nDEWS: a New Domains Early Warning System for TLDs

Giovane C. M. Moura, Moritz Müller, Maarten Wullink, and Cristian Hesselman
SIDN Labs
Stichting Internet Domeinregistratie Nederland (SIDN)
Arnhem, The Netherlands
Email: {firstname.lastname}@sidn.nl

***Abstract*—We present nDEWS, a Hadoop-based automatic early warning system of malicious domains for domain name registry operators, such as top-level domain (TLD) registries. By monitoring an entire DNS zone, nDEWS is able to single out newly added suspicious domains by analyzing both domain registration and global DNS lookup patterns of a TLD. nDEWS is capable to detect several types of domain abuse, such as malware, phishing, and allegedly fraudulent web shops. To act on this data, we have established a pilot study with two major .nl registrars, and provide them with daily feeds of their respective suspicious domains. Moreover, nDEWS can also be implemented by other TLD operators/registries.**

I. Introduction

Since their inception, domain names have been used to provide a simple identification label for hosts, services, applications, and networks on the Internet [1]. In the same way, domains and the Domain Name System (DNS) infrastructure have also been misused in various types of abuses, such as phishing, spam, malware distribution, among others.

In [2], Hao *et al.* have investigated the initial lookup behavior of malicious domains under the .com and .org top-level domains (TLDs), immediately after their registration. They showed that malicious domains have different global DNS lookup patterns compared to legitimate ones, having an abnormal higher number of lookups.

We observed a similar behavior for the .nl country code TLD (ccTLD), which we manage as part of our role as .nl registry. The right part of Figure 1 shows the average number of DNS requests that we received on one of the .nl authoritative name servers for all phishing domains reported by Netcraft [3] (Jan–Aug 2015), while the left part of the figure shows the same metric for a random sample of 20,000 new domains (Jan-March 2015). As can be seen, the differences are significant with more than 30 times more queries for phishing domains.

We assume that this behavior is a consequence of the "business model" used by phishers: whenever they register a malicious domain name, they try to scam the largest number of people before the domain/website is taken down. To do so, they very quickly resort to spam campaigns, reaching users all over the world, which ultimately leads to a high volume of global DNS lookups from more diverse sources compared to legitimate domains.

978-1-5090-0223-8/16/$31.00 © 2016 IEEE

In this paper, we take this business model and behavior into account and use our position of a registry (which is that it has a global view on the DNS traffic for its respective TLD) to build an early warning detector of newly registered malicious domains, as they are added to a DNS zone (.nl in our case). We refer to it as nDEWS (new Domains Early Warning System). nDEWS differs from traditional DNS-based detection systems (e.g.: [4], [5]), since it analyzes only DNS traffic from and to TLD authoritative servers, as well as domain registration data.

We show that nDEWS is capable to detect not only phishing, but any sort of domain abuse that may employ spam/social networking as a domain dissemination method. To perform such classification, nDEWS makes use of the k-means clustering algorithm [6] (no a-priori knowledge on the query pattern and adaptable to seasonal/diurnal patterns [7]). nDEWS classifies *every new domain* added to the DNS zone (instead of a ample captured by a spam trap as in [2]).

We evaluate nDEWS against 8-month historical datasets from the .nl country-code TLD (ccTLD) for the period between January and August, 2015, and employ ENTRADA [8], our Hadoop-based data streaming warehouse (DSW), which we have recently open-sourced [9]. nDEWS has been running for several months in a testing phase for the .nl zone. We have then setup a pilot study with two major .nl registrars to provide them with daily feeds of their respectively newly registered suspicious domains, so they can take action based on it.

We make the following contributions: we present nDEWS, an early-warning system for TLD registry operators that evaluates an entire DNS zone and singles out newly registered malicious domains used for several types of abuse (phishing, scams, malware, etc.). We discuss our design choices and feature selection, and carry out an evaluation and validation using 8 months of historical data. We also show the emergence of fraudulent websites (the so-called concocted websites [10], [11]) that, differently from traditional phishing sites, evade well-established industry blacklists (e.g., Google SafeBrowsing [12]), ultimately leaving users vulnerable to these scams [13], [14], [15].

