In the search of resolvers

Jing Qiao 乔婧, Sebastian Castro – NZRS

DNS-OARC 25, Dallas

Background

- Domain Popularity Ranking

  Derive Domain Popularity by mining DNS data

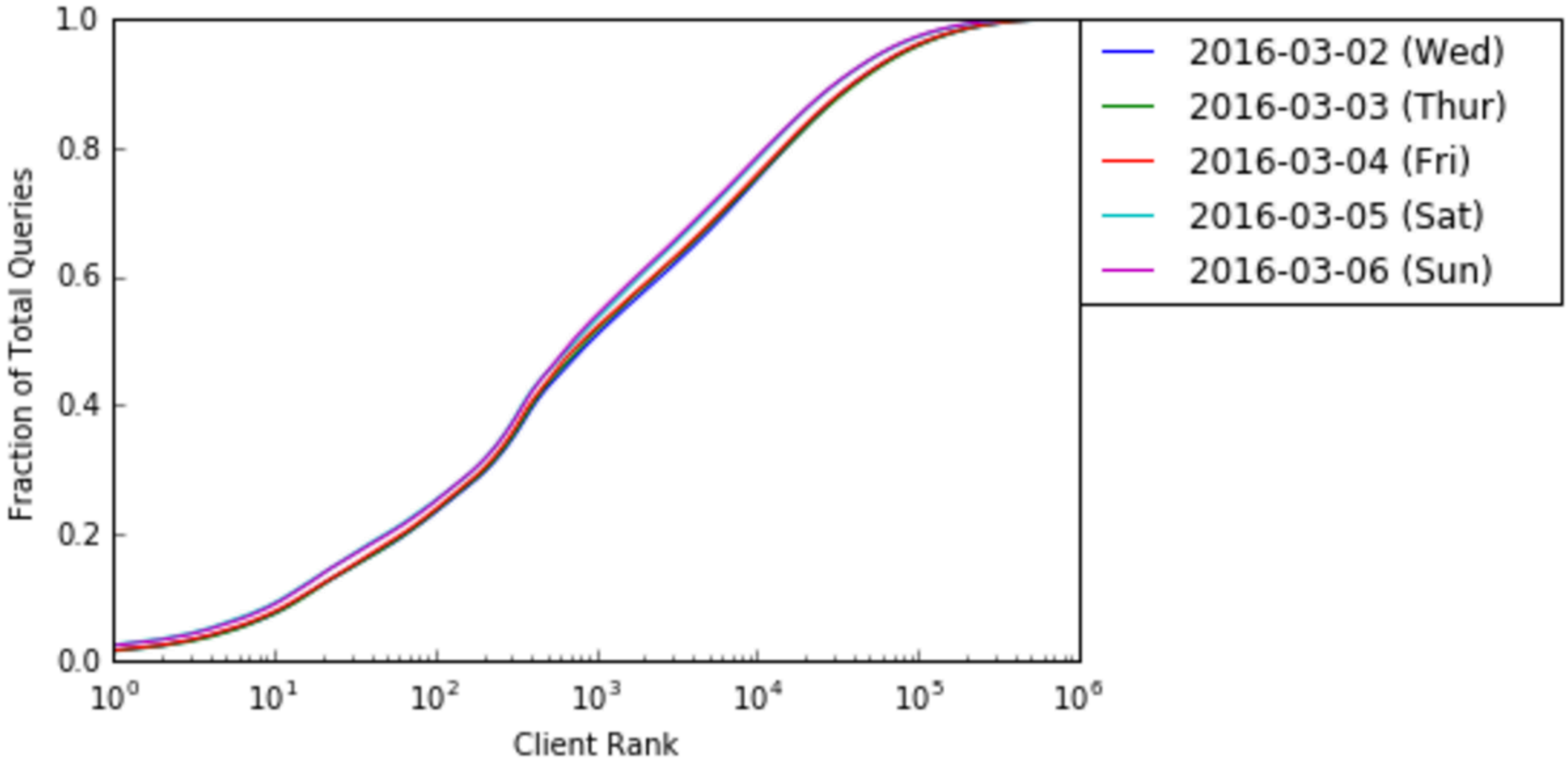
  Noisy nature of DNS data

  Certain source addresses represent resolvers, the rest a variety of behavior

  Can we pinpoint the resolvers?

