SIDN Labs

https://sidnlabs.nl

February 25, 2020

Peer-reviewed Publication

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

Authors: Thymen Wabeke, Giovane C. M. Moura, Nanneke Franken and Cristian Hesselman

Venue: In: Proceedings of the Passive and Active Measurement Workshop (Eugene, OR, USA, Mar. 2020)

DOI: TBD
Conference dates: March 30-31, 2020

Citation:

- Thymen Wabeke, Giovane C. M. Moura, Nanneke Franken and Cristian Hesselman . Counterfighting Counterfeit: detecting and taking down fraudulent webshops at a ccTLD. Proceedings of the Passive and Active Measurement Workshop (Eugene, OR, USA, Mar. 2020
- Bibtex:

```
@inproceedings{Wabeke20a,
  author = {Wabeke, Thymen and Moura, Giovane C. M. and
  Franken, Nanneke and Hesselman, Cristian},
  title = {{Counterfighting  Counterfeit:  detecting and taking down
  fraudulent webshops  at a ccTLD.}},
  booktitle = {Proceedings of the Passive and Active  Measurement Workshop},
  year = {2020},
  address = {Eugene, OR, USA},
}
```

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

Thymen Wabeke[1], Giovane C. M. Moura[1,3],
Nanneke Franken[2], and Cristian Hesselman[1,4]

[1] SIDN Labs, Arnhem, The Netherlands
[2] SIDN, Arnhem, The Netherlands
{firstname}.{lastname}@sidn.nl
[3] TU Delft, Delft, The Netherlands
[4] University of Twente, Enschede, The Netherlands

Abstract. Luxury goods such as sneakers and bags are in high demand. Many websites offer them at high discounts, which, in many cases, are simply cheap counterfeit versions of the original product. Online shoppers, however, may be unaware they are buying a counterfeit product and end up being scammed and having to deal with financial losses, as has been widely reported by various news outlets. This work presents a multiyear effort of The Netherlands' `.nl` country-code top-level domain (ccTLD) in detecting and removing counterfeit online shops from the `.nl` DNS zone. We have developed two detection systems and partnered with registrars and a large credit card issuer, which ultimately led to more than 4,400 counterfeit online shops being taken down.

1 Introduction

Counterfeit or fake goods are unauthorized replicas of products that attempt to pass as legitimate ones. They cover a large array of goods, such as pharmaceuticals [17], electronics [1], aircraft parts [37], and books [31].

Luxury goods, from brands such as Nike and Louis Vuitton, are among the most popular counterfeit products. Their popularity originates from the consumer's high demand, leading to high-profit margins [37] for those who sell them. In the U.S. alone, seizures at the border of counterfeit goods in 2017 had an estimated value of US$1.2 billion [36], had these products been genuine. In the EU, 2016 border seizures were valued at €670 million (US$ 743 million) [33]. In both cases, most shipments originated from China, which has been also found as a major source of counterfeit shoes [28].

To be able to sell online, counterfeiters first have to attract potential buyers, and they have been using various tactics. In a previous study, Wang *et al.* [38] have shown how counterfeiters often employ search engine optimization (SEO) in an attempt to improve rankings in search engines. In addition, social networking websites have been also employed [22]: in 2016, a large number of Instagram accounts were dedicated to disseminating counterfeit luxurious goods—roughly 20% of the 150k analyzed posts [35], which contained links to stores dedicated to selling these type of products. Last, market places such as Amazon [31] and Ebay have been exploited by counterfeiters.

Buyers of these goods are often *unaware* that they are buying from a counterfeit webshop and in many cases, they end up not receiving any product, or receiving a lower quality version—being scammed either way. Moreover, they may become victims of ID theft, given that they have to provide their credit card details and address information. Financial losses to online shoppers have also been widely reported by several media outlets in The Netherlands [21–23], where they have been known to exist since 2016 in the `.nl` zone (Figure 11 and Figure 12 in §A and [5]). This is not only observed in `.nl`: Germany's `.de` was found to have more than 16.000 counterfeit shops, many active for several years [24].

In this paper, we focus on a subset of the counterfeit industry—the so-called *luxury goods* that are sold *online*, that often leads to shoppers experiencing financial loss. We leverage our centralized vantage point as the country-code top-level domain (ccTLD) registry for The Netherlands (`.nl`), operated by SIDN [30]. Centralized, in this context, refers to access we have to historical registration data of all `.nl` domain names, which also includes registrants' contact details. Given that most webshops in The Netherlands are registered under the `.nl` ccTLD (and are available in Dutch language), counterfeiters would have incentives to register their domains under `.nl` as well, to mimic what most legitimate webshops do. As such, our centralized vantage point allow us to leverage this strong association between ccTLD, country, and language.

This paper presents the results of a multiyear effort in detecting such webshops, which led to 4455 domain names being removed from the `.nl` zone. We present two detection systems—BrandCounter (§3) and FaDe (§4), which have been used in production over the past three years by our Abuse Handling Analysts to evaluate `.nl` domain names and notify registrars and/or registrants. BrandCounter, the first system from 2017, employs a very *simple* but *effective* heuristic. We used its results in a case study with Registrar A, that ultimately led to the removal of ∼3.7k counterfeit webshops from the `.nl` zone (§3.1). FaDe (§4), in turn, was developed in early 2019 to cope with the new tactics employed by counterfeiters (§4.2), who adapted after the initial take downs based on BrandCounter's results. We carried out another case study with the results from FaDe together with International Credit Cards (*ICS*, [11]), a major credit card issuer in The Netherlands with more than 3.5 million clients. This study led to the removal of an additional 747 domain names (§4.1). Lastly, we infer the popularity of the counterfeit domains among users by analyzing the volume of DNS queries to the `.nl` authoritative servers (§5).

