



# Characterizing DNS Client Behavior Using Hierarchical Aggregate Entropy

---

2010/2/1

Keisuke Ishibashi, NTT Information Platform Labs

Masaharu Sato, NTT Communications

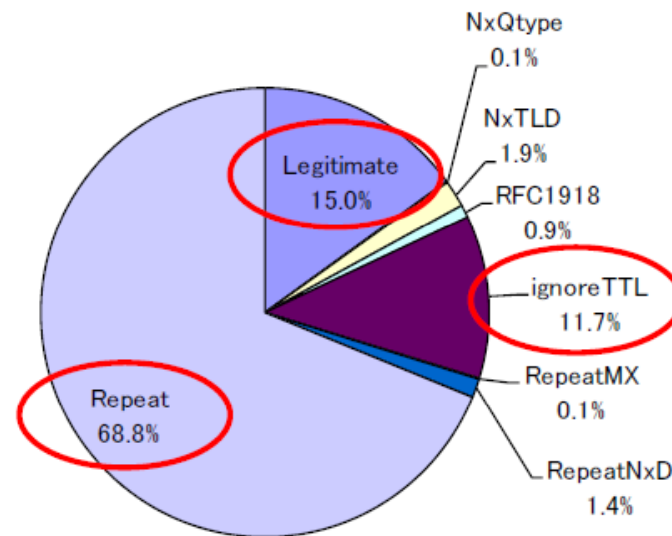
# Motivation

- Bogus queries are consuming resources of both DNS authoritative servers and caching servers

Type	Count	Percent
Unused Query Class	36,313	.024
A for A	10,739,857	7.03
Unknown TLD	19,165,840	12.5
Nonprintable in query	2,962,471	1.94
RFC1918 PTR	2,452,806	1.61
Identical Query	38,838,688	25.4
Repeated Query	68,610,091	44.9
Referral Not Cached	6,653,690	4.36
<b>Legitimate</b>	<b>3,284,569</b>	<b>2.15</b>

TABLE II

QUERY CLASSIFICATION RESULTS (24-HOUR PERIOD ON 4 OCTOBER 2002 AT THE F-ROOT DNS SERVER).



To root servers[wessels]

To caching servers[toyono]

[Wessels] D. Wessels et.al., "Wow, That's a Lot of Packets, " PAM 2003.

[Toyono] T. Toyono et. al., "An analysis of the queries from the view point of caching servers," 2007 DNS-Operations Workshop.

## Motivation cont'd

- Most of bogus queries are sent by small number of heavy clients [wessels][toyono]
  - Filtering queries sent by those heavy clients is efficient to protect DNS server resources

type \ rate	100qps	200qps	300qps	400qps	500qps	(Percentage of total queries)
<b>Legitimate</b>	<b>0.09%</b>	<b>0.01%</b>	0%	0%	0%	
NxQtype	0%	0%	0%	0%	0%	
NxTLD	0%	0%	0%	0%	0%	
RFC1918	0.80%	0%	0%	0%	0%	
ignoreTTL	1.63%	0.05%	0.01%	0%	0%	
RepeatMX	0.01%	0%	0%	0%	0%	
RepeatNxD	0.64%	0%	0%	0%	0%	
<b>Repeat</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	

# Motivation cont'd

- However, not all heavy clients send only bogus queries!!
  - PTR queries from web servers (analog)
  - Aggregated queries from DNS proxies
  - Prefetch queries
- Needs to classify heavy clients into normal (legitimate) clients and abnormal (bogus) clients
- Classify heavy clients by their query patterns
- How to characterize the query patterns?