Noisy DNS

- Long tail of addresses sending a few queries on a given day



Data Collection

- To identify resolvers, we need some data
- Base curated data

  836 known resolvers addresses
  - Local ISPs, Google DNS, OpenDNS

  276 known non-resolvers addresses
  - Monitoring addresses from ICANN
    - Asking for www.zz--icann-sla-monitoring.nz
  - Addresses sending only NS queries

Exploratory Analysis

- Do all resolvers behave in a similar way
  http://blog.nzrs.net.nz/characterization-of-popular-resolvers-from-our-point-of-view-2/

- Conclusions
  There are some patterns
  - Primary/secondary address
  - Validating resolvers
  - Resolvers in front of mail servers

Supervised classifier

- Can we predict if a source address is a resolver?
- 14 features per day per address
  - Fraction of A, AAAA, MX, TXT, SPF, DS, DNSKEY, NS, SRV, SOA
  - Fraction of NoError and NxDomain responses
  - Fraction of CD and RD queries
- Training data
  - Extract 1 day of DNS traffic (653,232 unique source addresses)
  - Base data

Training Model

LinearSVC

```
Training:
LinearSVC(C=1.0, class_weight='balanced', dual=True, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
     verbose=0)
train time: 0.003s
Cross-validating:
Accuracy: 1.00 (+/- 0.00)
CV time: 0.056s
test time:  0.000s
accuracy:   1.000
dimensionality: 14
density: 1.000000
classification report:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00        73
          1       1.00      1.00      1.00       206

avg / total       1.00      1.00      1.00       279
```

Other Learning Algorithms

- For the classification problem with less than 100K samples, LinearSVC is the first choice
- We also benchmarked some other algorithms such as K-Neighbors and Random Forest.

  All of them achieved 100% accuracy

Test the model

- Apply the model to three different days

  Resolver is represented as 1, and non-resolver as 0.

```
df = predict_result(model, "20160301")
df.isresolver_predict.value_counts()
1    645060
0      8172

df = predict_result(model,"20160429")
df.isresolver_predict.value_counts()
1    529757
0      6243

df = predict_result(model,"20151212")
df.isresolver_predict.value_counts()
1    453640
0      9279
```

