Counterfighting Counterfeit: detecting and taking down fraudulent webshops at a ccTLD

Thymen Wabeke[1], **Giovane C. M. Moura**[1],
Nanneke Franken[2], Cristian Hesselman[1,3]

Guest Lecture – Internet Security
University of Twente
2020-03-02

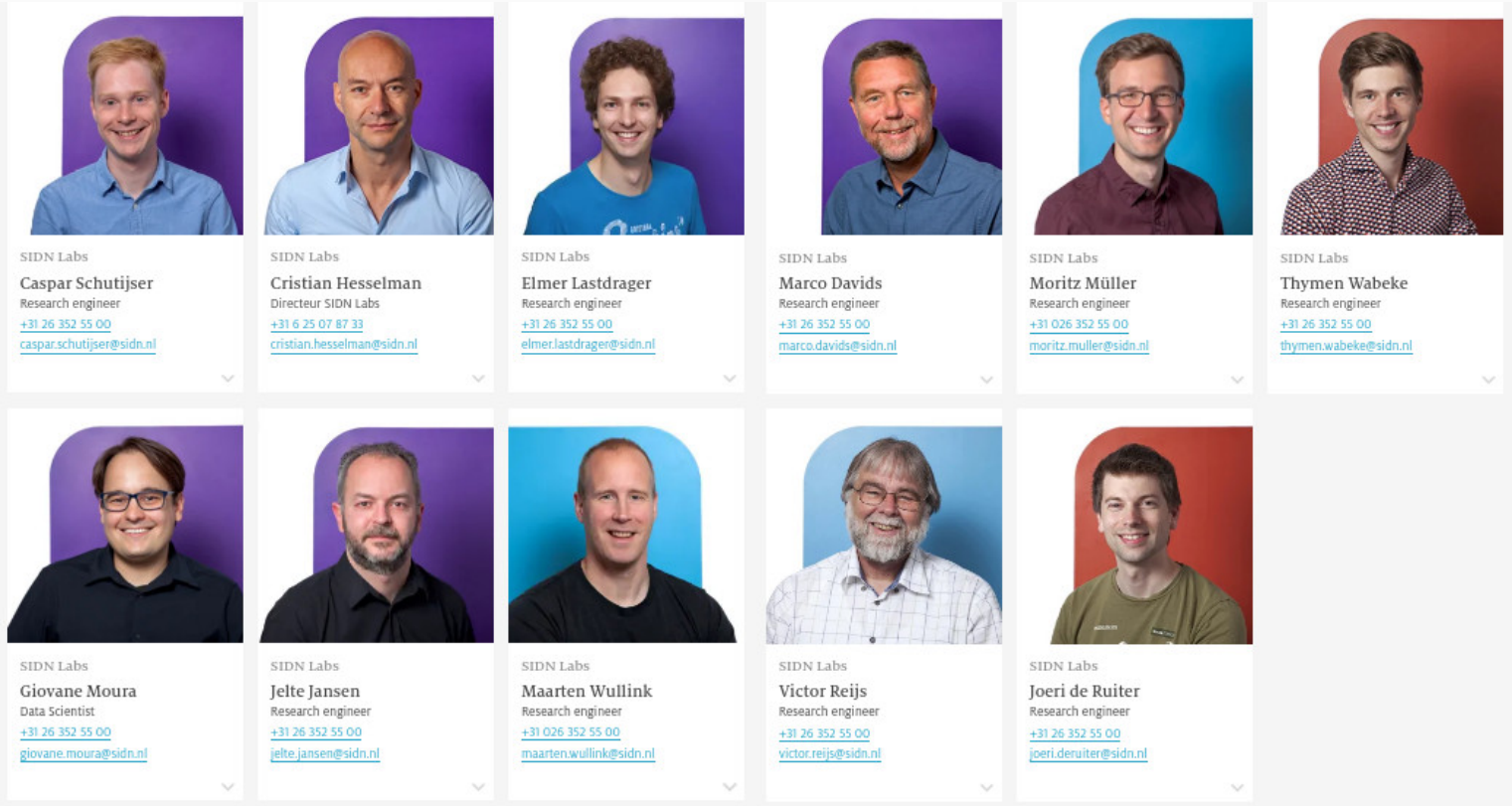
[1]SIDN Labs, [2]SIDN, [3]University of Twente

$whoami

- Data Scientist at SIDN Labs (Since 2015)
- Before:
  - PostDoc at TU Delft
  - Ph.D. from DACS/Twente (2013)
  - MSc from UFRGS, Brazil
- Activities
  - Research projects on both security and network engineering
  - Foot on academia and industry: IMC, IETF, RIPE

SIDN Labs

- Research arm of SIDN (`.nl` registry)
- Goal: advance security and resilence of the Internet
- A research team in the industry



What do we do?

1. Software and tools
   - https://entrada.sidnlabs.nl
   - https://spin.sidnlabs.nl
2. Papers, IETF drafts, and data
   - https://sidnlabs.nl/en/publications
   - https://stats.sidnlabs.nl
3. Services
   - NTP: https//time.nl
   - DoH resolver : https://doh.sidnlabs.nl/

SIDN

- `.nl` **registry**
- But what is a DNS registry?
  - Contact list for `.nl`
  - List of all `.nl` domains, and handles registrations and authoritative DNS servers
- **not-for-profit**, foundation
  - incentives and mission to invest on security and stability of the Internet

This research

- Will apear on the forthcoming PAM2020 conference
  - https://pam2020.cs.uoregon.edu/
- Covers our efforts over the last 3 years in finding and taking down scam sites
- Paper (PDF) at : https://tinyurl.com/yxx4tnwl
- Today: interactive lecture, with questions to you
- Think as a counterfeiter: how can I optimize my process to make more money?

