A Case of Identity: Detection of Suspicious IDN Homoglyph Domains Using Active DNS Measurements

Ramin Yazdani
University of Twente
Enschede, The Netherlands
r.yazdani@utwente.nl

Olivier van der Toorn
University of Twente
Enschede, The Netherlands
o.i.vandertoorn@utwente.nl

Anna Sperotto
University of Twente
Enschede, The Netherlands
a.sperotto@utwente.nl

***Abstract*—The possibility to include Unicode characters in domain names allows users to deal with domains in their regional languages. This is done by introducing Internationalized Domain Names (IDN). However, the visual similarity between different Unicode characters - called homoglyphs - is a potential security threat, as visually similar domain names are often used in phishing attacks. Timely detection of suspicious homoglyph domain names is an important step towards preventing sophisticated attacks, since this can prevent unaware users to access those homoglyph domain that actually carry malicious content. We therefore propose a structured approach to identify homoglyph domain names based not on use, but on characteristics of the domain name itself and its associated DNS records. To achieve this, we leverage the OpenINTEL active DNS measurement platform, which performs a daily snapshot of more than 65% of the DNS namespace. In this paper, we first extend the existing Unicode homoglyph tables (confusion tables). This allows us to detect on average 2.97 times homoglyph domains compared to existing tables. Our results show that we are able to identify suspicious domains on average 21 days before those appear in blacklists.**

***Index Terms*—homoglyph, IDN, homograph attacks, suspicious domains, active DNS measurements**

1. Introduction

Domain names in the Domain Name System (DNS) are encoded using the American Standard Code for Information Interchange (ASCII). The standard uses 8 bits to encode alphanumeric characters. The Unicode standard uses a maximum of four bytes to perform the encoding, allowing for a much larger character set to be encoded, e.g., Greek, Cyrillic, Arabic, or Chinese characters. The Internationalized Domain Name (IDN) [1] provides a method for using Unicode characters in domain names, allowing the usage of regional alphabet in domain names. However, a major security risk is introduced along with IDNs. The Unicode system contains characters that are visually similar to other Unicode or ASCII characters, called homoglyphs. An attacker can register a domain visually indistinguishable from an ASCII counterpart using homoglyphs, for example to perform a phishing attack [2]. In this paper we investigate the size of the problem in the case of Unicode – ASCII homoglyphs. We propose a simple method for proactively detecting suspicious IDNs. Since our approach is based not on use, but on characteristics of the domain name itself and its associated DNS records, we are able to detect such domains before they are involved in malicious activities. The main contributions of this paper are that we:

- propose an improved Unicode Confusion table able to detect 2.97 times homoglyph domains compared to the state-of-the-art confusion tables;
- combine active DNS measurements and Unicode homoglyph confusion tables to detect suspicious IDN homoglyphs. In doing so we introduce a potential time advantage between the moment of detecting suspicious domains and being listed by publicly available blacklists.

The remainder of this paper is organized as follows. In Section 2, the background of IDNs in DNS and homoglyph domains are given. Section 3 discusses the related works in the literature. Section 4 introduces the datasets used in this research. In Section 5, the proposed methodology is presented. Results of our study are presented in Section 6. Ethical considerations are discussed in Section 7. A discussion around drawbacks which still need to be addressed is given in Section 8. Finally, Section 9 concludes the paper.

2. IDN Primer

The Unicode system incorporates numerous writing systems and languages, in which many homoglyph characters exist, such as the Greek capital letter omicron “O” (U+039F), Latin capital letter “O” (U+004F), and Cyrillic capital letter “O” (U+041E). These letters are assigned to different code points, but visually appear to be indistinguishable, or very similar. The DNS is designed with ASCII in mind. In order to keep backwards compatibility and avoid the need to upgrade existing infrastructure, IDNs are converted into an ASCII-Compatible Encoding (ACE) string, which is done using the ‘Punycode’ [3] algorithm. This algorithm keeps all ASCII characters and encodes the non-ASCII characters alongside their position in the original string using a generalized variable-length integer for each non-ASCII character. Finally, an ‘xn--’ prefix is added to indicate the use of Punycode. This process allows the DNS to accept IDNs without any upgrade and is typically reversed before the domain name is presented to a user, by a browser for example.

