Kindred Domains: Detecting and Clustering Botnet Domains Using DNS Traffic

Matt Thomas

Data Architect, Verisign Labs

About the Author

Matthew Thomas

Data Architect

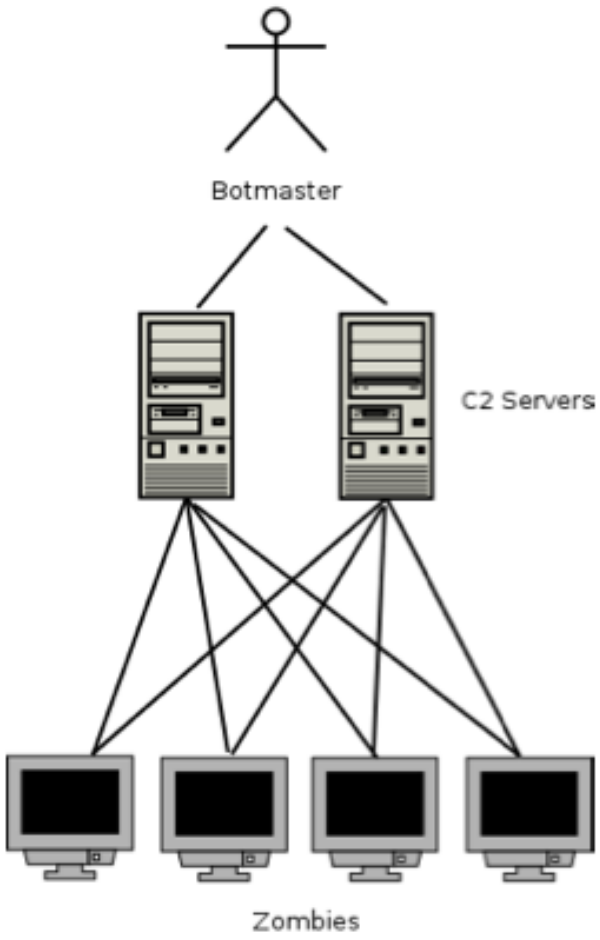
Verisign Labs

Aziz Mohaisen

Sr. Research Scientist

Verisign Labs

An Overview of Command & Control Botnets



- Malware commonly uses Domain Fast-Fluxing or Domain Generation Algorithms (DGA)
- Typically seeded by system clock
- Domains span many TLDs
- DNS traffic lookup patterns will emerge as all infected hosts resolve the same set of domains
  - Most result in NXDomain

Conficker: The Quintessential DGA

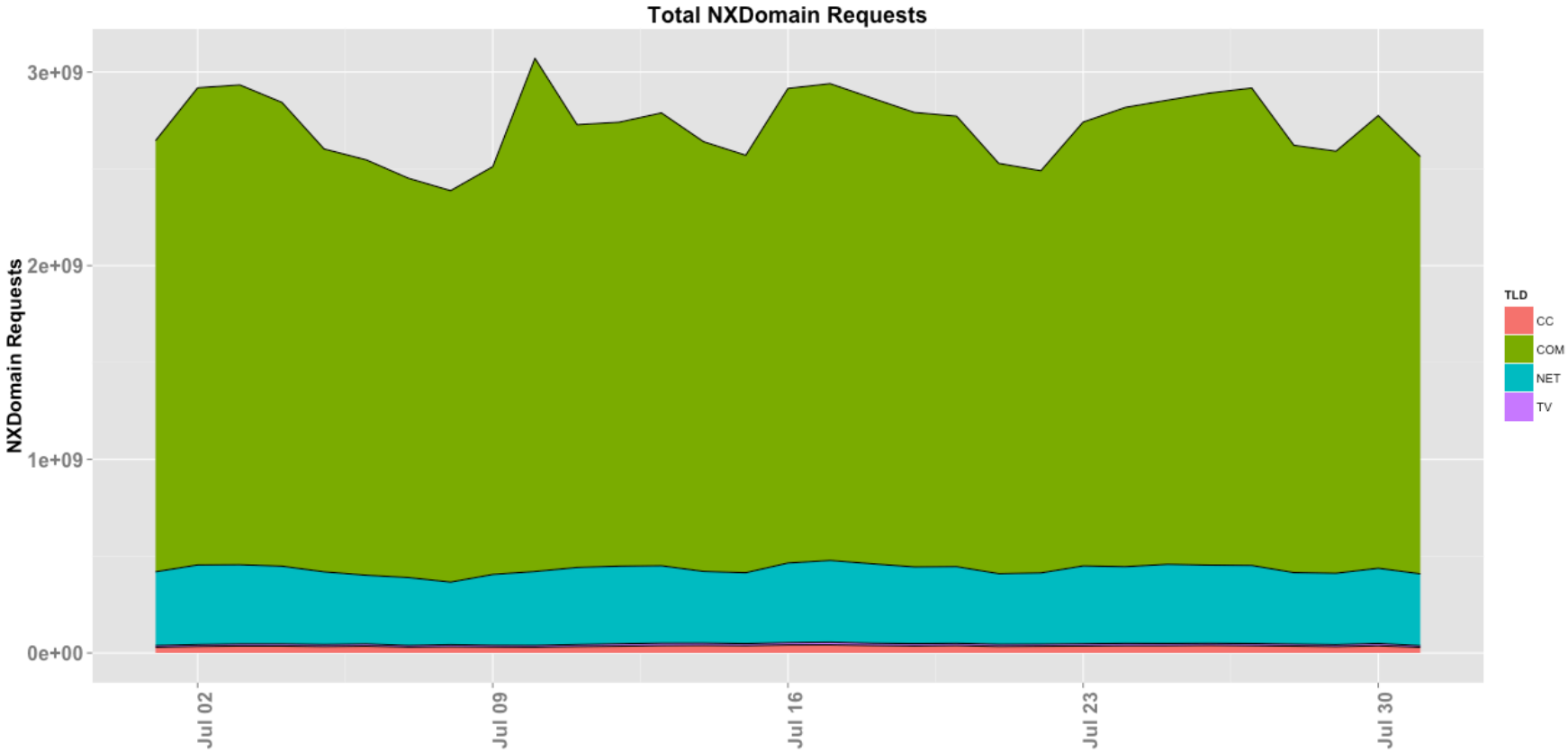
- A malware family that popularized the concept of DGA
- Thousands of infected machines still exist
- Many variants – each generates different set of domains

Variant	Domains / Day	TLDs
A	250	biz, info, org, net, and com
B	250	biz, info, org, net, com, ws, cc, cn
C	50k	110 ccTLDs not including tv or cc

- DGA Algorithm was reverse engineered
  - Provides set of domains generated for each variant for a given day
  - Useful “ground truth” for detection

