OpenINTEL

digging in the DNS with an industrial-size digger :-)
(or: I queried 60% of the DNS, and I found this)

UNIVERSITY OF TWENTE.

DACS
Design and Analysis of Communication Systems

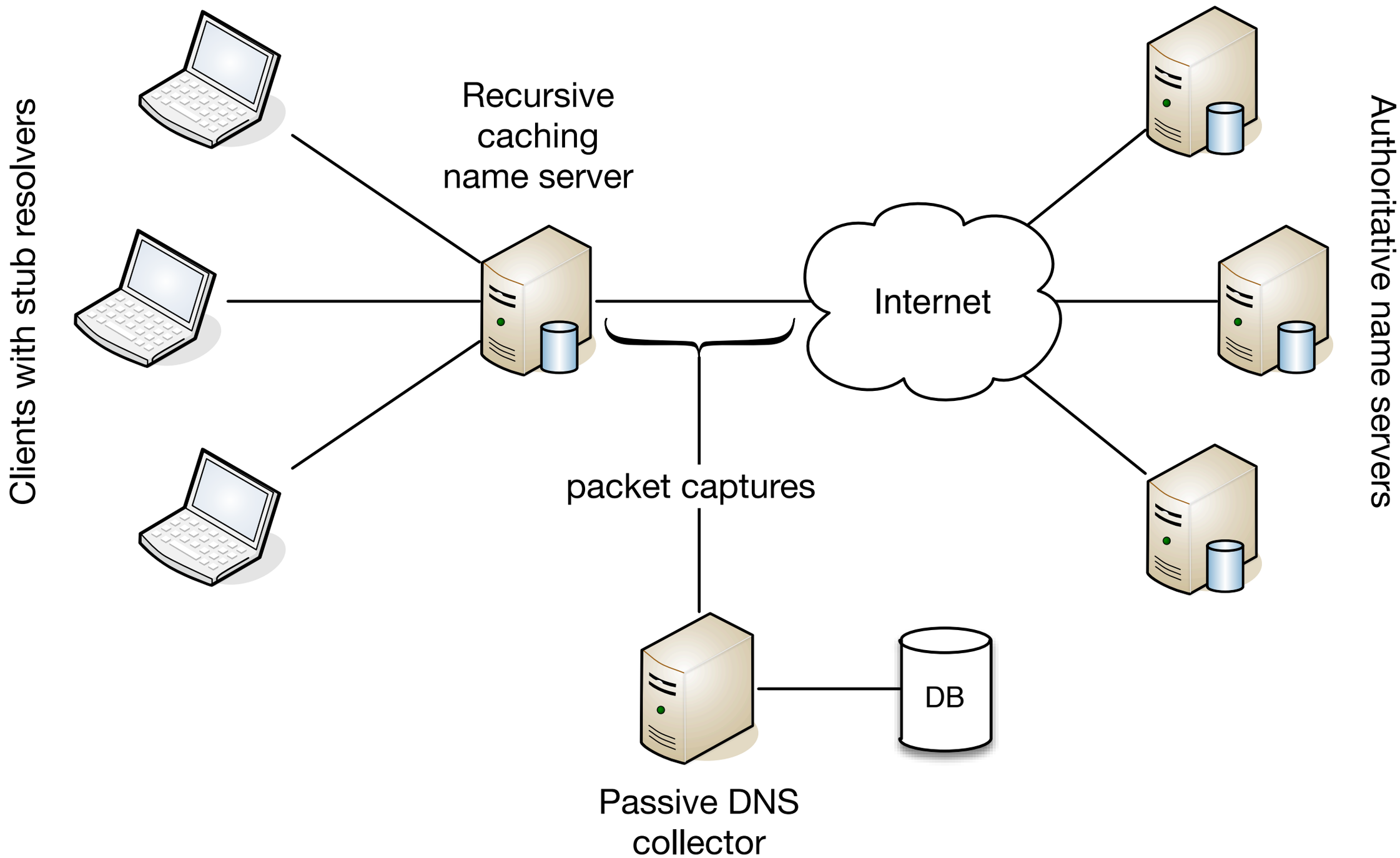
SIDN LABS

SURF NET

Why measure DNS?

- (Almost) **every networked service relies on DNS**
- DNS translates **human readable** names into **machine readable** information
  - e.g. IP addresses, but also: mail hosts, certificate information, …
- Measuring **what is in the DNS over time** provides information about the **evolution of the Internet**

Passive DNS


Clients with stub resolvers
Recursive
caching
name server
Internet
Authoritative name servers
packet captures
DB
Passive DNS
collector

Passive DNS

- pDNS suffers from **bias** that makes it unsuitable for reliably tracking DNS changes over time
- pDNS will only see data for domains that clients of the resolvers behind which pDNS data is collected are interested in
- This means that **pDNS will only see a domain** when it has been **used and observed at a sensor**
- pDNS gives **no control over the query frequency**, so the data is **unusable for e.g. time series**

Active DNS measurements

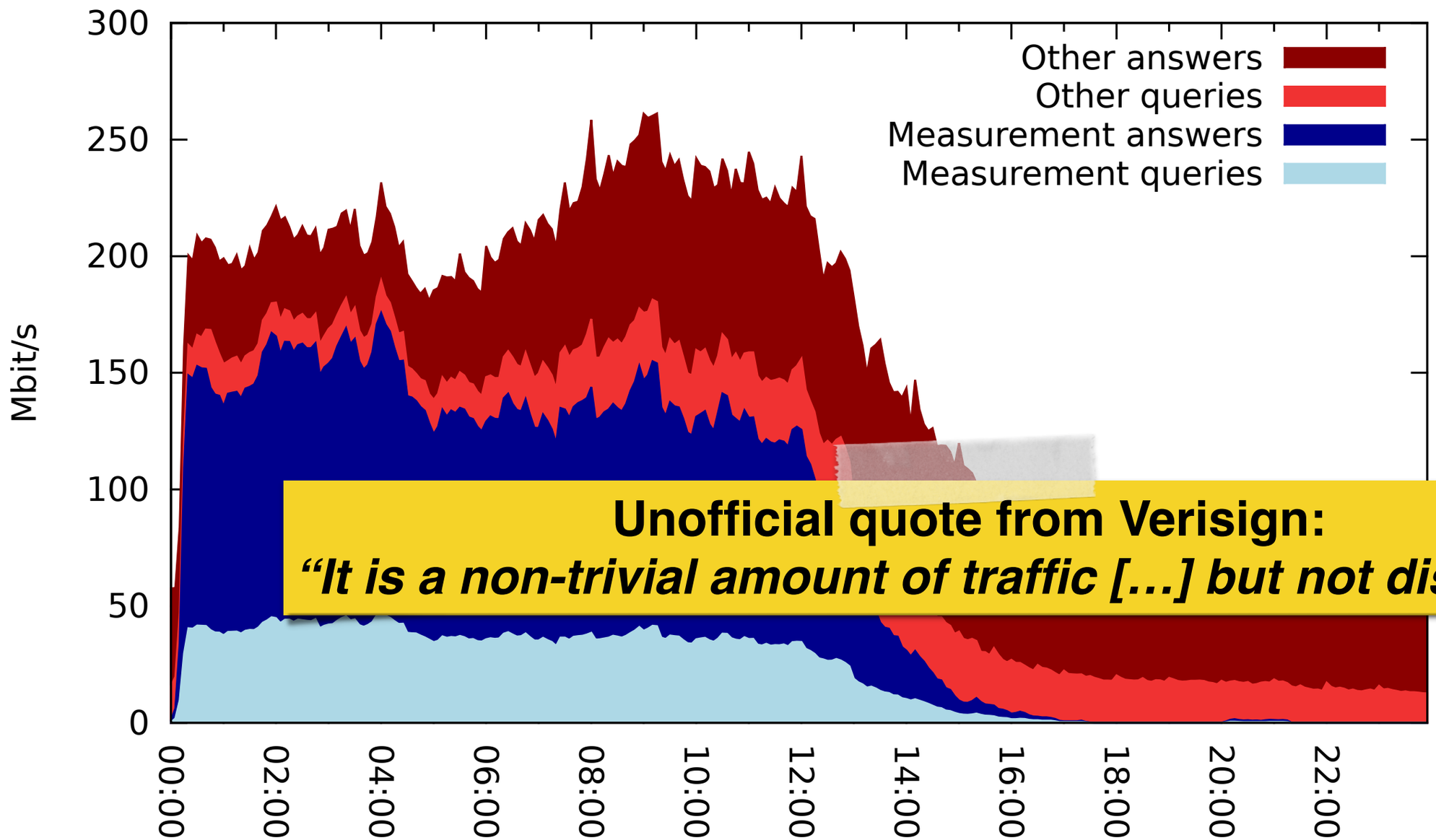
- We send a **comprehensive set of DNS queries for every name** in a TLD, **once per day**
- We do this at **scale**, our current measurement covers around **60% of the global namespace**:
  - .com, .net, .org, .info, .mobi, .gov
  - .nl, .se, .nu, .ca, .fi, .at, .dk, .ru, .рф, .us
  - ±1200 new gTLDs (e.g. .amsterdam, .frl, .xxx, …)
  - in total **almost 210M domain names**
- We need to **store and analyse** this data **efficiently**
- We **must not overburden the global DNS!**

What do we query and store?

- We ask for:
  - SOA
  - A, AAAA
    - (apex, and ‘www’)
  - NS
  - MX
  - TXT
  - CAA (new)
  - DS
    + CDS (new)
  - DNSKEY
    + CDNSKEY (new)
  - NSEC(3)
- We store:
  - All records in the *answer* section
  - CNAME expansions
  - DNSSEC signatures (RRSIG)
  - Metadata (Geo IP, AS)
- Separate “infrastructure” measurement
  - Collect A/AAAA for NS and MX names

Impact on the global DNS

- Our measurement is clearly visible in SURFnet's traffic flows:



Unofficial quote from Verisign:
"It is a non-trivial amount of traffic [...] but not disruptive"

Big data? Yes!