2 Background

Domain name registration: Registering a domain name is the process of creating a unique name that is added to a DNS zone file. Next, we describe this process under `.nl`. It usually involves a *registrant*, *registrar* (or reseller), and *registry*. The registrant (a user) requests an accredited registrar to register an available domain name at the registry. The registrar only executes this request once certain requirements are met, such as registrant information and payment being cleared, as shown in the left part of Figure 1.

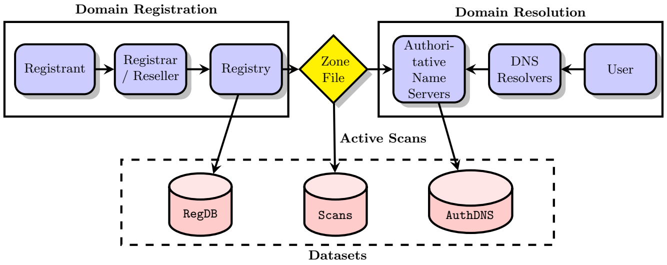

Fig. 1. TLD operations: registration (left), domain resolution (right), and datasets.

Domains are registered for a period of one year, which will be automatically renewed at `.nl`. If the domain is cancelled, it will expire and is put on hold for 40 days and right after that made available for a new registration by any registrant. The list of valid domain names is then used to generate a DNS *Zone File* (Figure 1) that contains the list of all domains under `.nl`, and their respective DNS records. These Zone Files are used as input on the *authoritative name servers*, which are used to answer queries on `.nl` domain names.

Domain name resolution: Domain name resolution consists of resolving a domain name into, ultimately, its IP address or other specific types of DNS records [18]. To do that, a user's application contacts the stub DNS resolver (Figure 1) on his/her computer, which, in turn, sends a DNS request to its DNS *resolver* [10]. The DNS resolver will, on behalf of the user, recursively resolve the requested domain name, and ultimately contact the appropriate authoritative name server. Caching on DNS resolvers [19, 20] is used to eliminate frequently issued queries, improving response times.

2.1 Datasets

We leverage three types of datasets available at the `.nl` registry. Two of them are passive data, while one is obtained through active measurements:

- `RegDB`: We have access to the historical database of registration and removal of `.nl` second-level domains (such as `example.nl`), which covers 20+ years. This dataset contains complete information about registrant and registrar (and resellers, if applicable), as well as some of the DNS records of the respective domains [18].

- `Scans`: We crawl all domains under `.nl` on a monthly basis. We scan for four types of application: DNS records, HTTP pages, SMTP and TLS (and its certificates on web pages). We employ DMap [40], an application we have developed to carry out these scans. Besides that, the `.nl` zone is scanned daily by OpenIntel [26], a research project that crawls daily multiple TLD zones for various DNS record types.
- `AuthDNS`: We have access to historical query data from two out of the four authoritative name servers for `.nl`. This data provides a centralized but sampled view (due to caching on the resolvers) of all queries issued to `.nl`. We use our open-source Hadoop-based ENTRADA [41] to store and process this dataset.

3 BrandCounter

While detecting phishing domains in the `.nl` zone in 2016 [5], we came across the first suspicious luxury goods webshops, which advertised goods at high discount, as shown in Figure 11, in §A. Upon inspection, we observed that they shared one common feature: long page titles (HTML element `<title>`) that listed a series of luxury brands—in an attempt to improve rankings on search engines [38].

That provided us with a simple but effective way to detect such shops in the entire `.nl` zone: we crawl the zone for web pages and, for each page, we compare how many words in the page title match the words from our 1,100+ precompiled list of luxury brands and discount-related words, such as "discount", "sale", in both English and Dutch. We determined empirically a threshold of $t >= 5$ matching words to classify webpages as suspicious. We automated this process into a single tool (BrandCounter), and ran it roughly once a month, for over 1.5+ years, as shown Figure 2. In total, BrandCounter detected 18952 suspicious webshops.

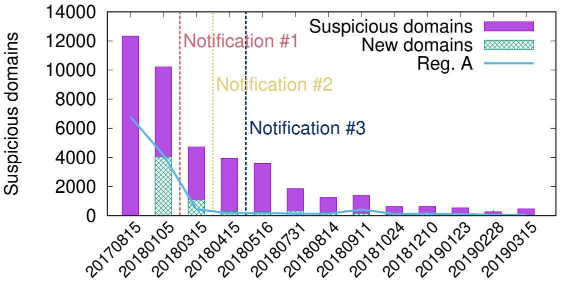

Fig. 2. BrandCounter suspicious domain results for `.nl` zone.

Results and Analysis: Eighteen thousand allegedly counterfeit webshops seems like a large number—0.3% of the entire of `.nl` zone. We analyzed these domains, and observe the following characteristics:

Domains are cheap and disposable: Given that it is relatively cheap to register a `.nl` domain (less than €10 in 2020), counterfeiters may choose to register a large number of domains, and even if some are taken down, the profits made from the remaining ones are enough to sustain the operation. The relatively short lifetime also indicates that domains are disposable (Figure 6).

