Heuristic methods for efficient identification of abusive domain names

Kidmose, Egon; Lansing, Erwin; Brandbyge, Søren; Pedersen, Jens Myrup

Published in:
International Journal On Cyber Situational Awareness (IJCSA)

DOI (link to publication from Publisher):
10.22619/IJCSA.2018.100123

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Kidmose, E., Lansing, E., Brandbyge, S., & Pedersen, J. M. (2018). Heuristic methods for efficient identification of abusive domain names. *International Journal On Cyber Situational Awareness (IJCSA)*, *3*(1), 121-142. https://doi.org/10.22619/IJCSA.2018.100123

Intl. Journal on Cyber Situational Awareness, Vol. 3, No. 1, 2018

Heuristic Methods for Efficient Identification of Abusive Domain Names

Egon Kidmose[*][†], Erwin Lansing[‡], Søren Brandbyge[†], Jens Myrup Pedersen[*]

[*]*Department of Electronic Systems, Aalborg University,*
Fredrik Bajers Vej 7, DK-9220 Aalborg Øst, Denmark.
[†]*LEGO System A/S, Aastvej, DK-7190 Billund, Denmark.*
[‡]*DK Hostmaster A/S, Ørestads Boulevard 108, 11. sal, DK-2300 København S, Denmark.*

ABSTRACT

Domain names and the Domain Name System (DNS) are essential to the Internet, but unfortunately cyber-criminals also make use of these to fulfil their nefarious agenda and gain illicit profit. In this work we survey known forms of domain and DNS abuse from the criminal business point of view. We relate this to abusive techniques, which we also survey. Based on the theoretical understanding of the abusive techniques, we devise a set of practical heuristics for recognising said techniques. This enables a focused and efficient manual analysis of heuristically ranked domains, with the goal of identifying abusive domains. As the .dk Country Code Top-Level Domain has received little scrutiny in the past, but is believed to see only limited abuse, it represents a relevant and presumably challenging case for identifying abuse, and we therefore use it for evaluation. WHOIS data is collected for 10.000 second level domains for 66 days, heuristics are applied, and the resulting rankings guide a manual vetting. Our findings are that with automated heuristics we can limit the manual investigative effort to hours, and still identify 5 domains which are actively abused during our observation period.

Keyword: DNS, Domain Name, Abuse, Heuristics, Top-Level Domain.

1 INTRODUCTION

As links between the cyber and physical realms have grown in numbers and strength, the potential profit and impact of cybercrime has grown too. Domain names are a prime example of this, as businesses, organisations, and private persons use domain names not only for technical administrative purposes, but also very much for branding themselves on the Internet. As domain names and the Domain Name System (DNS) are ubiquitous, cyber criminals have naturally found it useful too, both for technical purposes and especially due to it being trusted by users and organisations.

One example of cybercrime involving domains is phishing, where attackers pretend to be a trustworthy third party and lure the victim into disclosing confidential information such a credentials, which can be exploited for profit. In phishing schemes, it is common to use domain names mimicking the third party to gain the trust of the victim. This type of attack targets the discrepancies between human fuzzy interpretation of domain names and the exact mapping provided by DNS. It can be combined with similar techniques, such as mimicking visual identity and logos, which are not pertaining to DNS and domain names. In addition to these human-targeted techniques there are also examples of abuse that are purely technical. Just like legitimate organisations rely on the DNS infrastructure, so do the criminals. This can be observed when bots or other malware attempt to establish a Command and Control (CnC) channel from compromised victim machines to the attacker, when spam emails are sent, or when scam web shops present themselves like ordinary, legitimate web shops. There are many more examples of how domain names and DNS can be abused, but we defer a more extensive survey to Section 2.

The global losses caused by cybercrime in 2017 have been estimated by Lewis (2018) to be between 445 and 608 billion USD, while McGuire (2018) estimated the annual global revenue for cybercrime to be 1.5 trillion USD. Given the size of the cybercriminal underground economy, it is clear that it is beneficial for the criminals to specialise in different roles. Examples of such roles include finding exploits, writing malware, infecting victims, operating CnC infrastructure, and laundering money (Sood, Bansal, and Enbody, 2013). While the individuals of a group of criminals can specialise accordingly, the reality is that criminals even specialise to the extent that they provide their services on well-established underground markets. Pay Per Install (PPI) is an example of this, where the operators of botnets sell or rent access to victim machines. Another example is how credentials and personal information are traded on well-organised underground and online marketplaces, where customer satisfaction is ensured by providing merchant ratings and escrow services. As a final example, malware and exploit kits come nicely packaged with 24/7 phone support.

These phenomena are clearly undesirable for the general, law abiding society, so luckily there are means to counter them. As already exemplified, cybercrime relies extensively on domain names and DNS, which therefore is an interesting means for solving the problem. Victims can defend themselves with blacklists of bad domains and with detection methods based on both heuristic algorithms and data driven machine learning. An alternative, complementary approach is for the registries operating Top-Level Domains (TLD), such as *.com* or *.dk*, and related DNS infrastructure, to block queries or reject registrations for abusive domains. The registries are not necessarily impacted directly by abuse, but if a TLD is widely and commonly abused, the reputation is impacted negatively, and thereby also the value of the products offered by the registry. Furthermore, there are legal aspects that motivate registries. Registries are thus both capable and motivated to combat abuse.

Our primary contributions are a method to heuristically rank domains, such that manual effort invested towards identifying abusive domains can be used efficiently, and the first published scientific study on abuse in the *.dk* TLD. We also contribute with an overview of malicious techniques, with heuristics motivated by an understanding of the techniques, and with the detailed results of applying the heuristics.