- Calling your research "big data" is all the rage
- So would our work qualify as big data?
- One **human genome** is about $\mathbf{3 \cdot 10^9}$ base pairs

- We collect **over** $\mathbf{2.2 \cdot 10^9}$ DNS records **per day**
- Since February 2015, we collected $\mathbf{2.3 \cdot 10^{12}}$ results (2.3 **trillion**) or **over 781 human genomes**

Big data? Yes!

- Calling your research "big data" is all the rage
- So would our work qualify as big data?
- One **human genome** is about $\mathbf{3 \cdot 10^9}$ base pairs

- We collect **over** $\mathbf{2.2 \cdot 10^9}$ DNS records **per day**
- Since February 2015, we collected $\mathbf{2.3 \cdot 10^{12}}$ results (2.3 **trillion**) or **over 781 human genomes**

Big data? Yes!

- Calling your research "big data" is all the rage
- So would our work qualify as big data?
- One **human genome** is about $\mathbf{3 \cdot 10^9}$ base pairs

- We collect **over** $\mathbf{2.2 \cdot 10^9}$ DNS records **per day**
- Since February 2015, we collected $\mathbf{2.3 \cdot 10^{12}}$ results (2.3 **trillion**) or **over 781 human genomes**

Big data? Use the right tools

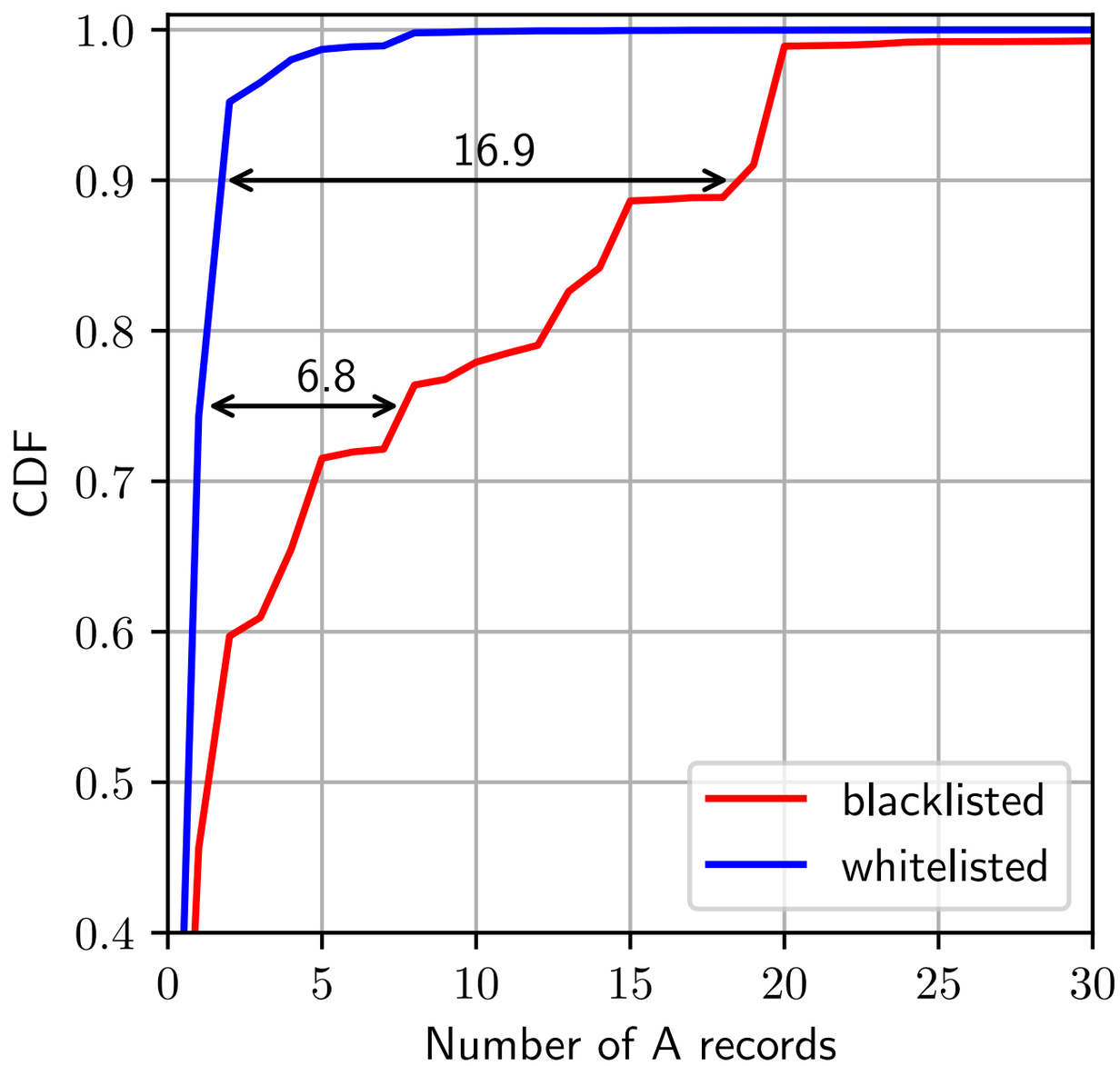
- Dedicated Hadoop cluster
- Latest & greatest tools for analysis; **Impala, Jupiter**

Example 1: Snowshoe spam

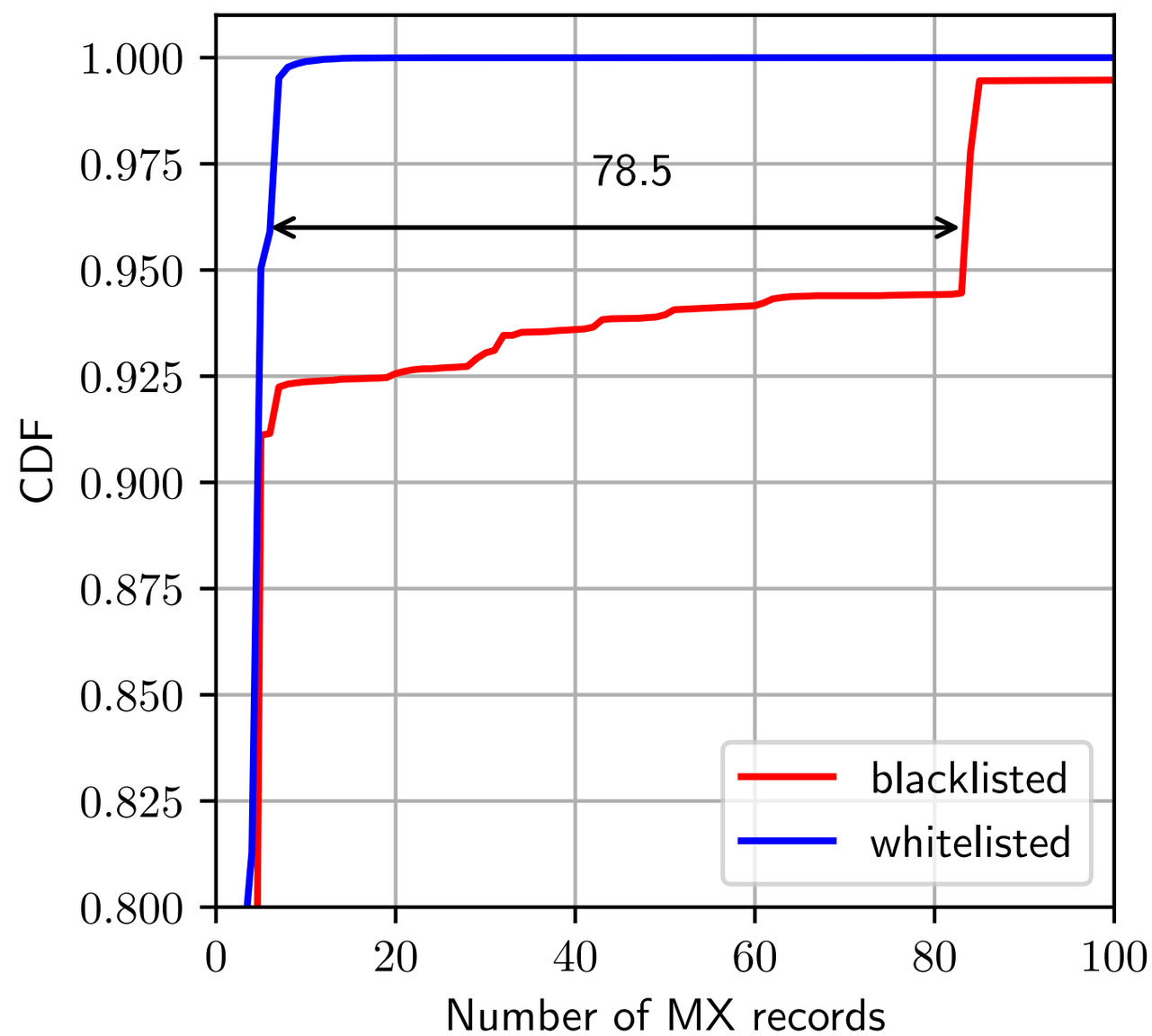
- Snowshoe spam is a form that is hard to filter out
- Spammers spread sending of spam across many IPs, in different prefixes and linked to different domains
- This makes it hard to blacklist “bad” IP blocks or domains
- Example pattern: many domains with e.g. 50 different IPv4 addresses linked to the name

