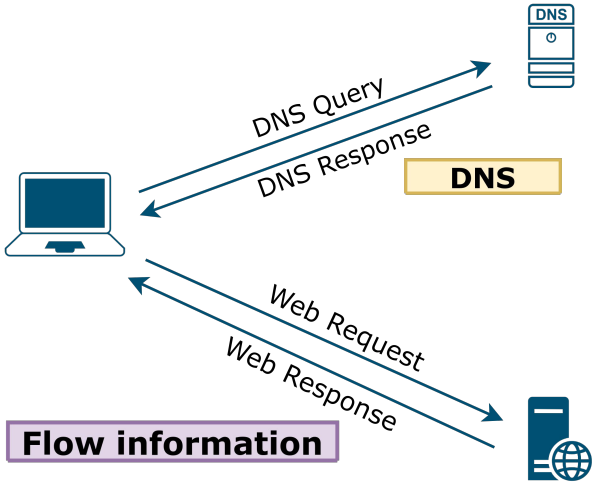
FlowDNS:
Correlating Netflow and DNS Streams at Scale

Aniss Maghsoudlou*, Oliver Gasser*, Ingmar Poese^, Anja Feldmann*
*Max Planck Institute for Informatics, ^ BENOCS Gmbh

CoNEXT 2022

Why Correlating?

- ISPs want to know the services used in their traffic
  - Better negotiation knowing the traffic volume of a service
  - Services cannot be distinguished only by IP if served by CDNs
  - Detect spam traffic
- ISPs gather flow information of their network traffic
- Flow information do not contain domain names
- Flow information + service/domain name

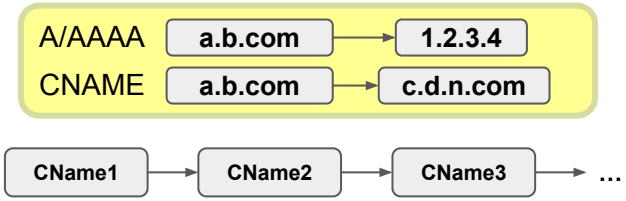


FlowDNS

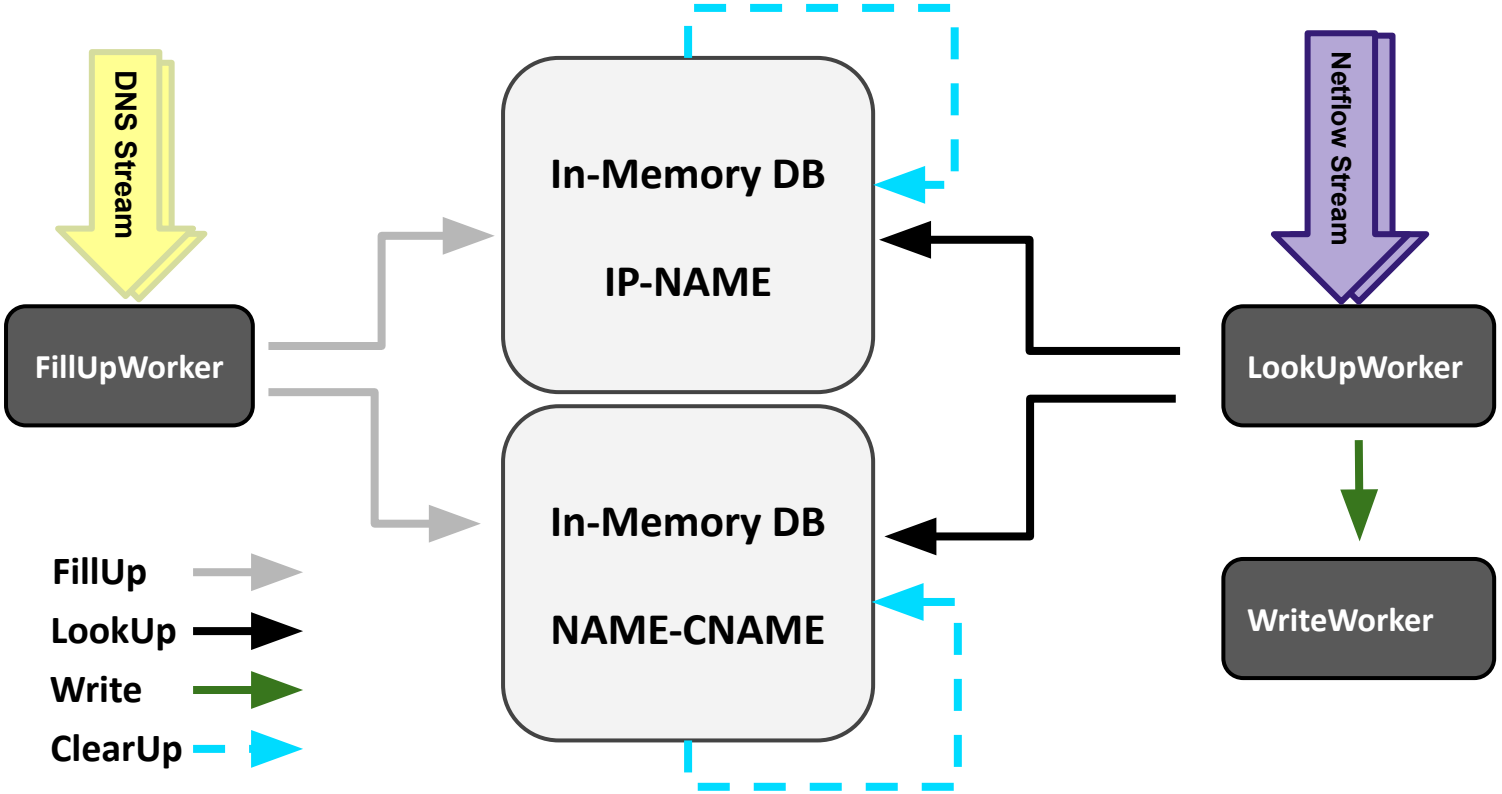
➔ Idea: combine Netflow and DNS live streams

