Understanding and Characterizing the Adoption of Internationalized Domain Names in Practice

Yunyi Zhang, Chengxi Xu, Fan Shi, Miao Hu, Min Zhang, Yuwei Li, and Zhijie Xie

Abstract—**Internationalized Domain Names (IDNs) allow users to access the internet using domain names in their native languages. This technology provides significant convenience for non-English speaking users. However, despite the widespread acceptance and use of IDNs, the risks associated with using IDNs remain unclear in practice, such as the IDN homograph problem. To address this issue, we conduct a systematic analysis of the IDN homograph problem and explore the adoption characteristics of IDNs in practice. Specifically, we design and implement an effective IDN analysis framework, named as IDNMon. We perform a large-scale measurement study covering 863 top-level domain zone files and historical top lists based on IDNMon. Our findings indicate that the IDN registration and usage in Europe exceeds that in East Asia. Our results confirm that the IDN homograph problem is universal (12.32% of 2,623,161 IDNs face this problem), which raises serious challenges when designing protection strategies for browsers. Our work provides new insights into the adoption of IDNs in practice, contributes to a better understanding, and promotes the development of IDNs.**

Index Terms—**Domain name system, IDN ecosystem, IDN homograph, internationalized domain name, measurement.**

I. Introduction

CURRENTLY, an increasing number of users from different regions access the Internet. However, for a long time, users could access the Internet using domain names composed of only English characters, hyphens, and digits [1]. Internationalized domain names (IDNs) were introduced and implemented in 2003 to satisfy the needs of users belonging to regions with different languages [2]. IDNs enable global users to access the Internet by applying domain names that contain characters belonging to their native languages. After 20 years of development, most registrars now support IDN registration, and most web browsers now accommodate IDNs.

However, in practice, potential security issues can hinder the development of IDNs. For example, an attacker can use IDNs to impersonate other domain names to deceive users; this attack is called a homograph attack [3], and it relies on the fact that there are different characters that look similar in different languages. For example, the Latin character “a” (U+0041) looks similar to the Cyrillic character “a” (U+0430). Another study reported that a domain name that was highly similar to *apple.com* was used for phishing attacks in 2017 [4]. Most browsers implement rules to detect homograph IDNs, and two IDN implementation standards, IDNA2003 [2] and IDNA2008 [5], have also been established. IDNA2008 is a revised version of IDNA2003 that addresses some of the design flaws in IDNA2003; however, it also suffers from many incompatibility issues present in IDNA2003. Hu et al. [6] showed that the rules of major browsers have weaknesses and can be bypassed. Despite the efforts of the Internet community to enhance the security of IDNs, the risk of IDN adoption is still unclear. In addition, ICANN is preparing to open the second round of new gTLD registration. It is important to understand the development of IDN over the past 20 years.

To address this problem, this paper aims to answer the following research questions: (1) Has IDN been adopted by more regions? (2) What factors influence the development of IDNs, and are top companies taking advantage of IDNs? (3) What factors affect the IDN homograph problem? (4) Does the browser and registrar have a common mitigation strategy for IDN incompatibility?

Identifying the language of an IDN is the first challenge in understanding the issue with the adoption of IDNs. Liu et al. [7] detected the IDN language using the language detection model LangID [8]. However, there are considerable differences between natural languages and IDNs, such as the use of short labels and language confusion, which result in poor detection accuracy. Finding IDNs with a homograph problem is another challenge in terms of exploring the usage of IDNs. Most IDN homograph detection studies [7], [9], [10] prepare a list of popular domain names as reference domain names and calculate visual similarity between the collected IDNs and reference domain names. However, the collected reference domain names directly limit the scope of the analysis and cannot provide a holistic analysis.

Our Study: We designed and implemented a generic framework referred to as IDNMon, which uses a novel homograph IDN discovery and language detection approach to overcome these challenges and measure the adoption state of IDNs worldwide. IDNMon can find all potential IDNs with the homograph problem; this overcomes the dependence on reference domain names. It can also analyze the IDN homograph problem in the entire domain name space. Through investigating the nature and purpose of domain labels, IDNMon proposes the language

Manuscript received 16 January 2024; revised 5 April 2024; accepted 6 April 2024. Date of publication 10 April 2024; date of current version 16 January 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100500 and in part by the National Social Science Fund of China under Grant 2023-SKJJ-C-063. *(Corresponding author: Min Zhang.)*

The authors are with the College of Electronic Engineering, Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, National University of Defense Technology, Hefei, Anhui 230037, China (e-mail: zhangyyzyy@nudt.edu.cn; xuchengxi@nudt.edu.cn; shifan17@nudt.edu.cn; jackyhumiao@163.com; zhangmindy@nudt.edu.cn; liyuwei@nudt.edu.cn; xzj9510@nudt.edu.cn).

Digital Object Identifier 10.1109/TDSC.2024.3386905

definition of IDNs and implements a new language detection approach based on hierarchical label and script correction. Moreover, it solves the problems of short domain name labels and language confusion.

Findings: We used the IDMon framework to understand the IDN ecosystem in terms of IDN popularity, homograph problem, and incompatibility. We obtained 4,705,632 IDNs (including subdomains) under 863 top-level domains (TLDs) by collecting the public zone files using the centralized zone data service (CZDS) [11] and public projects. Then, we collected auxiliary data including WHOIS, DNS records, and websites to study the characteristics of IDNs. We collected five top list history lists, such as Alexa and Umbrella, to display the actual use of IDNs by Internet users. As a result, we discovered that IDNs are gaining popularity in Europe (3/5 of the top 10 regions are in Europe), which covers 41.23% of all IDNs (Section VI-A). Russia is dominant not only in terms of the number of IDNs but also in the rankings. Moreover, we analyzed the domain names of 50 top companies from different industries, including finance, manufacturing industry, and information technology. The results illustrate that the number of registrable homograph IDNs is considerable in these companies; however, it has been overlooked so far (Section VI-C). We discovered that 323,135 IDNs (12.32% of SLDs) suffer from the homograph problem and have not been analyzed before (Section VI-C). Some financial domain names are not highly ranked; however, they have a large number of homograph domain names, such as *aresmgmt.com*. Finally, our findings revealed that registrars and browsers still lack common mitigation measures for addressing deviation characters (Section VI-B). We will open-source our code and datasets at the GitHub[1] to facilitate future research.

Actionable Suggestion: The adoption of IDNs has grown across different regions due to the collective efforts of many parties. However, IDNs are confronted with severe security issues. The new insights we uncover provide practical recommendations for browsers and registrars. Moreover, as ICANN is promoting the second round of top-level domain registration [12], our work also aims to provide support for the registration of IDN top-level domains.

The main contributions of this work are as follows:

1) We perform a large-scale measurement study on the adoption of IDNs in practice, presenting a thorough analysis of the IDNS ecosystem. We observe the regional rise of IDNs, with it becoming gradually more popular in Europe than in East Asia. Moreover, our work offers new insights into the evolution of IDNs to help registries, registrars, and registrants better understand and use IDNs.
2) We design and implement the new analysis tool IDNMon, which is an effective IDN analysis framework that supports homograph IDN discovery without the need for reference domains and accurate language detection, offering valuable resources for future IDN research and facilitating the advancement of IDN studies.
3) We analyze the adoption of IDNs by 50 top companies in different industries for the first time and demonstrate that the value of IDNs is still not widely recognized, which provides attackers with an opportunity to abuse homograph IDNs.

Roadmap: The rest of the paper is organized as follows. Section II introduces the background of our study. Section IV discusses the design and implementation of IDNMon. Section III presents the collected data. Section V provides the evaluation between IDNMon and other approaches. Section VI demonstrates our measurement findings. Section VII discusses our actionable suggestion and some issues requiring on-going attention. Section VIII describes the related work, and Section IX concludes our work.

II. Background

This section introduces an overview of the domain name and IDN and provides a brief introduction to IDN security risks.

A. Domain Name and IDN

Domain Name: A domain name consists of hierarchical labels, where each label is related to a zone. It can be divided into the top-level domain (TLD), second-level domain (SLD), and subdomain; TLD includes generic TLD (gTLD), country-code TLD (ccTLD), and sponsored TLD. The Internet Corporation for Assigned Names and Numbers (ICANN) allowed the registration of new TLDs since 2003 because of the increasing demand for domain names; after 2003, the number of TLDs gradually increased to 1,580. In 2009, after the introduction of IDN, ICANN also introduced an IDN TLD called iTLD. The introduction of new TLDs disrupts the domain name order and introduces new threats [13].

IDN: Traditional domain names only support English characters, digits, and hyphens, which does not satisfy the needs of linguistically diverse populations. Therefore, IDNs were successfully standardized and implemented in 2003 and updated in 2010 to satisfy these new requirements and challenges. IDNs were converted to ASCII-compatible encoding strings by Punycode [14] to retain backward compatibility in many network protocols; for example, domain name 例子.中国 was converted to *xn--fsqu00a.xn--fiqs8s*. The Internet community has invested significant efforts to promote the wide adoption of IDNs.