Signatures for snowshoe spam

Anomalous #A records



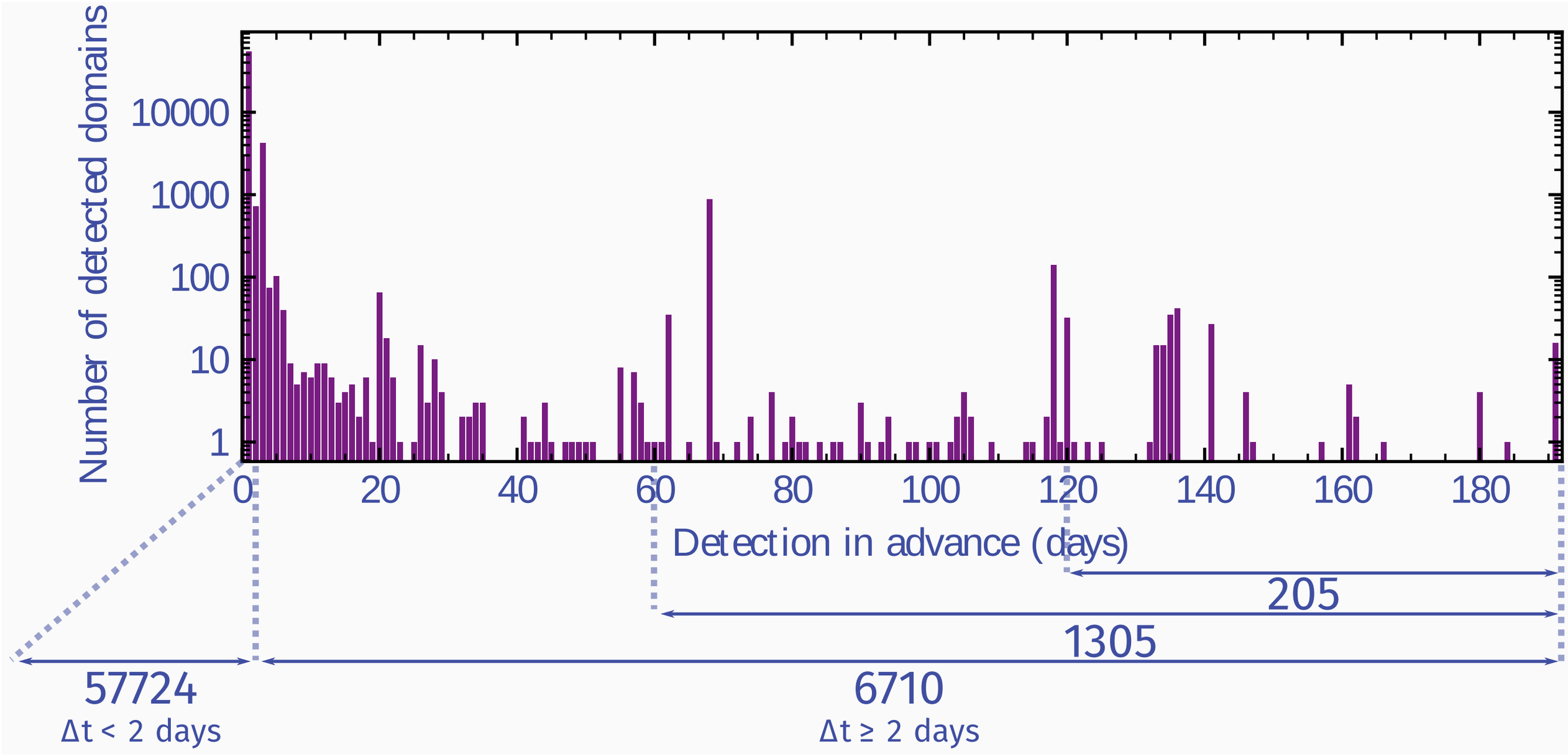
Anomalous #MX records



Snowshoe spam

- Project by an **M.Sc. student** (now a Ph.D. student in our group)
- Collaboration between university and SURFnet
- Used **real world mail filtering data** from SURFnet's **SURFmailfilter** service
- With **research** we **can improve real-world e-mail security** for SURFnet's constituency!

Significant improvement

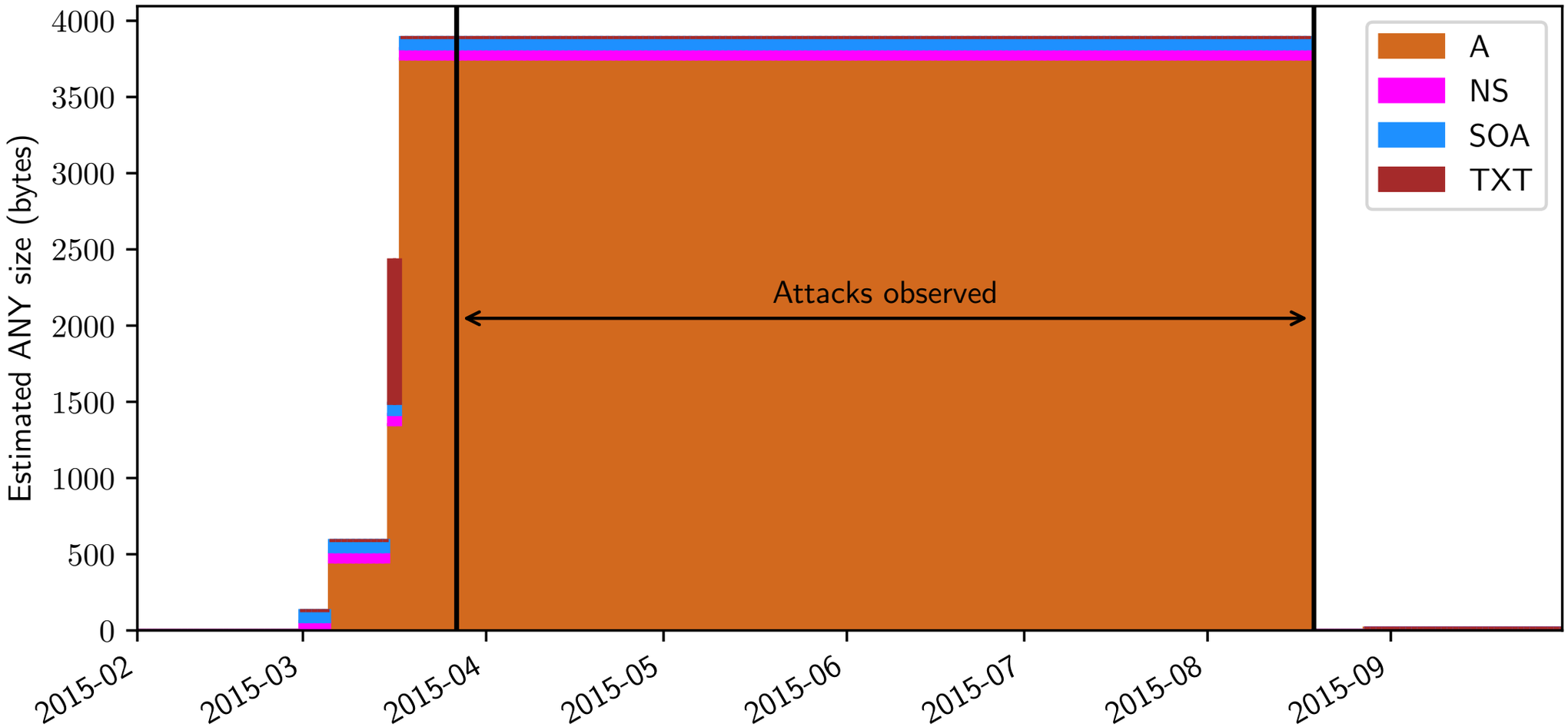


Example 2: crafted domains

- DNS amplification is (still) one of the most frequently used means for volumetric DDoS attacks
- An attacker basically has two options:
  - Abuse a DNSSEC-signed domain (large responses due to signatures)
  - Craft a domain with guaranteed “bang-for-your-buck”
    —> large TXT records, many A records, …

Crafted domains (2)

- While we don't find hundreds of domains, we do find some very creative ones that **have actually been abused***



*With thanks to Christian Rossow and Johannes Krupp from Uni Saarland for AmpPot data that confirms attacks

Examples of what we found

Domain ID	Description
#1	Has parts of a speech by President Obama on net neutrality in `TXT` records.
#2	Has one `TXT` record filled with random garbage.
#3	Has two `TXT` records filled with a mildly offensive repeating word. Has `NS` records that point to CloudFlare name servers.
#4	Has a high number of `A` records in `1.1.1.0/24`.
#5	Have a high number of `A` records in `111.111.0.0/16`.
#6	Has a high number of `AAAA` records in `2001:cafe::/32`.
#7	Has a high number of `A` records reserved for private networks (RFC 1918 [39]).
#8	Same pattern as number #5 but in a different TLD.
#9	Many `A`, `AAAA` and `MX` records, also observed on a Spamhaus blacklist.