Registrar concentration: Out of 18952 domains, 16512 are registered by 10 registrars, as can be seen in Figure 3. The top registrar—Reg. A—is alone responsible for 8017 (42.3%) of all detected shops. One of the reasons for that may be the fact that Reg. A ranks among the cheapest registrars and provides an API that allows for bulk registration of domains, which is very handy in case of automated registrations. Given such concentration, we carried a case study with Reg. A (§3.1), in which a large part of these domains were suspended.

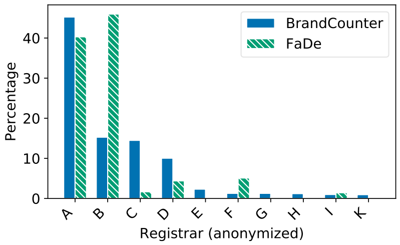

Fig. 3. Top 10 registrars with suspicious domains.

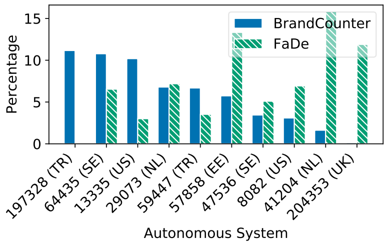

Fig. 4. Top 10 ASes (countries) hosting suspicious domains.

Similar but yet different website templates: We analyzed the home pages of some of these webshops and found out that they are different, but seem to be using a few content-management systems (CMS). The webshops do not support HTTPS, and have a single image in the page footer that contains icons of most credit card companies with no link or a broken link. Such designs also suggest use of automated tools to create such websites. Wang *et al.* [38] describe many *doorway pages*, which are non-shopping sites that are specifically designed to improve SEO results and *redirect* users to the real websites. In our work we do not see such pages since we do not rely on search engine results—we see the actual automatically generated pages listing the counterfeit goods, always with large discounts.

Most domains were drop-catch: 15242 shops are hosted on domains that expired and were re-registered by the counterfeiters (80.4%). The majority of these domains are immediately registered when they became available (Figure 5), a practice known as "drop-catch" [7]. By registering freshly expired domains to host conterfeit webshops, counterfeiters can benefit from their previously built reputation [14]. This timely precision in registering domains—and the fact that they seem indifferent to the name of the domain itself, as many were previously used by small businesses such as bakeries, beauty parlors—supports the idea of automation in the registration process.

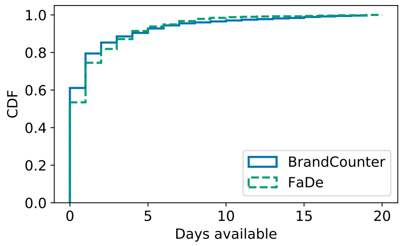

Fig. 5. Suspicious domains: days in between domain expiration and re-registration.

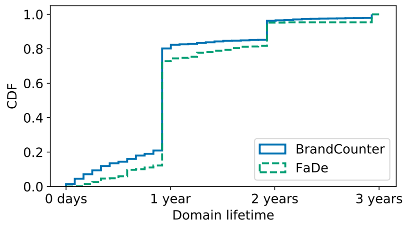

Fig. 6. Suspicious domains lifetime: most domains are not renewed after one year—the registration period.

Chinese e-mails and Chinese diurnal registration timing: Registrants are required to provide their e-mail address to register a domain with `.nl`. Out of 18925 suspicious domains, 4696 are registered using `163.com` (24.81%), a well-known Chinese e-mail provider which is particularly not popular in The Netherlands (Figure 7). Moreover, the registration diurnal patterns coincide with east China working hours (Figure 8).

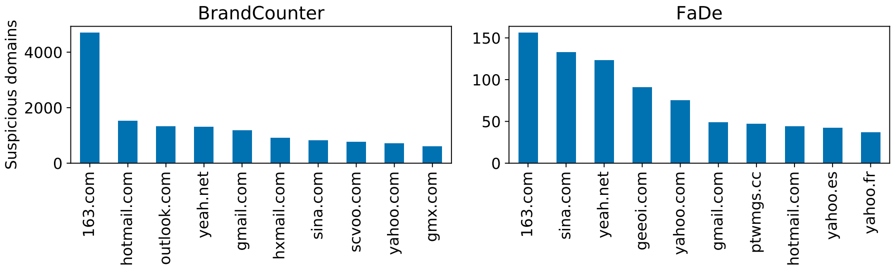

Fig. 7. Number of shops by the registrant's e-mail domain.

Hosting provider concentration: We see that 66.59% of the counterfeit webshops are hosted in 10 ASes—as can be seen in Figure 4—and none of them are located in China. We also see that most of them, however, use default DNS services provided by their registrars during registration. We inspected a sample of websites from the `.com` zone hosted under some of the same IP addresses of AS197328. Some were counterfeit webshops in other languages, but we also found websites that seemed legitimate, such as small businesses in Turkey.