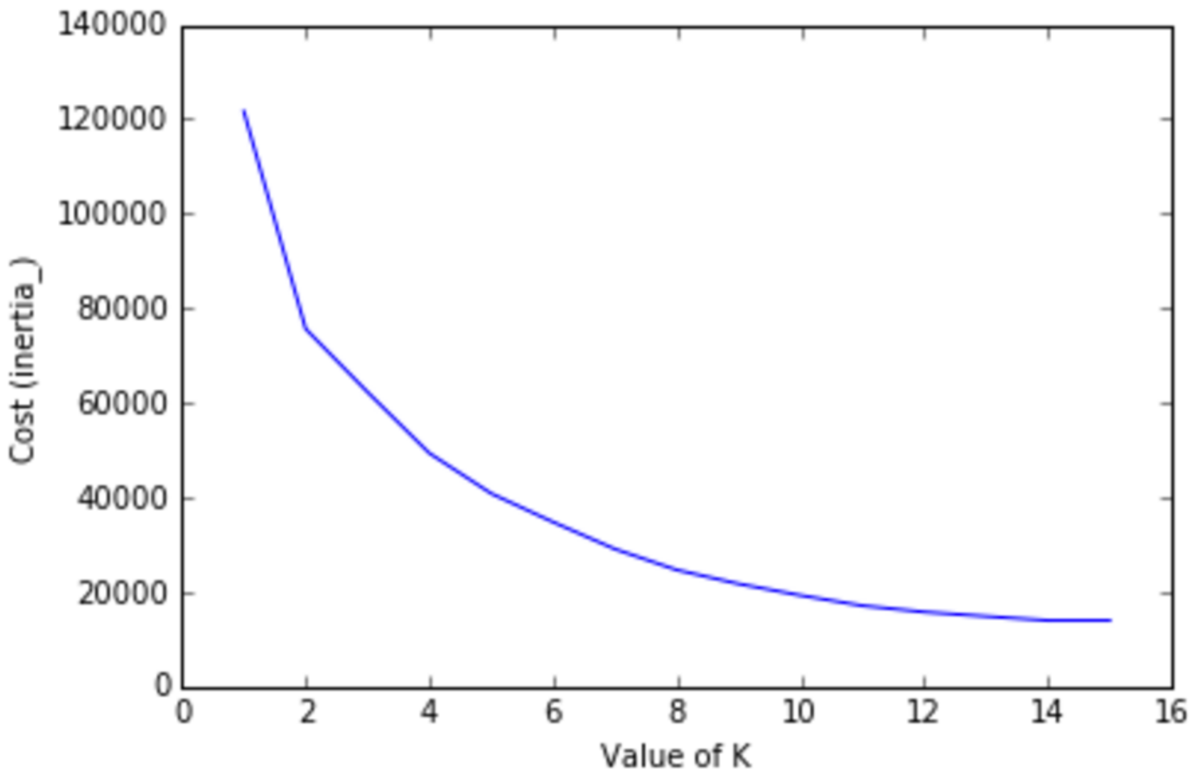
Preliminary Analysis

- Most of the addresses classified as resolvers
  - Possibly because the list of non-resolvers show a very specific behaviour
  - Model fitting specific behaviour, leaving any other address as resolver.
- The next iteration of this work should to improve the training data to include different patterns.