B. IDN Security Risk

The emergence of IDNs has enhanced user experience in terms of accessing the Internet from different regions; however, it has also introduced new risks such as IDN homographs. The IDN homograph problem has existed for a long time and is a type of cybersquatting, where malicious users register and use an Internet domain name with a malicious intent to profit from the goodwill of a trademark belonging to someone else. In general, attackers attempt to deceive users using a confusing domain name. IDNs have become a natural attack vector for homograph problems because they contain homoglyphs or characters that visually resemble other characters in different languages. In 2002, Gabrilovich and Gontmakher [15] reported homograph domain names with non-Latin letters. Further, in

[1] https://github.com/sy-yunyi/IDNMon

TABLE I
SUMMARY OF TOP LIST, IDN, AND BLACKLIST

Top list				
	#IDN	#IDN (unique)	Start date	End date
Alexa	6,593,673	276,130	2009/1/29	2021/8/9
Openpagerank	381,231	1,660	2019/9/9	2021/8/9
Majestic	3,763,464	16,865	2017/6/6	2021/8/9
Umbrella	588,674	14,896	2016/12/15	2021/8/9
Quantcast	27,719	58	2018/5/22	2020/4/2
IDN				
	#IDN	#gTLD	#iTLD	
Zone file	1,459,145	396	85	
FDNS	3,726,075	575	147	
domains	997,508	566	147	
Total (unique)	4,705,985	712	151	
Blacklist				
	#Domain	#IDN	Rate	
Phishing.Database	325,073	228	0.07%	

2017, another study reported that it was possible to deploy a domain name highly similar to *apple.com* to perform phishing attacks [4].

There are no effective countermeasures against the threats of IDN homograph problems; however, their necessity has been presented earlier. Hu et al. [6] systematically tested browser-level defenses against homograph IDNs and showed that all tested browsers have weaknesses in their policies and implementations. This could be attributed to IDNs not being popularized as quickly as expected, which resulted in these problems being overlooked. However, IDNs are widely adopted now, and thus, IDN homograph problems have become an urgent issue that needs to be understood and resolved.

III. Data Collection

In this section, we describe the data collection of IDNMon in detail. Table I summarizes our data.

TLD Zone File: TLD zone files record all registered domain names, using which we identified IDNs by searching the substring *xn--*. We downloaded 1,138 zone file snapshots from CZDS, which include 1,052 gTLDs and 86 iTLDs. In the end, we discovered 1,459,145 IDNs under 481 TLDs. We observed that there is no increase in the number of IDNs extracted from the zone files compared with those reported in the previous work [7]. This can be attributed to the slowing down of the growth trend of IDNs in recent years.

Top List: The ranking of domain names in the top list represents their popularity. Although some studies have shown that almost all top lists have vulnerabilities that can be exploited by attackers [16], a few abnormal domain names will not cause the measurement result to differ significantly. We collected five top lists, including Alexa (2009–2021), Openpagerank (2019–2021), Majestic (2017–2021), Umbrella (2016–2021), and Quantcast (2018–2020) [17]. These top lists are updated daily, which provides us with the changing trend of IDN rankings.

Public Project: Prior studies were limited by the perspectives of TLD zone files and focused only on IDN embedding Unicode at the second-level and top-level, while ignoring the subdomains. We collected more diversified domain names from two public projects, *FDNS* [18] and *domains* [19]. These projects apply an active scanning approach to construct the freely available list of domain names (not only second-level but also subdomains), which provides a new direction to analyze the adoption of IDNs. We collected 3,726,075 IDNs from *FDNS* and 997,508 from *domains*.

Blacklist: We collected the blacklist from a public project *Phishing.Database* [20] to explore which domain names can be easily targeted by attackers for homograph domain name attacks. We find that there exist a few IDNs in the blacklist, which account for only 0.07%; the detail is presented in Table I.

WHOIS: We designed and implemented a crawler to crawl the WHOIS records published by registrars to obtain the registration information of IDNs; we collected WHOIS information of 1,269,545 (47.92%) IDNs. There are two major reasons for failing to obtain the WHOIS of the remaining IDNs: (1) some registrars block the request, particularly for iTLDs, and (2) many IDNs set up a privacy protection strategy.

DNS Records: The domain name response status is an important criterion for assessing whether a domain name is registered. We considered the domain name as an unregistered domain name when we received a *NXDOMAIN* response. Thus, we applied ZDNS [21] to lookup `A` records and `NS` records for the collected IDNs.

Website: Website data show us the purpose of an IDN; it can be a normal site or a parked page. We leveraged our crawler to collect site data such as title and page content to assist us in determining the owner consistency of the IDN with its prototype domain name.

Advantage and Limitation of Data: Diversified data from many sources provide a new and complete analytical perspective. Our IDN list introduced subdomains and IDNs under ccTLDs. Subdomains are an important part of DNS but have not been focused on in previous studies. Further, IDNs under ccTLDs supplement the description of IDN usage in different regions. Thus, the measurement result helped us understand the adoption of IDNs in practice. We also collected top lists over 12 years. The top lists often represent the popularity of a domain name, which can illustrate the changing trend of IDN rank over time. Although we attempted to make the study as comprehensive as possible, the blacklist we collected was limited because of the limitation of data sources. However, a previous study showed that the proportion of IDNs in blacklists is very small [7], and our goal was not to detect more malicious IDNs in blacklists but to analyze what domain names are more likely to be targeted. Therefore, our measurement results do not differ significantly.

IV. IDNMon Framework

This section introduces the main challenges we face, and provides a high-level overview of the IDNMon framework and briefly describes the data collector. Further, this section presents the core idea of the homograph IDN discovery approach, which plays a key role in the IDNMon framework. Finally, the hierarchical language detection approach is introduced.

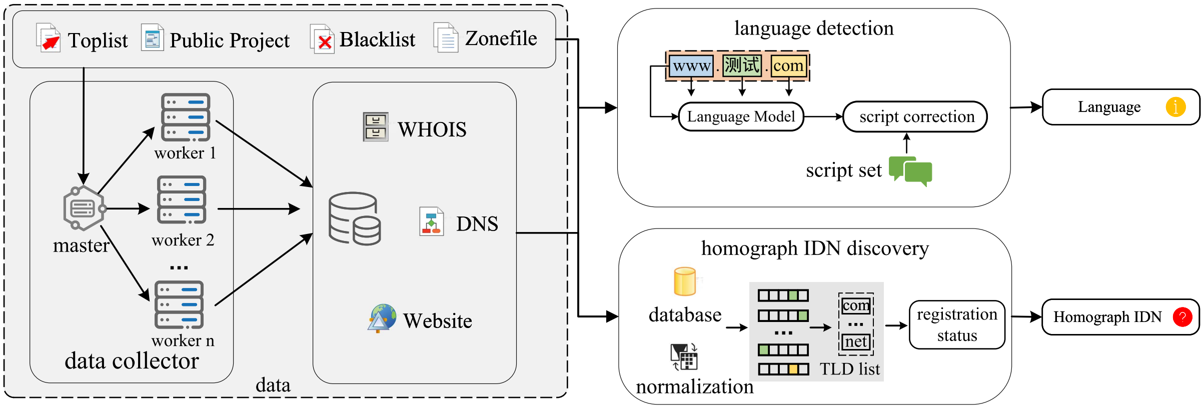


Fig. 1. Overview of the IDNMon framework.

A. Challenges

Our primary research goal is to measure the IDN ecosystem systematically. On the technical level, there are two challenges to overcome.

C1 *Proactively and highly scalable evaluation of IDN homograph problems:* Prior studies of the homograph IDN problem have involved the image similarity of domain names, which is only applicable to the detection of specific domain names. Moreover, these methods rely on comparisons with target domain names, thus lacking scalability and involving significant computational costs, making them challenging to apply in large-scale assessments.

C2 *Accurately detecting the language of IDN:* It is a common practice to leverage the published language detection model directly, but the issues of short labels and language confusion hamper the detection accuracy.

To overcome *C1*, IDNMon introduces active homograph IDN generation. Relying on domain name similarity comparisons to identify IDN homograph problems passively only allows for assessing specific domain names, significantly limiting the scope of research. This approach is incapable of conducting a comprehensive assessment of the domain name space, thus failing to achieve the goals of this paper. Starting from the root of the IDN homograph problems, we conducted an in-depth analysis of the existing homoglyph character sets, highlighting their advantages (comprehensive coverage) and drawbacks (severe fragmentation, difficult to use directly). Then, we propose a method for integrating homoglyph character sets, creating a more effective and user-friendly set. Building on this, we design an algorithm for generating homograph IDN domains, proactively creating candidate lists, thus freeing itself from the constraints of target domain lists and enabling a comprehensive analysis of the domain name space.

To overcome *C2*, we introduce the relationship between language and script and effectively correct erroneous language detection results. By thoroughly analyzing the character set differences between various languages, we construct a mapping between languages and script sets, which aids in identifying incorrect language results from language detection models.

B. Overview

In this paper, we consider the prototype character of a Unicode character to be a Latin character with the same appearance; the prototype character for the Latin character is the Latin character itself. The homograph database stores this relationship between Unicode characters and their prototype characters. Further, the prototype domain name of an IDN is the domain name that contains the prototype characters of all its characters. Fig. 1 presents a high-level overview of the IDNMon framework.

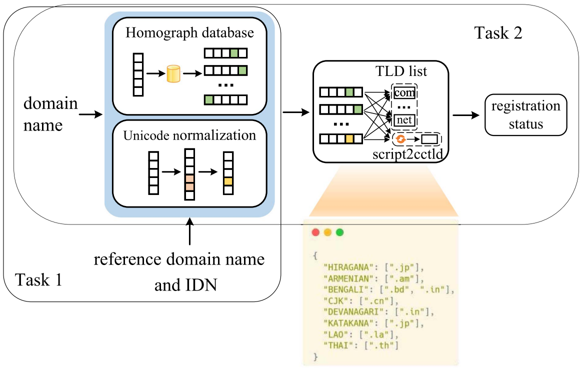
Data: The IDNMon collects IDNs from multiple data sources, which include TLD zone files, public projects of existing domain names, domain name top lists, and blacklists of malicious domain names. Further, the *data collector* module of IDNMon crawls auxiliary data including the WHOIS, DNS records, and website data. The details of the dataset are provided in Section III.