The rest of this paper is organized as follows: in Section II, we provide background information on DNS and TLDs. Then, in Section III, we describe the architecture of nDEWS, while in Section IV we carry out an evaluation of nDEWS. After that,

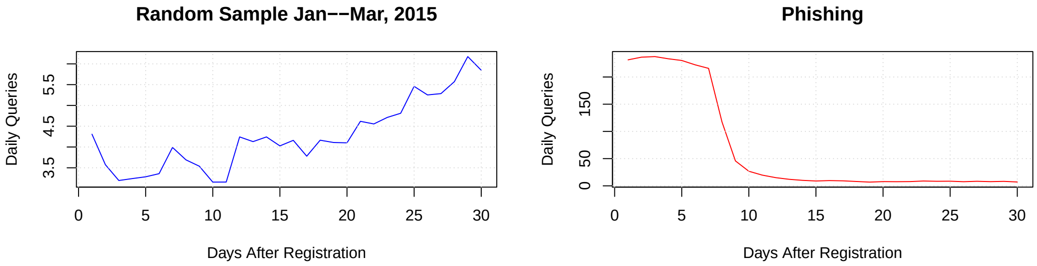

Fig. 1. .nl Random vs Phishing new domains average daily queries

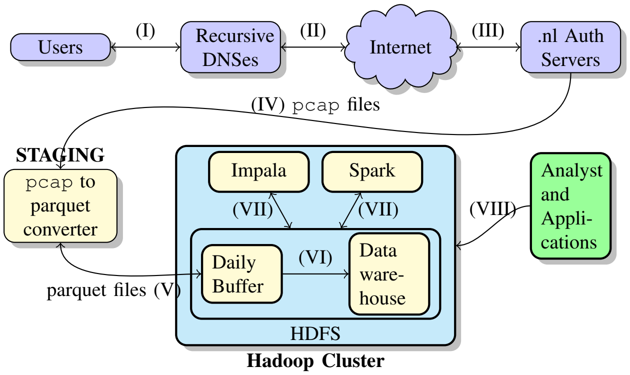

Fig. 2. ENTRADA data sequence flow

in Section V we cover the related work, while conclusions and future work are finally presented in Section VI.

II. Background

SIDN is the registry for .nl, which is the ccTLD of the Netherlands. As part of this registry[1] role, SIDN manages the authoritative DNS servers for .nl, as well as the domain registration system. Currently, more than 5.5 million domains are registered in the .nl zone [16].

Authoritative DNS data and domain resolution: to understand the data we analyze for this paper, we first have to understand the domain resolution process, which we simplify here in an example – we refer the reader to [1] for a more detailed explanation.

Consider a user trying to reach a website (e.g.: www.example.nl). This domain must be therefore resolved into an IP address. First, the user's computer (which runs a stub DNS resolver) connects to a recursive DNS resolver (I in Figure 2), which is usually located at the Internet Service Provider's (ISP) network. This resolver, will, in turn, start a recursive process on behalf of the user, asking for the IP address of .nl authoritative servers to the root servers (the "." zone, II in Fig. 2, not shown).

The DNS recursive resolver will then ask one of the .nl authoritative servers for the IP address of example.nl (III in the same figure). The .nl name server will refer the recursive server to the name server of example.nl, which knows the IP address of www.example.nl and returns it to the recursive server. The recursive server will ultimately send the requested IP address to the user, whose browser will then be able to reach www.example.nl.

To improve performance, recursive resolvers employ local caches to store responses [17]. As a consequence, just part of queries received by a recursive resolver makes it to the respective authoritative servers.