In Section 2 we survey different types of abuse, with an outset in the business models and the criminal economy in which the abuses took place. Based on these types of abuses we identify malicious techniques used by the criminals in relation to domain names and DNS. In Section 3 we apply the insights into the techniques to develop a set of heuristics for identifying domains where the techniques are applied. We collect data for a subset of the *.dk* country code TLD and apply the heuristics to search for abusive second-level domains (2LD), with Section 4 describing the outcome. We offer our interpretation of the results in Section 5 describe the future direction in Section 6 and conclude on the current study in Section 7

2 BACKGROUND

In this Section we survey abuse from a criminal business perspective, identifying different schemes, with clear links to how domain names are abused, and how the illicit profit is generated. This is followed by a survey of techniques that enables or improves the schemes. The distinction between schemes and techniques is important, because schemes allow us to understand the motivation of the criminals, while understanding of the techniques enables us to look for technical artefacts that can be searched for at a large scale.

Abuse Schemes

Phishing has already been described in the Introduction as a scheme where criminals rely on luring victims into disclosing confidential information. In this scheme the business model can be as simple as: Register a domain similar to the trusted third party, e-mail victim(s) misleading instructions, and receive credentials from the victims that succumb to the attack. Obtaining victim e-mail addresses, sending large amounts of phishing emails, and exploiting the results can be delegated, hence the criminal value added in phishing comes from tricking users. It is relevant to discern between 0-day phishing domains registered with intentions of abuse and compromised phishing domains that are registered with good intentions, but later compromised by criminals and abused for phishing (Moura, Müller, Davids, Wullink, and Hesselman, 2017). 0-day phishing domains should ideally be detected before registration, or briefly thereafter, as criminals register, exploit, and discard domains rapidly (Hao, Thomas, Paxson, Feamster, Kreibich, Grier, and Hollenbeck, 2013). Registries are challenged when it comes to addressing abuse with compromised domains without involving the registrant owning the domain, as any action by the registry will likely interfere with legitimate use of the domain. On the other hand, involving the rightful, exploited registrant can be slow and time consuming.

E-mail spam is the well-known malpractice of sending bulk, unsolicited e-mails, where profit can be made, for instance, by infecting victims (PPI), or through ads and referrals to dubious web shops (Kanich, Kreibich, Levchenko, Enright, Voelker, Paxson, and Savage, 2008). While the problem is well known, there is plenty of evidence that the abuse of domain names for spamming has not been handled yet (Hao, Thomas, Paxson, Feamster, Kreibich, Grier, and Hollenbeck, 2013).

Scam web shops use domain names for landing customers, just like legitimate web shops, but the shipped goods might be counterfeit, if it is even shipped at all, and the customer credit card details might be stored and abused (Abbasi and Chen, 2009). If the criminals accept payments but never ship the goods, or ship cheaper counterfeits, they profit. If the consumer is knowingly buying counterfeit goods, the criminals are profiting from facilitating the illicit trading of counterfeit goods. In all cases, domain name abuse enables illicit profit.

Domain parking is the practice of registering a domain without developing it and without providing genuine content, but rather redirecting traffic to a parking service, which generates generic content, typically advertisements, in order to monetise from users who, by mistake, point their web browser to the domain (Vissers, Joosen, and Nikiforakis, 2015). Parking is clearly not aligned with the users' intent, but as stated by Moura et al. (2017), it is not necessarily illegal, hence it can make sense to distinguish between legal parking with ads and illegal, malicious parking, where users are diverted to scams, exploits, or other attacks. In either case the revenue stems from selling redirections of users, regardless of the user's intentions.

Web spam is a scheme that lends itself to e-mail spamming, as it uses a bulk of useless or misleading content, but instead of distributing this via e-mail it presents itself as web pages (Gyöngyi and Garcia-Molina, 2005). Abusing multiple domains to host content that refers to each other only, the criminals seek to entrap search engine web crawlers and boost their own malicious content into search results, which is why this is also referred to as Blackhat Search Engine Optimisation (Moura et al., 2017). Profit comes from serving victims with ads or malicious content. While this form of abuse is fully dependent on domain name abuse, it is obvious that search providers also have motivation and options to combat this.

Botnet CnC can abuse domain names as rendezvous points, where victim machines infected with bot malware can reach the bot master's infrastructure. With an established CnC channel, the bot master obtains scalable remote control and data exfiltration capabilities, which can enable other schemes, including harvesting of banking credentials or credit card details, sending of e-mail spam, or PPI.

Abuse Techniques

Mimicking is a technique intended to make the human victim confuse a malicious domain for a legitimate and trusted third party. This can be achieved with slight alterations from the third-party domain name, e.g. barely noticeable spelling errors, minor edits, insertion of hyphens, substitution with homoglyphs (I.e. similar looking characters), and more. Attackers can both use mimicking domains actively, e.g. when sending targets phishing e-mails, and passively, such as with typosquatting where attackers rely on victims to mistype a domain name, so they end up at a parked domain.

Malicious re-registration, also known as drop catching, is when an attacker registers an expired domain for abuse. Browser bookmarks, hyperlinks, user-remembered domain names, residual search engine results, based on the previous content, and general system configurations rarely reflect that a domain has expired. Instead trust in a domain persists if it re-registered by a criminal. Users browsing the web can be redirected to parking services, bot-infected victims can be re-enrolled by a new botmaster, and DNS infrastructure can be hijacked just like any account recoverable through an e-mail address in the domain (Lever, Walls, Nadji, Dagon, McDaniel, and Antonakakis, 2016).

Bulk registration refers to the practice of registering many domain names in bulk. While Hao et al. (2013) described how e-mail spammers employ this technique, it is clearly relevant for web spammers too, and also when employing Domain Flux (See below). The motivation for criminals to do bulk registrations lies in the convenience and scalability, and in the discounts offered by registrars.

