A Reexamination of Internationalized Domain Names: the Good, the Bad and the Ugly

Baojun Liu*, Chaoyi Lu*, Zhou Li†, Ying Liu*✉, Haixin Duan*, Shuang Hao‡ and Zaifeng Zhang§
* Tsinghua University, † IEEE Member, ‡ University of Texas at Dallas, § Netlab of 360

***Abstract*—Internationalized Domain Names (IDNs) are domain names containing non-ASCII characters. Despite its installation in DNS for more than 15 years, little has been done to understand how this initiative was developed and its security implications. In this work, we aim to fill this gap by studying the IDN ecosystem and cyber-attacks abusing IDN.**

In particular, we performed by far the most comprehensive measurement study using IDNs discovered from 56 TLD zone files. Through correlating data from auxiliary sources like WHOIS, passive DNS and URL blacklists, we gained many insights. Our discoveries are multi-faceted. On one hand, *1.4 million* IDNs were actively registered under over 700 registrars, and regions within east Asia have seen prominent development in IDN registration. On the other hand, most of the registrations were opportunistic: they are currently not associated with meaningful websites and they have severe configuration issues (e.g., shared SSL certificates). What is more concerning is the rising trend of IDN abuse. So far, more than 6K IDNs were determined as malicious by URL blacklists and we also identified 1,516 and 1,497 IDNs showing high visual and semantic similarity to reputable brand domains (e.g., *apple.com*). Meanwhile, brand owners have only registered a few of these domains.

Our study suggests the development of IDN needs to be re-examined. New solutions and proposals are needed to address issues like its inadequate usage and new attack surfaces.

I. Introduction

Domain Name System (DNS) provides translation between domain names and IP addresses and is one of the cornerstones in the Internet infrastructure. In the beginning stage of Internet, only letter, digits, and hyphen were allowed and most of the domain names came from English words. To build a multilingual Internet and make the access easier for people around the globe, especially from eastern countries, IETF proposed Internationalized Domain Name (IDN) initiative and established standard to support domain names encoded with Unicode characters.

Despite its long history (more than 15 years after the first IDN installation), we still lack a good grasp of how IDN is positioned in the contemporary Internet ecosystem. So far, there is no comprehensive study to answer basic questions like how many IDNs are registered and what are their main usage. In fact, IDN has been constantly receiving criticisms. Prior works about this topic all focused on the security issues it brings in [21, 25, 35, 37]. Since an IDN registrant is free to choose characters of any language, she can create an IDN looking almost the same as a brand domain by replacing certain ASCII characters with Unicode characters. Such attack is called homograph attack. Interestingly, though this attack is known for a decade, it hasn't caught people's attention till recently as researchers demonstrated that a nearly perfect phishing attack against *apple.com* is possible with the help of IDN, and several major browsers were vulnerable [36]. Despite the security issues, we believe it is still too early to claim failure of the IDN initiative. Instead, we need to revisit the development of IDN and examine the scale of IDN abuse.

Our study. In this paper, we performed a measurement of IDN ecosystem from both business and security perspectives. Different from prior works constructing IDN dataset from network traffic [25, 37], we obtained IDNs by *scanning the entire zone files* of popular gTLDs and iTLDs. We were able to compile a list containing *1.4 million* IDNs registered under 56 TLDs. Compounding this list, we collected auxiliary data including three blacklists, WHOIS, passive DNS and SSL certificates to study the characteristics of IDNs (Section III). In particular, we looked into the languages associated with IDNs, registration statistics, presence in terms of DNS traffic, domain usage and security enforcement in terms of HTTPS deployment (Section IV). These results were also compared to non-IDNs. To assess how IDN is abused now, we first performed an empirical analysis on the malicious IDNs labeled by blacklists (Section V). In addition to the known homograph attack, we also identified a new type of IDN attack which exploits the semantic similarity between IDN and brand domain (called *semantic attack*). We developed two methods to identify IDNs potentially used for homograph and semantic attacks (Section VI and Section VII).

Findings. Putting together, our study shows a large volume of IDNs have been registered under many registrars but their value to Internet users is limited so far. The issue of IDN abuse indeed requires more attention from our community. Here we highlight some of the findings: 1) The 1.4 million IDNs we identified are provided by over 700 registrars. However, only a small proportion (below 20%) serves meaningful websites and mis-configuration exists in almost all HTTPS-enabled IDNs (over 97%). 2) Most of the mainstream browsers have responded to the latest homograph attack but several browsers are still vulnerable. What's more, through our detector, 3,013 registered IDNs were found to have high visual or semantic similarity with known brand domains. Only 6.0% of them were registered pro-actively by brand owners. The space for IDN abuse is substantial, as at least 47K IDNs (most of them unregistered) could be used for homograph attack.

II. Background

In this section, we briefly overview how domain names and IDNs are created, followed by the translation mechanisms and homograph attacks powered by IDN.

Domain names. A domain name is presented in a hierarchical string with each level related to a zone. DNS root zone (represented as a dot) is the top of the domain hierarchy. Under DNS root zone is Top-Level Domain (TLD), including generic TLD (gTLD), country-code TLD (ccTLD) and sponsored TLD (sTLD), which are managed by registries like Verisign. Under TLD, Second-Level Domains (SLD) are offered to public by different registrars, like GoDaddy. As an example, the TLD and SLD of $www.example.com$ are com and $example.com$.

Internationalized Domain Name (IDN). As mentioned, domain names in the beginning only allowed English letters, digits and hyphens. To enable people around the world using domain names in their native languages, like Chinese and Russian, ICANN issued guidelines and instituted a program to support the development and promotion of IDN, which encodes language-specific script or alphabet in multi-byte Unicode. So far, many efforts have been devoted by the Internet community to regulate IDNs and push for wide adoption [2, 12, 16, 24, 31–33].