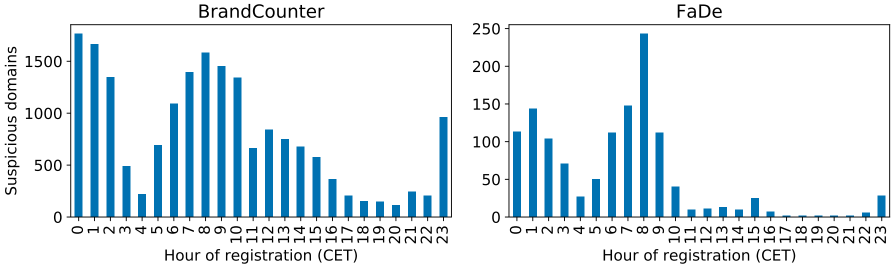

Fig. 8. Number of shops by registration hour.

3.1 Registrar Notification Case Study

Counterfeiters employed Registrar A to register 8017 suspicious domains (Figure 2), from the more than 15k detected, for the entire period covered. Given this concentration, we partner with Reg. A in a case study of three months in which we provided them with a list of domains that were labeled as suspicious by BrandCounter. In these three months, we sent 4106 domains to Reg. A.

Reg. A, in turn, would verify the identity of the registrants and take appropriate measures according to their regulations. While other registrars were also notified—-and many also removed suspicious domain names—we single out Reg. A in this section, because we only tracked results for this registrar.

Table 1 shows the number of domains we notified to Reg. A—more than four thousand in the three notifications. Upon receiving the list of domains, Reg. A determined the accuracy of the registrant data and judged each domain individually. The column "Suspended" shows the number of suspended domains by Reg. A—meaning they changed their NS records to sinkhole-like authoritative name servers (*e.g.,* `sinkhole.example.nl`), which they typically use for their suspended domains. To determine when the suspension occurs, we use daily crawls provided by OpenIntel [26].

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.

We can see the effects from this notification in Figure 2: first, a drop in the number of domains labeled as "suspicious" originated from Reg A, followed by an overall drop of detected suspicious domains. We also see the effectiveness of this intervention in the same figure: as domains started to be suspended by registrars, and we see a drop in the number of domains classified as suspicious. After October 2018, we see very little change in the volume of such domains. Overall, our notification study lead to more than 3708 domains being ultimately suspended by Registrar A, potentially protecting users from scams.

4 FaDe

BrandCounter was initially effective in detecting counterfeit webshops, but after a first round of takedowns, we observed a sharp decrease in the number of suspicious domains (Figure 2). Why was that? Have the counterfeiters given up or have they learned to avoid detection by BrandCounter? Given that, we set out to develop a new detector FaDe—Fake Detector—which does not rely on the words in the web page title. Instead we utilize a Support Vector Machine (SVM) [32] that employs nine features related to the registration itself and the infrastructure. We chose SVM because it is a robust method that has been successfully applied to classify various types of malicious activity [4, 12, 13].

SVM is a supervised learning method and relies upon labeled data for training. For that, we collaborated with the Abuse Department of *ICS*, a major credit card issuer in The Netherlands. *ICS* provided us with a list of 231 `.nl` domains labeled as fraudulent (Nov 2018 – Jan 2019). We also randomly sampled 229 webshops from our zone which we manually labelled as a trustworthy webshop. This resulted in a data set of 460 samples.

Feature selection: We employ nine features in FaDe that characterize counterfeit webshops (Table 2). The first three were inspired on the work by Hao *et al.* [6]—which we also observed with BrandCounter (§3). Re-registration indicates if the domain has been previously registered or not, Registration Hour represents the *hour of the day* in which the domain was registered, and the third was the registrar used.

The remaining six features (highlighted in Table 2) are based on other patterns we have seen with the domains detected by BrandCounter (§3) and the training set provided by ICS. E-mail provider indicates whether a suspicious e-mail domain is used by the registrant, given we have seen a high concentration of unusual mail providers (Figure 7). The fifth feature—reported domains score—is the ratio of malicious domains reported via the Netcraft abuse list [15] divided by all the domains registered by a given registrar in 2018 on the `.nl` zone. The sixth feature captures the ratio of lowercase characters in the registrant's name, given that we noticed that many counterfeit webshops register with lowercase only. We observed that 227 of the 231 webshops reported by ICS did not configure mail servers (defined by their MX record [18]), which we also then use as a feature. The eighth feature is the issuer of the TLS certificate, because we observed that 3 issuers are responsible for 156 of the 183 webshops that were labeled by ICS as fraudulent and have TLS configured (websites have also been found employing TLS [27]). Finally, we consider the autonomous system of the A record and of the domain, *i.e.,* the AS of the hosting provider, given the high concentration of certain ASes (Figure 4). All features are normalized to the same scale ($[0, 1]$) to ensure they all have the same influence on the distance metric.

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.

Model Training: To train our model, we start by randomly splitting our dataset with 460 samples into two categories: training set (367 samples, 80%) and test set (93 samples, 20%). We then use grid search [2] to find the optimal SVM parameters (*i.e.,* kernel, C and γ). We employ cross-validation [8] so that we can use the full training set for both training and validation. The best scores over all folds—mean precision of 0.98 and mean recall of 0.97—were obtained using the RBF kernel with $C = 10$ and $\gamma = 0.1$. Next, we train our final model using these parameters and the full training set. This model was then applied to the test set yielding a precision of 1.00 and recall of 1.00. Although the test set is small, it at least indicates that our model performed well.