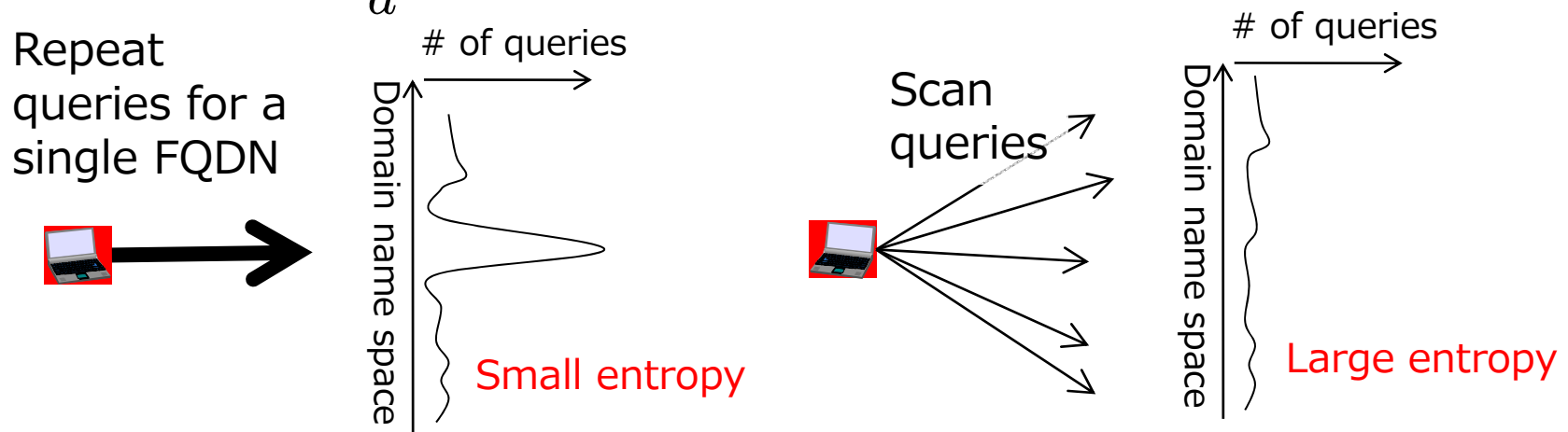


type \ rate	100qps	200qps	300qps	400qps	500qps	(Percentage of total queries)
<b>Legitimate</b>	<b>0.09%</b>	<b>0.01%</b>	0%	0%	0%	
NxQtype	0%	0%	0%	0%	0%	
NxTLD	0%	0%	0%	0%	0%	
RFC1918	0.80%	0%	0%	0%	0%	
ignoreTTL	1.63%	0.05%	0.01%	0%	0%	
RepeatMX	0.01%	0%	0%	0%	0%	
RepeatNxD	0.64%	0%	0%	0%	0%	
<b>Repeat</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	<b>59.69%</b>	

# Entropy based characterization

- Use of entropy of queries
  - Entropy: represents how queries disperse in name spaces

$$H(D) = \sum_d p_d \log_2(p_d) \quad p_d : \text{frequency of queries for domain name "d"}$$