While Unicode characters have been allowed at second and deeper levels since long time ago, it was until 2009 that the use of Unicode at top-level (called iTLD) was approved. Now both ccTLD and gTLD allow Unicode characters. The support from the domain industry is broad: all popular registries (e.g., com, net and org registries) accept registration of IDN below TLDs, and 150 iTLDs have been installed into the DNS root zone, such as ·中国(*xn–fiqs8s, China*)

For a domain registrant, getting an IDN SLD from a registrar is straightforward, with only one more step than registering a non-IDN SLD. According to Verisign [51], upon receiving a registration request, the registrar should first convert the requested domain into an ASCII-compatible encoding (ACE) string, and subsequently submit the ACE string to the Shared Registration System (SRS) for validation. When the domain name is valid and not registered, the requested IDN will be installed into the corresponding TLD zone. At the top level, the process of applying for an iTLD is similar to a new gTLD application, in which ICANN takes a thorough review and the whole process usually takes 20 months in average.

Punycode. Although IDNs with non-ASCII characters are supported by DNS, they have to be converted to ASCII characters to retain backward compatibility in many network protocols. Internationalizing Domain Names in Applications (IDNA) is such a mechanism that defines the translation between IDN and its corresponding ACE string [16], and has been adopted by major browsers and email applications. For these applications, before issuing a DNS request of an IDN, the domain name is translated into its ASCII version, or *Punycode* [12]. Specifically, Punycode uses an algorithm called *Bootstring* for such conversion, which keeps all ASCII characters, encodes the location of non-ASCII characters, and re-encodes the non-ASCII characters with generalized variable-length integers. A prefix $xn--$ is added to the converted Punycode after the above process. When an IDN is displayed by applications, the Bootstring algorithm is reversed to compute the Unicode values from ACE.

Homograph domain name spoofing attack. As different languages may have characters with similar shapes, attackers can construct an IDN with high visual resemblance to a known brand domain, in an attempt for phishing. Such attack is called *homograph domain name spoofing*, which was known at the beginning of IDN implementation [25]. However, even 10 years later, the problem still exists and plagues major browsers. In April 2017, a security researcher demonstrated that it is possible to create a phishing webpage highly similar to *apple.com*, using an IDN which visually resembles the brand domain when displayed in the Google Chrome address bar [36]. The trick is to replace the ASCII "a" (U+0041) in *apple.com* with Cyrillic "а" (U+0430) in the registered IDN. This attack raised broad attention and led to quick fixes from major browsers, some even terminating the support of IDN. However, this issue is not entirely addressed, as described later (Section VI-A).

III. Data Collection

Previous studies collected IDNs from network traffic from users [25, 37] and the data volume is small. On the contrary, we collected IDNs by *scanning zone files of TLDs*. In addition, we utilized auxiliary data like WHOIS and passive DNS to learn the development and distribution of IDNs. Below we elaborate each source and Table I summarizes the statistics.

TLD zone files. While Unicode is allowed to appear within any level of domain name hierarchy, we focus on the IDNs embedding Unicode at 2nd-level and top-level, because they can be obtained from zone files available to public. For 2nd-level IDNs, we downloaded three zone file snapshots from Verisign (for com and net) [52] and PIR (for org) [44], and identified IDNs using the prefix $xn--$. For top-level IDNs, we also searched substring $xn--$ in TLDs, and collected 53 zone files regarding iTLD [26] (all domains under these TLDs are IDNs). In the end, we scanned over 154 million domain names from three gTLDs and 53 iTLDs, and were able to extract *1, 472, 836 IDNs*, making the data scale several orders of magnitude higher than prior studies. Among these IDNs, more than two thirds are registered under com TLD.

To compare the characteristics of IDNs to those of non-IDNs, we also randomly sampled 1M, 100K and 100K non-IDNs from com, net and org zone files.

WHOIS database. To obtain the registration information of IDNs, we leveraged the WHOIS records published by registrars. Our industrial partners helped us to obtain WHOIS information of 739, 160 (50.19%) IDNs and parse them using a variety of tools, like python-whois. The two major reasons for missing WHOIS of the remaining IDNs are the request block from some registrars and parsing failures from the WHOIS crawler. In fact, the support of iTLD is very poor from WHOIS parsers: only 1.1% IDNs under iTLDs are correctly parsed.

TABLE I: Datasets collected

TLD	Snapshot on	# SLD	# IDN	Domain WHOIS	Blacklisted			
					VirusTotal	360	Baidu	Total
com	2017/09/21	129,216,926	1,007,148	590,542	3,571	1,807	26	5,284
net	2017/09/21	14,785,199	231,896	131,573	661	91	1	746
org	2017/10/05	10,390,116	25,629	19,271	56	2	1	59
iTLD (53)	2017/10/03	208,163	208,163	2,226	90	63	2	152
Total	-	**154,600,404**	**1,472,836**	**739,160**	4,378	1,963	30	**6,241**

Passive DNS. For each IDN, we are interested in the volume of network traffic it received and the time period of the user visits. To this end, we leveraged the passive DNS data provided by 360 DNS Pai Project [46] and Farsight Security [17]. The DNS Pai project has been collecting DNS logs from a large array of DNS resolvers since 2014, which now handles 240 billion DNS requests per day. Because our account under DNS Pai has no query limit, we submitted all 1.4 million IDNs for their DNS logs. On the other hand, the passive DNS database from Farsight has better coverage of resolvers outside China, but has a query limit of only a thousand domains per day. As a result, we only requested DNS logs of abusive IDNs detected by our system. Both data sources provide statistics of DNS look-ups aggregated per domain, which contain the number of look-ups and timestamps of the first and last look-up. To notice, as listed in Table I, our collected data from DNS Pai spans from 2014/08/04 to 2017/10/13. From Farsight, our collected data spans from 2010/06/24 to 2017/12/03.

URL blacklist. Since IDNs can be abused to launch homograph attacks, we want to learn whether IDN abuse is pervasive and if there are other attack vectors originated from IDN. We leveraged three URL blacklists from VirusTotal, Qihoo 360 and Baidu. If an IDN is alarmed by any of the blacklists, we considered the IDN as malicious. In the end, our blacklists contain 6,241 IDNs (0.42%), the details shown in Table I. Most of malicious IDNs are under normal gTLDs and only 152 IDNs are under iTLDs.