Zooming in on #1

- That first one (records below) was observed in over 8000 attacks over more than a year by AmpPot

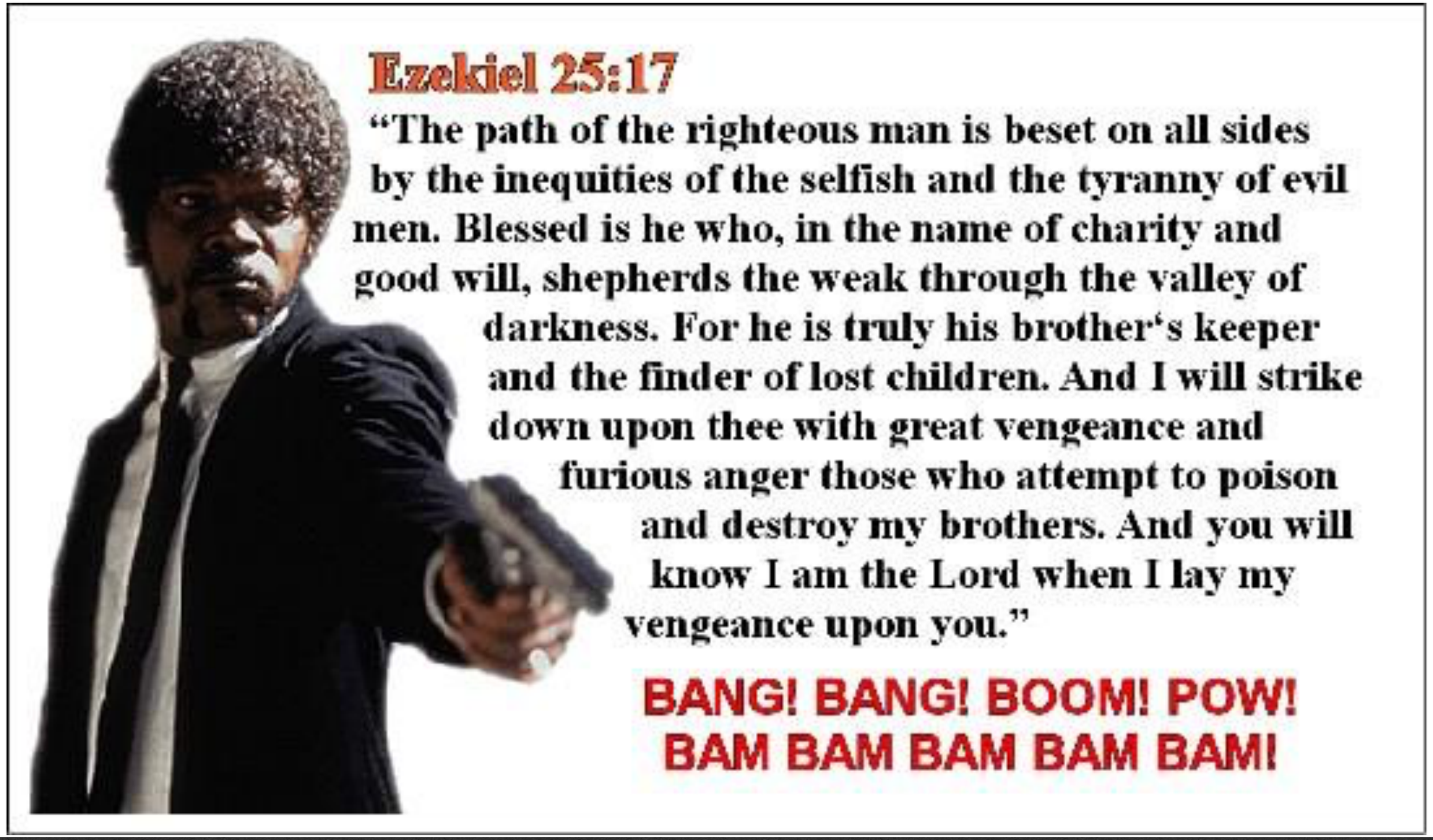
"More than any other invention of our time, the Internet has unlocked possibilities we could just barely imagine a generation ago. And here's a big reason we've seen such incredible growth and innovation: Most Internet providers have treated Internet traff""ic equally”
"That's a principle known as net neutrality and it says that an entrepreneur's fledgling company should have the same chance to succeed as established cor”
"porations, and that access to a high school student's blog shouldn't be unfairly slowed down to make way for advertisers with more money”
"That's what President Obama believes, and what he means when he says there should be no gatekeepers between you and your favorite online sites and services"
"When I was a can didate for this office, I made clear my commitment to a free and open Internet, and my commitment remains as strong as ever. Four years ago, the FCC tried to implement rules that would protect net neutrality with little to no impact on th""e telecommunications companies that make important investments in our economy. After the rules were challenged, the court reviewing the rules agreed with the FCC that net neutrality was essential for preserving an environment that encourages new investmen""t in the network, new online services and content, and everything else that makes up the Internet as we now know it. Unfortunately, the court ultimately struck down the rules not because it disagreed with the need to protect net neutrality, but because ""it believed the FCC had taken the wrong legal approach"
"To be current, these rules must also build on the lessons of the past. For almost a century, our law has recognized that companies who connect you to the world have special obligations not to exploit the monopoly they enjoy over access in and out of your ""home or business. That is why a phone call from a customer of one phone company can reliably reach a customer of a different one, and why you will not be penalized solely for calling someone who is using another provider. It is common sense that the same ""philosophy should guide any service that is based on the transmission of information"

More recently…

- At the beginning of September, another one of these domains popped up. Apparently by someone who likes the bible.

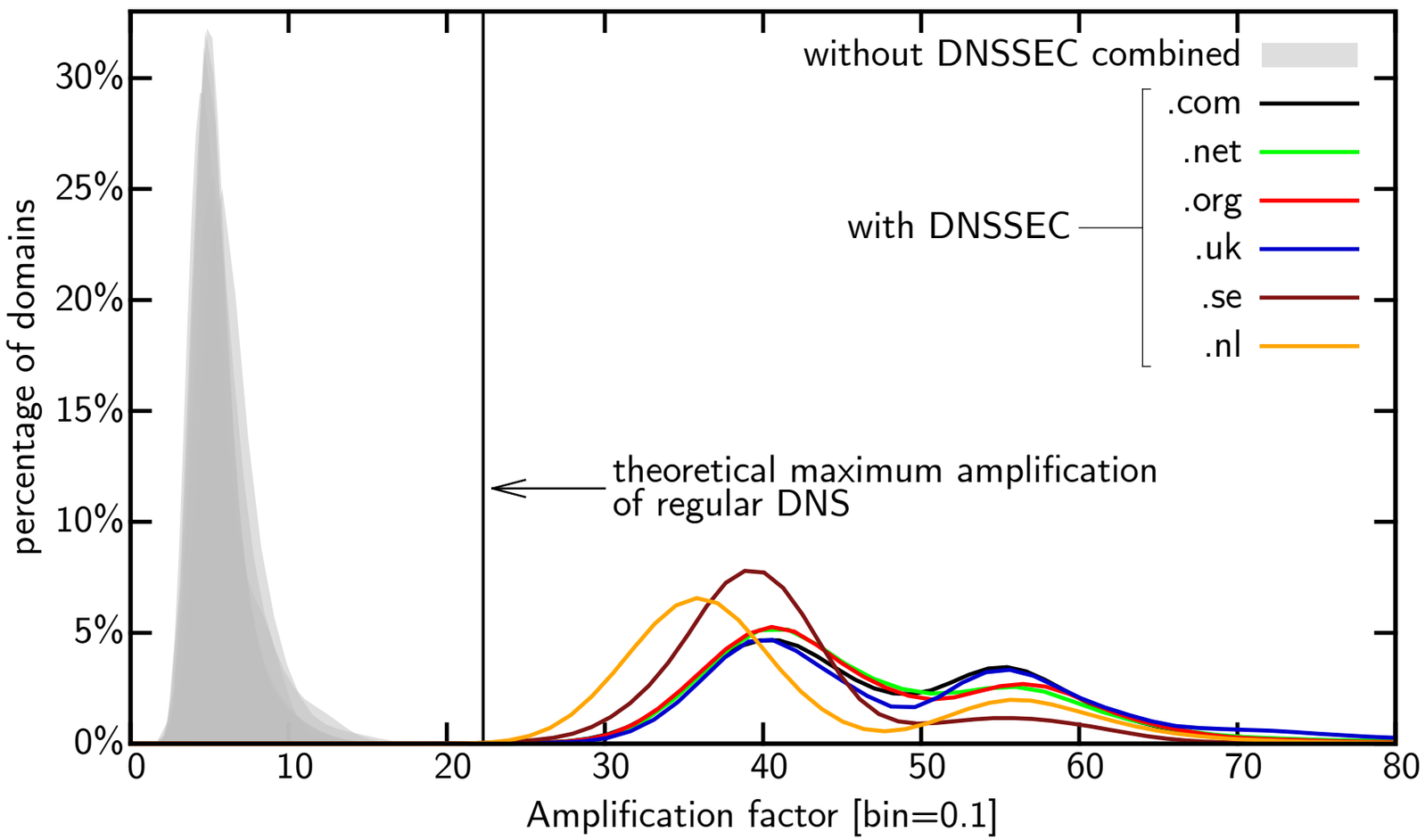
“ButNaomisaidReturnmydaughtersWhyshouldyougowithme?HaveIyetsonsinmywombthattheymaybeyourhusbands?Returnmydaughters!
GoforIamtoooldtohaveahusbandIfIsaidIhavehopeifIshouldevenhaveahusbandtonightandalsobearsonswouldyouthereforewaituntiltheyweregrown?
Wouldyouth""ereforerefrainfrommarrying
NomydaughtersforitisharderformethanforyouforthehandoftheLORDhasgoneforthagainstmeAndtheyliftedup theirvoicesandweptagainandOrpahkisse
dhermotherinlawbutRuthclungtoherThenshesaidBeholdyoursisterinlawhasgonebacktoherpeopleandhergods""returnafteryoursisterinlawButRuths
aidDonoturgemetoleaveyouorturnbackfromfollowingyouforwhereyougoIwillgoandwhereyoulodgeIwilllodgeYourpeopleshallbemypeopleandyourGodm
yGodWhereyoudieIwilldieandthereIwillbeburiedThusmaytheLORDdotomeandworseifanythingbutdeat""hpartsyouandmeWhenshesawthatshewasdeter"
"NowitcameaboutinthedayswhenthejudgesgovernedthattherewasafamineinthelandAndacertainmanofBethleheminJudahwenttosojourninthelandofMoa
bwithhiswifeandhistwosonsThenameofthemanwasElimelechandthenameofhiswifeNaomiandthenamesofhistwosonswereMahlonandChilionEphra""thites
ofBethleheminJudahNowtheyenteredthelandofMoabandremainedthereThenElimelechNaomishusbanddiedandshewasleftwithhertwosonsTheytookforthe
mselvesMoabitewomenaswivesthenameoftheonewasOrpahandthenameoftheotherRuthAndtheylivedthereabouttenyearsThenbothMahlon""andChilionals
odiedandthewomanwasbereftofhertwochildrenandherhusbandThenshearosewithherdaughtersinlawthatshemightreturnfromthelandofMoabforshehadh
eardinthelandofMoabthattheLORDhadvisitedHispeopleingivingthemfoodSoshedepartedfromtheplacewhereshewasandhertwo""daughtersinlawwithhe
randtheywentonthewaytoreturntothelandofJudahAndNaomisaidtohertwodaughtersinlawGoreturneachofyoutohermothershouseMaytheLORDdealkindly
withyouasyouhavedealtwiththedeadandwithmeMaytheLORDgrantthatyoumayfindresteachinthehouseofherhusbandThe""nsh"
"The words of Jeremiah the son of Hilkiah of the priests who were in Anathoth in the land of Benjaminto whom the word of the LORD
came in the days of Josiah the son of Amon king of Judah in the thirteenth year of his reignIt came also in the days of Jehoia""kim
the son of Josiah king of Judah until the end of the eleventh year of Zedekiah the son of Josiah king of Judah until the exile of
Jerusalem in the fifth monthNow the word of the LORD came to me sayingBefore I formed you in the womb I knew you And befo""re you
were born I consecrated you; I have appointed you a prophet to the nationsThen I said Alas Lord GOD! Behold I do not know how to
speak Because I am a youthBut the LORD said to me Do not say I am a youth Because everywhere I send you you shall go An""d all that
I command you you shall speakDo not be afraid of them For I am with you to deliver you declares the LORDThen the LORD stretched out
His hand and touched my mouth and the LORD said to me Behold I have put My words in your mouthSee I have appoint""ed"
"Then God said Let the earth sprout vegetation plants yielding seed and fruit trees on the earth bearing fruit after their kind with
seed in them; and it was so. The earth brought forth vegetation plants yielding seed after their kind and trees bearing fru""it with
seed in them after their kind; and God saw that it was good. There was evening and there was morning a third day. Then God said Let
there be lights in the expanse of the heavens to separate the day from the night and let them be for signs and for ""seasons and for
days and years; and let them be for lights in the expanse of the heavens to give light on the earth; and it was so. God made the two
great lights the greater light to govern the day and the lesser light to govern the night; He made the sta""rs also. God placed them
in the expanse of the heavens to give light on the earth and to govern the day and the night and to separate the light from the
darkness; and God saw that it was good. There was evening and there was morning a fourth day. Then God"" sai"

