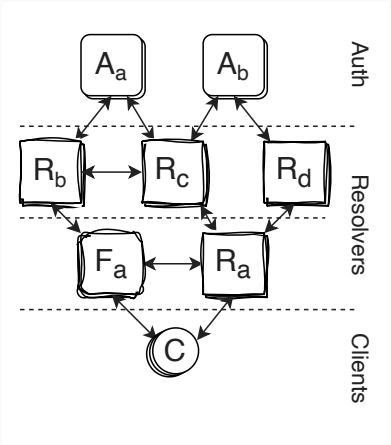
Challenges on Working with DNS Data

Alfred Arouna[1,2] • Mattijs Jonker[3] • Ioana Livadariu[1]
alfred@simula.no

[1]Simula Metropolitan • [2]Oslo Metropolitan University • [3]University of Twente

February 22, 2023 • DNS and Internet Naming Research (DINR) 2023 • Virtual Workshop

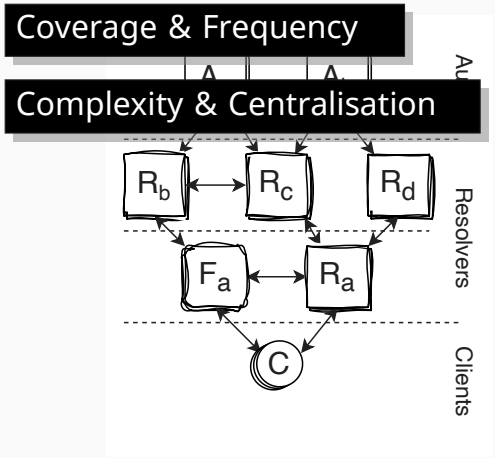
DNS Infrastructure: Highly Distributed and Verbose



(a) Distributed

Simplified Overview of the DNS Infrastructure

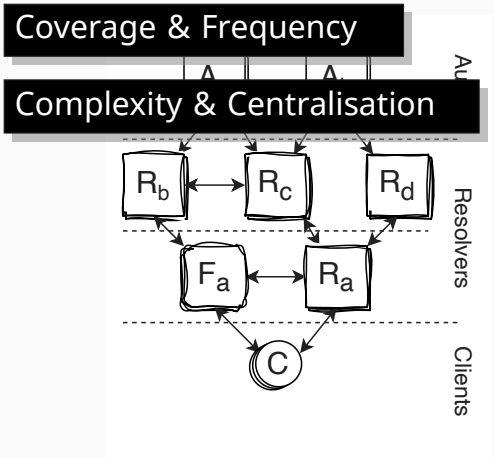
DNS Infrastructure: Highly Distributed and Verbose