Alexa Top Sites. Attackers abusing IDNs usually target well-known brand domains, like *apple.com*. In this study, we selected the top 1K SLDs based on Alexa website ranking as the potential victims of IDN abuse.

SSL certificates. Finally, we collected SSL certificates associated with IDNs to study whether security practices, i.e., traffic encryption, are followed and how they are executed. In particular, we used OpenSSL to connect to port 443 of remote hosts associated with IDNs and fetch the certificate chains of IDNs. The validity of all certificates were checked by OpenSSL as well. Similarly, we collected SSL certificates from our sampled 1.2 million non-IDNs for comparison.

Limitations of data. Although we tried to make the study as comprehensive as possible by using data from many sources, there are still limitations. First, our IDN list does not contain IDN domains with Unicode characters at 3rd level or deeper, due to the limitation of zone files. Nevertheless, previous study showed that IDNs of those cases only account for 6.05% of all IDNs they observed [37]. As a result, the measurement result would not differ significantly. Second, we did not collect IDNs under country-code TLDs (ccTLD) because most of the zone files are kept private by their registries. Additionally, including IDNs under a subset of ccTLDs could introduce bias regarding registrants' language and geographic location. For example, nearly all IDNs under *cn* ccTLD contain Chinese characters. Finally, false positives and false negatives are unavoidable in the blacklists we use. Regarding false positives, their quantity should be very small, as reflected from our manual analysis on a sample of domains. False negative is a bigger issue as many of the IDNs were not even encountered by security companies. As such, we scanned all IDNs using an in-house detector based on visual resemblance (elaborated in Section VI and VII), detected many new malicious IDNs and augmented the blacklisted IDNs.

IV. Overview of IDN Characteristics

The introduction of IDN enables domain names to contain characters from languages other than English. Registration of IDN has been growing substantially since its birth (more than 1.4 million IDNs are currently listed by the TLDs we surveyed), making the Internet more accessible to users worldwide, which meets its original design purpose. However, problems do exist. Despite the sheer volume of domains, the fact that only a small proportion of IDNs are actually in use implies the less value provided by them to the Internet community than non-IDNs. Other than benefits, IDN opens up new attack surface (e.g., homograph attack) and incurs more cost on the side of brand owners for protection. In this section, we describe our examination of the ecosystem around IDN and give quantitative analysis regarding each finding.

A. Language

The first question we ask is what languages are favored by IDN registrants. By looking into the language distribution, we are able to learn which countries are actively promoting IDN development, as well as attackers' preferences. In particular, we leveraged a tool called *LangID* to identify the most likely language of each IDN [40, 41]. LangID uses a multinomial Bayes learner trained by five language-labeled datasets to predict the probability of each language on an IDN. LangID could achieve reasonable accuracy for this task: as demonstrated in the prior works [40], the accuracy ranges from 0.904 to 0.992 on different datasets. In the end, we were able to identify the language for all IDNs. Below, we describe our findings.

Finding 1. More than 75% of all IDNs are registered in languages spoken in east Asian countries. East Asian countries turn out to stay at the forefront of IDN promotion.

TABLE II: Languages of all and malicious IDNs (Top 15)

Language	IDN		Blacklisted	
	Volume	Rate	Volume	Rate
Chinese	766,135	**52.03%**	3,495	**56.02%**
Japanese	191,058	**12.97%**	238	**3.81%**
Korean	128,291	**8.71%**	902	**14.46%**
German	72,110	4.90%	119	1.91%
Turkish	43,100	2.93%	196	3.14%
Thai	36,660	2.49%	357	5.72%
Swedish	32,275	2.19%	51	0.82%
Spanish	25,310	1.72%	97	1.55%
French	24,771	1.68%	56	0.90%
Finnish	17,609	1.20%	36	0.58%
Russian	13,972	0.95%	96	1.54%
Hungarian	11,969	0.81%	36	0.58%
Arabic	12,419	0.84%	43	0.69%
Danish	8,544	0.58%	22	0.35%
Persian	7,976	0.54%	28	0.45%
Total	1,392,199	**94.54%**	5,772	**92.22%**

As shown in Table II, more than 75% of all IDNs are related to Chinese, Japanese, Korean and Thai. Meanwhile, among all 150 iTLDs approved by ICANN, more than 60% are in east-Asian languages (e.g., 62 in Chinese, 9 in Japanese, 4 in Korean and 4 in Mongolian). In fact, countries including China, Japan and Korea have launched many promotional programs to push IDN registration [10]. Our results suggest these efforts are well rewarded so far. Another explanation could be that compared to western countries whose residents are more familiar to English, IDN is more attractive to Internet users in east Asia.

Next, we looked into the languages presented by the 6,241 malicious IDNs labeled by blacklists. We found the distribution is similar to the overall: languages having more IDNs are more likely to contain malicious domains. Chinese tops the chart for both overall and malicious IDNs. By investigating the semantic meanings of malicious IDNs in Chinese, we found that underground business (e.g., online gambling, which is illegal in China) is using IDN to promote their illegal products and services. Different from homograph IDN used for phishing, IDNs registered under this setting do not impersonate brand domains.

B. Registration Characteristics

We identified more than 1.4 million IDNs from 154 million SLDs. Though the ratio of IDN is still small (only 1%), suggesting non-IDNs are dominating the domain business, its absolute number proves the business value of this IETF initiative. By correlating IDNs with WHOIS data, we studied how they are distributed across registrars, registrants and registration time windows. We report our key findings below.