➔ Challenges:

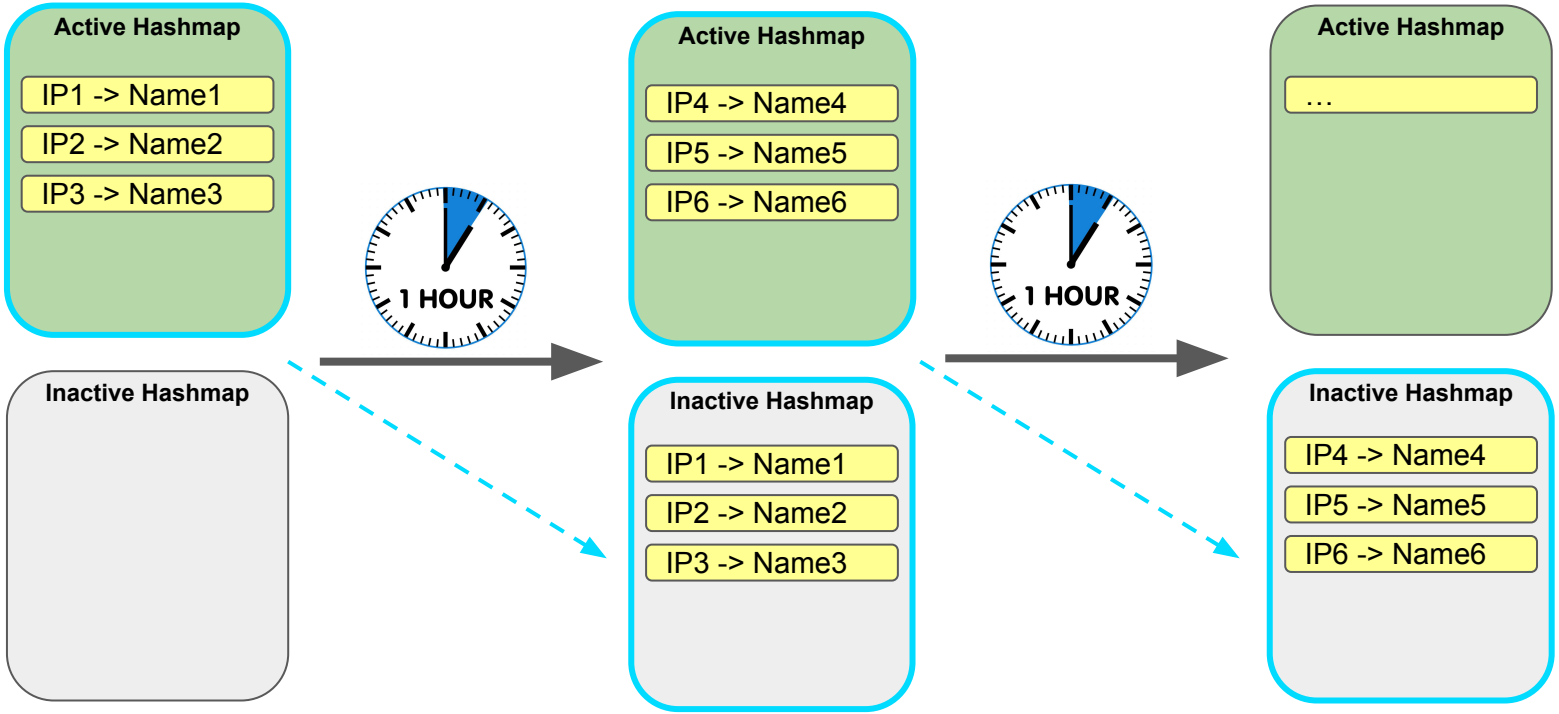
- ◆ DNS records
  - TTL
  - CNAME Chains
- ◆ Infrastructure
  - Live streams buffer overload
  - Limited memory resources



