Needles in a Quadrillion-Straw Haystack

DNS-OARC Fall 2013

Sam Bretheim, Paul O'Leary
firstname.lastname@nominum.com

Introduction

Nominum works with a lot of data

- Subset of a trillion-query-per-day flow
- Terabytes of new data per day
- Hundreds of terabytes stored
- ~400 feeds in ~60 different formats

Here are some interesting things we've found

Why do we need data?

- Security services
  - Nominum Network Protection Service
    - DNS-resolver-based malware C&C and DoS detection and remediation
  - Nominum Subscriber Safety
    - Network-based malware and phishing protection
- Client support (troubleshooting)
- Product performance tuning

Why not just buy some feeds?

- We tried!
- Vendor data is a critical asset, but...
- There was no single feed vendor that sufficed
  - Unacceptably high false-positive rates (ours needs to be exceptionally low), or unacceptably low coverage
  - Limited relevance to real DNS usage in ISP networks
- Combining vendor feeds still didn't do the job
- ... So we need to do our own research

We use data, but we're not the NSA

- We use anonymized client IP addresses
- We have no IP-to-identity mappings anyway
- We whitelist, filter, and purge aggressively
- We do not disclose *anything* that isn't a common trend on multiple ISP networks

How do we handle all the data?

- Query data collected using Nominum RTV
- Custom crawling and ETL frameworks
- Summary data imported into Postgres
- Filtered query data imported into Hadoop
- Map/reduce jobs (Pig, Java, Python) for summarization, further filtering, and heuristic-based discovery
- Meta-learning and classification using Weka and scikit-learn
- Graph-based NoSQL data warehouse

Filtering

- Whitelisting is hard!
- Most queries are not for Web-site/email names
  - Web backend stuff (images, CSS/JS, ads, ...)
  - Application API calls
  - OCSP, CRLs
  - Software updates (OSes, apps, antivirus, ...)
  - RBLs
  - Reverse DNS
  - ...

Filtering

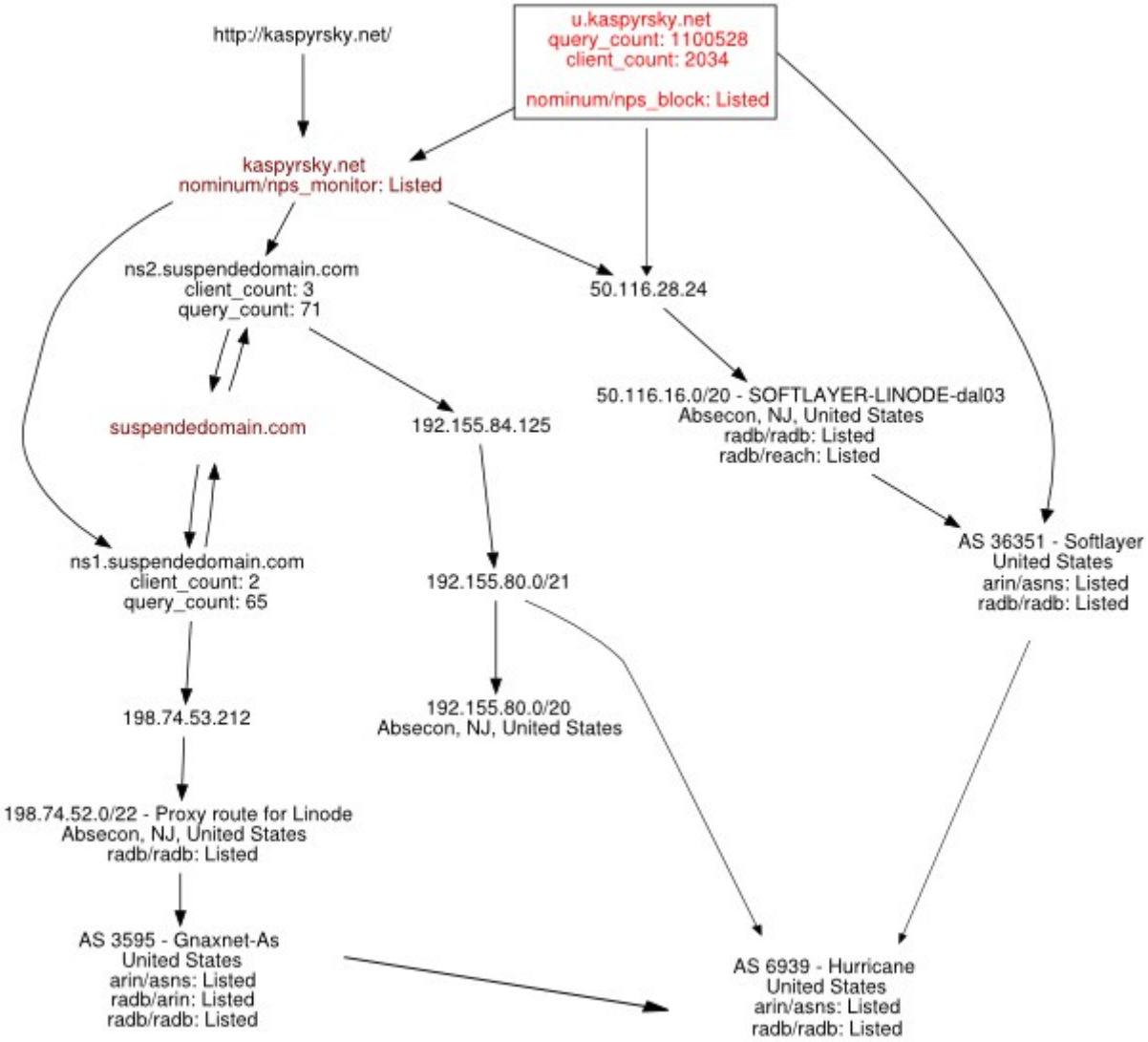
- Internet traffic changes constantly
- Nobody makes a decent feed of this machine-generated traffic
- None of it is self-documenting and a lot of it is deliberately obscured
- Malware authors usually forget to set the evil bit

