Identifying DNS Scanners

Master Thesis





Topic

- ZDNS¹ and MassDNS² were published around 2016
 - Allow resolution of millions of domain names per minute
- OpenINTEL was not
 - "[...] if lots of researchers were to set up similar infrastructures this would have a significant and possibly disruptive impact on the Internet." [3]
- Does availability of those tools really pose a risk to DNS infrastructure?

Research Questions

- Can scans be found in the requests to authoritative ccTLD name servers?
- How can they be distinguished from normal traffic?
- Who performs scans and for what purpose?
- What portion of traffic is due to scans?
- Does availability suffer?

Research Questions

- Can scans be found in the requests to authoritative ccTLD name servers? Yes
- How can they be distinguished from normal traffic? Many patterns
- Who performs scans and for what purpose? Diverse groups, diverse purposes
- What portion of traffic is due to scans? Approx. 30 %
- Does availability suffer? No

Relevance

- Little research/analysis is done on the composition of DNS traffic, especially at TLD level
 - IP/Port scanning, Root servers, Bogus traffic, Performing scans, Resolver classification (e.g. monitoring or public resolvers)
- Reasons for scanning: domaining, academics, protection of trademarks, monitoring, finding vulnerabilities, web scraping, bulk email sending, and more!
- Results could help...
 - understand incoming traffic
 - substantiate best practices and protection mechanisms, if necessary
 - sanitize data (DNS-based popularity)



Data

- Individual requests recorded by all .nl AuthNS Anycast sites
- Processed by SIDN's Entrada¹
- Stored in Apache Hadoop for querying with the Spark query engine
- Around 5 billion queries from 1 million different IP addresses each day

root |-- id: integer (nullable = true) |-- time: long (nullable = true) |-- qname: string (nullable = true) |-- ttl: integer (nullable = true) |-- ipv: integer (nullable = true) |-- prot: integer (nullable = true) |-- src: string (nullable = true) |-- srcp: integer (nullable = true) |-- dst: string (nullable = true) |-- dstp: integer (nullable = true) |-- aa: boolean (nullable = true) |-- tc: boolean (nullable = true) |-- z: boolean (nullable = true) |-- rcode: integer (nullable = true) |-- qtype: integer (nullable = true) |-- country: string (nullable = true) |-- asn: string (nullable = true) |-- labels: integer (nullable = true) |-- proc_time: integer (nullable = true) |-- server_location: string (nullable = true) |-- pub_resolver: string (nullable = true) (60 columns in total)

Methodology

1. Manual work:

- 1. Understand and examine the data
- 2. Find out how to classify a scan as a human
- 3. Collect some ground-truth data
- 2. Machine Learning:
 - 1. Implement features describing resolver behavior
 - 2. Apply a clustering algorithm
 - 3. Evaluate
- 3. Use to answer research questions

Results

Results - 1

- More than 67 scanning operations identified and confirmed
 - 714 different sources
 - 437 million queries per day (10.3 % of all traffic)
 - Just part of the total scan traffic
- Scans show distinct behavior
 - Sometimes more obvious, sometimes less obvious
 - Most relevant:

Query distribution over time, percentage of NXDOMAIN answers, alphabetical ordering, question types