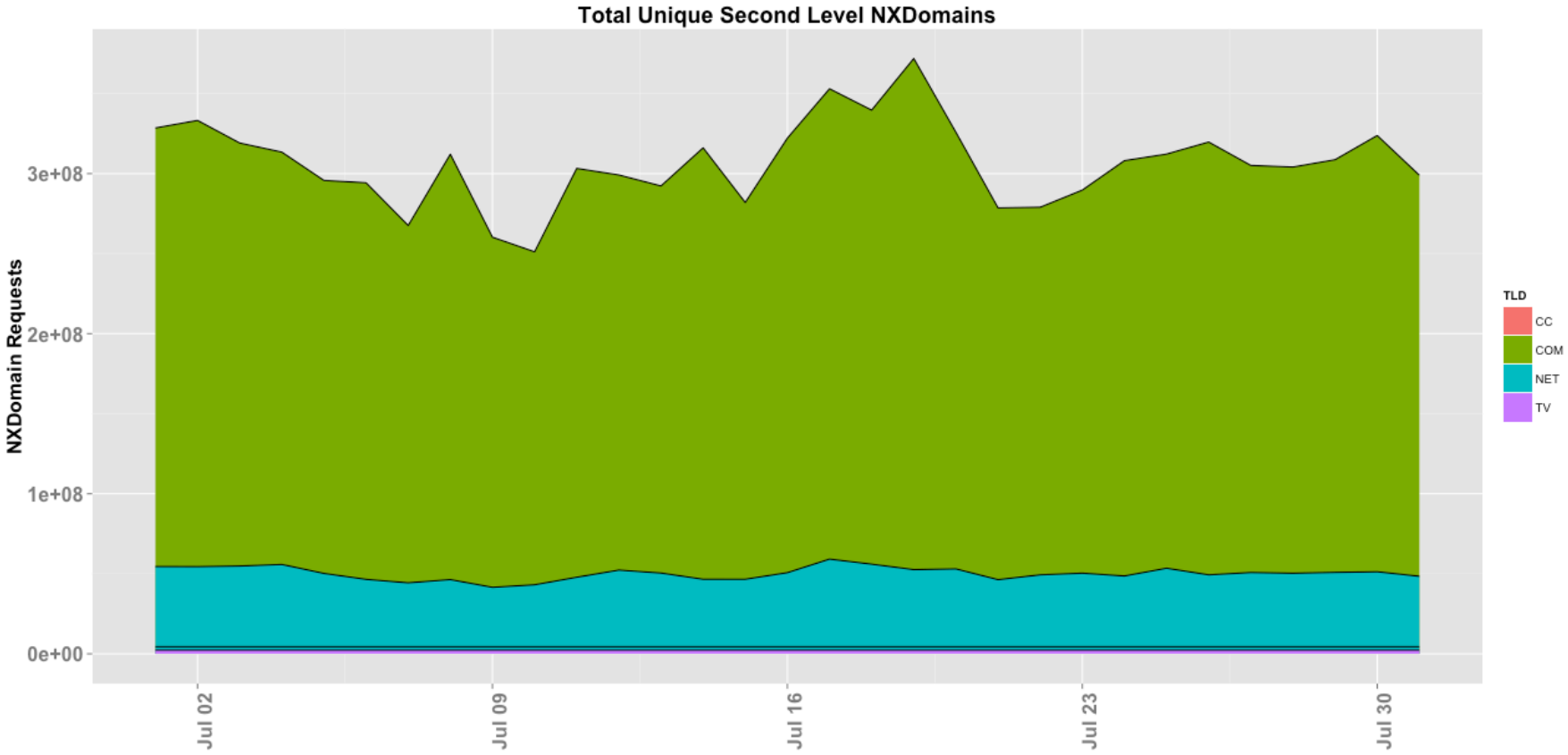
How does Conficker traffic differ from typical NXD traffic?

NXDomain Traffic Data



- Daily volume of NXD traffic for CC/TV/NET/COM
- NXD traffic is associated with domains not registered, mistyped, etc

NXDomain Traffic Data

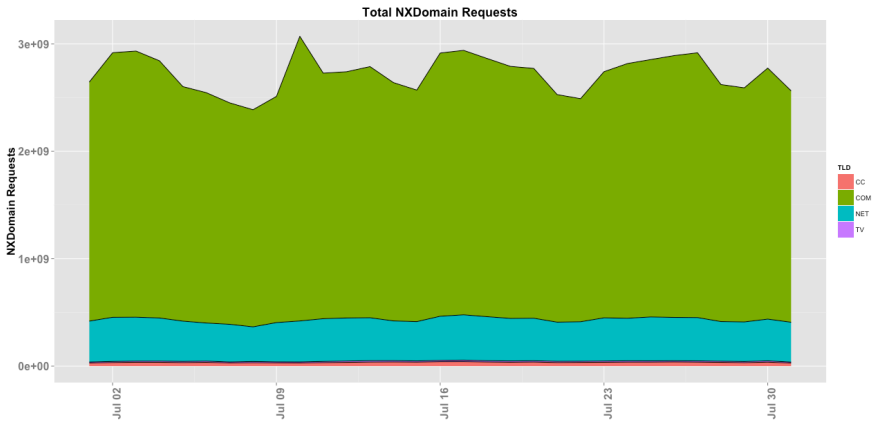


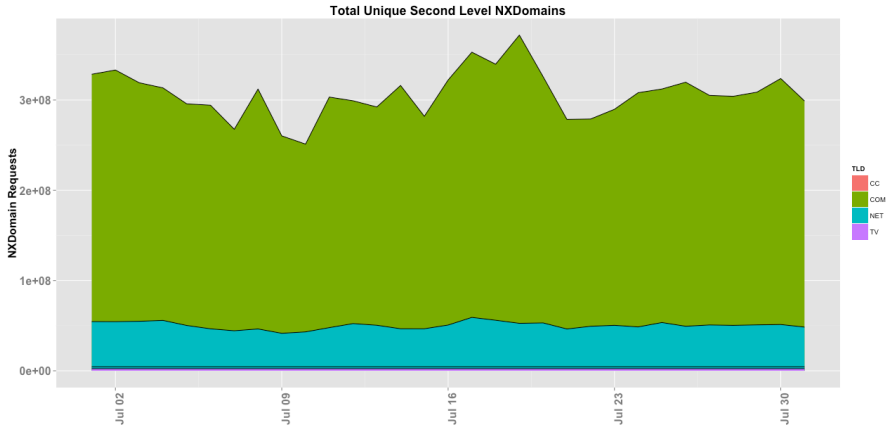
- Daily number of unique Second Level Domains (SLDs)
- The traffic is NXD (not registered domains, mistyped, etc)

NXDomain Traffic Data

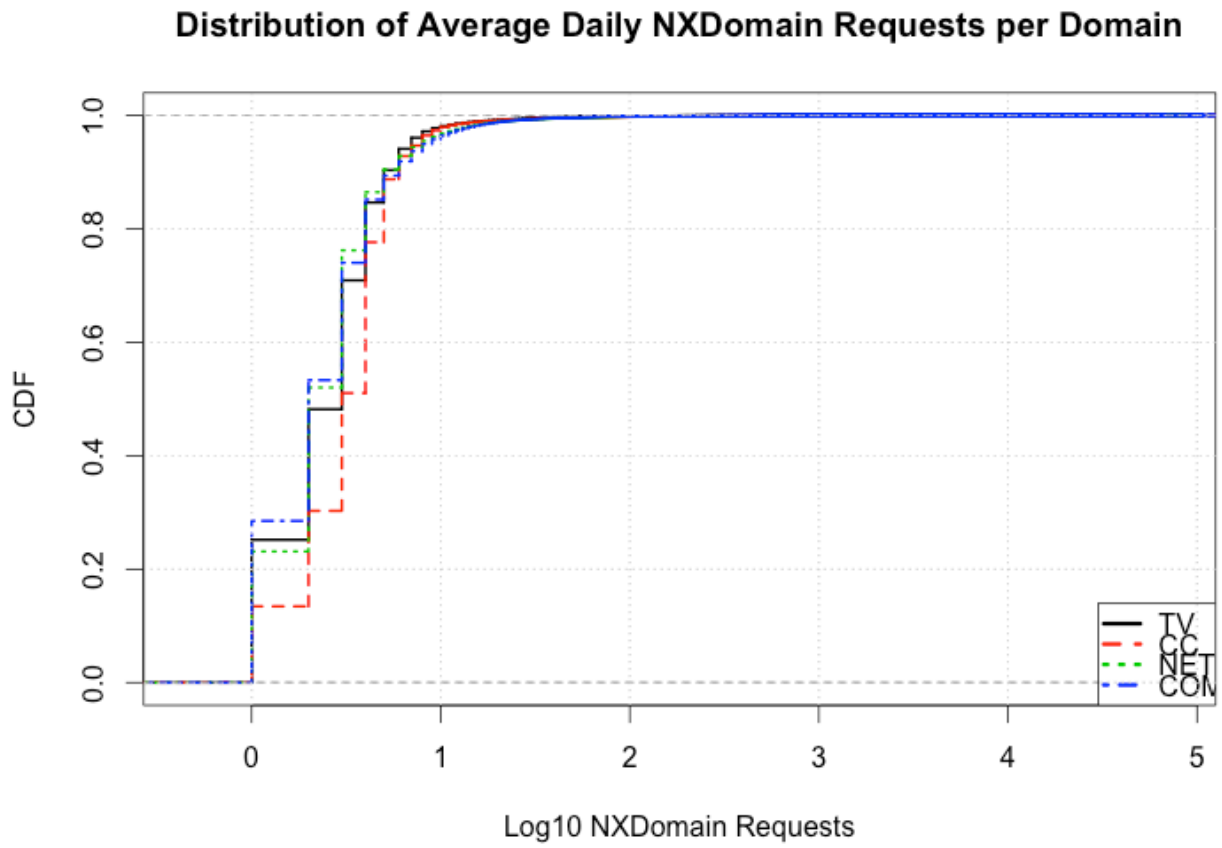
- COM typically sees ~2.5 billion NXD requests per day and spans over 350 million unique SLDs
- NET receives 500 million NXD requests over 60M SLDs (SLDs that are not registered, or mistyped)
- ccTLDs receive significantly less traffic, and less NXD

What's the distribution of requests per SLD?