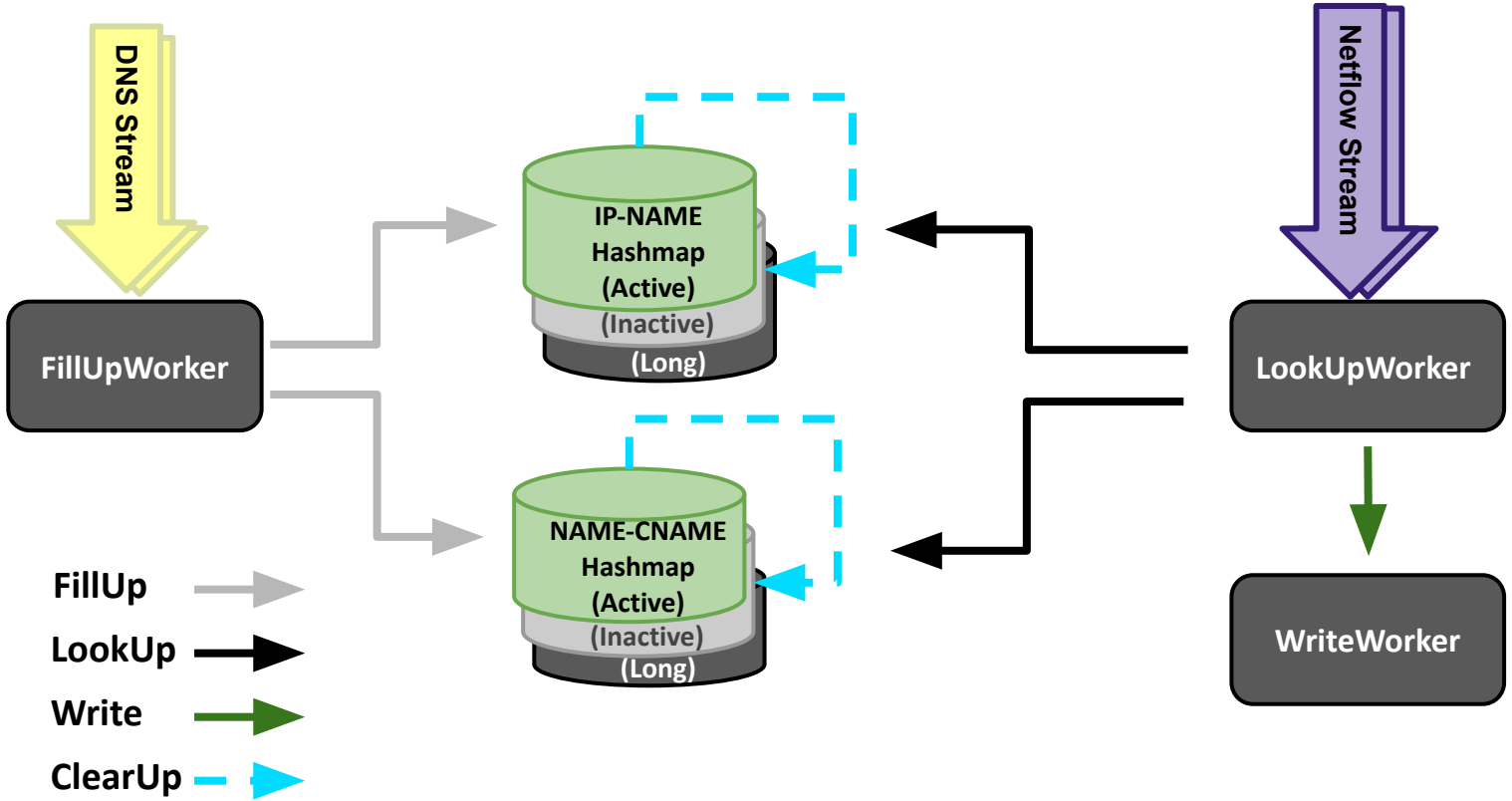
FlowDNS Architecture



Buffer Rotation



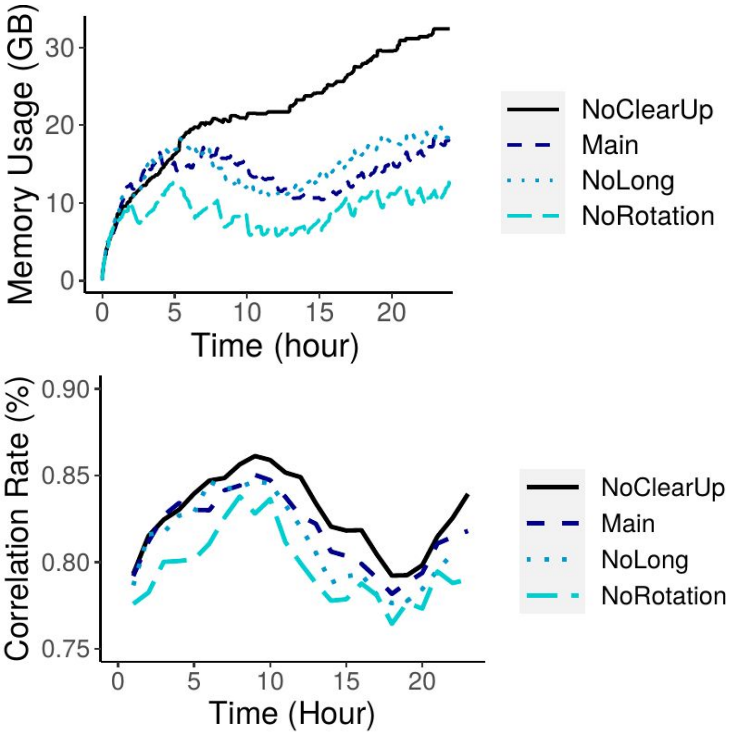
FlowDNS Architecture



Evaluation

- Live Netflow and DNS streams from a large European ISP
- Removing mechanisms 1 by 1
  - Main, NoRotation, NoClearUp, NoLong [, NoSplit in paper]
- Memory usage
- DNS-Netflow Correlation rate

Evaluation



$$Correlation\ Rate\ (CR) = \frac{Correlated\ Traffic\ Volume}{Total\ Traffic\ Volume}$$

