Enumerating Privacy Leaks in DNS Data Collected above the Recursive

Basileal Imana[1], Aleksandra Korolova[1] and John Heidemann[2]

[1]University of Southern California [2]USC/Information Science Institute

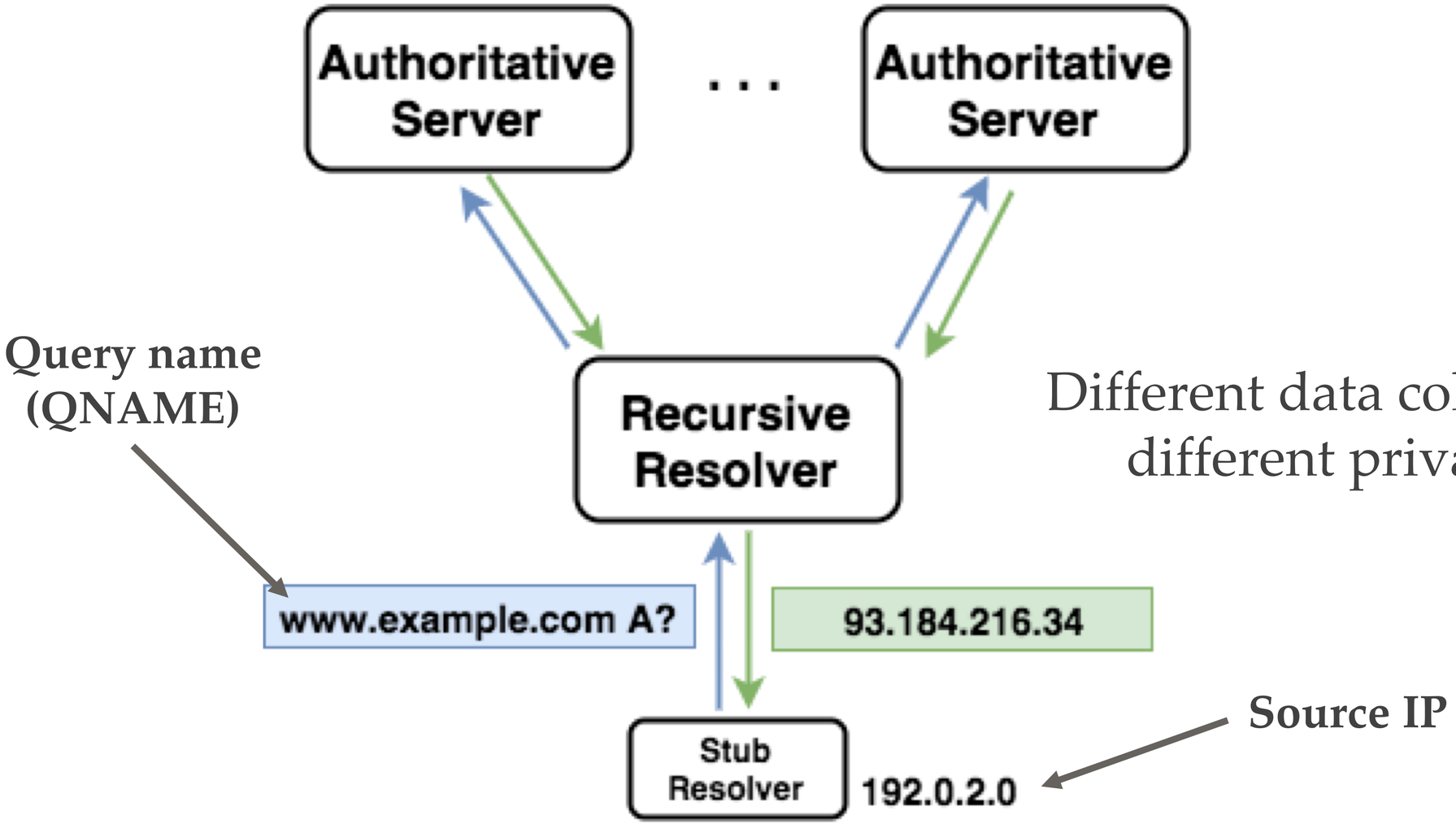
NDSS DNS Privacy Workshop

February 2018

Your DNS queries can say a lot about you!

Almost all activities on the Internet start with a DNS query

Data in DNS queries



Different data collection points have different privacy implications

Why study DNS privacy?

- Researchers and operators analyze and share DNS data
- Queries in data often represent end-users actions
- Privacy risks not fully understood
- Some users may care about their privacy

Our Contribution

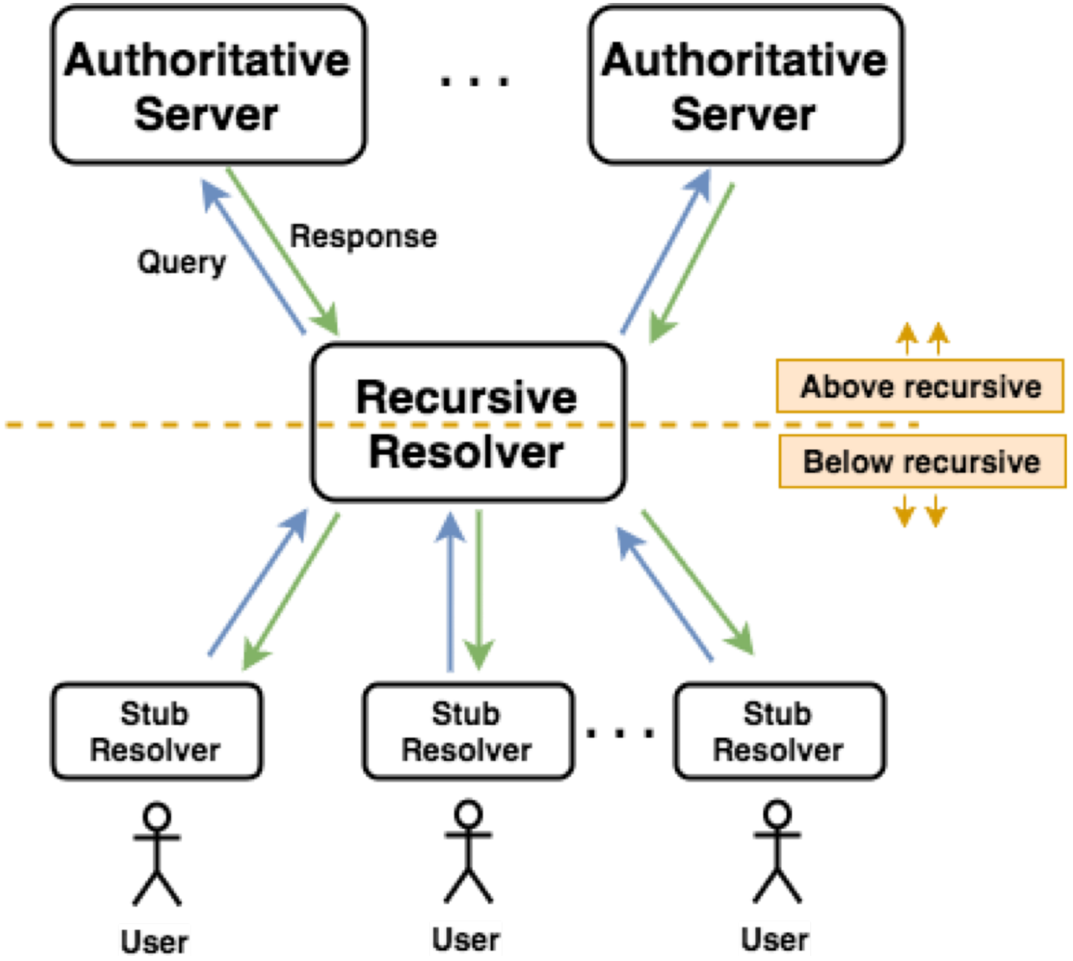
- Enumerate classes of privacy leaks in query names **above the recursive**
- Examine root DNS data to measure **how often** two types of leaks appear in real-world traffic