Filtering

- Filter out:
  - Whitelisted things
  - Things that we already block
  - Noise
    - Names that aren't within a valid TLD (typos, wpad, isatap, _tcp, eth0, br0, lo, ...)
    - Other queries that return errors (NXDOMAINs, ...)

Classification

- Look for features that correlate with known malware activity
  - Name-structure features
    - Does the name match known malware-related patterns?
    - Does the name look machine-generated?
    - String features (label count, ...)
    - ...
  - Server-side features
    - Nameserver reputation
    - Hosting resource reputation (CNAME, IP, ASN, ...)
    - ...


http://kaspyrsky.net/
u.kaspyrsky.net
query_count: 1100528
client_count: 2034
nominum/nps_block: Listed
kaspyrsky.net
nominum/nps_monitor: Listed
ns2.suspendedomain.com
client_count: 3
query_count: 71
50.116.28.24
suspendedomain.com
192.155.84.125
50.116.16.0/20 - SOFTLAYER-LINODE-dal03
Absecon, NJ, United States
radb/radb: Listed
radb/reach: Listed
AS 36351 - Softlayer
United States
arin/asns: Listed
radb/radb: Listed
ns1.suspendedomain.com
client_count: 2
query_count: 65
192.155.80.0/21
198.74.53.212
192.155.80.0/20
Absecon, NJ, United States
198.74.52.0/22 - Proxy route for Linode
Absecon, NJ, United States
radb/radb: Listed
AS 3595 - Gnaxnet-As
United States
arin/asns: Listed
radb/arin: Listed
radb/radb: Listed
AS 6939 - Hurricane
United States
arin/asns: Listed
radb/radb: Listed

Classification

- Look for features that correlate with known malware activity
  - Client-side features
    - Client reputation
    - Queries per client
    - Periodicity
    - Query header values
    - ..
  - Feed-based features
    - Is it on feed X?
    - What's the value of field A for this name's entry on feed X?