Finding 2: 6.16% (90,708) IDNs were created before 2008, registered for at least ten years. Although our snapshots of zone files were all collected recently, a large number of IDNs registered 10 years ago were still recorded, suggesting there are serious registrants willing to keep IDNs for long-term business. Figure 1 presents the creation dates of IDNs, with malicious ones shown separately. In general, the number of registrations rises along the timeline (similar for malicious IDNs) but we did observe several spikes. The

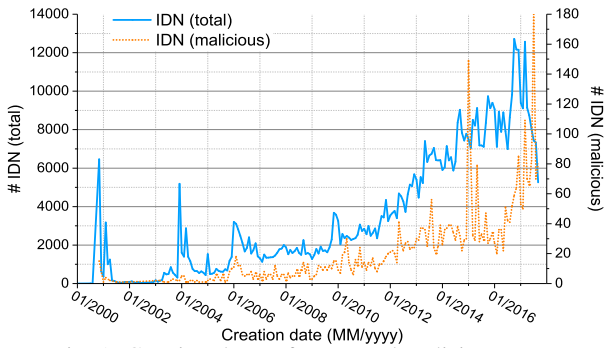


Fig. 1: Creation dates of IDNs and malicious IDNs

TABLE III: Top 5 IDN registrants

Email Account	# IDN	IDN Characteristics
776053229@qq.com	2,609	All are southwest city names in China.
daidesheng88@gmail.com	1,562	All are about online gambling.
tetetw@gmail.com	1,453	All are short words in Chinese.
840629127@qq.com	1,324	All are related to Chongqing, China.
776053229@163.com	1,178	All are southwest city names in China.

spikes of overall registration seem to be relevant to big events in domain community: the spike in 2000 overlaps with the launch of IDN testbed by Verisign GRS [28] and the spike in 2004 follows the introduction of German and Latin characters in domain names [27]. For malicious registrations, we also found two spikes in 2015 and 2017. After inspecting the registrants' emails, we found the spikes were caused by cybersquatting by a few registrants. As an example, a registrant under *13779950000@139.com* registered 126 IDNs in Chinese in Mar. 2017, with all being related to online gambling.

Finding 3. A few registrants performed large-scale opportunistic registrations and grabbed 29,318 (4%) IDNs. A registrant performing opportunistic registration grabs many domains they believe are attractive to other buyers or can be monetized through parking. They usually have no intention in developing websites on top of the domains. Often, opportunistic domains under one registrant are of one specific topic (e.g., online gambling and shopping) or pattern (e.g., short words and city names). Through manual analysis, we observed a few registrants performing opportunistic registration extensively. Since personal emails are used for all such registration, the registrations are unlikely to be defensive (i.e., registration performed by a reputable company for brand protection). Table III shows the top 5 registrant emails ordered by number of IDNs. We found all IDNs owned by a registrant are dedicated for the same purpose, by analyzing the meaning of domain names. As an example, all 1,562 IDNs registered by *daidesheng88@gmail.com* are related to online gambling.

Finding 4. At least hundreds of registrars offer IDN registrations. 55% IDNs were registered by top 10 registrars. We identified over 700 registrars through clustering IDNs by their registrar field, showing IDN registration is an indispensable business category for many registrars. As shown in Table IV, more than half of IDNs belong to 10 registrars and 70% belong to top 20. Interestingly, though

TABLE IV: Top 10 most active registrars offering IDNs

Registrar	# IDN	Rate
GMO Internet Inc.	155,491	22.99%
HiChina Zhicheng Technology Limited.	73,439	10.86%
Name.com, Inc.	28,906	4.27%
Gabia, Inc.	27,201	4.02%
Dynadot, LLC.	21,578	3.19%
1&1 Internet SE.	19,512	2.89%
Chengdu West Dimension Digital. Technology Co., Ltd.	18,641	2.76%
eNom, LLC.	16,002	2.37%
DomainSite, Inc.	15,687	2.32%
GoDaddy.com, LLC.	12,717	1.88%

GoDaddy dominates the global domain market, it only takes a small share of 1.88% when it comes to IDNs. Registrars facing east-Asian markets are more active: for example, GMO, a Japanese Internet company, accounts for 23% IDNs and HiChina Zhicheng accounts for 10.86% IDNs. By contrast, the sampled 1.2 million non-IDNs belong to more than 1,500 registrars, suggesting not all registrars offer IDN service.

C. DNS Statistics

Our passive DNS dataset allowed us to assess the visits flowing to IDNs and estimate their popularity among Internet users. Here we consider two metrics, **active time** and **query volume**, for this measurement. Active time is the time span in which DNS requests are observed (the time difference between the first and last requests). Query volume is the total number of requests. As comparison, we also computed the same metrics for non-IDNs. Lastly, we extracted the observed IP addresses in their DNS responses to learn how IDNs are located.

Finding 5. IDNs have significantly shorter active time than non-IDNs, except for malicious IDNs. Illustrated in Figure 2, the distributions of active time between IDN and non-IDN are clearly separated. As an example, 60% of *com* IDNs stayed active for less than 100 days, while 40% non-IDNs under *com* have the same property. The differences become even larger under other TLDs. However, we notice that malicious IDNs tend to have longer active time, which are even close to non-IDNs (legitimate for most of them).

Finding 6. IDNs are visited less frequently than non-IDNs, except for malicious IDNs. Illustrated in Figure 3, query volume differs significantly between IDNs and non-IDNs: 88% *com* IDNs were queried less than 100 times, the rate being 74% for non-IDNs under *com*. Again, malicious IDNs witnessed larger traffic volume, even more than non-IDNs in average.

The observations on malicious IDNs indicate attackers are effective in trapping visitors. As an example, we found 波色.com (*xn--0wwy37b.com*), an illegal Chinese gambling site flagged as malicious, received 3,858,932 queries, the largest among all IDNs, and stayed active for 118 days. Meanwhile, malicious parties often choose to register IDNs which are highly deceptive to attract visitors. By contrast, benign IDNs are struggling to attract visits.

Finding 7. IP addresses of IDNs are concentrated. 106,021 IP addresses are identified from passive DNS, which are further mapped to 43,535 /24 network segments. We computed ECDF of IDNs over segments, as illustrated in Figure 4, and found that 80% IDNs are hosted by servers in 1,000 /24 network segments. Among the top 10 network segments hosting 24.8% IDNs, four belong to web hosting services (e.g., Linode), four belong to parking services (e.g., GoDaddy parking), one belongs to Akamai and the remaining one is a private network segment.