Fluxing refers to a collection of techniques that abuse the capabilities of DNS to make the criminal’s infrastructure more resilient to take-downs and/or harder to track, investigate, and block. Fast Flux is the first variant, where a domain name maps to many IP addresses that all provide identical service or content, possibly by proxying, with the mappings changing rapidly (Holz, Gorecki, Rieck, and Freiling, 2008). The benefit for the attacker is that forensic analysis on the victim only provides one of the many redundant IPs, that IP might only point to another victim unknowingly proxying for the attacker, and all this information is rapidly outdated. This thwart blacklisting and take-down efforts.

Double Flux extends Fast Flux by also applying the same approach to how authoritative name servers are found (Fast Flux applies to A records in DNS. Double Flux applies to NS records.) (Nazario and Holz, 2008).

Domain Fluxing can be seen as the inverse of the above: Fast Flux and Double Flux enables a domain name to point to many IP addresses. With Domain Flux, a large set of domain names can be used to point to a single IP. This is achieved with Domain Generation Algorithms (DGAs) that produce large sets of pseudorandom domain names, of which the attacker can chose to register any one to establish a CnC channel (Porras, Saidi, and Yegneswaran, 2009), (Schiavoni, Maggi, Cavallaro, and Zanero, 2014). This makes it infeasible to block or sink-hole all the domains.

RELATED WORK

An empirical study of spammers advertising for scam web shops, focusing on the value chain for the criminal operation has been conducted by Levchenko, Pitsillidis, Chachra, Enright, Félegyházi, Grier, ... and McCoy (2011), providing the insight that payment processor represents a bottleneck and thereby an interesting point for disrupting the illicit business. This study is different to ours in that it goes towards the physical realm, studying the criminal business operations extensively. This clearly has benefits, but also some drawbacks, such as the ethical issue of completing business with the criminals, the inertia of operations (e.g. shipping), and poorer scalability compared to the cyber realm. Our proposal does not require business interactions, can be conducted with the speed and scale common to the cyber realm, and spans more forms of abuse.

A study on abuse in generic TLDs (gTLDs) was presented by Korczynski, Wullink, Tajalizadehkhoob, Moura, and Hesselman (2017), with an emphasis on how abuse differs between legacy and new gTLDs. In contrast, our work is concerned with a country-code TLD (ccTLD). Data used in their study included WHOIS information, blacklists, DNS zone file, and active measurements, whereas our study only relies on WHOIS data for the heuristics. Their observations indicated that with the emergence of new gTLDs, abusive registrations have shifted from the legacy to the new gTLDs. A strong concentration within particular new gTLDs was observed, while no proof of abuse was found for some new gTLDs. As criminals have demonstrated the ability to migrate to domains better meeting their demands, it is important to study different TLDs, including ccTLDs, to understand if and how abuse differs.

Korczyński, Tajalizadehkhoob, Noroozian, Wullink, Hesselman, and Eeten (2017) also presented a study on “badness” across TLDs, taking both the number of domains found in blacklists and the time until cases of abuse were remediated into account. In this study it was found that abuse was a more significant problem for large gTLDs and less so for new gTLDs, when considering the size of the domains. Abuse was found to be correlated with domain pricing models, DNSSEC[i] deployment rates, and strict registration policies. These findings suggest that the varying traits of TLDs affect the prevalence of abuse, making it relevant to analyse specific TLDs in depth.

Another example of how specific practices in a given TLD can impact abuse can be found in the work of Lauinger, Chaabane, Buyukkayhan, Onarlioglu, and Robertson (2017), who explored “questionable practices” related to re-registrations in legacy gTLDs. These practices are strongly coupled with the registration processes that differ among TLDs.

Finally, we point to a prior study into a specific ccTLD, namely *.eu*, conducted by Vissers, Spooren, Agten, Jumpertz, Janssen, Wesemael, ... and Desmet (2017). Information on registrations for 14 months was compared to blacklists, and blacklisted domains was analysed manually. The authors attributed 80.04 % of the abusive registrations to 20 campaigns. This demonstrates how analysis of domains registered within a ccTLD can provide insight into the activities of cybercriminals.

3 METHODS

With a solid understanding of abuse from the preceding surveys of schemes and techniques, we now move on, towards a method for identifying abuse. The goal is to efficiently identify abusive domain names. First, we reason for an approach of applying heuristics to domain names, as they can be applied at scale. We then present a set of concrete heuristics that can be used to rank domains by how likely it is that they are employing specific techniques. Finally, we describe an approach for manual vetting, which can be applied to the domains that are ranked the highest by the heuristics. By relying on heuristics to focus the manual effort where it is most likely to provide positive identification of abuse, we seek to optimise our method to identify domains that require action, with manual capacity as the constraining resource.

We observe that domain names are a frequent component among the surveyed schemes, and a component of all the techniques. This implies that domains have substantial potential for abuse, but also that analysis of domain names can provide insights to multiple forms of abuse, and that mitigations based on domain names can have serious impact on criminal activity. At the same time domain names can be processed as purely digital entities at a scale and speed that is significantly higher than if the analysis was to also encompass the criminal activities in the physical world.

When analysing abusive domains, a known problem is that the set of abusive domains is not clearly isolated in the set of all domains. This is caused by criminals having the choice to register any free domain, and at the same time they are naturally interested in blending in with legitimate registrations, to improve their chances of success.

A common practice is to rely on blacklist as a mechanism to identify abusive domains. This has the benefit of being a practical solution to the problem, but it relies on the blacklists to be correct, and it does not provide for identifying abusive domains that were not previously blacklisted.

Furthermore, different blacklists often target different types of abuse, based on vaguely defined or unspecified criteria, making them difficult to understand. An additional concern is the extensive discrepancies between blacklists, which adds to the ambiguity. In this work, we define domain abuse as any use that violates accepted terms, applicable law, or the non-conflicting interest of non-criminal users of the Internet.