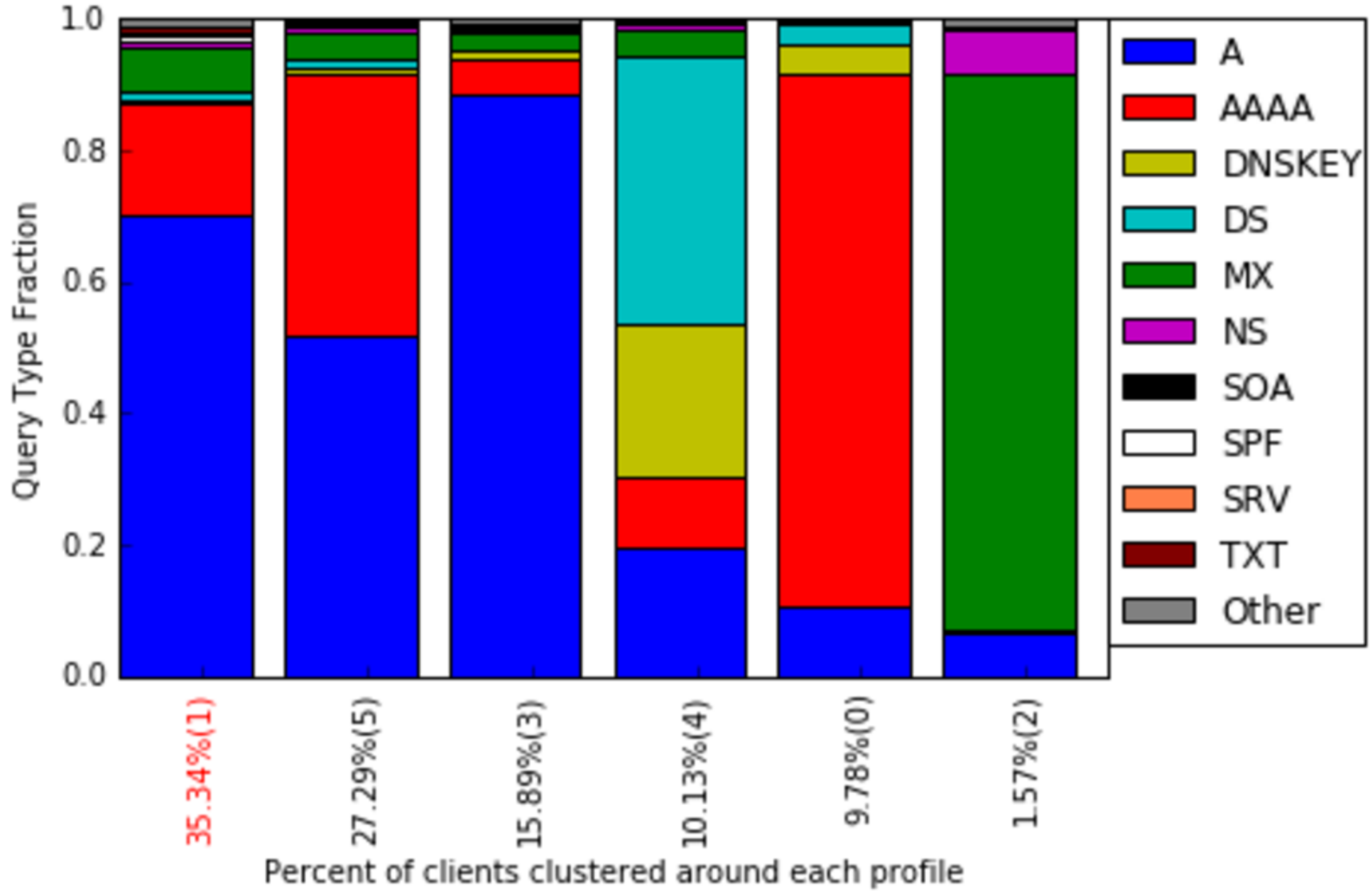
Unsupervised classifier

- What if we let a classifier to learn the structure instead of imposing
- The same 14 features, 1 day's DNS traffic
- Ignore clients that send less than 10 queries
  Reduce the noise
- Run K-Means Algorithm with K=6
  Inspired by Verisign work from 2013
- Calculate the percentage of clients distributed across clusters

