AI CENTER
FEE CTU

DNS4EU

From Billion Queries to Action: How DNS4EU Transforms Threat Defense

Sebastián García - Tigran Oganesian

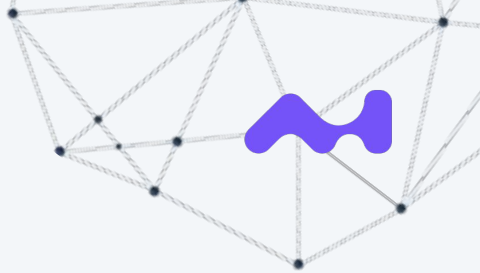
Stratosphere Laboratory

Artificial Intelligence Center

Faculty of Electrical Engineering

Czech Technical University in Prague

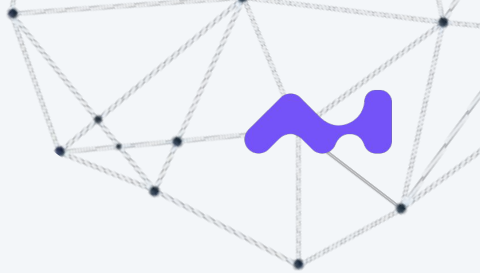
The Problem



How to find new malicious domains in all

the DNS traffic of the DNS4EU project?

What is DNS4EU?



- The official DNS resolution infrastructure of EU
- So far DNS traffic ended in the hands of commercial companies outside EU
- EU wants digital sovereignty, privacy and application of EU regulations
- DNS4EU was born to offer a private and secure DNS infrastructure

Join! https://www.joindns4.eu/

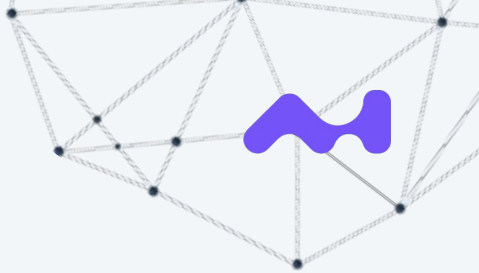
Type	IP address	DNS over HTTPS	IPv6	Description
Protective resolution	86.54.11.1 86.54.11.201	protective.joindns4.eu/dns-query	2a13:1001::86:54:11:1 2a13:1001::86:54:11:201	Avoid access to websites with fraudulent or malicious content.
Protective resolution with child protection	86.54.11.12 86.54.11.212	child.joindns4.eu/dns-query	2a13:1001::86:54:11:12 2a13:1001::86:54:11:212	Avoid any access to websites inappropriate to children like sexual content, violence or drugs on top of the protective functionality.
Protective resolution with ad blocking	86.54.11.13 86.54.11.213	noads.joindns4.eu/dns-query	2a13:1001::86:54:11:13 2a13:1001::86:54:11:213	Hide the advertisement on the websites and in the applications on top of the protective functionality.
Protective resolution with child protection & ad blocking	86.54.11.11 86.54.11.211	child-noads.joindns4.eu/dns-query	2a13:1001::86:54:11:11 2a13:1001::86:54:11:211	Avoid any access to websites inappropriate to children like sexual content, violence or drugs. Plus filter the advertisement on top of the protective functionality.
Unfiltered resolution	86.54.11.100 86.54.11.200	unfiltered.joindns4.eu/dns-query	2a13:1001::86:54:11:100 2a13:1001::86:54:11:200	Option for the users who are confident about security of their devices and connection and looking for fast, reliable and anonymized resolution service.