ENTRADA: to store, analyze, and continuously process authoritative DNS data, we employ ENTRADA [8], [9], a Hadoop-based high performance data streaming warehouse that we open-sourced. It enables us to analyze vast amounts of network traffic and measurement data within interactive response times (seconds to few minutes, depending on the query). For example, for aggregation type queries (e.g., average packet size), it takes 3.5 minutes to process the equivalent of 52 TB of pcap in a small 5 node-cluster with 1.9 GHz 6-core Xeon processors.

ENTRADA delivers such performance by (i) employing Parquet [18], an optimized column-based file format based on Google's Dremel [19], and (ii) using Impala [18], an open-source massively parallel processing SQL-query engine, as can be seen in Figure 2. We convert the pcap files from the .nl authoritative servers to Parquet (IV, Fig. 2) and store these files on the Hadoop file system (HDFS, IV in Fig. 2), where they are then accessible by any Parquet-compatible tool. In our case, we employ Impala, which allowss query execution to be parallelized. nDEWS is developed as an application (VIII) that connects to ENTRADA through Impala.

Currently, ENTRADA stores all the DNS traffic from two .nl authoritative servers, starting from May 2014. This, in turn, corresponds to ∼25% of the total DNS traffic to all the .nl authoritative servers (∼400 M daily queries stored). We refer the reader to [8] for more details on ENTRADA.

Additionally, we have developed, together with our legal department, a publicly available data privacy framework that conforms to both EU and Dutch laws [20].

III. nDEWS Architecture

Figure 3 shows the architecture of nDEWS. In the first step (I), we obtain all domains that have been added to the

[1] The complete list can be found at https://www.iana.org/domains/root/db.

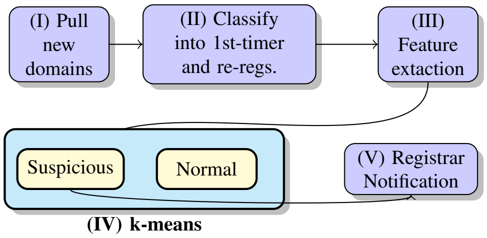

Fig. 3. nDEWS Architecture

.nl zone, we obtain $< d, ts >$ in which d is the domain name and ts is the registration timestamp, using ENTRADA, via the Impyla Python API[2].

In step II, we filter these domains and discard domains that have been re-registered. When domain names are abandoned by users, we keep them in "quarantine" for 40 days. After this time, we allow them to be re-registered. We ignore these domains because they do not follow the pattern shown in Fig. 1, since many crawlers and domain drop catchers keep on querying in the period before they are made once again available for registration.

A. Feature Selection

After that, we move to the feature selection step (III). To determine these features, we analyzed the behavior of malicious and benign domains for a period of several months ([21] and Fig. 1). We observed that besides having a smaller volume of queries, benign domains are less distributed geographically (e.g.: most resolvers IP are geolocated in the Netherlands, US, and a handful of EU countries in the first 24 hours after the creation of the domain)[3].

Malicious new domains, on the other hand, allegedly relying on spam campaigns, are prone to hit mailboxes of users located in a larger number of countries/networks/ASes. These users, in turn, after resolving the domain (Fig. 2), end up increasing the values for the features we evaluate.

Given these empirically observed differences, we chose four DNS-based features for nDEWS: total number of DNS requests ($\#Req$), total number of unique source IP addresses ($\#IPs$), unique countries ($\#CC$), and unique Autonomous Systems (ASes) ($\#ASes$). We chose a 24 hour period for this paper in order to have a good compromise between early detection and enough data points. However, the algorithm can be adjusted to any time window.

The output of (III) is a a tuple as follows: $< d, ts, \#Req$, $\#IPs$, $\#CC$, $\#ASes >$. In a typical day, more than 2,000 new domains are added to the .nl zone. First-timer domains (never previously registered), on which we focus in this paper, make up ~80% of all new daily domains. We employ ENTRADA to extract these features.