Where did we see that before?



DNSSEC for DDoS?

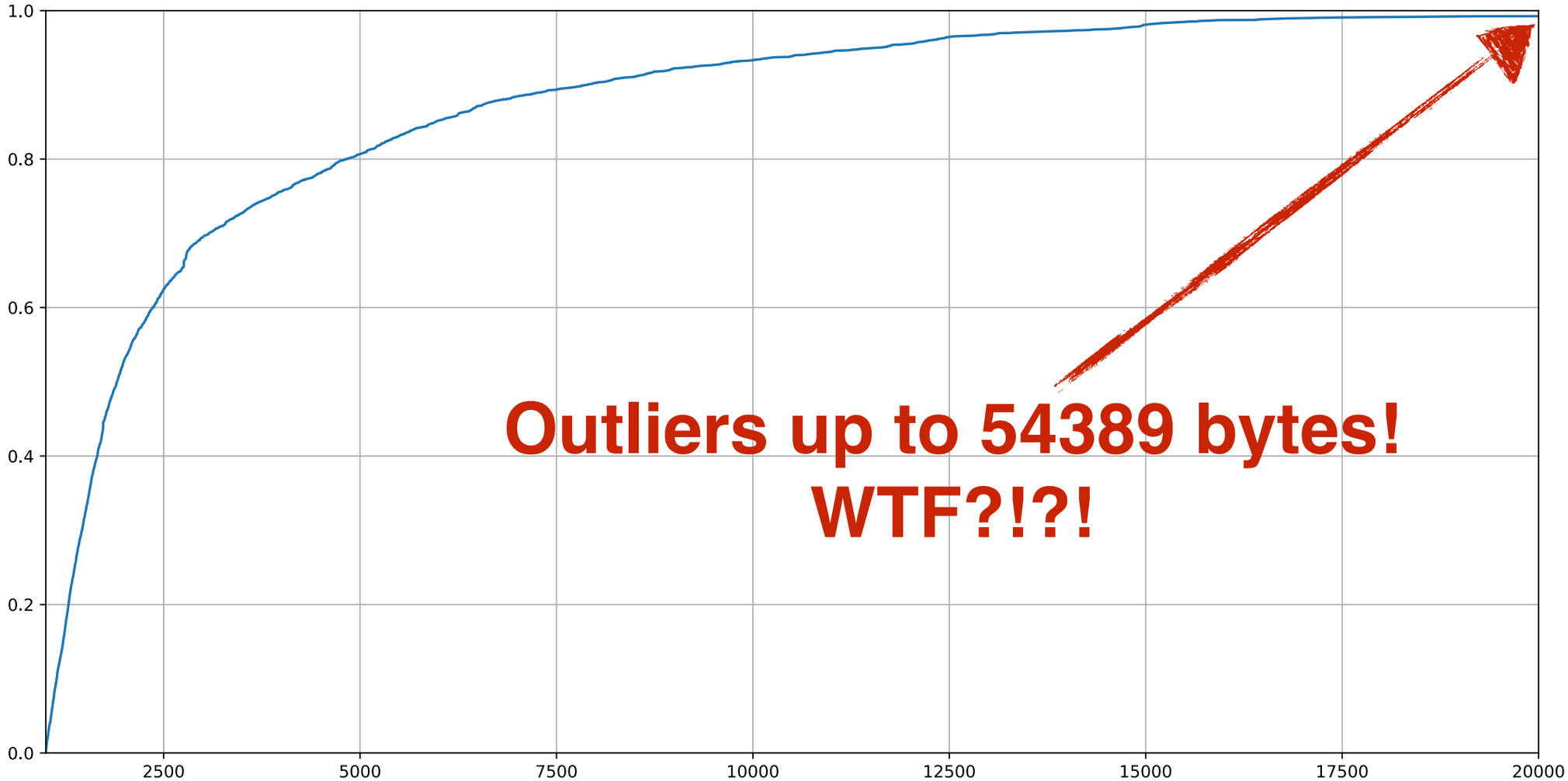
- Many people claim DNSSEC is an amplification attack nightmare. No need to craft domains, just use what is out there:



- Some claim this is a reason not to deploy DNSSEC

But who needs DNSSEC?

- **If we have ".tel" domains?!**



- **3488** domains with **over 1000 bytes** of TXT records
  1288 domains with **over 2500 bytes** of TXT records

Never attribute to malice…

Hanlon's maxim:

"*Never attribute to malice, that which can adequately be explained by stupidity*"

In TXT records we find:

- HTML snippets
- JavaScript
- Windows Powershell code to configure the built-in DNS server
- PEM-encoded X.509 certificates
- Snippets of DNS zone files
- … (you literally can't make this shit up)

And the winner is…

```
-----BEGIN RSA PRIVATE KEY-----
MIIEowIBAAKCAQEA4ggO1HUSc5PscySd74FFDZwWZVxSbg1QlWhlWlqXYzlsCGHD
OoPAXEccE1bia6zqnj7GY9C72i4/ixKp4KcYG74PZXmnmWZ4M9WFkpDlJjTbN1cr
27iHV9wLd8RN1z5ag+0bXrAuD+KkMnT1fSwtDCe5fI2UDJLhb/5TGE2xvXhYl6rw
UpukfTf7QYDO0ekJpKv4XQVkLX0I"
"..." <- I left this part out...
"my5KO0 -----END RSA PRIVATE KEY-----"'
```

- **Why, oh why, oh why…**
- And this is just one example, we've seen quite a few of these.

Example 3: CEO fraud

- August 30, 2016, SURFcert reports incident with CEO fraud, targeting SURFnet among others
- Uses domain names that look like real domain names in e-mails pretending to be from the CEO with instructions to aid in funds transfers
- e.g.:
  “surfnet-nl.net”
  “utwente-nl.net”
  …

More CEO domains?

- Later that day, reports start trickling in that others in the SURF community have seen similar e-mails
- Then SURFcert reports having received a longer list of domain names including ones that look like names used by the SURF community
- When we saw the reports on the SCIRT mailinglist, we decided to see what we could find in OpenINTEL

Digging around

- Let's see record types we find for 'surfnet-nl.net':

```
SELECT DISTINCT response_type
FROM openintel.net_warehouse_parquet
WHERE year="2016" AND month="08" AND day="30"
AND lower(query_name) LIKE '%surfnet-nl.net.';
```

```
+---------------+
| response_type |
+---------------+
| MX            |
| NS            |
| TXT           |
| TXTHASH       |
| NSHASH        |
| SOA           |
| MXHASH        |
+---------------+
```

Who’s handling their e-mail?

```
SELECT DISTINCT mx_address
FROM openintel.net_warehouse_parquet
WHERE year="2016" AND month="08" AND day="30"
AND lower(query_name) LIKE '%surfnet-nl.net.'
AND mx_address IS NOT NULL
```

```
+-----------------------------------------+
| mx_address                              |
+-----------------------------------------+
| surfnetnl-net01i.mail.protection.outlook.com. |
+-----------------------------------------+
```

Hmm… they use Office 365

…

Oh wait, they use Office 365!