What has been done?

- IETF DPRIVE working group
- Understanding risks
  - DNS Privacy Considerations (RFC 7626): eavesdropping and data misuse
- Mitigations
  - DNS over TLS (RFC 7858): encryption
  - Query minimization (RFC 7816): reduce information disclosure

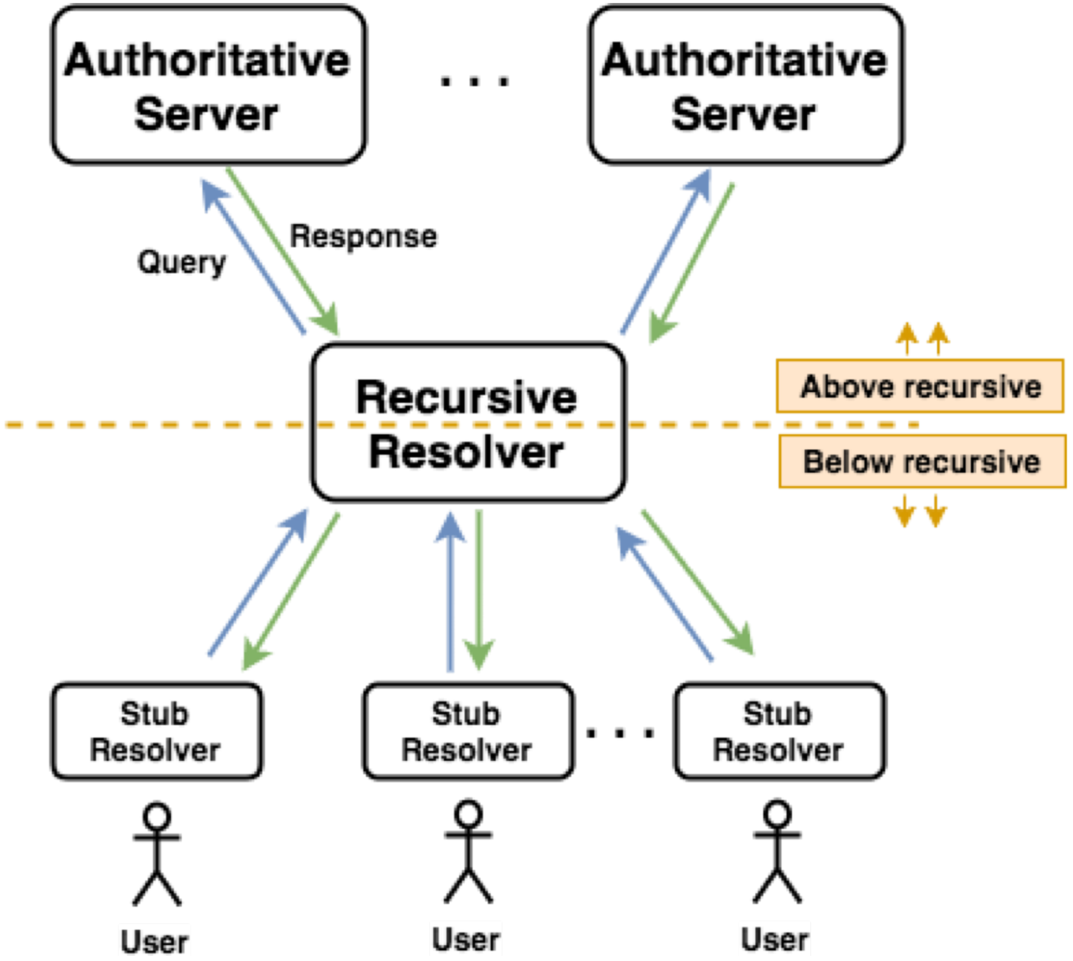
Above vs. below the recursive

- Prior work focused on securing data below the recursive
  - E.g., Stubby
- Does data above the recursive pose minimal privacy risk due to aggregation?
  - What types of queries leak information despite aggregation?
  - We re-examine this assumption