Homograph IDN Discovery (Section IV-D): IDNMon generates the prototype domain name for each IDN by applying the homograph database and Unicode normalization to find all IDNs with the homograph problem. Next, IDNMon introduces the common TLD list to extend the scope of the generated domain names. Finally, IDNMon checks their registration status to filter the invalid prototype domain names (i.e., unregistered domain names).

Language Detection (Section IV-E): IDNMon splits domain names into different labels using a "." and conducts detection on each label separately. IDNMon introduces the relationship between languages and scripts to rectify the wrong results obtained from natural language processing models for enhancing the accuracy of IDN language detection.

C. Data Collector

We implement a distributed crawler module to collect dynamic data efficiently *data collector*. To this end, we customize the uniform data format for different source data and then implement special task crawlers, such as *Webcrawler*, *DNScrawler*, and *Whoiscrawler*. The master node schedules and distributes data collection tasks. Finally, worker nodes asynchronously perform tasks to collect the target data and save it to the database.



```
{
  "HIRAGANA": [".jp"],
  "ARMENIAN": [".am"],
  "BENGALI": [".bd", ".in"],
  "CJK": [".cn"],
  "DEVANAGARI": [".in"],
  "KATAKANA": [".jp"],
  "LAO": [".la"],
  "THAI": [".th"]
}
```

Fig. 2. Process of IDN homograph discovery.

D. Homograph IDN Discovery

IDNs with the homograph problem (hereafter, homograph IDNs) are those with valid prototype domain names. Unlike previous studies that focused only on popular domain names, IDNMon can discover all homograph IDNs in the entire domain name space.

The process of IDNMon's homograph IDN discovery approach is presented in Fig. 2. The homograph IDN discovery can be summarized into two tasks according to the difference in input: Task 1 detects the homograph IDNs for given reference domain names and an IDN list, and Task 2 attempts to find all potential IDNs with the homograph problem. Prior studies focused on Task 1; the common approach in these studies was to compare the visual similarity between the reference domain names and collected IDNs. However, these studies were limited by reference domain names, and therefore, Task 1 cannot analyze domains besides the reference domain names. Task 2 can easily break this barrier, but there exist two challenges: (1) how to generate prototype domain names of the input IDNs, and (2) how to check the validity of the generated prototype domain names.

Task 2 Workflow: Fig. 2 exhibits the process of IDN homograph discovery. We divided Task 2 into three parts. The first step is to generate prototype domain names (the blue part) by using a homograph database and Unicode normalization. Next, to enhance the coverage of prototype domain names, we introduce a TLD list to enrich the variety of domain TLDs. Finally, we use the resolution status of the domain names to determine the validity of the prototype domain names.

Generation of Prototype Domain Names: IDNMon implements a generation method based on the homograph database and Unicode normalization to generate prototype domain names of input IDNs. The Unicode Technical Standard (UTS) [22] provided a mapping file from Unicode characters to their prototype characters to resolve the issue of confusable strings. ShamFinder [23] proposed an approach for the automatic construction of a homograph database. The mapping of UTS provides a comprehensive guide for confusable characters. However, it is not sufficiently friendly for homograph problem detection. As indicated in Fig. 3, h and ħ are divided into two groups (green and orange). ShamFinder can generate the homograph databases automatically, but it is difficult to evaluate its accuracy.

To apply the homograph database to the generation of prototype domain names, IDNMon restructures the homograph database by combining the advantages of UTS and ShamFinder. The principle of reorganization suggests aggregating characters in the same family. We constructed associations centered around ASCII characters based on UTS and ShamFinder, and then clustered them according to these relationships. For example, as shown in Fig. 3, the character ħ is a combination character of h and –, and thus, ħ and h are associated with each other and belong to one family, group 1 and group 2. Fig. 3 right (green, group 1^*) shows an example of restructuring the homograph database. The left shows two similar character sets (green and orange, each representing a set), while the right depicts the combined similarity character set in IDNMon (green). Finally, the homograph database reconstructed by IDNMon contains 7,249 homograph characters, which is a 40-fold increase compared to ShamFinder and a 32% reduction relative to UTS, eliminating some ineffective homoglyph characters.

On the other hand, the homograph database misses some cases of character equivalence, such as the combining sequence and the ordering of combining marks [24]. Therefore, IDNMon applies Unicode normalization, which is used to determine whether any two Unicode strings are equivalent to each other, to perfect these cases. Fig. 4(a) and (b) display the conversion processes of applying the homograph database and Unicode normalization, respectively. In Fig. 4(a), we replace the character in the original IDN with the character a to obtain its prototype domain name. In Fig. 4(b), through Unicode normalization, we discover that the character $\bar{a}$ is composed of a and -, hence we remove the additional composite characters and retain the ASCII character a.

Conversion From non-IDNs to IDNs: Further, IDNMon supports the generation of corresponding IDNs for specified non-IDNs (i.e., domain names that only consist of English characters, digits, and hyphens) to discover potential registrable IDNs. For a non-IDN, we mainly focus on the SLD (i.e., apple in apple.com). By using a similarity character database, we identify characters that could be confused with each character in the non-IDN and replace the original characters with them. This process ultimately generates homograph IDNs for the target non-IDN. As shown in Fig. 4(c), we can replace a with a and a, and use e to replace e. Moreover, we need to overcome the domain name explosion problem to accomplish this task. For the domain name *example.com*, we can obtain 78,125 potential IDNs if each character responds to 5 confusable characters. Two factors affect this issue: the number of confusing characters per character and the number of Unicode characters contained in the domain name. The former is fixed, and therefore, the latter needs to be controlled. We need to choose the number of Unicode characters in the domain name carefully because the size of this problem grows exponentially. An IDN with fewer Unicode characters can be better masqueraded as another domain name. Moreover, in *Finding 3*, we found that the number of Unicode characters in IDNs shows a long-tailed distribution and nearly 50% of IDNs contain only two Unicode characters. Thus, in

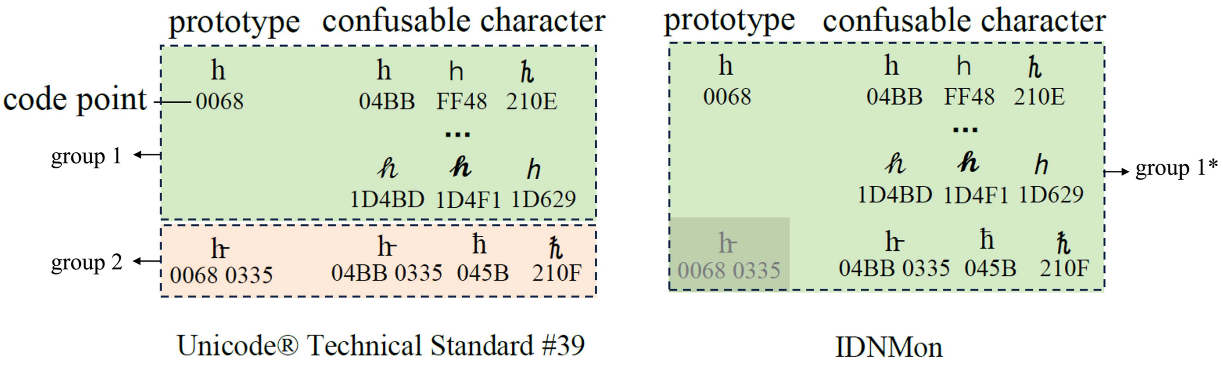
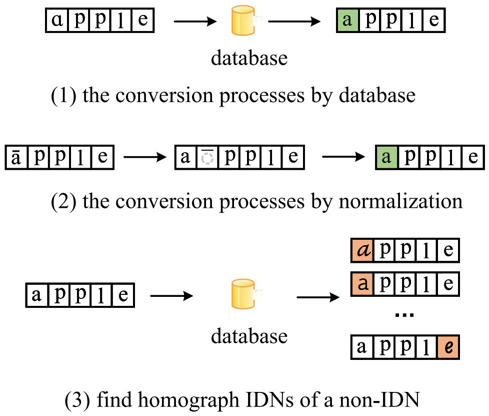


Fig. 3. Examples of mapping in the Unicode Technical Standard and IDNMon. The left shows two similar character sets (green and orange, each representing a group, i.e., group 1 and group 2), while the right depicts the combined similarity character group in IDNMon (group 1*).



(1) the conversion processes by database

(2) the conversion processes by normalization

(3) find homograph IDNs of a non-IDN

Fig. 4. Conversion processes between IDNs and non-IDNs.

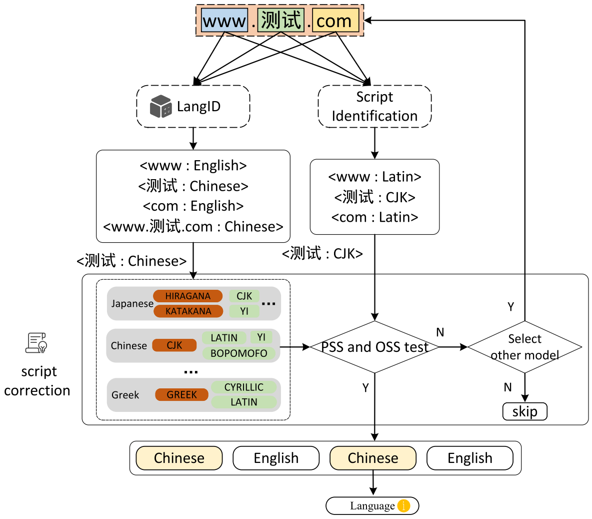


Fig. 5. IDN language detection.

combination with our analysis of *Finding 3*, we fix the number of Unicode characters contained in the domain name to less than 3. Although this can result in missing some IDNs, it does not affect our measurement results; instead, it reduces the time complexity of the method.