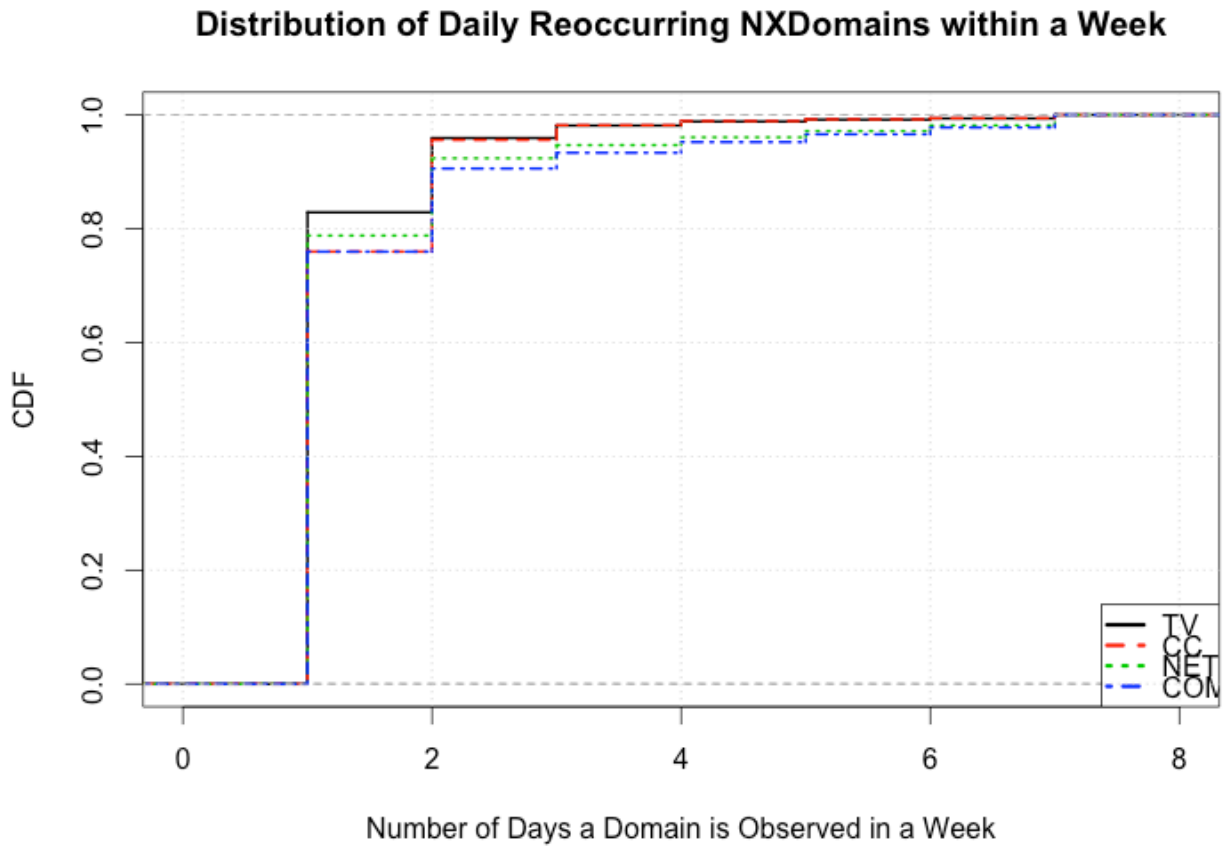


NXDomain Traffic Data



- CDF of average number of NXD requests per SLD

NXDomain Traffic Data

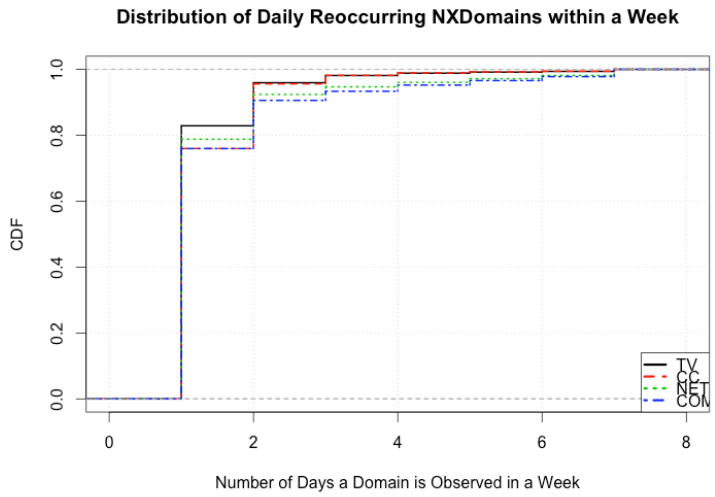


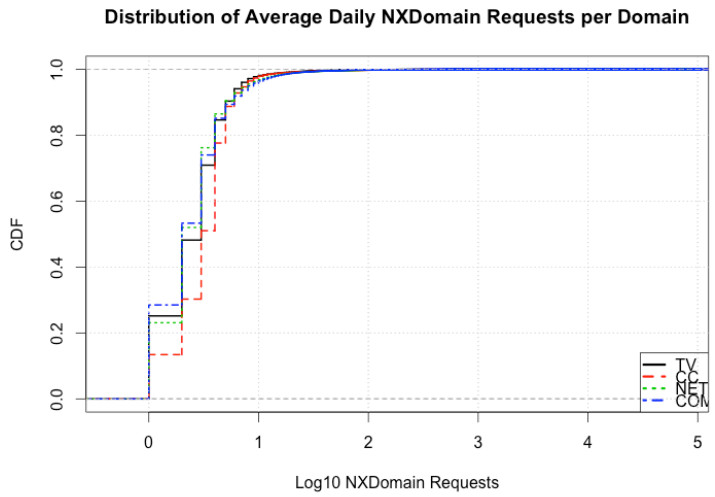
- CDF of number of days an SLD appears in a week

NXDomain Traffic Data

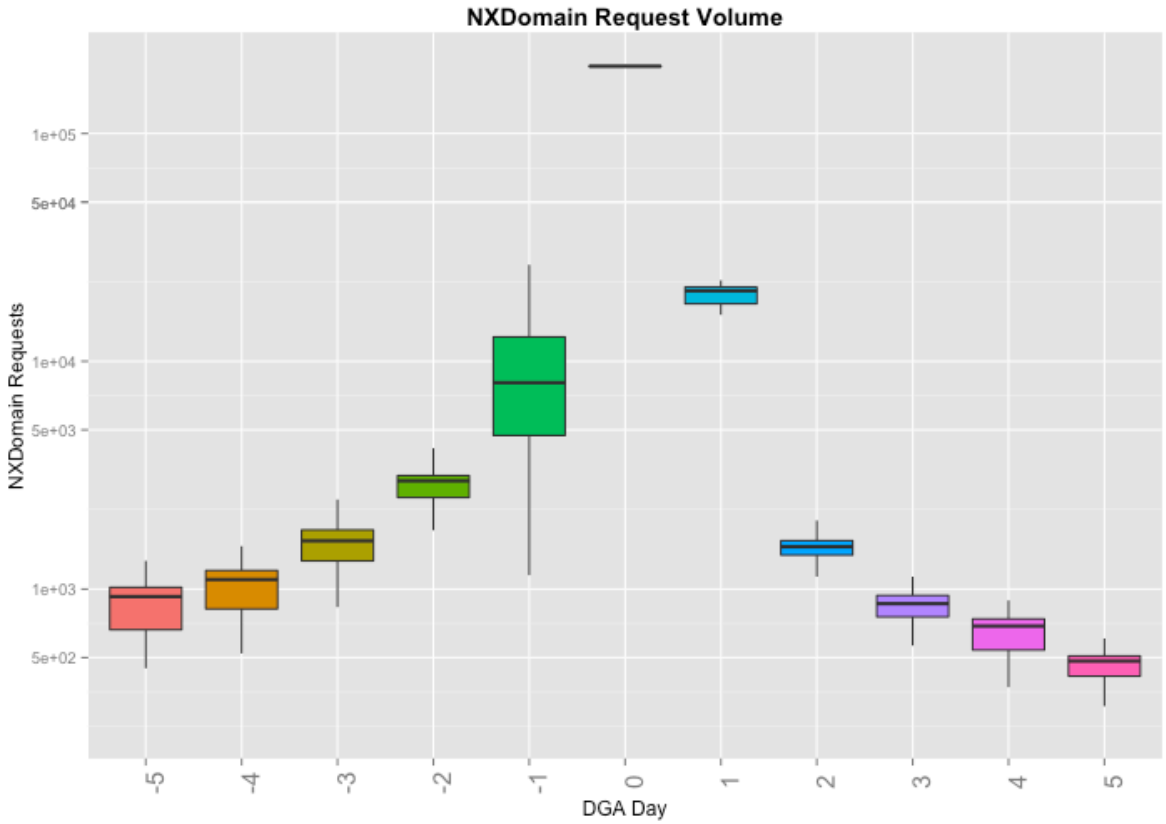
- Average SLD receives minimal amount of NXD traffic
  - 95% of SLD's NXD receive less than 10 requests within 24 hours
- High "churn" rate within the SLD set during a week

How does Conficker's NXD traffic compare?



