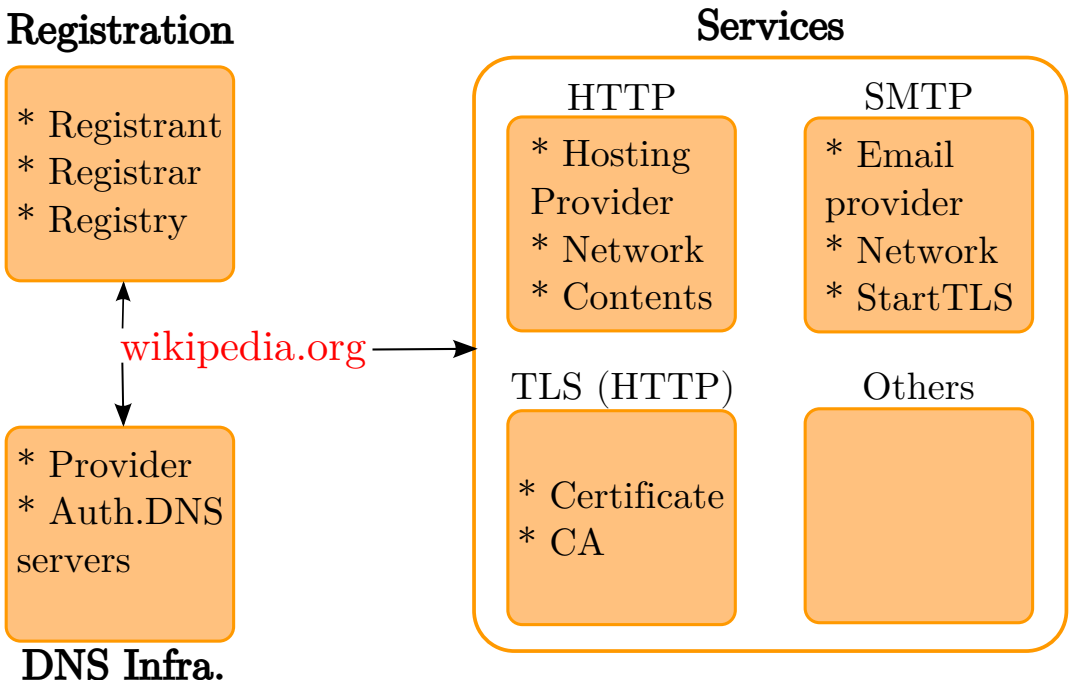
Dmap: Automating Domain Name Ecosystem Measurements and Applications

Maarten Wullink, **Giovane C. M. Moura**,
and Cristian Hesselman
SIDN Labs
Arnhem, the Netherlands
giovane.moura@sidn.nl
giovane.moura@tudelft.nl

IEPG @ IETF102
Montreal, Canada

The *situation*

- **Question:** how to measure all these properties with current measurement tools?



The *situation*

- **Answer:** `zmap + dig + masscan + ... + ad gustum`

you can do it, but it ain't pretty...

The *situation*

The problem:

1. **wasted time** spent on repetitive tasks
   - execution, coordinating different tools, error handling, retries.
2. **heterogeneous data formats** per tool
   - good luck in parsing them
3. **more complexity** $\rightarrow$ **more errors**
   - As we know, one little mistake may void entire experiments [1]
4. **hard to reproduce studies**

Can we do better than this?

- Complexity *begs* for automation
- We faced the same problem at `.nl`
  - a DNS zone with 5.8m domain names
  - imagine periodically crawling it ...
  - for multiple applications

We decided to build a new tool: `Dmap`

- `Dmap`: **Domain** name ecosystem **map**per
- current tools do not fulfill our requirements
- neither current data repositories

We open it for researchers:

- `https://dmap.sidnlabs.nl`

Our requirements

1. Domain-centric (scan list of domains, no IP addresses)
2. Automate the execution of measurements
3. Crawl multiple protocols (DNS, HTTP, TLS, SMTP, etc)
4. Support both IPv4 and IPv6
5. Completeness: fail-safe, retry mechanism
6. Derived features (not only raw datasets)
7. SQL-based interface
8. Modular
9. Distributed
10. Open source

