NZRS
an InternetNZ company

Domain Popularity Ranking Revisited

Sebastian Castro
Technical Research Manager, NZRS Ltd
CENTR Jamboree, Brussels 2016

Domain Popularity Ranking

- By mining DNS data, derive popularity for domain names
- Uses a technique from text mining called TF-IDF

  Term Frequency – Inverse Document Frequency
- Presented at DNS-OARC Amsterdam 2015
- Some interest from a few ccTLDs

Quick methodology recap

- Use DNS queries
- Extract query name, IP address
- Per day calculate

$$\mathrm{tf}(d, a) = \frac{\sum \mathrm{queries}(d \text{ from } a)}{\sum \mathrm{queries}(a)}$$

$$\mathrm{idf}(d, A) = \log \frac{|A|}{|a \in A| : a \text{ asked for } d}$$

$$\mathrm{tfidf}(d, a) = \mathrm{tf}(d, a) \times \mathrm{idf}(d, A)$$

Recap

$$\mathsf{rank}(d, a) = \#d \text{ sorted desc by } \mathsf{tfidf}(d, a)$$

$$\mathsf{rank}(d, A) = \#d \text{ sorted desc by}$$
$$\sum \mathsf{rank}(d, a)\mathsf{weight}(\mathsf{rank}(d, a))$$

Open questions

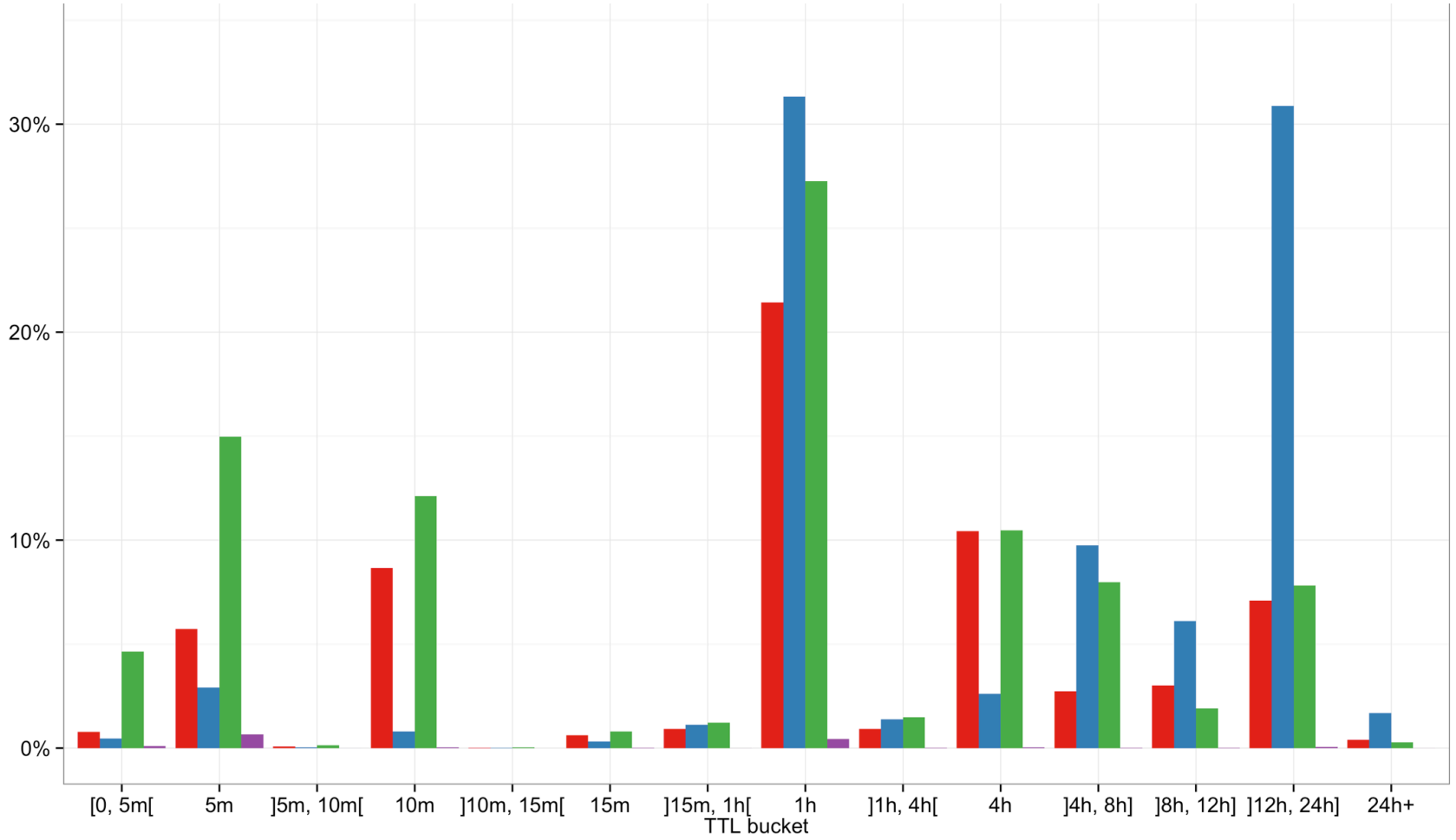
- How TTL affects the result
- How to weight traffic coming from certain addresses
- How this compares to “ground truth”
- How to handle the noisy nature of DNS data.
- Checking for bias

TTL

- Not resolved yet
- DNS traffic may follow a Poisson distribution, where lambda=f(record_TTL)
- Distribution of TTL values

  We have it, from our zone scan

TTL distribution



Weight function?

- Not all source addresses are equal
  70 different address asking for www.zz--icann-sla-monitoring.nz.
- Known resolvers
  Local ISPs, Google DNS, OpenDNS
- Source address characterization
- Population estimate based on variety of query names, query type/names, other
  http://blog.nzrs.net.nz/characterization-of-popular-resolvers-from-our-point-of-view-2/

Ground Truth

- Have access to Google Analytics for a few .nz domains

  Can compare User views against number of DNS lookups

  Can generate ranking in both cases
- High level check: compare ranking from GA against DNS traffic
- Finer grain check: Estimated number of visits versus real number of visits

Noisy DNS

- Started using DNS queries
  A lot of noise, a lot of rubbish
- Moved to use only NOERROR DNS responses
  Only consider names that are in the registry
- Some artifacts were eliminated
  DDoS reflection attack traffic

Checking for Bias

- Our data comes mainly from New Zealand
  Got a sample from an offshore provider to compare
- TF-IDF is useful for finding relevant words in a document
  Stop words don't exist in the DNS
  IDF punishes very popular sites, like google.co.nz
- Simplified calculation as fraction of population and fraction of volume

Looking for collaboration

- Review
- Different datasets
- Replicate our algorithm with your data

Contact: sebastian@nzrs.net.nz
www.nzrs.net.nz

NZRS
an InternetNZ company