Finding similar domains

- **Office 365** use **requires** you to set a domain **validation token** in a TXT record. But this token is **linked to** your **account** not to the domain!

```
SELECT txt_text
FROM openintel.net_warehouse_parquet
WHERE year="2016" AND month="08" AND day="30"
AND lower(query_name) LIKE '%surfnet-nl.net.'
AND txt_text IS NOT NULL
```

At least their fraud is protected against forgery :-)=)

```
+-------------------------------------------------------------------------------------------------+
| txt_text                                                                                        |
+-------------------------------------------------------------------------------------------------+
| "v=spf1 include:spf.protection.outlook.com -all"                                                |
| "mscid=4XkWBaUB7vjkfgFZJpNTnEfgrQYWwGUpm3av8QfuHAPwfT8r0LMLCxlD7mK2S0StiLCy55d0p9n1B5qfB/gC2Q==" |
+-------------------------------------------------------------------------------------------------+
```

OK, can we find others?

```
SELECT COUNT(DISTINCT lower(query_name))
FROM openintel.net_warehouse_parquet
WHERE year="2016" AND month="08" AND day="30"
AND txt_text LIKE '%4XkWBaUB7vjkfgFZJpNTnEfgrQYWwGUpm3av8QfuHAPwfT8r0LMLCxlD7mK2S0StiLCy55d0p9n1B5qfB/gC2Q==%'
```

```
+-----------------------------------+
| count(distinct lower(query_name)) |
+-----------------------------------+
| 17                                |
+-----------------------------------+
```

- Wow, so we found an additional 16 domains with this token!

What about another TLD?

```
SELECT COUNT(DISTINCT lower(query_name))
FROM openintel.com_warehouse_parquet
WHERE year="2016" AND month="08" AND day="30"
AND txt_text LIKE '%4XkWBaUB7vjkfgFZJpNTnEfgrQYWwGUpm3av8QfuHAPwfT8r0LMLCxlD7mK2S0StiLCy55d0p9n1B5qfB/gC2Q==%'
```

```
+-----------------------------------+
| count(distinct lower(query_name)) |
+-----------------------------------+
| 199                               |
+-----------------------------------+
```

- Holy sh*t, we found another **199** domains!

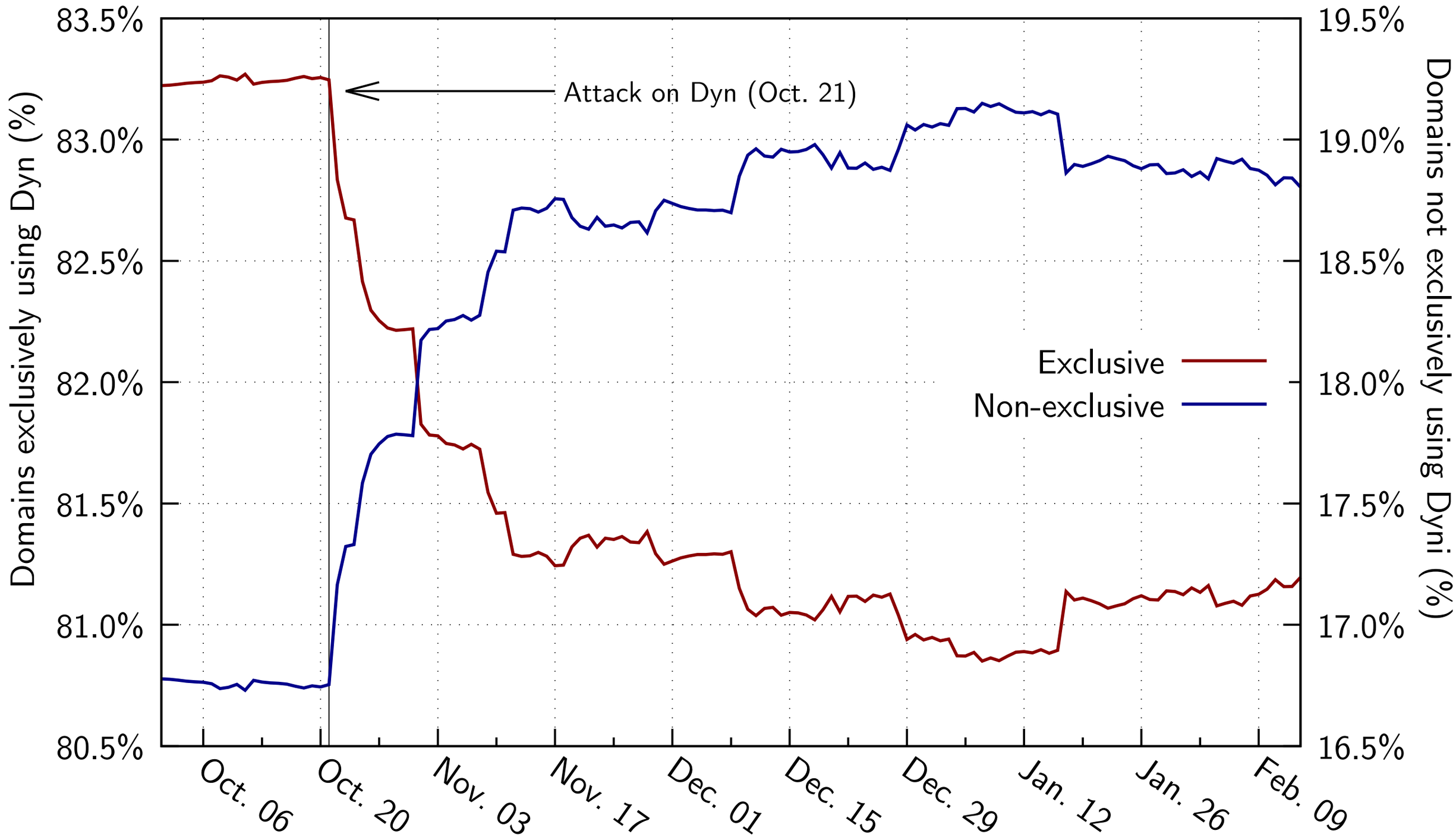
We scripted this

- Based on an input list of **867** domains, we found an additional **1375** domains, so a total of **2242** domains.
- Also found new patterns with letter/digit substitutions:
  "gr0up0n.com"
  "0verstappen.com"
  …
- Conclusion: this kind of **data has direct operational applications**
- Data also shared with NCSC

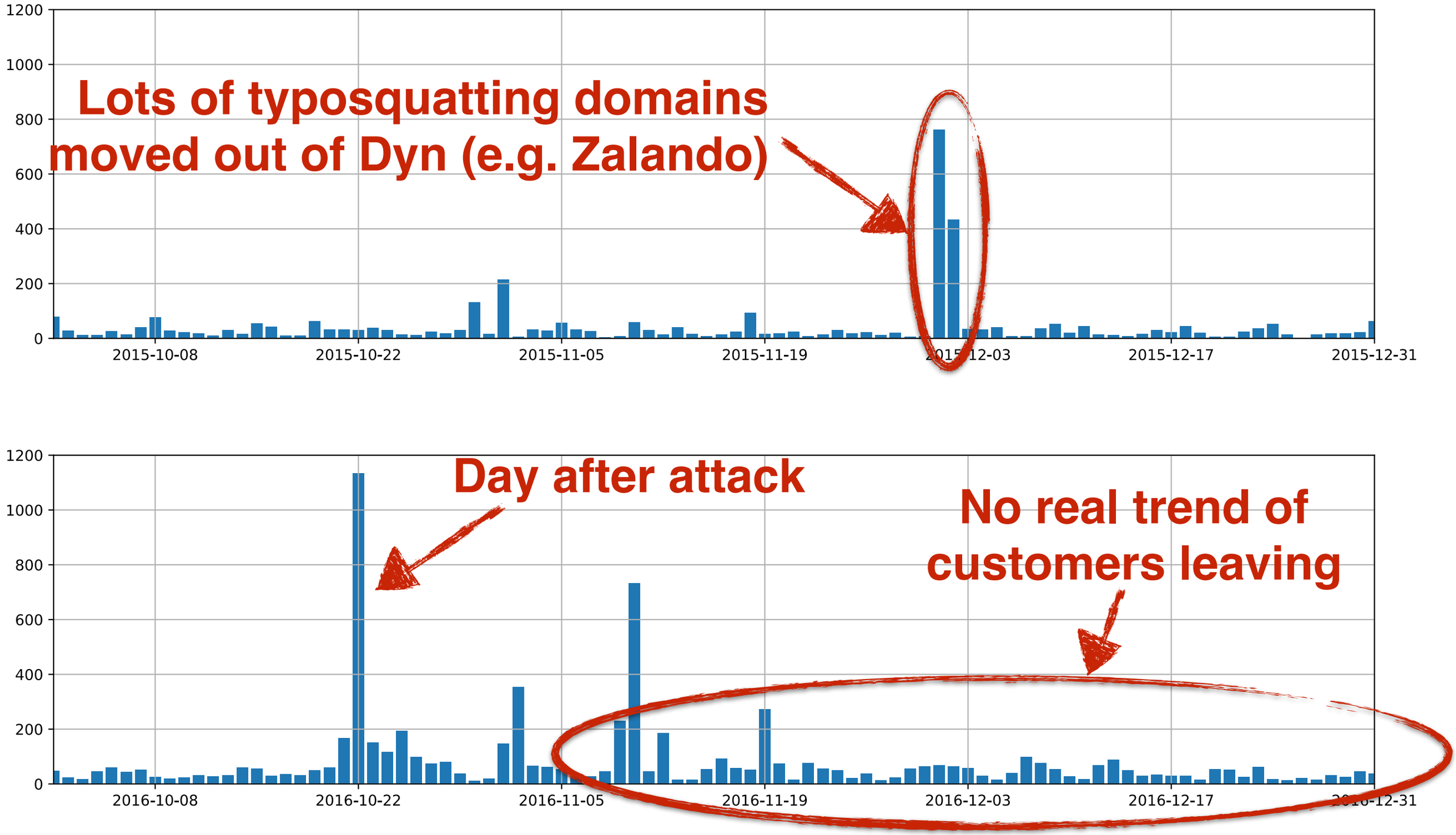
Example 4: Dyn, attack resilience

- On **October 21, 2016, massive DDoS attack** on US East Coast services of **Dyn Inc.** (now Oracle)
- Attack used Mirai botnet — “IoT” (or “Internet of Shit” as I like to call it) devices
- **Dyn** is a DNS **service provider** that people outsource their DNS to, for e.g. **DDoS protection**
- Attack **affected large Internet brands**, e.g. Netflix, Twitter, eBay, Paypal, LinkedIn, …
- Illustration of the risk of **putting all your eggs in one basket**

