Towards a Model of DNS Client Behavior

Kyle Schomp
CWRU

Michael Rabinovich
CWRU

Mark Allman
ICSI

Motivation

- Previous studies considered aggregate DNS behavior:
  - At root or TLD
  - Per-organization
  - Per-home
- What about per-device behavior?
  - Helps reasoning about a critical Internet component
  - E.g., may help with anomalous behavior detection
  - Helps with resolver dimensioning
  - Needed for other studies
- Ultimate goal: a model for DNS client behavior

Data Sources

- Packet trace of DNS traffic between resolvers and clients
  - Dorms and offices
  - No NATs per policy
- DHCP logs
  - Per MAC address behavior
- Resolver query logs
  - Sanity check

Types of Client Devices

Gaming consoles
Smart televisions
Laundry machines
Photocopiers
General purpose user devices (82% of all clients)

Identified by markers for browsing, searching, email, and Case's single sign-on portal

Datasets

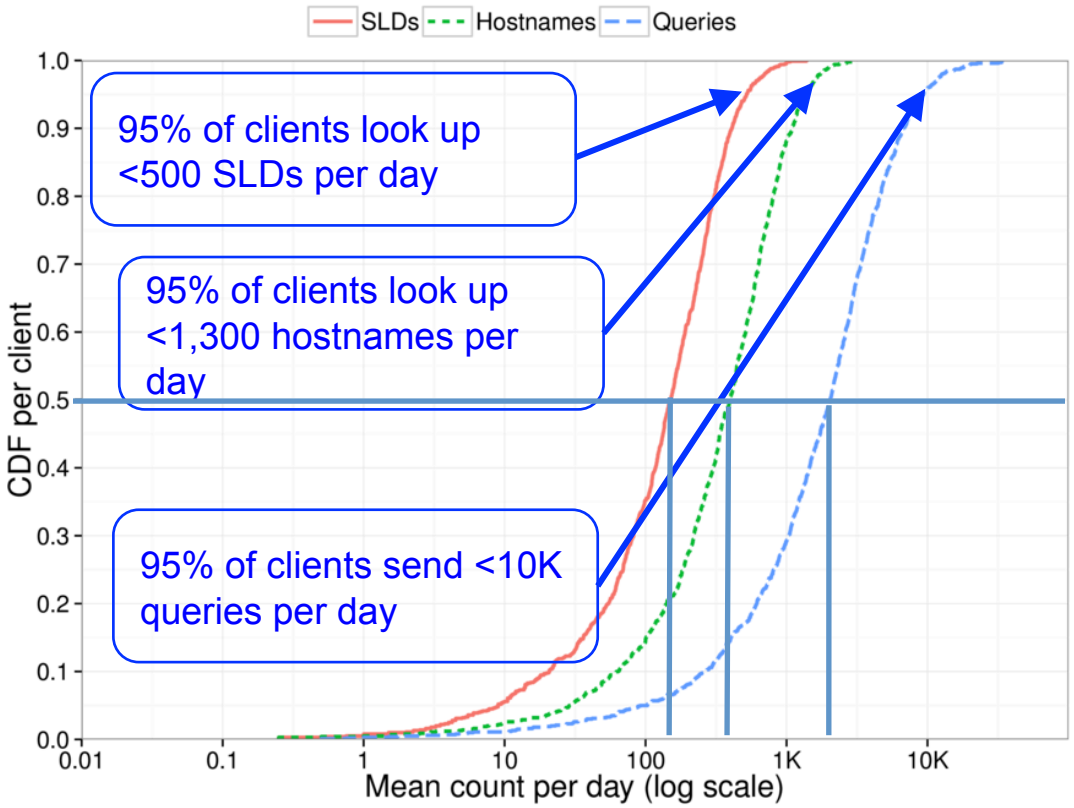
- A work-week of data (post-filtering)

Population	# MAC addresses	# queries	# hostnames
Dorms	1033	15.3M	499K
Office	5986	118M	1.52M

