Rebuilding zone files from passive DNS data

John-Paul Verkamp
Minaxi Gupta

Indiana University

Funded by NSF OCI Grant with kc Claffy of CAIDA
"SDCI Sec: Metadata Management Software Tools to Support Cybersecurity Research and Development of Sustainable Cyberinfrastructure".
Motivation

- Zone files for the largest gTLDs, especially .com are (non-commercially) available.
- Zone files for most ccTLDs aren't often available.
- Passive DNS data exists, we can use it to rebuild zone files for any TLD.
Data Sources

- DNSParses
- ISC/SIE
Data Sources - DNSParse

- **Per day:**
  - ~ 100 MB of data
  - ~ 4M DNS records
Data Sources - ISC/SIE

- Raw passive DNS payloads
- Unpacked and deduplicated DNS
- Further de-duplication
  - Removal of generically named records and high volume wildcards

- Per hour:
  - ~ 1 GB of data
  - ~ 23 million entries
Algorithm

- Parse records
- Merge timestamps
- Apply smoothing
- Local verification
- Reconstruct zone files
Algorithm - Parse Records

- **DNSParse**
  - Gzipped comma separated values
  - Contains: query, answer, rrtype, ttl, firstseen, lastseen, sensorID

- **ISC/SIE**
  - Binary format: libnmsg
  - Contains: section, qname, qtype, qclass, rrname, rrtype, rrclass, rrttl, rdata

- **We want:**
  - rrtype, query, response, first, last, ttl
Algorithm - Merge Timestamps

- Timestamps are stored as a binary tree with each leaf being a pair / time range
- Each new record has the time it was first seen, last seen, and a time to live
  - add (first, last + ttl) to the tree
- Merge overlapping records to save memory and insertion time, rebalance on update
Algorithm - Smoothing

- After each collection of records (hourly), timestamps are smoothed.
- Assumption is that domains that have been long lived but have short lapses before returning to the previous value remained active.
- Short lived domains and those that disappear for long periods of time are not smoothed.
- Parameters for "short lived" and "long periods of time" are still being tweaked.
Algorithm - Local Verification

- After local data has been added
- Attempted on domains that have been inactive for a long enough period of time
- If local verification matches the original record, the timestamps will be smoothed
- If it doesn't return or no longer exists, no further local verification will take place
Algorithm - Reconstruction

- Zone files can be reconstructed for any zone file for any day
- Scan for that day's time stamp for valid domains
Results - DNSParse

- Using 9 months of DNSParse data:
  - 6% of .com, 5% of .net
  - data is too sparse to accurately smooth, resulting in slowing growth
  - many ccTLDs are nearly empty
Results - Growth of

- Older results after 5 months:
  (number of unique domains in zone file)
Results - ISC/SIE

- Using 1 month of ISC/SIE data:
  - 52% of .com, 43% of .net
  - smoothing is accurate enough that the data is still growing, albeit slower
  - at current growth, estimates are ~70% of .com
  - even the smaller ccTLDs have tens of thousands of domains, still some variability within
Questions?