Above vs. below the recursive

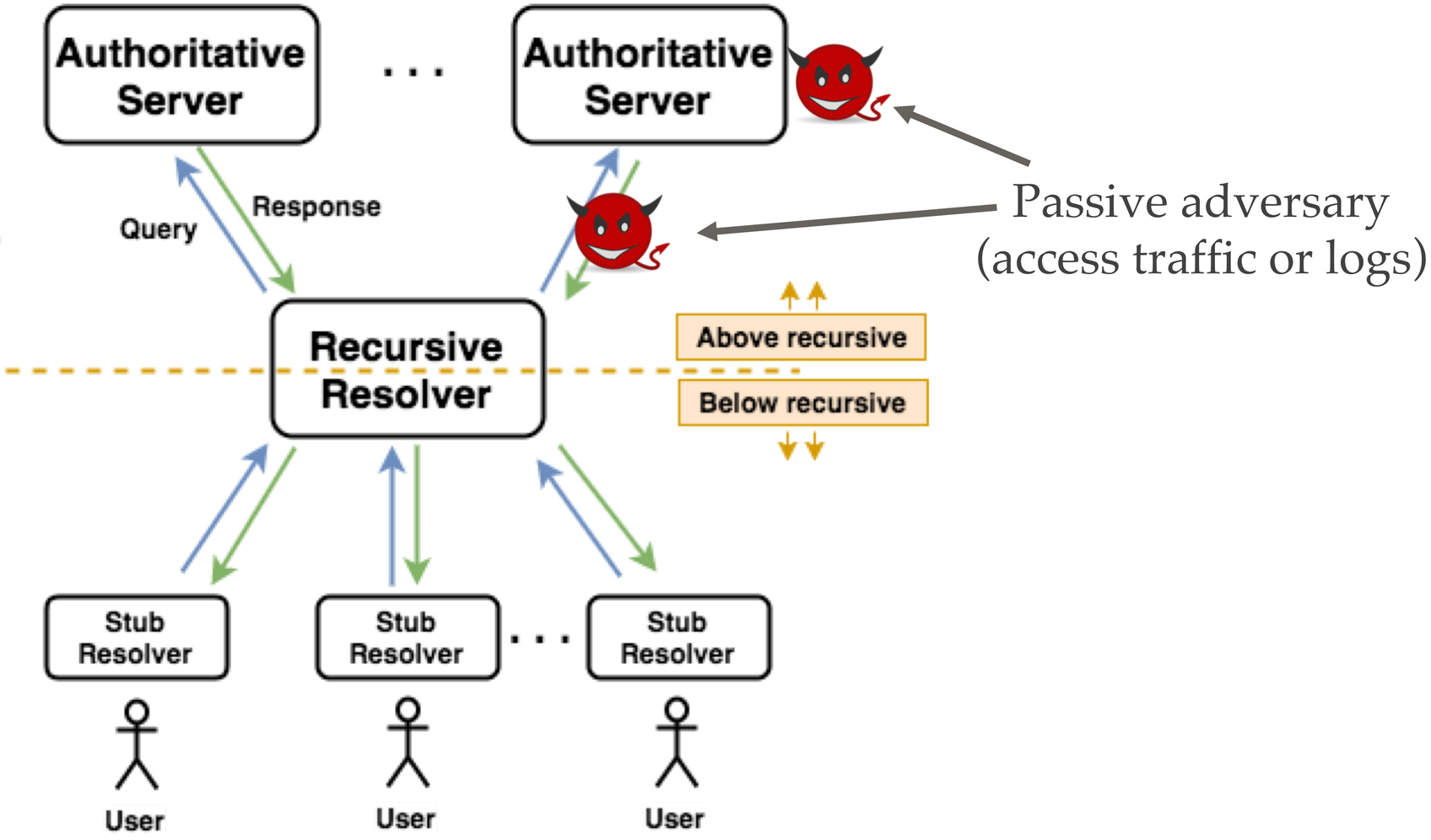
- Prior work focused on securing data below the recursive
  - E.g., Stubby
- Does data above the recursive pose minimal privacy risk due to aggregation?
  - What types of queries leak information despite aggregation?
  - We re-examine this assumption



Enumerating Leaks

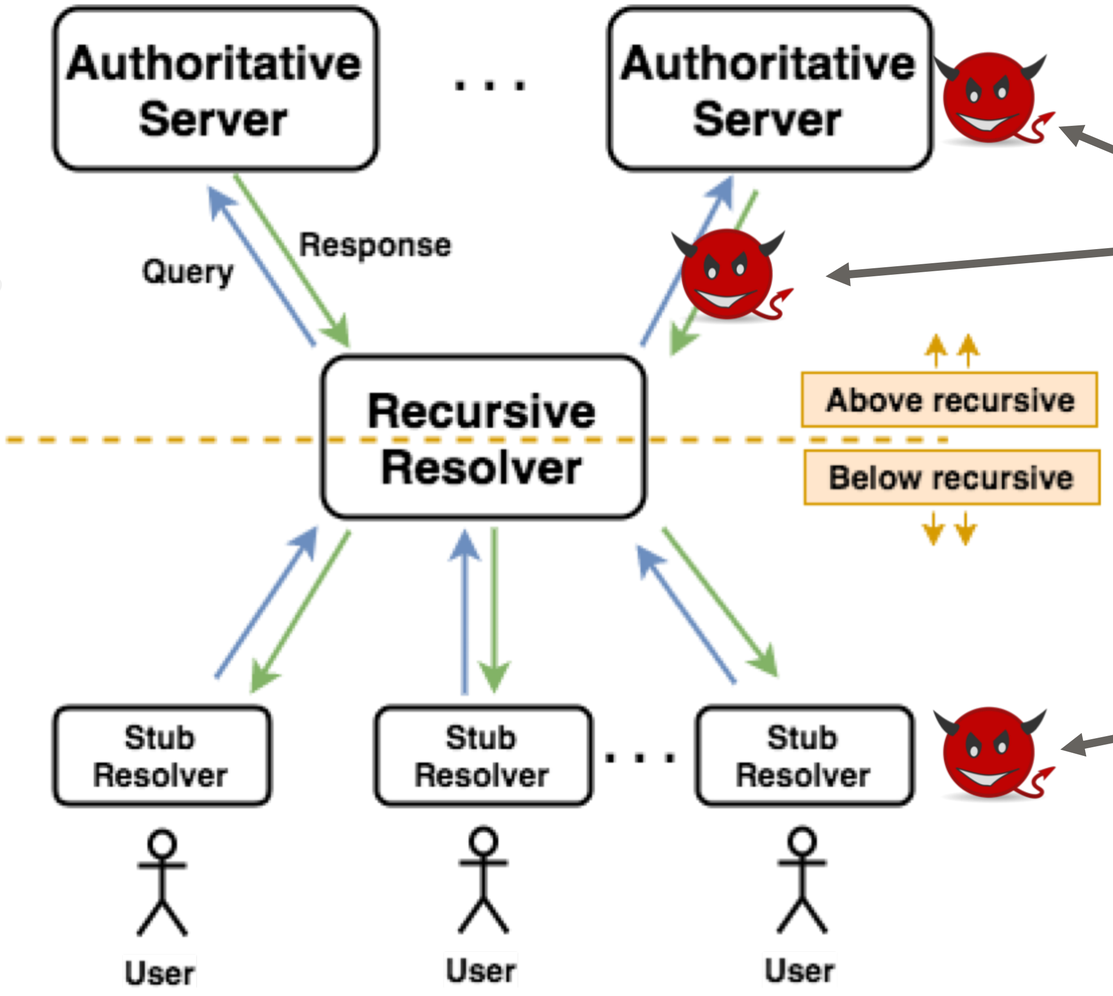
Threat Model

- ISPs, DNS operators, researchers, compromised servers
- Target an individual
- Target a group



Threat Model

- ISPs, DNS operators, researchers, compromised servers
- Target an individual
- Target a group