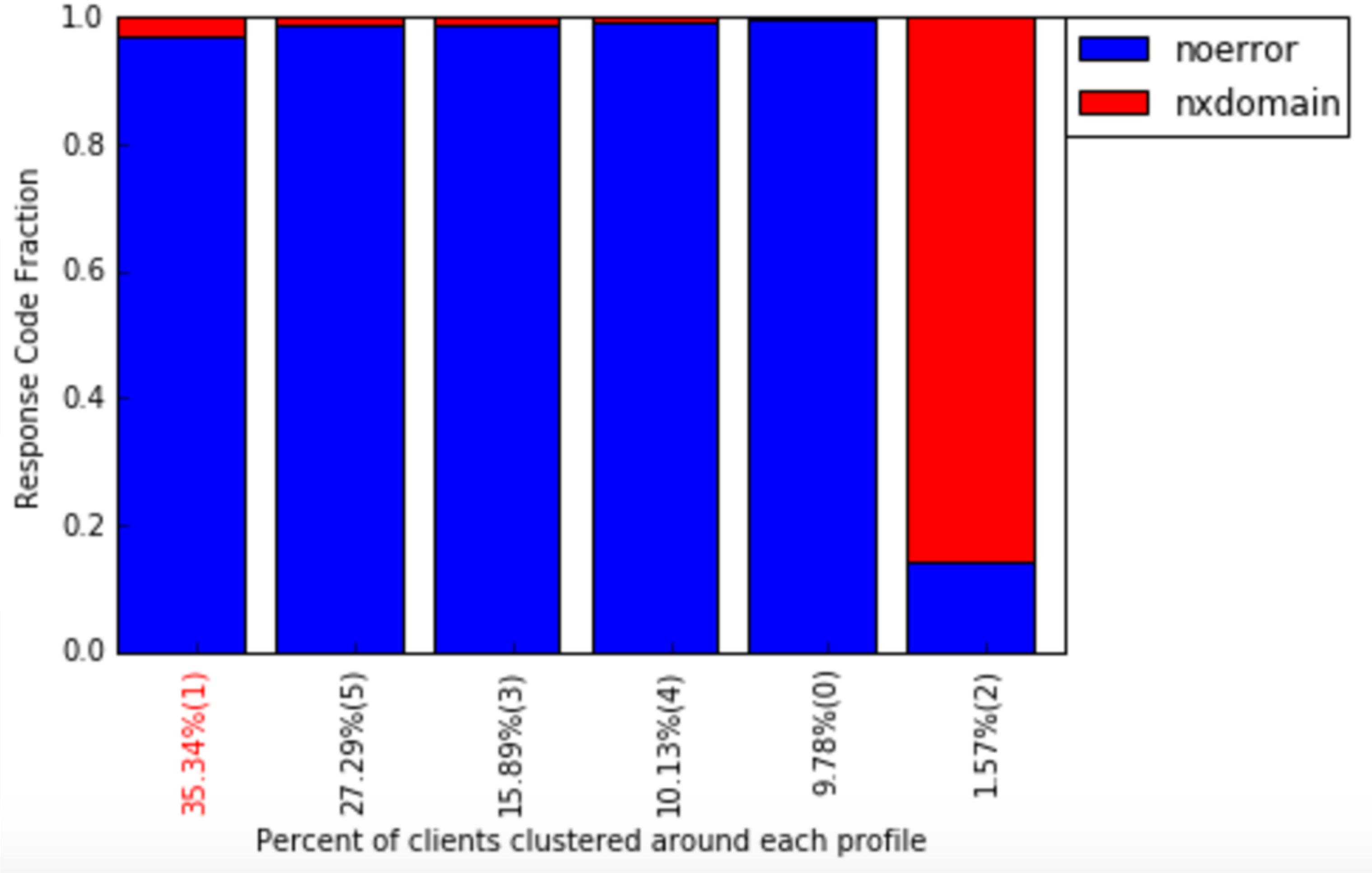
K-Means Cost Curve



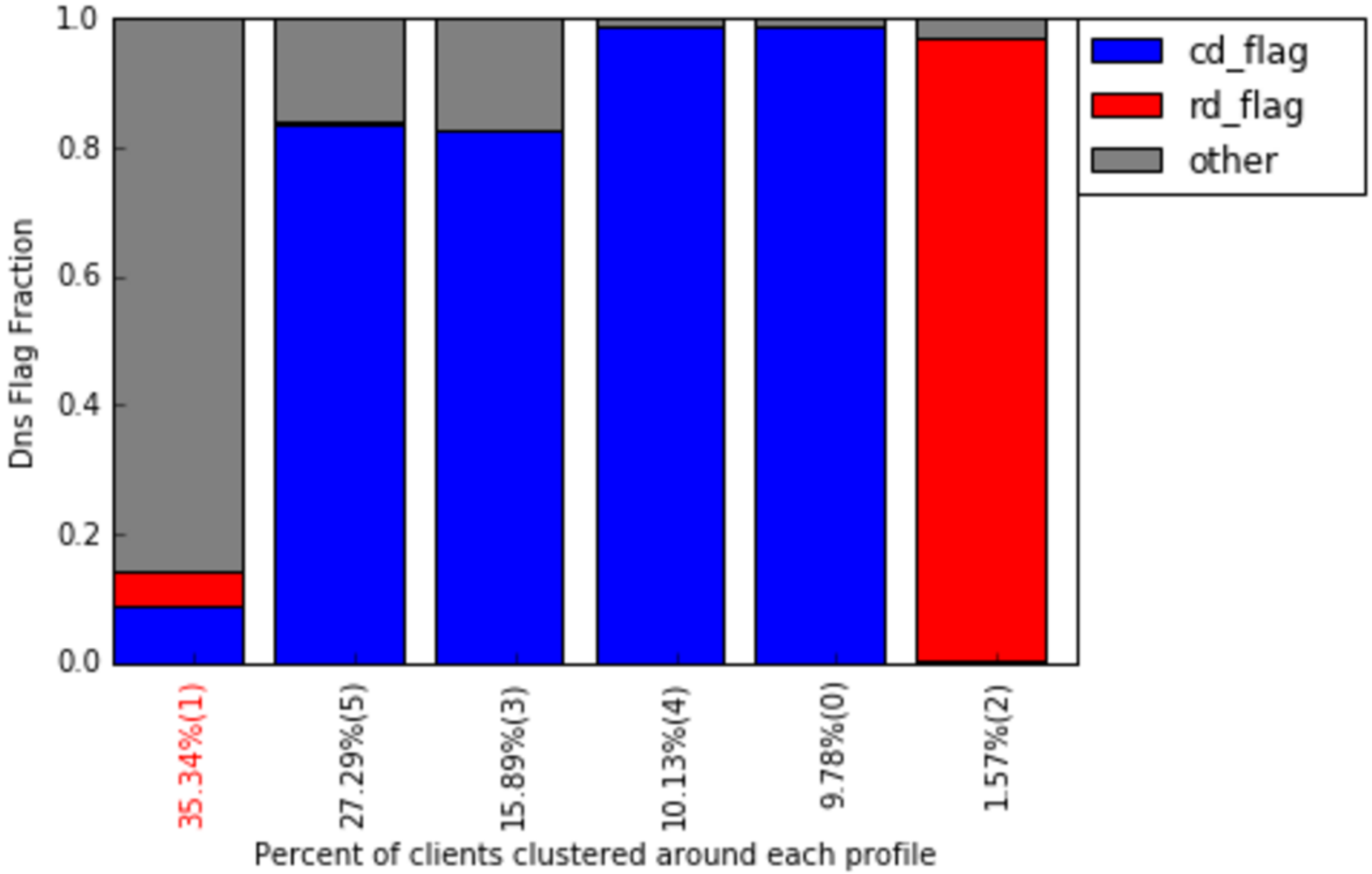
Query Type Profile per cluster



Rcode profile per cluster



Flag profile per cluster



Clustering accuracy

- How many known resolvers fall in the same cluster?
  - How many known non-resolvers?
- Tested on both week day and weekend, 98% ~ 99% known resolvers fit in the same cluster

```
df_res_label
```

	label	resolver_ip	total	percent
0	1	831	839	99.05%
1	3	4	839	0.48%
2	4	3	839	0.36%
3	5	1	839	0.12%