DNS4EU

whalebone

STRATOSPHERE LABORATORY

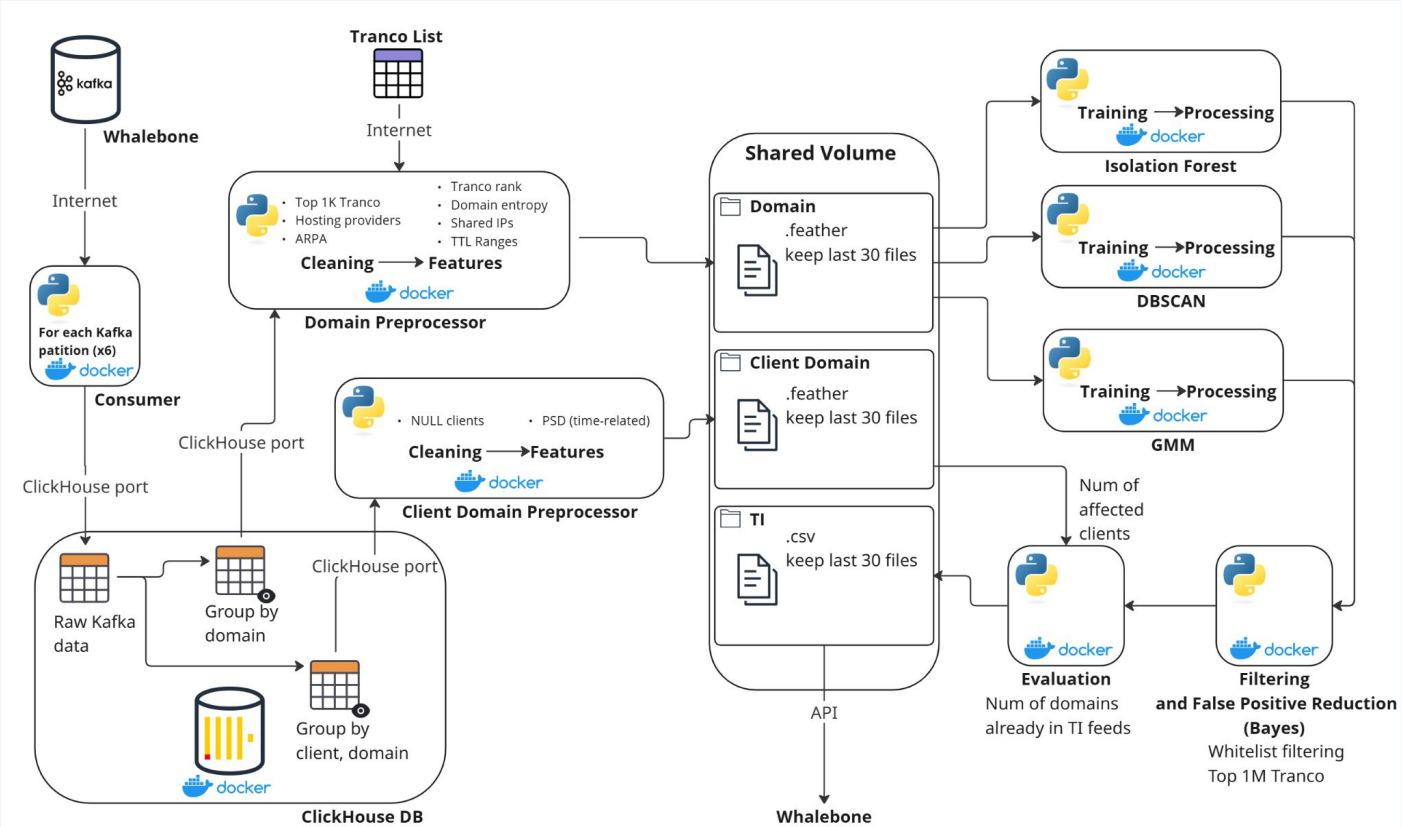
⚖️ The Scale Issue of DNS4EU



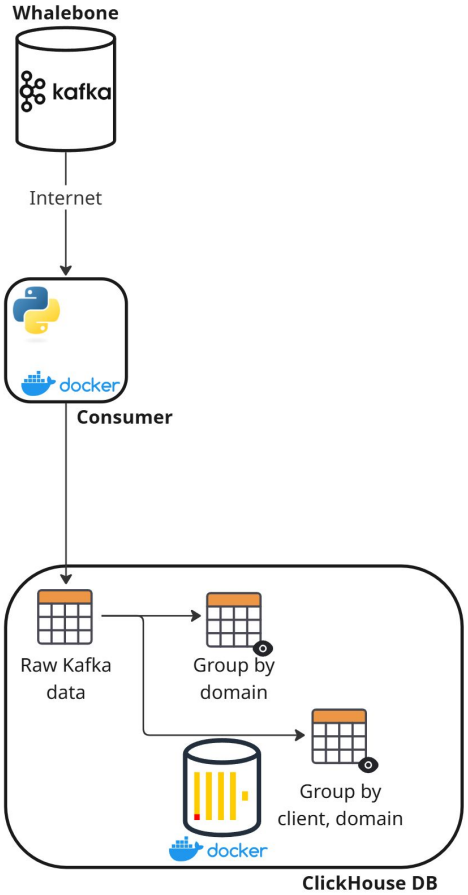
- Planned to service **100 million** users
- **Federated** architecture: Public Cloud + ISP on-premise + Gov resolvers
- Guarantees **low-latency** and availability in EU
- Balance between storage, analysis and infrastructure cost
- Should scale up with future use

The Real Scale Issue

In Stratosphere



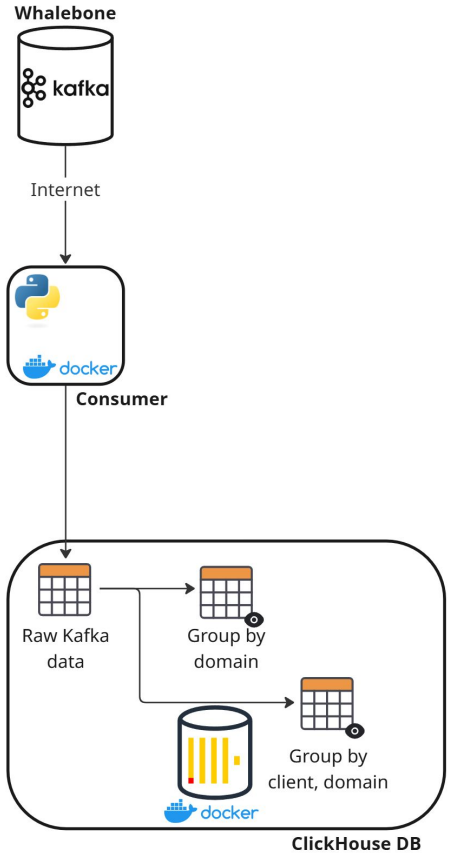
Consumer



Kafka Raw Data we Use:


```
{
  "answer": "74.123.123.12",
  "answer_ip": "74.123.123.12",
  "client_ip": "3fff:ffff:6bf1:...:...:...:",
  "domain_l1": "org.com",
  "domain_l2": "new.org.com",
  "query": "new.organization.com.",
  "query_type": "A",
  "timestamp": "2024-10-01 12:34:56",
  “ttl": "300"
}
```

Consumer



Kafka Consumers:

Six consumers, **each to a Kafka broker partition**.

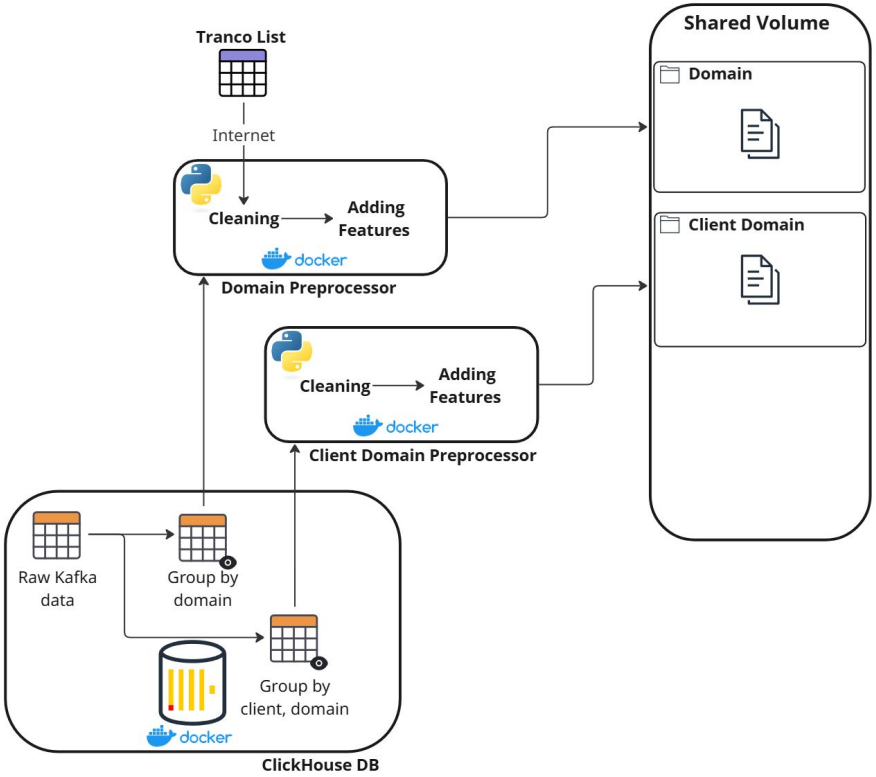
ClickHouse Storage:

Raw Data Table: All DNS traffic data.

