Evaluation of domains using third party data

Siôn Lloyd

ICANN OCTO-SSR

TLDCON

September 2024

Introduction

External data

Many forms of third party data available:

- Reputation Block Lists
- APIs
- Passive DNS
- etc.

We use many different sources in many different projects

Different uses of data

- Broader statistics vs focused view
- Historic vs current
- Consistency vs best available

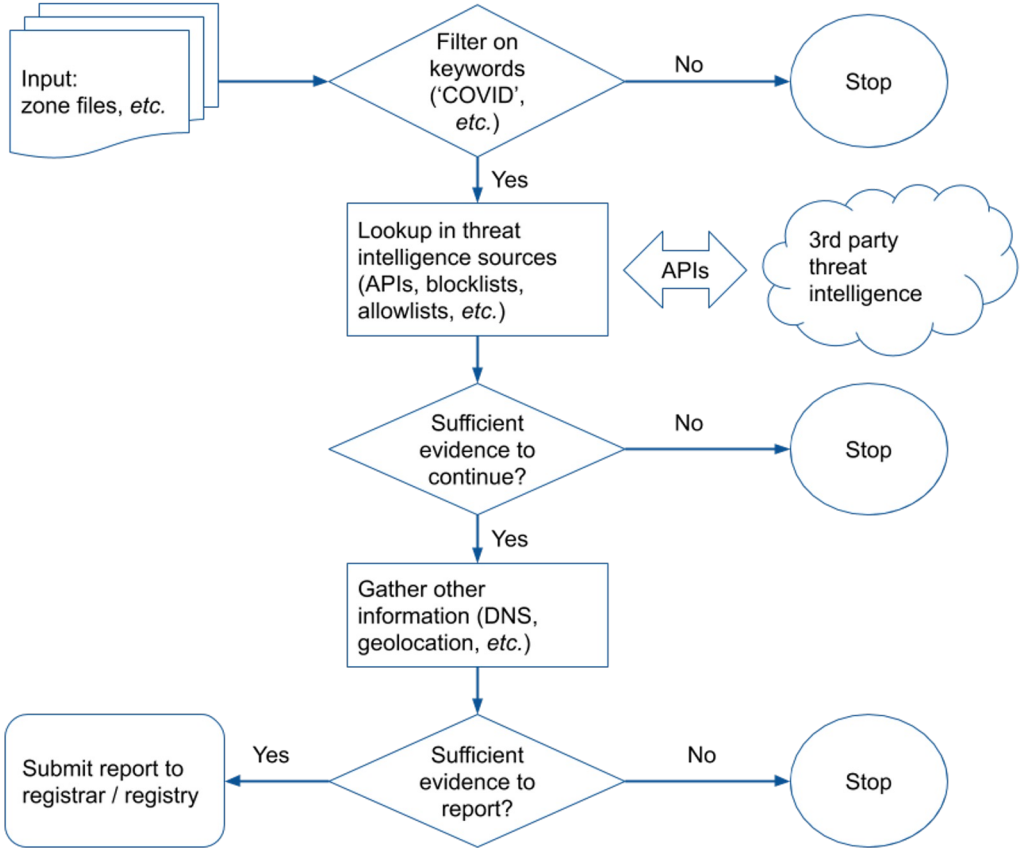
Different requirements

Different issues

Different modes of “failure”

Different costs of “failure”

DNSTICR - Data to Intelligence



Data “issues”

Many different providers, with different collection methods, different focus

We don’t collect this data ourselves

We don’t control the collection methodology

Our use cases may not be those imagined by the producers

Mitigation of mismatches is down to the consumer.