We analyse domains from blacklists as well as domains sampled randomly from the relevant zone-file, thereby providing insights on the zone and not just on the blacklisted part of it.

We hypothesize that given a solid understanding of the abusive techniques, and of the technical aspects of DNS and domain names, it is possible to describe artefacts of the techniques, which in turn enables us to define heuristics that can be applied automatically and at scale to highlight domains that are likely abusive, such that they can be subjected to deeper manual vetting process to identify abusive domains.

Data collection

This study is enabled by a unique access to information from the *.dk* zone, specifically a list of all 2LDs. Previous studies on abuse are largely focussed on *.com* and *.net*, with some other TLDs also receiving some attention, but this is to the best of our knowledge the first published study of different forms of abuse in the *.dk* TLD.

To gather sufficient details for identifying as many forms of abuse as possible the list of domain names is enriched with a data from the public WHOIS service[ii]. The *.dk* WHOIS service applies a rate limit of 1 query per second, meaning that collecting data for all *.dk* 2LDs would take weeks. As abuse for a given domain might be limited to hours or days this poses a problem. We solve this by limiting our study to 10.000 domains. First, to increase the likelihood of finding abuse, we use all *.dk* 2LDs found in a set of 31 retrievable, public blacklists (Kidmose et al, 2018), as we have higher expectancy of these being abused. Second, we sample the list of all *.dk* 2LDs up to 10.000 total domains. For our subset of *.dk* domains we collect WHOIS data once a day. This strikes a balance between coverage of the zone and frequency of updates, given the boundaries of the rate limit.

Heuristics

Mimicking domains are intended to look similar to other domains, hence our heuristic for this is targeting similarity. Similarity can be expressed in many ways, but as domain names are essentially text strings it is obvious to apply the Levenshtein distance, which describes the minimum number of edits that transforms one string to another (Moore and Edelman, 2010). Clearly the similarity measure needs to be applied to a potentially mimicking domain and a potential target domain. For target domains we expect that criminal focus on popular domains, as they are more likely to impact a larger group of victims. Consequently, our evaluation uses the Alexa top 1 million of popular domain names as targets. We analyse both 2LD labels, such as the *example* part of *example.dk* and the Fully Qualified Domain Name (FQDN) of the *.dk* domains against the 2LD label and FQDN, as the TLD might and might not be part of the mimicry. We expect small editing distances for longer labels/domains to be less likely to occur naturally, meaning they are more interesting if they occur, so we normalise editing distance by label/domain length.

Malicious re-registration occurs very rapidly according to Lauinger, Chaabane, Buyukkayhan, Onarlioglu, and Robertson (2017). While their study is subject to the specific conditions for the *.com* domain, the logical reasoning is generally applicable. Hence, our heuristic for malicious re-registrations is to select the re-registrations that follows the closest after deletion for further inspection.

Double flux is defined as a domain having rapidly changing nameservers, so our heuristic is to rank domains by how frequent the set of nameserver hostnames change in the WHOIS data.

Domain Flux relies on DGA’s to generate domains that appear pseudorandom, as seen in the examples provided by Schiavoni et al., (2014). Our heuristic is the entropy of the distribution of letters within the 2LD labels, which is expected to be high when the letters are pseudo-randomly distributed.

Manual vetting

For each of the above heuristics we obtain a ranked list of most likely abusive domains (Based on the heuristic). Given the lack of ground truth, the top ranked domains are vetted manually, based on the following procedure.

Additional information on the domain is retrieved. This includes any homepage hosted via HTTP(S). The domain is checked against a more extensive set of blacklists[iii], including some that are query-only and not retrievable in full. If the registrant appears to be a Danish company, the official Central Business Registry (CVR) is queried[iv], as for instance missing records are highly suspicious, while long-lived companies are assumed to be less prone to register a domain for abuse. The google search engine is queried for the 2LD and the first 10 results are inspected for any obvious relations to abuse. Finally, the history of WHOIS data is inspected for any content or change that could be relevant to understand if the domain is abused.

4 RESULTS

The following describes the details of the data collection operation, the results of applying heuristics, and the outcome of the subsequent manual vetting. In summary, the heuristics based on editing distance extracted 78 domains which were vetted manually, leading to the conclusion that 5 were actively abused during our observation period, while 22 were cases of defensive registrations. The manual vetting took approximately 4 hours.

Data collection

In accordance with the data collection procedure described in the previous section, 269 *.dk* domains were found on the monitored blacklists, and the remaining 9.731 domains were selected at random among all *.dk* domains. The WHOIS information for these 10.000 domain was retrieved once a day from March 23rd, 2018 to May 29th, 2018 (66 days).

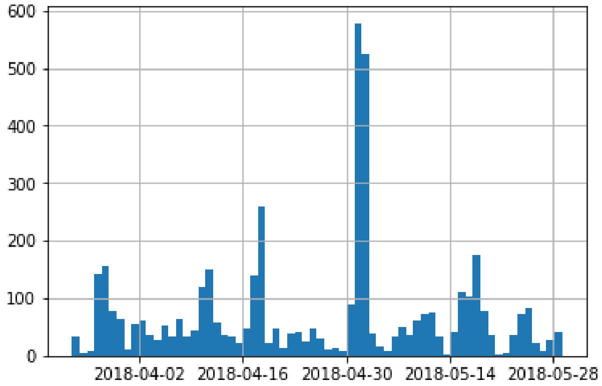

Figure 1: Number of daily updated records.

Figure 1 shows the number of changed WHOIS records per day. A spike at the end of April and the beginning of May was caused by a failure where the records for some domains were flapping to and from empty. Figure 2 represents the same information but cleaned for this error. The cause has not been identified, but as evident from Figure 2 most records resumed the same value after said incident. The WHOIS data for 9.051 domains was unaffected, of the 949 affected domains, 846 domains saw exactly one failure, and six was the highest number of failures for any domain.