`Dmap` Architecture

- `Dmap` **is not** a stand-alone tool, like `nmap`, `zmap`...
- **It is** a RESTFul web service developed in Java, based on Spring Boot [2]
- Can be run as a single `.jar` file
- Can be distributed over multiple servers for scalability
- Uses PostgreSQL to store results and metadata (not raw measurements)

Dmap Architecture

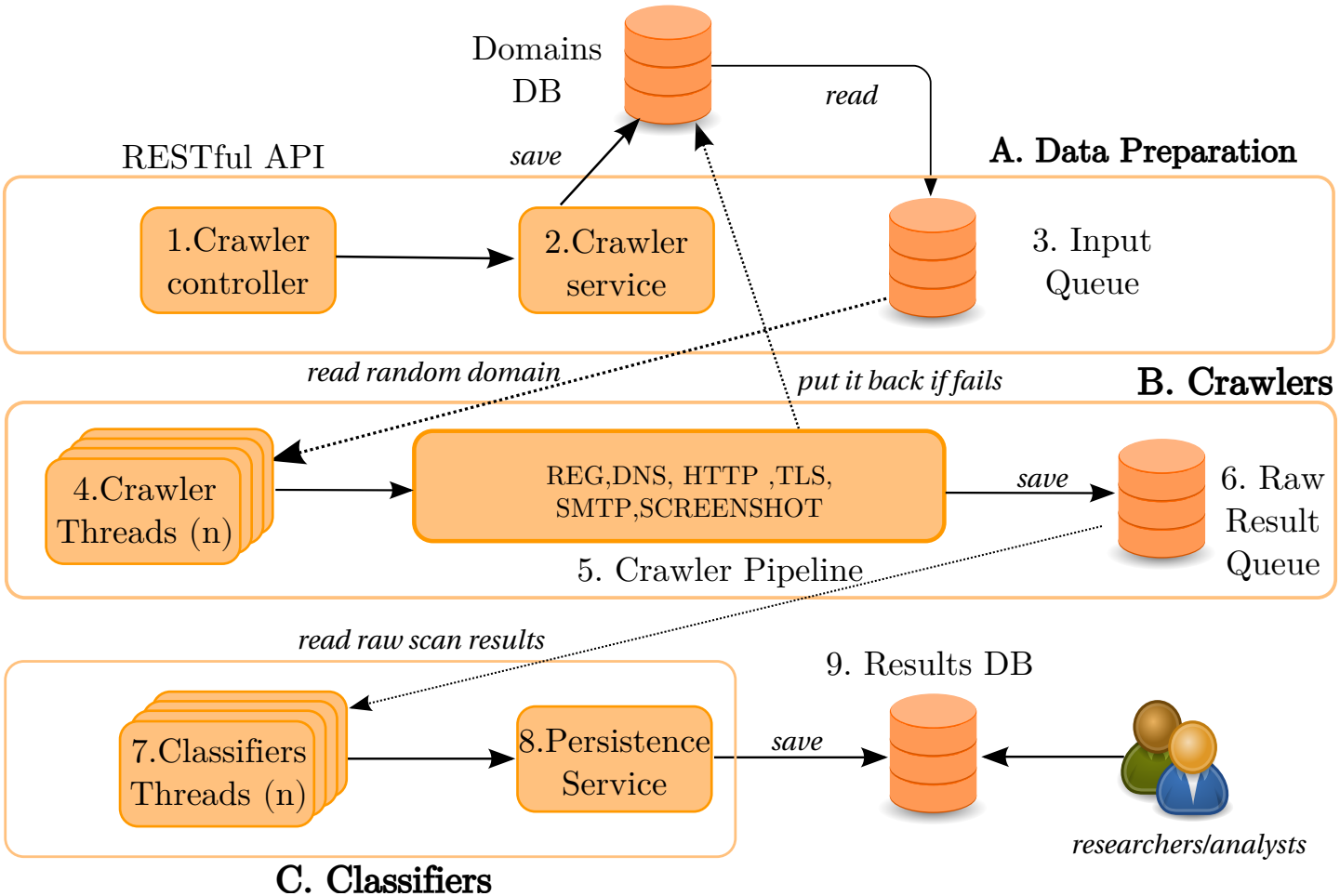
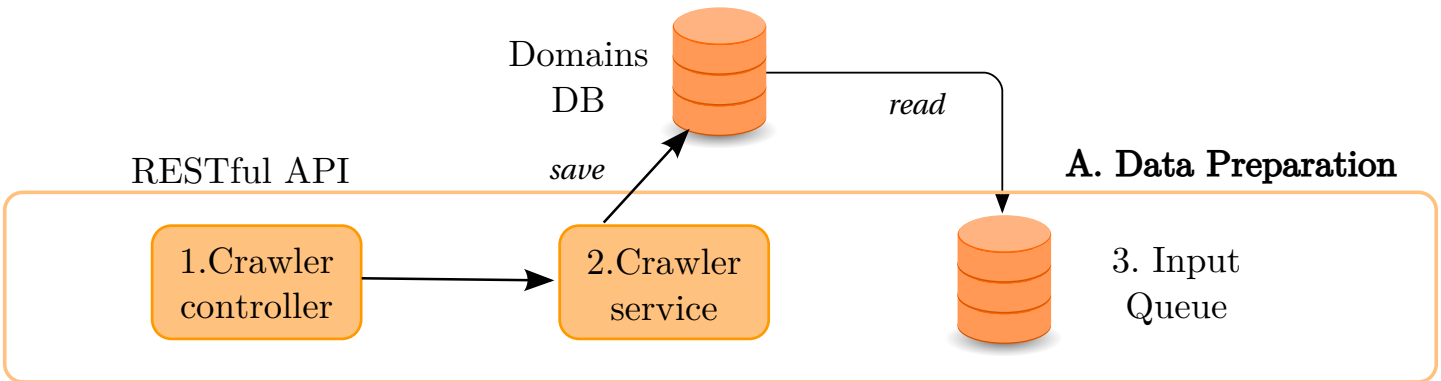