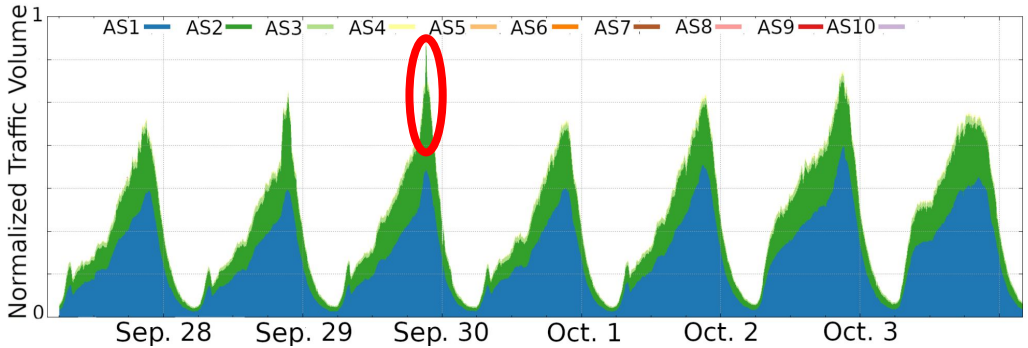
- Clear-up mechanism is necessary
- Buffer rotation increases CR with memory overhead
- Long Hashmaps increase CR without much overhead

Use Cases

- Netflow and DNS data from a large European ISP
- Service-based network provisioning (1 week)
- Spam traffic detection (1 day)

Use Case: Service-based Network Provisioning

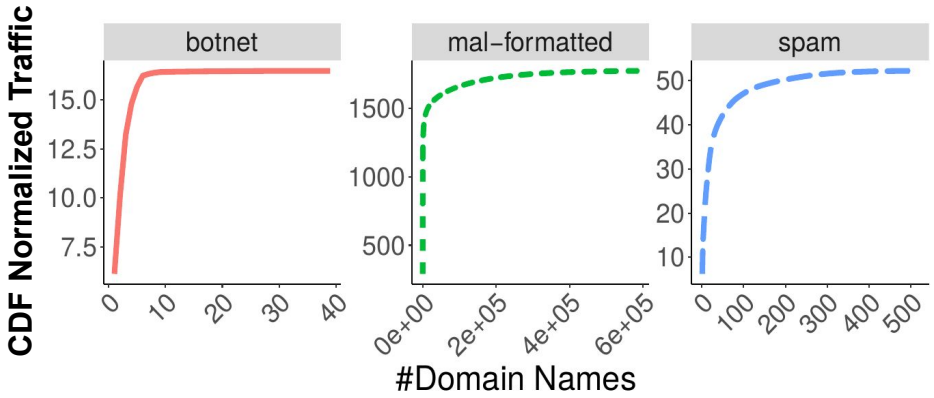
- Filtered traffic based on domain names of Service S1
- Correlated with BGP info
- Insights on how traffic is distributed, e.g. during peak hours



- S1 originated mostly by 2 ASes
- Only AS2 carries the peak on Sep. 29th

Use Case: Spam Traffic Detection

- Checking correlated traffic with
  - Spamhaus DBL domains **spam** **botnet**
  - RFC 1035: implementation and specification of domain names **mal-formatted**



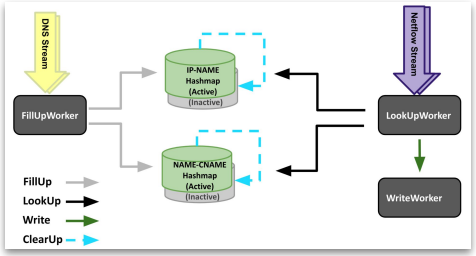
Limited #domains contribute to a large fraction of the traffic.

Lessons learned

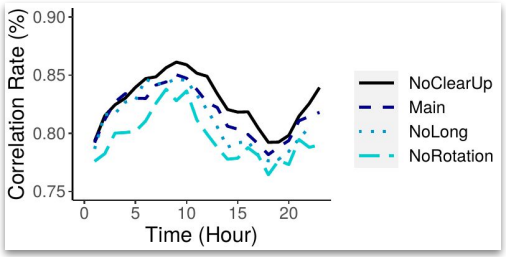
- Applying the exact TTLs leads to buffer overload and higher memory usage
- CNAME chain length needs to be limited
- Several splitting mechanisms may be used, depending on the data
- Buffer rotation and long hashmaps help increasing the correlation rate

Summary

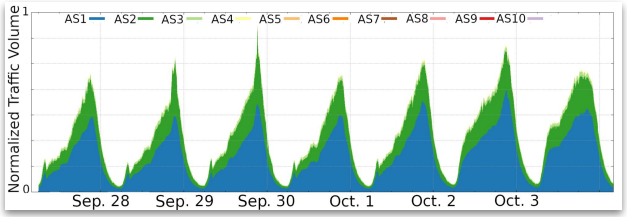
FlowDNS combines DNS and flow data.