unusual

Results – 2

expected



Constant traffic, more during the day



unusual

Results - 2

expected

Queried names by time ordered lexicographically

A SRV CNAME TXT

No relationship between time, domain name and query type







unusual Results - 2 Contraction of the local sector of the local s way way way The structure of the state of the FITI Destanta 1. Salari Latari Carta Salari Salari Salari Carta Salari Salari Carta Salari Salari Carta Salari Salari Carta Sa Salari Carta Salari Salari Carta Salari Cart Salari Carta C. Manufer and C. Martin, M. S. Martin, J. Lewis, M. S. W. Waldhall, and Yoshinkama, Annual Sciences and Sciences of the Statistical Conference on Conference on Conference on Con-temporation of Conference on Conference on Conference on Con-temporation of Conference on Conference on Conference on Con-temporation of Conference on Co 1 - 1 Indetectments a torono valorebar a and a straight of the Land and the second and 1 - El televidoren estat E. Totelevidoren -F. El televidoren -F. El televidoren and an and constant provers, the product of the call of the and <u>والدي المحاطبة والمحمد في مريدة المحمول منها.</u> - E Minicipanita לי על איר אינטאר אין איר איר אינטאראנערע אין איז אינטארע אונערט אינטארע איר אינטארע אינטארע איר אינטארע אינטאר in the second states in the second states of the i Geological grant i i NCOMIN No arciver NOCHAN No arciser . 1.322177.2 \$1 States and the of the states of the A State of the sta ____ 4 1000 Construction of the 5 25 25 新史。 Let A Barrier and A Barrier an A DELEY AND DEPARTURE B) Construction and the second sec ale and the state of the . A. . ÷4. 5 to satisfy the second second TARGET AND ADDRESS OF ADDRESS A STATE OF A STATE And Andrew States STREET, ST A 101 Notice of the second **5** 05 10110 DI10 100 and the second second Chine (1986) Chinese and an an an antipathon of the second s CHERONOM > in company And Constant Designed Constant Designed Constant Designed PARTICLES AND A DESCRIPTION PROCESSION A 41705 86 80 99/ 99/ 99/ 95 83 99/ 157 737 737

ta anasar and and and the sea and and and

a share a far and the part of a share the set of the

ware ware ware ware and and and and and

- How can we make sure we have found everything?
- How can we choose and group resolvers more systematically?

- How can we make sure we have found everything?
- How can we choose and group resolvers more systematically?
- 1. Create features
- 2. Calculate for each resolver
- 3. Clustering
- → Similarity measure, grouping

Features

- Different flags (truncation, recursion desired, ...)
- Basic statistics (query count, distinct percentage, country)
- Percentage of queries with each

response code, operation code, query type, query class, number of labels, UDP/TCP, target server, starting symbol for the domain name, IPv4/IPv6, punycode

Statistics about

query timestamps, domain name length, IDs, source port, EDNS UDP, packet TTL

Custom Features

- Distribution of queries among all names queried
- Histogram of queries in time
- Max/Min number of queries sent within 60 seconds
- Statistics about:
 - Repeated queries
 - Time between consecutive queries
- Intersection of names with name lists
 - Common Crawl
 - Certificate Transparency
 - Registered names from 1 year ago
- DNS2Vec features (learned from domain names)

Results - Clustering

- High accuracy when distinguishing scans and nonscans on manually curated dataset (97 %)
- Easily able to find typical domaining or similarly obvious scans
- Is explainable