Behavior Characterization

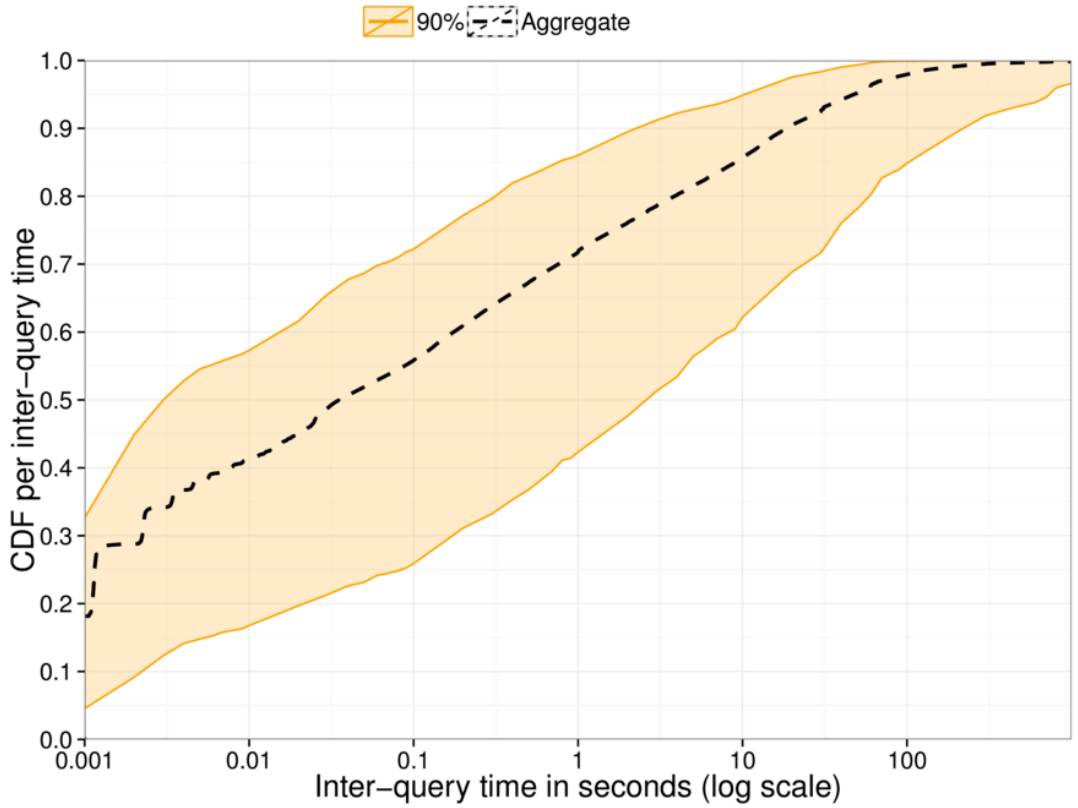
- Client activity level
  - How many queries
  - How many hostnames
- Query timing
  - Inter-arrival times from the same client
- Query targets
  - Name popularity and temporal locality
  - Name dependencies
- Client Similarity
  - Day-to-day similarity of the same client
  - Daily similarity of different clients

Average client activity per day



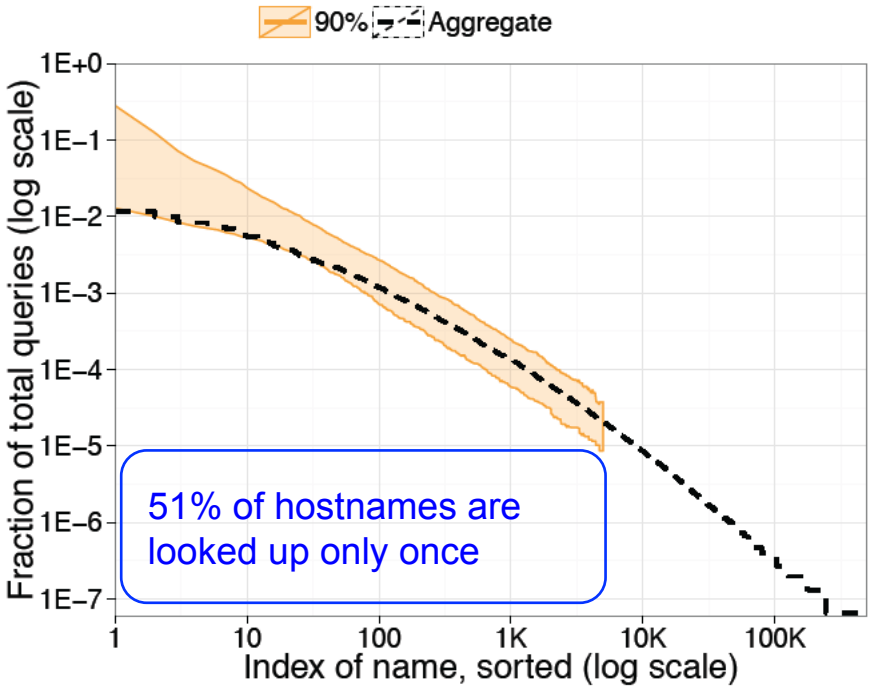
- Median client:
  - 149 SLDs
  - 393 hostnames
  - 2K queries