Figure: Dmap Architecture

Dmap Architecture

A. Data preparation:

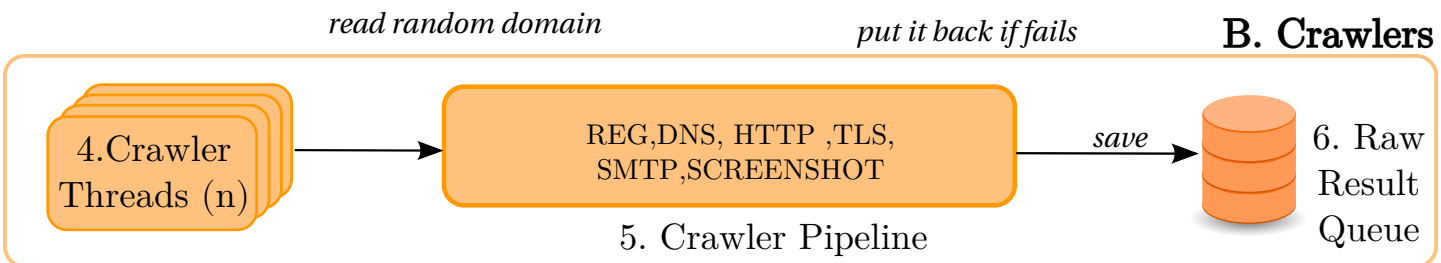
- Crawler service reads csv file with list of domains into PostgreSQL
  - can be used with million domains (5.8m on .nl)
- Crawler service moves 15,000 domains at a time



Dmap Architecture

B. Crawlers modules:

- Crawler reads domains from Input Queue, start crawler threads
- We wrote each crawler – based on libraries
- Each thread uses this domain in a pipeline
  - actual measurements of DNS, HTTP, TLS, SMTP ,Screenshot
- Users can activate/deactive crawlers per measurement
- Dmap keep state of each crawler for each domain
  - fail-safe mechanisms
  - in a PostgreSQL table – so you can check it later