Validation of the Generated Prototype Domain Names' Validity: IDNMon checks the validity of prototype domain names using the domain name registration status. IDNMon introduces additional TLDs to augment the generated prototype domain names to discover valid prototype domain names efficiently. The additional TLDs consist of the five most popular gTLDs and the ccTLDs corresponding to IDN scripts. The number of domain names under gTLDs conforms to the Zipf distribution, and thus, we select the five most popular gTLDs, including *.com*, *.net*, *.org*, *.info*, and *.top*. The prototype domain name of an IDN is likely to register under its ccTLDs. For example, the script set of domain name *example.*中国 is *{CJK}*, and the ccTLD corresponding to the scripts is *.cn*; therefore, there is a good chance the domain name *example.cn* is registered. The corresponding yellow block in Fig. 2 shows an example of the IDN script and its ccTLD.

E. Language Detection

Language distribution helps us understand the popular areas of IDNs, which is the foundational but challenging work necessary for analyzing IDNs. The common approach in prior studies was to apply language detection models to identify the likely language of each IDN, such as LangID [8] and Polyglot [25]. However, these models cannot handle short labels or domain names with multiple language scripts such as *xn--1iu.cn* (慧*.cn*). To mitigate these problems, IDNMon introduces the language definition of an IDN and proposes a new language detection approach based on hierarchical label and script correction, shown in Fig. 5.

According to [1], a domain name consists of multiple labels, i.e.,

$$< subdomain, domain, tld, root >$$

The difficulty of registration increases in order from the left to the right. If you own a *domain*, you can set up an arbitrary *subdomain*, or you can apply the *subdomain* you want from a third-party service, such as *webhost.com*. *Domain* registration is managed by different registrars, and you need to provide your personal information for registering a *domain*. Since 2003, ICANN has been allowed to register a new TLD, and the application process is rigorous and complex. Further, the language between different labels is independent. For example, you can register an IDN 测试*.com*. The language of the label *com* is

English and the language of the label 测试 is Chinese, which is the language of the IDN 测试*.com*. The language of an IDN can be determined by its independent label, which can help reduce language confusion during detection. Therefore, IDNMon proposes the language definition of an IDN in terms of the label hierarchy.

Definition 1. Language of an IDN: The language of an IDN is the language of the non-English label closest to the root. If the language cannot be identified from the individual labels, it is the language of the fully qualified domain name (FQDN).

IDNMon explores the relationship between languages and scripts to reduce the false positives of language models and constructs the primary script set (PSS) and optional script set (OSS) for each language.

Definition 2. Primary script set (PSS): PSS is the fundamental script set for a language used by all words of the language.

Definition 3. Optional script set (OSS): OSS is the extended script set for a language and appears only in certain words.

To explore the relationship between languages and scripts, we collected corpora of various languages from Wikipedia, and then extracted the script used in each word. Specifically, for each language, we collected 20 articles, from which we selected 10 K distinct words. We then extracted the scripts used in these words and tallied the frequency of each script's occurrence. Scripts that appeared more than 90% of the time were labeled as the PSS for that language, while others were designated as OSS. Moreover, to ensure the accuracy of the results, we manually assessed the automatically constructed PSS and OSS for each language. Ultimately, we obtained the PSS and OSS conforming to Unicode standards for 155 languages.

Detection Workflow: IDNMon first uses a traditional language detection model (LangID (preferred) and Polyglot (alternative)) to get the base language of each label for a given IDN, and then, it uses PSS and OSS to verify it. The language of the label is determined if the validation is successful. Otherwise, IDNMon detects it again using other models or skips it. Finally, IDNMon determines the language of the IDN according to the definition of IDN language. In the script correction part shown in Fig. 5, the red block represents PSS and the green block represents OSS, i.e., the PSS of Chinese is CJK, and the OSS is LATIN, BOPOMOFO, and YI.

As shown in Fig. 5, the process of the IDN language detection approach is as follows:

Step 1. Detection Baseline: For a given IDN, IDNMon extracts its labels as the input of the language detection model LangID and identifies the script set for each label.

Step 2. Script Correction: IDNMon applies the PSS and OSS and rectifies the wrong results obtained from Step 1. IDNMon determines that the detected language is wrong when the PSS of the detected language is not a subset of the scripts of the IDN, or when there exists a script in the script set of the IDN but not in the PSS and OSS of the detected language. IDNMon changes the language detection model in Step 1 and repeats Step 2 if other models (e.g., Polyglot) can be selected. Otherwise, IDNMon skips this label.

TABLE II
COMPARISON RESULTS OF THE FOUR WORKS DISCUSSED IN TERMS OF DATASET, CONDITION, CAPACITY, AND METHOD

		IDNMon	DomainScouter	ShamFinder	Liu et al. [7]
Dataset	#TLD	863	570	1	56
	#IDN	4,705,632	4,426,317	952,352	1,472,836
Condition	reference domain	optional	necessary	necessary	necessary
Capacity	subdomain analysis	♦	◊	◊	◊
	language detection	♦	◊	◊	♦
	homograph generation	♦	◊	◊	◊
Method	brand feature	♦	♦	▲	▲
	TLD feature	♦	♦	◊	◊
	script feature	♦	◊	◊	◊

♦ : *Fully Covered* ▲ : *Partially Covered* ◊ : *Not Covered*

Step 3. Language Decision: IDNMon chooses the language of the given IDN according to *Definition 1*.

V. EVALUATION

This section describes the results of comparing the proposed IDNMon framework with prior studies in terms of properties, language, and homograph detection. There is no previous work related to detect IDNs with the homograph problem without reference domain names, and therefore, we introduce it in Section VI. Comparative results demonstrate that IDNMon holds advantages in multiple areas, effectively enhancing the capabilities in evaluating IDNs.

A. Comparison of Properties

We compared the properties of IDNMon and those of three previous works [7], [23], [26]. DomainScouter [26] evaluated the suspiciousness of a deceptive IDN that attempts to deceive users. It extracted the visual similarity, brand, and TLD features, and then used random forest to detect the deceptive IDNs. ShamFinder [23] is an automated scheme to detect homographs IDNs, and it compares the similarity between the glyphs of corresponding characters. Liu et al. [7] utilized the visual resemblance between reference and homograph domain names to detect homograph IDNs. Table II summarizes the comparison results of the four works in terms of dataset, condition, capacity, and method.

Dataset: In terms of data volume and data source, there are clear gaps between IDNMon and the other three previous works. IDNMon contains 863 TLDs, which include 712 gTLDs and 151 iTLDs. In contrast, DomainScouter contains 570 TLDs, Liu et al. [7] explored only 56, and ShamFinder analyzed only the TLD . *com*. Further, IDNMon contains more IDNs than the three previous works. IDNMon also introduces the *subdomain* for the first time, which provides a new perspective for analyzing IDNs.

Condition: Previous studies need reference domain names as a necessary condition to discover homograph domain names; their core approach is to calculate the similarity between the reference domain names and IDNs. Further, our IDNMon proposes an approach that does not require reference domain names.

Capacity: IDNMon introduces new capabilities such as subdomain analysis and homograph generation for a non-IDN to provide a comprehensive analysis of the current adoption state of IDNs. DomainScouter and ShamFinder focused on the similarity between reference domain names and IDNs, and Liu et al. [7] analyzed the language distribution of IDNs by LangID.

TABLE III
NUMBER OF IDNS UNDER DIFFERENT LANGUAGES DETECTED BY DIFFERENT METHODS

IDNMon		Polyglot		LangID	
Language	#IDN	Language	#IDN	Language	#IDN
Chinese	1,106,898	Chinese	699,943	Chinese	1,119,159
Russian	808,356	Russian	634,263	German	627,439
German	700,825	German	622,096	Russian	569,856
Japanese	338,748	Japanese	519,401	Japanese	328,456
Korean	243,513	English	447,933	Swedish	263,987
Swedish	187,350	Korean	221,573	Korean	241,837
Danish	94,754	Danish	202,470	Finnish	136,795
Thai	93,575	Swedish	133,003	Turkish	110,525
French	82,221	unknown	109,443	Danish	105,865
Turkish	67,429	Thai	91,989	Hungarian	94,918

TABLE IV
PART OF THE RESULTS OF THE LANGUAGE RANDOM SAMPLING TEST

Examples		IDNMon	LangID
punycode	unicode		
01hotel.xn--3ds443g	01hotel. 在线	✓ Chinese	✓ Chinese
cleverbees.x--55qx5d	cleverbees. 公司	✓ Chinese	✗ Latin
guru.xn--vuq861b	guru. 信息	✓ Chinese	✗ Japanese
xn--2ckyd.com	モゼ.com	✓ Japanese	✗ Chinese
xn--2s2bpa.com	룰루.com	✓ Korean	✗ Chinese
xn--b1aeca2ch.com.ua	вперед.com.ua	✓ Russian	✗ Serbian
xn--80aedunrvgkc.online	автоуслуги.online	✓ Russian	✗ Mongolian
xn--glasknstler-xhb.com	glaskünstler.com	✓ German	✗ Turkish
xn--eckbnke-8wa.ch	eckbänke.ch	✗ German	✓ Finnish

TABLE V
SOME CASES OF HOMOGRAPH DOMAIN NAMES NEWLY DETECTED BY IDNMON

Domain name	Rank	Homograph IDN
nature.com	1041	ñature.com
		naturê.com
		nâture.com
anydesk.com	2461	anydẹsk.com
		ạnydesk.com
4shared.com	4174	4sharéd.com

Method: ShamFinder focused on visual similarities between single characters for building the homograph database and Liu et al. [7] calculated visual similarities between IDNs and reference domain names using the structural similarity index. For visual similarity, DomainScouter introduced TLD and brand features. IDNMon introduces script features to enhance the accuracy of language detection, which fixes the false positives of language detection models.