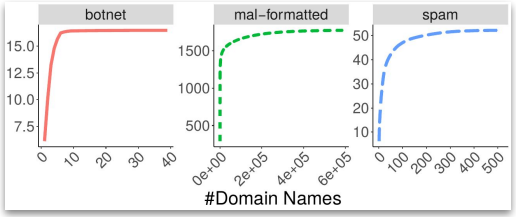
Buffer rotation increases correlation rate.



FlowDNS enables service-based network provisioning.



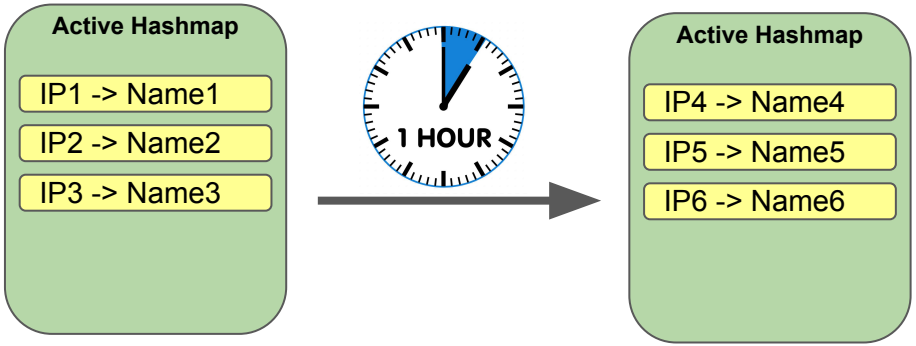
FlowDNS enables spam traffic detection.