It is important to notice that $\#Req$ alone does not provide a good indication if the websites is malicious or not. We have seen over the testing period several false positives, in which politically or socially motivated websites, spread with the help of social networks, have a large number of requests. These, in turn, are typically more geographically concentrated than malicious domains having smaller values for $\#IPs$, $\#CC$ and $\#ASes$ than typical malicious websites.

B. Classification and Implementation

To classify these domains into "suspicious" and "normal", we employ the k-means clustering algorithm [6], which aims at partitioning the dataset into n clusters in a way that minimizes the total distance between the data points and the cluster's corresponding centroid. The advantages of k-means is that it does not require any a-priori knowledge about the domains and does not required labeled datasets/training, therefore being able to better cope with seasonal/diurnal patterns [7].

Another possibility would be developing a classification system trained based on the ground truth provided, for example, by Netcraft [3]. This has been done by one of the authors in [21], found effective in phishing detection. However, for nDEWS, we intentionally did not want to tailor to phish – but to *any* malicious activity, such as malware, fraudulent websites, etc.

Therefore, we employ the k-means algorithm on the features described in Section III-A, on a daily basis, and produce a daily list of suspicious domains. The daily analysis is independent from the results obtained in the previous days. In the current setup, it takes less than 45 minutes to run nDEWS daily, which include all steps in Fig. 3, including access to the registration database, feature extraction using ENTRADA, and classification using the R statistical software in a local server.

IV. Evaluation

A. Datasets

We apply nDEWS to our historical record, as shown in Table I. We analyze 8 months of data (Jan-Aug 2015) from both DNS registrations and DNS traffic stored by ENTRADA, from one of the .nl authoritative servers. In this period, the .nl zone had an average of ~5.5 million domains registered. In the same period, more than 586 thousand new domains were registered – and more than 80% of those were never registered before in the .nl zone history (first-timers).

In the same period, more than 32B DNS requests/responses were stored at ENTRADA. Out of those, 1.7 million DNS requests were related to the new domains, for the first 24 hours after their creation. Out of those, 476K were request to first-timer domains[4], which we use in this evaluation.

B. Results

We employ k-means on the aforementioned datasets using the approach described in Section III. Before proceeding with the clustering of these domains, we exclude first-timer domains

[2] https://github.com/cloudera/impyla

[3] It is important to notice that .nl is a TLD open to registrations from anyone in any country. However, most of the registrations are in some way related to the Netherlands. It is not clear if the same would occur to other ccTLDs.

[4] For other statistics on .nl DNS traffic and registration, please refer to our DNS statistics website [16].

Key	Value
Interval	Jan 1st, 2015 to Aug 30th 2015
Average .nl zone size	∼ 5,500,000
# new domains	586,201
New domains - first timers	476,040(81.2%)
New domains - re-registered	110,161 (18.8%)
Total DNS Requests	32,864,402,270
DNS request new domains (24h)	826,740
DNS request new domains - first-timers (24h)	420,362

TABLE I
EVALUATED DATASETS (FROM ONE .NL AUTH SERVER)

Cluster	Size	$\#Req$	$\#IPs$	$\#CC$	$\#ASes$
Normal	132,425	4.31	3.06	1.64	1.43
Suspicious	2,956	55.03	27.87	4.99	7.43

TABLE II
MEAN VALUES FOR FEATURES AND CLUSTERS - EXCLUDING DOMAINS WITH 1 REQUEST - 1ST TIMERS

that had only one DNS request – they would automatically belong to the "normal" cluster, since malicious domains are likely to have more queries. Out of 476,040 domains, 139,103 had more than 1 DNS request on the first 24 hours, which we then classify using k-means.

Table II shows the results for the first-timer domains. As can be seen, the clusters have dissimilar mean values for the evaluated features. The suspicious cluster is also far smaller than the "normal" one, representing 0.068% of all first-timer domains, having an average of 12.2 new "suspicious" first-timer domains daily. Figure 4 shows the ECDF for the number of requests for both clusters. As can be seen, more than 90% of the "normal" domains had less than 10 requests on the first 24 hours, while more than 90% of the "suspicious" domains had more than 10 requests.