Query Inter-arrival Time

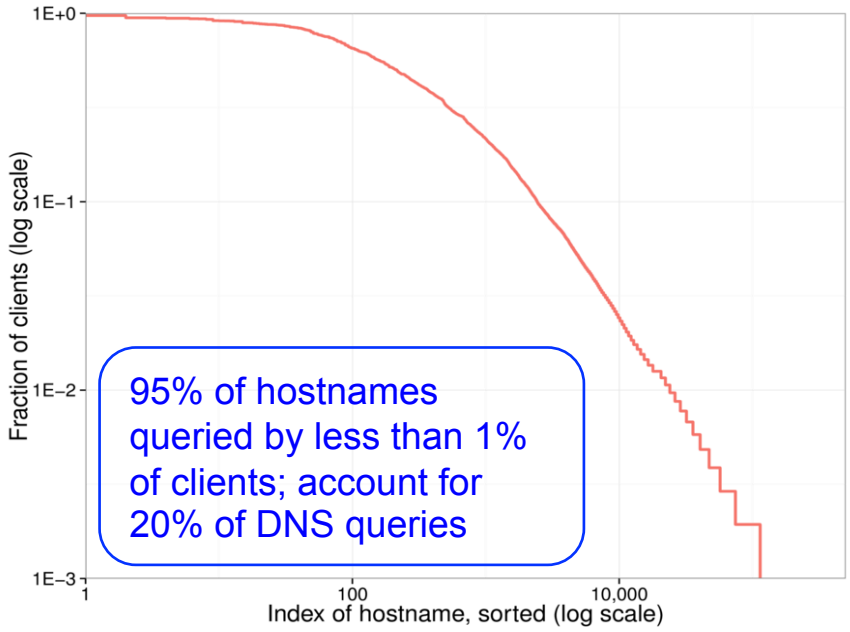


- Modeled well analytically
  - Weibull for body (up to 22s)
  - Pareto for tail (over 22s)
  - Common switch over point
  - Different parameters