Domain View: Aggregates per-domain metrics: request counts, TTL statistics (min, max, avg, stddev), and unique IPs in last day 24-hour window.

Client Domain View: Aggregates client-domain metrics: request counts and timestamps for pattern analysis in last day 24-hour window.

Preprocessors



Get Tranco list

Domain Preprocessor: Get 24 hours of domains.

Cleans data: Removes invalid entries + Tranco Top 1000.

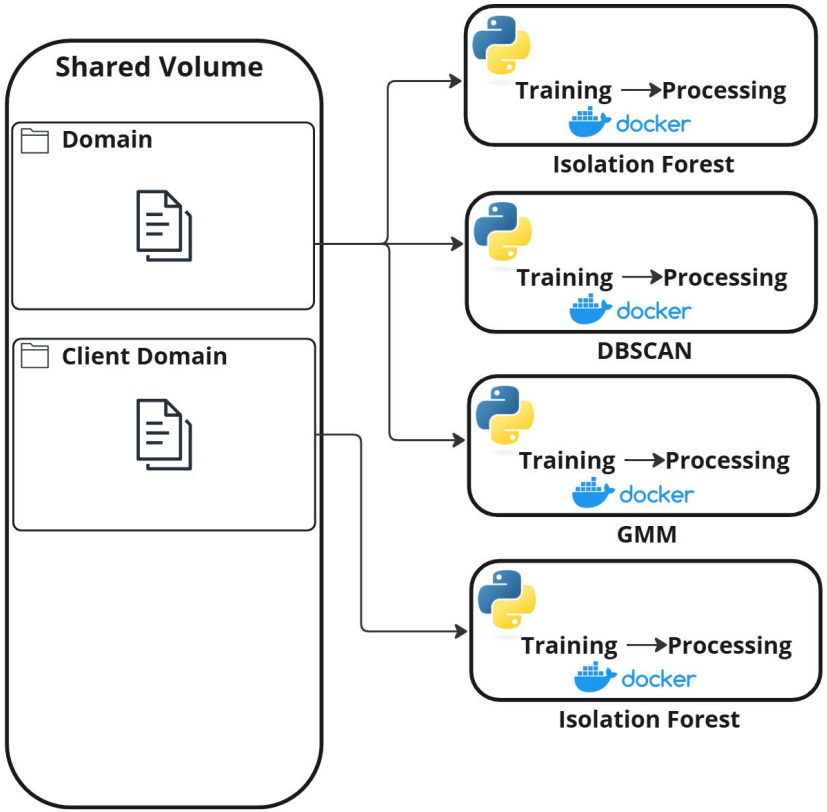
Extracts features: Request counts, TTL statistics, aggregate metrics, domain name entropy.

Client-Domain Preprocessor: Gets 24 hours of client-domain

Cleans data: Removes invalid entries, Tranco Top 1000.

Extracts features: Applies Power Spectral Density (PSD) for temporal pattern analysis.

Models



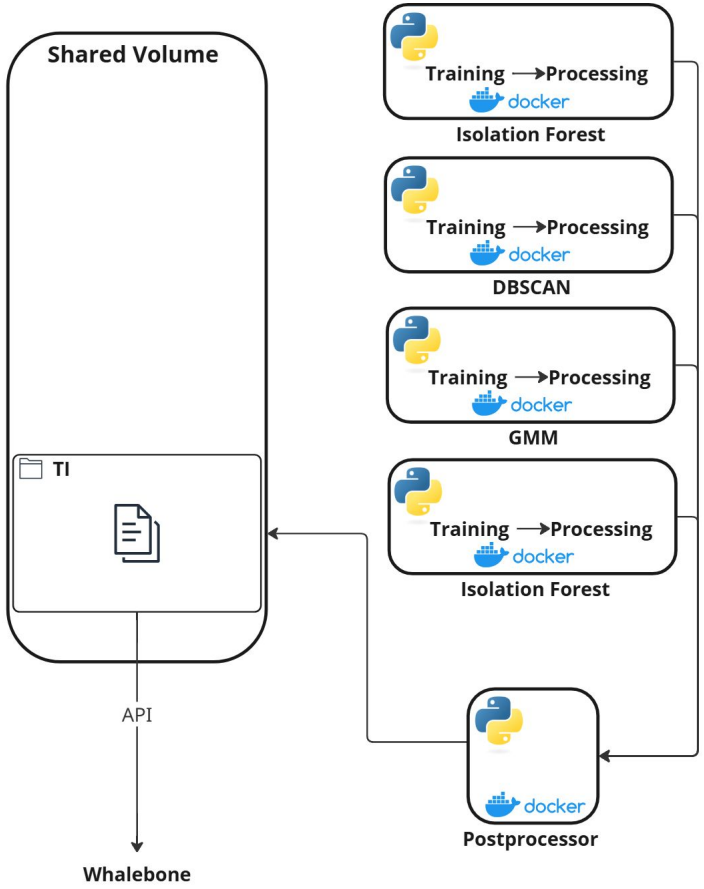
Models:

Isolation Forest: Detects anomalies via tree-based isolation.

DBSCAN: Clusters data, flags outliers as potential threats.

GMM: Models data distribution, identifies low-probability domains.

Postprocessor



Postprocessor:

Merges CSV outputs from Isolation Forest, DBSCAN, and GMM models.

Generates **daily Threat Intelligence feed**.