C. Validation

"Suspicious" domains exhibit a distinct behavior in comparison to "normal" ones. However, this classification by itself does not imply that those domains were, in fact, malicious. Validating our results without historical data is a challenge, since there is no way to be sure about the content of a suspicious website months or week before. In absence of a ground truth dataset, we apply different approaches to verify our results: content analysis and comparison against well-known abuse lists, as we discuss next.

Content Analysis: assuming a domain has a A or AAAA record associated with it, and it has a webserver running on port 80/443, we can, potentially, analyze their website content. Therefore, we employed PhantomJS[5], a headless browser used for automating interaction with web pages. We capture these screenshots on October 9th, 2015 and manually analyzed them. Although we cannot guarantee that the content of these domains were the same as of their creation it may still suggest malicious activities.

First, we remove domains we cannot analyze: out of 2,956 domains in the "suspicious" cluster, 479 had no A/AAAA record, while other 82 had no content in their pages. We also noticed that many websites employed the same "landing" page – e.g.: Google's or another hosting provider's error page. By hashing the respective PNG images, we determine which websites were exactly the same, which amounts to 288. A landing page on a website can indicate that it hosted malicious content earlier, which has now been removed.

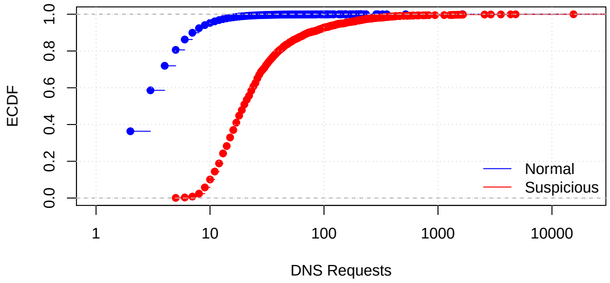

Fig. 4. ECDF Number of Requests on the first 24 hours – 1stTimers

We also removed domains that have been re-registered within the evaluated time interval, i.e., belonging to another registrant within the 8 month period. In total, 85 fell into this category. After that, 2,227 domains were left for analysis. Out of those, we found that 148 contained very similar suspicious shoe stores, while 13 were related to adult content. By manually inspecting the remaining ones, we were not able, using their current content (and not the one when they were created) to determine if they were malicious or not.

Netcraft phishing blacklist: Netcraft collects the phishing domains via user submission through their toolbar. They claim that they have detected over 17.5 million unique phishing sites for all TLDs (until October 2015) [3]. For the same period (Tab. I), Netcraft listed 1,672 malicious phishings URLs on .nl zone. Out of those, 1,553 were first-timer domains – i.e., domains that have never existed in the .nl zone before.

Since we evaluate the first 24 hours of each domain, we compare our results against the set of domains reported by Netcraft within 24 hours of the creation of the domain: 47 out of 1,553. nDEWS classified 19 of the 47 phishing domains as malicious. The majority of the domains detected by nDEWS were used to target Dutch banks and credit card companies (18 domains) and 1 domain targeted users of the online payment service PayPal. The remaining domains that were not detected by nDEWS indicate that their dissemination method did not led to a suspicious query pattern at our .nl authoritative server. This can be improved by running nDEWS in adaptive time windows for newly created domains and analyzing traffic from other authoritative servers altogether. Moreover, nDEWS was able to detect 3 phishing domains at least one day *earlier* than Netcraft.

Google Safe Browsing: Google Safe Browsing provides a constantly updated list of domains that are suspected of hosting phishes, malware or other unwanted software. Google only stores information for the last 90 days, therefore covering our dataset only partially. As a consequence, domains that have only been malign before this period are not listed anymore.