3. Related Work

Existing research about IDN homoglyphs can be divided in two main groups: the studies trying to construct Unicode confusion tables, and the ones on detecting homoglyph domains.

3.1. Unicode Confusion Tables

Fu et al. [4] have constructed a Unicode Character Similarity List (UC-SimList) using a visual similarity formula based on pixel overlap, covering English, Chinese and Japanese scripts. A similarity threshold is considered to select characters considered as homoglyphs. Roshanbin et al. [5] propose a comparable method to create a similarity list using the Normalized Compression Distance (NCD) metric to determine the similarity between Unicode characters. Suzuki et al. [6] build a Unicode homoglyph table called 'SimChar' using the pixel overlap of the characters. They apply this table to IDNs in '.com' Top Level Domain (TLD) to extract homoglyph domains of the top-10K Alexa list. To the best of authors' knowledge there is no prior work evaluating the quality of existing Unicode confusion tables when applied to domain names.

3.2. Homoglyph Detection

Liu and Stamm [7] use the UC-SimList to detect Unicode Obfuscated spam messages. Alvi et al. [8] focus on detecting plagiarism where Unicode characters are used for obfuscation. Their method uses the 'Unicode Confusables' list[1] and the normalized hamming distance. A tool called REGAP is proposed in [9], where a keyword level Non-deterministic Finite Automaton (NFA) is used to identify potential IDN-based phishing patterns. Contrary to our work, this approach requires manual intervention limiting the number of investigated domains.

Krammer et al. [10] and Al Helou et al. [11] propose improved user interfaces for browsers to defend against phishing attacks. The client-side anti-phishing browser extension prints characters of the Unicode subsets in multiple colors in the address bar. Although browser based solutions are helpful, they do not prevent naive users from clicking on a malicious homoglyph domain link on a webpage or an email.

Shirazi et al. [12] propose a phishing domain classification strategy which uses seven domain name based features, modeling the relation between the domain name and the visible content of a Web page. Considering the fact that not all homoglyph domains have a Web page, this method cannot be applied at scale. Holgers et al. [13] perform a measurement study by first passively collecting a nine-day-long trace of domain names accessed by users in their department and then generating corresponding homoglyph domains. The subsequent step in their work was to perform active measurements against the confusable domains to determine if they are registered and active. Both ASCII and Unicode homoglyphs of characters were investigated in their study. However this is possible when dealing with a limited number of domains and is computationally expensive otherwise. Qiu et al. [14] propose a Bayesian framework to calculate the likelihood a character in a domain name is suspicious (visual spoofing).

A group of studies [15]–[17] investigate homoglyph IDNs targeting top brand domains by processing the similarity of the domain name images. Although image-based methods bypass the problem of needing a homoglyph table, this approach is limited to protecting a number of brand domains. Elsayed et al. [18] extract newly registered Unicode domains from DNS zone files for '.com' and '.net' TLDs and replace the Unicode characters by their ASCII homoglyph counterparts based on the 'Unicode Confusables' list, to determine possible phishing domains. They also make use of the WHOIS data to differentiate between malicious and protective domains. Quinkert et al. [19] extract IDN homoglyphs targeting top 10K from the Majestic top 1 million domains [20] using the 'Unicode Confusables' list. Our method differs from this last group of studies because it does not make assumption on the nature of the name, but considers all existing domain names. Besides, we replace the use of WHOIS data with DNS measurement, as WHOIS crawling is notoriously error-prone and sometimes not even feasible due to WHOIS privacy protection.

