TLD Registry Data

an unstructured wander through the zoo

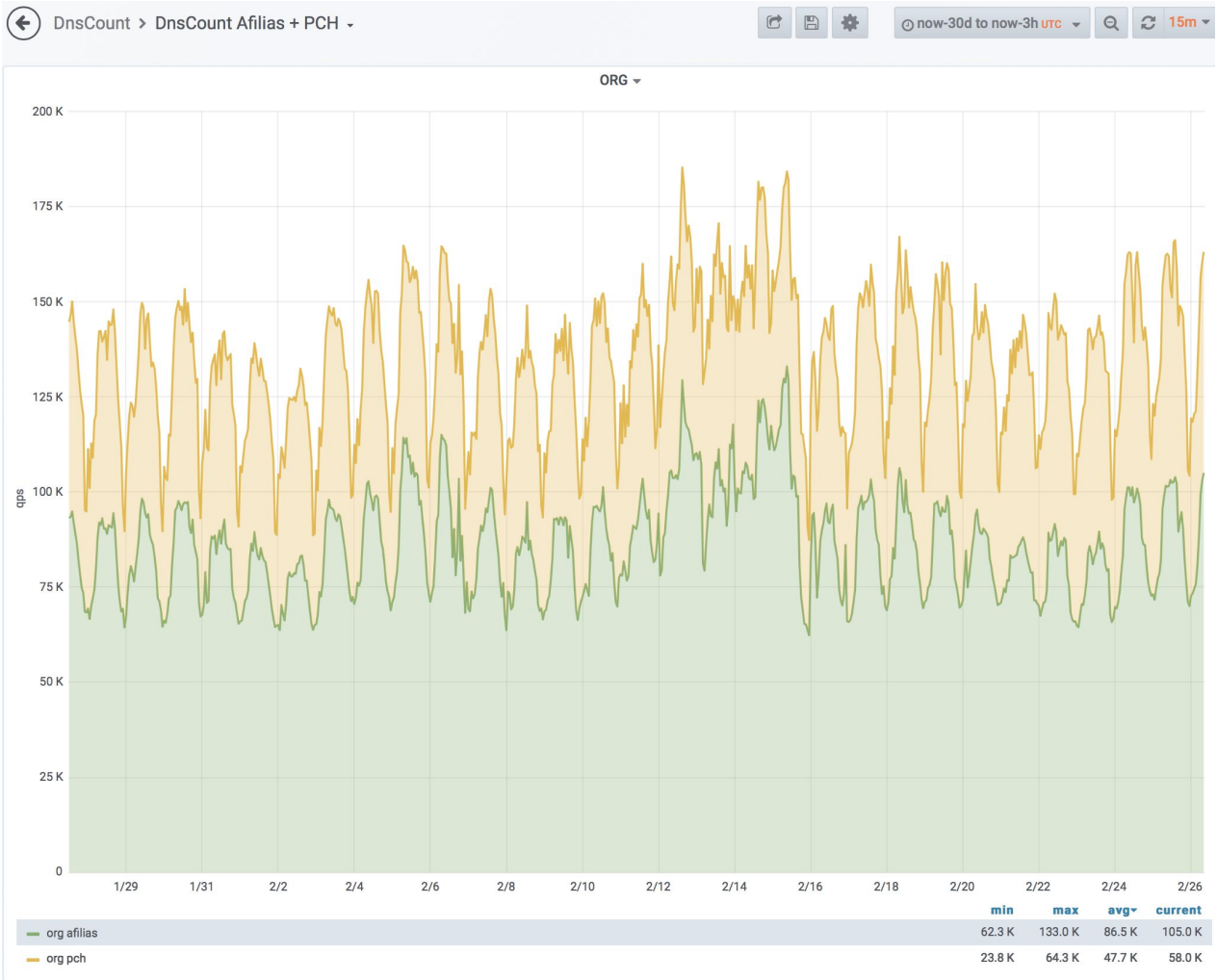
Joe Ably
Public Interest Registry

DINR2020

What is a TLD Registry

DNS Answer

- We publish the ORG zone in the DNS
  - Highly delegation-centric zone
  - ~10M delegations
  - DNSSEC-signed (RSASHA1, NSEC3, opt-out)
  - Without signing operations, changes frequently (non-zero deltas every minute)
- We operate the authoritative servers for ORG
  - Six dual-stack servers
  - Those servers receive and respond to DNS queries (mainly, overwhelmingly) from recursive nameservers


DnsCount > DnsCount Afilias + PCH
now-30d to now-3h UTC
15m
ORG
200 K
175 K
150 K
125 K
100 K
75 K
50 K
25 K
0
qps
1/29
1/31
2/2
2/4
2/6
2/8
2/10
2/12
2/14
2/16
2/18
2/20
2/22
2/24
2/26
min
max
avg
current
org afilias
62.3 K
133.0 K
86.5 K
105.0 K
org pch
23.8 K
64.3 K
47.7 K
58.0 K

Database Answer

- We ensure uniqueness of names under .ORG by running a centralised database
  - Source of truth for what domains exist and don’t exist
  - Thick registry; full registrant information is included
- We provide an authenticated interface that allows qualified clients to interrogate the contents of and make changes to the registry
  - Clients systems are operated by registrars
  - Extensible Provisioning Protocol (RFC 5730)
  - Various basic primitives (check, info, poll, transfer, create, delete, renew, transfer, update)
- We provide a terrible, ancient interface that allows people to get variously-redacted information out of the registry
  - Whois (RFC 3912)
  - RDAP (RFC 7482) seeks to provide authorisation and privacy-sensitive redaction

Abuse Answer

- We provide a clearing house for reports of abuse relating to registered .ORG names from trusted notifiers
  - Response is almost always to escalate to the sponsoring registrar
  - Other actions are possible with a court order
  - In a very small set of cases (e.g. CSAM) we may take unilateral action
  - Other registries have different policies and practices
  - This whole area is sensitive, since through the lens of free speech it can look like censorship
- This is not my area
  - Don’t necessarily expect good answers from me on this
  - Others here know far more
  - I care though, obviously. I’m not a monster.