Feature importance: To estimate feature importance, we use the coefficients of the best SVM classifier with a linear kernel. We omit the exact coefficients because we do not want to help counterfeiters with exact values and show the relative importance in Table 2.

Results and Analysis After training our model, we apply it to a subset of the `.nl` zone: only domains that are automatically classified as *eCommerce* by our crawler DMap [40]. We focus on this subset to prevent many false positives that could discourage abuse analysts. For this purpose, the crawler extracts technologies used on webpages using Wappalyzer [39] and some regular expressions that look at specific HTTP headers, HTML content and cookies. A domain is classified as eCommerce if it has at least one eCommerce related technology (e.g., Zen Cart or WooCommerce).

Ultimately, we evaluated 30k domains of our zone that were classified as eCommerce and were registered at most 365 days ago, using data crawled in January 2019. Table 3 shows the results. In total, FaDe classifier detected 1407 suspicious domains.

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.

To validate the results, we shared the lists of suspicious domains with *ICS*, where analysts manually verified every single domain in the period between 2019-01-29 and 2019-02-04—including evaluating the payment provider used by the website. Out of the 1407 domains, 181 domains (Table 4) were not reachable anymore by the time of the validation—in 14 cases analysts report a DNS error, 167 domains are annotated with a generic 'no response' label which could indicate failure at the DNS or server level. This left us with 1226 domains that were both suspicious and reachable. Out of these, 894 were confirmed as true positives (72.92% precision). *ICS* analysts reported notes on a few false positives: 38 were redirects to legitimate webshops and 8 were adult websites.

4.1 Registrar notification and takedown

Being able to detect these counterfeit webshops is just the first step. To protect `.nl` users, we need to act upon these domains, and preferably take them down. We then split the true positives per registrar, as can be seen in Table 4, and notified the respective registrars of these domains, via two channels: *ICS* carried out their notifications and the registration department at SIDN also notified registrars. After receiving notifications, registrars can individually decide, according to their policies and processes, to suspend the domain—a process that we were not involved in.

To determine which domains were taken down, we could use the same approach shown in §3.1. However, different registrars may employ different take down methods: web page content changes, domain suspension, DNS records changes, among others. Given that we notify multiple registrars, we analyze changes in the content of web pages, domain cancellations, and nameserver changes in the period starting from the notification date until 01-05-2019. We use RegDB data and Scans data (§2.1) for this purpose.

Table 4 shows the results. Out of the 894 domains that we notified to registrars, 747 (83.56%) were effectively taken down, as measured by a change of webpage content. We can also see in the same table the method employed by the registrar: 133 (14.88%) domains are cancelled resulting in an NX domain and 713 (79.75%) changed their NS records [18], which point to the authoritative name server of a domain. We manually checked the name server changes. 677 domains changed to a sinkhole name server and 36 to a regular name server. This indicates that registrars employ different strategies to take down counterfeit webshops. For example, Reg. B suspended most domains by changing name servers whereas all Reg. F domains were cancelled. 147 (16.44%) of the notified domains were not taken down. In the majority of those cases the registrar did not respond and the registrant details were legitimate, giving us no ground to remove the domain from our zone.

4.2 BrandCounter vs FaDe compared: evolving tactics

Given that BrandCounter was effective with such a simple heuristic, we can deduct that counterfeiters were likely facing very little defensive pressure—they did not seem to make any efforts to hide the suspicious characteristics of their websites, or at least not in early 2017. We could expect counterfeiters to *adapt* to our detection methods, especially because thousands of domains were taken down.

To determine why BrandCounter's performance reduces over time (Figure 2), we apply BrandCounter to the true positives generated by FaDe. Out of the 894 domains, 707 had a score of 0 matching words—and no domain had a score above 3. Given we use a threshold of $t > 5$, counterfeiters evaded BrandCounter detection. In other words, they *adapted* to BrandCounter. Upon inspection, we see that they have essentially removed references to popular brands and inserted generic product titles, colors, type of garment, and targeted age group/gender, ultimately evading BrandCounter—which is surprising, given that up to that point we have not disclosed how we detected these websites.

Registrar and email provider diversification: We have shown in §3.1 how Registrar A took down more than 3700 domain names upon our notifications. We could expect counterfeiters to respond to that. We see that in Figure 3, in which registrar B becomes the number 1 registrar employed by counterfeiters. More prominently, we see a diversification of e-mail providers used by registrants (Figure 7)—moving from the dominant 163.com for BrandCounter detected domains, to a more diverse distribution for domains detected by FaDe.

Hosting diversification: We still observe that counterfeit webshops are hosted on a small number of ASes. However, the ASes themselves did change over the years as can be seen in Figure 4. AS 41204 and AS 204353 were frequently observed during the second study based on FaDe, while no shops were later hosted on AS 197328. Interestingly, the hosting infrastructure still does not map to Chinese IP addresses.

5 How popular are the counterfeit webshops?

Our notification campaigns led to 4.5k domains being removed or suspended. In this section, we explore the popularity of these counterfeit webshops.

We can indirectly infer a counterfeit webshop popularity by analyzing incoming queries for the `.nl` authoritative server—leveraging our `AuthDNS` dataset described in §2.1. For each domain name d, we extract the number of queries and unique IP addresses of resolvers we observed one week before the notification dates. (we chose one week given the known weekly diurnal patterns of Internet traffic [25]). While the number of queries and resolvers do not correspond to the number of unique shoppers (due to caching at DNS resolvers), it provides an indication of how diverse the population of the resolver is.