Warm-up questions

1. Who wears sneakers/running shoes?
2. Who buys them online?
3. Who has bought them on “shady” web sites?

Warm-up questions

1. Who wears sneakers/running shoes?
2. Who buys them online?
3. Who has bought them on “shady” web sites?

Warm-up questions

1. Who wears sneakers/running shoes?
2. Who buys them online?
3. Who has bought them on “shady” web sites?

Sale: Nike Air Max 60% Off

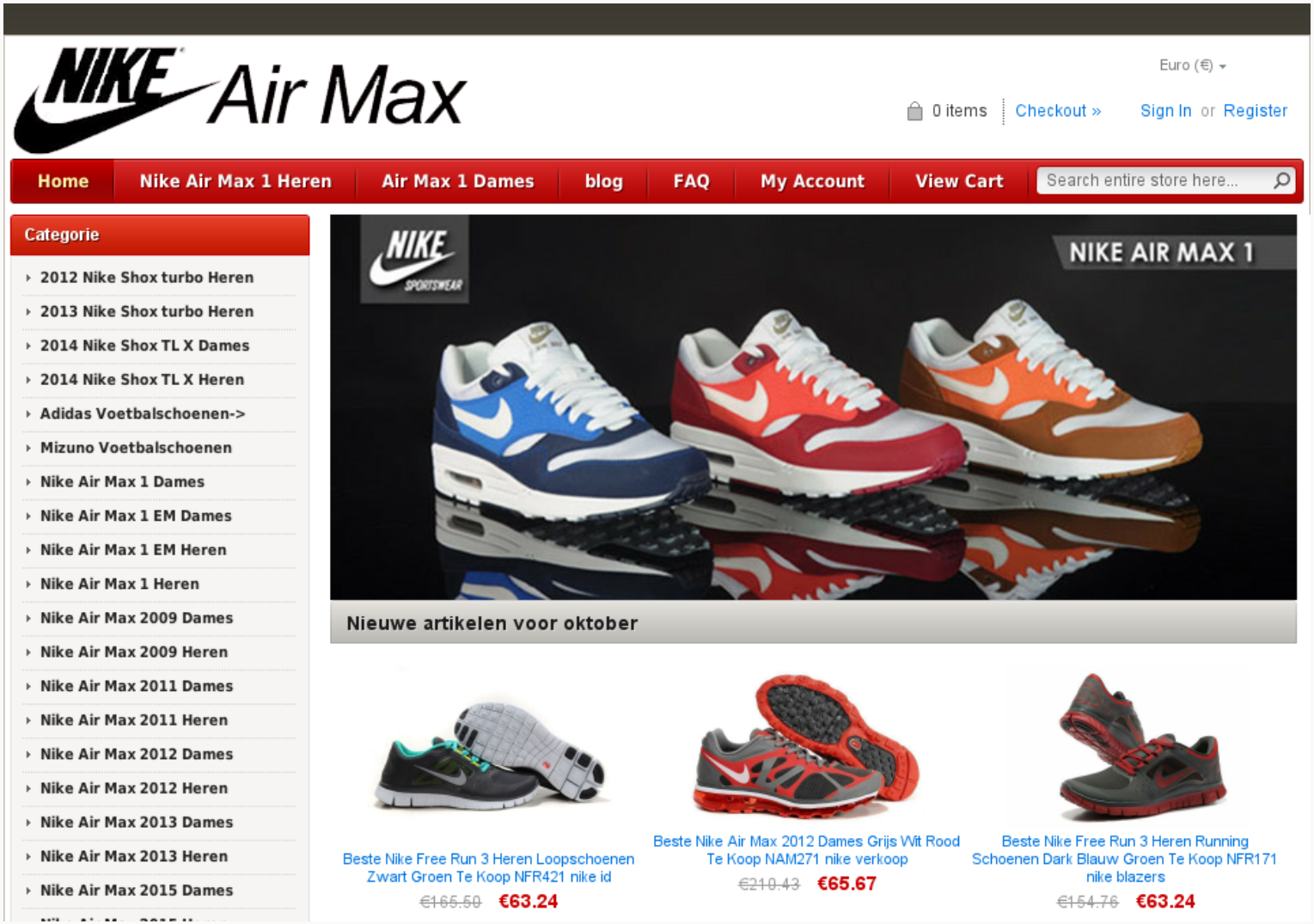


Figure 1: Screenshot of a `.nl` website (2016)

Back in 2015

- We stumbled on these websites while looking for phishing [1]
- They were rather *odd*
- We had many questions:
  1. does anyone even *buy* from them?
  2. what is their *business model*?
  3. how many they were (on .nl)?
  4. what can we do about it?

1. Does anyone even *buy* from them?

Figure 2: NOS news (2018) – https://nos.nl/artikel/2258095-consumenten-voor-5-miljoen-euro-opgelicht-via-nepwinkels-op-sociale-media.html

- So yeah, people were buying from them
- Just be scammed: getting fake or no product
- Dealing with financial losses

1. Does anyone even *buy* from them?

Figure 2: NOS news (2018) – https://nos.nl/artikel/2258095-consumenten-voor-5-miljoen-euro-opgelicht-via-nepwinkels-op-sociale-media.html

- So yeah, people were buying from them
- Just be scammed: getting fake or no product
- Dealing with financial losses

2. What is their *business model*?

- Counterfeit (fake) industry is **huge**: books, computers, shoes, bags, electronics
  - EU borders seizures 2016: 670 miliion EUR
  - US 2017: US$ 1.2 Billion
- Luxury goods have a massive demand

2. What is their *business model*?

- Penalty risks are rather low (compared with drugs)
- They are sort of sweet spot:
  - large demand
  - large profit margins [4]
  - low risk of getting caught (on the Internet) [4]
- Online you may be scammed
  - (differently from buying from a street vendor)
- That's why we got involved

2. What is their *business model*?

- The business model goes like this:
  1. Consumer demand [4]
  2. Manufacturing in China [3]
  3. These webshops connect both of them
- It's not only a .nl problem:
  - .de, .be, .com, and many others have the same issue
- We are dealing with *pros* here

3. How many there are on .nl

- Back to 2016: we stumbled on them
- We realized they all share a similar pattern:
  1. long html <title> tags

```
<title>Vans Schoenen On Sale 70% OFF |Geen
    verzendkosten</title>

```

  2. tags listing many brands (Nike, Reebok, Gucci, you name it..)
- Question: Why this tactic?

3. How many there are on .nl

- Back to 2016: we stumbled on them
- We realized they all share a similar pattern:
  1. long html <title> tags

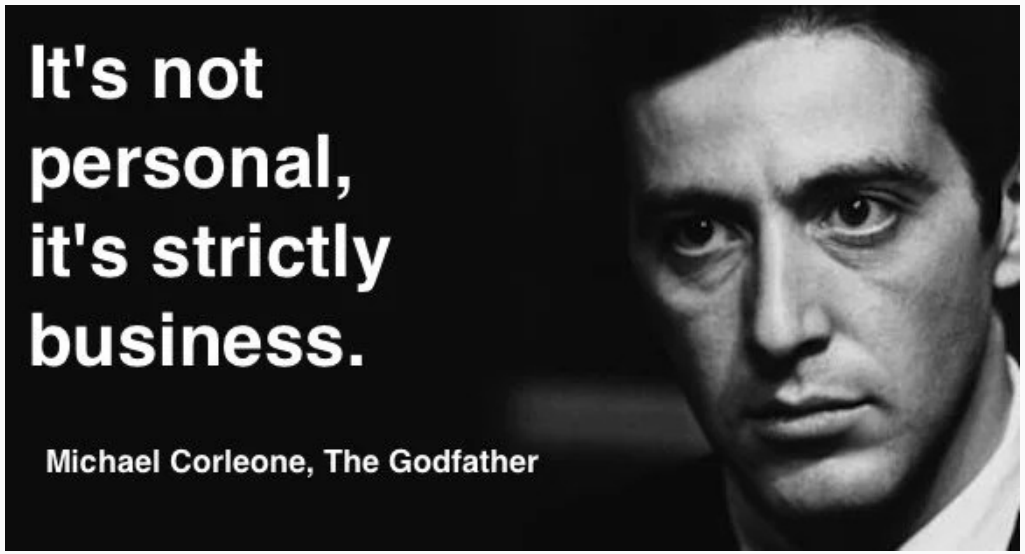
```
<title>Vans Schoenen On Sale 70% OFF |Geen
    verzendkosten</title>

```

  2. tags listing many brands (Nike, Reebok, Gucci, you name it..)
- **Question: Why this tactic?**

Search Engine Optimization (SEO)

- To rank high on Google
  - more clicks $\rightarrow$ more $
- Listing brands may help them rank high [5] (SEO)



Can we used that to detect these shops?

Any ideas on how to do it?

- Consider we *know* all active `.nl` domain names
- We need the `html` `<title>` tags from 5.8 million domains
- **How to get that information?**
  - Luckily, we had Dmap (`https://dmap.sidnlabs.nl`)
  - Crawls HTTP, DNS, TLS, MX and screenshot
  - Run once a month on SIDN

Can we used that to detect these shops?

Any ideas on how to do it?