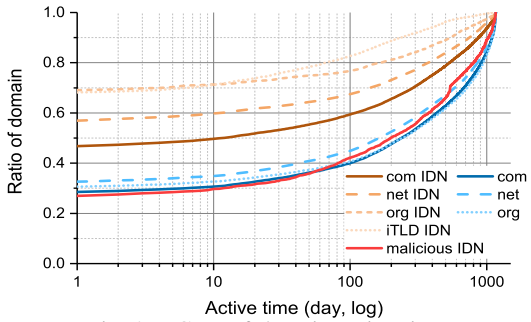


Fig. 2: ECDF of domain active time

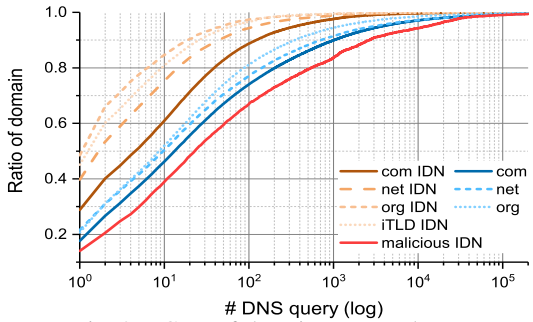


Fig. 3: ECDF of domain query volume

D. Content and Intention

To understand the motivation of registration, we performed content analysis on IDNs, using the homepages fetched by our web crawlers. As accurate content-based classification is challenging for a large volume of websites, we chose to sample a number of IDNs (500) and manually examine their content. We divided them into 7 general categories, as listed in Table V. In addition, we sampled the websites of non-IDNs and performed the same classification.

Finding 8. Visiting an IDN leads to meaningless content and resolving errors with much higher probabilities. We found that over 45% of the sampled IDNs are not resolved, while only 19% are running meaningful websites. The ratios for non-IDNs are 15.2% and 33.6% respectively. To notice, all

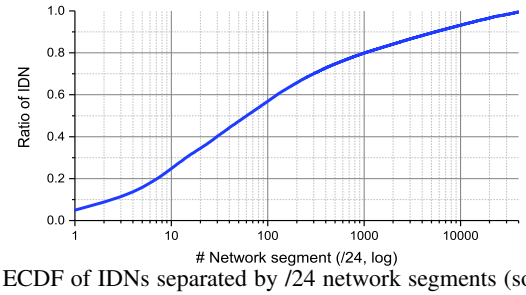


Fig. 4: ECDF of IDNs separated by /24 network segments (sorted by volume of IP addresses hosting IDNs)

TABLE V: Usage of domain names

Type	IDN	Non-IDN
Not resolved	**228 (45.6%)**	**76 (15.2%)**
Error	65 (13.0%)	74 (14.8%)
Empty	16 (3.2%)	43 (8.6%)
Parked	56 (11.2%)	107 (21.4%)
For sale	8 (1.6%)	16 (3.2%)
Redirected	28 (5.6%)	16 (3.2%)
Meaningful content	**99 (19.8%)**	**168 (33.6%)**
Total	500	500

IDNs in zone files have associated NS records so all resolution errors come from name servers (e.g., DNS REFUSED error). The large volume of resolution errors on IDNs means their owners are not even making the right DNS configuration. When a domain is not in active use, its owner can park the domains and gain earnings from advertisements. However, it turns out IDN owners prefer to leave it without any monetization intention. Among the IDNs with meaningful content, most are in Japanese and Korean, implying registrants in these countries are more serious in IDN registration and use IDN to deliver language-specific content.

E. SSL Certificate

HTTPS is the key to secure today's Internet traffic between users and web sites, and to mitigate the threats from eavesdroppers. Recent study shows the adoption of HTTPS grows steadily among reputable sites. We are interested in the trend of HTTPS adoption in IDN domains and whether SSL certificates are installed in the correct way. To this end, we queried for SSL certificates from 737,269 IDNs (50.06%) and 816,882 non-IDNs (68.07%) which can be resolved, and downloaded 67,087 (4.55%) and 35,028 (2.92%) certificates from IDNs and non-IDNs for further analysis.

Finding 9. The management of SSL certificates is very problematic among IDNs and more than 97% of them have security issues. Considering most IDNs are inactive, the result on SSL certificates is actually not as negative as the ratio depicts. However, a close look at the configurations of certificates revealed that they are oftentimes poorly installed. In fact, 12.54%, 18.14% and 67.28% certificates are expired, self signed and shared in an invalid way as shown in Table VI. The analysis on non-IDNs revealed the similar issue, but the ratio of expired certificates is higher and there are less shared certificates. Though surprising, we found that our results in fact resemble the prior studies in HTTPS measurement. Liang et al. [39] studied 10,721 DNS-CDN-Enabled sites and found 68.8% websites were using invalid SSL certificates mainly offered by CDN service providers. For the HTTPS ecosystem, Durumeric et al. [15] showed 4.6% of all domains had invalid certificates. We found 4.5% (65,713) invalid certificates among all 1.4M IDNs. Since most of our studied IDNs and non-IDNs are less known, this finding indicates there is still a long way ahead for deploying SSL certificates correctly in long-tail websites.

Next, we elaborate our observations around certificate sharing. In this case, a number of websites deploy the same SSL certificate whose owner field is inconsistent with the domain names of websites. This security issue broadly exists for both IDNs (67.28%) and non-IDNs (45.47%). Table VII presents the Common Names of top 10 certificates shared among IDNs, and we found that most of shared certificates belonged to domain parking and hosting services. When a domain is parked, parking service modifies the DNS response IP to one under itself [54]. Most owners have no motivation to do advanced configuration on parked domains, like installing their owned SSL certificates (if any), resulting in certificate sharing. Similarly, several hosting service providers are providing shared SSL certificates, such as *cafe24.com* based in Korea, which are also extensively used. Domains associated with identical IPs are more likely to share certificates.