Classification

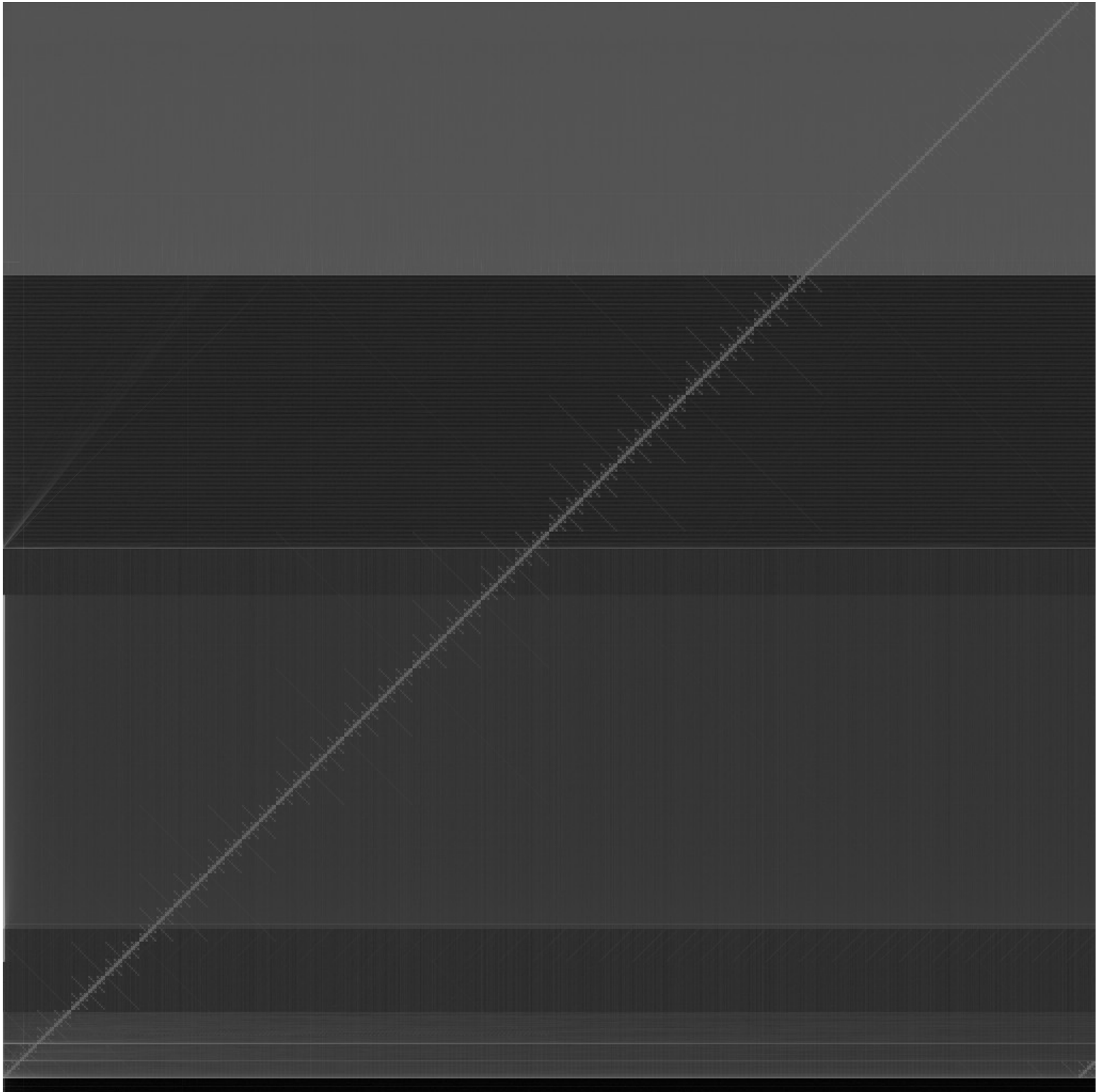
- Feed extracted features into heuristic and ML algorithms
  - Manually determined thresholds/rules are sometimes as useful as machine learning, but *much* faster
  - Use meta-learning to select and parameterize ML classifiers
- Continuously validate output using human review, and feed the results back into classification algorithms