Figure 9 shows the average number of daily queries for the domains taken down before the notification, while Figure 10 shows the average daily number of resolvers. The baseline consists of a random set of 500k domain names that serve a website (defined by a `200 OK` HTTP status code). We see a significant discrepancy in counterfeit webshops popularity: 50% of them have, on average, 100 daily queries prior to the notification, from ~ 70 unique resolvers. However, there are *some* domains that are *very* popular: 55 domains had an average 1000 daily queries from 653 resolvers. We manually analyzed the queries of the top 10 counterfeit webshops and found that most queries originated from public resolvers and local ISPs, which is similar to normal query behavior. This suggests variability in domains' popularity, which may coincide with their advertisement strategies.

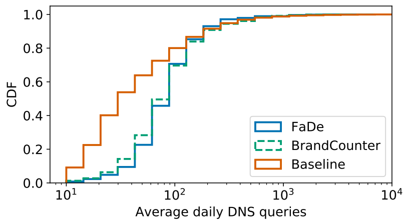

Fig. 9. Average number of daily DNS queries for counterfeit shops one week prior notification and a random subset of 500k domains that serve a website.

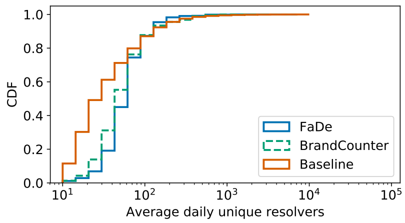

Fig. 10. Average number of daily unique resolvers for counterfeit shops one week prior notification and a random subset of 500k domains that serve a website.

6 Privacy and Legal Considerations

Together with our legal department, we have developed a publicly available data privacy framework [3] that conforms to both EU and Dutch [3,9] legislation. This framework has been implemented, including a privacy board that oversees SIDN

Labs' research. For the purpose of this research, only domain names and their associated labels—either legitimate or suspicious—were shared between SIDN Labs research and respectively *ICS* and the registrars. This collaboration was formalised using a data sharing agreement.

Note that domains with counterfeit webshops were mostly taken down by registrars. SIDN only takes down domains based on content if it is clearly criminal or unlawful. However, `.nl` regulations [29] determines that registrant data must be legitimate. Failure to conform to the regulation may result in domain name removal from the zone—the legal instrument that has been used in some take down procedures.

7 Related Work

Counterfeit market: Counterfeit industry has been previously studied by criminology researchers [37]. However, they focus on sales in the streets and not online. The online world of counterfeit stores have been extensively studied and mapped by [38]. The authors' starting point was Google search results. Our work, however, is based on 5.8M domains issued by `.nl`, and with a focus on non-English results. Besides, we cover years of continuous efforts to mitigate such webshops and we carry out notification campaigns with domain registrars and a credit card issuer, which lead to 4.5k domains being taken down (and more belonging to other registrars, which our colleagues of the registration department notified but we did not cover in this study). We also show how counterfeiters adapted to our first classifier, once their domains started being taken down.

Payment systems: McCoy *et al.* [16] cover payment systems in abuse-advertised goods, and in 2018 they focused on bullet-proof payment systems [34]. We do not cover payment systems in this paper, but we collaborated with *ICS*, which is a major credit card provider that deals with payment systems themselves.

8 Conclusions

Counterfeit luxury goods are a very profitable business, and employ high levels of automation in both registration and hosting. Our results suggest most registrations are supposedly done from China, but most hosting is not. We show that counterfeiters operate not only in English and in `.com`, as in previous works, but also in Dutch and on `.nl`, which illustrates how professional this industry is.

We have developed and used two systems to detect counterfeit webshops in production at `.nl`, detecting more than 20k suspicious webshops over a period of more than two years. By notifying registrars and teaming up with *ICS*, we carried out notification campaigns that resulted in 4455 domains being suspended, ultimately protecting users of the `.nl` zone from possible scams. Both detectors are relatively simple but at the same time effective, suggesting that counterfeiters were suffering little defensive pressure. As such, we can expect they will try to evade our detection systems again—as they have done with BrandCounter—which requires us to continuously adapt to evolving tactics.

Acknowledgments

We thank very much the collaboration involved in this study: the (anonymized) registrars that collaborated in removing counterfeit webshops, as well as ICS and their analysts for manually validating our results.

We also would like to thank Geoff Voelker, Moritz Müller, Damon McCoy, Elmer Lastdrager for reviewing on various paper drafts, as well as the anonymous reviewers of PAM2020, and our shepherd, Dave Levin.