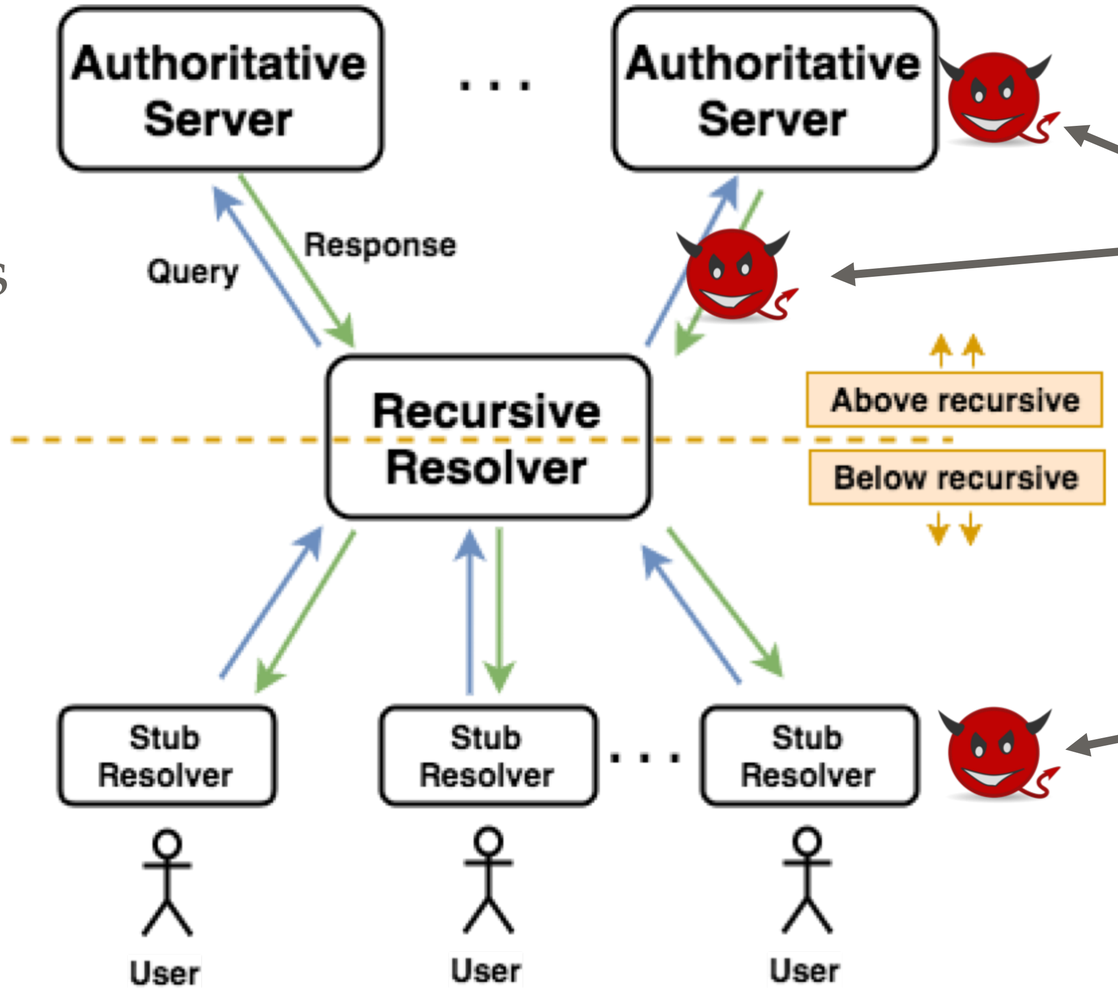


Passive adversary (access traffic or logs)

Active adversary (make a user inject queries)

Threat Model

- ISPs, DNS operators, researchers, compromised servers
- Target an individual
- Target a group

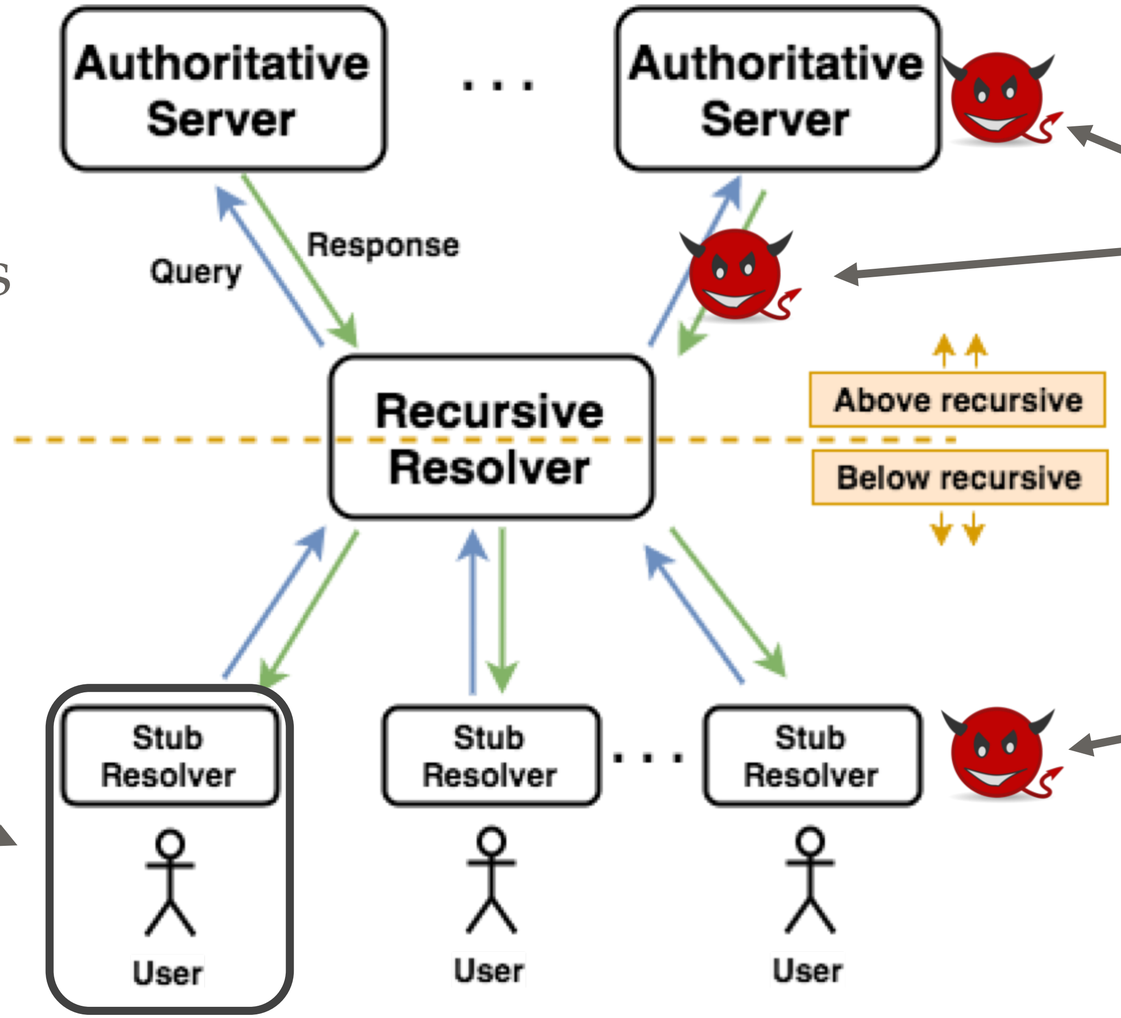


Passive adversary (access traffic or logs)

