 2	58.52 %

What did we learn about the IPs

- Each provider **reuses** IPs among different paths (and possibly users)
- Different providers **share** pools of IPs
- The total amount of RESIP IPs is **smaller** than advertise values
- IP changes, **/24 vary less**
- Can we **leverage** this information?
  - Tracking /24 and associate the ones where RESIPs appear to a **risk score**
  - Genuine users share their devices -> **Whitelisting** to reduce FPs
    - Association of IPs completing a confirmed human action (e.g. booking) to the corresponding fingerprint
- **Next step:** track the coverage with the RESIP IPs detected in Amadeus

Conclusions

Conclusions

- RESIP are largely used for **sophisticated bot attacks**
- New technique based on **RTT measurement** to differentiate between RESIP and direct connections
  - Working request by request
  - Not impacted by packet speed, TLS version, client processing time and only small false negative increases in case of network delays and parties close in location
  - Difficult to evade
- New directions to **track /24** used by RESIP for detection
  - Risk score when an IP of the /24 acts as RESIP
  - Whitelisting of genuine user fingerprints

amadeus

Any question?

Offline questions? elisa.chiapponi@amadeus.com

Presentation based on:

- E. Chiapponi (2023). Detecting and Mitigating the New Generation of Scraping Bots. In Ph.D. Dissertation, Sorbonné Université, Cryptography and Security.
- E. Chiapponi et al. (2022). BADPASS: Bots taking ADvantage of Proxy AS a Service. In ISPEC 2022.
- E. Chiapponi et al. (2023). Inside Residential IP Proxies: Lessons Learned from Large Measurement Campaigns. In WTMC 2023.
- E. Chiapponi et al. (2023). Towards Detecting and Geolocalizing Web Scrapers with Round Trip Time Measurements. In TMA 2023.
- E. Chiapponi et al. (2023). Poster: The Impact of the Client Environment on Residential IP Proxies Detection. In IMC 2023.

Check them here:

amadeus