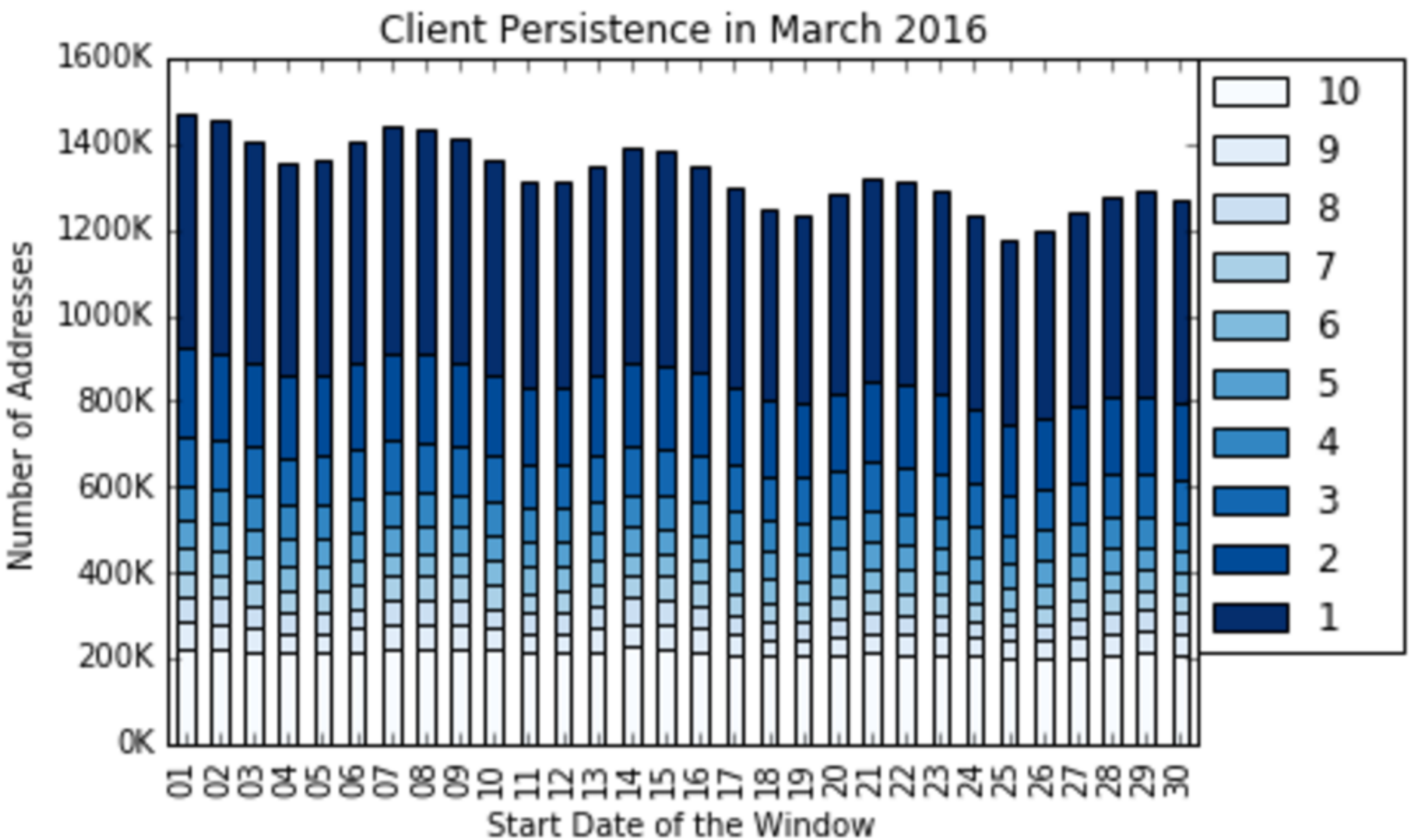
```
df_nonres_label
```

	label	nonres_ip	total	percent
0	1	74	275	26.91%
1	2	200	275	72.73%
2	4	1	275	0.36%

Client persistence

- Another differentiating factor could be client persistence
- Within a 10-day rolling window, count the addresses seen on specific number of days
- Addresses sending traffic all the time will fit into known resolvers and monitoring roles