Name Popularity



Queries per hostname

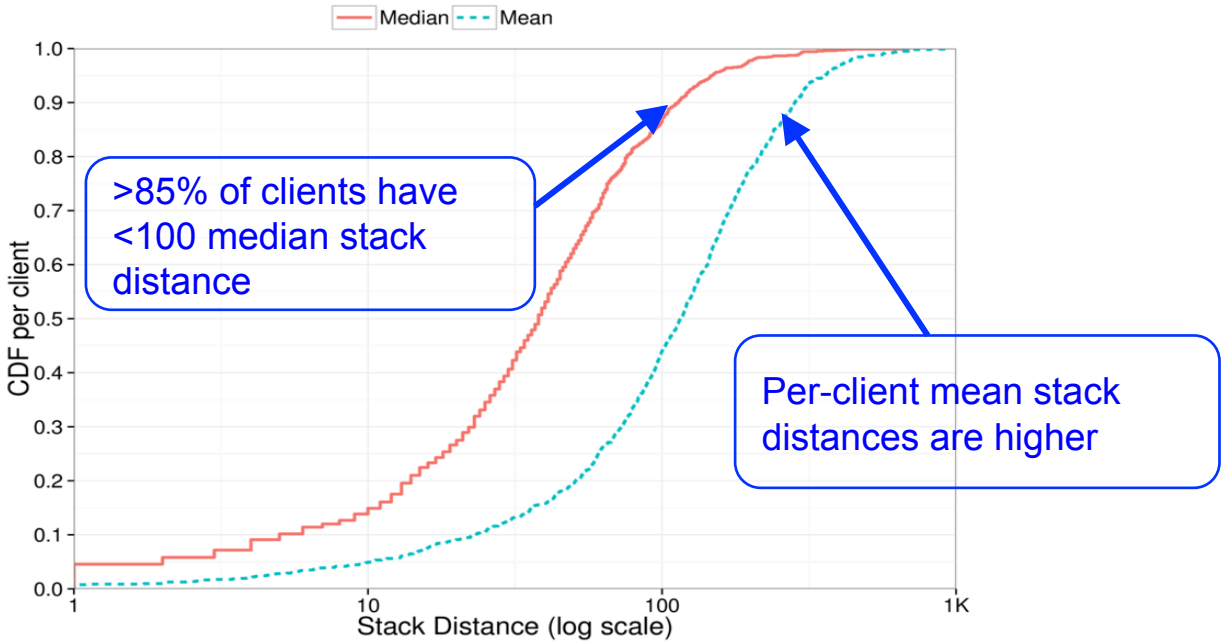


Clients per hostname

- The two popularity metrics are weakly correlated (ρ=0.51)
- Unpopular names account for significant part of DNS activity

Stack Distance

Temporal locality: How quickly a client reissues a query?



Client similarity

Daily query vectors:

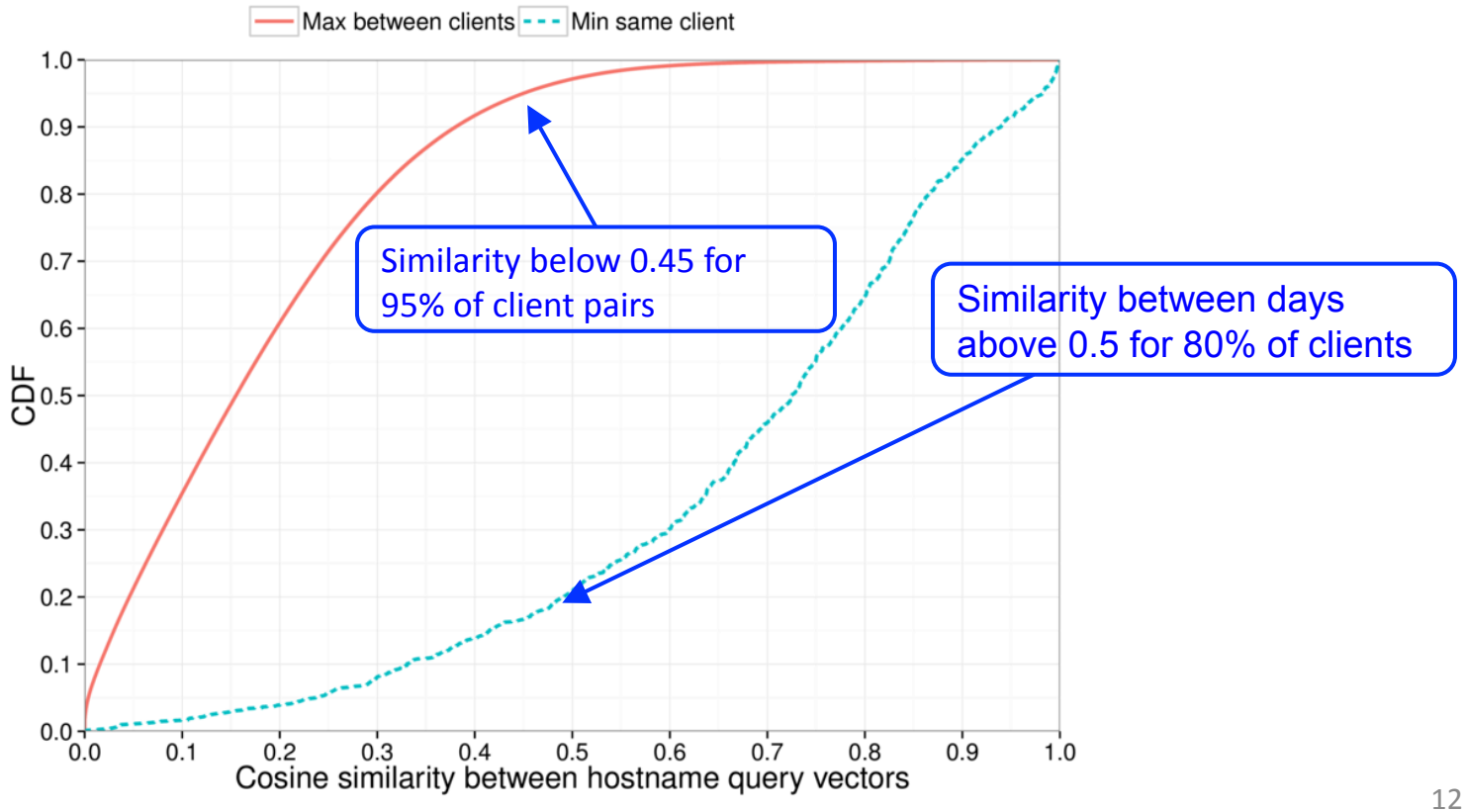
Client A on day D queries:
foo.com 1 time
bar.com 2 times
foo.bar.com 0 times
xyz.com 2 times