[5] http://phantomjs.org/

During our evaluation, Google Safe Browsing classified 6 out of the 2,956 (Tab. II) domains detected by nDEWS as malicious. 5 were observed during a phishing campaign and 1 took part in the distribution of malware.

Virustotal aggregated blacklist: VirusTotal provides an aggregated blacklist with regards to the domains' reputation [22]. 25 domains were classified by at least one of the Virustotal data sources as malign. 14 were classified as phishing, 6 as malware and 4 as malicious website. At the time of our analysis, 3 of these 25 domains hosted suspicious shoe stores, as described in the following section. These stores were either classified as malicious or as malware site, which indicates that those sites not only sell fake goods but might be responsible for the distribution of malware as well.

Ultimately, nDEWS was capable to detect roughly 1/3 of the 1st-timer first-day phishing domains notified by Netcraft, but only few from Google SafeBrowsing, which is also designed for phishes not older than 90 days, and 25 from VirusTotal. Ultimately, we performed a manual content analysis for each suspicious domain. From the 2,227 still active domains, we were able to confirm that 148 are still involved in some suspicious activity. However, this does not mean that the remaining ones were not involved at an early point in time.

D. Validation on current data

Currently, we run nDEWS on a daily basis to evaluate all the domains added to the .nl zone. We capture their screenshot, download their main page, and check all DNS records available for the domain. On average, we find 12 suspicious domains a day. By performing content analysis, we see that most of these domains are indeed concocted websites – and 1 or 2 of the remaining ones seem to be false positives. By using nDEWS we can potentially stop these websites within 24 hours after their registration. In this sense, nDEWS fills a void left by the current industry blacklists and complements it as a method to prevent fraud.

Of course, besides detecting these websites, we need to take action on that. Therefore, we have set up a pilot study (Dec. 2015) with two major .nl registrars. After domains are classified as malicious, we sent to these registrars, on a daily basis, domains that have been registered by their respective customers. In many cases, phishing is the most common type of abuse. We are currently evaluating this pilot and deciding which actions we can take based on our results.

E. Concocted websites

In the previous section, we showed that 148 domains were found hosting suspicious big discount online stores. At a first glance, they look like any other online store, except for large discounts, as can be seen in Figure 5. One may even underestimate the risks posed by these websites in comparison with "traditional" banking/credit card phishing.

Before dismissing such websites, it is important to understand the counterfeit industry. According to the World Customs Organization, counterfeit goods account for nearly 10% of worldwide trade, an estimated $500 billion annually [23]. Sneakers are the *number one* products seized by the U.S. Customs and Border Protection – amounting to a total of 40% of all seizures [24], [23].

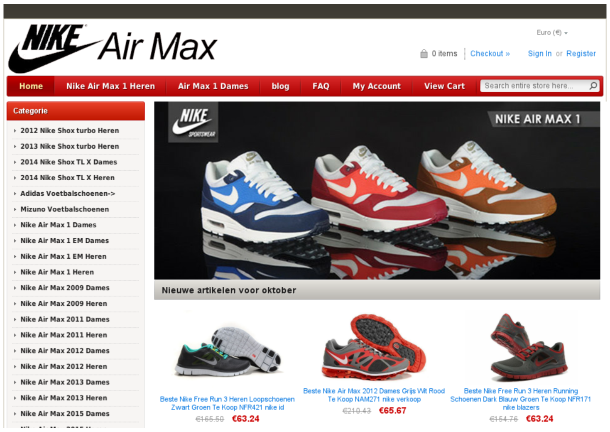

Fig. 5. Example of a concocted shoe store

In the literature, such websites are referred to as concocted stores, i.e., deceptive websites that appear to be legitimate commercial ones [11], and either fail to ship their ordered goods or ship different/counterfeit products. They differ from spoof sites, which are intended to deceive authentic site's costumers [10]. It is hard to determine with 100% certainty if a website is fraudulent or not – ultimately this involves a trial purchase.