Dmap Architecture

C. Classifiers:

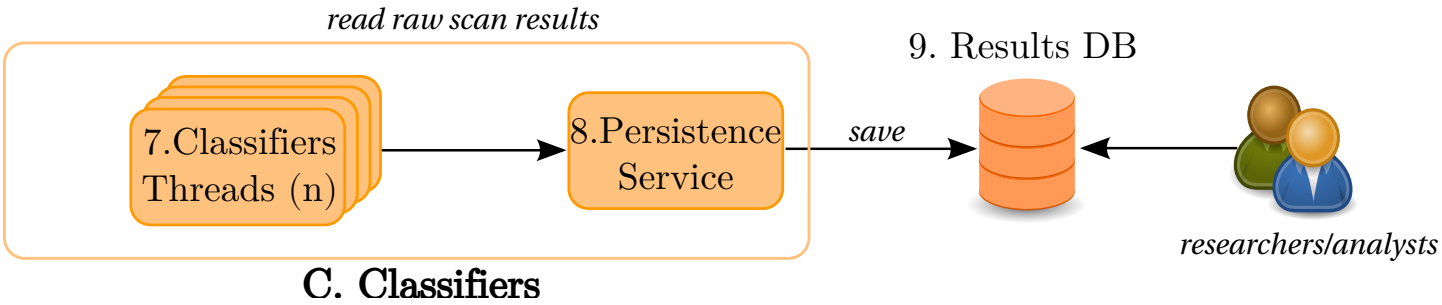


Table: Six out of thirty Dmap classifiers and features

Classifier	Features
TLSCert	extract X.509 cert. info and type of cert
ServerClass	OS and Web server fingerprint
Parking	if a domain is parked and/or for sale
CMSClass	CMS, shopping cart and forum detection
	if a domain has a registrar placeholder page
Business	type of site (online shop)
LetsEncrypt	if domain uses Let's Encrypt Certs

Data Model

- 166 features
- in SQL format, also JSON
- See https://dmap.sidnlabs.nl
- Other tables used for metadata and control
- Demo ?

```
domainname:"wikipedia.org",
crawlRun:764,
ipVersion:4,
url:null,
crawlName:"www.wikipedia.org",
crawlUrl:"https://www.
wikipedia.org/",
crawlDomain:"wikipedia.org",
crawlDate:"2018-05-15
T09:28:48.642+0000",
crawlStatus:0,
crawlPages:2,
crawlRetries:0,
networkLoadTime:24,
networkLoadTimeAll:[
pageLangProb:100,
pageFingerprint: (ommit)
httpRedirect:false,
httpRedirectCount:1,
httpRedirectChain:"www.wi
pedia.org,www.wikipedia.c
httpRedirectHttps:true,
httpRedirectTld:false,
tldStart:"org",
tldEnd:"org",
htmlLinksAll:320,
htmlLinksInt:0,
htmlLinksExt:320,
htmlLinksImg:0,
htmlVersion:"HTML 5",
```

Figure: Sample HTTP features for wikipedia.org

Evaluation

- Input: 5,766,118 domains from `.nl` DNS zone
- Date: September 3rd, 2017
- 8x 1.4GHz VM, 32 GB of RAM, 1Gbps shared line
- 6 crawlers (screenshot off), 300 crawler threads, 50 classifier threads

Evaluation

Precision:

- We compared DNS results with OpenIntel[1]
  - OpenIntel Crawls only DNS records; `Dmap` does multi protocols
- Different vantage points, different software
- Same input: `.nl` domains on Sep. 3rd, 2017
- Metric : number of AAAA records (IPv6 addresses)
  - `Dmap`: 913,593
  - OpenIntel: 913,511