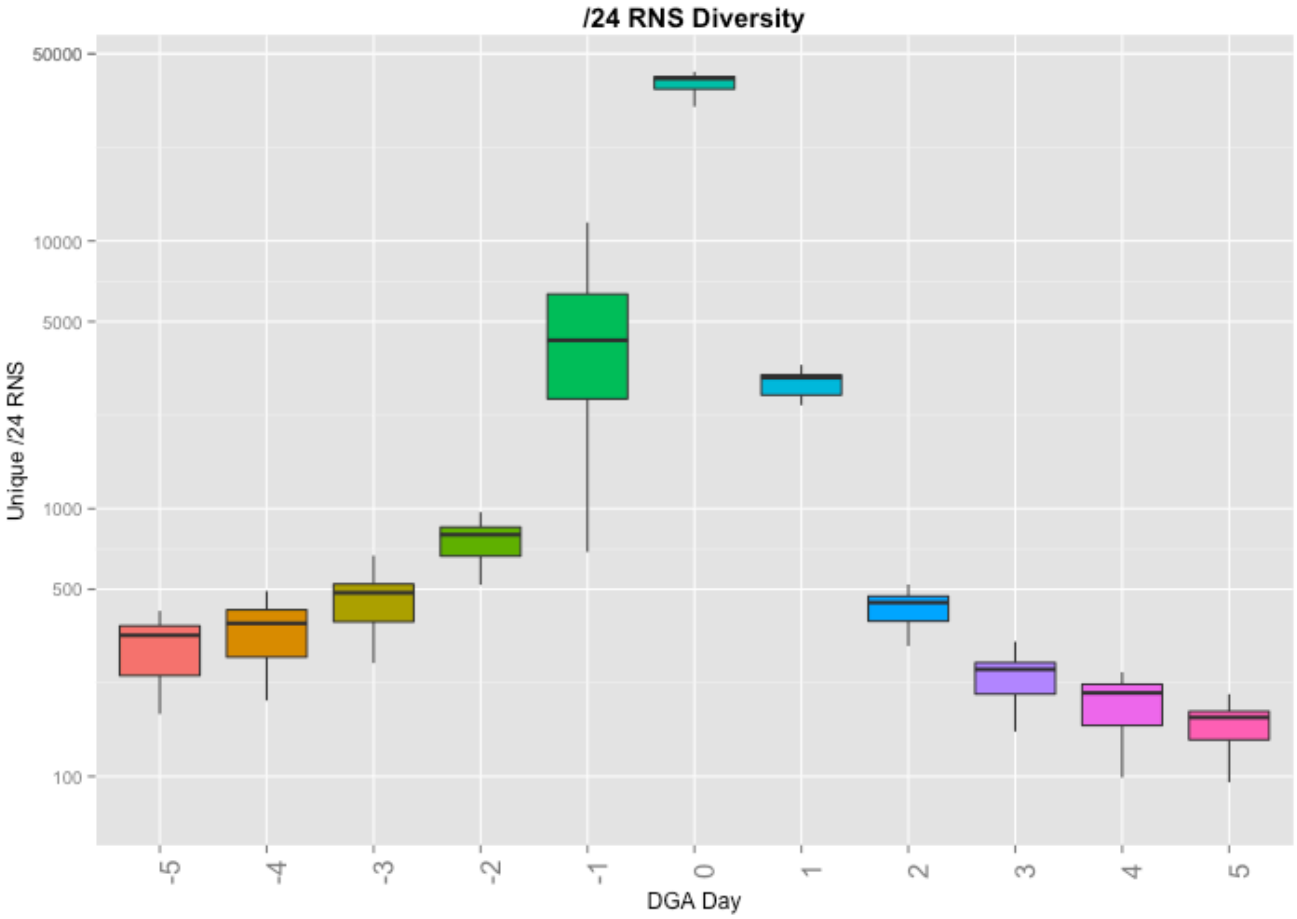


Conficker's NXDomain Traffic Data



- NXD traffic volume for DGA domains prior, during and after their expected generation date

Conficker's NXDomain Traffic Data

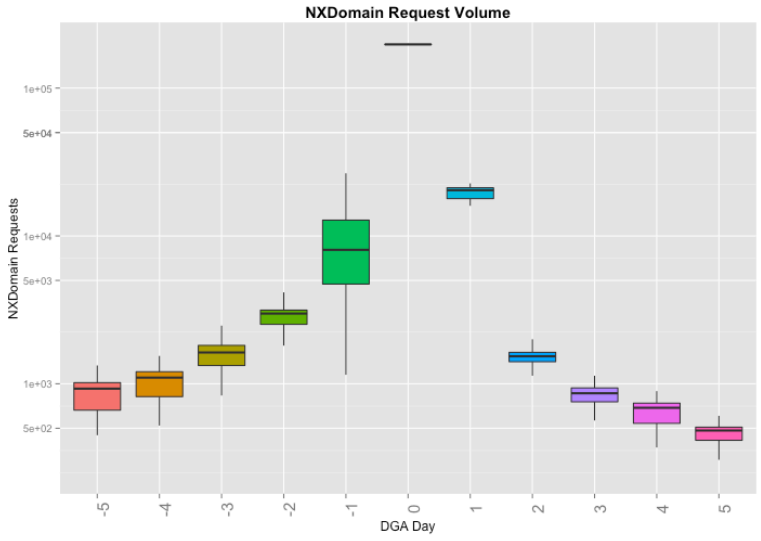


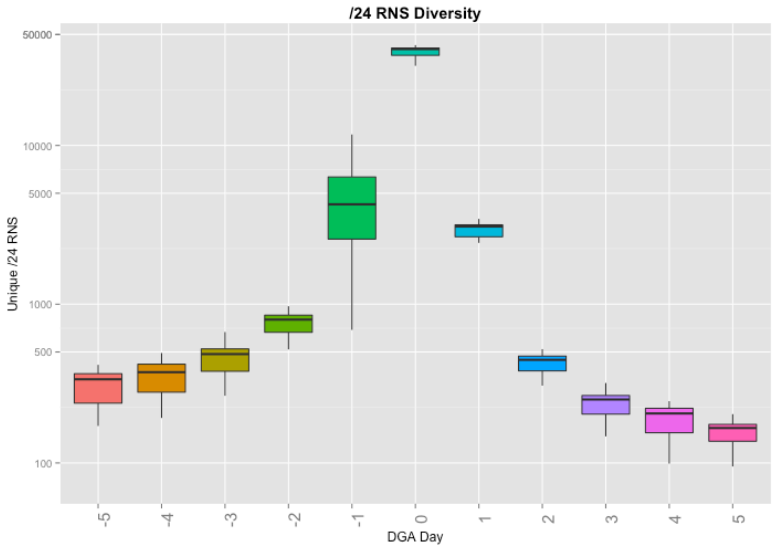
- NXD traffic diversity (/24 of the RNS) for DGA domains prior, during and after their expected generation date

Conficker's NXDomain Traffic Data

- Despite specific generation date, DGA domains receive traffic pre and post its specific generation date
  - Possible global clock skew; also possible misconfiguration
- Large amount of traffic from diverse set of RNS for SLDs
  - Statistical abnormal compare to whole NXD population

How can we detect and associate malware domains?