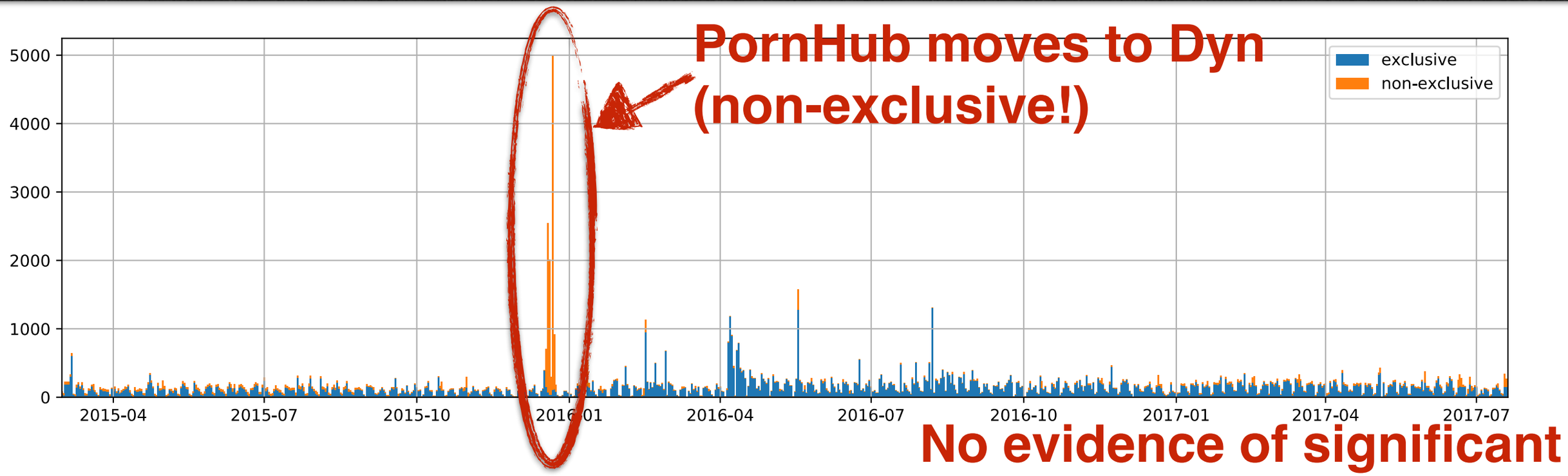
Aftermath of the attack

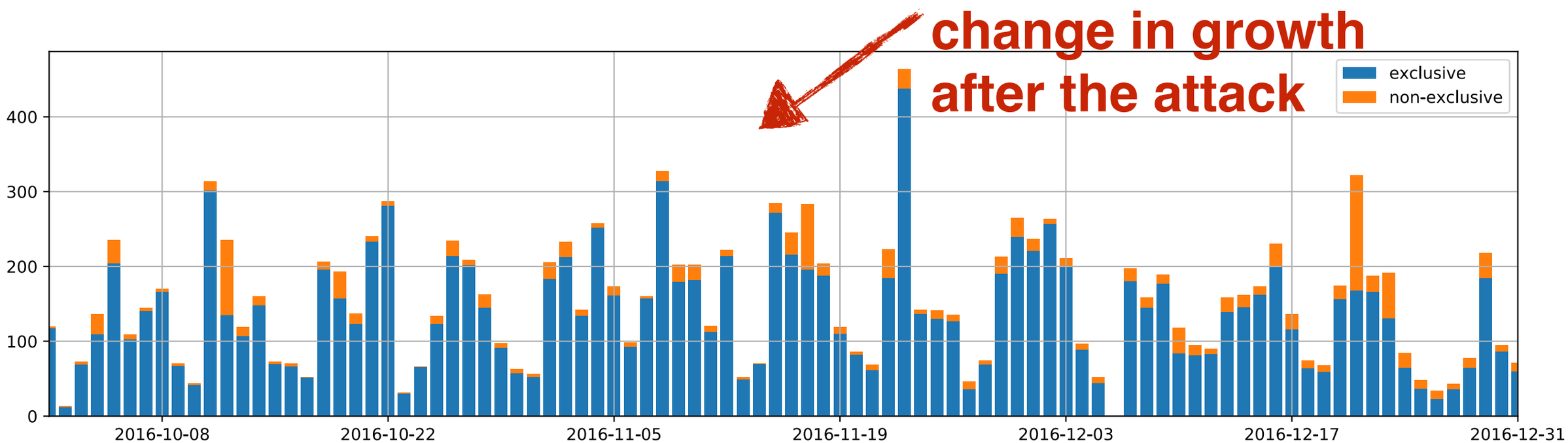


Does it cost Dyn customers?

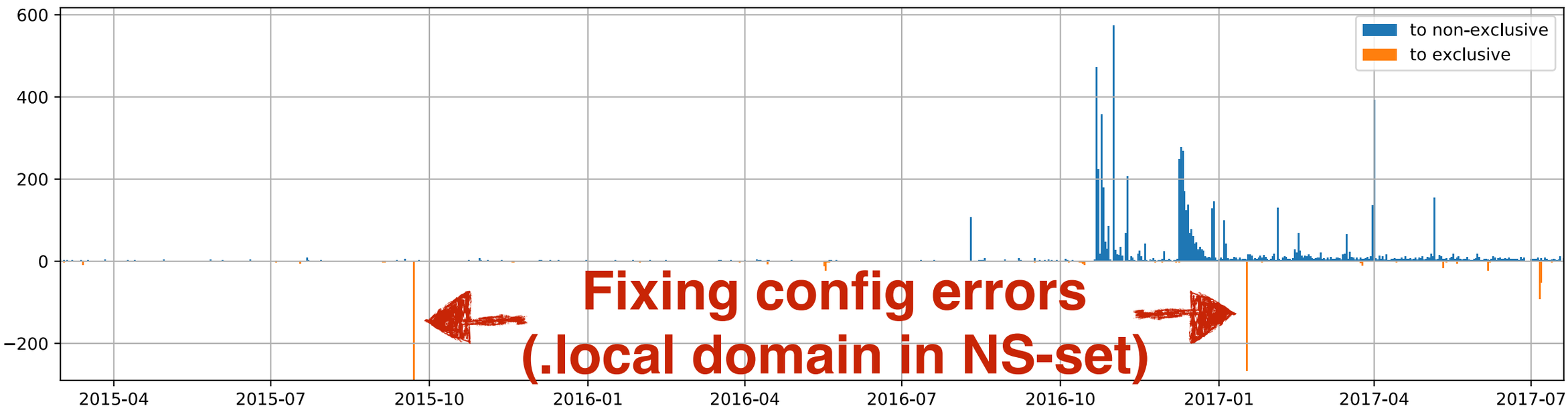


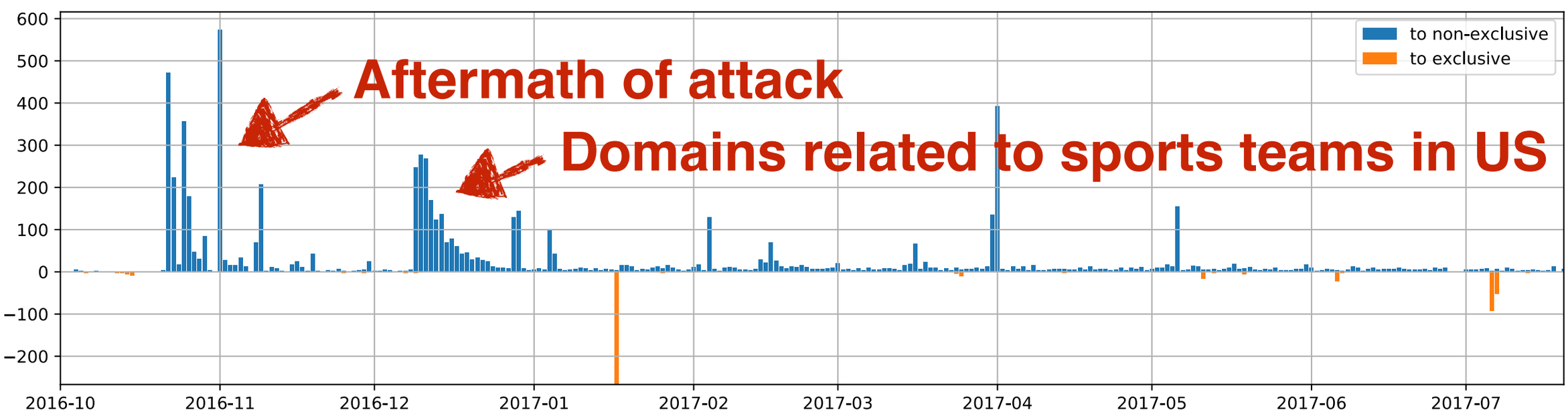
New customers





Switching to non-exclusive





Dyn takeaways

- Our goal is **not to bash Dyn**; this **can happen** to even **the largest providers, through mis-management or attacks**
  (Amazon, OVH, …)