[1] `http://openintel.nl`

Evaluation

HTTP Crawler: *we know when and why it fails, if it fails*

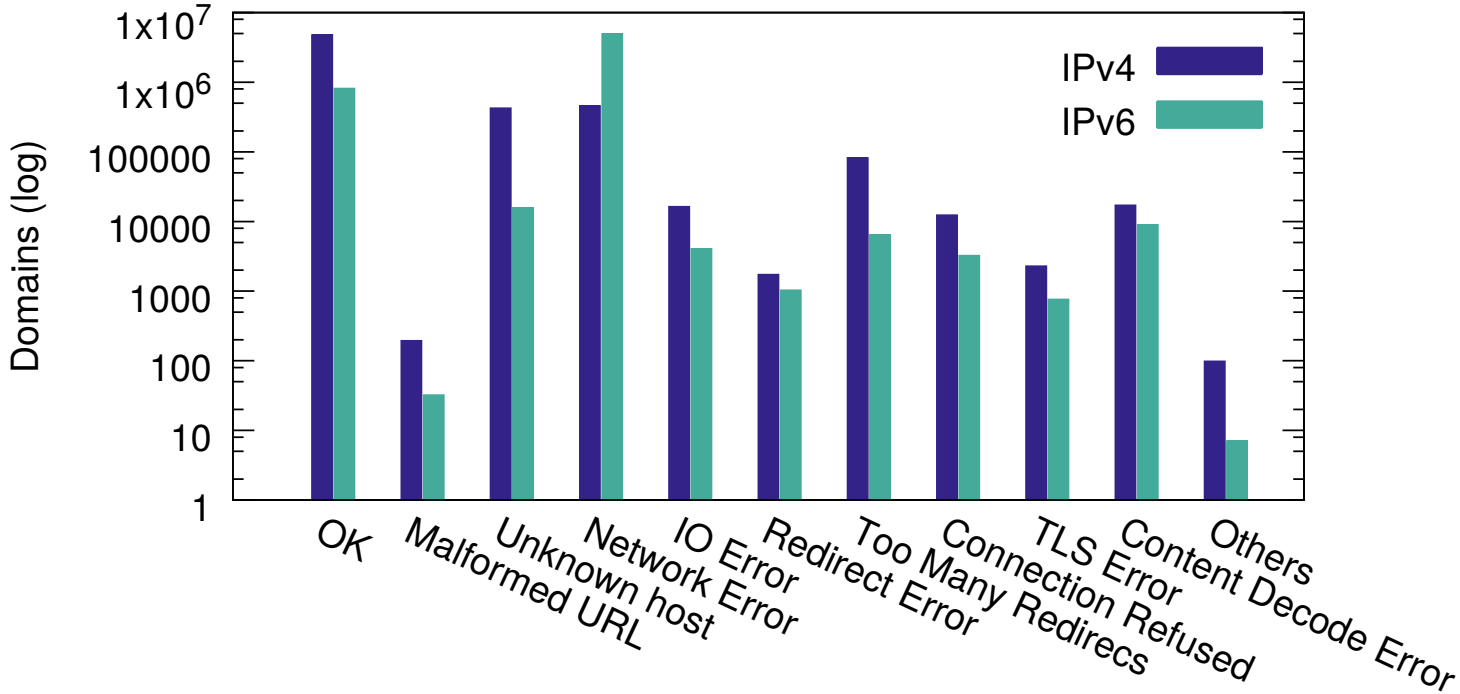


Figure: HTTP crawler result codes

Evaluation

Performance

- This is not a stateless wide scan tool
- It is way more expensive: DNS, HTTP, TLS, SMTP, over both IPv4 and IPv6
- In our setup, `Dmap` crawled 1 million domains a day
  - 8x 1.4GHz VM, 32 GB of RAM, 1Gbps shared line
  - 6 crawlers (screenshot off), 300 crawler threads, 50 classifier threads
- Run it on multiple machines, designed to scale

Applications

- We present three applications for `Dmap`
- They are not an exhaustive list
- We are using several others internal projects
- We hope others build new applications using `Dmap`

Profiling Alexa 1 million domains

Applications

Profiling Alexa 1 million

- **Try it** ! dataset and SQL code available at https://dmap.sidnlabs.nl
- Each result here obtained with SQL
- Hypothesis tests within seconds

query to count # IPv4 HTTP domains

```
select count(1) from crawl_result_http where
crawl_run=67 and ip_version=4 and crawl_status=0;
```