$V_{A,D} = <⅕, ⅖, 0, ⅖>$

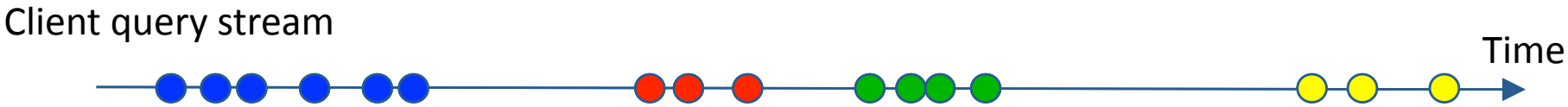
Client B on day D queries:
foo.com 2 time
bar.com 0 times
foo.bar.com 1 times
xyz.com 1 times

$V_{B,D} = <½, 0, ¼, ¼>$

Client similarity

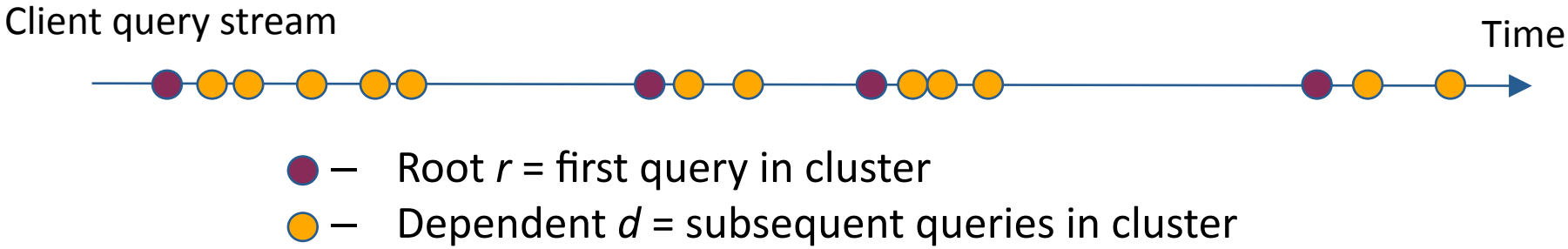


Queries occur in clusters



- DBSCAN clustering algorithm
  - 80% of all queries are in clusters
  - Median cluster size 5, mean 12
- Clusters are short
  - 99% of clusters are less 20 seconds
  - 72% of queries in clusters less than 20 seconds

Co-occurrence of queries



- (# of clusters with *r* and *d*) / (# of clusters with *r*)
  - High co-occurrence indicates a relationship
- Find many frequently occurring pairs of hostnames
  - e.g., *www.gmail.com* and *oauth.googleusercontent.com*
  - *www.reddit.com* and *www.google-analytics.com*
  - *www.buzzfeed.com* and *www.google-analytics.com*
  - Estimate that at least 21% of queries are co-occurrence

Summary

- Initial step towards a model of per-client DNS behavior
- Query arrival process is well modeled by combination of Weibull and Pareto distributions
- Clients exhibit working set of hostnames
  - Stable for client across time
  - Distinct across clients
- Most of DNS activity is due to unpopular names
- Clients emit queries in short bursts