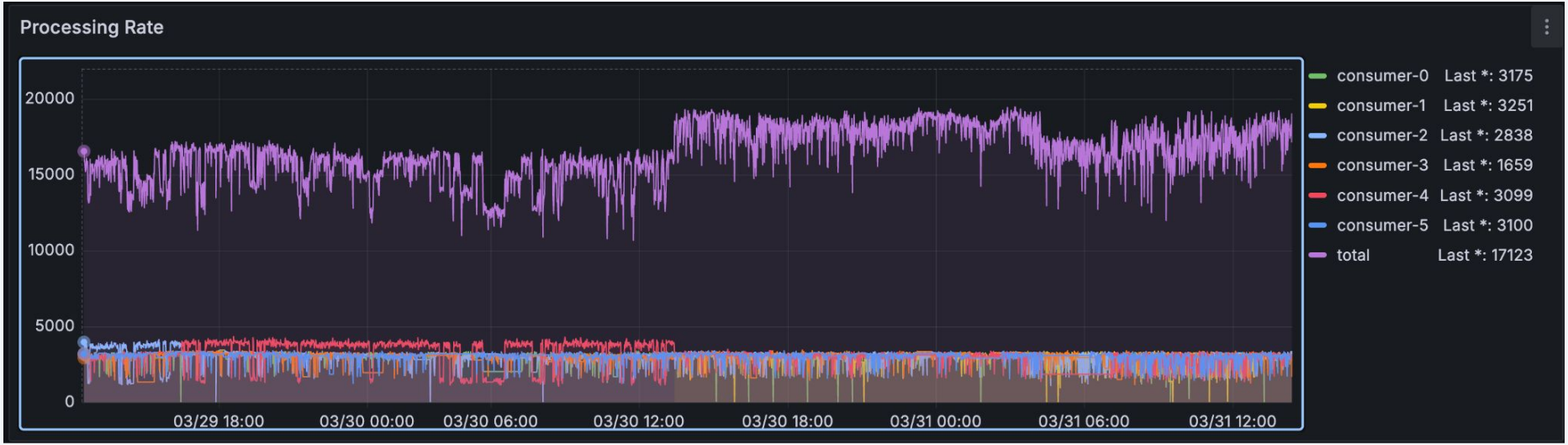
Data Consumption

- **Min performance**
  ~15,000 queries per second

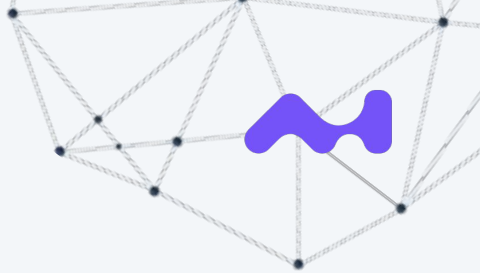
- **Min Vol per Hour**
  15,000 × 3,600 = **54,000,000 q/hour**

- **Daily Min Vol Mínimo Diario**
  54,000,000 × 24 = **1.296 billion queries/day**

- **Estimated Storage**: **~1 TB per week**

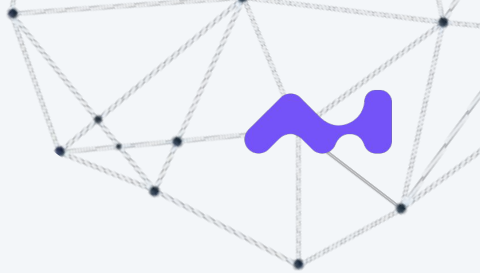


The Real Problem. Detect New Malicious DNS Domains



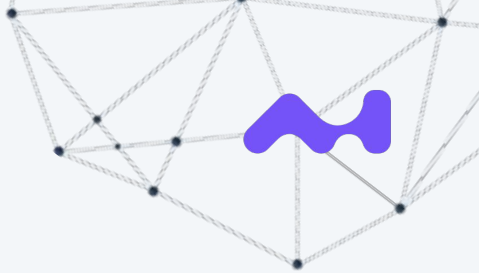
- New.
  - Not in Threat Intelligence lists.
  - Not in Blacklists, or historic reports.
  - No previous knowledge
- Malicious.
  - No DGA
  - No Phishing
  - No DoS
- What is left:
  - Malware C&C and delivery
  - Exfiltration/Tunneling
  - Botnets
  - Spyware
  - Infrastructure for targeted operations

Features. The Hard Problem



- How to find the correct set of features from the available data?
- The client ID changes every 24hs
- Client IDs are not IPs, so no geolocation or TI.

Features. The Hard Problem



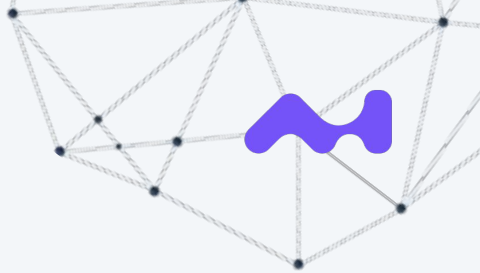
- `query`: Domain name (used only for identification)
- `num_requests`: Number of DNS queries for this domain in the time window
- `ttls`: Raw TTL values (list)
- `min_ttl`: Minimum TTL
- `max_ttl`: Maximum TTL
- `avg_ttl`: Mean TTL
- `stddev_ttl`: Standard deviation of TTLs
- `ttl_unique_count`: Number of unique TTLs
- `ttl_range`: Range (max - min) of TTLs
- `ttl_entropy`: Shannon entropy of TTL values
- `ttl_iqr`: Interquartile range (75% - 25%) of TTLs
- `answer_ips`: Raw IPs (list)
- `num_ips`: Count of unique IPs seen
- `ips_entropy`: Entropy of IP distribution (e.g., random/rotating vs. stable)
- `ip_sharing_count`: Number of other domains sharing any of the IPs
- `dominant_frequency`: Most significant frequency (from FFT of query timestamps)
- `total_power`: Total signal power from FFT
- `peak_magnitude`: Peak magnitude in FFT
- `mean_magnitude`: Average FFT magnitude
- `spectral_entropy`: Entropy of FFT magnitudes — how spread the signal is
- `domain_entropy`: Shannon entropy of the domain string (e.g., to detect DGA-like domains)

If the features do not split the data somehow, models will have a hard time

AI Models in Two Steps

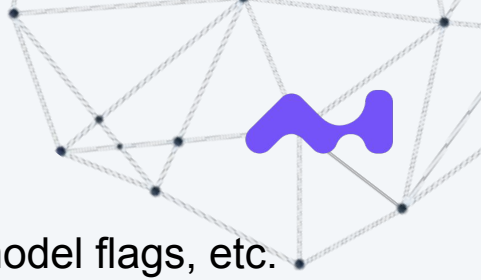
1. **Clustering** to group suspicious domains and find relationships. Many Assumptions.
    a. **Anomaly detection in the Clusters**
        i. Isolation Forest
            1. Detects anomalies by **isolating points** through random feature splits. Anomalies have shorter path lengths than normal data.
        ii. Gaussian Mixture Model (GMM)
            1. Learns the shape by fitting probability distributions. Identifies the likelihood of a new point for each group model.
        iii. DBSCAN
            1. Detects anomalies by grouping points **based on density**; domains not fitting into any dense region are labeled anomalies.
2. **False positive reduction** with Bayesian Inference

FP Reduction With Bayes Inference



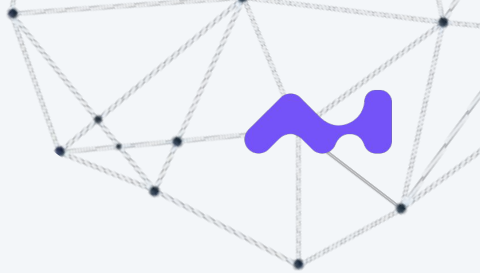
- A prior belief is the starting **assumption** we make before seeing new data.
    a. “Only about 5% of domains are malicious”
    b. “This threat-intel feed usually lights up on bad domains.”
- For each assumption, we record both the **percentage** itself and how **confident** we feel
    a. Strong confidence behaves like lots of past experience
    b. Weak confidence is easy to change
- These **priors** give the model a sensible baseline so it isn’t guessing in the dark.

