# Towards Detecting and Geolocalizing Web Scrapers with Round Trip Time Measurements

Elisa Chiapponi [1], Marc Dacier [2], Olivier Thonnard [3]

[1]EURECOM, FR [2]RC3, CEMSE - KAUST, SA [3]Amadeus IT Group, FR

## Background

- A **persistent battle** takes place between e-commerce websites detecting scrapers and scraping bots trying to evade detection [1].
- Lately, scrapers exploit **Residential IP provider** (RESIP) services.
- RESIP providers supply **tens of millions** residential IPs as exit points, **shared with real users**. → Risk to block legitimate users that share IPs with scrapers.

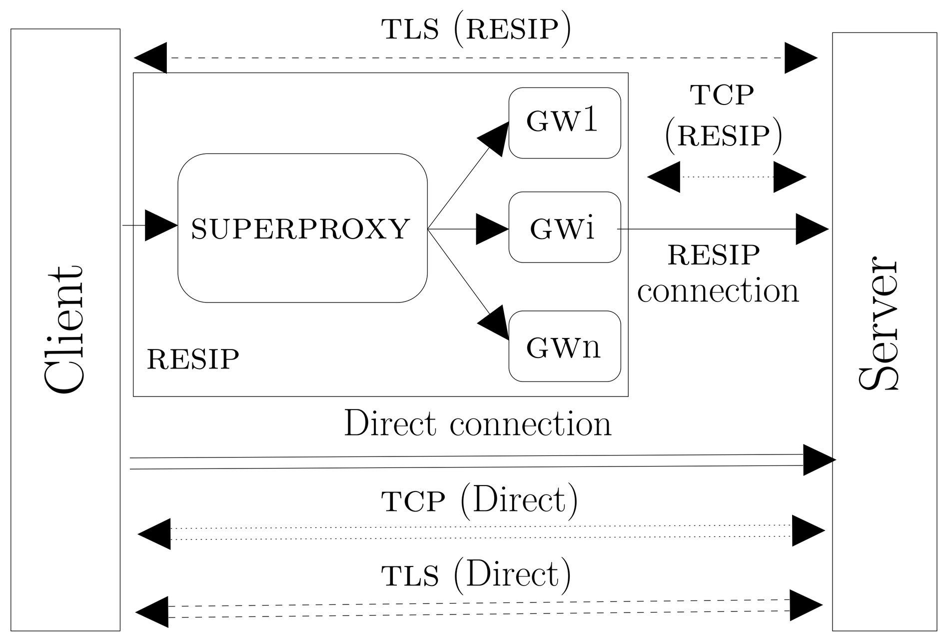



Figure: Client sending direct and RESIP connections. The TCP and TLS sessions are built between different parties in the two scenarios.

- In [2], we proposed RTT_DETECTION a **server-side detection method** based on the **difference** in the Round Trip Times (RTTs) at the TLS and TCP layer of RESIP connections.

$$\delta_{RTT} = \mathrm{RTT}_{TLS} - \mathrm{RTT}_{TCP} > 50ms \implies \mathrm{RESIP}$$