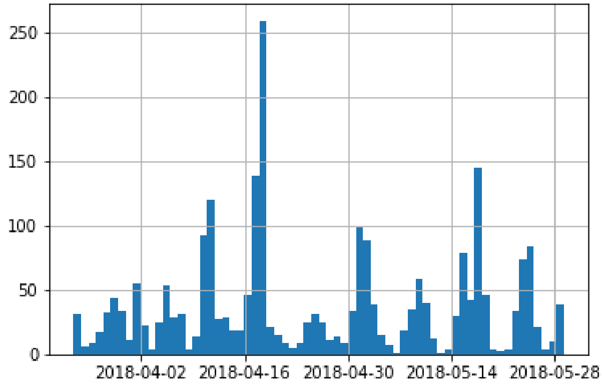

Figure 2: Number of daily updated records with empty results removed.

Editing distance: .dk 2LD labels against Alexa 2LD labels

The editing distances between the 10.000 2LD labels from *.dk* and the 888.876 unique 2LD labels from Alexa were naturally 0 for the 554 labels that was present in both. As lower editing distance is more interesting, we prioritised to analyse these 554. The 2LDs in *.dk* corresponding to 12 of these labels were never active during our observation period, so they were ignored. Reasons for this inactivity was that they expired before our observation, meaning they were observed as “Deactivated” in WHOIS for a while before disappearing, or that the registration never completed, leaving them in a “Reserved” state. In either case, they were not in the *.dk* zone, and not resolvable via DNS, could not have been abused, and therefore they were ignored. Of the remaining 542 labels, the corresponding *.dk* 2LD for 506 labels had been registered for more than a year, which is the minimum registration period. This contradicted the expectation that the period of active abuse is short, so these were ignored, in order to focus on the remaining 36 domains. For 34 domains the manual vetting procedure yielded no indications of abuse. For one label, which coincides with a British menswear brand, the *<LABEL>.com* 2LD was found in Alexa and was registered to the company behind the brand, but *<LABEL>.dk* was registered to an individual with no apparent affiliation to the brand, and the domain is parked (The authoritative nameservers were at *sedoparking.com*). This appeared to be a parked, cybersquatted domain. The remaining one label was related to online marketing, and a marketing company owned the Alexa-listed *<LABEL>.com*. The *<LABEL>.dk* domain was to expire between the end of observations and the time of writing. It was not listed in

WHOIS at the time of writing but was still reported as a phishing domain by Fortinet[v]. The Internet Archive[vi] has a single capture during the suspicious registration period, which shows a minimal frontpage consisting of nothing but some JavaScript and an iframe. It appeared very likely that this was a domain abused for phishing.

TABLE 1: OVERVIEW OF THE 554 2LD LABELS FOUND IN BOTH THE 10.000 *.DK* DOMAINS AND IN THE ALEXA TOP 1 MILLION. FILTERED DOMAINS HAVE NOT BEEN SUBJECT TO MANUAL VETTING, BECAUSE THEY WERE NEVER OBSERVED AS ACTIVE OR WERE OBSERVED AS ACTIVE FOR +1 YEAR.

	# 2LD labels
Never active (Filtered)	12
Active for +1 year (Filtered)	506
Manually vetted	36

In summary, comparing 2LD labels, ignoring inactive and long-lived domains, we manually analysed 36 domains, and found two domains that were likely seeing active abuse. One was a parked cybersquatting of a domain related to a menswear brand. The other had suspicious content and was blacklisted as a phishing site.

TABLE 2: OVERVIEW OF ABUSIVE DOMAIN FOUND BY COMPARING 2LD LABELS AND SUBSEQUENT MANUAL VETTING.

	# 2LDs
Cybersquatting brand/web shop, parked	1
Phishing	1

Editing distance: .dk 2LD FQDNs against Alexa FQDNs

The normalised editing distance between FDQNs highlighted 31 *.dk* domains as they were also found in Alexa, resulting in a distance of zero. This is of course not abuse and these were filtered out. Subsequently, we manually analysed the 50 pairs of *.dk* and Alexa FQDN's that were closest, without being exact matches. Seven of the top 50 pairs were the legitimate domain *triumphmotorcycles.dk* paired with the same 2LD label in *.de, .be, .ch, .es, .fr, .in,* and *.it* TLDs. Similarly, three pairs were the legitimate *blogspot.dk* paired with the *.de, .mk,* and *.sk* variants.

For all 50 pairs, we subjected the 2LD from *.dk* to our manual vetting process, with the following results. Data on the Alexa domain from the same sources was also considered where relevant. 20 domains appeared to be used for hosting legitimate websites of enterprises, smaller companies, and private persons, with no suspicion raised through the manual vetting process. 22 domains appeared to be defensive registrations by the owners of the corresponding Alexa domain, either redirecting accordingly or parked with an apparently conscientious, add-free provider. Of these 22, 14 where defensive registrations for other *.dk* domains, six where for domains in the *.de* TLD, and remaining two are those pertaining to *blogspot* label as mentioned above.

For three of the 50 pairs, the *.dk* 2LD was registered, was not found to be related to abuse, nor did we find any indication that it was actively used, e.g. with a webpage or Google results that include an e-mail address at the domain. These domains appeared to be unused.

A summary of the non-abusive domains captured by the normalised editing distance between FQDNs is as follows: 20 domains seeing ordinary use, 22 cases of defensive registrations, and three passive domains.