TABLE VI: Security problems of IDN related to SSL certificates

Security Problem	IDN	non-IDN
Expired Certificate	8,411 (12.54%)	8,730 (24.92%)
Invalid Authority	12,169 (18.14%)	5,801 (16.56%)
Invalid Common Name	45,133 (67.28%)	19,527 (45.47%)
Total	65,713 (97.95%)	34,058 (97.23%)

TABLE VII: Analysis of shared certificates

Common Name (CN)	Volume	Description
sedoparking.com	27,139	Parking service.
cafe24.com	4,024	Hosting service provider.
ovh.net	3,691	Webmail service provider.
bizgabia.com	3,271	Hosting service provider.
03365.com	449	Same DNS resolution.
ihs.com.tr	314	Parking service.
seoboxes.com	230	Hosting service provider.
nayana.com	137	Hosting service provider.
suksawadplywood.co.th	123	Parking service.
ssl-sys.jp	117	Hosting service provider.

V. An Empirical Analysis of IDN Abuse

IDN abuse has been brought to discussion since the beginning of IDN implementation. However, only a few incidents, i.e., homograph attacks, have been reported so far [23, 36]. Little has been done to fully understand adversaries, including their registration patterns and intentions behind IDN abuse. In addition, a comprehensive analysis regarding how applications (e.g., browsers) handle IDN abuse is also missing. We aim to fill these missing pieces in this study.

By matching 1.4 million IDNs with blacklists, we identified 6,241 malicious IDNs. While homograph attack is a natural way to abuse IDN and many domains in our dataset are under this category, we discovered another category of abuse, which exploits the semantic similarity between IDNs and brand domains. These two attack categories are described below.

Homograph attack. Homograph attack exploits the visual resemblance of different characters, i.e., homoglyphs. Our dataset of malicious IDNs contains a large number of homographic IDNs and we select 12 malicious IDNs impersonating *facebook.com* as examples to highlight attackers' registration patterns (listed in Table VIII). In these cases, attackers replace 1 to 3 ASCII letters with characters of similar shapes from Vietnamese, Arabic, Icelandic and Yoruba.

Semantic attack. Besides homographic domains, there are malicious IDNs attempting to fool users using semantic sim-

TABLE VIII: Examples of malicious homographic IDNs

facebọọk.com	faċebook.com	faceḅook.com	faceboôk.com
facebọọk.com	fácebook.com	fâcêbook.com	facebook.com
facebóók.com	fạcẹbook.com	fạcebook.com	facebȯok.com

TABLE IX: Examples of Type-1 semantic abuse

IDN		Description
Punycode	Unicode Characters	
xn--icloud-uz2li34m.com	icloud 登录.com	icloud login
xn--icloud-1u6oy84r.com	icloud 登陆.com	icloud login
xn--apple-rq8mk98i.com	apple 邮箱.com	apple email
xn--apple-4l0lt9l.com	apple 激活.com	apple activate

ilarity. Such semantic-based abuse has never been reported before. We divide IDNs of this category into two types.

In Type-1 attack, adversaries combine brand domain names with keywords from another language to create IDNs. For example, adversaries may register *apple 售后.com* to impersonate Apple's customer services. A user could be fooled when she forgets to check SSL certificates or owner information of the domain. In Table IX we list four other malicious IDNs attacking Apple and iCloud. This attack vector is similar to combosquatting attacks which combine brand names with English keywords [30].

In Type-2 attack, IDNs are created by translating English brand names to other languages. An example we discovered was an IDN defrauding a Chinese company, Gree Air Conditioner. Though the company has registered the English domain name *gree.com.cn* long time ago, its Chinese version *格力空调* was registered by attackers. Table X lists several such IDNs.

Confirming whether domains are Type-2 abuse is challenging, as mapping a potential Type-2 abuse to its targeted brand is not always feasible. In this work, we focus on homograph attack and Type-1 attack.

VI. Homograph Attack

In this section, we first examine how latest browsers tackle homograph attacks. Then, we propose a method to detect registered homographic IDNs and estimate their scale (including unregistered ones).

A. Browsers

Following the homograph attack this year [23, 36], many browsers have upgraded their policies of IDN display. As an example, in Firefox, if all characters within one IDN label belong to a single character set, the IDN is displayed in Unicode characters [42]; Chrome adopts a similar policy with more restrictions [9]. As such, many of the homographic domains we found (see Table VIII) will be rendered in Punycode form, because each domain contains characters from at least two character sets. Alternatively, showing Punycode under all circumstances should mitigate the issue entirely, which is in fact the default option of some browsers. Nevertheless, this policy runs opposite to the IETF requirements [16] and we do not recommend this solution. In the end, we want to understand how IDN policies are enforced by browsers and how far it is till solving the entire problem. As such, we carried out a survey study of a set of browsers.

TABLE X: Examples of Type-2 semantic abuse

IDN		Description
Punycode	Unicode Characters	
xn--tfr361cl2mbrq.net	格力空调.net	Gree Air Conditioner
xn--tlqpa605apxhf4c468i.com	北京交通大学.com	Beijing Jiaotong University
xn--ztsu95bbqz6hj.com	奔驰汽车.com	Mercedes Benz Automobile

Specifically, we manually tested ten widely-used browsers on three different platforms (PC, iOS and Android). We inputted Unicode characters of homographic SLDs and checked how they are displayed in regions like address bar, status bar and title bar. Besides, we tested how IDNs under iTLDs are supported in the same experiment settings (e.g. testing *央视网.中国*, *xn--wss800gp5g.xn--fiqs8s*).

Our survey result is shown in Table XI. We found that browsers treat IDNs differently. Our first observation is that except one browser (Sogou PC), all others could address certain homograph attacks (examples in Table VIII). However, their security policies are not consistent. As an example, *soso.com* (all characters are from Cyrillic, with punycode being *xn--n1aa1eb.com*, mimicking *soso.com* which ranks 96 in Alexa) bypasses the policy of Firefox as *all* characters are in the same set. In the end, we found five browsers on PC and one on Android are vulnerable. Moreover, some mobile browsers (five browsers on iOS and three on Android) choose to display webpage titles in address bars when visiting IDNs. This setting is quite problematic, as adversaries can use a title which is identical to a brand domain's. Among all browsers, QQ browser is particularly interesting as it redirects user to *about:blank* for some IDNs (and displays Punycode for others). The reason behind this design is unclear.