One way to estimate their impact is by analyzing reported losses. According to FraudHelpdesk, a Dutch foundation against online fraud, 1.1 million Dutch citizens have been scammed in 2013/2014, totalling a damage of €5.3 billion [13]. Germany and Australia [15], [14] present similar figures.

We found that the IP addresses of 103 of the 148 stores were geolocated to Russia, following by 17 in the U.S. and 13 in the Netherlands. Out of those, 101 IP addresses belong the same Autonomous System. By analyzing the registration database, we found that 118 of them used the same registrar (one that provides whois anonymization), located outside the EU. This suggests that the same gang might be responsible for all these websites. We are currently investigating how to curb these types of suspicious websites.

V. Related Work

To the best of our knowledge, this is the first work that provide an early warning of newly malicious registered domains for an entire DNS zone. We employ global DNS lookups from a TLD and employ a machine learning (k-means) algorithm to monitor and classify an entire DNS zone.

This work is inspired by the work of Hao *et al.* [2], in which the authors have analyzed the initial lookup behavior of malicious domains under .com and .org TLDs. The authors set of malicious domains is obtained from a spam trap (they obtained 2,045 malicious domains for March 2011). Combined with an external source, they employed roughly 6,000 malicious domains in their study.

We, on the other hand, classify *every single new domain* added to the .nl zone, and run nDEWS on a daily basis. Per day, we have more than 2,000 new domains. For these domains, we analyze their global lookup patterns (Section III) and employ a k-means-based classifier. Moreover, we not only analyze the number of queries, but how many distinct source IP addresses, ASes, and countries have queried these domains. In the 8-month period in which we have evaluated nDEWS, we found 2,956 suspicious domains.

Other works, such as [25], while also monitoring DNS traffic from authoritative servers (or top-level domains), do not focus on newly registered domains. Also, different methods exist to classify malicious websites. For example, Abbasi and Chen [11] present a comparison for tools to detect fake websites. They perform content analysis to classify the websites. Bilge et al. [4] use, next to DNS traffic characteristics collected at resolvers, also domain name based features. We, on the other hand, rely upon the domains' global DNS lookup patterns. Moreover, nDEWS can be evaluated and used by other TLDs in order to monitor the behavior of their domains in their own DNS zone and it would be interesting to observe if the same patterns hold for those TLDs as well.

VI. Conclusions and Future work

We present nDEWS, a Hadoop-based early warning automatic detector of newly registered malicious domains on a top-level domain (TLD) zone. By monitoring an entire DNS zone and analyzing data from global DNS lookups, nDEWS is capable not only of detecting phishing domains, but also other type of abuses, such as malware, phishing, and allegedly fraudulent web shops, filling a void left by traditional phishing blacklists. To classify these domains, nDEWS employs the k-means clustering algorithm.

We developed nDEWS as an application atop of ENTRADA, our open-source DSW, and currently use it on a daily basis to evaluate every domain added to the .nl zone. We are currently conducting a pilot study with two major .nl registrars by sending abuse notifications for .nl domains under their management.

As future work, we intend to asses the impact of including more features in the analysis (IP reputation, type of DNS records, html and content analysis, and domain registration information) to improve accuracy, as well as running nDEWS on both short and longer time windows. Moreover, we are also working on a new ENTRADA anomaly detection application to evaluate all domains in the .nl zone, and not only the newly registered domains. We hope that the findings here presented can help other TLD registries to deploy similar systems.

Acknowledgments: The authors would like to thank Bert Ten Brinke for his valuable comments.

References

[1] P. Mockapetris, *RFC 1034 Domain Names - Concepts and Facilities*, Internet Engineering Task Force, 1987.

[2] S. Hao, N. Feamster, and R. Pandrangi, "Monitoring the Initial DNS Behavior of Malicious Domains," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 269–278.

[3] Netcraft, "Phishing Site Feed," http://www.netcraft.com/anti-phishing/phishing-site-feed/, 2015.

[4] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis." in *NDSS*, 2011.