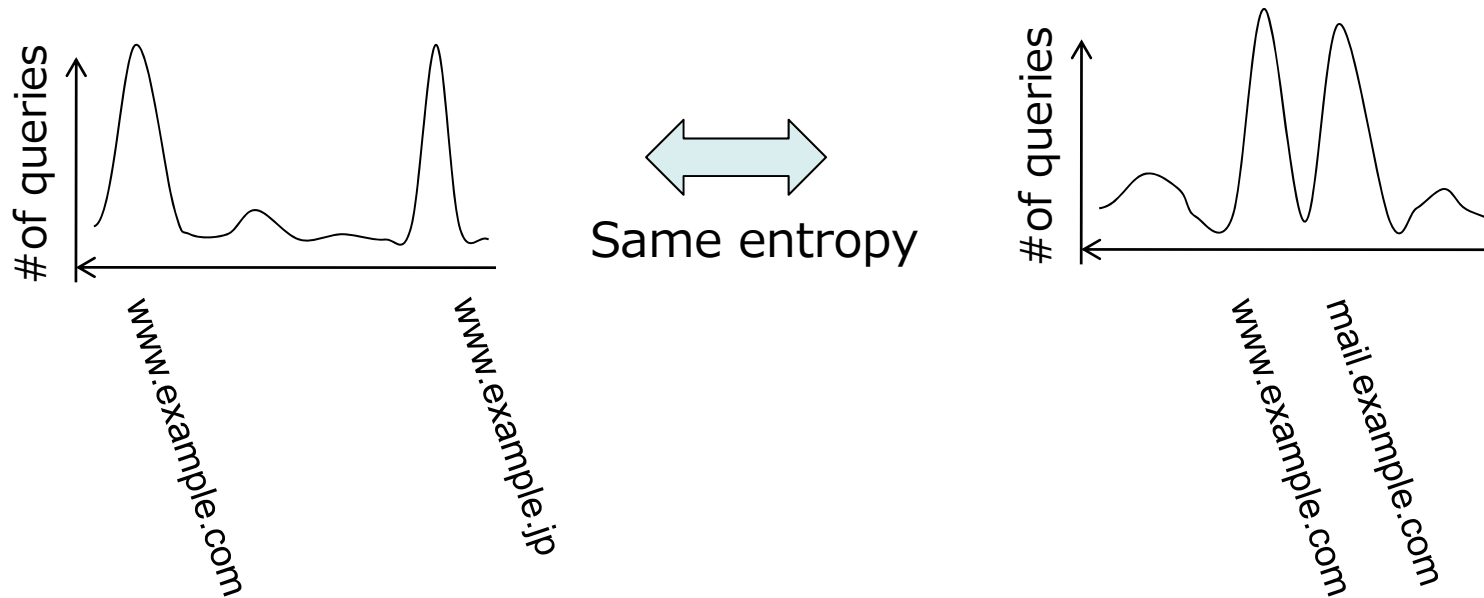


- Entropy of legitimate queries: expected to lie between them
- Calculate query entropies for heavy clients, and classify them using their entropies

Kazuya Takemori, et. al., "Entropy Study on A Resource Record DNS Query Traffic from the Campus Network," IEICE Tech. Rep. IA2008-84, Mar. 2008.

# Drawback of entropy based characterization

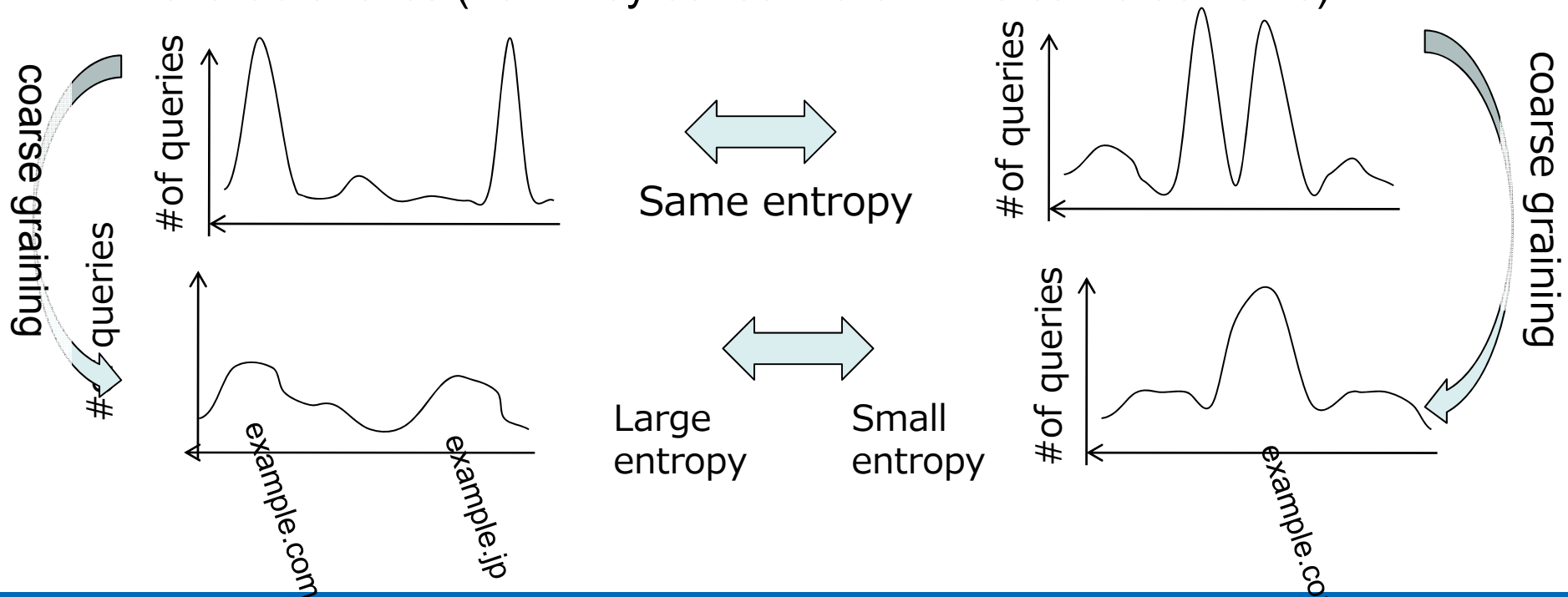
- Entropy does not tell information on spatial characteristics
  - Independent on where queries concentrate or diverse in domain name spaces
  - Only depends on how queries concentrate or disperse