Summary: IDNMon effectively enhances the evaluation capabilities for homograph IDNs in multiple aspects. Compared to existing methods, IDNMon has advantages in aspects such as dataset, application conditions, capabilities, and methodology. IDNMon introduces new features, enhancing the recognition capabilities for IDNs. Moreover, IDNMon is free from reliance on target domain name lists, enabling systematic analysis of IDNs within the domain name space. Furthermore, IDNMon expands the analytical perspective by adding analysis of subdomains and the capability to identify homograph domain names of non-IDNs proactively.

B. Comparison of Language Detection

We compared our detection approach and the common language detection models, as shown in Table III. The model polyglot failed to detect the language of IDNs, and the results contain many uncertain outputs, such as English. IDNMon and LangID obtained similar results. However, it showed significant differences in some languages. For example, IDNMon labeled 808,356 IDNs as Russian, while LangID found only 569,856 Russian IDNs. As far as we know, no public datasets are currently available that provide accurate language mapping of IDNs, which hinders us from directly evaluating the accuracy of our method. Thus, we conducted a random sampling test on IDNs with conflicting results of IDNMon and LangID language detection to evaluate the improvement effect of IDNMon on the accuracy of language detection. We randomly selected 100 IDNs and manually verified their language. The results of the test show that the language detection accuracy of IDNMon is 3 times better than that of LangID. IDNMon can reduce the false positives of models and overcome the short label problem and language confusion. Table IV lists part of the test results. For example, for domain name .com, the output language of LangID is Chinese, which is incorrect and may be due to the domain name being too short. Following the process outlined in Section IV-E, IDNMon identifies that Chinese is an incorrect result by comparing the script set. Subsequently, another language detection model is selected, which identifies the output as Korean and passes the script set check, indicating that it is the correct result. We acknowledge that the additional checking steps we introduce will impact the overall performance of the method. However, our test results demonstrate that our method can complete the language detection of 1 million internationalized domain names in 10 minutes, which is acceptable.

C. Comparison of Homograph Detection

The homograph database restructured by the IDNMon redefines the mapping between the Unicode characters and their prototype domain name. We applied the restructured database and original mapping provided by UTS to find homograph domain names of Alexa's Top 10 K domain names. UTS detected 1,254 popular domain names with homograph IDNs, and IDNMon used the restructured database and discovered 2,290, which fully covers the result of UTS, suggesting that the homograph database reconstructed by IDNMon does not miss any information. We performed a random sampling check of newly discovered homograph domain names to verify the accuracy of IDNMon's homograph domain name discovery. Table V presents some cases of homograph domain names newly detected by IDNMon. Analyzing these newly detected homograph IDNs, we found that the homograph characters in these IDNs were not present in UTS's homograph database but were instead combinations of different characters. *This indicates that IDNMon's introduction of Unicode normalization effectively compensates for the deficiencies of traditional methods, enhancing detection capabilities.*

VI. MEASUREMENT STUDY

This section discusses the measurement of the IDN ecosystem in terms of IDN popularity, IDN incompatibility, and IDN homograph problem to gain an understanding of the adoption of IDNs.

TABLE VI
IDN and TLD Distribution for the Top 10 Languages

Language	#iTLD	#IDN under iTLDs	#IDN	Rate
Chinese	60	385,623	1,106,898	34.83%
Arabic	24	3,987	33,462	11.91%
Japanese	9	8,195	338,748	2.41%
Russian	7	743,422	808,356	91.96%
Hindi	6	886	4,166	21.26%
Persian	5	1,491	10,883	13.70%
Korean	4	10,382	243,513	4.26%
Tamil	3	119	764	15.57%
Bangla	3	827	1,829	45.21%
Ukrainian	2	9,342	14,240	65.60%

A. IDN Popularity

Finding 1

Europe has emerged as the new popular region for IDNs, with over 41.23% of IDNs registered, compared to that for East Asia.

Our measurement results helped update our understanding of the popular regions of IDNs reported in previous studies. We observed that the growth rate of IDNs in East Asian countries has decreased; the total number of IDNs has been surpassed by that in European countries, particularly Russia, whose number of IDNs has risen to second place behind that of China, as summarized in Table III. These results indicate that an increasing number of countries are recognizing the value of IDNs and that IDNs are gaining wider popularity.

To explore where IDNs are adopted well, we first analyzed the top 1 IDN every day in Alexa and found a high frequency of iTLDs in European countries, especially Russia. Then, we detected languages for IDNs in Alexa and found that the top 5 languages in the top 1 IDN are Russian, Japanese, Korean, Thai, and Chinese. The results show that Russia has not only caught up with the East Asian countries in terms of the number of IDNs but is also more active than those countries.

Finding 2

Different regions have different TLD preferences when using IDNs such as gTLDs, ccTLDs, or iTLDs.

Among the 863 TLDs measured, *. com* was the most popular TLD, with nearly 30% of IDNs registered on it. Notably, two TLDs *xn--p1ai* and *de* are the iTLD of Russia and the ccTLD of Germany, respectively. Russia prefers to use iTLDs, with 91.96% of IDNs registered under iTLD, while Germany uses IDNs directly under ccTLD. The results reveal behavioral differences in the use of IDNs among regions. We discovered that most countries prefer to register IDNs under the original ccTLD, and even in China, which has the largest number of iTLDs, only 34% of IDNs are registered under iTLDs, as indicated in Table VI.

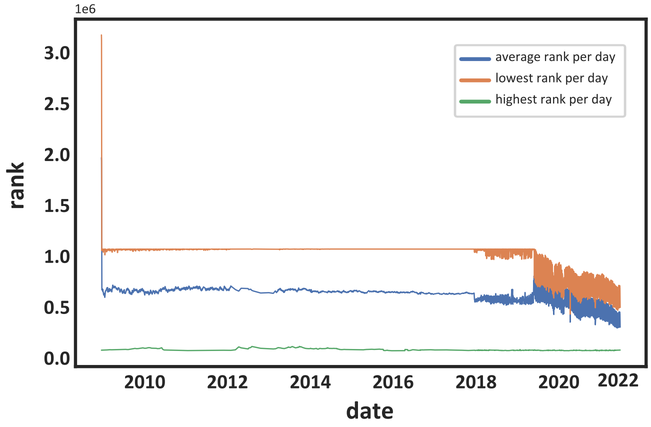


Fig. 6. Distribution of IDN rankings in Alexa.

Finding 3

Most IDNs are a mixture of Unicode and ASCII characters, especially subdomains; however, nearly 50% of IDNs contain only two Unicode characters.

To the best of our knowledge, this is the first study that analyzes subdomains and Unicode characters of IDNs. As defined in RFC1034 [1], the domain names consist of different levels of labels. The ideal scenario for IDNs would be one where each label is a Unicode character; this would mean that IDNs are widely recognized. We observed that the ideal scenario is rare (less than 20%), and most people remain accustomed to a mix of IDN and non-IDN labels. We also counted the number of Unicode characters in IDNs and found that it shows a long-tailed distribution, and nearly 50% of IDNs contain at most two Unicode characters. In addition, IDNs where Unicode characters appear in subdomains are mostly third-party hosting service providers, like seesaa.net and mixh.jp. On these platforms, users can get their own IDN through simple configuration without registering with a registrar.

Finding 4

The overall IDN ranking shows an upward trend; however, more than 75% of IDNs in the top lists appear for only 4 days.

We analyzed 5 top lists (Alexa, Umbrella, Openpagerank, Majestic, and Quantcast). We focused our analysis on Alexa because there exists a small-time span for the other lists that are not representative. We found that the number of IDNs peaked around 2019, after which the popularity decreased gradually. The overall IDN ranking showed an upward trend. We collected and analyzed the daily ranking data of IDNs from Alexa, calculating the highest, average, and lowest rankings for each day. Fig. 6 shows the ranking curves of IDNs in Alexa from 2009 to 2021, which include the average (blue), lowest (orange), and highest (green) rank per day. Except for a few IDNs with high

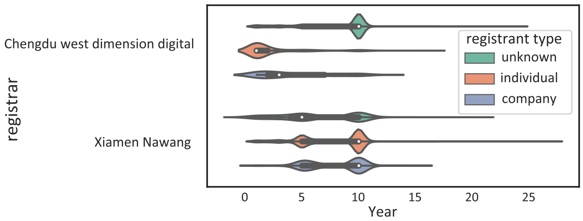


Fig. 7. IDN's lifetime for Xiamen Nawang and Chengdu west dimension digital.

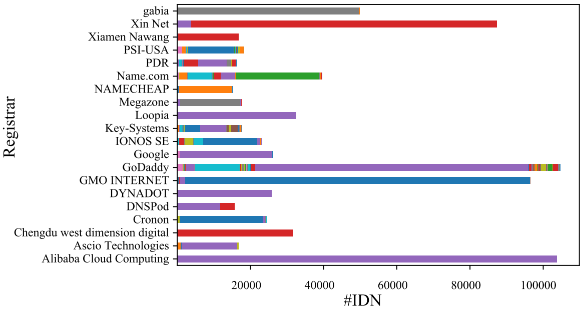


Fig. 8. Distribution of registrant country for the top 20 registrars.

rankings (green line), the rankings of most IDNs are below 500,000. However, IDN activity has gradually increased, and the overall increase has been substantial since 2018.