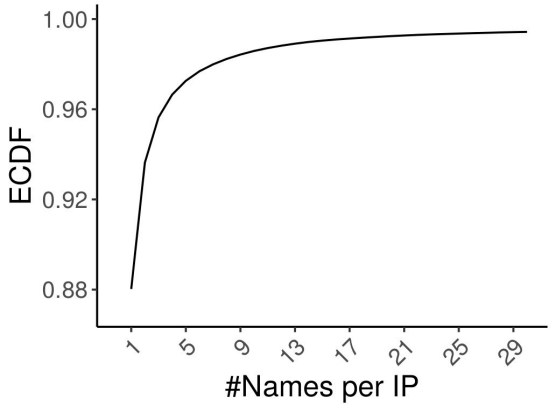
github.com/maganiss/FlowDNS

Back-up Slides

Simple ClearUp



Evaluation

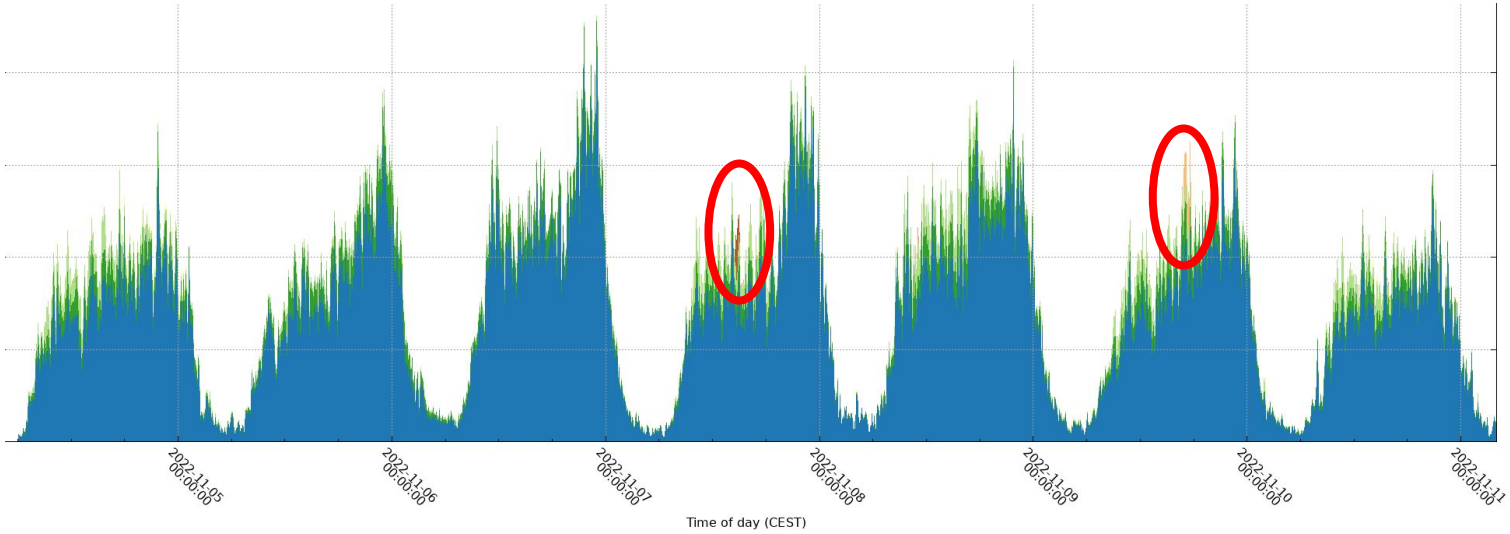


- 88% IPs map to only one domain name
  - min. Accuracy: 88%
- 1 out of every 20 DNS packets is sent to a public DNS resolver