Active adversary (make a user inject queries)

Threat Model

- ISPs, DNS operators, researchers, compromised servers
- Target an individual
- Target a group

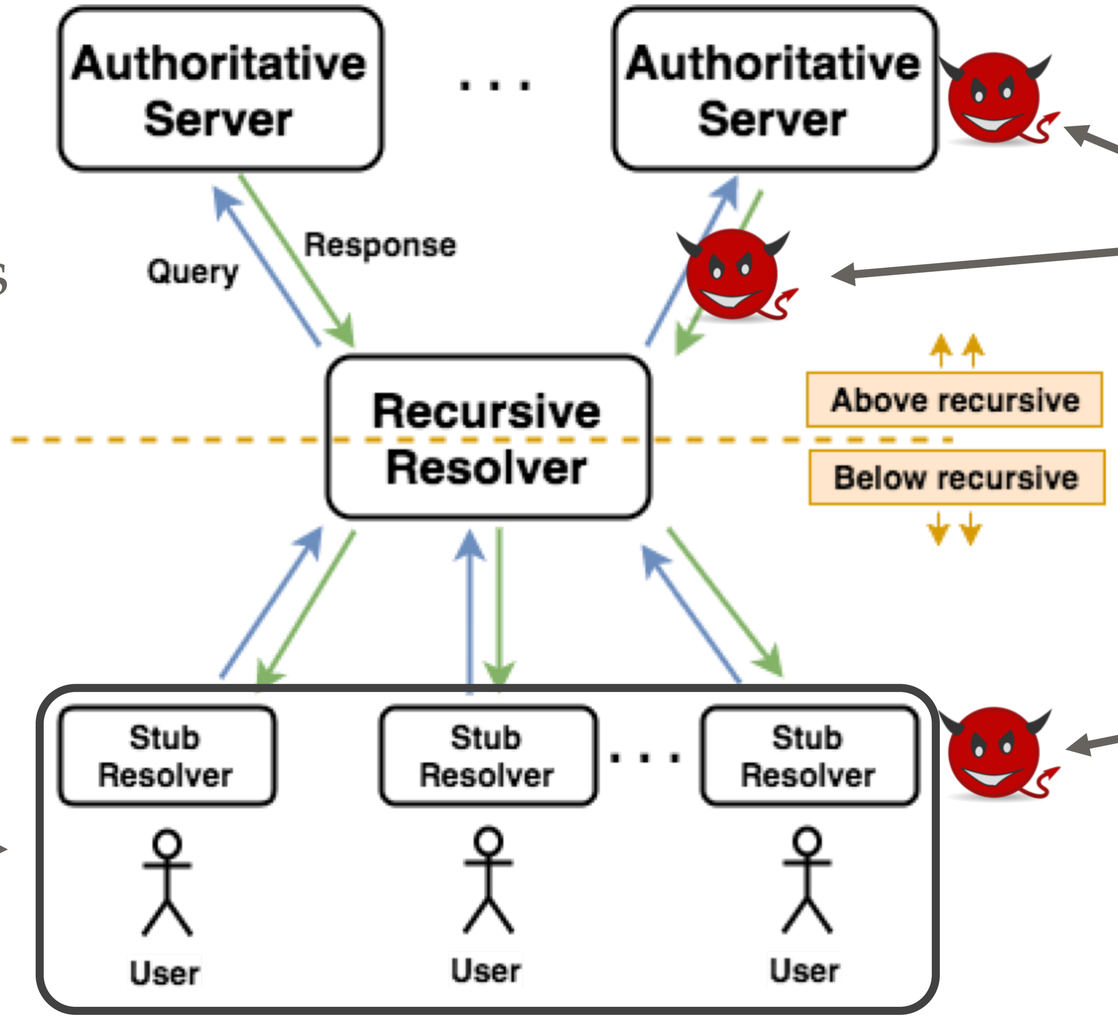


Passive adversary (access traffic or logs)

Active adversary (make a user inject queries)

Threat Model

- ISPs, DNS operators, researchers, compromised servers
- Target an individual
- Target a group



Passive adversary (access traffic or logs)

Active adversary (make a user inject queries)

Enumerating Leaks

1. Trackable names
2. IP addresses in QNAMES
3. Sensitive domain names

} Passive

4. Query injection → Active

Passive Adversary

1. Trackable Names

- A unique identifier associated with an individual/group
- E.g., a user who uses **last-name.example.com** to host email or calendar services
  - **clintonemail.com** was Hillary Clinton's private server

- Leaks despite aggregation at the recursive
- Such attack possible when association of domain to individual is known

2. IP Addresses in QNAMEs

- Not all IP addresses are equally sensitive (e.g., static vs. dynamic)
- Common examples
  - Reverse DNS queries (rDNS), **0.2.0.192.in-addr.arpa.**
  - IP based reputation system (DNSBL), **0.2.0.192.sbl.spamhaus.org.**
  - Customer provided equipment (CPE), **192-0-2-0.dedicated.static.sonic.net.**
- Privacy implications depend on
  - how often addresses change
  - availability of ISP data that maps IPs to individuals

3. Sensitive Domains Names

- Use domains pertaining to health, lifestyle, ethnicity, etc., to profile users
- Examples:
  - Alcoholic Anonymous (**aa.org**)
  - Sexual preference (**gaycities.com)**
  - Lifestyle (**veggieboard.com**)
- Happens when there is insufficient aggregation

Active Adversary

4. Query Injection

- Cause a user to perform a query
- Query that pierces through the recursive and reaches attacker's server
- A similar technique used by Netalyzr [Kreibich2010]
  - e.g. **369839a0-32153-dcf252d3-821e-46e1b706**.netalyzr.icsi.berkeley.edu.