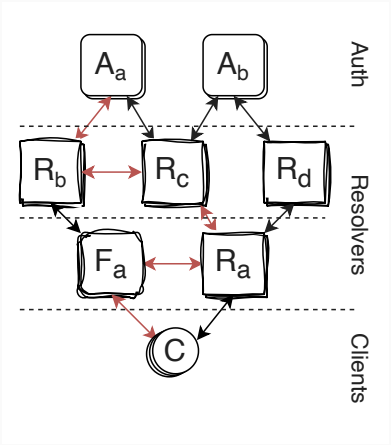
(a) Distributed

Simplified Overview of the DNS Infrastructure

DNS Infrastructure: Highly Distributed and Verbose



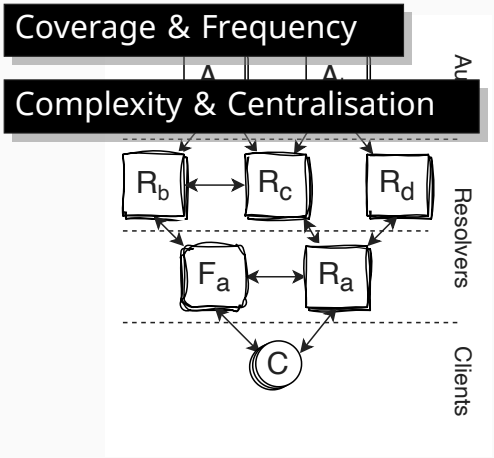
(a) Distributed



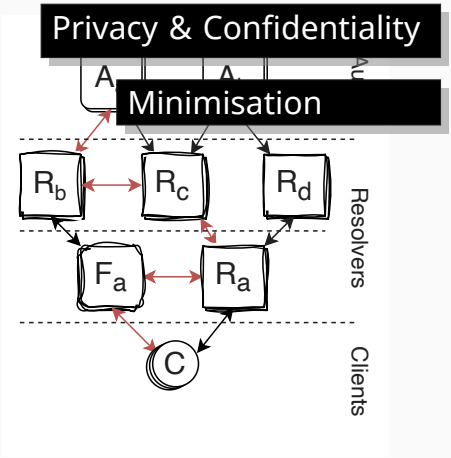
(b) Verbose

Simplified Overview of the DNS Infrastructure

DNS Infrastructure: Highly Distributed and Verbose



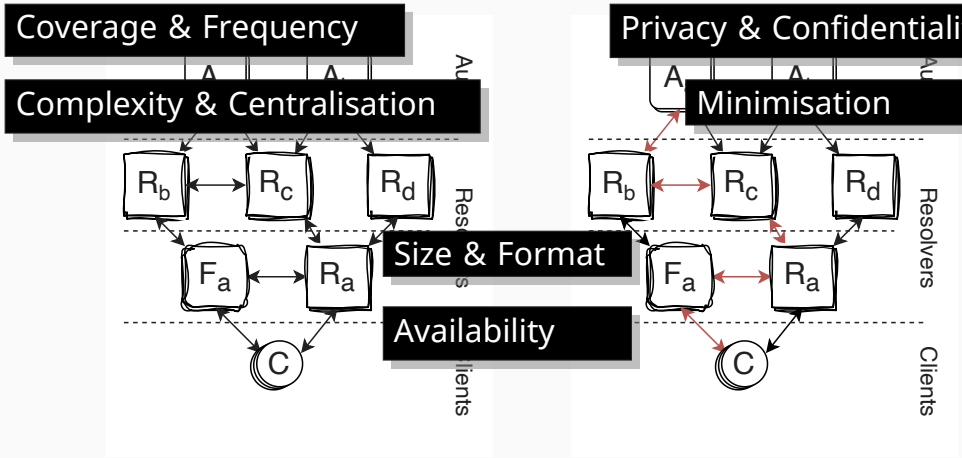
(a) Distributed



(b) Verbose

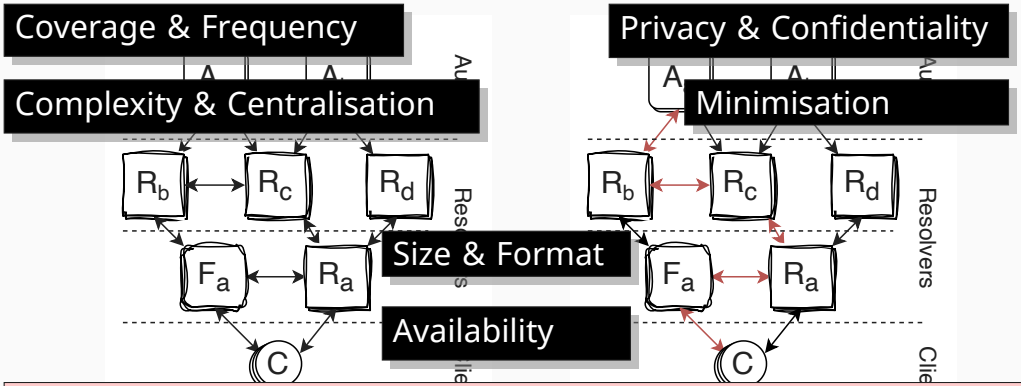
Simplified Overview of the DNS Infrastructure

DNS Infrastructure: Highly Distributed and Verbose



Simplified Overview of the DNS Infrastructure

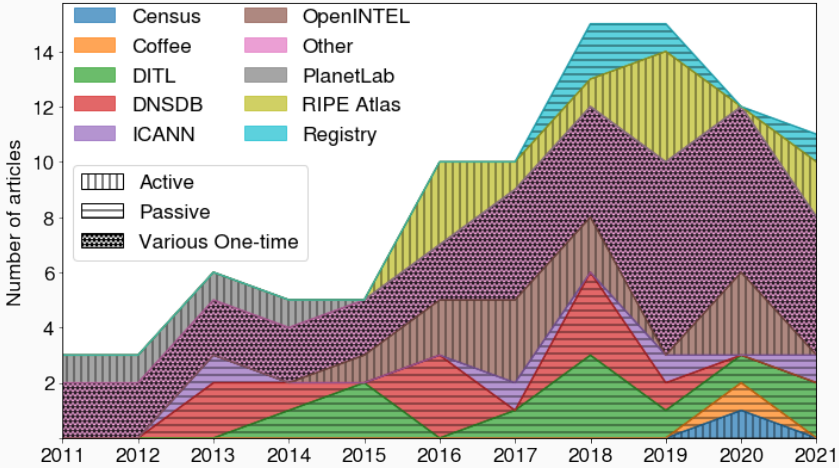
DNS Infrastructure: Highly Distributed and Verbose



How has the DNS research community so far managed to address these data collection challenges?
Evaluate existing datasets usage from most impacting publications on DNS.