  - 95% coverage

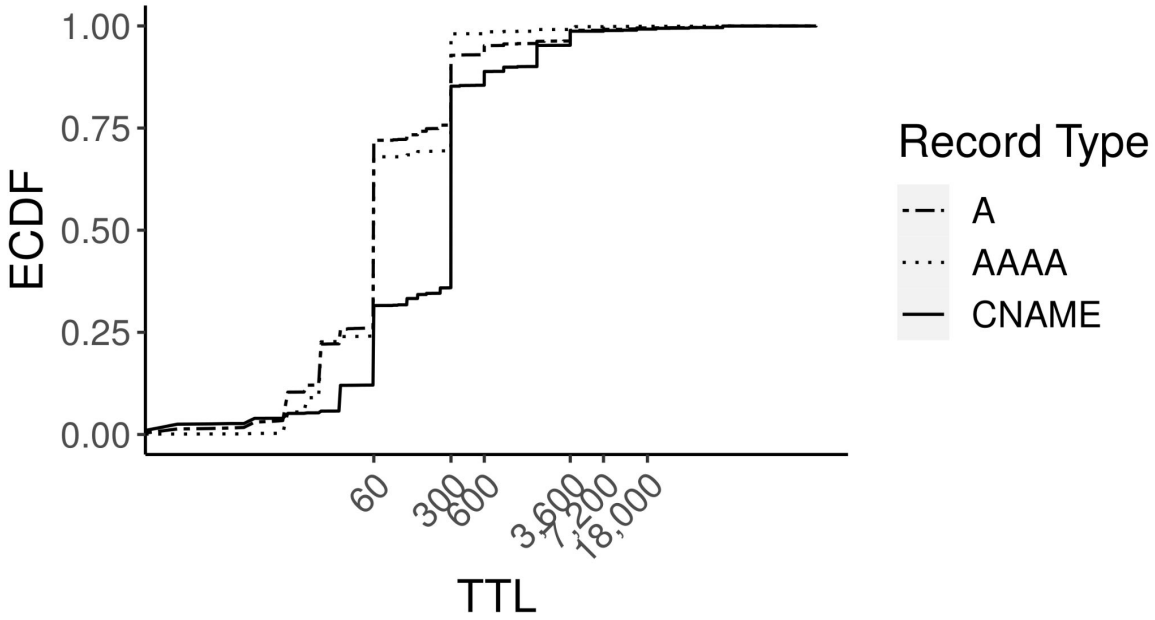
Why not other approaches

- Passive DNS correlation
  - Expired DNS records
  - CDN usage and frequent change of IP-name mapping
- SDN and P4
  - Possible architectural modifications needed
  - Domain name restrictions
  - Encrypted traffic