Further, we observed that most IDNs exist only for a short period of time in the top lists. 75% of IDNs in the top lists appear for only 4 days and only one in ten appears for more than a month.

Finding 5

The lifetime of IDNs is influenced by both the registrant and the registrar.

The lifetime of IDNs represents the time interval between the registration and the expiration of a domain name. We determined the types of registrants (individual, company, and unknown) by utilizing the *organization* and *email* in the registration information. We observed that there are two significant aggregation points in the IDNs' lifetime for *individual* IDNs: one year and ten years. One year means that the user is simply trying to register an IDN out of curiosity. When the user recognizes the value of an IDN, the user will directly register for ten years to reduce operational costs. We discovered that there are significant differences in the IDN lifetime among registrars as well. Fig. 7 displays the IDN lifetime for two registrars. The graph contains two sets of data, each originating from a different service provider. Each color represents a type of registration, including *unknown*, *individual*, and *company*. The difference lies in the IDNs registered by individuals, which are distributed around one and ten years for registrars *Chengdu west dimension digital* and *Xiamen Nawang*, respectively. This can be related to the promotion strategies of registrars. *Chengdu west dimension digital*'s marketing strategy focuses on attracting more registrants without concerning itself with the duration of their registrations, while *Xiamen Nawang* promotes longer registration periods to its registrants.

Finding 6

Approximately 60% of IDNs are serviced by the top 20 registrars; however, there are significant disparities in their user base, geographic coverage, and privacy protection policies.

Only partial domain name registration data are available because of privacy protection policies such as the European Union's General Data Protection Regulation. However, our research still yielded some intriguing results. Overall, registrars (*gabia, Xiamen Nawang, PSI-USA, Megazone,* and *IONOS SE*) deployed relaxed privacy policies, which enabled us to speculate the identity of their registrants. Our analysis revealed two distinct distributions of registrant types. For registrars (*gabia, Xiamen Nawang,* and *Megazone*), the majority of IDN registrants were individuals, while for registrars (*PSI-USA* and *IONOS SE*), most registrants were companies. The former are regional registrars, while the latter are international registrars. The observed difference in distribution suggests that regional registrars have made considerable efforts to promote IDNs to individuals, which has resulted in positive outcomes. Conversely, companies appear to prefer international registrars.

Registrars exhibit regional distribution: We collected and analyzed registrant countries and provided direct evidence of the regional orientation of the registrars. Fig. 8 shows the distribution of registrant countries for the top 20 registrars, where different colors represent different countries; purple indicates that no country information was obtained. A clear distinction can be observed in the service range of registrars, with some regional registrars providing services to users in specific regions only, while some global registrars offer services to more than dozens of countries. For example, the two Korean companies, *gabia* and *Megazone*, primarily serve Korean registrants, while the three Chinese companies, *Xiamen Nawang*, *Chengdu west dimension digital* and *Xin Net*, have a user base mainly from China. However, GoDaddy has a global user base, with registrants from all over the world. We were surprised to find that the three US companies, *PSI-USA*, *Name.com* and *NAMECHEAP*, had a significant number of registrants from Germany, Singapore and Iceland, respectively. We speculate that this may be attributed to their local market dominance. In addition, we also inferred the deployment of privacy protection policies in purple. *Alibaba Cloud Computing* hides all user information while *gabia* does not appear to have any such measures in place.

Summary: IDNs have gradually gained wider recognition among internet users, covering more extensive areas and exhibiting higher levels of activity. Compared to previous research, there has been a noticeable shift in the active regions of IDNs, indicating a gradual expansion in the adoption region of IDNs. Moreover, observing the ranking trends of IDNs, it is evident that

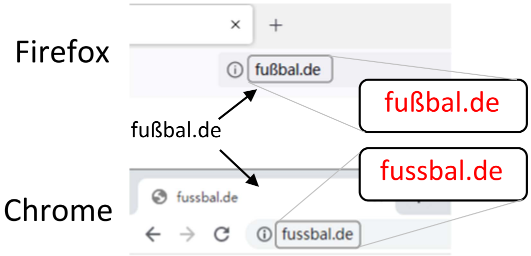


Fig. 9. Resolution strategy on deviation character ß for Google Chrome and Mozilla Firefox.

their activity levels are steadily increasing. From the perspective of registrations, IDN registrants come from a wide range of sources, with both enterprises and individuals increasingly recognizing the importance of IDNs.

B. IDN Incompatibility

Finding 7

Browsers and registrars still lack a uniform strategy for IDN incompatibility.

IDNA2008, as a revised version of IDNA2003, is intended to solve the major problems in IDNA2003. It can automatically extend the future versions of the Unicode Standard and place strict constraints on possible risk points. For example, it can limit the registration of domain names that contain specific symbols. In our analysis of IDNs with incompatibility issues in zone files and top lists, we identified 1,478 and 1,672 IDNs, respectively. The major problems include *IDNAError*, *InvalidCodepoint*, *round-trip*, and so on, which suggests that some registrars adopt an accommodative registration policy instead of IDNA2008. Further, 12.38% of IDNs with incompatibility problems involved deviation characters, particularly German ß. Le Pochat et al. [27] exposed a phishing vulnerability in iOS Mail caused by the German ß being supported as-is in IDNA2008 but being converted to *ss* in IDNA2003. Our findings revealed that registrars and browsers still lack common mitigation measures for addressing deviation characters. For registrars, GoDaddy converts the input Unicode ß to “ss”, whereas Hexonet does not. For punycode encoded strings input, they both decode them to Unicode. For browsers, Google Chrome converts the Unicode ß to “ss”, whereas Mozilla Firefox uses ß, as shown in Fig. 9. For example, when a user visits the site *fu* ß *bal.de* using Google Chrome, he or she will eventually visit *fussbal.de*. If he or she wants to visit the site *fu* ß *bal.de*, he or she needs to use Firefox. Attackers can exploit this inconsistency to hijack the access requests of users.

Summary: The standardization of IDN applications still lacks attention. There is an absence of uniform handling standards across different software (such as Firefox and Chrome) and registrars, posing a risk of misuse.

C. IDN Homograph Problem

We analyzed the distribution of homograph IDNs for popular domain names using Alexa Top 10 K domain names. We also conducted a comprehensive analysis of the potential homograph IDNs without popular domain names to illustrate the current status of the IDN homograph problem. Finally, we explored malicious IDNs in blacklists.

Finding 8

Brand value is not the sole determinant of IDN registrants’ decisions; the length of the domain name is also a significant factor that needs to be considered.

For the given popular domain names (Top 10 K in Alexa), IDNMon discovered 114,473 homograph IDNs, which cover 2,290 popular domain names. For the identified homograph IDNs of popular domains, we initially assess their usage and determine whether they are being utilized as intended (e.g., being redirected to the site of their prototype domain) or are being directed to parked sites [28], [29]. Parked sites are sites employed to showcase advertisements or sell domain information. We conducted a manual analysis of the characteristics of typical parked sites and developed their page patterns to identify parked pages. We classified popular domain names into two categories based on their own attributes, namely short domain name and brand domain name to highlight the differences in the characteristics of homograph IDNs.

Short Domain Name: Short domain names, which are domain names with very short lengths, e.g., i.ua, ae.com, are called platinum domain names in the domain name market. Their scarcity has created a certain brand identity. Original short domain name resources have been divided, and the introduction of IDN has opened up new possibilities for obtaining short domain names. Despite their low Alexa rankings, there are numerous homograph IDNs registered for short domain names. Our analysis revealed that approximately half of the short domain names are parked sites, which indicates that there is a lucrative market for the sale of these domains.

Brand Domain Name: Brand domain names are domain names of companies that are well known to the public, such as Facebook’s facebook.com. Various companies have implemented distinct strategies to combat the misuse of homograph IDNs. *Coinbase.com* has opted to park the majority of its IDNs, while a majority of companies have chosen to render their IDNs invalid. Only a few IDNs are actively used and redirected to the original website.

Despite i.ua (ranking 2,055) and a.com.cn (ranking 3,919) ranking lower than google.com (ranking 1) in Alexa, they possess a greater number of homograph IDNs, which indicates that ranking is not the sole factor influencing IDN registrations. The impact of the characteristics of the domain name outweighs the previously perceived brand effect.

Finding 9

The IDN homograph problem is common but neglected, with 323,135 IDNs facing the problem.

Our findings confirm that IDNs are subject to a serious homograph issue; this has not been previously highlighted in prior studies. To find all IDNs with homograph problems, we first generated 87,367,927 prototype domain names for 2,623,161 IDNs using IDNMon. We ignored the subdomain. Then, we filtered unregistered domain names and identified 323,135 IDNs with homograph problems, accounting for 12.32% of the total IDNs. Moreover, 95.66% of these IDNs include less than 3 Unicode characters.

Our research reveals a substantial number of new IDNs affected by the homograph problem. Besides the popular domain names, we identified some less well-known domain names with a greater number of homograph IDNs, like *csileasing.com*, *allsta.ecorporation.info*, which highlight the limitations of the analysis methods employed in prior studies. From an industry perspective, the majority of newly discovered brands are in the insurance and finance sectors, with the exception of *cavinklein.com*, which is a clothing trade corporation.