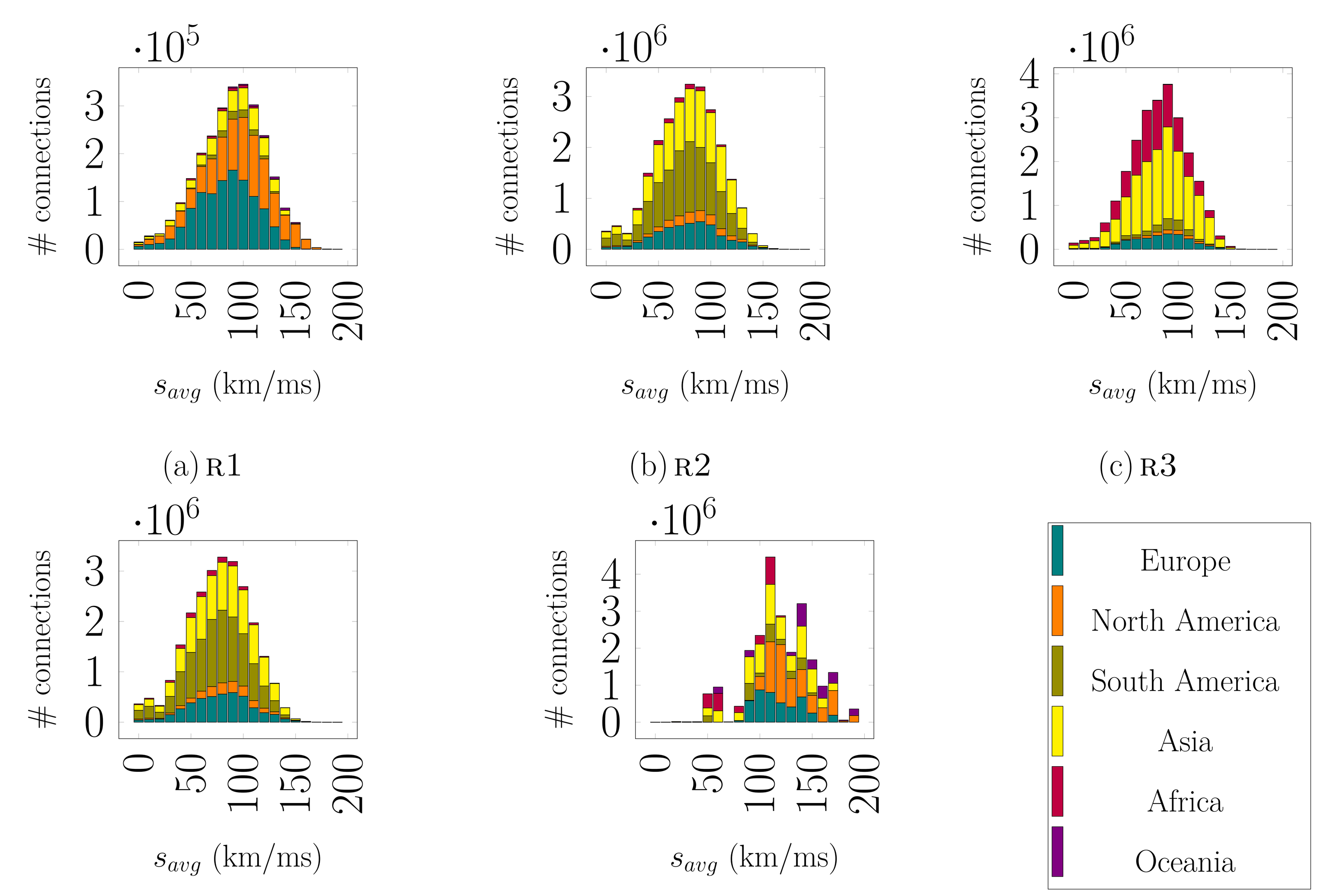
- We run a **4-months measurement campaign** (92M+ connections). The technique showed **99.01%** accuracy.

## Machines Proximity Impact

- **Close proximity**: client, server, SUPERPROXY and GW not further than 1,000km from each other (**0.07%** of RESIP connections in our experiment).
- $\delta_{RTT} > 50$ms in **3 out 4** of cases.
- Only **3.07%** of connections show $\delta_{RTT} < 20$ms where 97% of direct connections.
- The machines proximity **influences** the technique but there is still a **significant difference** between RESIP and direct connections.

## Measurements Representativeness

- RTT =measure of time → we need the **speed** to find the distance.
- An **idealized common value** for the average packet speed ($s_{avg}$) **does not exist** for connections across different areas of the world.
- **Does our data reflect this?**
- $s_{avg}$ (RESIP) $= d_{server-\mathrm{GW}} / (1/2\ \mathrm{RTT}_{TCP})$.
- $s_{avg}$ (Direct) $= d_{server-client} / (1/2\ \mathrm{RTT}_{TCP})$.




(a) R1 (b) R2 (c) R3 (d) R4 (e) Direct Connections (f) Legend

Figure: Distribution of the mean speed of packets for each RESIP provider and direct connections.

## Real World Implementation

- Implementation of the RTT_DETECTION in front of **real-world domains** suffering from web scraping.
- Early results: the $\delta_{RTT}$ is a **strong parameter** to check when a connection passes through a RESIP.
- In two representative months, the detection was used in **74.32%** of investigations.
- **Ongoing study** of the flagged connections to assess the impact of the detection and possible false positives.

## Geolocalizing Behind the RESIP

- **Idea:** using the $\delta_{RTT}$ to **geolocalize the client**.
- The $\delta_{RTT}$ gives information about the **"distance" client-GW**.
- If the same client uses multiple GWs to send requests to the same server, we can find the **intersection of the circles** whose centers are the GWs locations and whose radii are half of the $\delta_{RTT}$ multiplied by the average packet speed ($s_{avg}$).