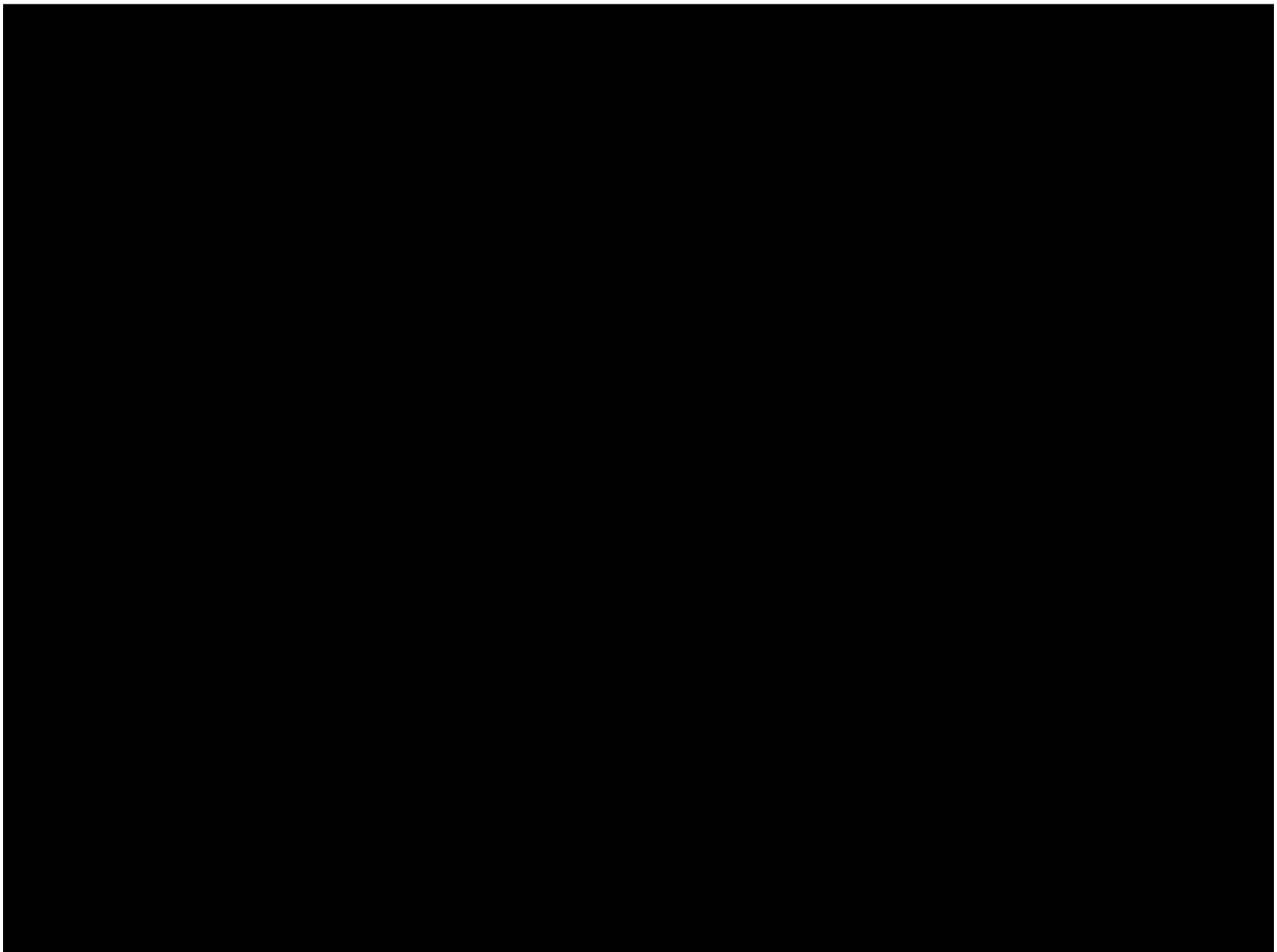
Results

- Many newly discovered malware domain names
- Productive use of feeds with high FP rates

Weird Stuff

- In order to filter out the noise, you need to understand it
- Some of the noise is interesting