[5] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, "Building a Dynamic Reputation System for DNS." in *USENIX security symposium*, 2010, pp. 273–290.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[7] L. Quan, J. Heidemann, and Y. Pradkin, "When the Internet Sleeps: Correlating Diurnal Networks with External Factors," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, ser. IMC '14. New York, NY, USA: ACM, 2014, pp. 87–100.

[8] Maarten Wullink, Giovane C. M. Moura, Müller, M, and Cristian Hesselman, "ENTRADA: a High Performance Network Traffic Data Streaming Warehouse," in *Network Operations and Management Symposium (NOMS), 2016 IEEE (to appear)*, April 2016. [Online]. Available: https://www.sidnlabs.nl/downloads/sidn-noms2016_EN.pdf

[9] SIDN Labs, "ENTRADA homepage," http://entrada.sidnlabs.nl/, 2015.

[10] N. Chou, R. Ledesma, Y. Teraguchi, J. C. Mitchell *et al.*, "Client-Side Defense Against Web-Based Identity Theft," in *NDSS*, 2004.

[11] A. Abbasi and H. Chen, "A comparison of tools for detecting fake websites," *Computer*, no. 10, pp. 78–86, 2009.

[12] Google, "Google safe browsing api," https://developers.google.com/safe-browsing/, January 2015.

[13] FraudHelpdesk.nl, "Ruim miljoen Nederlanders opgelicht (in Dutch)," https://www.fraudehelpdesk.nl/nieuws/ruim-miljoen-nederlanders-opgelicht-2/, Dec 2014.

[14] P. Hornung, N. Walker, and H. Maassen, "Wie Kriminelle gestohlene Identitäten missbrauchen – NDR.de (in German)," http://www.ndr.de/nachrichten/netzwelt/Identitaetsdiebstahl,identitaetsdiebstahl100.html, Oct 2015.

[15] ScamWatch, "Online shopping scams," https://www.scamwatch.gov.au/types-of-scams/buying-or-selling/online-shopping-scams, Oct 2015.

[16] SIDN Labs, ".nl stats and data: Insight into the use of .nl," http://stats.sidnlabs.nl/, 2015.

[17] M. Andrews, "Negative Caching of DNS Queries (DNS NCACHE)," RFC 2308, Internet Engineering Task Force, Mar. 1998.

[18] M. Kornacker, A. Behm, V. Bittorf, T. Bobrovytsky, C. Ching, A. Choi, J. Erickson, M. Grund, D. Hecht, M. Jacobs *et al.*, "Impala: A modern, open-source SQL engine for Hadoop," in *Proceedings of the Conference on Innovative Data Systems Research (CIDR'15)*, 2015.

[19] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, "Dremel: Interactive Analysis of Web-scale Datasets," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 330–339, Sep. 2010.

[20] C. Hesselman, J. Jansen, M. Wullink, K. Vink, and M. Simon, "A privacy framework for DNS big data applications," Tech. Rep., 2015. [Online]. Available: https://www.sidnlabs.nl/uploads/tx_sidnpublications/SIDN_Labs_Privacyraamwerk_Position_Paper_V1.4_ENG.pdf

[21] M. Müller, "SIDekICk: SuspIcious DomaIn Classication in the .nl Zone," Master's thesis, University of Twente, the Netherlands, 2015.

[22] VirusTotal, "About VirusTotal," https://www.virustotal.com/en/about/, 2015.

[23] B. Burnsed, "The Most Counterfeited Products – Bloomberg Business," http://www.bloomberg.com/ss/08/10/1002_counterfeit/1.htm, 2015.

[24] N. Schmidle, "Inside the Knockoff-Tennis-Shoe Factory - The New York Times," http://www.nytimes.com/2010/08/22/magazine/22fake-t.html, 2010.

[25] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou II, and D. Dagon, "Detecting malware domains at the upper dns hierarchy." in *USENIX Security Symposium*, 2011, pp. 16–32.