Regarding iTLD IDNs, browser policies also differ. Firefox treats an iTLD IDN as a valid domain only if a protocol prefix (e.g., *http://*) is present. Though a browser should handle both Unicode and Punycode TLD based on standard, we found that three browsers on iOS and two on Android only recognize Unicode iTLDs. We speculate the TLD lists used by these browsers only contain the Unicode version of iTLDs. On the other hand, one Android browser only supports Punycode iTLDs. Surprisingly, Baidu browser on Android does not support iTLD at all, regardless of the format.

B. Detecting Homographic IDN

We identified some homographic IDNs and their targeted brands through manual analysis. This approach cannot scale on the blacklisted IDNs, not to mention the entire IDN dataset (1.4 million). To address this problem, we developed an approach to automatically detect homographic IDNs given a set of brand names.

In essence, our approach leveraged the visual resemblance between brand and homographic domain names. We first rendered the image of every IDN (1.4 million) and brand domain (Alexa Top 1k SLDs), and then measured their pair-wise visual resemblance. To calculate similarity between domains, we adopted a metric called Structural Similarity (SSIM) Index, which compares luminance, contrast and structure between

TABLE XI: Surveyed browsers under homograph attack

Platform / Browser	PC			iOS			Android		
	Ver.	iTLD IDN Supported	Homograph Attack	Ver.	iTLD IDN Supported	Homograph Attack	Ver.	iTLD IDN Supported	Homograph Attack
Chrome	62.0			61.0			61.0		
Firefox	57.0	Need prefix	Bypassed	10.1			57.0	Need prefix	Bypassed
Opera	49.0		Bypassed	16.0			43.0		
Safari	11.0			11.0			/	/	/
IE	11.0			/	/	/	/	/	/
QQ	9.7			7.9	Unicode only	Title	8.0	Unicode only	about:blank
Baidu	8.7		Bypassed	4.10	Unicode only	Title	6.4	Not supported	Title
Qihoo 360	9.1			4.0		Title	8.2	Punycode only	
Sogou	7.1		Vulnerable	5.10		Title	5.9	Unicode only	Title
Liebao	6.5		Bypassed	4.18	Unicode only	Title	5.22		Title

Need prefix: iTLD IDN supported only with a protocol prefix.
Unicode only / Punycode only: iTLD IDN supported only in one encoding.
Not supported: iTLD IDN not supported.
Vulnerable (bypassed): displaying (certain) homographic IDNs in Unicode.
Title: displaying web page titles in address bar.
about:blank: certain homographic IDNs lead to blank pages.
Blank cells: full support of iTLD IDN; homographic IDNs displayed in Punycode.

TABLE XII: Examples of IDNs and maximum SSIM Indices, all of which has a maximum SSIM Index (locally) with *google.com*

Maximum SSIM Index	IDN	
	Punycode	Unicode Characters
1.00	xn--ggle-55da.com	google.com
	xn--oole-z7bc.com	google.com
0.99	xn--googl-r51b.com	googlẹ.com
	xn--googl-n0a.com	googlę.com
0.98	xn--googe-95a.com	googĺe.com
	xn--oole-cxa13q.com	goo̧gle.com
0.97	xn--gogle-e7b.com	goȏgle.com
	xn--gogle-dua.com	gõogle.com
0.96	xn--gle-dub9525aa.com	ɢoogle.com
	xn--gogl-qqa1s.com	góoglĕ.com
0.95	xn--ggle-0qaa.com	gõõgle.com
	xn--goggl-mza.com	gogglē.com
0.94	xn--ggle-bqaa.com	gȏȏgle.com
0.93	xn--ggle-qoa8i.com	gåøgle.com
0.92	xn--bgle-5qaa.com	böögle.com
0.91	xn--bggie-g9a.com	bõggie.com
0.90	xn--donol-fsa.com	donolé.com

two images [56]. Compared to traditional similarity metrics like MSE [57], SSIM strikes a good balance between accuracy and runtime performance. Taking two images as input, this algorithm outputs a decimal index in the range of [-1, 1], with 1 implying perfectly identical.

More specifically, an IDN image is compared to each image of brand domain, to generate 1,000 SSIM Indices. If the *maximum* SSIM Index exceeds a certain threshold, the IDN is considered as homographic to a brand domain. Here we assume a homographic IDN should not impersonate more than one brand domain. We executed the experiment on a CentOS machine with 4GB memory, and the whole process was completed within 102 hours.

Selection of threshold. To determine the threshold of SSIM Index, we sampled several brand domains, replaced some letters with homoglyphs and reviewed the similarity from the perspective of normal users. We found the threshold works best when set to 0.95. As shown in Table XII, when the index drops below 0.95, the difference becomes quite prominent.

C. Registered Homographic IDNs

In total, **1,516** IDNs (out of 1.4 million registered IDNs) are considered homographic to Alexa Top 1k SLDs, including **91** domains which appear identically as their corresponding

TABLE XIII: Top 10 brand domains ordered by homographic IDNs

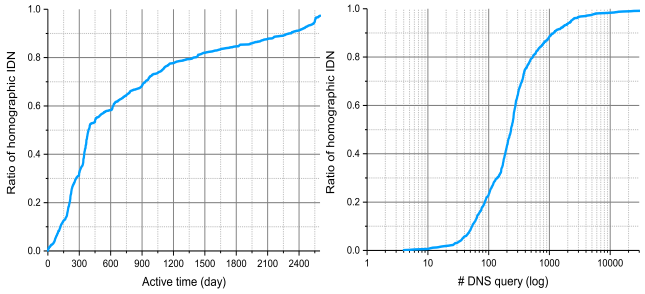
Domain	Alexa	# IDN	Rate	Protective Registrations
google.com	1	121	8.0%	19
facebook.com	3	98	6.5%	0
amazon.com	11	55	3.6%	14
icloud.com	372	42	2.8%	0
youtube.com	2	41	2.7%	0
apple.com	55	39	2.6%	0
sex.com	537	36	2.4%	0
go.com	391	29	1.9%	0
ea.com	742	28	1.8%	0
twitter.com	13	25	1.6%	5
Total		**514**	**33.9%**	**38**

brand domains. Among them, only 100 (6.6%) have been blacklisted. The registration intention of the remaining ones could be legitimate (defensively registered by brand owners), malicious or unknown (e.g., unresolved).