qtype_DS_percentage -	1.4	-0.99		-0.97	0.39	-0.95	1.4	-0.92	1.2		1.3	-0.98	1.2	-0.73	-0.38	-0.96	-0.79
rcode_minus_1_percentage -			-0.27	0.48	2.5	-0.049	-0.044	-0.23	-0.25		0.017	-1.1	-0.36	-1.1	-0.11	-1.1	1.4
rcode_0_percentage -	-0.64	0.38		-0.96	-1.2	1.2	-0.83	-0.75	-1.1	-0.81	-0.37	1.5	1.3	-0.22	1.9	-0.049	-0.39
qtype_NS_percentage -		-1.1	-0.31	-1.1		-0.91		-0.88			-0.49	-0.49	-0.38	-0.061	-0.27		0.27
distinct_percentage -	0.25	0.35	-0.51			0.19	0.35		0.39	0.34	0.24	0.36	-0.0069	0.19	-3.8		0.19
cd_percentage -	-0.52	-0.96	-0.46	-0.94			-0.74	1.3	-0.6		1.3	-0.97	1.6	-0.6		-1.1	-0.92
qtype_TXT_percentage -	-0.17		-0.99	-1.5		-1.3	0.49	-0.72			-0.18		-0.75	0.15	-1.4		-0.014
qtype_HTTPS_percentage -		-0.78	-0.6	-0.82	0.081	-0.31	-0.18	0.26	-0.65	0.11	-0.32	-0.77	-0.43	2.2	-0.65	-0.84	
qtype_AAAA_percentage -			-0.73	-2.1	0.42	-0.021		0.49			0.15		-0.12		-0.98	-2.4	
qtype_CNAME_percentage -	-0.23	-0.98	-0.081	-0.86		-0.88		-0.61	-0.19	2.1	-0.27	-0.93	-0.38		-0.49	-0.93	
qtype_SOA_percentage -	-0.015	-0.66	-0.44	-0.66		-0.56		-0.5	3		-0.44	-0.56	-0.56	-0.26	-0.56	-0.66	-0.31
rcode_3_percentage -			-0.85			-1.2					0.41	-1.1	-1.2	0.36		0.21	0.29
qtype_SRV_percentage -	-0.3	1.8	-0.68	-0.69	-0.58	-0.59	-0.65	-0.45	2.3	-0.51	-0.62		-0.6		-0.65	-0.69	
qtype_A_percentage -	-0.41	-0.43	-0.43	3.6	-0.81		-0.4	-0.067	-0.41	-0.49	-0.32	-0.39	-0.32	-0.33		0.34	-0.33
qtype_DNSKEY_percentage -	0.9	-1.1	0.36	-1.1	0.89	-0.96	1.8	-0.93	0.0051	0.48			1.6	-0.6	0.48	-1.1	-0.73
_over_short_timeframe_difference -	-1.1	-0.76	-0.084	0.26	0.19	0.006	0.5	-0.065		0.1	-0.36	-0.81	-0.54	-0.72	-0.069	3.2	-0.82
repeat_percentage -	-0.26	-0.34	0.49	-0.61	-0.57	-0.19	-0.36	-0.19	-0.42	-0.33	-0.25	-0.34	-0.061	-0.19	3.8	0.049	-0.18
name_starts_with_x_percentage -	1	0.025	-2.2	-0.092		-0.11		0.49	-0.14	0.26	0.092	-0.092	-0.22		-2.6		
name_starts_with_1_percentage -	0.4	-0.12	-1.8				0.034	0.15			0.15	0.24	0.11	0.048	-1.9	-1.9	-0.069
qtype_CAA_percentage -	1.5	-0.65	-0.29	-0.65		-0.64	2.5	-0.63		0.073	-0.21	-0.65	-0.55	-0.65	-0.49	-0.65	-0.65
rd_percentage -	-0.4	-0.4	-0.33	3.7	-0.064	-0.26	-0.4		-0.4	-0.16	-0.35	-0.33	-0.36	-0.4	0.041	-0.28	-0.39
name_starts_with_j_percentage -	0.46	0.38	-2.2	0.75	0.45	0.22		0.13		0.43	0.44	0.34	0.17	0.085	-2.9	-0.26	0.19
name_starts_with_q_percentage -	0.096		-2.4	0.12	0.33	0.054	0.24		0.26	0.35	0.19	0.21	-0.19		-2.7		
name_starts_with_k_percentage -		0.48	-2.1			0.32	0.43	0.12		0.34	0.35	0.46	0.29	0.1	-3.1	0.39	0.12
name_starts_with_u_percentage -	0.15		-2.3	0.41	0.25	0.21		0.33	0.41	0.39	0.22	0.49	0.045		-2.9	0.48	
name_starts_with_2_percentage -	-0.019	0.34	-2.4	0.11	0.097		-0.28			0.24	0.097		0.095	0.35	-2.6		0.26
name_starts_with_y_percentage -	0.11	0.32	-2.4		0.48	0.26		0.19		0.26	0.31	0.29	0.086		-2.7		-0.0028
name_starts_with_f_percentage -	0.38	0.36	-2.3		0.29	0.35		0.4	0.36	0.28	0.36	0.35	0.33	0.39	-3	0.39	0.34
name_starts_with_r_percentage -			-2.1	0.46	0.35			0.24	0.49	0.35	0.36	0.47	0.28	0.39	-3.1	-0.26	0.44
tc_percentage -	-0.49	-0.88	-0.53	0.44	3.1	-0.71		0.15		0.37	-0.44	-0.83	-0.73	-0.62	-0.2	0.29	-0.72
tcp_percentage -	-0.49	-0.88	-0.54	0.45	3.1	-0.73		0.11			-0.43	-0.82	-0.71	-0.6	-0.28	0.25	-0.71
name_starts_with_p_percentage -	0.37		-1.8	0.39	0.31	0.32	0.36	0.23	0.4		0.34	0.41		0.26	-3.3	0.42	0.27
name_starts_with_l_percentage -	0.43	0.46	-2.1		0.38	0.36	0.45	0.26		0.35	0.36	0.46	0.33	0.14	-3.1	-0.096	0.17
name_starts_with_n_percentage -	0.32	0.35	-2.2	0.18	0.25		0.21	0.41	0.24	0.33	0.31	0.34	0.31		-3		0.48
name_starts_with_z_percentage -			-2.3		0.23			0.29	0.28	0.41	0.31		0.39	0.24	-2.8	-0.35	
name_starts_with_e_percentage -			-2.2	0.42	0.29	0.38	0.33	0.34	0.41		0.36	0.48	0.31	0.44	-3	-0.24	0.43
punycode_percentage -	2.2	-0.52	-0.46	-0.092		-0.52	2.6	-0.4	-0.36	-0.23	-0.42	-0.52	-0.52	-0.49	-0.52	-0.49	-0.49
name_starts_with_v_percentage -	0.36	0.5	-1.9	0.36	0.29	0.3	0.3	0.21	0.38	0.33	0.32	0.46	0.24	0.34	-3.2		0.36
name_starts_with_3_percentage -		0.14	-1.9	0.48		0.085	-0.18	-0.02	1.9	0.21	-0.074	0.21	-0.33	-0.32	-2	1.8	-0.67
name_starts_with_i_percentage -		0.41	-2.1	0.26	0.26	0.33	0.3	0.32	0.3	0.33	0.28	0.35	0.25	0.44	-3.1	0.63	0.4
percentage	0.54	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
		-	-	-					-				-	-		-	