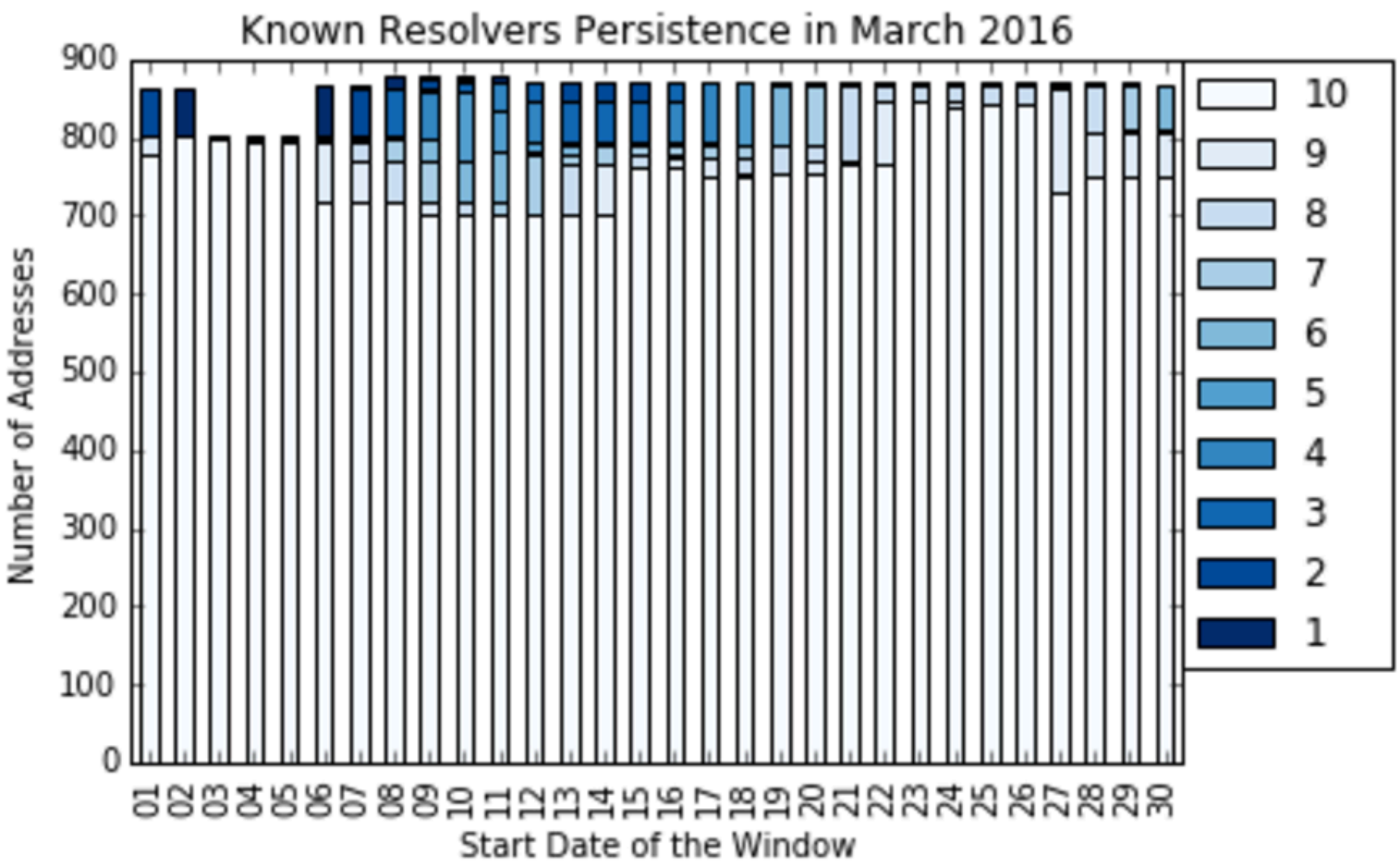
Client Persistence



Resolvers persistence

- Do the known resolvers addresses fall into the hypothesis of persistence?
- What if we check their presence in different levels?

Resolvers persistence



Future work

- Identify unknown resolvers by checking membership to the “resolver like” cluster
- Exchange information with other operators about known resolvers.

- Potential uses: curated list of addresses, white listing, others.

Conclusions

- This analysis can be repeated to other ccTLDs
- Using open source tools
- Code analysis will be made available
- Easily adaptable to use ENTRADA

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