4. Datasets

This section discusses the details of the dataset used in this work. Our dataset comes from the OpenINTEL, which is an active DNS measurement platform, measuring more than 65% of the entire DNS namespace on a daily basis. The platform queries domains for their 'A', 'AAAA', 'MX', 'NS' records and more. In this paper we use the data from 2018-01-01 through 2019-11-30 for the '.com', '.net' and '.org' TLDs and the '.se', '.nu', '.ca', '.fi', '.at', '.dk' and '.рф' country-code Top Level Domain (ccTLD)s. For comparison purposes we use data from publicly available blacklists which is measured by OpenINTEL, namely 'Hostfile', 'hpHosts', 'Ransomwaretracker', 'Openphish', 'Malwaredomainlist', 'Joewein', 'Threatexpert', 'Zeustracker' and 'Malcode'[2]. In the rest of the paper we refer to this set as 'RBL'. The date range of blacklist data is from 2018-01-01 till 2019-12-15 to give domains registered at the end of our dataset a chance of appearing on a blacklist. As Unicode Confusion tables, the 'Unicode Confusables' list, published by the Unicode Consortium (version 12.0.0) and the 'UC-SimList0.8' [4], together with an improved confusion table based on these two tables are used.

5. Methodology

In this section we present our methodology. A high-level view of the proposed detection mechanism for suspicious IDN homoglyphs is shown in Fig. 1. The approach is divided in five major steps, (1) through (5). We have applied the above process on each day of our dataset. Details of these steps are elaborated in the following subsections.

5.1. IDN Extraction

All IDNs start with a 'xn--' prefix. In the first step we filter out the domains from our dataset that contain this

1. https://unicode.org/Public/security

2. https://www.tide-project.nl/blog/unicode_homoglyphs

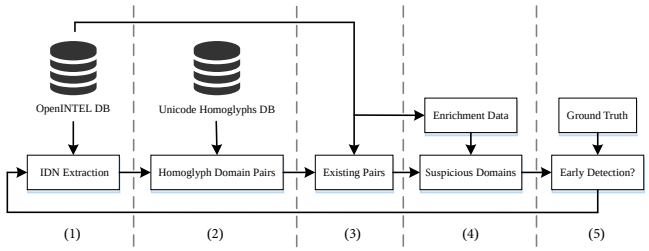

Figure 1. High-level overview of the proposed method

prefix. This filtering gives us a first indication on how large the problem of suspicious Unicode homoglyph domains could be at a maximum. According to the 2019 IDN report provided by EURid [21], there were approximately 7.5 and 9 million IDNs by the end of 2017 and 2018, respectively, which accounts for approximately 2% of the entire DNS namespace. In 2018 there were 7 million (78%) IDNs registered under a ccTLD, confirming the importance of including ccTLD in the investigation of suspicious homoglyph domains.

5.2. Homoglyph Domains

In this step we filter the Unicode domains containing homoglyph characters from the set of Unicode domains extracted in the previous step using the Confusion tables mentioned in Section 4. We have noticed some irregularities in the "UC-SimList0.8". For example, the Latin small letter dotless I 'ı' (U+0131) is considered a homoglyph of exclamation mark '!' (U+0021). However, we argue that the Latin small letter I 'i' (U+0069) would be a better choice for our purpose, since the exclamation mark is an illegal character in DNS labels. This finding urged us to explore the quality of the confusion tables, with regard to our goal. We introduce a third table, which mixes the 'Unicode Confusables' and 'UC-SimList0.8', but replaces irregularities and adds missing characters. The proposed table is publicly available online[3]. Specifications of each Unicode confusion table is given in Table 1. Comparing the two existing homoglyph tables, we observe that while the 'UC-SimList0.8' contains more characters in total, it covers much fewer characters with an ASCII homoglyph than the 'Unicode Confusables' table. This is because the 'UC-SimList0.8' covers many characters from the Chinese and Japanese alphabet for which an ASCII homoglyph does not exist. Another major difference between these two tables is that the 'Unicode Confusables' table provides homoglyph strings for Unicode characters that can not be replaced by a single character. On the other hand, the 'UC-SimList0.8' does provide multiple homoglyphs for each Unicode character, if the homoglyphs exist, ordered by the degree of similarity. In this paper we use the ASCII homoglyph with the highest similarity score. We realize that a Unicode character may have Unicode homoglyphs, but due to computation restraints we focus on ASCII homoglyphs in this paper. We compare the performance of the three confusion tables, with respect to our goal, in Section 6.2.