Lifecycle of a Domain

Ante-Natal

- Domains that are not (and have never been) registered can still be used
  - Queries for non-existent names still arrive at the ORG authoritative servers
  - Names that will be provisioned, but not yet (e.g. product pre-positioning, malware signalling through DGAs)
  - Names that are actively being provisioned
  - Names that are intended for internal use but which are leaking to the Internet
  - Typos, bit-flips, other?
- What do we see on the ORG servers?
  - If we see a query, chances are good that some end system triggered a query
  - We assume some names are actively suppressed in resolvers
  - Aggressive NSEC caching and negative response TTLs can mask query frequency
  - Retry frequency might tell us something

Birth

- Newly-registered domains appear in the registry
  - Exposed through whois, RDAP (to the degree that anything is exposed through whois, RDAP)
  - Also in zone file repositories, e.g. CZDS
  - Also in blacklists of recently-registered domains
  - Birth potentially reflected in different query patterns (certainly in response patterns)
- We see patterns in domain registration and renewal
  - Speculative registration of portfolios of names, refinement, branding
  - Bundles of domains registered using the same DGA
  - No doubt many more

Adulthood

- Domains are used in different ways
  - Investments, sometimes parked to support pricing
  - Brand protection, often not well-delegated
  - Domains in more deliberate use, perhaps reflected in query patterns (e.g. domains that support mail are sticky)
- Registry data changes
  - Nameservers, transfers, registrant data
- Datasets exist that attempt to categorise domain names along particular slices through the Internet
  - Web crawls, e.g. DataProvider, DomainsBot/Pandalytics, CENTR
  - Mail domains
- Presence or absence of a delegation does not mean domains exist
  - Lame delegations, much of the Internet is broken, film at 11

Death

- Domain expiry is somewhat registry-specific
  - From a registry perspective, always renew until delete? Expire unless renewed? Other?
  - Elaborate set of policy-based timers determine when domains are able to be renewed for normal fee, renewed for higher fee, allowed to expire, released for re-registration
- Domains can also disappear from the DNS for other reasons
  - Various registry flags can suppress publication in the zone
- DNS queries don’t necessarily disappear just because a delegation disappears
  - But responses change

Re-Birth

- Domains can be resurrected after their delegations have been pulled
  - Sometimes managing to pay for something that is cheap is difficult to remember to do
- Domains can be re-registered after they expire
  - There’s an industry in registering domains within milliseconds of them becoming available after being deleted
- Queries that arrive for a domain that has been reborn might correspond to old management or new management
  - Geoff Huston has also observed zombie queries that seem to persist for unnatural lengths of time, for unique names at third and lower level labels

Datasets

A Note on the Privacy of Individuals

When it comes to data sharing, PIR is constrained and motivated by such things as:

- its privacy policy and various privacy legislation
- its various contracts with ICANN and others
- a strong sense of common decency

We will not knowingly compromise the privacy of individuals.

DNS Data

- Zone data, e.g. CZDS, DNS-OARC
  - Oddities (e.g. orphan glue)
  - zone size
  - Patterns in delegation data
  - macro change sets
  - may or may not include DNSSEC artefacts, e.g. opt-out sections
- Query data
  - DITL collections at DNS-OARC
  - Query rates
  - Complete query collections
  - Response data (e.g. name errors)
- Non-DNS traffic
  - e.g. backscatter

Registry Data

- The Registry Itself
  - Mapping domains to sponsoring registrar
  - Keyed retrieval by more than just domain or host name
  - Contains more domain names than exist as DNS delegations
- EPP Logs
  - Record of every registry transaction that represents a data change
  - Create, update, transfer, delete
  - Enables a view of the registry over a time axis
- Whois/RDAP Logs
  - Records of whois/RDAP transactions

Applications

- Business Intelligence
  - Renewal prediction
  - Domain spinning (e.g. NIC.AT)
  - Channel services
- Abuse Detection and Mitigation
  - Minority Report (e.g. EURID)
  - Patterns in registrar behaviour
  - Policy development
- Infrastructure Scaling
  - Anomaly detection
  - Forecasting, scaling, provisioning

So now what?

- We have (access to) data.
  - What data didn’t I think of?
- How should we make it available? Under what terms?
  - If you have good ideas about how to use this data, what terms can you tolerate?
  - Note again that we operate under a robust privacy regime, and we will not compromise the privacy of individuals
- How can we tell if the data is useful?
- What business case makes sense to a registry to counterbalance the costs of making data available?
  - We are not the only TLD registry in the world
    - Who else can we learn from?
    - How could we provide a good model for others to follow?