- Consider we *know* all active `.nl` domain names
- We need the `html` `<title>` tags from 5.8 million domains
- **How to get that information?**
  - Luckily, we had Dmap (`https://dmap.sidnlabs.nl`)
  - Crawls HTTP, DNS, TLS, MX and screenshot
  - Run once a month on SIDN

Can we used that to detect these shops?

Any ideas on how to do it?

- Consider we *know* all active `.nl` domain names
- We need the `html` `<title>` tags from 5.8 million domains
- **How to get that information?**
  - Luckily, we had Dmap (`https://dmap.sidnlabs.nl`)
  - Crawls HTTP, DNS, TLS, MX and screenshot
  - Run once a month on SIDN

Can we used that to detect these shops?

Any ideas on how to do it?

- Consider we *know* all active `.nl` domain names
- We need the `html` `<title>` tags from 5.8 million domains
- **How to get that information?**
  - Luckily, we had Dmap (`https://dmap.sidnlabs.nl`)
  - Crawls HTTP, DNS, TLS, MX and screenshot
  - Run once a month on SIDN

Can we used that to detect these shops?

You have all HTML titles... how would you detect these shops?

- Deploy the *most advanced* ML algorithm ever 🙂
  1. *count* the number of brands in the `html` `<title>`
- From a precompiled list with 1100 brands and discount words
- If has more than 5 words (arbitrary), marks as suspicious
- **For real?**

Can we used that to detect these shops?

You have all HTML titles... how would you detect these shops?

- Deploy the *most advanced* ML algorithm ever 🙂
  1. *count* the number of brands in the `html` `<title>`
- From a precompiled list with 1100 brands and discount words
- If has more than 5 words (arbitrary), marks as suspicious
- **For real?**

Can we used that to detect these shops?

You have all HTML titles... how would you detect these shops?

- Deploy the *most advanced* ML algorithm ever 🙂
  1. *count* the number of brands in the `html` `<title>`
- From a precompiled list with 1100 brands and discount words
- If has more than 5 words (arbitrary), marks as suspicious
- **For real?**

Can we used that to detect these shops?

You have all HTML titles... how would you detect these shops?

- Deploy the *most advanced* ML algorithm ever 🙂
  1. *count* the number of brands in the `html` `<title>`
- From a precompiled list with 1100 brands and discount words
- If has more than 5 words (arbitrary), marks as suspicious
- For real?

Can we used that to detect these shops?

You have all HTML titles... how would you detect these shops?

- Deploy the *most advanced* ML algorithm ever 🙂
  1. *count* the number of brands in the `html` `<title>`
- From a precompiled list with 1100 brands and discount words
- If has more than 5 words (arbitrary), marks as suspicious
- **For real?**

BrandCounter

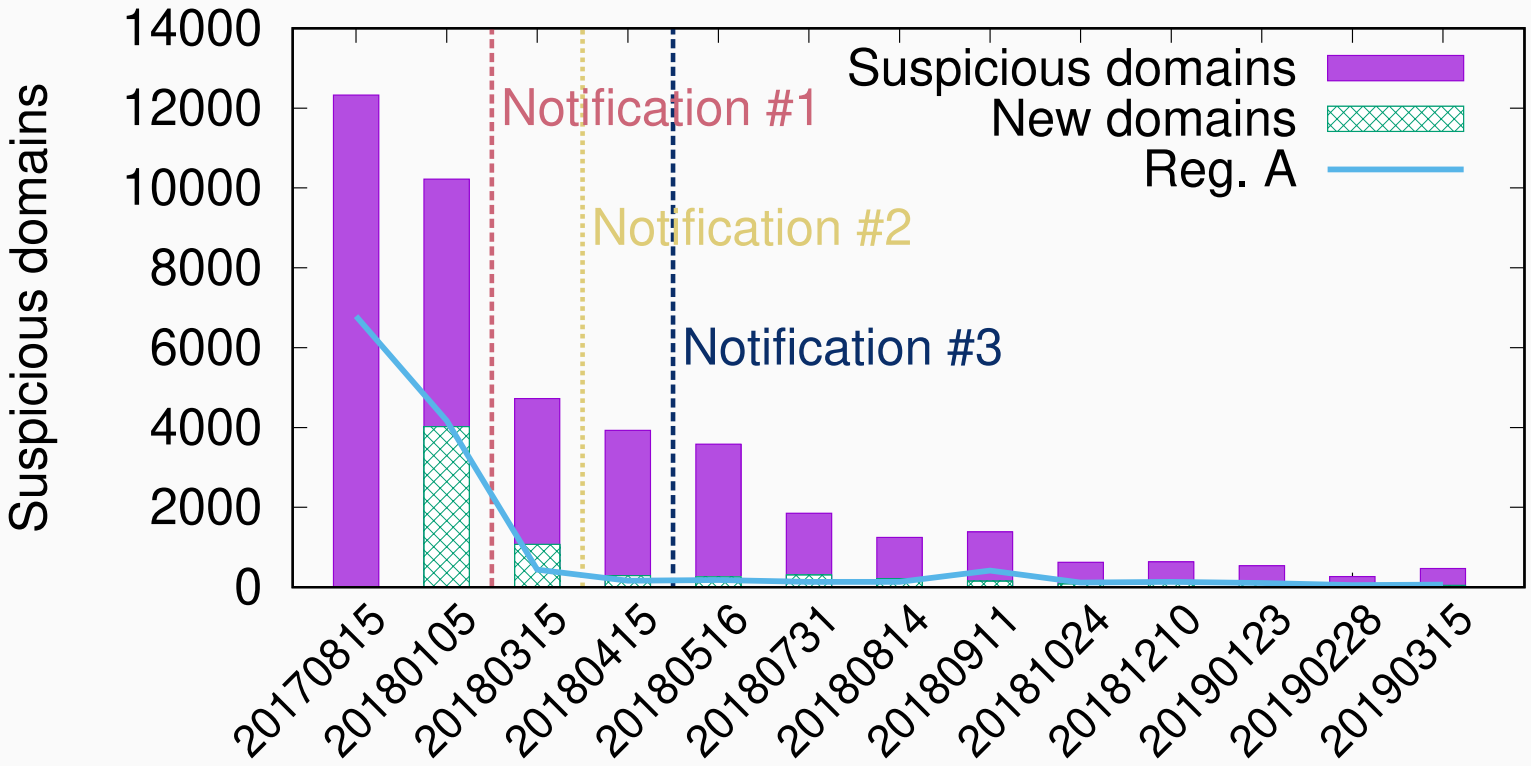
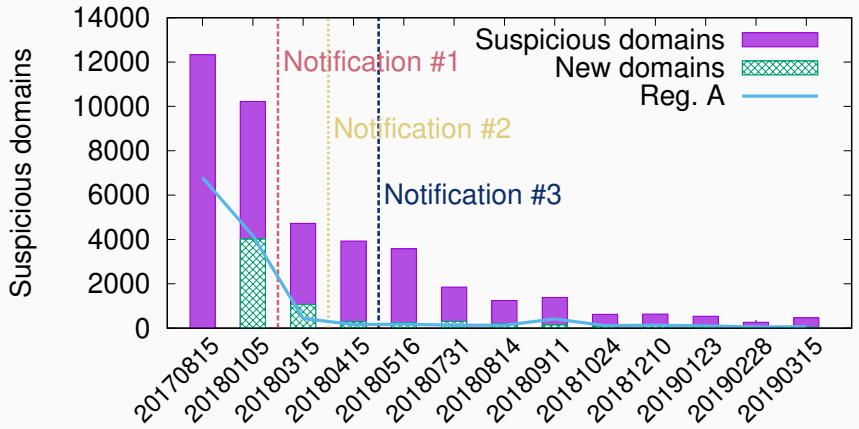


Figure 3: BrandCounter suspicious domain results for `.nl` zone.