Detection and Clustering

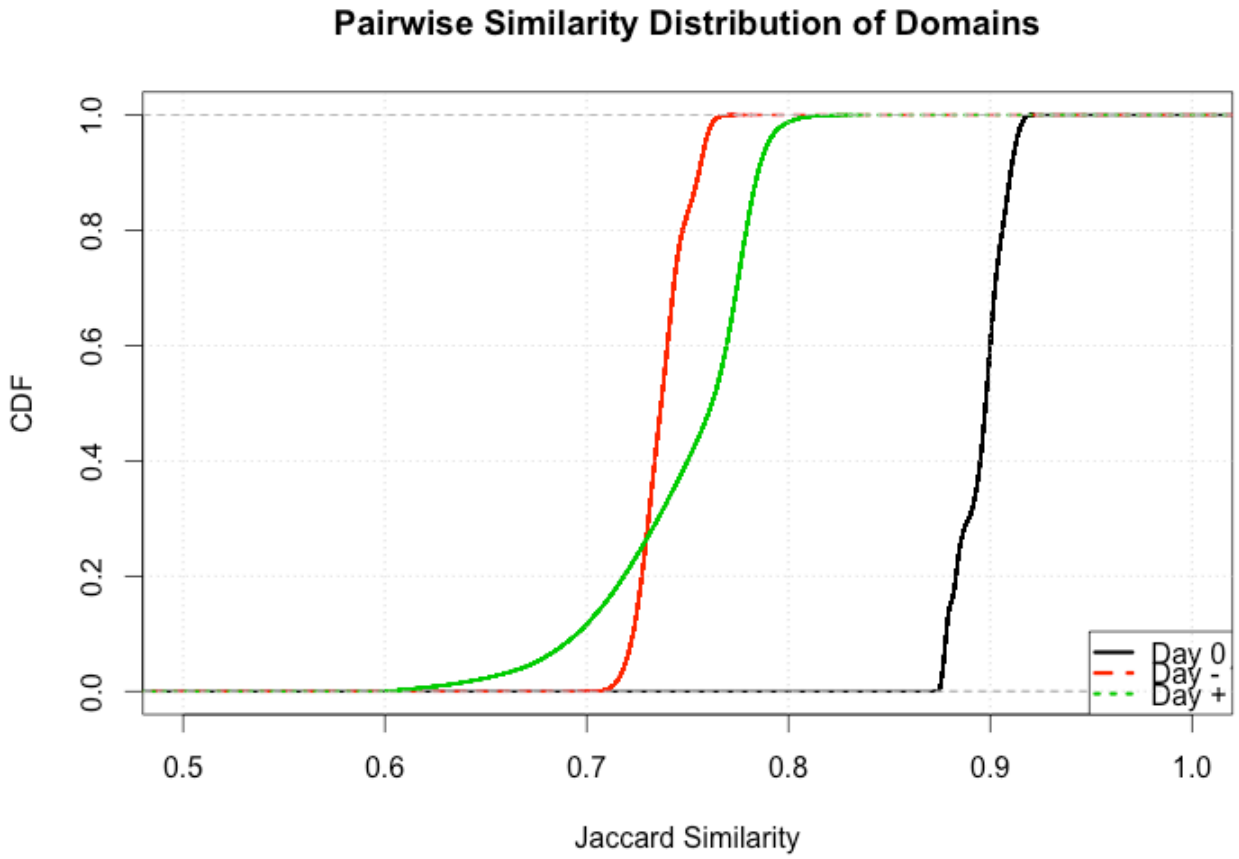
Computing Traffic Similarity

- We try to address the following questions:
  - How similar are traffic streams to each other?
  - Can the similarity be used to group different traffic streams?

- The similarity function is a real-valued function that quantifies the *similarity* between two entities
- Jaccard Index is a statistic for comparing the similarity and diversity of sample sets (one among many)

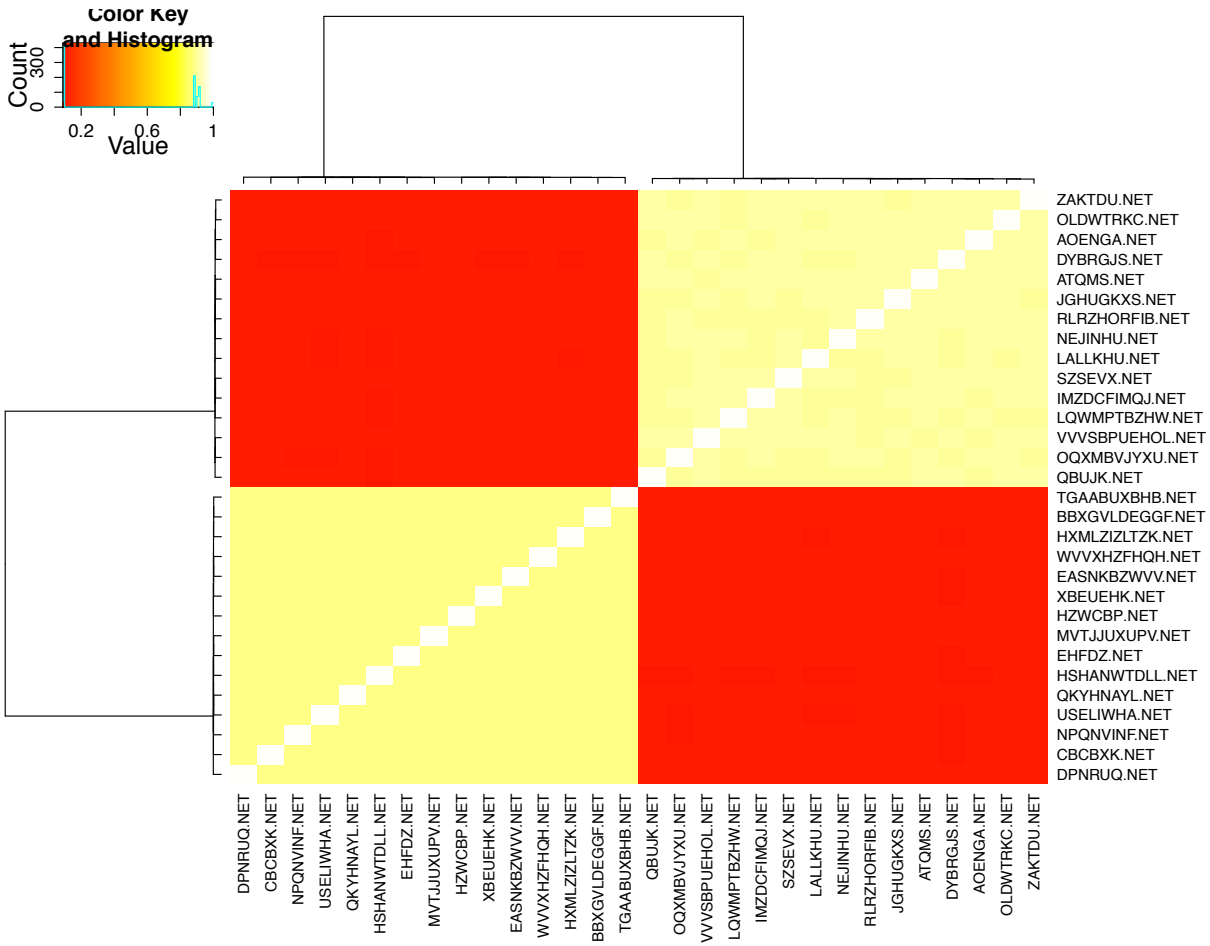
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \qquad 0 \leq J(A,B) \leq 1.$$

Conficker's Traffic Similarity



- CDF of pairwise domain similarities for a set of DGA domains based on on their /24 RNS set for a given day