3. https://www.tide-project.nl/blog/unicode_homoglyphs

TABLE 1. UNICODE HOMOGLYPH CHARACTER TABLES

	Unicode Confusables	UC-SimList0.8	Proposed table
Total character pairs	6296	29880	2627
Characters with an ASCII homoglyph	2236	536	2627
Character to string mapping	✓	✗	✓

5.3. Existing Homoglyph Pairs

At this stage we have a list of Unicode domains with their ASCII counterparts. Since these ASCII counterparts are 'fabricated' domains, we need to determine if it concerns registered domain names. We do so by querying our dataset for these ASCII homoglyph domains. If the ASCII homoglyph domain is not present in our dataset we discard the entire pair.

5.4. Suspicious Homoglyph Detection

In this step we investigate if the Unicode homoglyph domain has the same owner as the ASCII counterpart, as the Unicode domain may be a protective registration. For this purpose we use the 'AS' number, and the 'A', 'AAAA', 'NS' and 'MX' records of the two domains retrieved from the OpenINTEL platform. We use these record types since these are frequently used by domains, and will likely point to the same addresses in the case of a protective registration. A normalized 'suspiciousness' score is calculated by counting the number of differences between the existing parameters divided by the number of existing records. However, a record is counted as existing only if both sides have an entry for it. Hence, the normalized suspiciousness score will be a discrete real value in [0,1], where 0 means the Unicode domain is likely a protective registration and a score of 1 suggests suspiciousness. If the score exceeds a threshold we mark the Unicode domain as suspicious. We determine the threshold by calculating the suspiciousness score for the detected domains which have appeared on the blacklists. Based on the distribution we choose a cut-off point, the threshold, which captures more than 75% of the blacklisted domains.

5.5. Time Advantage

To explore the potential achievable time advantage we run our method on historic OpenINTEL data and compare this to historic blacklist data, considering the dates when we detect suspicious homoglyph domains and when these appear on any of the blacklists. Although our method does not aim to detect malicious domains, it gives an early notification on domains that need to be further investigated to discover possible malicious activity. One way to do this is studying the web pages of these domains (if existed), however this is beyond the scope of this paper.

6. Results

6.1. IDN growth

The growth of IDNs in '.com', '.net', '.org' and ccTLDs are depicted in Fig. 2. Two deeps are visible