- Learn about user's resolver or when certain activity happens on user's machine

Analyzing root DNS data

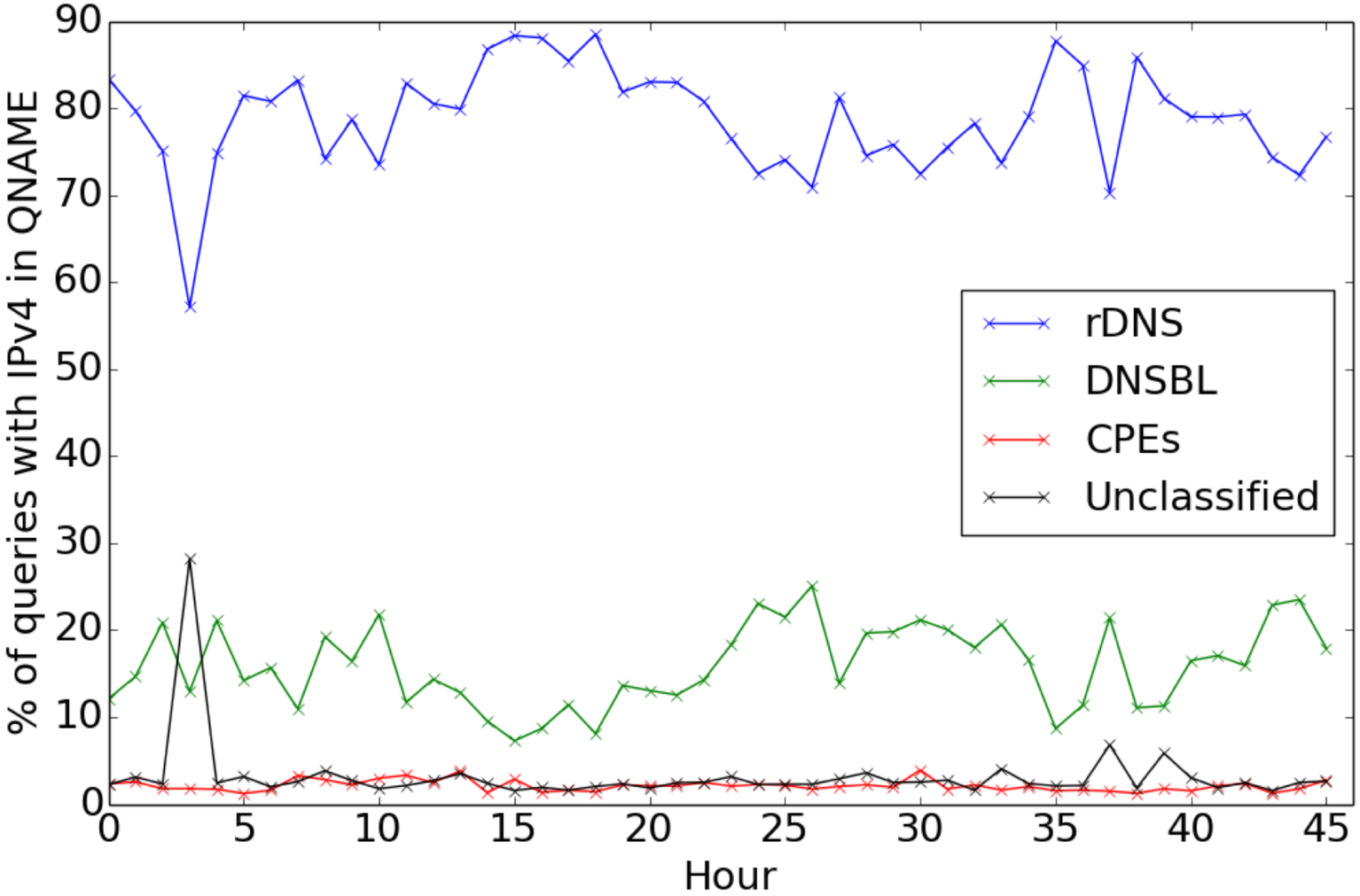
Data

- 48 hours b-root DITL data from April 2017
- Sampled ~100k DNS messages from approximately every hour

Dataset	Duration	Queries	Sampled and filtered
B-ditl-2017	48 hours	5.7×10^9	1,085,703

- Questions
  - How often do IP addresses appear in QNAMEs?
  - How common are queries to sensitive domain?

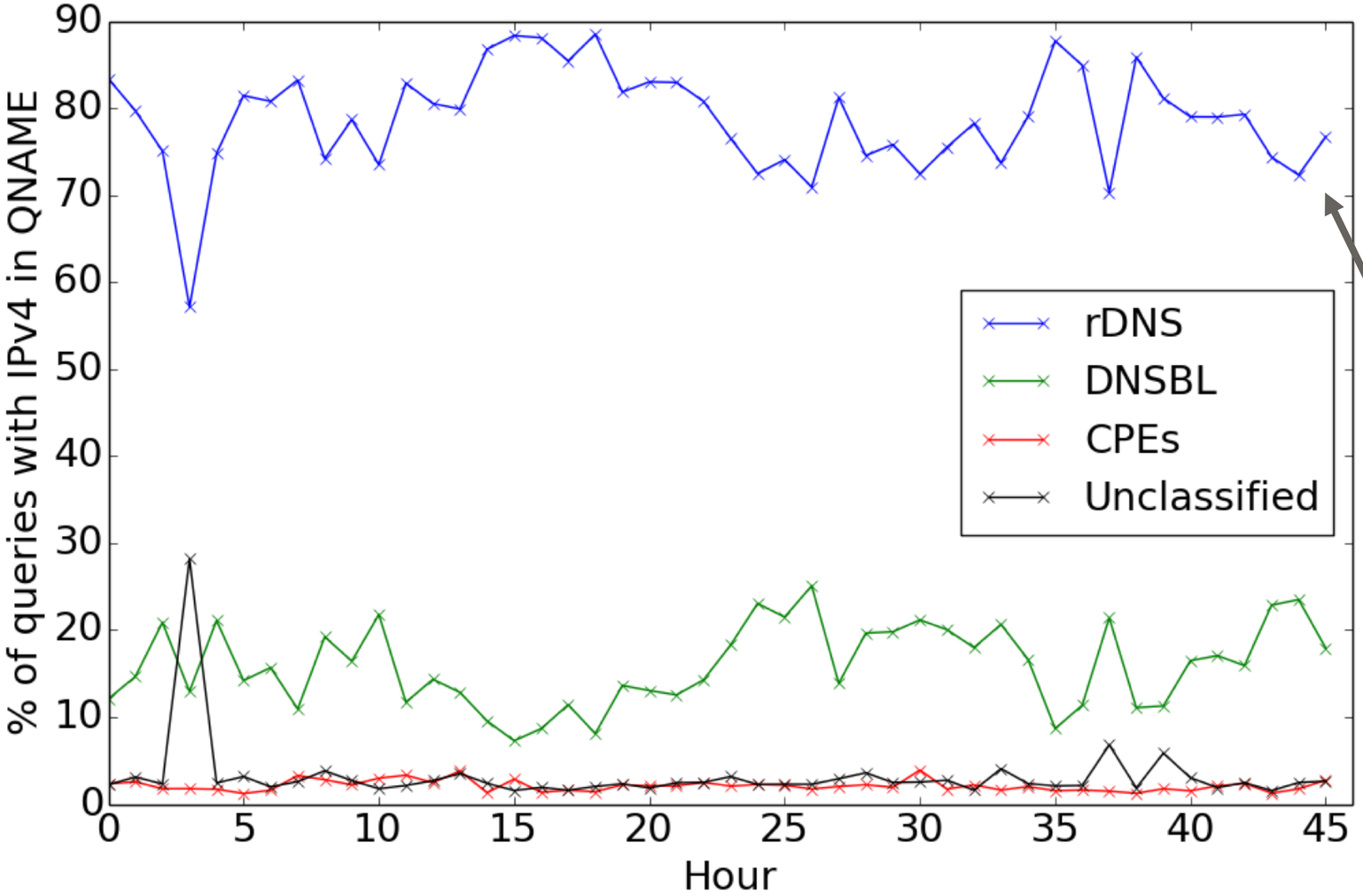
Result: IP Addresses in QNAMEs



Queries with IPs in QNAMEs

- In sample:
  - IPv4: 42,846 queries (3.9%)
  - IPv6: 863 queries (0.08%)
- Estimate for total traffic
  - **~57 million queries/day**