BrandCounter

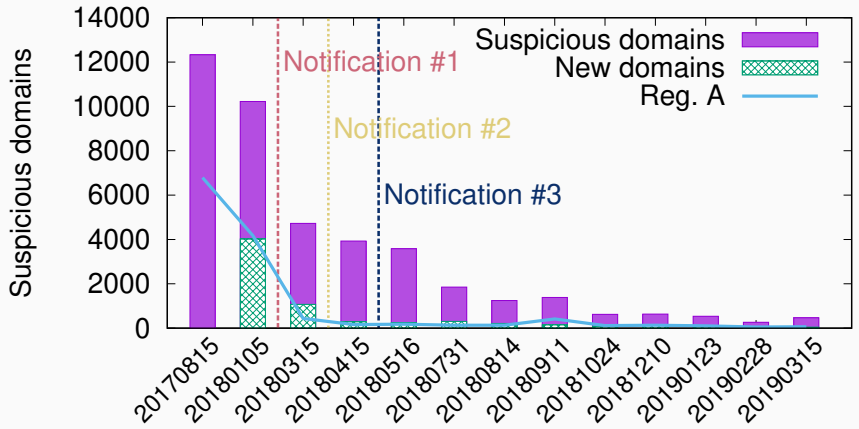


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter

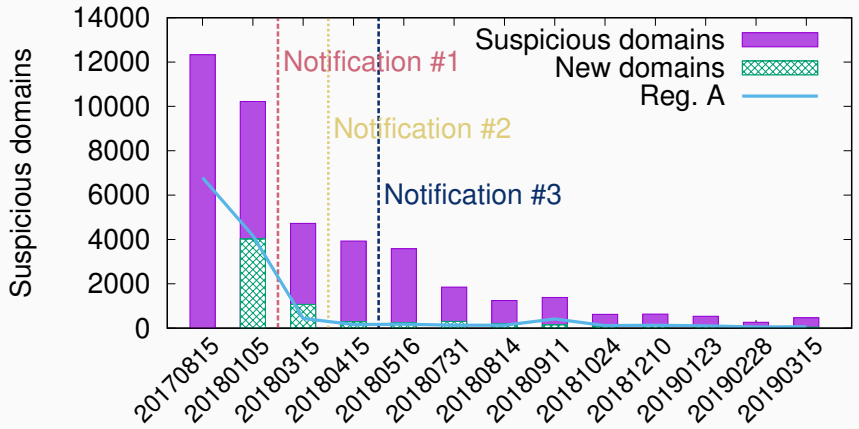


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter

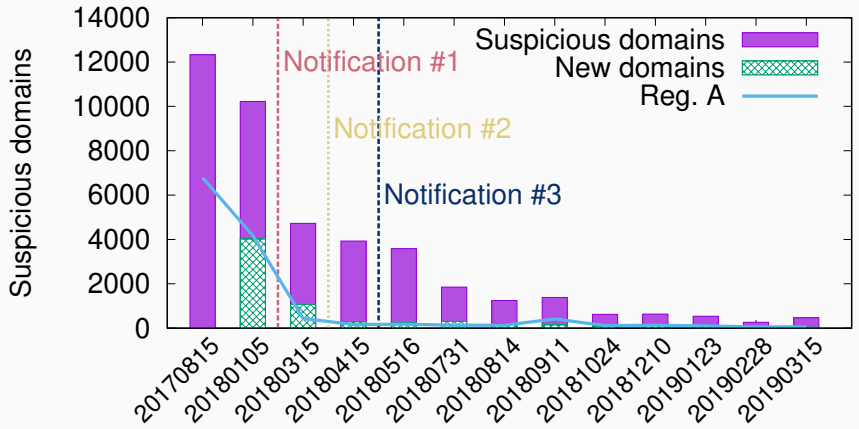


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter

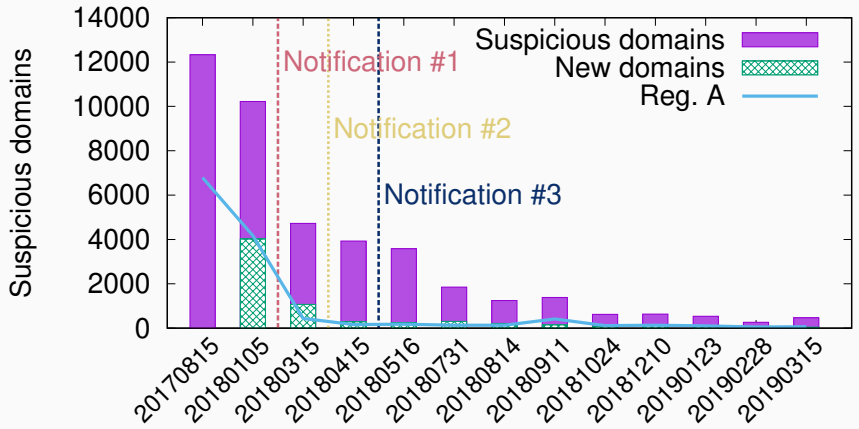


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter

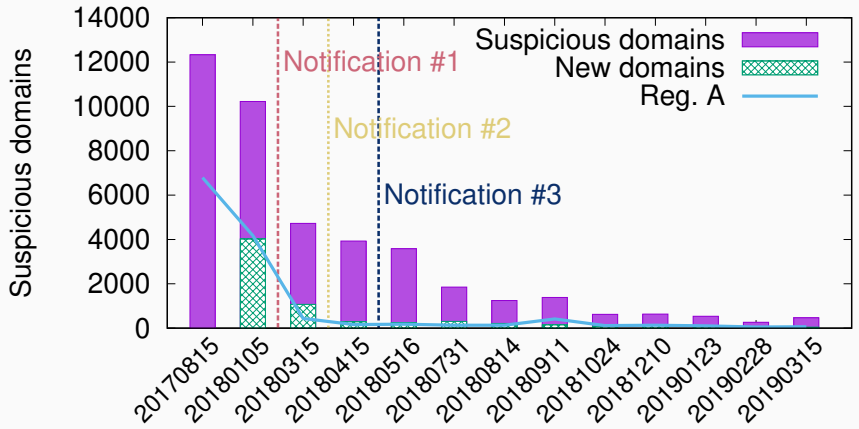


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter

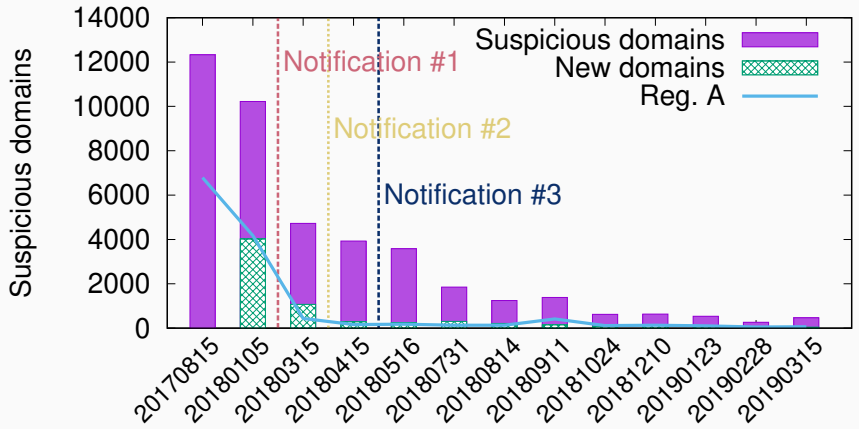


Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

BrandCounter



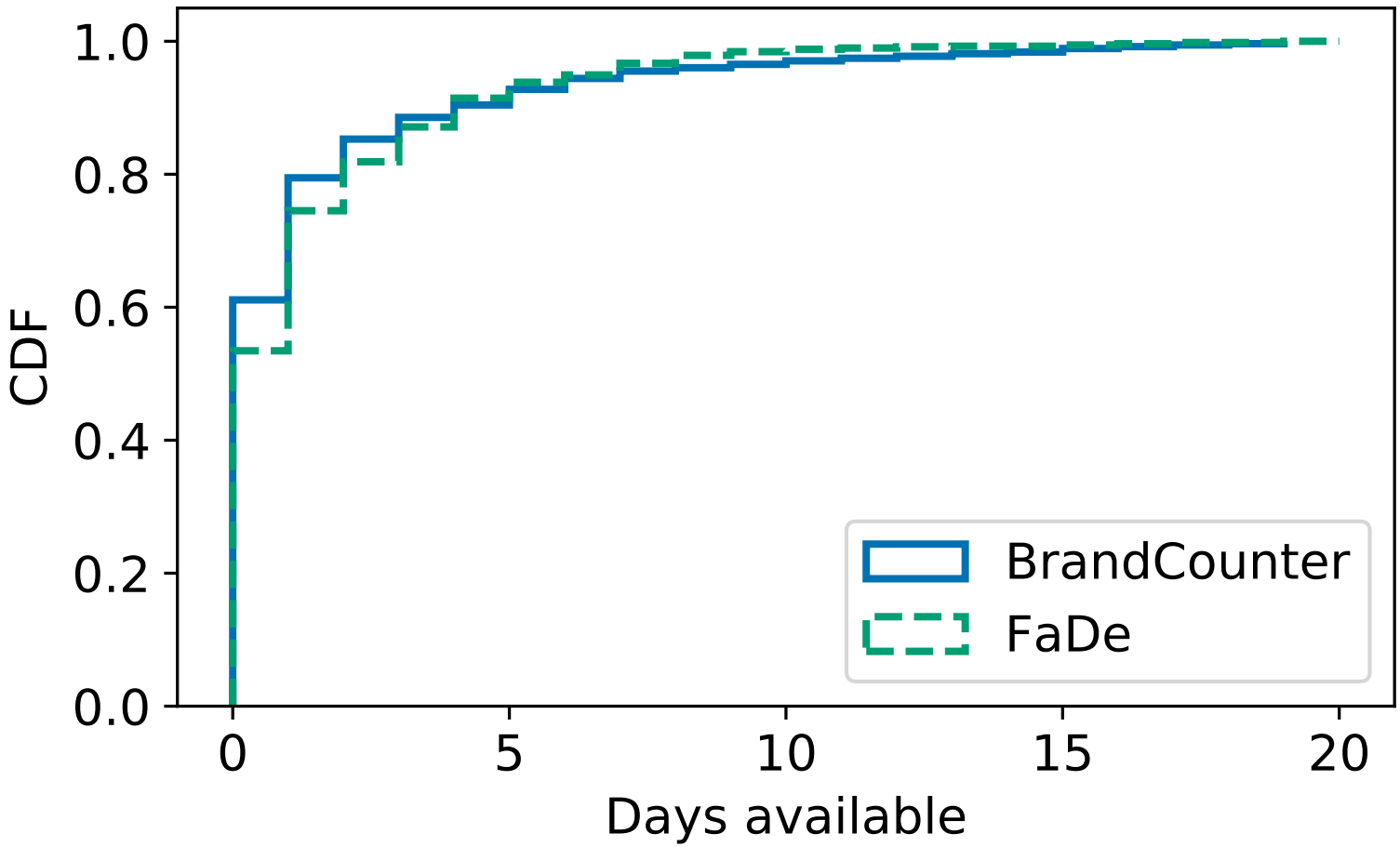
Wait, what?

1. How come does this even work?
   - This is to show they suffered little pressure
2. Why so many of these webshops?
   - it's unlikely there are that many counterfeiters
   - *Domains are cheap and disposable*
   - automation heavily used
   - 10 down does not even make a difference
3. Why 6K were registered with only one registrar?
   - API for automatic registration & good price