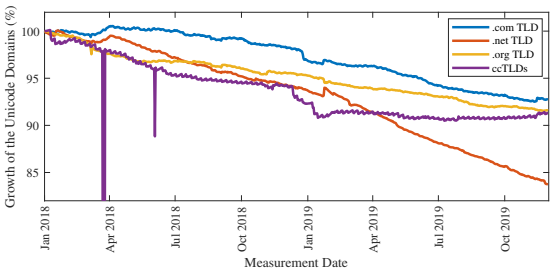

Figure 2. Relative growth of IDNs in '.com', '.net', '.org' and ccTLDs

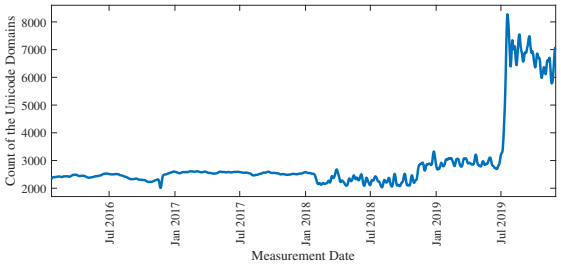

Figure 3. IDNs in Alexa top 1M domains

for ccTLDs which are due to measurement errors in the OpenINTEL platform. A negative growth trend is seen for IDNs in any of the four sets of domains. Specifically, IDNs in '.com' have the least decrease and shrink to 92.7% at the end of the study period compared to the beginning and the IDNs in '.net' have the steepest descent by shrinking to 83.7% at the end of the period. This is in line with the 2019 EURid report, showing a negative growth of 13% for IDNs under generic TLDs. Considering the 7.5 million IDNs on December 2017 as reported by EURid, our dataset contains approximately 30% of all Unicode domains. Additionally, EURid reported in 2017 that 48% of IDNs use Han script for which no ASCII homoglyph exists, making our dataset extremely valuable for performing detection of suspicious IDN homoglyph domains.

Fig. 3 depicts the number of IDN domains on the Alexa top 1 million from 2016-01-22 till 2019-11-30. From 2018-02-01 onwards extreme oscillations can be observed. The reason behind this behaviour is the change of rank calculation policy used by Alexa from one month to a one day average [22]. The average number of IDN on Alexa roughly contributes to 0.3%, which leads us to conclude that IDN domains are not as popular as ASCII domains. However, the number of IDNs on Alexa has significantly increased from July 2019.

6.2. Comparison of Confusion Tables

Extraction of homoglyph domain pairs is done using the three Unicode confusion tables discussed in Section 5.2. Fig. 4 and Fig. 5 depict the number of IDNs added per day and the total number of Unicode characters in the added domains in '.com' TLD, respectively. To improve readability of the figures we applied a moving average filter and log-scaled the y-axis. Fig. 4 shows that, on average, the confusion table we propose extracts up to 6 times homoglyph domains compared to 'Unicode Confusables' and up to 1.5 times the 'UC-SimList0.8' since the proposed list covers both tables with additional missing characters. Similar results are achieved considering the number of domains and characters corresponding to '.net' and '.org' TLDs.

Table 1 draws the expectation that the 'Unicode Confusables' table is able to extract more domains than the 'UC-SimList0.8' since the list contains more characters with an ASCII homoglyph. However, the 'UC-SimList0.8' outperforms the 'Unicode Confusables' table in this case. The main cause is the punctuation characters from the 'Unicode Confusables' table which rarely appear in a domain name.

Furthermore, we observe that the addition of a modest set of missing characters to our proposed table makes a large difference in the number of extracted domains, implying that frequently used Unicode homoglyphs are not covered by existing tables. This observation can be further quantified by investigating the amount of Unicode characters present in newly observed IDNs per day compared to the number of characters covered by the tables. Fig. 5 plots the number of total Unicode characters and the Unicode characters covered by each table for IDNs added per day in '.com' TLD.

6.3. Detection results

Since a Unicode homoglyph domain does not automatically mean it is suspicious, we compute a suspiciousness score to differentiate between protective registrations and suspicious domains. For this purpose we queried the extracted homoglyph domain pairs for the additional records (see Section 5.4 for details). The normalized score of suspiciousness is calculated by summing the number of mismatches between record values divided by the number of records which exist on both sides. If one of the two sides does not have a value for a particular record type that record type is ignored in the calculation of the score. We have chosen for this approach to error on the side of caution in case of missing records. A higher score makes the Unicode domain more suspicious. Fig. 6 shows the distribution of these scores for the detected homoglyph domain pairs. In the first glance it is seen that the homoglyph domain pairs are mainly concentrated in two ends of the score range, achieving a score of zero corresponding to no difference in records or a score of one corresponding to all of the records being different. Besides, a large portion of the homoglyph domain pairs from the ccTLDs achieve a score of zero (66.4%), while '.com' has a large group of domain pairs with a score of one (61.5%). This suggests a relatively higher chance of malicious intent behind IDNs extracted from '.com' compared to the ccTLDs.