Result: IP Addresses in QNAMEs

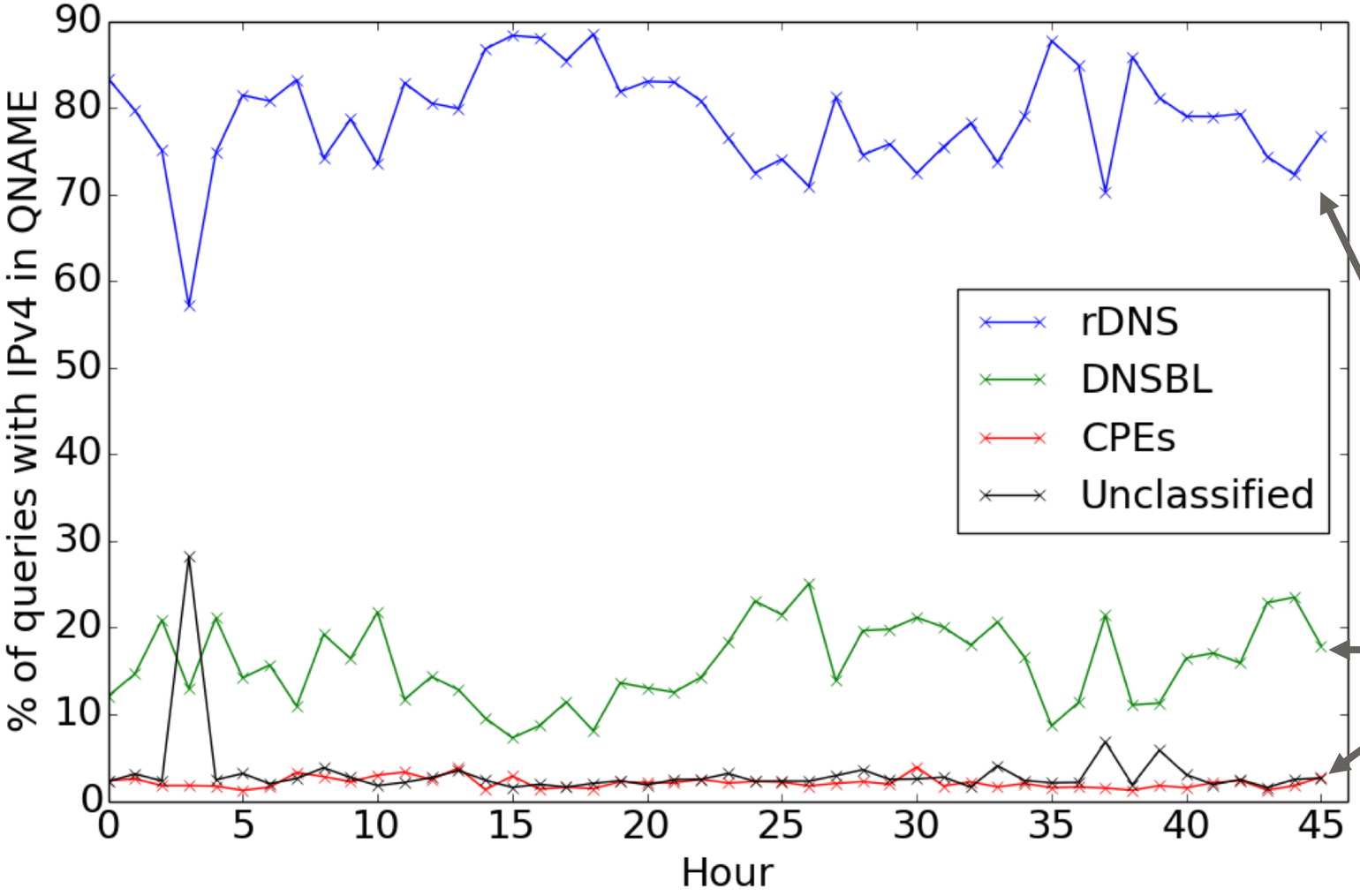


Queries with IPs in QNAMEs

- In sample:
  - IPv4: 42,846 queries (3.9%)
  - IPv6: 863 queries (0.08%)
- Estimate for total traffic
  - **~57 million queries/day**

rDNS has largest percentage (less privacy sensitive)

Result: IP Addresses in QNAMEs



Queries with IPs in QNAMEs

- In sample:
  - IPv4: 42,846 queries (3.9%)
  - IPv6: 863 queries (0.08%)
- Estimate for total traffic
  - **~57 million queries/day**

rDNS has largest percentage (less privacy sensitive)

Smaller fraction of DNSBL and CPE queries

Categorizing Sensitive Domains

- Used 5 out of 17 top-level categories from Alexa top domains

Category	Subcategories	Domains
Religion	62	2158
Ethnicity	30	859
Lifestyle	7	265
Health	37	1621
Gender	36	1126