FP Reduction With Bayes Inference



- A new domain arrives with evidences such as TI feed hits, anomaly model flags, etc.
- For each domain we model both the probability of being malicious and of being benign.
- We **compare** each **evidence** against our reliability **priors**: does it fit the “malicious” story better or the “benign” story?
- Every evidence nudges the belief up or down, and the nudges add together.
- Combine those nudges with the prior malicious-rate story and we obtain the posterior.
    a. The **updated** probability the domain is **malicious**, plus an honest view of remaining uncertainty.

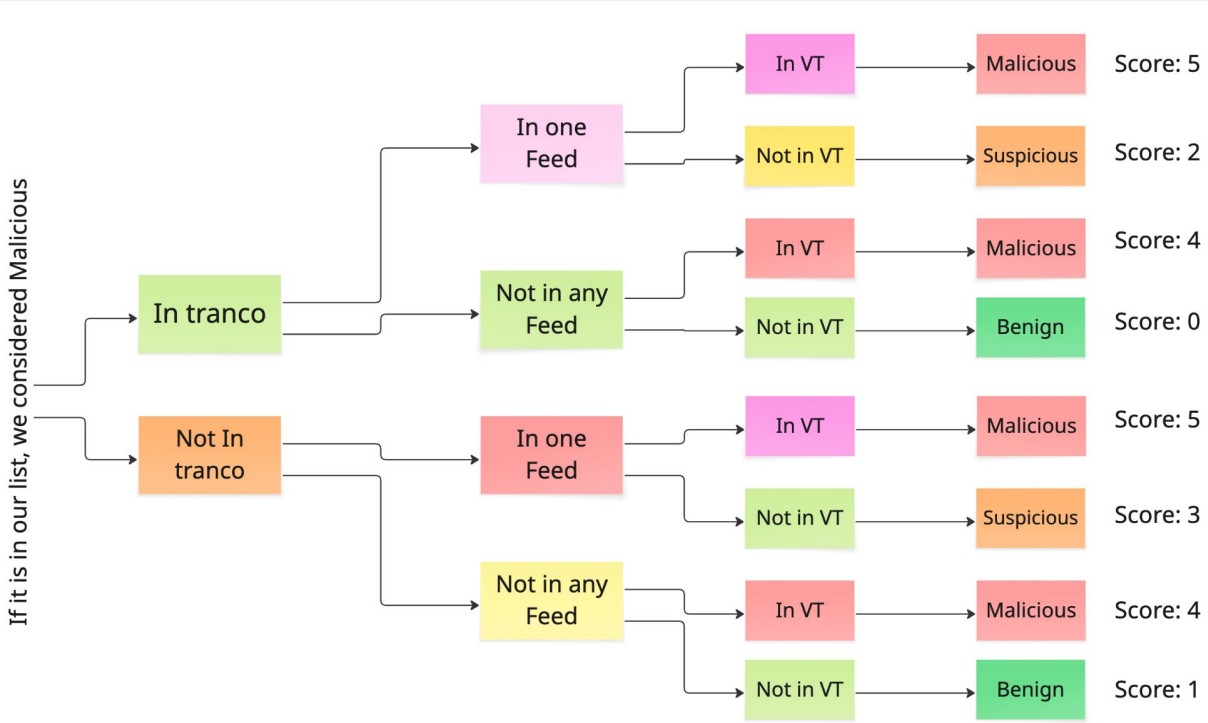
How to Validate if it works?



- There are no labels, no confirmations.

- For each domain proposed to be blocked, we search about it ***the next day***.
  - Position in tranco list
  - Check 15 TI feeds. Daily updated
  - Consult APIs, e.g.
    - VirusTotal
    - Urlscan
    - OTX
    - Greynoise
    - Crowdsec

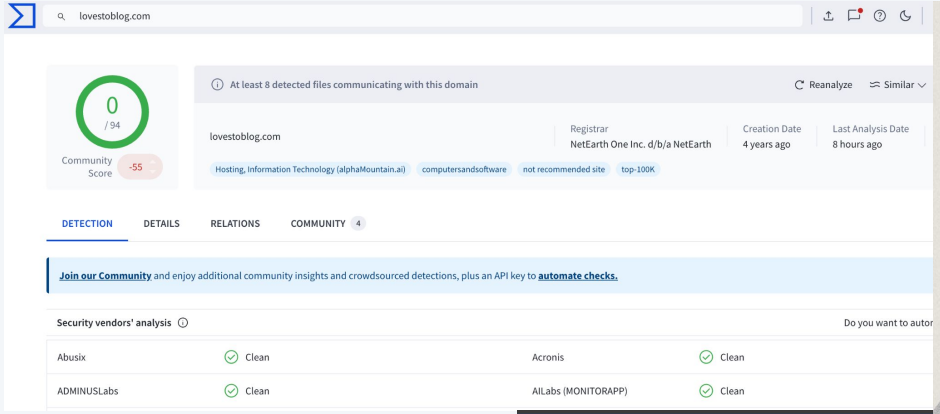
Validation Methodology



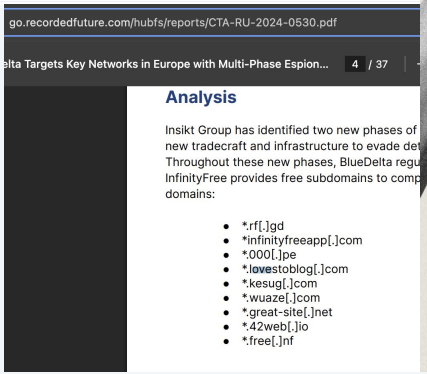
Malicious	Score: 5	16.4%
	Score: 4	0.1%
Suspicious	Score: 3	5.9%
	Score: 2	0.1%
Benign	Score: 1	77.4%
	Score: 0	0.1%