Take downs were effective, in partnership with our registrars

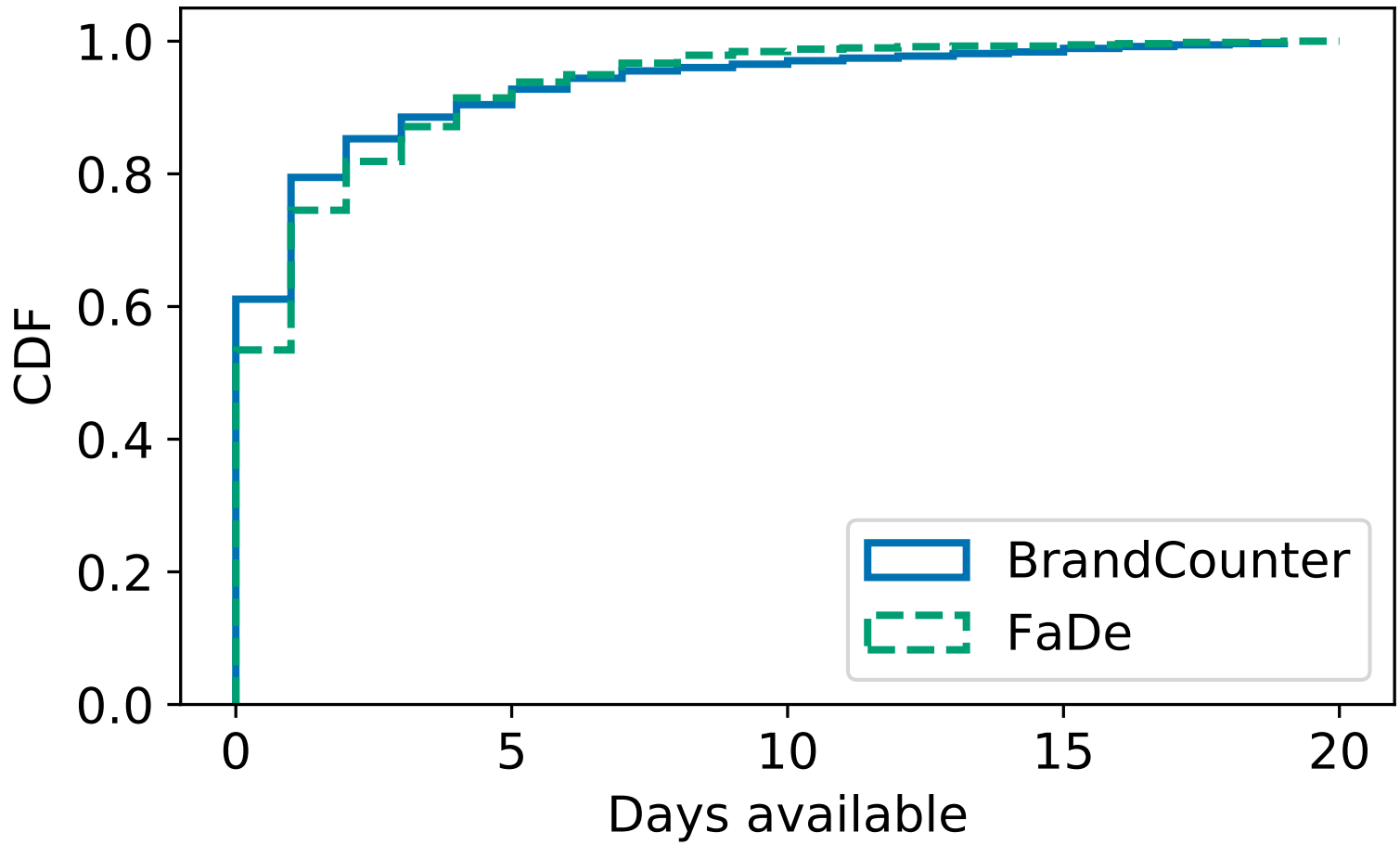
Automated Domain Registration

- **Question: why registering freshly released domains?**
  - Tap on their “residual reputation” [2]
  - And on their previous traffic



Automated Domain Registration

- **Question: why registering freshly released domains?**
  - Tap on their “residual reputation” [2]
  - And on their previous traffic



Automated Content Creation

- use of CMSes
- Automate page creation and hosting
  - Same CMS
  - None support HTTPs
  - Same ugly payment methods image – with all CC flags
- It’s economics: create 100s of pages, a few sales cover the costs

Registrations point to China

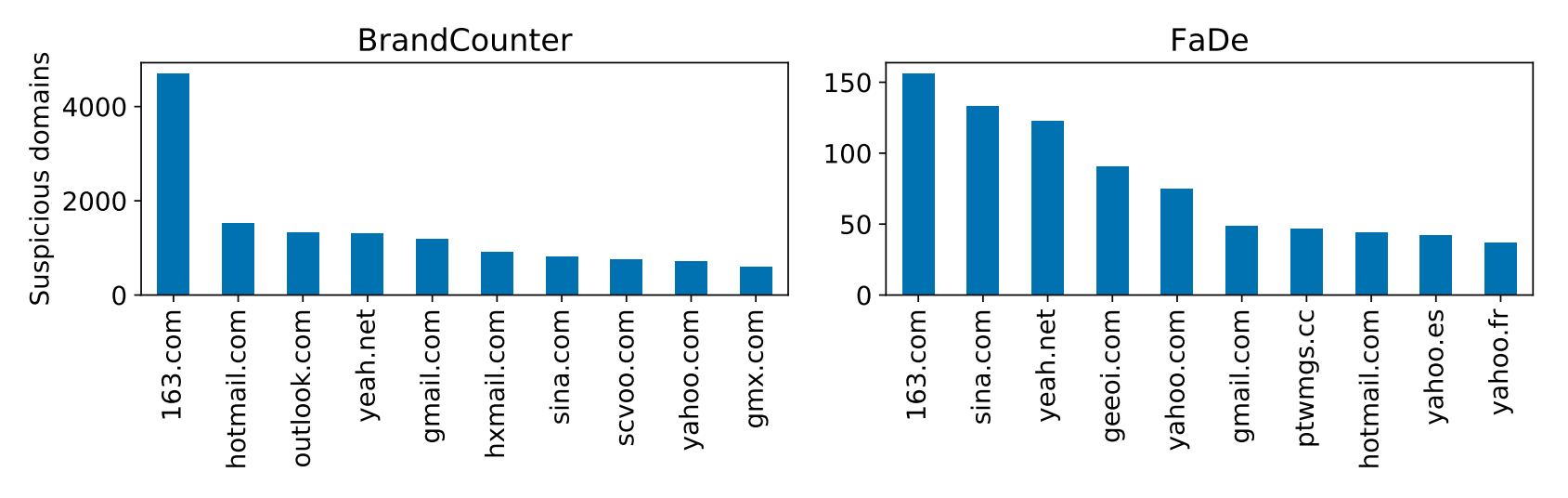


Figure 5: Number of shops by the registrant's e-mail domain.

Registrar A notification

- Roughly 50% of domains registrars with one registrar
- We teamed up with them and notified then of scans
- They had the power to contact registrants
- And took down thousands of domains
- We (SIDN) cannot legally take down such domains directly

Date	Domains	Suspended-NS	Online
2018-01-18	3560	3174 (89.16%)	386 (10.84%)
2018-03-16	399	387 (97.24%)	12 (3.02%)
2018-05-02	148	147 (99.32%)	1 (0.68%)
Total	4107	3708 (90.31%)	398 (9.69%)

Table 1: Registrar A notification and suspension results.

So problem solved?

We can all go home, everyone's safe now

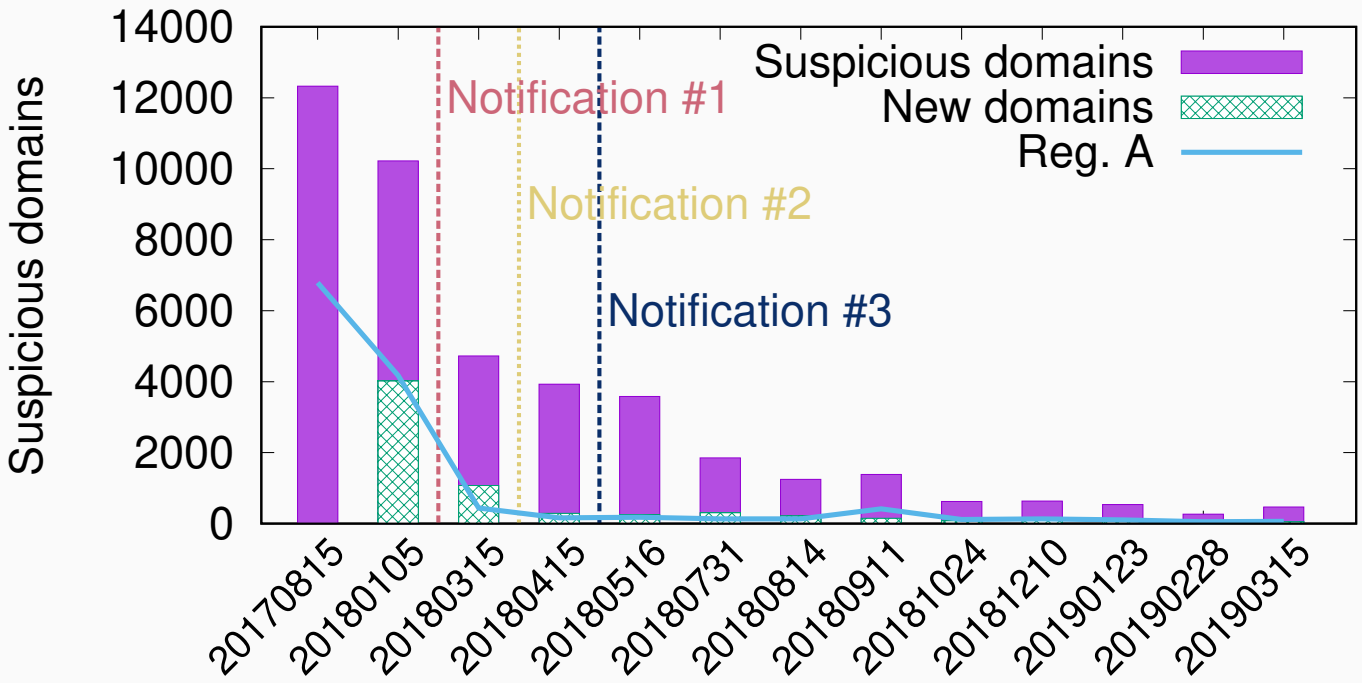
Right?