# Hierarchical Aggregate Entropy

- Hierarchical Aggregate Entropy

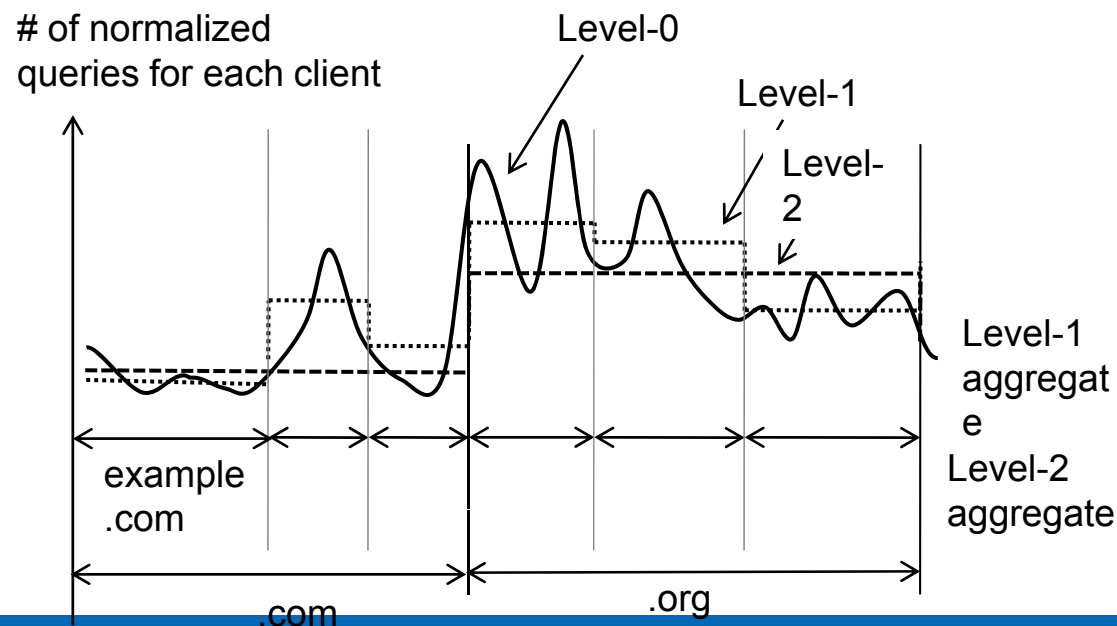
- Aggregating queries accordance to its hierarchical structure and calculate entropy for each hierarchy
- Decrease of query entropy with aggregation tells spatial characteristics (how they concentrate in the same domains)