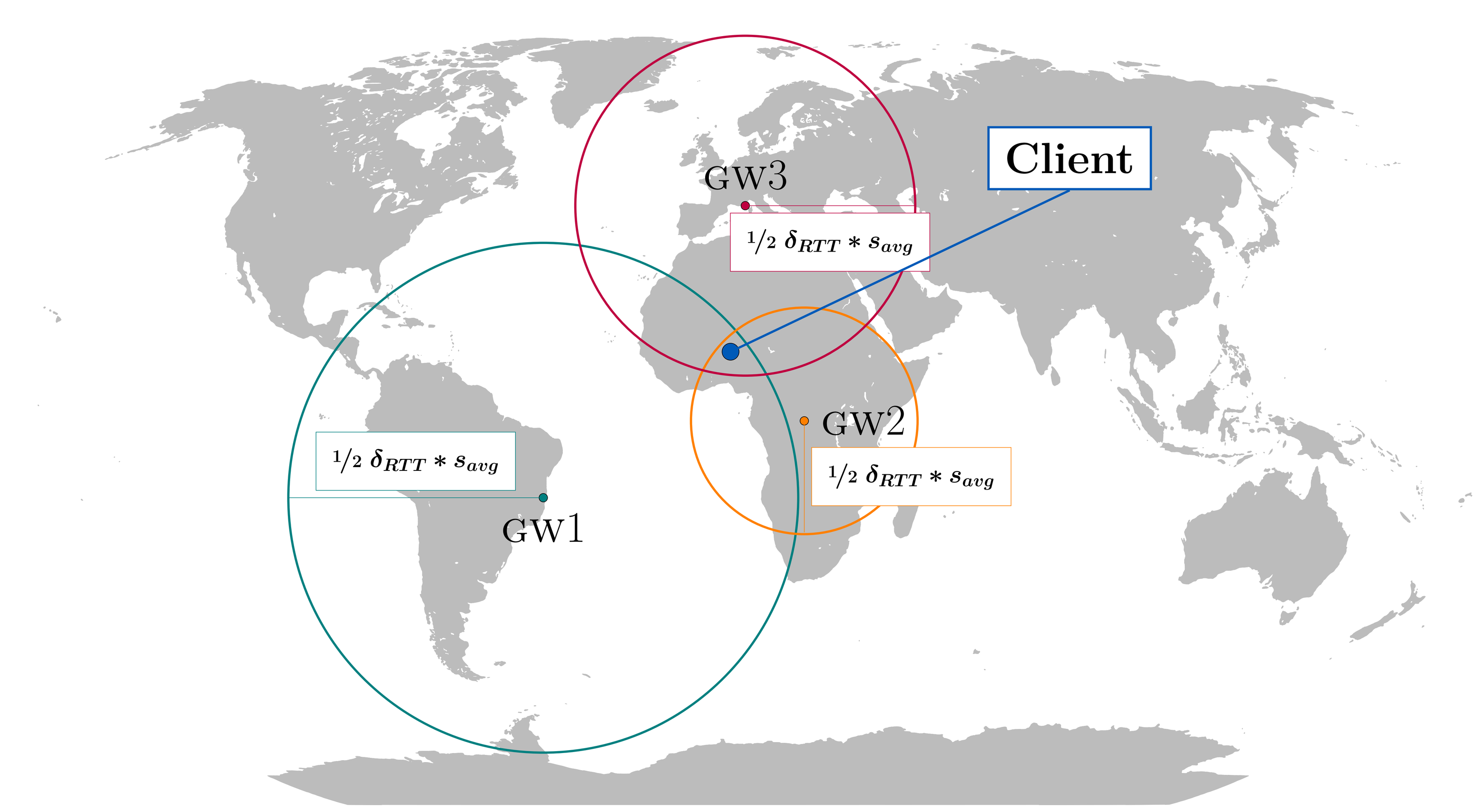



Figure: Example of geolocalization of a client that uses 3 GWs.

- **Challenges** in achieving our goal:
  - The $s_{avg}$ has **no average value**.
  - **Current geolocalization algorithms** are **not able** to correctly put into practice our theoretical idea [3].
- **Ongoing implementation** of a new algorithm that overcomes previous limitations.

[1] E. Chiapponi, M. Dacier, O. Thonnard, M. Fangar, M. Mattsson, and V. Rigal, "An industrial perspective on web scraping characteristics and open issues," in *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*, pp. 5–8, 2022.

[2] E. Chiapponi, M. Dacier, O. Thonnard, M. Fangar, and V. Rigal, "BADPASS: Bots Taking ADvantage of Proxy as a Service," in *Information Security Practice and Experience: 17th International Conference (ISPEC 2022)*, p. 327–344, 2022.

[3] M. Champion, M. Dacier, and E. Chiapponi, "ImMuNE: Improved Multilateration in Noisy Environments," in *2022 IEEE 11th International Conference on Cloud Networking (CloudNet)*, pp. 1–6, 2022.