Applications

1. Profiling Alexa 1 million

	IPv4	IPv6	IPv6/IPv4
DNS			
# Domains (OK)	972,155	153,485	0.16
# Unique NSes	289,014	26,127	0.09
# Unique IP	210,650	19,754	0.09
# Unique ASes	18,418	3,178	0.17
# CDN Cloudflare	117,538	115,396	0.98
HTTP			
# Domains (OK)	968,338	153,485	0.16
# HTML 5	681,757	116,066	0.17
Bytes (median)	53,889	64,735	1.20
External links (median)	7	8	1.14
Internal links (median)	67	75	1.12
Cookies (median)	1	1	1.00
TLS			
# Domains (OK)	772,455	129,443	0.17
# Let's Encrypt	165,526	10,466	0.06
SMTP			
# Domains (OK)	843,126	190,736	0.23
# Unique SMTP	501,848	24,311	0.05
# Unique IP	286,504	10,113	0.04
# Unique StartTLS	302,871	8,016	0.03

Applications : Alexa profiling

IPv6 adoption is slightly faster on SMTP

	IPv4	IPv6	IPv6/IPv4
DNS			
# Domains (OK)	972,155	153,485	0.16
# Unique NSes	289,014	26,127	0.09
# Unique IP	210,650	19,754	0.09
# Unique ASes	18,418	3,178	0.17
# CDN Cloudflare	117,538	115,396	0.98
HTTP			
# Domains (OK)	968,338	153,485	0.16
# HTML 5	681,757	116,066	0.17
Bytes (median)	53,889	64,735	1.20
External links (median)	7	8	1.14
Internal links (median)	67	75	1.12
Cookies (median)	1	1	1.00
TLS			
# Domains (OK)	772,455	129,443	0.17
# Let's Encrypt	165,526	10,466	0.06
SMTP			
# Domains (OK)	843,126	190,736	0.23
# Unique SMTP	501,848	24,311	0.05
# Unique IP	286,504	10,113	0.04
# Unique StartTLS	302,871	8,016	0.03

Applications : Alexa profiling

77% support HTTPS, 1 in 5 are Let's Encrypt

	IPv4	IPv6	IPv6/IPv4
DNS			
# Domains (OK)	972,155	153,485	0.16
# Unique NSes	289,014	26,127	0.09
# Unique IP	210,650	19,754	0.09
# Unique ASes	18,418	3,178	0.17
# CDN Cloudflare	117,538	115,396	0.98
HTTP			
# Domains (OK)	968,338	153,485	0.16
# HTML 5	681,757	116,066	0.17
Bytes (median)	53,889	64,735	1.20
External links (median)	7	8	1.14
Internal links (median)	67	75	1.12
Cookies (median)	1	1	1.00
TLS			
# Domains (OK)	772,455	129,443	0.17
# Let's Encrypt	165,526	10,466	0.06
SMTP			
# Domains (OK)	843,126	190,736	0.23
# Unique SMTP	501,848	24,311	0.05
# Unique IP	286,504	10,113	0.04
# Unique StartTLS	302,871	8,016	0.03

Measure Web Encryption Adoption on .nl

Applications

2. Measure Web Encryption Adoption (on .nl)

- Dmap supports longitudinal studies

Table: TLS/SSL Web deployment on .nl zone over IPv4

	20170903	**20180201**
Zone size	5,766,118	5,801,191
TLS/SSL	2,595,281 (45.10%)	2,674,877 (46.10%)
Types of Certificates		
Unknown	382 (0.01%)	407 (0.01%)
self-signed	883,844 (34.05%)	749,033 (28.01%)
DV	1,406,072 (54.17%)	1,618,575 (60.5%)
OV	269,284 (10.37%)	272,318 (10.18%)
EV	35,699 (1.37%)	34,544 (1.29%)
Let's Encrypt DV deployed certificates share		
Let's Encrypt (%DV)	366,623 (14.12%)	523,029 (32.31%)

Applications: web encryption evolution on .nl

Small growth on Web encryption over 5 month period

Table: TLS/SSL Web deployment on .nl zone over IPv4

	20170903	**20180201**
Zone size	5,766,118	5,801,191
TLS/SSL	2,595,281 (45.10%)	2,674,877 (46.10%)
Types of Certificates		
Unknown	382 (0.01%)	407 (0.01%)
self-signed	883,844 (34.05%)	749,033 (28.01%)
DV	1,406,072 (54.17%)	1,618,575 (60.5%)
OV	269,284 (10.37%)	272,318 (10.18%)
EV	35,699 (1.37%)	34,544 (1.29%)
Let's Encrypt DV deployed certificates share		
Let's Encrypt (%DV)	366,623 (14.12%)	523,029 (32.31%)