Conficker's Traffic Similarity

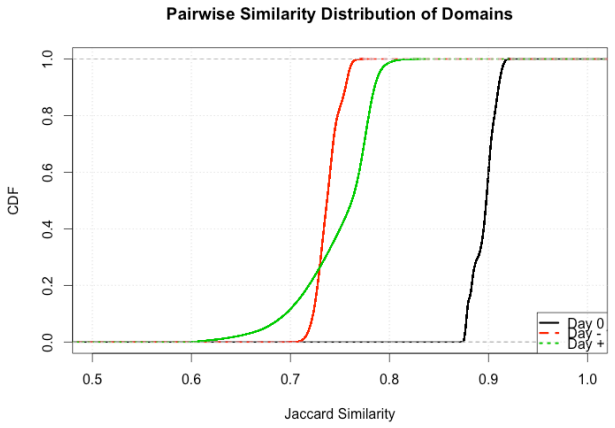


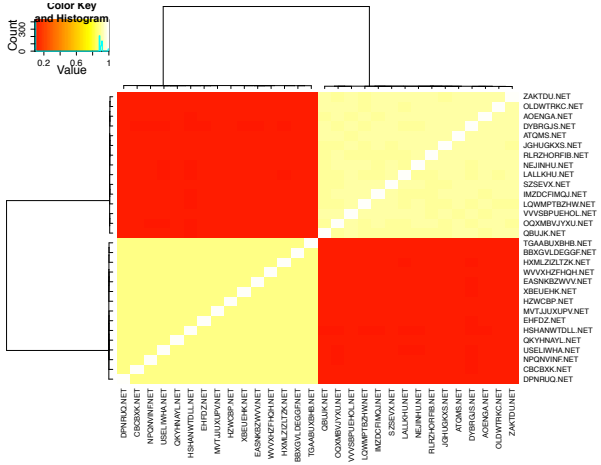
- Subset of domains from Conficker A & B clustered based on similarity using single-linkage algorithm

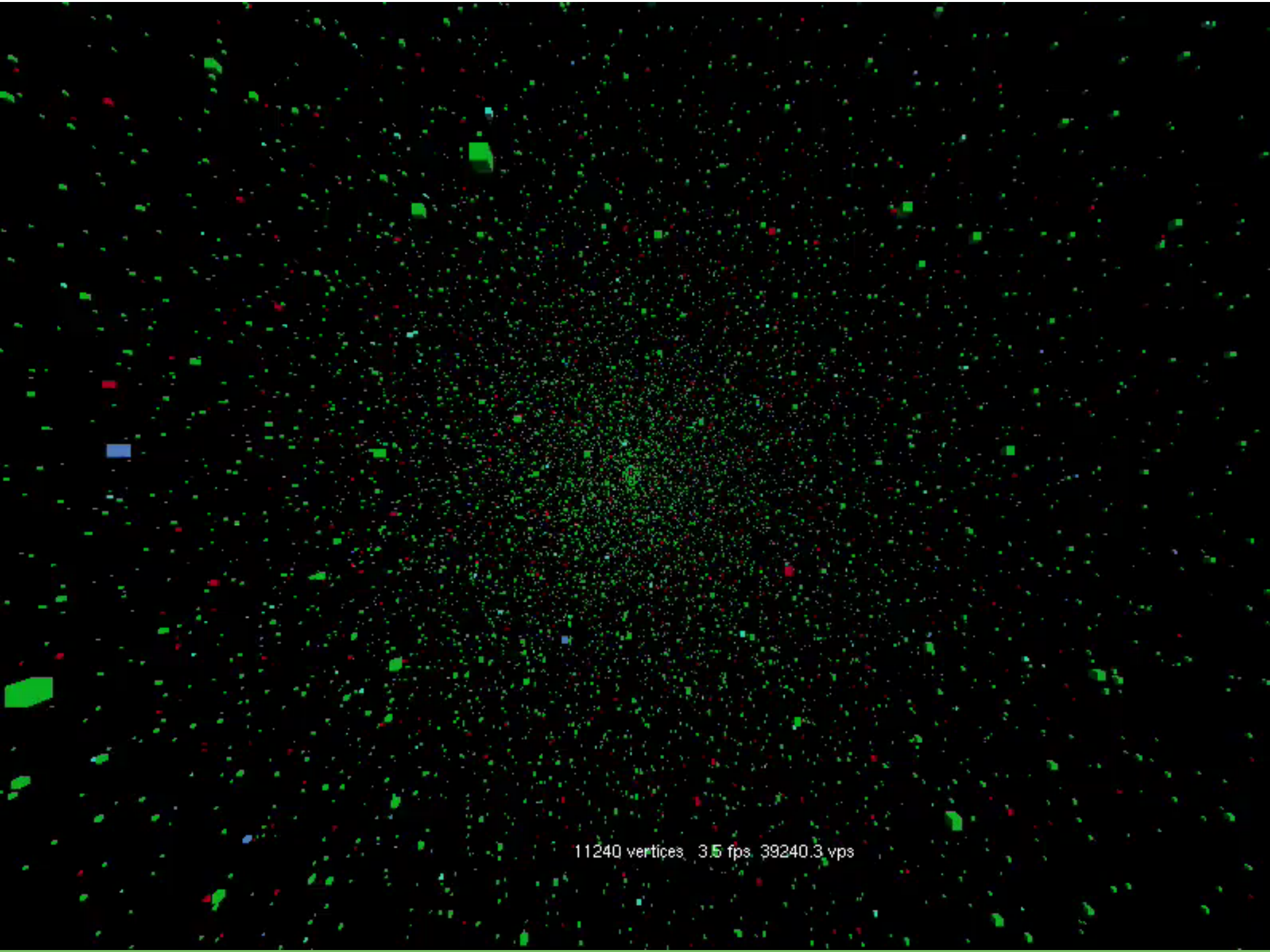
Conficker's NXDomain Traffic Data

- Domains on a specific DGA date have very high similarity measures, most measuring higher than 0.9
- Techniques such as hierarchical clustering could potentially group domains from a specific DGA into distinct clusters based on DNS traffic similarity

How do various similarity thresholds affect the cluster?