# Hierarchical Aggregate Entropy (Cont'd)

- $H(D^{(0)})$  : Entropies of non-aggregate (FQDN) queries sent by a clients
- $H(D^{(0)})$  can be represented as the sum of following terms:
  - $H(D^{(2)})$ : Entropies of queries aggregated into TLD level
  - $H(D^{(1)}|D^{(2)})$  : Conditional SLD entropies of queries aggregated into TLD
  - $H(D^{(0)}|D^{(1)})$  : Conditional FQDN entropies of queries aggregated into SLD

$$H(D^{(0)}) = H(D^{(2)}) + H(D^{(1)}|D^{(2)}) + H(D^{(0)}|D^{(1)})$$

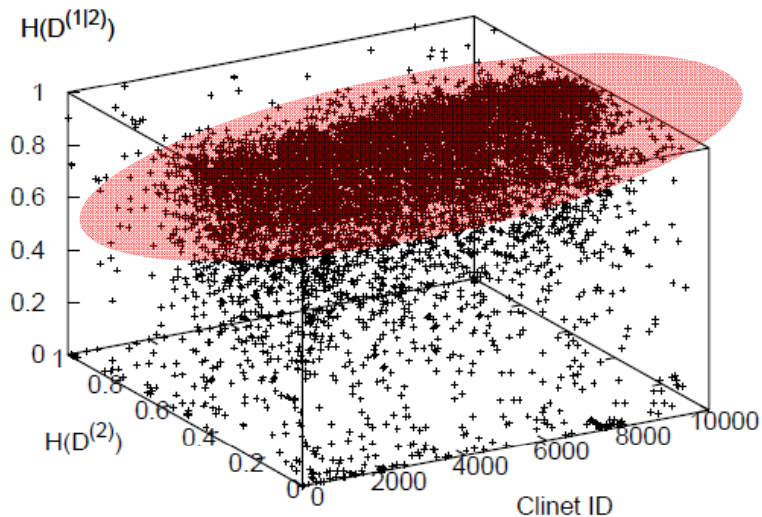


Can simultaneously capture dispersions in terms of FQDN/SLD/TLD

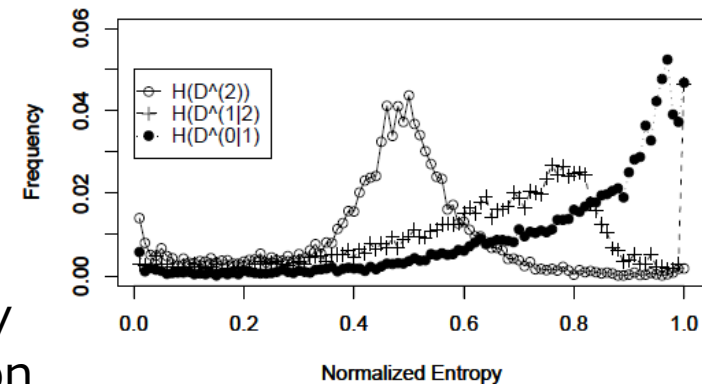


# Experimental results

- Calculate hierarchical aggregate entropies of top 10,000 heavy clients for DNS traffic monitored at DNS caching servers
  - Entropies from normal clients concentrated in a specific region
- ⇒ Clients whose entropies are out of the region can be