Understanding our reputation data

Understanding our reputation data – general

- Entry Types

  Spam, Phish, *etc.*

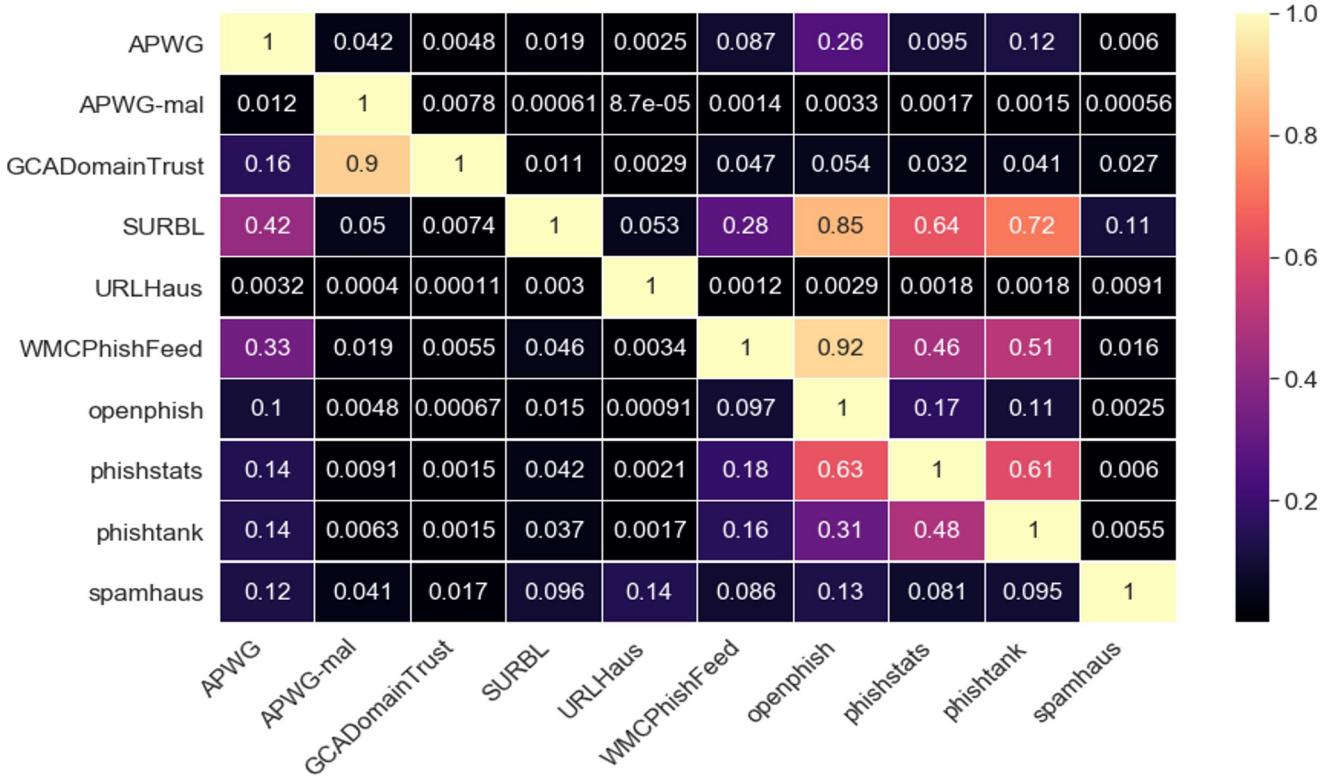
- Metadata

  Malware family, phished brand, *etc.*

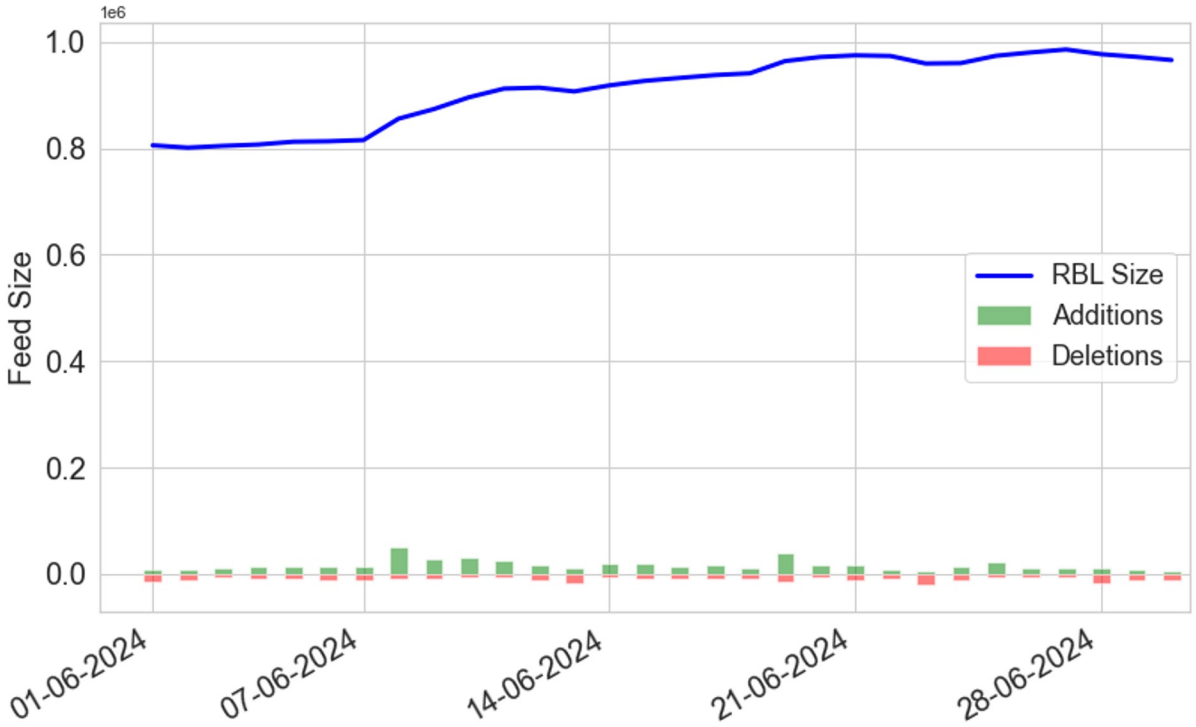
- Entry provenance

  Observations or predictive

Understanding our reputation data – Overlap



Understanding our reputation data – Churn



Understanding our reputation data – other checks

- Do entries resolve
- Detect parked/suspended pages
- Does categorization match reality

Cross-reference with high reputation sources (*e.g.* TRANCO)

Or other RBLs

General Observations

- We often rely on data collected by third parties
  - Think about your specific needs
  - The costs of various “failures”, *e.g.*
    - false positives (false negatives?)
    - …
  - Can the data be improved?
- Understand how datasets complement each other
  - Be prepared to read multiple RBLs
- Don’t stop testing
  - Things change

One World, One Internet

Visit us at **icann.org**

- @icann
- facebook.com/icannorg
- youtube.com/icannnews
- flickr.com/icann
- linkedin.com/company/icann
- soundcloud.com/icann
- instagram.com/icannorg