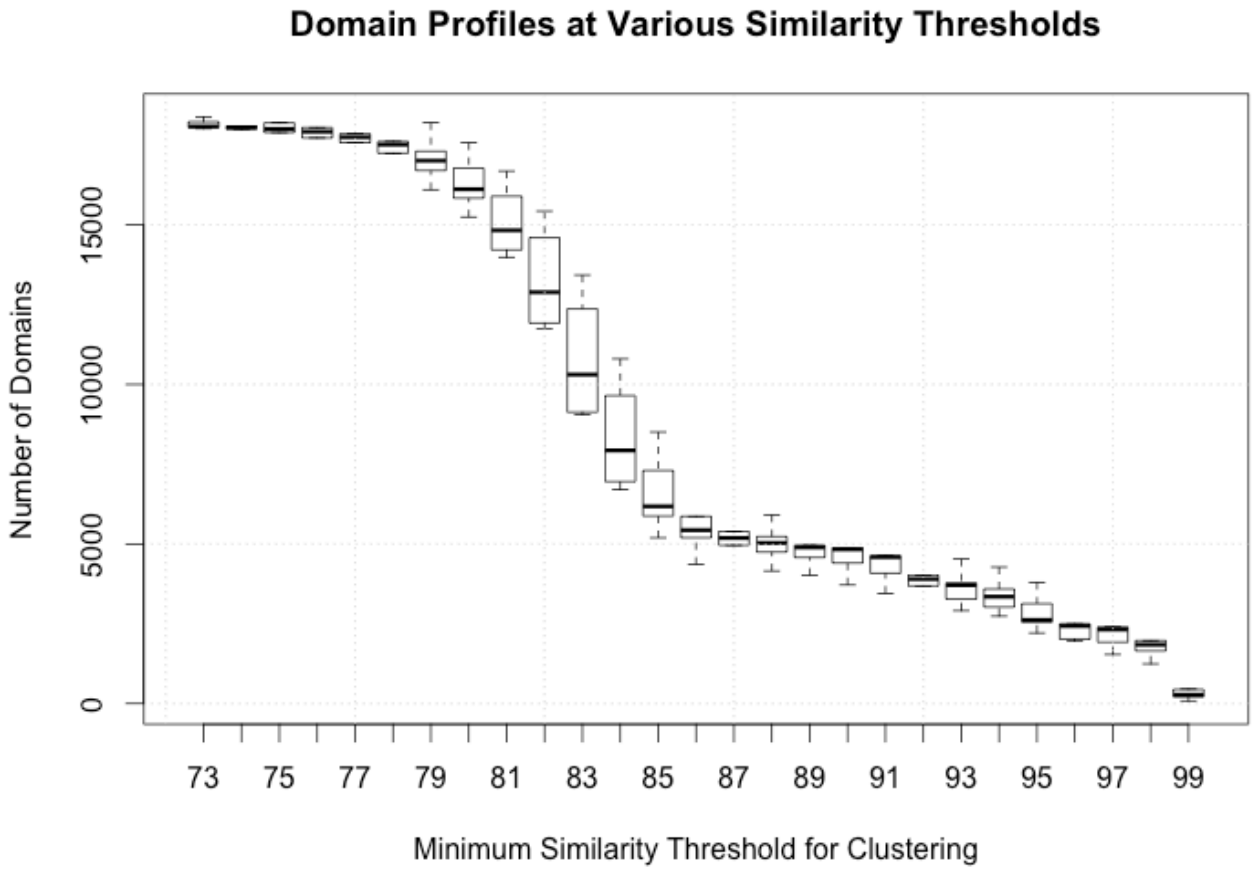





11240 vertices 3.5 fps 39240.3 vps

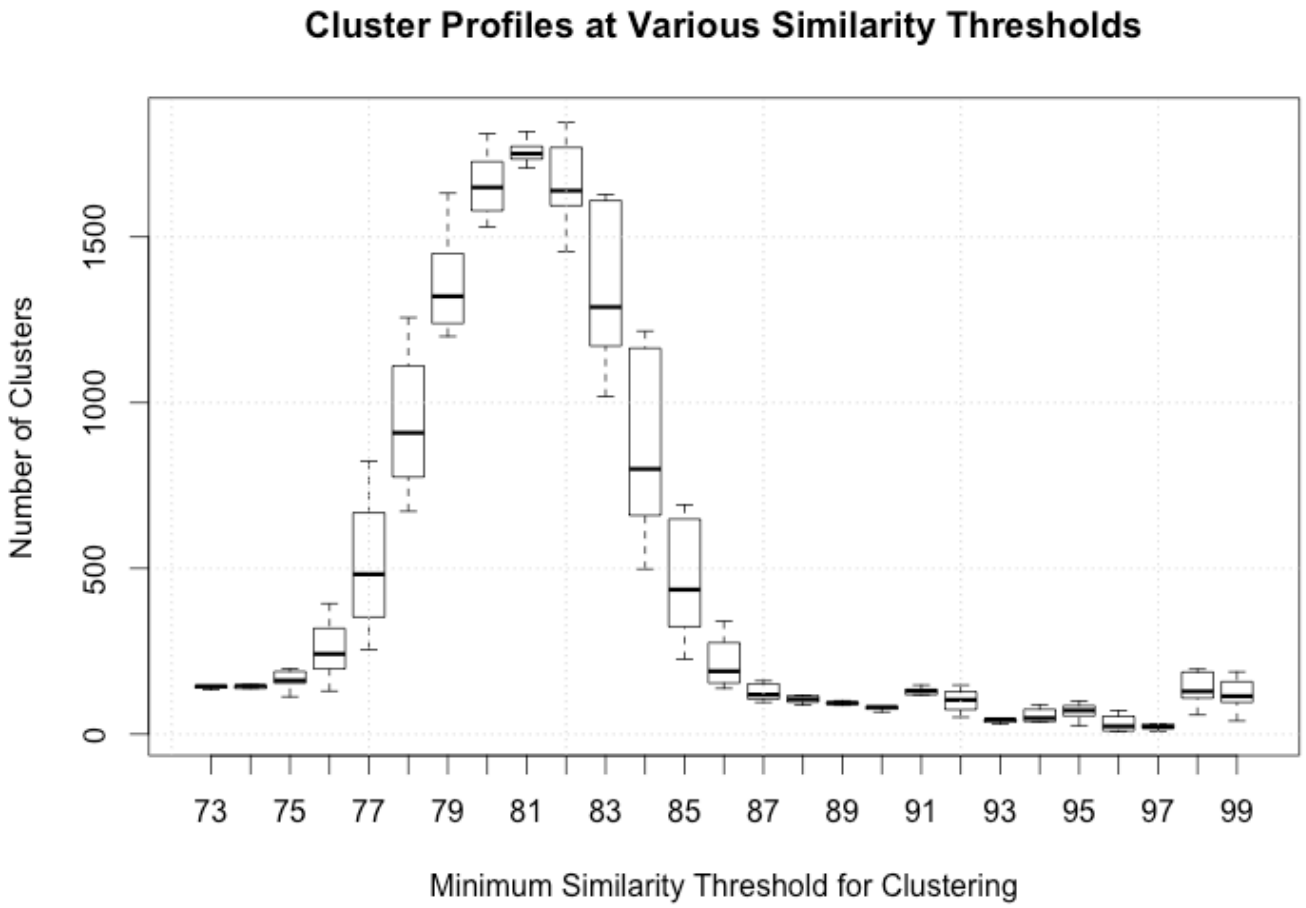
Clustering – Similarity Thresholds

Clustering – Similarity Thresholds



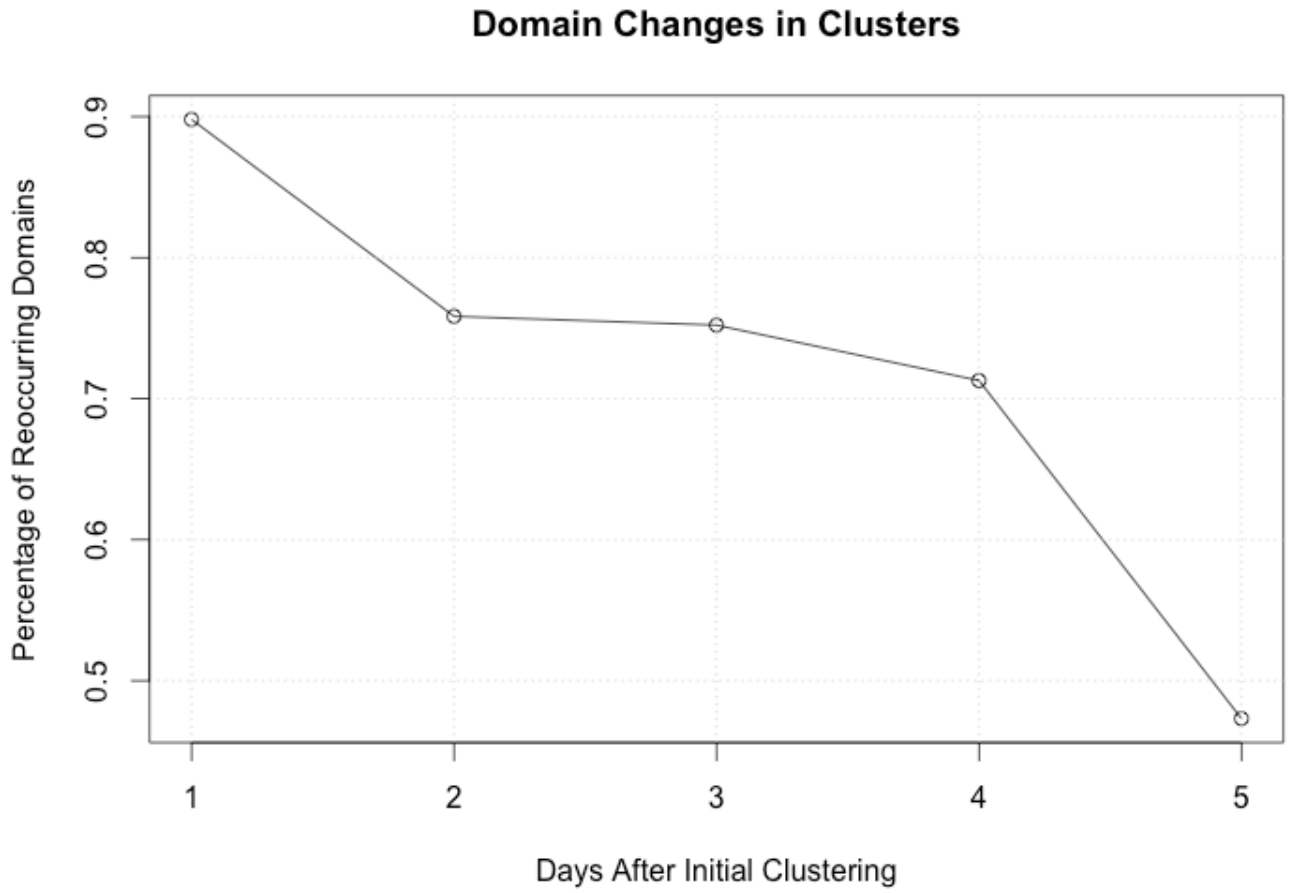
- The number of SLDs contained in a cluster at various similarity threshold levels