So problem solved?

We can all go home, everyone's safe now

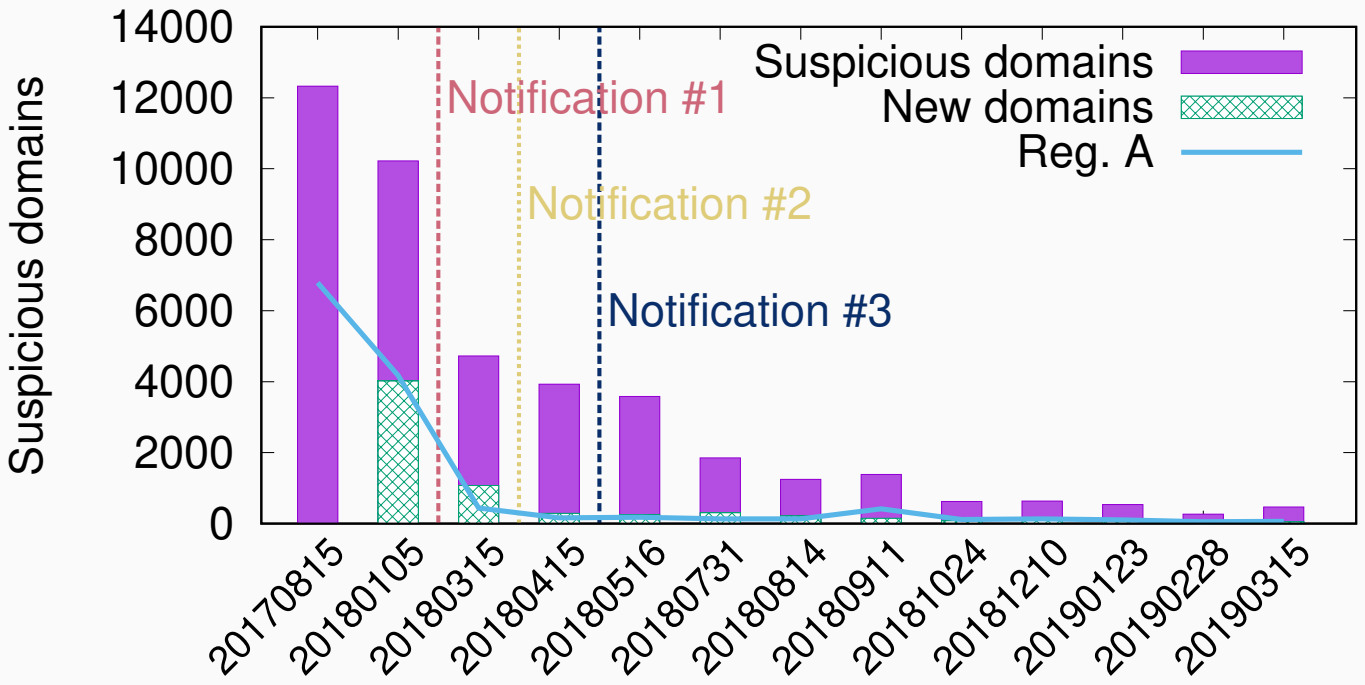
Right?

Looking closer



Were the **counterfeiters gone** or have they learned to dodge BrandCounter?

Looking closer



Were the **counterfeiters gone** or have they learned to dodge BrandCounter?

Security is always a cat-and-mouse game

- To be sure, we had to come up with a new classifier
- We teamed up with ICS, a credit card issuer in the Netherlands
  - `https://www.icscards.nl`
- They gave us a list of 231 .nl shops involved in scams
  - so we had some **ground truth**
  - **how would you use it?**
  - Use some supervised learning method on it – training set vs test set
  - Then use the output model to detect other shops

Security is always a cat-and-mouse game

- To be sure, we had to come up with a new classifier
- We teamed up with ICS, a credit card issuer in the Netherlands
  - `https://www.icscards.nl`
- They gave us a list of 231 .nl shops involved in scams
  - so we had some **ground truth**
  - **how would you use it?**
  - Use some supervised learning method on it – training set vs test set
  - Then use the output model to detect other shops

Datasets we have

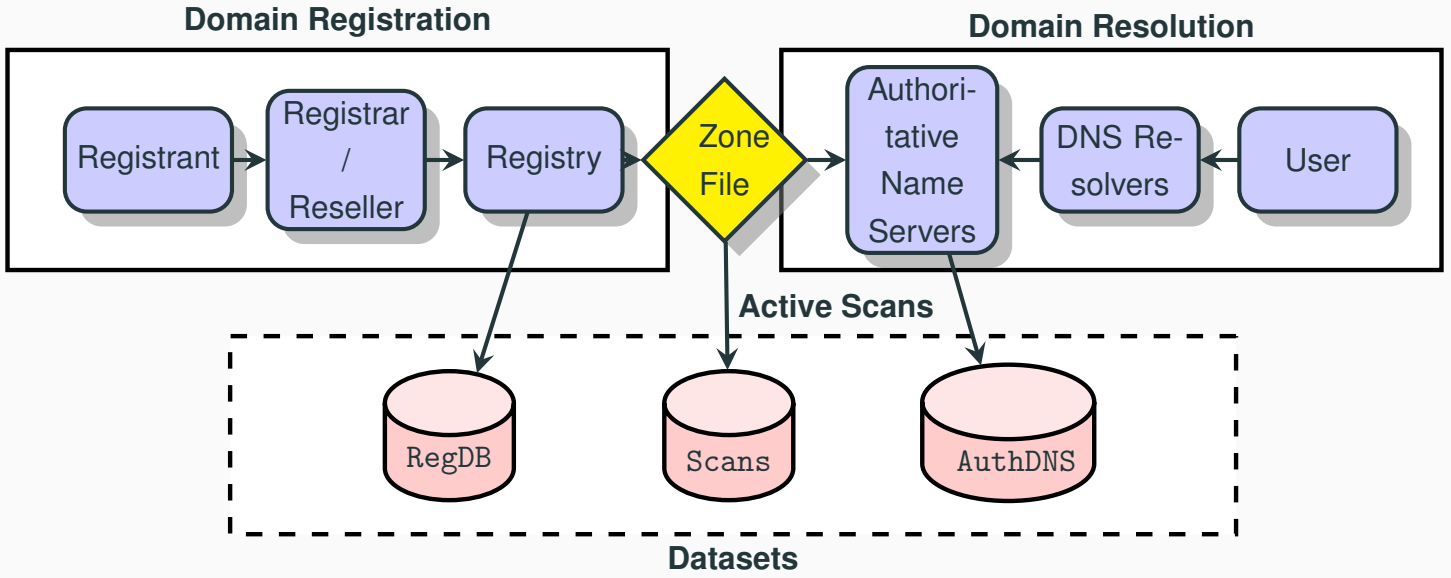


Figure 6: TLD operations: registration (left), domain resolution (right), and datasets.