Findings

- Broad definition necessary, including bulk mail sending and monitoring
- Much of traffic unclear (shows signs of scan, but not definite)
 - Some scans are also *very* obvious
- About 30 % of traffic from scans (± 10 %)
- AuthNS manage just fine, not affected by extremely large scans, even

Groups

- 10 % of traffic *definitely* scanning
- Majority on one day from single subdomain scanning operation
- Most scans from networks of hosting providers
- Public resolvers: Most contain subdomain scanning or small scans





Interesting Findings

- Some resolvers have little similarity to others
 - Others have a clear group of similar ones
- A significant number of resolvers simply ignore NS3 or NS4
- Clustering well-able to find public resolvers with scans
- Few features necessary
 - Can be calculated from just traffic





Other patterns

- Duplicate queries (why?)
- Repetitions within a short time (poor caching?)
- Querying name servers far too often
- Insufficient caching, or no negative caching for NX 2LDs
- Often either NS3 or NS4 completely ignored (out of 3 servers)

Lessons

- Scanning is highly prevalent
- Subdomain scanning can be seen in TLD traffic
 - Mostly done through public resolvers

Limitations and Challenges

- Only considering the NL zone, results expected to generalize
- Broad definition of scanning: Includes domaining, monitoring, and basically everything causing many contiguous DNS requests
- Analysis done on two days of data (5 billion queries each)
- Wide range of behavior that is not easy to explain/grasp/classify

- Scans might not be visible when looking at individual IP addresses
 - Scans distributed among hundreds, even different organizations, do exist!
 - Particularly large numbers of scans probably do not
- Public resolvers difficult to analyze
- Small chance of scans escaping detection (false negatives) due to feature count