To determine the threshold of when to mark a domain as suspicious we calculated the normalized score for detected IDNs appearing on blacklists. In Fig. 6 the distribution of scores of these domains is shown. We selected a threshold of 0.9, because this selects 90% of the blacklisted domains while in line with our intuition that a higher score makes the domain more suspicious. However, we note that selecting a threshold of 0.1 selects 93% of blacklisted domains, showing that the threshold value selection has minimal effect on our method.

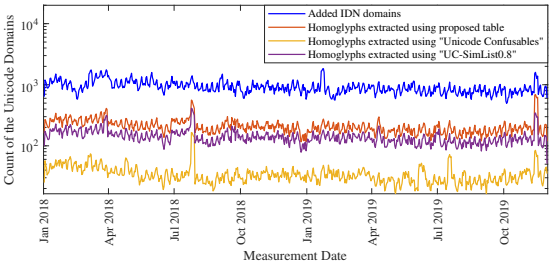

Figure 4. Extracted homoglyphs for added IDNs

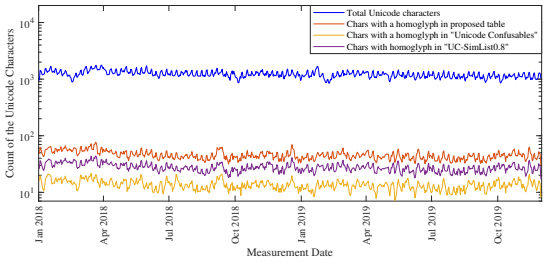

Figure 5. Characters with a homoglyph for added IDNs

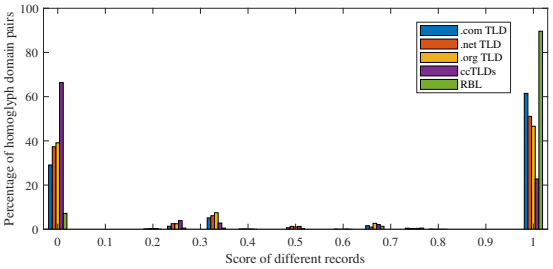

Figure 6. Normalized scores for extracted and blacklisted domains

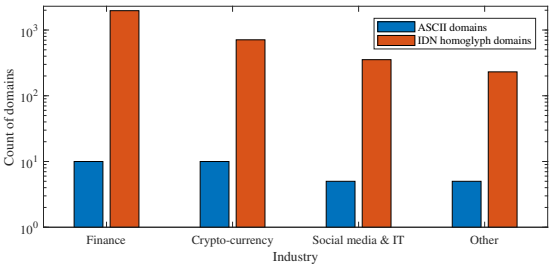

Figure 7. Top targeted domains on ".com" TLD

We have compared detected domains against domains from RBL between 2018-01-01 and 2019-12-15. During our detection period we have marked 53323 domains as suspicious for exceeding the normalized suspiciousness score threshold of 0.9. Of these domains 337 (0.63%) have appeared on one of the blacklists in RBL during the observation period. While there are many domains which are not blacklisted, we feel strongly that these remain suspicious, as these Unicode and ASCII domain pairs closely resemble each other visually while their DNS records are different. From the RBL listings 320 originate from '.com', 15 from '.net', one from '.org' and a single domain from the ccTLDs. However, from all of our detections 18% of suspicious domains originate from '.net', '.org' and ccTLDs. This suggests that the RBL is targeted more towards '.com' TLD and doesn't cover other TLDs proportionally. In Section 6.5 we discuss how much earlier some of our detections are when compared to these blacklists. 52986 (99.37%) of suspicious IDNs have not appeared on a blacklist during the observation period, considering a threshold of 0.9, which suggests that majority of detected suspicious domains are not actively used for a malicious intent yet. This is inline with the EURid 2019 report that 81% of generic TLD IDNs are parking pages.