One of the 42 *.dk* 2LD domains (50 pairs, 42 unique *.dk* domains) was found to be close to a service for price comparison service also in *.dk*, but with no apparent affiliation, and with the domain parked with *parkingcrew.net*. We strongly suspect this to be a typosquatting domain that was active in the period of our study, and still was at the time of writing. Two other pairs were cases of the *.dk* FQDN being similar to a *.de,* with 2LD labels matching, while the *.dk* domains redirected to parking services. We believe these two was typo-squatted domains, that monetized through parking. Finally, we also found one domain that was close to a domain with outdoor lifestyle content and a web shop. The suspected domain had been seized by authorities prior to our observations, the Internet Archive has records of a web shop which appeared highly suspicious, and it was blacklisted by Web of Trust[vii] as a scam/counterfeit web shop.

TABLE 3: OVERVIEW OF PAIRS WITH LOW EDITING DISTANCE (LEVENSHTEIN) BETWEEN ONE OF 10.000 *.DK* 2LD DOMAINS AND A FQDN FROM THE ALEXA TOP 1 MILLION. RIGHTHAND COLUMN LIST COUNT OF UNIQUE 2LDS FROM *.DK*. WHERE RELEVANT NUMBER OF UNIQUE PAIRS ARE LISTED IN PARENTHESIS. NOTE THAT A .DK DOMAIN CAN BE IN MULTIPLE PAIRS.

	# Unique 2LDs (pairs)
Manually vetted	42 (50)
Legitimate use (excluding defensive)	20
Defensive registrations	(22)

- Defensive, paired with Alexa domain from**.dk*	8 (8)
- Defensive, paired with Alexa domain from **.de*	6 (6)
- Defensive, paired with Alexa domain from **.mk* or **.sk*	1 (2)
Not in use	3
Typosquatting of price comparison service, parked	1
Typosquatting (*.de* vs. *.dk*)	2
Typosquatting (web shop, Seized prior to our study)	1

Reregistration and High Entropy

Among the 10.000 domains that was observed, three were successfully reregistered during our 66 days observation period, with lags of 14, 15, and 152 days respectively. We found no evidence of abuse and assume all cases to be legitimate registrations.

For the 50 domains that had the highest entropy, we found no indications of active abuse. One domain had previously been seized by the authorities, but before our observations started, and due to trademark infringements and scams, which appears unrelated to the entropy. All of the 50 domains appeared human readable.

5 DISCUSSION

We were able to apply theory about how criminals operate their business and what techniques they use to devise heuristics that automatically can prioritise domains for manual scrutiny. In the case of the Levenshtein editing distance as a heuristic for identifying the mimicking technique the automated procedures processed 10.000 domains and prioritised 78 domains. 36 of these domains were identified as 2LD labels that were found in both *.dk* and Alexa. 42 of these were identified as unique *.dk* FQDNS that was in the top 50 of similar FQDN pairs, compared to Alexa FQDNs. The discrepancy between the total of 78 domains and contributions of 42 plus 50 stems from the intersection between the two contributions. Among the 72 domains, we are confident that five domains were being actively abused during our observation period: Four cases of apparent typosquatting with redirection to parking services, and one case of phishing. Additionally, we identified 22 cases of defensive registrations, which either have been abused previously and then seized, or the current owner have deemed is so likely to be registered for abuse that it is worth acquiring. In either case, it supports that our heuristic is suitable for identifying relevant domains.

Considering that manual analysis of a domain takes a few minutes, we are able to identify five cases of abuse (and 22 defensive registrations) by investing about a half a working day (78 domains at 3 minutes per domains ≈ 4 hours). We hypothesize that if the entire procedure is applied to a larger set of domains, while the number of top-ranked domains subjected to manual vetting is kept fixed, the number of identified abusive domains is expected to be even better. This is based on the assumption that the expected prevalence of abuse increases with higher relative ranks, and the fact that with a larger set of domains the top-N (with fixed N) will correspond to a relatively smaller part of all the domains. This naturally prompts for validation through studies on larger scale, such as the entire *.dk* domain, in order to support or dismiss the hypothesis.

In our study, the heuristic for abusive re-registrations fails to capture any abuse. This can be because the manual vetting process fails to discover present abuse, or because abusive re-registration does not occur for the *.dk* zone. Another more likely explanation is that the data set is too small and therefore does not contain abuse. Among the 10.000 domains observed for 66 days, we only observed 3 re-registrations. No abusive re-registrations can be expected from such a small set. To evaluate this heuristic the data set must be extended, which can be done by including more domains and by observing for a longer period of time. Extrapolating the observed frequency of re-registrations to the 1.3 million domains in *.dk*, 390 re-registrations can be expected to occur in a 66 days period. While the ratio of abusive re-registrations is not known, 390 re-registrations still appear to be a low count, so expanding the observation period also appears necessary. Alternatively, evaluation can be extended to more or larger TLDs.

The heuristic for entropy also fails to yield any domains applying domain flux in the top-50, with one apparently irrelevant exception (A domain abused and seized prior to our observations, with a human readable label that did not appear pseudo random, i.e. apparently not from a DGA). Like for re-registrations, this can be explained by errors in the manual vetting process or by the evaluation data not containing examples of DGA domain. We are inclined to rule out errors in the vetting process, as DGA domains are expected to clearly stand out simply by the pseudo randomness apparent from the label, see for instance Schiavoni et al., (2014). It is possible that there are no cases of domain flux in the data, or perhaps even in the *.dk* zone. The price of registering a *.dk* domain, the registration process, which involves a process for registrant identity validation, the legal requirement for public whois, and other specifics of the *.dk* might divert certain forms of abuse to other TLDs. It is also possible that criminals have improved their DGAs to be more stealth. This appears to be possible by generating domains that are closer to legitimate domains in some lexical sense, such as character or N-gram distributions, or simply by combining random words. In this case, a new heuristic must be devised. As per Section 3 this prompt for a study of the novel techniques, if such exist, which would merely be guesswork without examples.