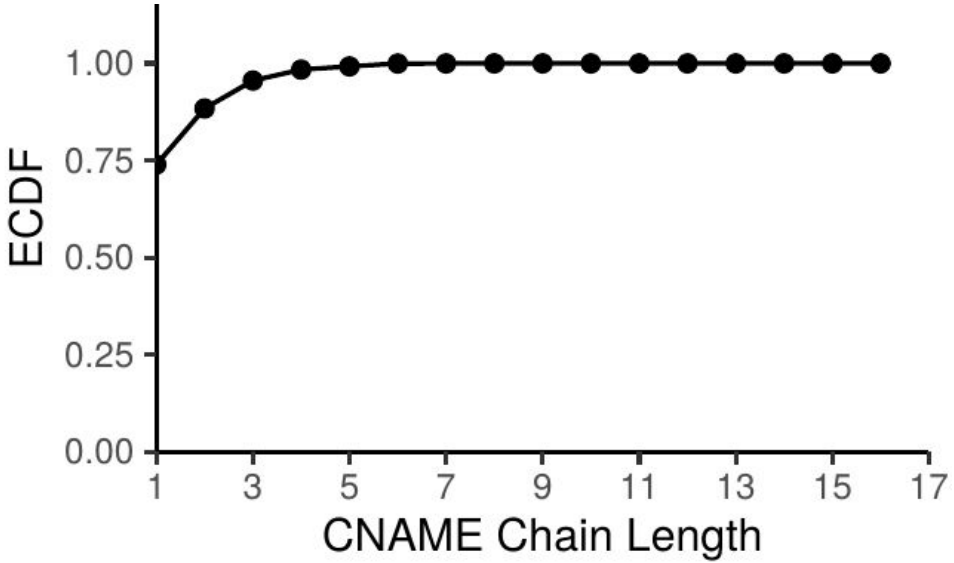
Specific AS showing up occasionally



TTL


1.00
0.75
0.50
0.25
0.00
ECDF
60
300
600
3,600
7,200
18,000
TTL
Record Type
A
AAAA
CNAME

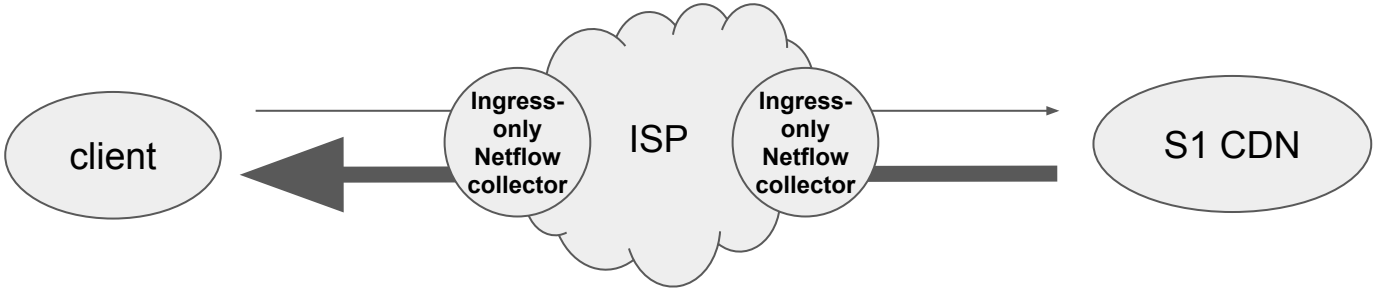
CNAME chain length



Why Spamhaus DBL

- Expiry window of 14 days
- Free (but rate-limited)

Why Correlating SrcIP?



FlowDNS

- Idea: combine Netflow and DNS live streams
- Challenges:
  - Loss on the streams
  - TTL
  - Limited memory resources
  - CNAME chains
- Our approach:
  - Multiple queues to read/write
  - Splitting the DNS records into different hashmaps
  - Clear-up mechanism: Buffer rotation
  - Limiting CNAME chain lookups