Clustering – Similarity Thresholds



- The number of distinct clusters formed at various similarity threshold levels

Clustering – Similarity Thresholds



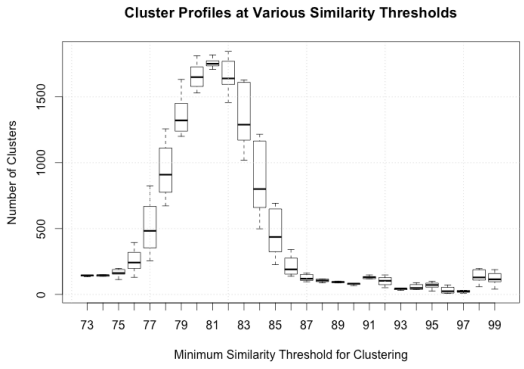
- Temporal evolution of clusters based on SLDs present

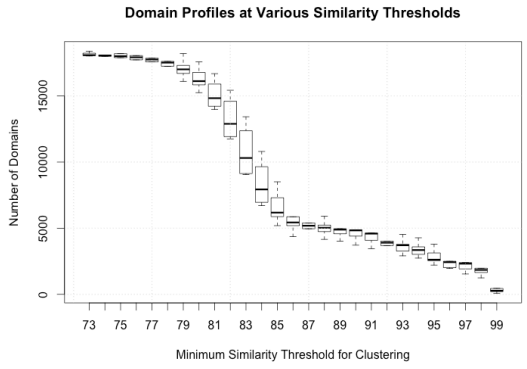
Global Malware Detection

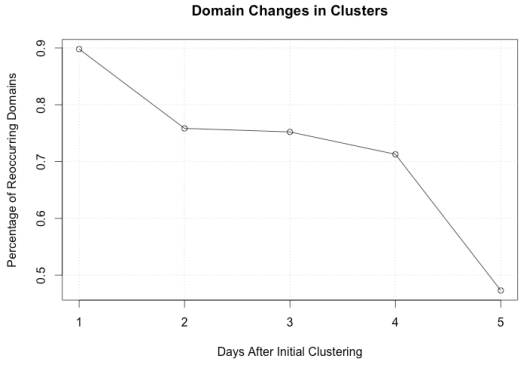
Global Malware Detection

- Similarity thresholds influence the number of clusters and the number of domains within each cluster

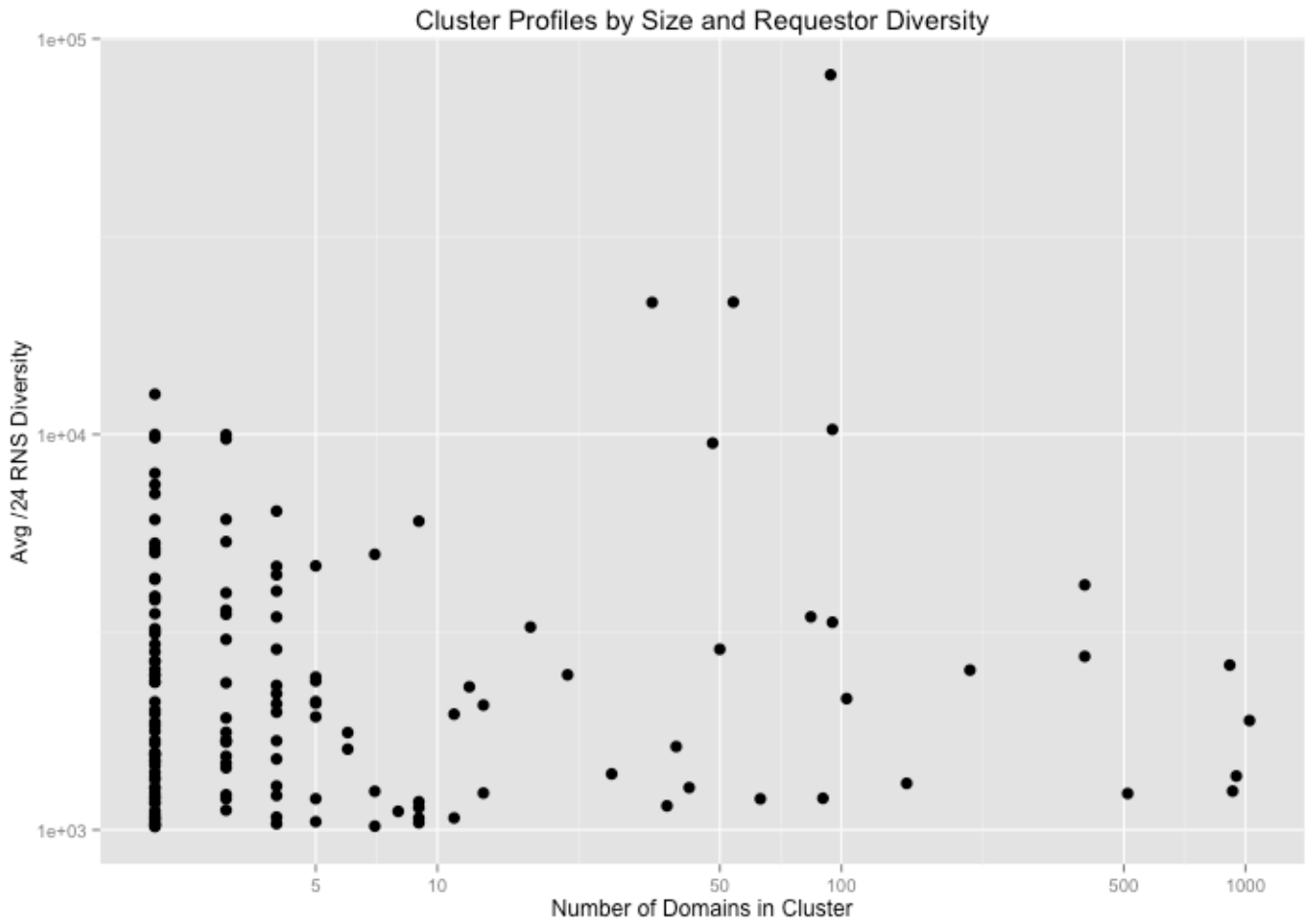
What resulting clusters appear when such a technique is applied to all NXD traffic?







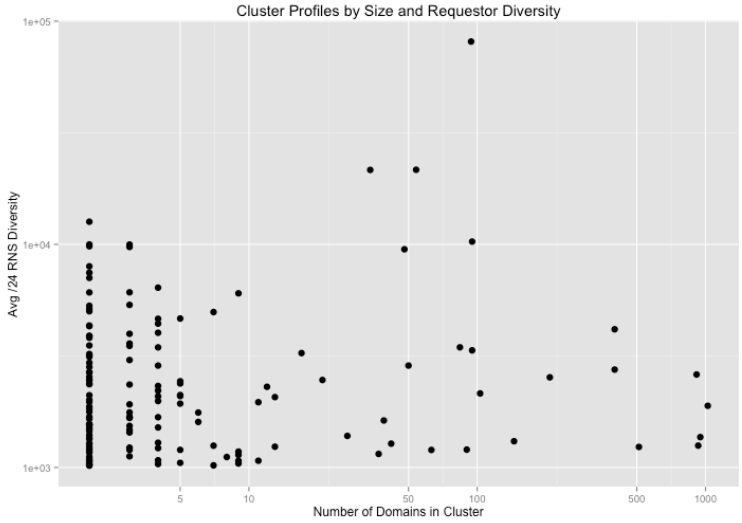
Global Malware Detection



- Clusters identified with a similarity threshold set to > 0.9
- Each point is a cluster – measures # domains and # RNS

Global Malware Detection

- Many detected clusters of malware/variants use a small amount of domains observable in our dataset
- A few clusters generated several hundred domains
  - May influence evasiveness and resilience of a botnet
- Several clusters have thousands of distinct /24s
  - Infection rate or prevalence of a malware



Concluding Remarks and Future Work

- We look at using the authoritative NXD traffic for identifying DGA's used as the C&C channel of malware
  - We use the largest dataset from com/net resolution
  - Domain names used for C&C are identifiable by their request pattern
  - Different generations of the Conficker malware family are identified
  - Clustering of traffic yields interesting structures identifying evasion
- Future works: We will look into extending the work
  - To other malware families using DNS for C&C
  - Highlight operation impact and evolution of evasion techniques
  - Explore spread of infections via remote sensing at the DNS level

powered by

VERISIGN™