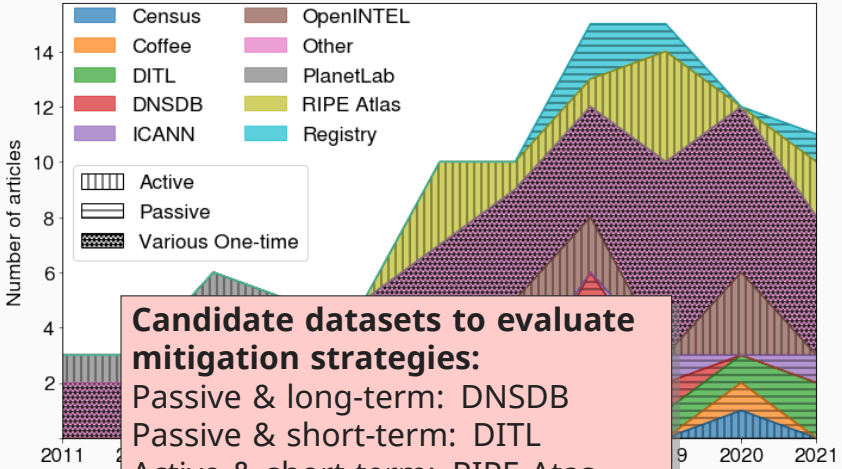
Simplified Overview of the DNS Infrastructure

Popular DNS Datasets



DNS datasets usage over the last 10 years. Although the increase of publications can be correlated with the rise of long-term datasets, DNS researchers relied in majority on one-time snapshot of the state of (parts of) the DNS.

Popular DNS Datasets



Candidate datasets to evaluate mitigation strategies:
Passive & long-term: DNSDB
Passive & short-term: DITL
Active & short-term: RIPE Atas
Active & long-term: OpenINTEL

DNS datasets usa crease of publications can be correlated with the rise of long-term datasets; DNS researchers relied in majority on one-time snapshot of the state of (parts of) the DNS.

Distributed: Addressing Challenges

Table: Mitigation approaches for challenges resulting from the distributed DNS infrastructure. Long-term datasets have large coverage based on consistent data collection frequency.

	Coverage	**Frequency**	**Complexity**	**Centralisation**
OpenINTEL	Large	Consistent	High: Workers	Low
RIPE Atlas	Limited	Variable	High: probes	Low
DNSDB	Large	Consistent	High: VPs	Medium
DITL	Limited	Variable	High:servers	Medium

Distributed: Addressing Challenges

Table: Mitigation approaches for challenges resulting from the distributed DNS infrastructure. Long-term datasets have large coverage based on consistent data collection frequency.

	Coverage	**Frequency**	**Complexity**	**Centralisation**
OpenINTEL	Large	Consistent	High: Workers	Low
RIPE Atlas	Limited	Variable	High: probes	Low

Dataset should apply a distributed systematic collection data at an established frequency with minimal burden on the DNS.

Verbose: Addressing Challenges

Table: Mitigation approaches for challenges resulting from DNS verbosity. Addressing privacy and confidentiality is still challenging. However, minimization can help to reduce privacy and confidentiality risks. Active dataset by controlling the resolver are less impacted by minimization.

	Privacy	**Confidentiality**	**Minimization**
OpenINTEL	High: zones	High: zones	Medium
RIPE Atlas	High: user	High: users	Medium
DNSDB	Medium[1]	Medium[1]	High
DITL	Medium[2]	Medium[2]	High

[1] resolver-authoritative
[2] root-servers

Verbose: Addressing Challenges

Table: Mitigation approaches for challenges resulting from DNS verbosity. Addressing privacy and confidentiality is still challenging. However, minimization can help to reduce privacy and confidentiality risks. Active dataset by controlling the resolver are less impacted by minimization.

	Privacy	**Confidentiality**	**Minimization**
OpenINTEL	High: zones	High: zones	Medium
RIPE Atlas	High: user	High: users	Medium

Dataset should consider the increasing adoption of the principle of minimum disclosure to minimize privacy and confidentiality risks

[1] resolver-authoritative
[2] root-servers

Distributed & Verbose: Addressing Challenges

Table: Mitigation approaches for challenges resulting from DNS distributed infrastructure and verbosity. The variability in data formats limits for large-scale and long-term analysis. Active dataset are publicly available while passive dataset access are on demand.

	Format	**Size**	**Availability**
OpenINTEL	Avro/Parquet	+10TB[3]	Public
RIPE Atlas	JSON	+25TB[4]	Public
DNSDB	ISC/dnsqr	n.a.	Restricted
DITL	PCAP	n.a.	Restricted

[3] 10TB of compressed data as of Feb. 2015 (1 year data).
[4] July 2015 in Hadoop/HBase (5 years of data).

Distributed & Verbose: Addressing Challenges

Table: Mitigation approaches for challenges resulting from DNS distributed infrastructure and verbosity. The variability in data formats limits for large-scale and long-term analysis. Active dataset are publicly available while passive dataset access are on demand.

	Format	**Size**	**Availability**
OpenINTEL	Avro/Parquet	+10TB[3]	Public
RIPE Atlas	JSON	+25TB[4]	Public

The variability in data formats limits for large-scale and long-term analysis. Compacted-DNS (C-DNS) seems to be a good candidate for a common data format. However, Avro/Parquet has been proven for large-scale analysis as part of DNS big data pipeline.

[3] 10TB of compressed data as of Feb. 2015 (1 year data).

[4] July 2015 in Hadoop/HBase (5 years of data).

Comments?

Ideas to improve the work?
Questions?