- The **Internet** was **designed to be dis-tributed**, so it is **resilient against attacks on a single part** of it

- Trend of **outsourcing to "the cloud"** is **breaking** that **assumption**

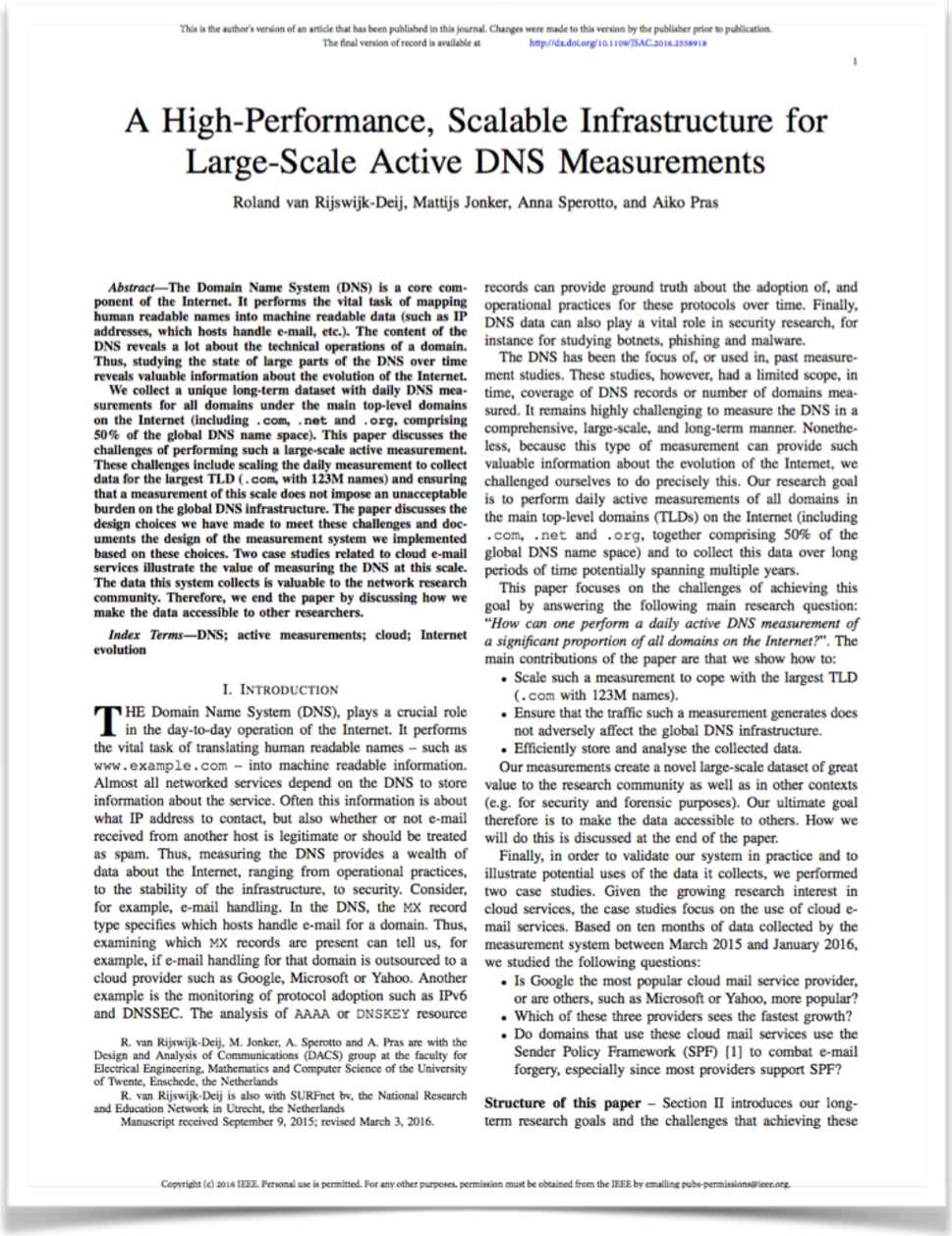
Data access

- We share data with other **academic** researchers
- We publish **open access data** through our webportal **https://www.openintel.nl/**
- Other data - limited access: **contracts for zone file access** (com/net/org/nl/…) **are (very) restrictive**
- Solutions:
  - Can run queries “on behalf”
  - Can provide access to some of the data under conditions of non-disclosure (should be good enough to publish results)

Further reading

van Rijswijk-Deij, R., Jonker, M., Sperotto, A., & Pras, A. (2016). A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements. IEEE Journal on Selected Areas in Communications, 34(7)
http://bit.ly/jsac-openintel

https://openintel.nl/



This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/JSAC.2016.2558918

A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements

Roland van Rijswijk-Deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras

***Abstract*—The Domain Name System (DNS) is a core component of the Internet. It performs the vital task of mapping human readable names into machine readable data (such as IP addresses, which hosts handle e-mail, etc.). The content of the DNS reveals a lot about the technical operations of a domain. Thus, studying the state of large parts of the DNS over time reveals valuable information about the evolution of the Internet.**

We collect a unique long-term dataset with daily DNS measurements for all domains under the main top-level domains on the Internet (including `.com`, `.net` and `.org`, comprising 50% of the global DNS name space). This paper discusses the challenges of performing such a large-scale active measurement. These challenges include scaling the daily measurement to collect data for the largest TLD (`.com`, with 123M names) and ensuring that a measurement of this scale does not impose an unacceptable burden on the global DNS infrastructure. The paper discusses the design choices we have made to meet these challenges and documents the design of the measurement system we implemented based on these choices. Two case studies related to cloud e-mail services illustrate the value of measuring the DNS at this scale. The data this system collects is valuable to the network research community. Therefore, we end the paper by discussing how we make the data accessible to other researchers.

***Index Terms*—DNS; active measurements; cloud; Internet evolution**

I. Introduction

THE Domain Name System (DNS), plays a crucial role in the day-to-day operation of the Internet. It performs the vital task of translating human readable names – such as `www.example.com` – into machine readable information. Almost all networked services depend on the DNS to store information about the service. Often this information is about what IP address to contact, but also whether or not e-mail received from another host is legitimate or should be treated as spam. Thus, measuring the DNS provides a wealth of data about the Internet, ranging from operational practices, to the stability of the infrastructure, to security. Consider, for example, e-mail handling. In the DNS, the `MX` record type specifies which hosts handle e-mail for a domain. Thus, examining which `MX` records are present can tell us, for example, if e-mail handling for that domain is outsourced to a cloud provider such as Google, Microsoft or Yahoo. Another example is the monitoring of protocol adoption such as IPv6 and DNSSEC. The analysis of `AAAA` or `DNSKEY` resource records can provide ground truth about the adoption of, and operational practices for these protocols over time. Finally, DNS data can also play a vital role in security research, for instance for studying botnets, phishing and malware.

The DNS has been the focus of, or used in, past measurement studies. These studies, however, had a limited scope, in time, coverage of DNS records or number of domains measured. It remains highly challenging to measure the DNS in a comprehensive, large-scale, and long-term manner. Nonetheless, because this type of measurement can provide such valuable information about the evolution of the Internet, we challenged ourselves to do precisely this. Our research goal is to perform daily active measurements of all domains in the main top-level domains (TLDs) on the Internet (including `.com`, `.net` and `.org`, together comprising 50% of the global DNS name space) and to collect this data over long periods of time potentially spanning multiple years.

This paper focuses on the challenges of achieving this goal by answering the following main research question: "*How can one perform a daily active DNS measurement of a significant proportion of all domains on the Internet?*". The main contributions of the paper are that we show how to:

- Scale such a measurement to cope with the largest TLD (`.com` with 123M names).
- Ensure that the traffic such a measurement generates does not adversely affect the global DNS infrastructure.
- Efficiently store and analyse the collected data.

Our measurements create a novel large-scale dataset of great value to the research community as well as in other contexts (e.g. for security and forensic purposes). Our ultimate goal therefore is to make the data accessible to others. How we will do this is discussed at the end of the paper.

Finally, in order to validate our system in practice and to illustrate potential uses of the data it collects, we performed two case studies. Given the growing research interest in cloud services, the case studies focus on the use of cloud e-mail services. Based on ten months of data collected by the measurement system between March 2015 and January 2016, we studied the following questions:

- Is Google the most popular cloud mail service provider, or are others, such as Microsoft or Yahoo, more popular?
- Which of these three providers sees the fastest growth?
- Do domains that use these cloud mail services use the Sender Policy Framework (SPF) [1] to combat e-mail forgery, especially since most providers support SPF?

Structure of this paper – Section II introduces our long-term research goals and the challenges that achieving these

R. van Rijswijk-Deij, M. Jonker, A. Sperotto and A. Pras are with the Design and Analysis of Communications (DACS) group at the faculty for Electrical Engineering, Mathematics and Computer Science of the University of Twente, Enschede, the Netherlands

R. van Rijswijk-Deij is also with SURFnet bv, the National Research and Education Network in Utrecht, the Netherlands

Manuscript received September 9, 2015; revised March 3, 2016.

Thank you for your attention!

Questions?

nl.linkedin.com/in/rolandvanrijswijk

@reseauxsansfil

roland.vanrijswijk@surfnet.nl
r.m.vanrijswijk@utwente.nl

UNIVERSITY OF TWENTE.

SURF NET