SIDN and the University of Twente received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No 830927. Project website: https://www.concordia-h2020.eu/

References

1. Ahi, K., Asadizanjani, N., Shahbazmohamadi, S., Tehranipoor, M., Anwar, M.: Terahertz characterization of electronic components and comparison of terahertz imaging with x-ray imaging techniques. vol. 9483 (04 2015). https://doi.org/10.1117/12.2183128
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(Feb), 281–305 (2012)
3. C. Hesselman, J. Jansen, M. Wullink, K. Vink, and M. Simon: A privacy framework for DNS big data applications. Tech. rep. (2014), https://www.sidnlabs.nl/downloads/yBW6hBoaSZe4m6GJc_0b7w/2211058ab6330c7f3788141ea19d3db7/SIDN_Labs_Privacyraamwerk_Position_Paper_V1.4_ENG.pdf
4. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Transactions on Neural Networks **10**(5), 1048–1054 (1999). https://doi.org/10.1109/72.788645
5. 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)
6. Hao, S., Kantchelian, A., Miller, B., Paxson, V., Feamster, N.: Predator: Proactive recognition and elimination of domain abuse at time-of-registration. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 1568–1579. CCS '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2976749.2978317
7. Hao, S., Thomas, M., Paxson, V., Feamster, N., Kreibich, C., Grier, C., Hollenbeck, S.: Understanding the domain registration behavior of spammers. In: Proceedings of the 2013 Conference on Internet Measurement Conference. pp. 63–76. IMC '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2504730.2504753
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer New York (2009). https://doi.org/10.1007/978-0-387-84858-7
9. Hesselman, C., Moura, G.C.M., d. O. Schmidt, R., Toet, C.: Increasing DNS Security and Stability through a Control Plane for Top-Level Domain Operators. IEEE Communications Magazine **55**(1), 197–203 (January 2017). https://doi.org/10.1109/MCOM.2017.1600521CM

10. Hoffman, P., Sullivan, A., Fujiwara, K.: DNS Terminology. RFC 8499, IETF (Nov 2018), http://tools.ietf.org/rfc/rfc8499.txt
11. ICS: International Credit Card Services. https://icscards.nl (2020)
12. Kazemian, H., Ahmed, S.: Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications **42**(3), 1166–1177 (Feb 2015). https://doi.org/10.1016/j.eswa.2014.08.046
13. Kruczkowski, M., Szynkiewicz, E.N.: Support vector machine for malware analysis and classification. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE (Aug 2014). https://doi.org/10.1109/wi-iat.2014.127
14. Lever, C., Walls, R., Nadji, Y., Dagon, D., McDaniel, P., Antonakakis, M.: Domain-z: 28 registrations later measuring the exploitation of residual trust in domains. In: 2016 IEEE Symposium on Security and Privacy (SP). pp. 691–706 (May 2016). https://doi.org/10.1109/SP.2016.47
15. Ltd., N.: Netcraft (Oct 10 2019), https://www.netcraft.com/
16. McCoy, D., Dharmdasani, H., Kreibich, C., Voelker, G.M., Savage, S.: Priceless: The role of payments in abuse-advertised goods. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security. pp. 845–856. CCS '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2382196.2382285
17. McCoy, D., Pitsillidis, A., Jordan, G., Weaver, N., Kreibich, C., Krebs, B., Voelker, G.M., Savage, S., Levchenko, K.: PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs. In: Proceedings of the 21st USENIX Security Symposium. USENIX Association, Bellevue, Washington, USA (Aug 2012)
18. Mockapetris, P.: Domain names - concepts and facilities. RFC 1034, IETF (Nov 1987), http://tools.ietf.org/rfc/rfc1034.txt
19. Moura, G.C.M., Heidemann, J., , de O. Schmidt, R., Hardaker, W.: Cache me if you can: Effects of DNS Time-to-Live. In: Proceedings of the 2019 ACM Internet Measurement Conference (Oct 2019). https://doi.org/https://doi.org/10.1145/3355369.3355568
20. Moura, G.C.M., Heidemann, J., Müller, M., de O. Schmidt, R., Davids, M.: When the dike breaks: Dissecting DNS defenses during DDoS. In: Proceedings of the ACM Internet Measurement Conference (Oct 2018). https://doi.org/https://doi.org/10.1145/3278532.3278534
21. Nieuws, R.: Dit jaar al 307 nep-webwinkels offline gehaald door politie *(in Dutch)* (Dec 12 2018), https://www.rtlnieuws.nl/geld-en-werk/artikel/4520646/dit-jaar-al-307-nep-webwinkels-offline-gehaald-door-politie
22. NOS: Consumenten voor 5 miljoen euro opgelicht via nepwinkels op sociale media *(in Dutch)* (Dec 12 2018), https://nos.nl/artikel/2258095-consumenten-voor-5-miljoen-euro-opgelicht-via-nepwinkels-op-sociale-media.html
23. NOS: Waar komen al die nep-webshops toch vandaan? *(in Dutch)* (May 5 2018), https://nos.nl/artikel/2230087-waar-komen-al-die-nep-webshops-toch-vandaan.html
24. Peter Hornung: Gefälschte Sneaker von der FDP? (*In German*). https://www.tagesschau.de/wirtschaft/fakeshops-plagiate-sneaker-china-101.html (2019)
25. Quan, L., Heidemann, J., Pradkin, Y.: When the Internet Sleeps: Correlating Diurnal Networks with External Factors. In: Proceedings of the 2014 Conference on Internet Measurement Conference. pp. 87–100. IMC '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2663716.2663721