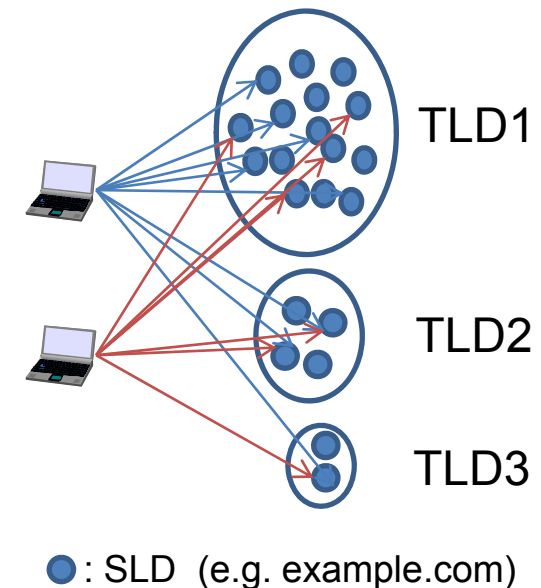
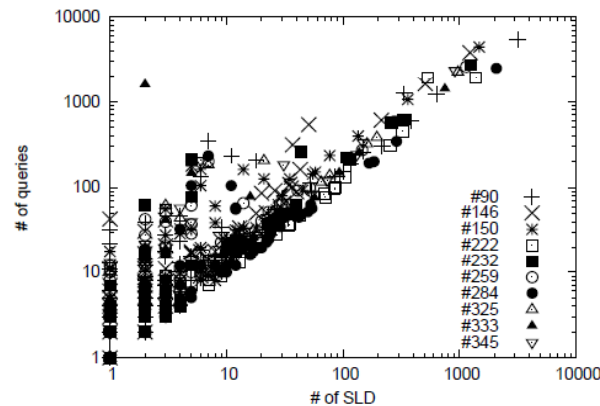
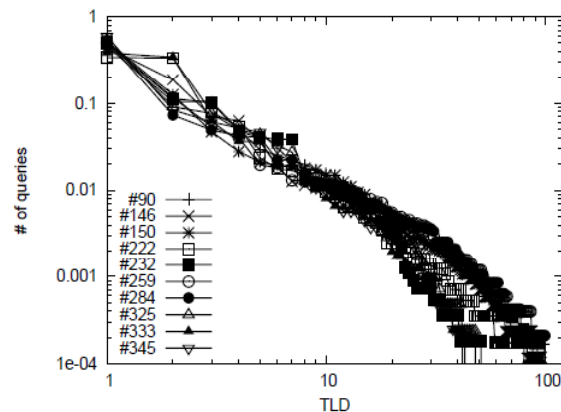


Frequency distribution



# Experimental results

- Why entropies from normal clients concentrated in a specific region?
- Investigation of top10 normal clients
  - Query distribution among TLDs: almost same Zipfian distribution
  - # of SLD in TLDs vs # of queries for the TLDs: almost linear
    - Large TLD attracts large number of queries (gravity model)



# Experimental results

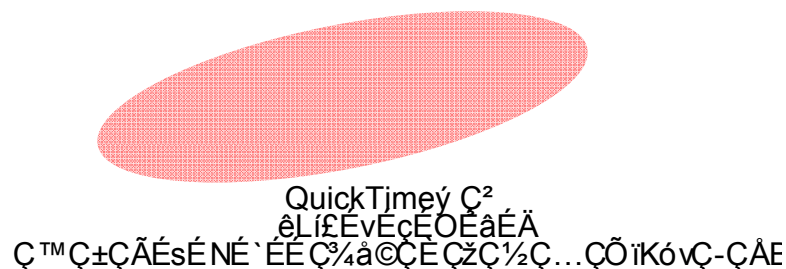
---

- Classifying clients that sent queries more than 1 qps by their hierarchical aggregate entropies
- Comparing to eyeballing classification (legitimate, mail sender, repeater, scanner, log analyzer)
  - ⇒ 80% accuracy
- Use of TLD level entropy or FQDN entropy
  - ⇒ 50-70% accuracy

Type	# of clients	$H^{(2)}, H^{(1 2)}, H^{(0 2)}$ (%)	$H^{(2)}$ (%)	$H^{(0)}$ (%)
Legitimate	114	81.6	86.8	49.1
Mail sender	50	84.0	88.0	78.0
Repeater	186	1.1	24.2	12.9
Scanner	8	12.5	14.3	85.7
Log analyzer	46	2.1	10.9	87.0

# PTR queries

- Hierarchical aggregation in TLD, SLD level cannot capture dispersion of PTR queries
  - 1.0.168.192.in-addr.arpa -> TLD: arpa, SLD: in-addr.arpa
  - Cannot distinguish between log-analyzer and scanner
- Apply hierarchical aggregation for IP address part!!
  - Entropies of dispersion in first octet, first+second octet...
  - Shows concentration to a specific region that reflects distribution of source IP addresses in IPv4 address spaces



## Conclusion

---

- Propose the use of hierarchical aggregate entropies to classify DNS heavy clients
- Can capture spatial dispersion of queries among domain name spaces
- Entropies from normal clients concentrated in a specific region
- Experimental results show that the proposed method achieve 10-20 % improvement in classification accuracy