Section 3 omits heuristics for some of the techniques that was discussed in Section 2 This is so because we believe the techniques to be relevant for designing heuristics, but we have not been able to devise heuristics for these techniques given that available data. As described, Fast Flux exploits rapid changing A records on the authoritative nameserver, which is chosen by the registrant and not operated by the registry, therefore we do not have access to the master data. The nameserver operators are likely bound by confidentiality to their client, the registrant, and are perhaps also accomplices to abuse. Possible approaches to overcome this are passive DNS traffic monitoring and active probing. In the case of bulk registrations, practical limitations lead us to select a subset of domains for observation, as described in Section 3 , which lead to our data not providing insights on this technique. The solution, which is in development, is to monitor the entire *.dk* zone, including new registrations.

6 FUTURE RESEARCH DIRECTIONS

Having demonstrated that the approach of analysing abuse techniques, devising heuristics, ranking domains, and manually vetting the domains ranked as most likely to be abusive is valid, at least for the used data set and some of the techniques considered, we would like to expand the study to obtain more general results. The most obvious first step is to analyse the entire *.dk* zone, as the limited number of domains is a recurring issue, as evident from the above discussion. This implies some practical challenges that we are currently working on. Similarly, expanding the duration of the observation is relevant, and this is more straightforward. Extending to other TLDs is also a possibility, but this is subject to the details available in the data. As registries and TLDs differ, the current heuristics might not apply, but this only means that it is a possibility to evaluate our entire approach, including the analysis of abuse techniques.

Some techniques are presumably not evident in the WHOIS data, as discussed above, so passive DNS traffic monitoring and active probing of recursive name servers are under consideration as data sources for further studies. Specifically, we are currently investigating the OpenINTEL framework[viii], which can enable heuristics for the Fast Flux techniques and more.

In general, we still see domain names as an essential link between the physical and cyber realms, which is enabling for the majority of both legitimate and criminal activity involving cyber, and therefore a key point for attacking the criminal activity which evidently persists.

7 CONCLUSION

We have described the extent of cybercriminal business and surveyed the abusive schemes and techniques that employ domain names to enable the crime, but also make domain names and the Domain Name System a choke point for combatting cybercrime. We have proposed an approach of defining heuristics for abusive domain names, based on the abusive techniques, and present a set of such heuristics. We have demonstrated that our heuristics can be applied to focus manual effort, allowing us to identify five abusive domains among 10.000, with four hours of manual effort. Specifically, we extracted 36 *.dk* 2LDs automatically with a heuristic based on matching *.dk* 2LD labels to 2LD labels of the Alexa top 1 million, in combination with WHOIS information on domain state. Through a manual vetting process, two of the 36 were identified as positive cases of abuse. Another heuristic based on editing distances between FQDNs from *.dk* and FQDNs from Alexa was used to obtain a top 50 of most similar FQDNs. These pairs represent 42 interesting domains, which was also subjected to manual vetting, and three was found to be active typosquatting of a malicious nature. Among the 50 pairs we also found 22 examples of defensive registrations and one seized typosquatting domain. A heuristic based on reregistration was proposed, but results were inconclusive as the data only holds a total of three reregistration. Expanding the data with more domains and a longer observation can possibly improve on this. A heuristic based entropy was applied, but no abuse was identified among the top 50 of heuristically ranked domains, either because the heuristic fails or because the targeted technique is uncommon in the *.dk* zone. In summary, we have contributed by detailing our efficient heuristic approach, and by providing the first scientific study on abuse in the *.dk* country code Top-Level Domain.

8 REFERENCES

Abbasi, A., & Chen, H. (2009). *A comparison of tools for detecting fake websites*. Computer, (10), 78-86.

Gyöngyi, Z., & Garcia-Molina, H. (2005). *Spam: It's not just for inboxes anymore*. IEEE Computer, 38(10), 28-34.

Hao, S., Thomas, M., Paxson, V., Feamster, N., Kreibich, C., Grier, C., & Hollenbeck, S. (2013, October). *Understanding the domain registration behavior of spammers*. In Proceedings of the 2013 conference on Internet measurement conference (pp. 63-76). ACM.

Holz, T., Gorecki, C., Rieck, K., & Freiling, F. C. (2008, February). *Measuring and Detecting Fast-Flux Service Networks*. In NDSS.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., & Savage, S. (2008, October). *Spamalytics: An empirical analysis of spam marketing conversion*. In Proceedings of the 15th ACM conference on Computer and communications security (pp. 3-14). ACM.

Kidmose, E., Gausel, K., Brandbyge, S., & Pedersen, J. M. (2018, November). *Assessing usefulness of blacklists without the ground truth*. Paper presented at 10th International Conference on Image Processing and Communications, Bydgoszcz, Poland.

Korczyński, M., Tajalizadehkhoob, S., Noroozian, A., Wullink, M., Hesselman, C., & Eeten, M. Van. (2017). *Reputation Metrics Design to Improve Intermediary Incentives for Security of TLDs*. Proceedings - 2nd IEEE European Symposium on Security and Privacy, EuroS and P 2017, 579–594. https://doi.org/10.1109/EuroSP.2017.15

Korczy'ski, M., Wullink, M., Tajalizadehkhoob, S., Moura, G. C., & Hesselman, C. (2017). *Statistical Analysis of DNS Abuse in gTLDs Final Report*. Technical Report. https://www.icann.org/en/system/files/files/sadag-final-09aug17-en.pdf.

Lauinger, T., Chaabane, A., Buyukkayhan, A. S., Onarlioglu, K., & Robertson, W. (2017, August). *Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers*. In Proceedings of the USENIX Security Symposium.

Lauinger, T., Chaabane, A., Buyukkayhan, A. S., Onarlioglu, K., & Robertson, W. (2017, August). Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In Proceedings of the USENIX Security Symposium.

Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Félegyházi, M., Grier, C., ... & McCoy, D. (2011, May). *Click trajectories: End-to-end analysis of the spam value chain*. In 2011 ieee symposium on security and privacy (pp. 431-446). IEEE.

Lever, C., Walls, R., Nadji, Y., Dagon, D., McDaniel, P., & Antonakakis, M. (2016, May). *Domain-Z: 28 registrations later measuring the exploitation of residual trust in domains*. In Security and Privacy (SP), 2016 IEEE Symposium on (pp. 691-706). IEEE.

Lewis, J. (2018, February). *Economic Impact of Cybercrime - No Slowing Down*. McAfee, LLC and Center for Strategic and International Studies, Available at https://www.mcafee.com/enterprise/en-us/assets/reports/restricted/rp-economic-impact-cybercrime.pdf. Retrieved on 18. September 2018.

McGuire, M. (2018, April). *Into the Web of Profit*. Bromium, Inc. Available at https://learn.bromium.com/rprt-web-of-profit.html. Retrieved on 18. September 2018.

Moore, T., & Edelman, B. (2010, January). *Measuring the perpetrators and funders of typosquatting*. In International Conference on Financial Cryptography and Data Security (pp. 175-191). Springer, Berlin, Heidelberg.

Moura, G. C., Müller, M., Davids, M., Wullink, M., & Hesselman, C. (2017, May). *Domain names abuse and TLDs: From monetization towards mitigation*. In Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium on (pp. 1077-1082). IEEE.

Nazario, J., & Holz, T. (2008, October). *As the net churns: Fast-flux botnet observations*. In Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on (pp. 24-31). IEEE.

Porras, P., Saidi, H., & Yegneswaran, V. (2009). *An analysis of conficker's logic and rendezvous points*. 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '09)

Schiavoni, S., Maggi, F., Cavallaro, L., & Zanero, S. (2014, July). *Phoenix: DGA-based botnet tracking and intelligence*. In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 192-211). Springer, Cham.

Sood, A. K., Bansal, R., & Enbody, R. J. (2013). *Cybercrime: Dissecting the state of underground enterprise*. IEEE internet computing, 17(1), 60-68.

Vissers, T., Joosen, W., & Nikiforakis, N. (2015, February). *Parking sensors: Analyzing and detecting parked domains*. In Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015) (pp. 53-53). Internet Society.

Vissers, T., Spooren, J., Agten, P., Jumpertz, D., Janssen, P., Van Wesemael, M., ... & Desmet, L. (2017, September). *Exploring the Ecosystem of Malicious Domain Registrations in the .eu TLD*. In International Symposium on Research in Attacks, Intrusions, and Defenses (pp. 472-493). Springer. https://doi.org/10.1007/978-3-319-66332-6

KEY TERMS

DNS: Domain Names System. A distributed, hierarchical database of, among other things, mappings from human readable domain names to machine readable IP addresses.

Domain Name: A string that identifies an autonomous subset of the Internet typically corresponding to a personal or organisational entity.

Abuse: Use that violates accepted terms, applicable law, or the non-conflicting interest of non-criminal users of the Internet.

Heuristics: Practical and applicable methods that are not necessarily optimal, perfect, or theoretically derived.

BIOGRAPHICAL NOTES

Egon Kidmose received the B.Sc.Eng. (Computer Engineering) in 2012 and his M.Sc.Eng. (Networks and Distributed Systems) in 2014, both from Aalborg University, Denmark. He is currently pursuing a Ph.D. in a cooperation between Aalborg University and LEGO System A/S, where he is employed to research topics within IT security, including a corporate and applied perspective. Interests and areas of research includes network security, incident detection, machine learning and application of Big Data methods for security.

Erwin Lansing received the M.Sc. in Biology in 1998 at the University of Utrecht. Since 2000, he has held several positions in the ICT industry, including network operations, server hosting and automation. Currently, he holds the position of Head of Security and Technical Advisor at DK Hostmaster A/S in Copenhagen, Denmark. His interests include network and information security, DNS and domain names, data analysis, and machine learning. He has also been a long-time open source software developer.

Søren Brandbyge received the M.Sc. in Food Science & Technology from The Royal Veterinary High School, Copenhagen, Dep. of Process Technology in 1988, specialized in mathematical modelling and simulation, and a M.Sc. in Information Technology from Center for Interactive Medias at Syddansk University & Dep. of Information and Media Studies at Aarhus University, in 2007. Since 2002 he has held various security & network roles as Infrastructure Engineer at LEGO System A/S, Billund Denmark. His interests include asynchronous process communication, network security & planning with focus on cryptography and anomaly-detection in highly heterogenous information networks.

Jens Myrup Pedersen received the M.Sc. in Mathematics and Computer Science in 2002, and the Ph.D. in Electrical Engineering in 2005 from Aalborg University, Denmark. He is currently Associate Professor at the Wireless Communication Section, Department of Electronic Systems, Aalborg University. His research interests include network planning, traffic monitoring, and network security. He is author/co-author of more than 120 publications in international conferences and journals, and has participated in Danish, Nordic, and European funded research projects. He is also board member of a number of companies within technology and innovation.

REFERENCE

Reference to this paper should be made as follows: Kidmose, E., Lansing, E., Brandbyge, S., & Pedersen, J. M. (2018). Heuristic methods for efficient identification of abusive domain names. *International Journal on Cyber Situational Awareness*, Vol. 3, No. 1, pp 121-142.

[i] RFC 3833 https://tools.ietf.org/html/rfc3833

[ii] https://github.com/DK-Hostmaster/whois-service-specification

[iii] https://www.urlvoid.com

[iv] https://datacvr.virk.dk/data/

[v] https://fortiguard.com/webfilter?q=<label>.dk

[vi] http://web.archive.org/

[vii] https://www.mywot.com

[viii] https://www.openintel.nl