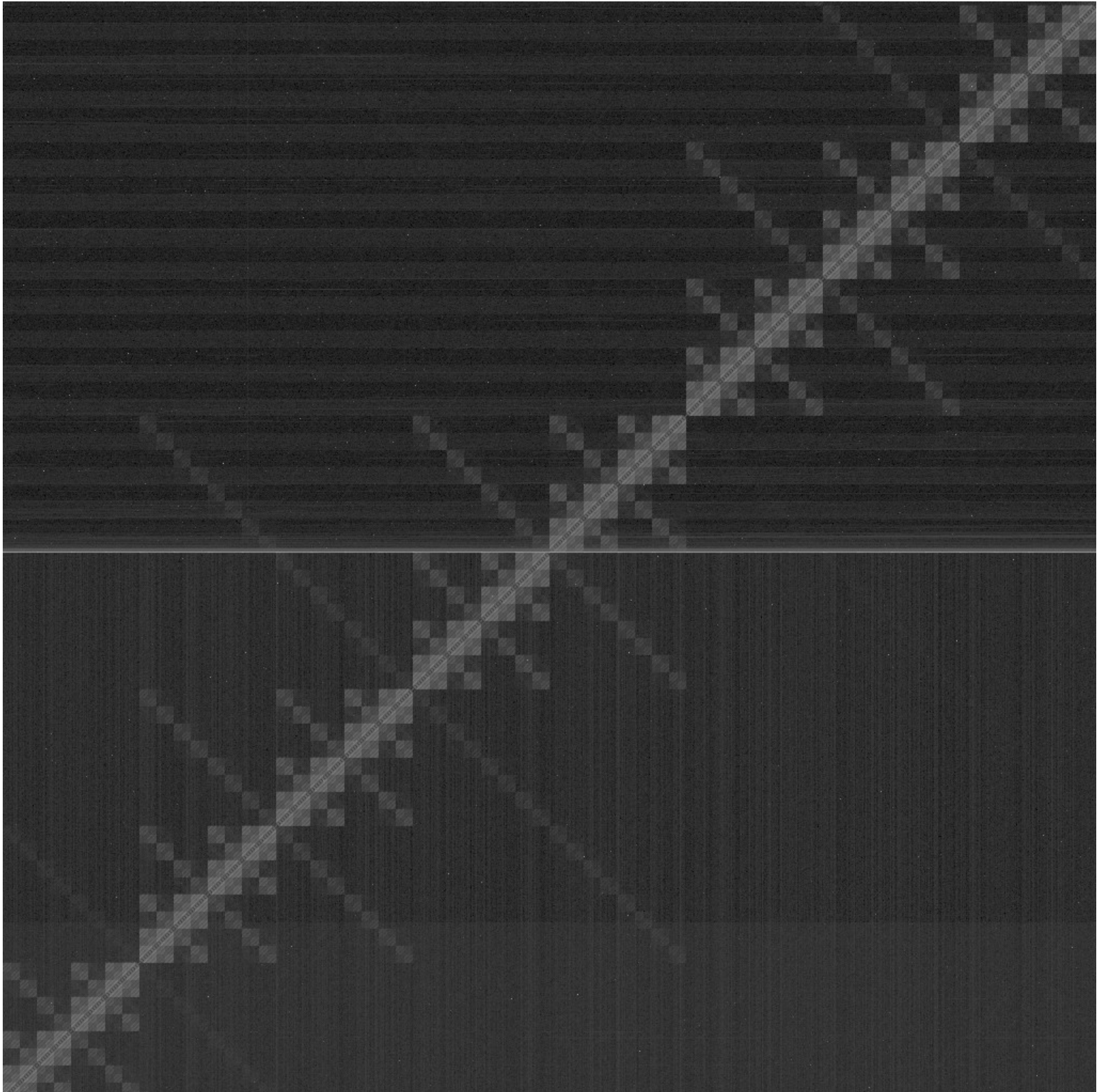




65535

IANA recommended port range: 49152-65535
Uniform, random port and QID selection

49152



Weird Stuff: RNG failures

- Mostly coming from mobile/embedded devices
- Mostly SRV, PTR, and other non-libc-supported stuff
- Different RNG/PRNG failure modes exposed by different applications
  - Android libc was improved - CVE-2012-2808
  - Third-party DNS libraries are still a problem

See also: Kai Michaelis, Christopher Meyer, Jörg Schwenk, “Randomly Failed! The State of Randomness in Current Java Implementations”, CT-RSA 2013

Weird Stuff: Bugs

- Lots of application programming errors
  - http://foo.com (as a domain name)
  - 8.8.8.8 (as a domain name)
  - <doctype public html ... (as a domain name)
- DNS implementation errors
  - Case-sensitivity-related bugs in homegrown/embedded DNS implementations

Weird Stuff: Deliberate NXDOMAINs

- 10-letter [pseudo]random strings
  - cowcggjwqr
  - dmgodynugx
  - dqdoiusymr
  - ebctragwlk
  - eibyqxkdny
  - enkusribxn
  - ewfjydcueh
  - fbjmpepaar
  - fdfrwmzugt
  - fdfrwmzugt
  - Chrome NXDOMAIN redirection detection