Finding 10

The scope of the IDN homograph problem is larger than expected; each enterprise has a considerable number of homograph domain names that can be registered.

Upon analysis of the 50 top companies, it was found that few have paid attention to IDNs. The top 50 enterprises in the Fortune ranking were selected as research subjects to analyze the attention of companies for IDNs in different industries. These companies encompass industries such as finance, energy, retail, manufacturing, and information technology. Two companies have applied their own top-level domain, i.e., Toyota Motor's *global.toyota* and Nippon Telegraph and Telephone's *group.ntt*. We collected their active domains in gTLDs for our analysis. IDNMon generated 4,690,829 valid IDNs by applying the domain names of 50 companies as input. Then, IDNMon analyzed their registration and resolution information and determined owner consistency between brand domain names and their IDNs by utilizing website data. The analysis results demonstrate that only a limited number of IDNs have been registered, with each company possessing a considerable number of registrable homograph IDNs. Fig. 10 shows the owner distribution of registered IDNs for each brand domain name. We observed that registered IDNs are not properly utilized, and their owners are not the owners of their corresponding brand domain names. Thus, it appears that most companies in various industries are yet to take advantage of the opportunities presented by IDNs, and they may be overlooking the associated challenges.

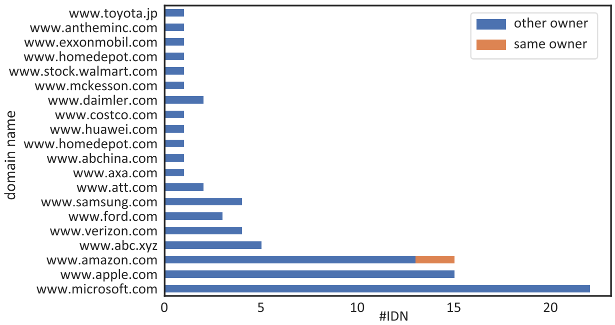


Fig. 10. Owner distribution of registered IDNs for each brand domain name.

Finding 11

Traditional IDN homograph protection strategies require increased attention to non-popular domain names.

Despite the limited size of the blacklist from the public project [20] attributed to the data source, an interesting finding is that more homograph domain names are non-popular domain names. Similar to extracting IDNs from zone files (Section III), we collected malicious IDNs by searching for *xn--* from the blacklist. Finally, we found 228 (0.07%) IDNs from the 325,073 blacklisted domain names. Using IDNMon to analyze the domain names in the blacklist, we observed that the majority of domain names that are blacklisted are not traditionally popular domain names. Conversely, they are two cryptocurrency sites, *blockchain.com* and *metamask.io*. Existing IDN homograph protection policies, which filter popular domains, may encourage attackers to target non-popular domains with IDN homograph attacks.

Summary: The issue of IDN homographs remains severe, with many well-known domain names having a large number of registrable IDN variants, especially those belonging to renowned enterprises. On the other hand, not just brand domain names, but also the homograph issues of certain special short domain names are particularly prominent, reflecting the active presence of IDNs in the domain name market.

VII. Discussion

This section briefly provides actionable suggestions and discusses some issues requiring ongoing attention.

A. Actionable Suggestion

Our research illustrates that the adoption of IDNs has grown across different regions due to the collective efforts of many parties. However, IDNs are confronted with severe security issues, including 12.32% of IDNs suffering from the homograph problem and a lack of uniform specifications between registries, registrars, and browsers. For registries and registrars, it is advised that the latest IDN standards be strictly followed, with a limitation on special characters or supplementary registration validation added. We found that. DE Registry (DENIC) has a policy concerning the bundling of all variants of β [30]. For browsers, we recommend that they refine their solutions to deal with the homograph problem and it is inadequate to focus on the list of popular domains. Furthermore, we think that it is imperative that the deviation characters in IDNs be managed correctly to prevent any potential misuse. For individual registrants, we believe it is important to embrace the advancements of the Internet, yet we must be wary of engaging in personal speculation. We should be mindful not to be swayed by deceptive propaganda and not impulsively to purchase IDNs. As an example, we have observed that certain scammers are known to inflate the value of IDNs in order to lure users to register for them [31].

B. Potential Issues

Emoji Domain Name: Emojis are widely adopted by users as an emerging thing on the Internet. Emojis were introduced into IDNs in IDNA2003; however, some recent studies [32], [33] showed that emojis pose additional risks at the same time. Thus, in IDNA2008, emoji domain names were prohibited. We discovered 30,913 emoji domain names in our data, which indicates that there are still registrars supporting the registration of emoji domain names. Most emoji domain names are used for adult sites. We recommend that registrars strictly comply with domain name registration specifications and gradually clean up existing non-compliant domain names to maintain the normal and stable growth of the domain name market.

New iTLD: The IDN exacerbates the problems faced by the new TLD. We should support enterprises to register the iTLD to protect their brands; however, we should also protect the interests of the region itself. For the domain name *xn--cckwcxetd* (.アマゾン means amazon), Amazon.com, Inc. has registered it to protect its brand. However, we are unaware if there are potential pitfalls with "Amazon" as a regional name registered by a company. However, if the regional name is arbitrarily registered, it can have disastrous consequences, as expected. We recommend that additional discussions be added for specific domain names, especially those with regional names, and that binding norms for domain name use be added to regulate the use of domain names and prevent misuse.

VIII. Related Work

This section discusses related work in terms of IDN measurement and cybersquatting.

IDN Measurement: Although IDNs have been accepted by an increasing number of countries and regions, only a few studies have focused on their development statuses and security threats. Conversely, the IDN homograph problem has received considerable research attention. A visual similarity-based approach should be applied to detect homograph IDNs based on the principle of IDN homograph problems. To this end, Sawabe et al. [9] implemented a method to find homograph IDNs by using optical character recognition. The SSIM index was also used for IDN homograph detection [7], [10]. In addition to the visual similarity-based approach, a few studies generated homograph domain names by applying homograph databases [13], [34]. Suzuki et al. [23] developed a framework to identify homograph IDNs; this framework can generate a new homograph database in an automated manner. However, the above detection approaches need a reference domain name list as input, and this can limit the scope of studies. IDNMon overcomes this reliance on reference domain names and provides a more comprehensive analytical perspective.

Several studies have measured IDN homograph attacks in the wild. In 2006, Holgers et al. [3] conducted a passive measurement study on a campus network to find IDNs impersonating Alexa's top 500 sites. Le Pochat et al. [27] explored IDNs that could be exploited by an attacker and found that 43% of them were available for registration. Chiba et al. [26] performed a measurement study for deceptive IDNs and demonstrated that there are many IDN homograph attacks that target non-English brands. Liu et al. [7] performed a measurement study using IDNs discovered from 56 TLD zone files, and presented information about the IDN ecosystem. They found that most registrations were opportunistic, and a few IDNs were registered by brand owners. These studies focused on the measurement of IDN homograph attacks for special domain names, and they ignored the other risks introduced by IDNs. IDNMon not only analyzes the IDN homograph problem but also provides a comprehensive analysis of the IDN ecosystem.

Cybersquatting: Cybersquatting refers to the malicious registration of domain names that are similar to popular domain names for borrowing their reputation [35]. According to prior studies, common cybersquatting attacks consist of typo [36], [37], [38], [39], bit [40], [41], homograph [3], combo [42], level [43], and wrong TLD [13].

IX. Conclusion

The IDNMon framework was proposed in this paper to understand the IDN ecosystem. IDNMon overcomes the dependence on reference domain names, and finds all potential IDNs with the homograph problem. Moreover, we performed a large-scale measurement study using IDNMon to show the following: (1) The popularity of IDN is expanding regionally, which has updated the understanding of prior studies; (2) The IDN homograph problem is universal, and we discovered 323,135 IDNs with homograph problems that had been overlooked in previous studies; (3) The development of IDNs remains limited by many factors, such as language environment and promotion strategies of registrars. We hope that our work helps users understand IDNs and promote the sustainable and secure development of the IDN ecosystem.

Ethics Statement

In this section, we discuss issues related to the ethical conduct of this research.

We carefully designed our study to identify and address potential ethical risks upfront. IDNMon primarily relied on publicly available datasets. All data does not involve user privacy information and is used only for academic research. In addition, we have followed the community practice of data crawling (such as PlayDrone [44]) and limited our number of requests per second so as not to affect the standard service of the target when collecting dynamic data.

ACKNOWLEDGMENT

Authors thank all anonymous reviewers for their helpful suggestions to improve the paper.

REFERENCES

[1] P. Mockapetris, “Domain names - concepts and facilities,” 1987. [Online]. Available: https://www.rfc-editor.org/rfc/rfc1034

[2] P. Faltstrom, P. Hoffman, and A. Costello, “Internationalizing domain names in applications (IDNA),” RFC 3490, 2003. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3490.txt

[3] T. Holgers, D. E. Watson, and S. D. Gribble, “Cutting through the confusion: A measurement study of homograph attacks,” in *Proc. USENIX Annu. Tech. Conf.*, 2006, pp. 261–266.

[4] M. Kumar, “Phishing attack is almost impossible to detect on chrome, firefox and opera,” 2023. [Online]. Available: https://thehackernews.com/2017/04/unicode-Punycode-phishing-attack.html

[5] J. Klensin, “Internationalized domain names for applications (IDNA): Definitions and document framework,” 2010. [Online]. Available: https://www.rfc-editor.org/rfc/rfc5890

[6] H. Hu, S. T. K. Jan, Y. Wang, and G. Wang, “Assessing browser-level defense against idn-based phishing,” in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 3739–3756.