Applications: web encryption evolution on .nl

Decline of self-signed, growth of DV certs

Table: TLS/SSL Web deployment on .nl zone over IPv4

	20170903	**20180201**
Zone size	5,766,118	5,801,191
TLS/SSL	2,595,281 (45.10%)	2,674,877 (46.10%)
Types of Certificates		
Unknown	382 (0.01%)	407 (0.01%)
self-signed	883,844 (34.05%)	749,033 (28.01%)
DV	1,406,072 (54.17%)	1,618,575 (60.5%)
OV	269,284 (10.37%)	272,318 (10.18%)
EV	35,699 (1.37%)	34,544 (1.29%)
Let's Encrypt DV deployed certificates share		
Let's Encrypt (%DV)	366,623 (14.12%)	523,029 (32.31%)

Applications: web encryption evolution on .nl

1/3 of .nl DV certs of domains are by Let's Encrypt

Table: TLS/SSL Web deployment on .nl zone over IPv4

	20170903	**20180201**
Zone size	5,766,118	5,801,191
TLS/SSL	2,595,281 (45.10%)	2,674,877 (46.10%)
Types of Certificates		
Unknown	382 (0.01%)	407 (0.01%)
self-signed	883,844 (34.05%)	749,033 (28.01%)
DV	1,406,072 (54.17%)	1,618,575 (60.5%)
OV	269,284 (10.37%)	272,318 (10.18%)
EV	35,699 (1.37%)	34,544 (1.29%)
Let's Encrypt DV deployed certificates share		
Let's Encrypt (%DV)	366,623 (14.12%)	523,029 (32.31%)

Applications

CMS/Web server usage on .nl

Applications

3. CMS/Web server profiling and usage

- We wrote classifiers that identify those
- Apache/Wordpress dominates it

Table: CMS and Web Server Usage `.nl` zone (20170903)

CMS	Total (%)	Web Server	Total (%)
Wordpress	848,083 (73.84%)	Apache	3,397,930 (77.05%)
Joomla	99,865 (8.69%)	nginx	696,697 (15.79%)
Drupal	40,798 (3.55%)	MS ISS	249,367 (5.65%)
Blogo	20,749 (1.80%)	cloud-nginx	60,530 (1.13%)
Wix	17,861 (1.53%)	Coyote	4,237 (0.01%)
Others	121,232 (10.43%)	Others	905 (0.00%)
Total	**1,148,408**	**Total**	**4,409,666**

Summary

- Complexity is bad for measurements
- `Dmap` automates and reduces complexity of multi-application measurements
- It has now 40 classifers (30 when we wrote the paper), that add rich set of features
- Besides, provides an SQL data model (easy to analyze)
- It's open, scalable, and distributed
- Potential to speed up your research/PhD/project
- We developed and use it in `.nl`
  - Why not share it with researchers?

Future work – you can help us

Nicolaes Visscher, 1658, cartographer from Amsterdam

Antarctica was only discovered in 1820! Let's map other protocols

https://www.newyorker.com/magazine/2018/02/12/the-white-darkness

Future work – you can help us

Nicolaes Visscher, 1658, cartographer from Amsterdam

Antarctica was only discovered in 1820! Let's map other protocols

https://www.newyorker.com/magazine/2018/02/12/the-white-darkness

Questions?

Download it!

- https://dmap.sidnlabs.nl
- contact: giovane.moura@sidn.nl
- **We are hiring!**
  - 2 Future Internet Research Positions, 1 Machine Learning Engineer
  - Talk to me if you're interested
  - https://sidnlabs.nl

Bibliography I

[1] V. Paxson, “Strategies for Sound Internet Measurement,” in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC ’04. New York, NY, USA: ACM, 2004, pp. 263–271. [Online]. Available: http://doi.acm.org/10.1145/1028788.1028824

[2] Spring, “Spring Boot,” Feb. 2018, https://projects.spring.io/spring-boot/.