Real Cases

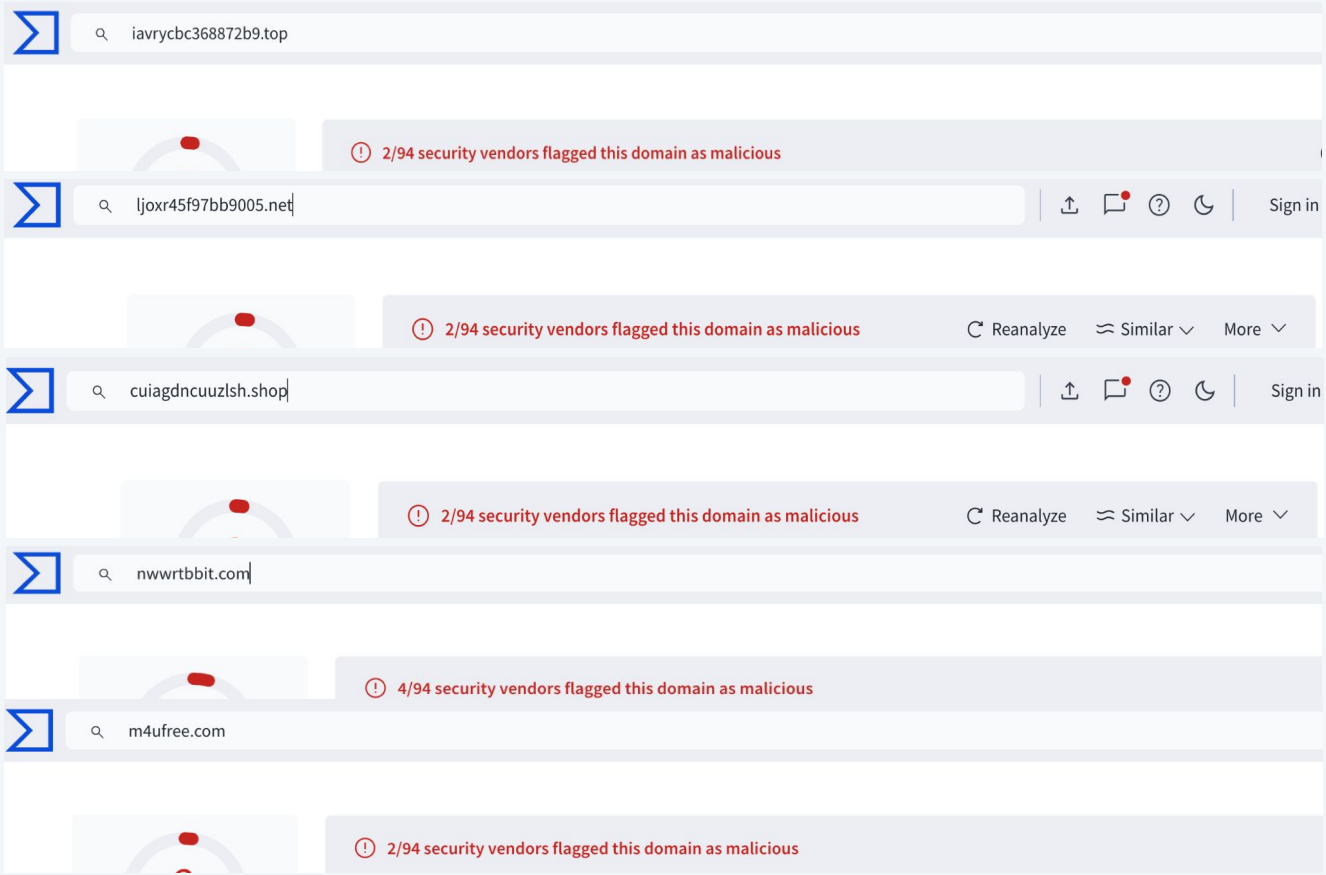
lovestoblog.com



It is in the top 100k of the Tranco whitelist!



Real Cases




```
ai-growth-matrix.net
parner-id-1381834.com
2x.si
olimpptsp.kz
premcogroup.com
atyurs.com
avastexodus.com
bdcvpn.com
burriton.ru
securedmicrosoft365.com
bizlawyer.org
cry-havok.org
flyxz.top
gandharaart.org
getbehavior.top
gliteam.net
nohelp.top
poracholly.ru
speedjc.top
angstromcom.com
antiquebotv3.com
asfaltwerk.com
awesometech.team
bnb-teiegram.com
eucudo.icu
eventq.io
expressoquiririm.com.br
gobelconstruction.com
inducleandecolombia.com
onyxshieldpro.de
servicescenter.net
tvowv55.cfd
voozaak.ru
zopoman.com
911concept.com
aktpl.com
```

Real Cases

Domain Queried

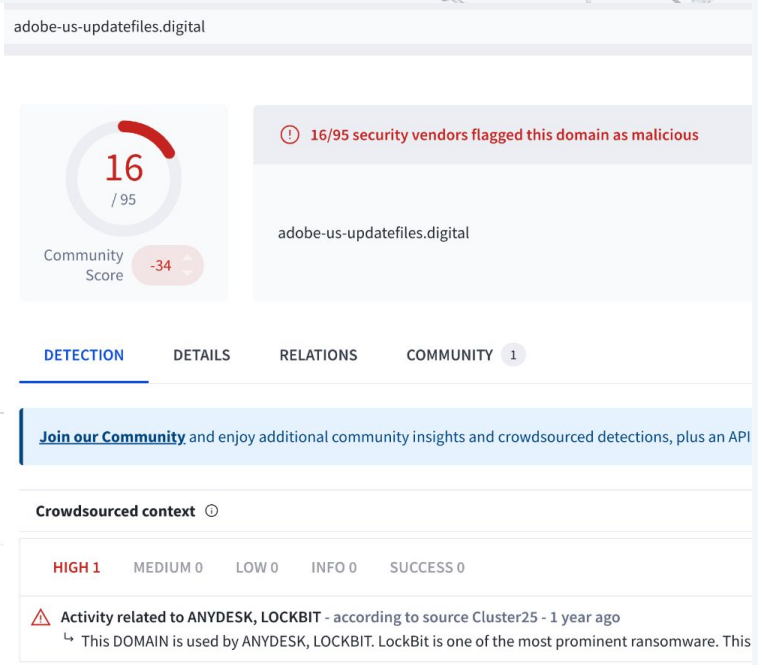
🔷 `adobe-us-updatefiles.digital.`

Answer (Resolved IPs)

- `159.100.251.128` (ASN: **61098**, Org: **Akenes SA**)