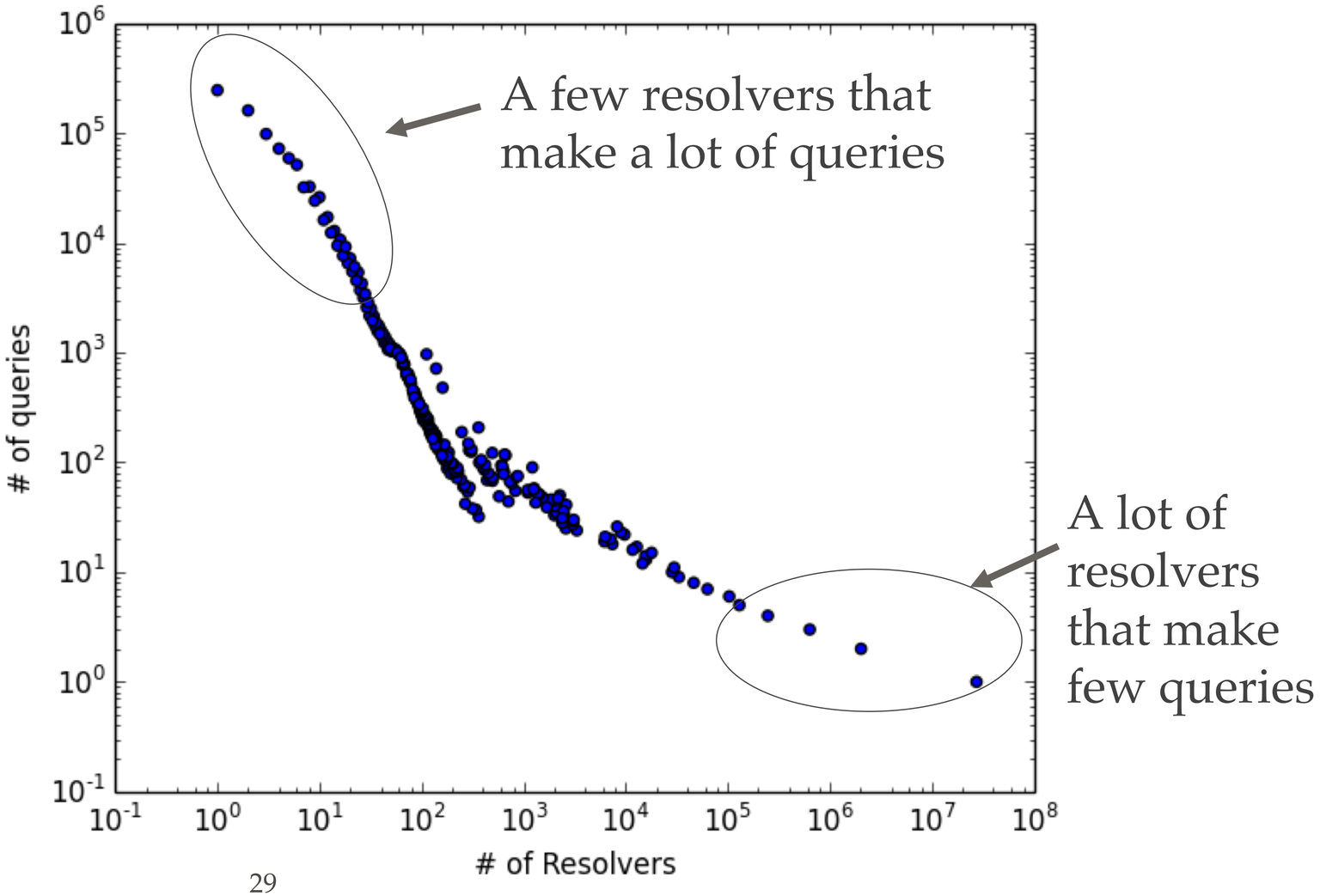
Result: Sensitive Domains

- Queries with sensitive domains
  - In sample: 12,752 queries (1.2%)
  - Estimate for total traffic: **~17 million queries/day**
- Small percentage but probably significant because of DNS traffic's long tail

Category	Count	% (out of 1.2%)
Religion	2437	19.1
Ethnicity	2030	15.9
Lifestyle	141	1.1
Health	1585	12.4
Gender	6559	51.4

Measuring Aggregation

- How many users share a resolver?
- Challenging because other factors affect number of queries
  - Multi-level caching
  - Diverse DNS clients
  - NAT addressing

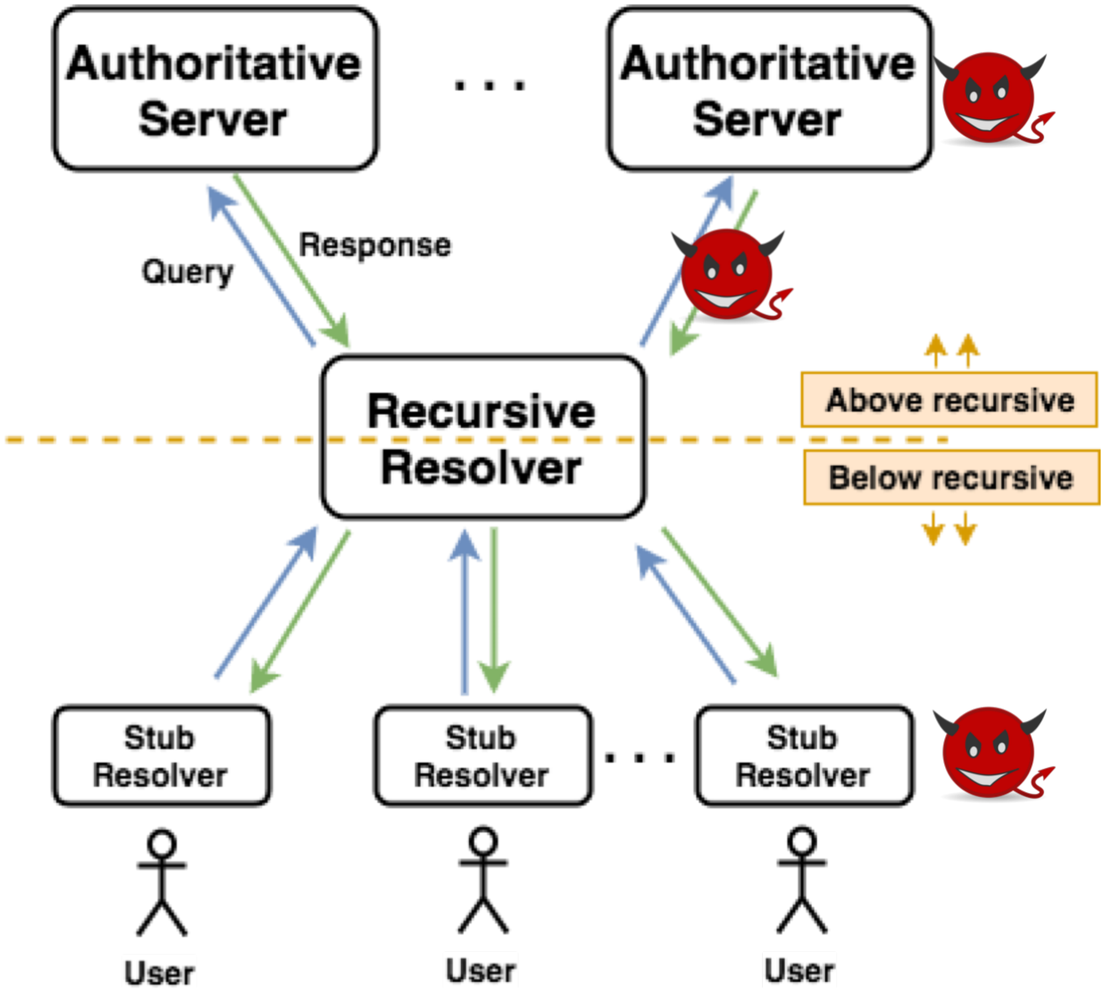


Future Work

- How much aggregation is there in the wild?
- Leaks at an organization level?

Conclusion

- Enumerate privacy leaks in DNS data above the recursive
- Root DNS data contains a notable fraction of queries that may leak information



- **Basileal Imana, imana@usc.edu**