Weird Stuff: Paul's Mystery Patterns

Type A	Type B	Type C	Type D
[a-z1-9:]{9}\.[0-9]{2}[a-z]{2}	[a-z1-9:]{9}\.[a-z][0-9][a-z][0-9]{2}[a-z]{2}[0-9]	[a-z1-9:]{9}\.[a-z][0-9]{2}[a-z][0-9][a-z][0-9][a-z]	_[0-9]{3}_[0-9]{2}_[0-9]
shjp2wi3f.75xc	p4zvsmddo.p4h77ej9	6f21ok7vr.s59r6n1x	_749_54_1
t8vuaxgxw.85xk	p65w7i4bp.s6d14qn1	6i561mf:f.r65q5u4t	_717_59_6
tkbwuajqm.03ls	p6p6csy11.d6n93fm7	6juuuieqg.u64y5n8t	_070_94_0
ws65q1go9.58gx	pf2h73jqi.j8x74ya6	6rs9yng6d.s84i6y7n	_895_19_2
xb2o36x5v.28pe	pib3gsuhk.c2a50qq1	75sjfpprj.d14y3z3x	_407_54_3
y6it46d5q.48lp	pj:pr476e.w1p87ml6	7be7g516e.x19v9m4f	_433_33_2
yltcqwcr9.36ib	pky6iuyne.m0k11oa6	7h7844cbz.x78y8a3b	_162_75_4
2v8739o7l.26tl	plv4yfdle.b9w93kl3	9emzo9bdz.m92l7c2m	_007_52_2
3sxs19zx1.47hy	ppc85k8qs.f9e71ug4	9m15wcw3j.n09n2z2t	_849_23_8
4r2zubv8w.35zy	qg2vdbswz.p1j84sd4	9v1x392hv.j89p4t4z	_404_49_8
5c7r8kn4s.27dd	qn2xolftu.o1w63it1	a5c3rtoyz.d64n8n3l	_433_51_7
dokhb2ozb.43cv	qusq1ru7y.e6l66at0	annsa8gwf.r58l4b1d	_891_43_6
e71h8d1t8.56rv	rgnyjx1ox.l5l88fn3	b77vze24m.e49b8i3r	_399_31_3
et9vxq45w.12hn	ruaewf8ay.o2x36jo2	bvoi8m2iw.s29l6t2i	_761_72_6
iynag6pdq.53jl	s9jj6ghb4.e1h76ca6	c5mphdar8.f38u5w2x	_524_14_7
ken6zg6hx.18rb	s:22ug73m.q8i66dc6	czhbq3x6z.y10e0t8j	_620_74_8
kkcpwah95.49li	sexqw51x:.p4p85kk5	d271rts57.z11b7v6k	_017_69_8
ownudwxvm.44ll	skbbupd4m.r4o26db0	dacd7txvx.w89g2l5b	_251_96_4
r8s4wpghq.68by	ssrgyygdn.l8i43yb2	ddk375doe.e37y7w1r	_486_04_4
t85jqrsdq.97bu	stdlrqv3i.s5c99ac7	djdmv556p.q35s0h0z	_247_40_2

Paul's Mystery Patterns: Properties

- Found globally on about 0.5% of all client IP addresses (so expected to be millions worldwide, making billions of queries)
- Usually (but not close to always!) issued in sequence by each client, every two minutes
- Query names appear to be a function of time
- Most common sequence is types A-C-D
- A-B and B are also reasonably common
- Has been happening for years

Paul's Mystery Patterns: Hypotheses

- DDoS?
  - Query volume is too low
- Botnet C&C DGA?
  - All NXDOMAINs; would not actually work as C&C unless someone poisoned the resolver
- NXDOMAIN redirection detection?
  - Seems far too weird and complex
- Sending a signal to authoritative servers?
- Sending a signal to SIGINT?

Paul's Mystery Patterns: Malware?

- Queries correlate (but nowhere near 100%) with queries to the following names that resolve to 208.87.149.250:
  - beattriangle.com
  - queryscanone.com
  - beatboxtriangle.com
  - questscantwo.com
  - batbeathit.com
  - boxtaphit.com
  - basicscanone.com
- and this one that resolves to 208.87.149.246:
  - upgradeservererror.com
- All of these appear on malware blacklists

Questions? Answers?