6.4. Top Targeted domains

In order to get a better insight on the domains which are highly targeted by suspicious IDN homoglyphs, we have counted the number of homoglyph IDNs targeting a specific ASCII domain. From the 30 most targeted domains (domains with the highest count of corresponding homoglyph IDNs), we have determined the industry category. Fig. 7 shows the count of domains summed per industry category of this top 30 in '.com' TLD. The domain name targeted the most in '.com', belongs to a financial service with 1250 added IDN homoglyphs targeting this ASCII domain throughout 2018 and 2019. Considering these 30 highly targeted domains in '.com', we observe that 10 domains (33.3%) are related to financial service providers with 1969 corresponding IDN homoglyph domains. The crypto-currency platforms are in the second rank targeting 10 ASCII domains (33.3%) with overall 710 IDN homoglyph domains. The social media and IT service providers achieve the third rank consisting of 5 ASCII domains (16.7%) with overall 354 IDN homoglyph domains. A similar behaviour is seen in '.net' TLD with comparably lower number of homoglyph domains per ASCII domain. This characteristic is not seen in '.org' TLD and ccTLDs where each ASCII domain has a handful of corresponding IDN homoglyphs.

6.5. Time advantage

In this section the potential achievable time advantage using our proposed detection approach against existing blacklists is investigated. We calculate the time advantage as the window (in days) between detection of suspicious domains by our method and the time by which these domains appear on the blacklists. Out of the 337 detected domains, 78 domains (23.2%) were detected on the same day as their registration, 179 domains (53.1%) were detected at least a day after their registration and at most in a month, 79 domains (23.4%) were detected with a difference between a month and a year, and a single domain (0.3%) was detected after a year. On average 21 days of early detection is achievable using the proposed method considering the domains that are already blacklisted.

7. Ethical Considerations

Although we believe that the new registerd IDNs detected by the proposed method in this paper are highly suspicious due to having a different source than the existing ASCII domain, it is not ethically acceptable to

publish the list of these domains, specially since we do not study the intent behind these domains. This paper is meant to provide awareness among both authorities and end-users with a simple method to help stop malicious usage of IDNs.

8. Discussion

The measurement data OpenINTEL provides for IDNs clearly shows the problem of IDN homoglyph phishing attacks. However, the used dataset shows only the tip of the iceberg. As of December 2018 there were 85 ccTLDs supporting IDNs [21]. In this work we have investigated 7 out of the 85 ccTLDs which support IDNs. This makes for a good start since we show the proposed methodology works in detecting homoglyph IDNs, but it does not show the complete picture of homoglyph IDNs. Extending the number of measured ccTLDs which support IDNs would increase the grasp we have of the problem. Additionally, our proposed Unicode confusion table only covers Unicode to single ASCII mappings that is relatively computationally inexpensive. With multiple homoglyphs for a single character, either Unicode or ASCII, the Unicode confusable table may be further improved and detect more suspicious homoglyph domains. We noticed, during this work, that the blacklists we have used do not focus on IDNs. Creating our own blacklist, which specifically focuses on malicious IDNs, may be beneficial for the security community at large. Since we have good indication that the domains we detect, are at least suspicious.

9. Conclusions

In this paper we investigate how suspicious domains using Unicode homoglyph characters can be detected using active DNS measurements. Combining the unique OpenINTEL dataset with our improved Unicode Confusables table we are able to detect 53323 domains from '.com', '.net' and '.org' TLDs and seven ccTLDs which exceed our 'suspiciousness' score threshold. Additionally, we have shown that our method can potentially detect suspicious domains on average 21 days earlier when compared to blacklists. Furthermore, we show that the suspicious IDNs frequently target domains in the finance and crypto-currency industries, followed by domains in social media and IT sectors.

Acknowledgments

We would like to thank the OpenINTEL project for providing us with valuable DNS data. This work has been funded by SIDN fonds, an independent fund on the initiative of SIDN, the registrar for '.nl' domains. This work has partially been funded by the EU H2020 project CONCORDIA (#830927).

References

[1] L. M. Masinter, M. J. Dürst, and M. Suignard, "Internationalized Resource Identifiers (IRI)," Internet Engineering Task Force, Internet-Draft draft-masinter-url-i18n-08, Nov. 2001, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-masinter-url-i18n-08

[2] E. Gabrilovich and A. Gontmakher, "The Homograph Attack," *Communications of the ACM*, vol. 45, no. 2, p. 128, 2002.