[7] B. Liu et al., “A reexamination of internationalized domain names: The good, the bad and the ugly,” in *Proc. 48th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, 2018, pp. 654–665.

[8] M. Lui, “Langid.py: An off-the-shelf language identification tool,” 2023. [Online]. Available: https://github.com/saffsd/langid.py

[9] Y. Sawabe, D. Chiba, M. Akiyama, and S. Goto, “Detection method of homograph internationalized domain names with OCR,” *J. Inf. Process.*, vol. 27, pp. 536–544, 2019.

[10] T. P. Thao, Y. Sawaya, H. Nguyen-Son, A. Yamada, K. Omote, and A. Kubota, “Hunting brand domain forgery: A scalable classification for homograph attack,” in *Proc. 34th IFIP TC 11 Int. Conf. ICT Syst. Secur. Privacy Protection*, Springer, 2019, pp. 3–18.

[11] ICANN, “Centralized zone data service (CZDS),” 2023. [Online]. Available: https://czds.icann.org/en/

[12] Karen Lentz, “New gTLD program next round update,” 2023. [Online]. Available: https://icann78.sched.com/event/1Sgpf/new-gtld-program-next-round-update

[13] K. Tian, S. T. K. Jan, H. Hu, D. Yao, and G. Wang, “Needle in a haystack: Tracking down elite phishing domains in the wild,” in *Proc. Internet Meas. Conf.*, 2018, pp. 429–442.

[14] A. Costello, “Punycode: A bootstring encoding of unicode for internationalized domain names in applications (IDNA),” Network working group, 2003. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3492

[15] E. Gabrilovich and A. Gontmakher, “The homograph attack,” *Commun. ACM*, vol. 45, 2002, Art. no. 128.

[16] V. L. Pochat, T. van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, “Tranco: A research-oriented top sites ranking hardened against manipulation,” in *Proc. 26th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2019.

[17] Q. Scheitle et al., “A long way to the top: Significance, structure, and stability of internet top lists,” in *Proc. Internet Meas. Conf.*, 2018, pp. 478–493.

[18] Rapid7, “Forward DNS (FDNS),” 2023. [Online]. Available: https://opendata.rapid7.com/sonar.fdns_v2/

[19] T. Bohdan, “Domains project, world’s single largest internet domains dataset,” 2021. [Online]. Available: https://github.com/tb0hdan/domains

[20] M. Krog, “Phishing domain database,” 2023. [Online]. Available: https://github.com/mitchellkrogza/Phishing.Database

[21] L. Izhikevich et al., “ZDNS: A fast DNS toolkit for internet measurement,” in *Proc. 22nd ACM Internet Meas. Conf.*, 2022, pp. 33–43.

[22] M. Davis and M. Suignard, “Unicode security mechanisms,” 2023. [Online]. Available: https://www.unicode.org/reports/tr39/

[23] H. Suzuki, D. Chiba, Y. Yoneya, T. Mori, and S. Goto, “ShamFinder: An automated framework for detecting IDN homographs,” in *Proc. Internet Meas. Conf.*, 2019, pp. 449–462.

[24] K. Whistler, “Unicode normalization forms,” 2023. [Online]. Available: https://www.unicode.org/reports/tr15/

[25] R. Alrfou, “Polyglot: A natural language pipeline that supports massive multilingual applications,” 2023. [Online]. Available: https://github.com/aboSamoor/polyglot

[26] D. Chiba, A. A. Hasegawa, T. Koide, Y. Sawabe, S. Goto, and M. Akiyama, “Domainscouter: Analyzing the risks of deceptive internationalized domain names,” *IEICE Trans. Inf. Syst.*, vol. 103-D, no. 7, pp. 1493–1511, 2020.

[27] V. L. Pochat, T. van Goethem, and W. Joosen, “Funny accents: Exploring genuine interest in internationalized domain names,” in *Proc. 20th Int. Conf. Passive Act. Meas.*, Springer, 2019, pp. 178–194.

[28] T. Vissers, W. Joosen, and N. Nikiforakis, “Parking sensors: Analyzing and detecting parked domains,” in *Proc. 22nd Netw. Distrib. Syst. Secur. Symp.*, 2015, pp. 53–53.

[29] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang, “Understanding the dark side of domain parking,” in *Proc. 23rd {USENIX} Secur. Symp.*, 2014, pp. 207–222.

[30] M. Davis and M. Suignard, “Unicode IDNA compatibility processing,” 2022. [Online]. Available: https://www.unicode.org/reports/tr46/

[31] SIQIHULIAN, “Chinese IDN scams are popping up all over the place,” 2022. [Online]. Available: https://zhuanlan.zhihu.com/p/507386354

[32] M. Liu, Y. Zhang, B. Liu, and H. Duan, “Exploring the characteristics and security risks of emerging emoji domain names,” in *Proc. 27th Eur. Symp. Res. Comput. Secur.*, Springer, 2022, pp. 186–206.

[33] ICANN, “Emojis in domain names: A security risk for everyone,” 2023. [Online]. Available: https://www.icann.org/en/system/files/files/idn-emojis-domain-names-13feb19-en.pdf

[34] F. Quinkert, T. Lauinger, W. K. Robertson, E. Kirda, and T. Holz, “It’s not what it looks like: Measuring attacks and defensive registrations of homograph domains,” in *Proc. 7th IEEE Conf. Commun. Netw. Secur.*, 2019, pp. 259–267.

[35] Y. Zeng, X. Chen, T. Zang, and H. Tsang, “Winding path: Characterizing the malicious redirection in squatting domain names,” in *Proc. 22nd Int. Conf. Passive Act. Meas.*, Springer, 2021, pp. 93–107.

[36] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, “Seven months’ worth of mistakes: A longitudinal study of typosquatting abuse,” in *Proc. 22nd Annu. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, February 8–11, 2015.

[37] M. T. Khan, X. Huo, Z. Li, and C. Kanich, “Every second counts: Quantifying the negative externalities of cybercrime via typosquatting,” in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 17–21, 2015, pp. 135–150.

[38] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Félegyházi, and C. Kanich, “The long ”taile” of typosquatting domain names,” in *Proc. 23rd USENIX Secur. Symp.*, K. Fu and J. Jung, Eds, San Diego, CA, USA, Aug. 20–22, 2014, pp. 191–206.

[39] Y. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels, “Strider typo-patrol: Discovery and analysis of systematic typo-squatting,” in *Proc. 2nd Workshop Steps Reducing Unwanted Traffic Internet*, San Jose, CA, USA, Jul. 7, 2006, pp. 2–2.

[40] T. Vissers, T. Barron, T. van Goethem, W. Joosen, and N. Nikiforakis, “The wolf of name street: Hijacking domains through their nameservers,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dallas, TX, USA, Oct. 30 - Nov. 03, 2017, pp. 957–970.

[41] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, “Bitsquatting: Exploiting bit-flips for fun, or profit?,” in *Proc. 22nd Int. World Wide Web Conf.*, Rio de Janeiro, Brazil, May 13–17, 2013, pp. 989–998.

[42] P. Kintis et al., “Hiding in plain sight: A longitudinal study of combosquatting abuse,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dallas, TX, USA, Oct. 30 - Nov. 03, 2017, pp. 569–586.

[43] K. Du et al., “Tl;dr hazard: A comprehensive study of levelsquatting scams,” in *Proc. 15th EAI Int. Conf. Secur. Privacy Commun. Netw.*, Springer, Orlando, FL, USA, Oct. 23–25, 2019, pp. 3–25.

[44] N. Viennot, E. Garcia, and J. Nieh, “A measurement study of Google play,” in *Proc. ACM SIGMETRICS / Int. Conf. Meas. Model. Comput. Syst.*, Austin, TX, USA, Jun. 16–20, 2014, pp. 221–233.

Yunyi Zhang received the BS degree in computer science and technology from the Qingdao University of Science and Technology, Qingdao, China, in 2018 and the MS degree in cyberspace security from Sun Yat-sen University, Guangzhou, China, in 2021. He is currently working toward the PhD degree in cyberspace security with the School of National University of Defense Technology, Hefei, China. His research interests include DNS measurement, network security, and cyber crime.

Chengxi Xu was born in China. He received the BSc degree in automation from Electronic Engineering Insitute in 2010 and the MSc degree in computer application technology from Electronic Engineering Insitute in 2013. He is currently working toward the PhD degree in cyberspace security in National University of Defense Technology. His research interests include Internet infrastructure security and network measurement.

Fan Shi was born in 1938, an associate professor with the College of Electronic Engineering, National University of Defense Technology. His research interests include construction of knowledge graph for cyberspace security, and cyberspace surveying and mapping.

Miao Hu received the BS degree in systems engineering from the PLA University of Science and Technology, NanJin, China, in 2006 and the MS degree in communication and information system from the PLA University of Science and Technology, NanJin, China, in 2009. His research interests include web security, construction of knowledge map, and neural language processing.

Min Zhang was born in 1966. He received the PhD degree from Anhui University, China. Now he is a professor with the College of Electronic Engineering, National University of Defense Technology. His research interests include communication network security, intelligent computation, and vulnerability analysis.

Yuwei Li received the PhD degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2021. She is now a lecturer of electronic engineering, National University of Defense Technology, Hefei, China. Her interested research directions include software security and cyber security.

Zhijie Xie received the BS degree in communication engineering from Xiamen University, Amoy, China, in 2018 and the MS degree in cyberspace security from the National University of Defense Technology, Hefei, China, in 2020. He is currently working toward the PhD degree in cyberspace security from the National University of Defense Technology, Hefei, China. His research interests include identity authentication security and data mining.