26. van Rijswijk-Deij, R., Jonker, M., Sperotto, A., Pras, A.: A high-performance, scalable infrastructure for large-scale active dns measurements. IEEE Journal on Selected Areas in Communications **34**(6), 1877–1888 (2016)
27. Roberts, R., Goldschlag, Y., Walter, R., Chung, T., Mislove, A., Levin, D.: You Are Who You Appear to Be: A Longitudinal Study of Domain Impersonation in TLS Certificates. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. p. 2489–2504. CCS '19 (2019). https://doi.org/10.1145/3319535.3363188
28. Schmidle, N.: Inside the Knockoff-Tennis-Shoe Factory - The New York Times. http://www.nytimes.com/2010/08/22/magazine/22fake-t.html (2010)
29. SIDN: General terms and conditions for .nl registrants (May 19 2019), https://www.sidn.nl/downloads/d_7zdiiDQvOGbSo1FGCcqw/6d8b113b06e293bd9af55fb11a66c499/General_Terms_and_Conditions_for_nl_Registrants.pdf
30. SIDN: Stichting internet domein nederland (Ago 30 2019), https://sidn.nl/en
31. Streitfeld, D.: What happens after amazon's domination is complete? its bookstore offers clues. New York Times (Jun 23 2019), https://www.nytimes.com/2019/06/23/technology/amazon-domination-bookstore-books.html
32. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural processing letters **9**(3), 293–300 (1999)
33. Taxation and Customs Union: Customs Union: EU customs seized over 41 million fake goods at EU borders last year . https://ec.europa.eu/taxation_customs/node/976_en (2016)
34. Tian, H., Gaffigan, S.M., West, D.S., McCoy, D.: Bullet-proof payment processors. In: 2018 APWG Symposium on Electronic Crime Research (eCrime). pp. 1–11 (May 2018). https://doi.org/10.1109/ECRIME.2018.8376208
35. Turner, K.: That Chanel bag on your Instagram feed may not be a Chanel bag. https://www.washingtonpost.com/news/the-switch/wp/2016/05/26/that-chanel-bag-on-your-instagram-feed-may-not-be-a-chanel-bag (2016)
36. U.S. Customs and Border Protection Office of Trade: Intellectual Property Rights – Fiscal Year 2017 Seizure Statistics. https://www.cbp.gov/document/stats/fy-2017-ipr-seizure-statistics (2017)
37. Wall, D.S., Large, J.: Jailhouse frocks: Locating the public interest in policing counterfeit luxury fashion goods. The British Journal of Criminology **50**(6), 1094–1116 – http://ssrn.com/abstract=1649773 (2010)
38. Wang, D.Y., Der, M., Karami, M., Saul, L., McCoy, D., Savage, S., Voelker, G.M.: Search + seizure: The effectiveness of interventions on seo campaigns. In: Proceedings of the 2014 Conference on Internet Measurement Conference. pp. 359–372. IMC '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2663716.2663738
39. Wappalyzer: Identify technology on websites (Oct 19 2019), https://www.wappalyzer.com/
40. Wullink, M., Moura, G.C., Hesselman, C.: Dmap: Automating domain name ecosystem measurements and applications. In: 2018 Network Traffic Measurement and Analysis Conference (TMA). pp. 1–8. IEEE (Jun 2018)
41. Wullink, M., Moura, G.C., Müller, M., Hesselman, C.: Entrada: A high-performance network traffic data streaming warehouse. In: Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP. pp. 913–918. IEEE (Apr 2016)

A Appendix: screenshots of counterfeit webshops

Figure 11 shows the screenshot of a counterfeit webshop captured in 2016 on the `.nl` zone, also shown in [5]. Figure 12 shows the screenshot of a counterfeit webshop captured in 2019.

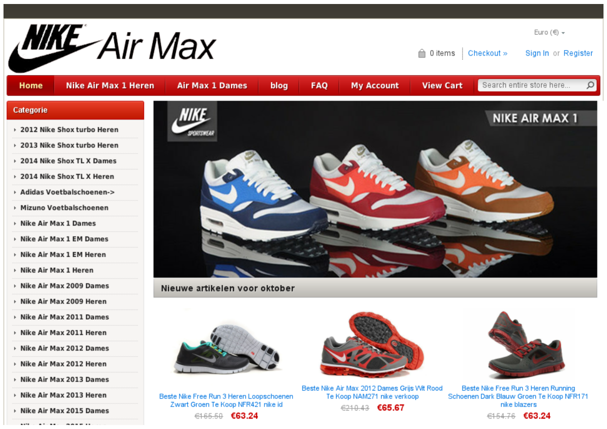

Fig. 11. Example of counterfeit webshop detected in 2016.

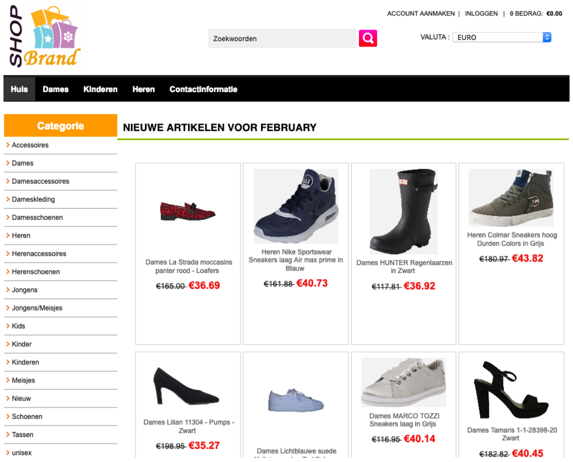

Fig. 12. Example of counterfeit webshop detected in 2019.