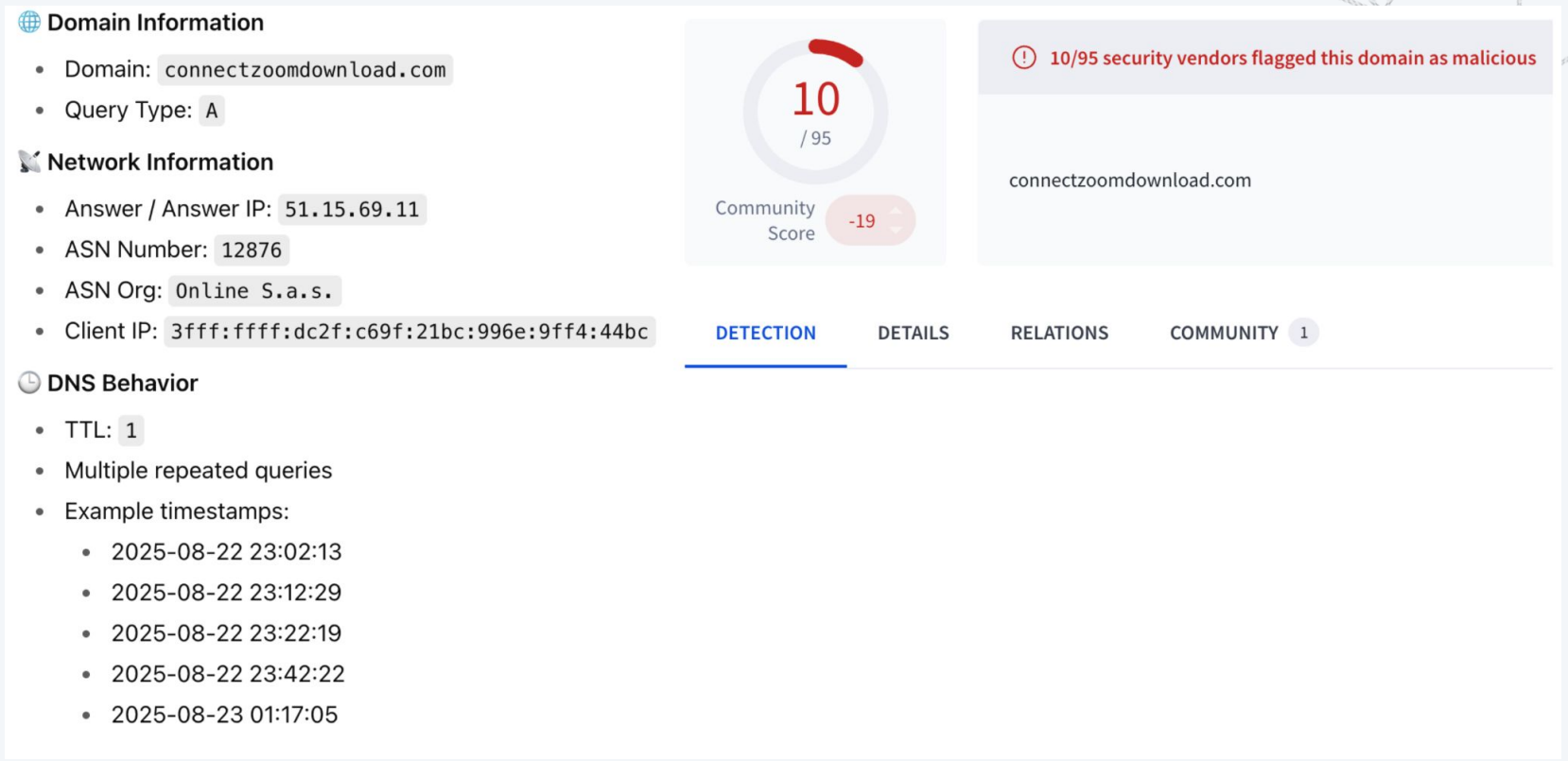
Queries Observed

Client IP (v6)	Timestamp	Resolved IP
`3fff:ffff:35ce:e41a:db47:bf8b:9257:c85d`	2025-08-18 20:20:46 UTC	`159.100.251.128`
`3fff:ffff:3e65:440e:52ac:27a5:a6ea:4749`	2025-08-11 06:43:56 UTC	`159.100.251.128`



Malware Distribution + C2 through Brand Impersonation (Fake Software Update).

Real Cases



Fake Software Update / Installer

Real Cases

🌐 Domain Information

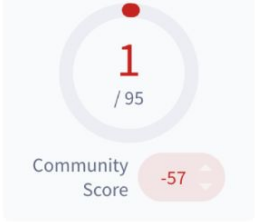
- Domain: `dns0.org`
- Queries contained **long, random-looking subdomains** (e.g., `onsjzji.1.0.komu5z3k5rbhaysn...dns0.org`)
- Query Types: `A`, `AAAA`, `NS`

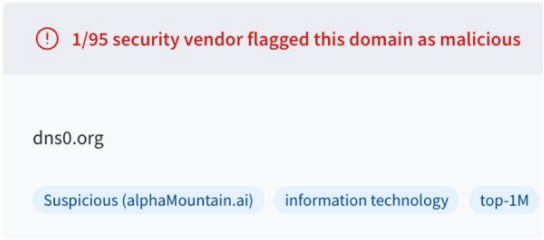
🛰️ Network Information

- Answer IPs:
  - `159.100.251.128` (ASN 61098 – Akenes SA)
  - `92.123.128.187`, `92.123.128.132`, `92.123.128.195` (ASN 16625 – Akamai Technologies, Inc.)
- Several queries returned `NXDOMAIN`
- Clients: IPv6 addresses (e.g., `3fff:ffff:...`)

🕒 DNS Behavior

- Many queries with **NXDOMAIN responses**
- TTL values: `0`, `1`, `19`, `20`
- Subdomains appear algorithmically generated, extremely long and complex
- Activity spans multiple days (Aug 14 – Aug 21, 2025)