[3] A. Costello, "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)," Internet Requests for Comments, RFC Editor, RFC 3492, March 2003.

[4] A. Y. Fu, X. Deng, L. Wenyin, and G. Little, "The Methodology and an Application to Fight against Unicode Attacks," in *Proceedings of the second symposium on Usable privacy and security*. ACM, 2006, pp. 91–101.

[5] N. Roshanbin and J. Miller, "Finding Homoglyphs - A Step towards Detecting Unicode-Based Visual Spoofing Attacks," in *International Conference on Web Information Systems Engineering*. Springer, 2011, pp. 1–14.

[6] H. Suzuki, D. Chiba, Y. Yoneya, T. Mori, and S. Goto, "ShamFinder: An Automated Framework for Detecting IDN Homographs," in *Proceedings of the 19th ACM Internet Measurement Conference (IMC 2019)*, 2019.

[7] C. Liu and S. Stamm, "Fighting Unicode-Obfuscated Spam," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*. ACM, 2007, pp. 45–59.

[8] F. Alvi, M. Stevenson, and P. Clough, "Plagiarism Detection in Texts Obfuscated with Homoglyphs," in *European Conference on Information Retrieval*. Springer, 2017, pp. 669–675.

[9] A. Y. Fu, X. Deng, and L. Wenyin, "REGAP: A Tool for Unicode-Based Web Identity Fraud Detection," *Journal of Digital Forensic Practice*, vol. 1, no. 2, pp. 83–97, 2006.

[10] V. Krammer, "Phishing Defense against IDN Address Spoofing Attacks," in *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*. ACM, 2006, p. 32.

[11] J. Al Helou and S. Tilley, "Multilingual Web Sites: Internationalized Domain Name Homograph Attacks," in *2010 12th IEEE International Symposium on Web Systems Evolution (WSE)*. IEEE, 2010, pp. 89–92.

[12] H. Shirazi, B. Bezawada, and I. Ray, "Kn0w Thy Doma1n Name: Unbiased Phishing Detection Using Domain Name Based Features," in *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies*. ACM, 2018, pp. 69–75.

[13] T. Holgers, D. E. Watson, and S. D. Gribble, "Cutting through the Confusion: A Measurement Study of Homograph Attacks." in *USENIX Annual Technical Conference, General Track*, 2006, pp. 261–266.

[14] B. Qiu, N. Fang, and L. Wenyin, "Detect Visual Spoofing in Unicode-based text," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 1949–1952.

[15] D. Chiba, A. A. Hasegawa, T. Koide, Y. Sawabe, S. Goto, and M. Akiyama, "DomainScouter: Understanding the Risks of Deceptive IDNs," 2019, pp. 413–426.

[16] B. Liu, C. Lu, Z. Li, Y. Liu, H.-X. Duan, S. Hao, and Z. Zhang, "A Reexamination of Internationalized Domain Names: The Good, the Bad and the Ugly." in *DSN*, 2018, pp. 654–665.

[17] Y. Sawabe, D. Chiba, M. Akiyama, and S. Goto, "Detection Method of Homograph Internationalized Domain Names with OCR," *Journal of Information Processing*, vol. 27, pp. 536–544, 2019.

[18] Y. Elsayed and A. Shosha, "Large Scale Detection of IDN Domain Name Masquerading," in *2018 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2018, pp. 1–11.

[19] F. Quinkert, T. Lauinger, W. Robertson, E. Kirda, and T. Holz, "It's Not what It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains," in *2019 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2019, pp. 259–267.

[20] "The Majestic Million." [Online]. Available: https://majestic.com/reports/majestic-million

[21] EURid, "IDN World Report." [Online]. Available: https://idnworldreport.eu/

[22] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," *arXiv preprint arXiv:1806.01156*, 2018.