Algorithm and Features

- We use Support-Vector Machine (SVM) algorithm
- Goal: classify domains into two classes:
  - "fake" webshops
  - non-fake webshops
- Math behind SVM

Training set:

- fake webshops: 231 domains provided by ICS
- non-fake: 229 random .nl domains (we manually verified btw)

Features for FaDe

Dataset	Feature	Importance
RegDB	1. Re-registration	2
	2. Registration Hour	4
	3. Registrar	6
	4. Suspicious e-mail provider of registrant	1
	5. Reported domains score	5
	6. Registrant name lowercase	9
Scans	7. Existence of a MX record	3
	8. Issuer of TLS certificate (if any)	7
	9. Autonomous System of A Record	8

Table 2: Features used by FaDe. Features are all normalized (0 to 1) to ensure they have the same influence

Model Training

- Split the dataset into two (random):
  - Training set: 367 domains (80%)
  - Test set: 93 samples (20%)
- Use grid search to find optimals SVM parameters (kernel, C and γ).
- Use cross-validation
- Best parameters: $C = 10$ and $\gamma = 0.1$
- mean precision of 0.98 and mean recall of 0.97

Applying the model

- .nl zone has 5.8 million domains
- We could apply FaDe to it, but we can reduce the search space
- DMap classifies sites into categories: we choose e-commerce
    - 30k domains , far smaller
    - They have shopping carts, and other specific stuff

FaDe Results

Category	Domains
Suspicious	1407
Unreachable	181 (13%)
Reachable	1226 (87%)
True Positive	894 (73%)
False Positive	332 (27%)

Table 3: FaDe results and validation.

Registrar	Notified	Webshop-Down	NX-domain	NS-change
A	505	248	57	244
B	576	433	9	438
C	21	11	12	0
D	55	31	0	31
F	64	11	39	0
Others	63	13	16	0
Total	894	747 (84%)	133 (15%)	713 (80%)

Table 4: Notification and take down results.

FaDe vs BrandCounter

- BrandCounter shows crooks were suffering little pressure
- We apply BrandCounter to the shops of FaDe, and none has more than 5
- They learned to evade BrandCounter

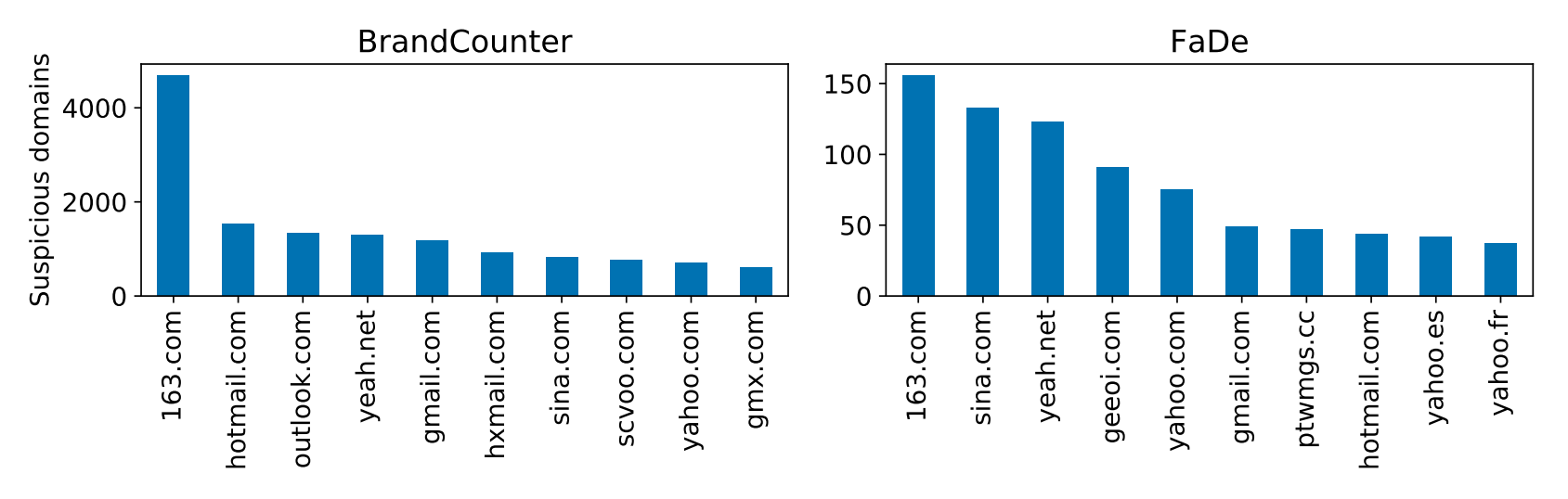


Figure 7: Number of shops by the registrant's e-mail domain.

Lessons learned

- In total, we helped to take down 4455 fake webshops over the last three years
  - Helping protecting .nl users, impact in real world

Net binnen Algemeen Economie Sport Tech Media en Cultuur Achterklap

Maandag 02 maart 2020 | Het laatste nieuws het eerst op NU.nl

Beheerder van het .nl-domein haalt 4.340 malafide webwinkels offline

17 februari 2020 17:20
Laatste update: 17 februari 2020 17:24

17

SIDN, de beheerder van het .nl-domein, heeft het afgelopen jaar 4.340 malafide webwinkels offline laten halen, maakt de stichting maandag bekend op zijn website. Volgens de organisatie gaat het vaak om websites met erg lage prijzen, gebrekkige vertalingen en zonder keurmerk

Lessons learned

- Registrars and ICS collaboration was key
- We already have a new system in place
- (demo)
- It's a ever going wack-a-mole game.
- Paper (PDF) at : https://tinyurl.com/yxx4tnwl
- **Multidisciplinary** example of security

Interested at SIDN Labs? https://sidnlabs.nl

- Internship, write your thesis, jobs
- Three SIDN Labs folks at DACS

Lessons learned

- Registrars and ICS collaboration was key
- We already have a new system in place
- (demo)
- It's a ever going wack-a-mole game.
- Paper (PDF) at : https://tinyurl.com/yxx4tnwl
- **Multidisciplinary** example of security

Interested at SIDN Labs? https://sidnlabs.nl

- Internship, write your thesis, jobs
- Three SIDN Labs folks at DACS

References i

[1] Giovane C. M. Moura, Moritz Muller, Maarten Wullink, and Cristian Hesselman.

nDEWS: a New Domains Early Warning System for TLDs.

In *IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2016), co-located with IEEE/IFIP Network Operations and Management Symposium (NOMS 2016)* (April 2016).

References ii

[2] Lever, C., Walls, R., Nadji, Y., Dagon, D., McDaniel, P., and Antonakakis, M.

Domain-z: 28 registrations later measuring the exploitation of residual trust in domains.

In *2016 IEEE Symposium on Security and Privacy (SP)* (May 2016), pp. 691–706.

[3] Schmidle, N.

Inside the Knockoff-Tennis-Shoe Factory - The New York Times.

http://www.nytimes.com/2010/08/22/magazine/22fake-t.html, 2010.

References iii

[4] Wall, D. S., and Large, J.

Jailhouse frocks: Locating the public interest in policing counterfeit luxury fashion goods.

The British Journal of Criminology 50, 6 (2010), 1094–1116 – http://ssrn.com/abstract=1649773.

[5] Wang, D. Y., Der, M., Karami, M., Saul, L., McCoy, D., Savage, S., and Voelker, G. M.

Search + seizure: The effectiveness of interventions on seo campaigns.

In *Proceedings of the 2014 Conference on Internet Measurement Conference* (New York, NY, USA, 2014), IMC '14, ACM, pp. 359–372.