DNS Tunneling / Data Exfiltration via DGA-Like Domains

Real Cases

🌐 Domain Information

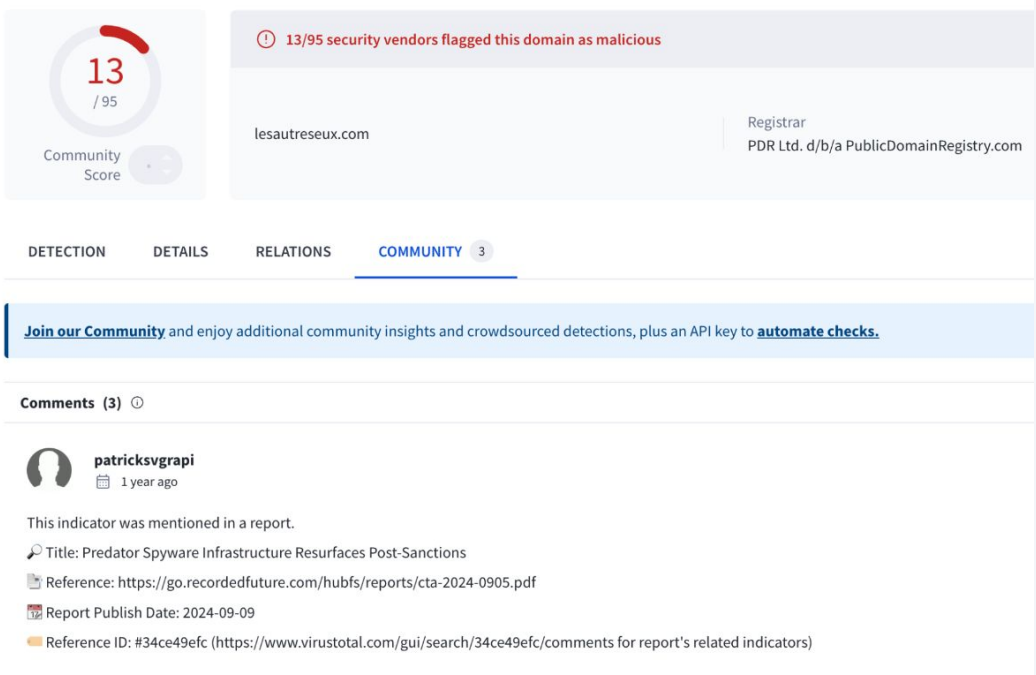
- Domain: `lesautreseux.com`
- Query Type: `A`

📡 Network Information

- Answer / Answer IP: `15.197.240.20`
- ASN Number: `0` (not attributed)
- ASN Org: *(empty)*
- Client IPs:
  - `3fff:ffff:d77c:94cd:71e9:9b37:cff6:13e0`
  - `3fff:ffff:77d5:c340:ca93:b1b7:6002:3684`

🕒 DNS Behavior

- TTL: `600` (10 minutes)
- Multiple queries across several days (Aug 21–23, 2025)



Predator Spyware Infra

Real Cases

Domain Information

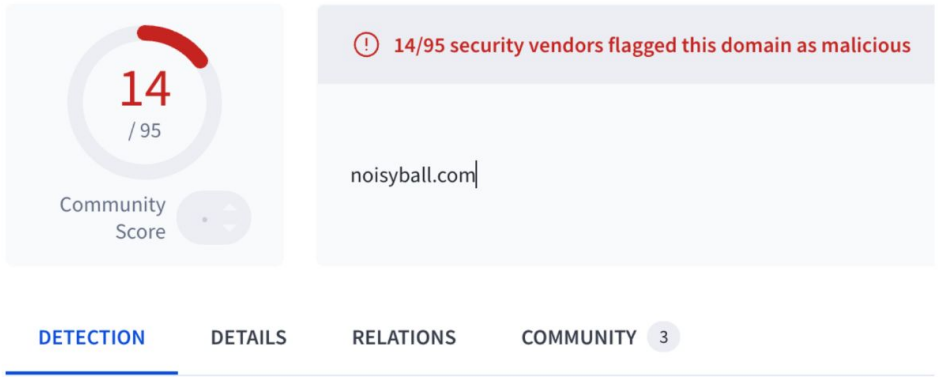
- Domain: `noisyball.com`
- Query Types: `A`, `AAAA`, `DS`
- Country Code: `US`

Network Information

- Answer / Answer IP: `15.197.240.20`
- ASN Number: `0`
- ASN Org: *(empty)*
- Client IPs:
  - `3fff:ffff:d77c:94cd:71e9:9b37:cff6:13e0`
  - `3fff:ffff:39ac:a722:28d0:172e:3fbc:722a`
  - `3fff:ffff:980a:af4b:2bc1:c07f:9ced:2753`
  - `3fff:ffff:1e8e:8807:58d8:d01b:4058:ab55`

DNS Behaviour

- TTL: `600` (10 minutes) for A records
- TTL: `0` for AAAA and DS records



Predator Spyware Delivery / Command & Control (C2)

Real Cases

🕵️ Domain Information

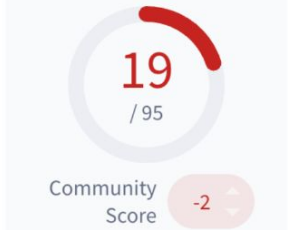
- Domain: `mercharena.biz`
- Query Type: `A`
- TTL: `1`
- Country Code: `LV`

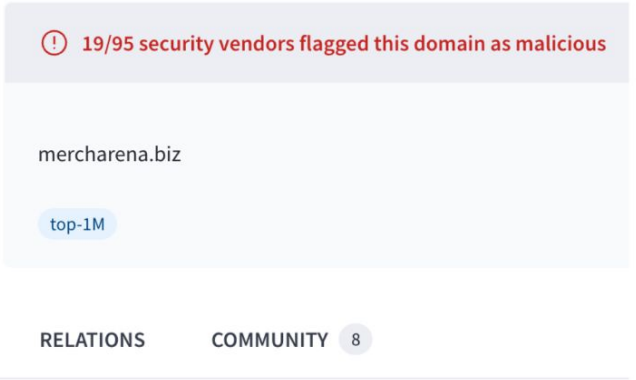
🌐 Network Information

- Answer / Answer IP: `212.93.97.109`
- ASN Number: `24921`
- ASN Org: *Latvijas Mobilais Telefons SIA*
- Client IP: *(not recorded in this log)*

🕒 DNS Behaviour

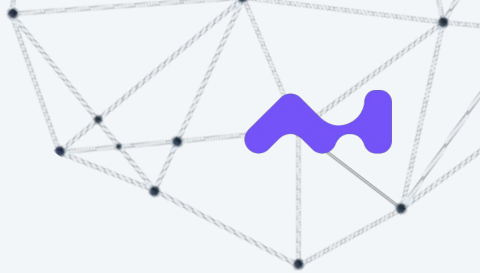
- Very low TTL (`1`) → indicates no caching, often used by malicious infra
- Single A record response
- Query observed on `2025-08-11 16:23:16`



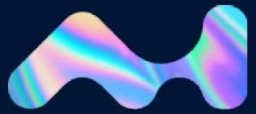


Lumma Stealer / LummaC2 malware

Conclusion



- A strong, resilient, automated and monitored infrastructure is key.
- Choose what to detect and what **not** to detect
    a. Each type is a different problem with different techniques to solve
- False Positive reduction is as critical as good detections



AI CENTER
FEE CTU

Thanks! Questions?

Sebastián García - Tigran Oganesian
sebastian.garcia@agents.fel.cvut.cz

Stratosphere Laboratory.

https://www.stratosphereips.org/

Artificial Intelligence Center

Czech Technical University in Prague