Behavior of DNS’ Top Talkers, a .com/.net View

Eric Osterweil - Verisign Labs
Danny McPherson - Verisign Labs
Steve DiBenedetto - Colorado State University (CSU)
Christos Papadopoulos - CSU
Dan Massey - CSU
DNS, the Internet’s Favorite Kitchen Sink…

- Some have long wanted to use DNS as a general distributed database
  - TLS certs, phone numbers, etc.
    (researchers, IETF, ops)

- But, DNS is a piece of Internet infrastructure that we would have trouble doing without
  - How much do we do today that doesn’t use DNS?

- Yet, what do we really known about its global query patterns?
Let’s Look Before We Leap

• Posit: before overloading DNS, we must understand *how it is being used*
  • How are resolvers behaving?
  • How dynamic are resolvers?
  • What are the most common query-types (qtypes)?
  • Who are the busiest resolvers?

• But when measuring, how do we separate wheat from chaff?
  • There’s a lot of strangeness in DNS measurements
Understanding How it’s Being Used

• We present a month-long study of traffic seen at one instance (of 13) of .com/.net
  • These are two of the largest TLDs in the world
• To our knowledge, this is the largest study of resolver traffic and query patterns to date
  • ~975 million queries/day and ~950 thousand unique sources/day

What we found:

• A lot of strangeness
  • “Nothing ever dies in the Internet”
  • Sources just keep growing
  • Wide range of query patterns
• But, we find some consistency in the top-talkers
Outline

• Background
• Our dataset (i.e. who we are seeing)
• Who top-talkers are and what they look like
• Conclusions and future work
Background

• DNS is best thought of as a two party system
  • Resolvers (the clients) and name servers (the data owners) have different challenges, behaviors, etc.

• Domain names are grouped into hierarchical zones
  • example.com is a child of .com, and .com is a child of the root

• Each zone is served by one or more name servers
  • Any name server can serve domain names for its zone(s)
  • .com has 13 name server instances

• But resolvers cache data, so who sees the most traffic?
  • Relatively few queries go to the root, *many* more go to TLDs

• Observations from large TLDs (*cough* .com/.net *cough*) see many more queries than even the root
Example

The first cold-cache query goes to the root

1: www.example.com?

2: www.example.com?

3: .com

4: www.example.com?

5: exmple.com

6: www.example.com

7: <ip>

8: <ip>
Example (2)

The next query goes straight to .com
Example (3)

Then the next query for another .com SLD also goes straight to .com
Our Dataset

• The .com/.net registry contains ~200 million delegations
• It is served by 13 name server instances
  • Each NS has a different domain name and a different IP address
  • Some of these are unicast, some are anycast
• These instances are served from over 70 sites, worldwide

• We limited our analysis to just one instance of .com/.net
  • g.gtld-servers.net (the G instance)
  • Unicast from a single site in California, USA
• This instance serves both .com and .net referrals
  • Verisign serves more TLDs, but this analysis was restricted
Who the G Instance Sees

- We saw clients from all over
- In just 10 minutes, large traffic volumes from all over the IP space
- We saw lots of strangeness
  - 7th most popular query type is for A6 records (a deprecated type)!
  - Number of unique source addrs didn’t stop growing during measurements
  - We saw *pinning and polling* behavior
    - Resolvers probing different name servers and swinging traffic back and forth! (see the paper) ;)
- Strangeness makes it tough to discern signal from noise
  - How does one determine when outliers are skewing...
Top-Talkers

• Starting point: what are some of the key behaviors of the most active query sources?

• # of sources grows without ebbing, but those responsible for 90% stabilizes

• We, therefore, classify most of our query traffic as coming from these top-talkers
  • In 1 day, out of 958,558 sources, we saw 39,936 top-talkers
  • Roughly 4 million unique sources in 30 days
How Much Traffic Makes a Source a Top Talker

- From the log-scale plot, we see the diminishing returns after including source beyond 90%

- Though not hard and fast, is it a good starting point?
What Resolvers are Querying For

- Resolvers seem to mostly just query for A/AAAA/MX RRs
  - There are other popular types, but not in comparison
- By looking at this sort of distribution, we cannot tell IPv6 adoption, but likely can gauge things like “happy eyeballs”
  - If concurrent IPv4/IPv6 queries are issued, our data should see both
Top Talkers: a Low Pass Filter

• This simple metric helps to separate some strange behaviors from what we might expect

• Whereas many sources asked for heterogeneous combinations of qtypes, top-talkers mostly asked for As, AAAAs, and MXs
  • Not all, but we wouldn’t expect all
Conclusions and Future Work

- The top-talkers approach helped us identify meaningful trends in our data, while systematically removing less active sources
  - We have not quantified the relative benefit of using different cut-offs 85%, 95%, etc.
- The *pinning and polling* behavior has given us a chance to begin classifying variations of resolvers
- We have begun to use the top-talkers filter to classify typical traffic vs. large-scale attack traffic
  - We have begun building automated attack defenses
  - He are investigating correlations between abnormal top-talkers and external events (MX spam campaigns)
  - Etc.
Thank You
Backup
Pinning and Polling

Pinned Resolver

Polling Resolver

Graphs showing the number of queries sent as a function of date/time for Pinned Resolver and Polling Resolver.
Full qtype distribution

Histogram of Counts of Query Types

# Queries

A AAAAA MX ANY TXT NS A6 CHAOS CNAME SRV DS SPF SOA PTR NAPTR NSEC DNSKEY HINFO

1e+07
1e+06
1e+05
1e+04
1e+03
1e+02
1e+01
1e+00
10
100
1000
10000
100000
1e+06
1e+07
Rolling top talkers
Qname distribution