Registrants. The first question we have about these IDNs is how many of them were registered out of brand protection purposes. Using WHOIS data of 1,111 out of 1,516 IDNs, we manually checked whether they were registered under email accounts of brand companies, and found only 73 (4.82%) registrations under this category. Among the remaining domains, 171 were registered by parties using personal email addresses and others were registered anonymously (protected by WHIOS Privacy). Though we could not verify their registrants, it is quite unlikely that brand companies are behind them.

Brand domains. Next, we clustered the IDNs by their corresponding brand domains. 255 SLDs within Alexa Top 1k are targeted by homographic IDNs, showing the diversity of registrations. Table XIII presents the top 10 brands ordered by the number of associated IDNs. Google, Facebook and Amazon are the top three which are also ranked very high by Alexa. As for brand protection, we found that only Google, Amazon and Twitter perform protective registrations, but most of the IDNs are out of their reach.

DNS statistics. Leveraging Farsight Passive DNS data (explained in Section III), we found that homographic IDNs tend to have long active time. Illustrated in Figure 5(a), homographic IDNs have 789 active days in average, with 40%



(a) Active time (b) Query volume

Fig. 5: ECDF of active time and query volume of homographic IDNs

being active for more than 600 days. Among the IDNs with long lifetime, several IDNs are used for security education (telling visitors the domain is a demonstration for homograph attack, e.g., *xn--fcebook-hwa.com*), which could be the reason of long lifetime.

Likewise, homographic IDNs tend to receive more DNS requests. As shown in Figure 5(b), 80% of homographic IDNs receive more than 100 queries, with 10% queried for over 1,000 times. While the top three IDNs have received considerable volume of requests (over 100,000), all of them are parked (e.g., *xn--instagram-5jf.com*, a homograph of *instagram.com*).

Usage of homographic IDNs. To understand how homographic IDNs are used, we manually classified their websites using the same methodology described in Section IV-D. Our observation here is consistent with prior result: only a low proportion of them are in active use. Among 100 sampled domains, 34 are not resolvable, 10 are returning errors, 16 are for sale, 14 are parked, and 11 are hosting test pages. However, we did identify one case of homograph attack (*xn--80aa1cn6g67a.com*, which mimics *alipay.com*, one of the largest online payment platform of China, and has already been blacklisted).

D. Availability of Homographic IDNs

Our prior study investigated 1,516 *registered* homographic IDNs. In this section, we further investigate the available space of IDN registration, i.e., how many homographic IDNs are still *unregistered*. From attackers' perspective, high availability makes domain abuse easier. To assess the availability, for each brand domain (also Alexa top 1K SLDs), we replaced its characters with homoglyphs to create a set of IDNs, and computed SSIM Indices subsequently. Similarly, IDNs with a maximum SSIM Index of over 0.95 are selected as homographic domains.

The key problem we need to solve is how to find homoglyphs for a character. Here, we leveraged a list called UC-SimList [8], which was composed based on pixel overlap between bitmaps of characters. To reduce the computation overhead, only one character was replaced at a time.

In the end, we created 128,432 new IDN domains, and discovered **42,671** of them to be homographic domains of Alexa Top 1k SLDs (among which 237 are registered). Figure 7 presents the number of homographic IDNs (both registered and unregistered) associated with Alexa Top 100 SLDs under

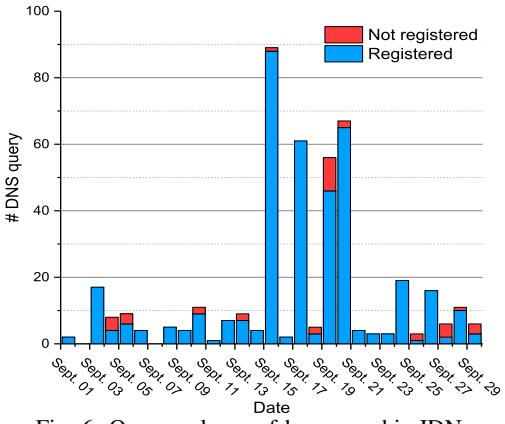


Fig. 6: Query volume of homographic IDNs

com, *net* or *org*. Clearly, attackers have lots of choices for phishing IDNs. To notice, the number of IDNs we found so far is just the lower-bound, as only one letter was replaced.

One may argue that not all homograph IDNs can be registered, as a registration undergoes name checks by registrars or registries. To assess how likely the registration succeeds, we sampled 10 homographic IDNs (e.g., *xn--eay-6xy.com* and *xn--sn-cxs.com*) and attempted to register them through GoDaddy. All our requests were approved.

Previous studies showed that by registering domains that are likely to be mistyped, attackers could harvest a huge amount of user traffic and launch attacks like name server hijacking [53]. We are interested in whether such traffic also flows to homographic IDNs. As such, we queried DNS Pai using the 42,671 homographic IDNs and counted the volume within Sept. 2017. The results are illustrated in Figure 6. Although queries to unregistered IDNs are observed, their proportion is very small. From user's perspective, mistyping a domain name with characters in another language is much rarer than normal typos.

E. Summary of Findings

- Most browsers have responded to the threat from homograph attacks. However, not all of them enforce the right policies and their implementations differ. Some browsers (e.g., Firefox) are still vulnerable even after the latest fix.
- 1,516 registered homographic IDNs are detected by our SSIM-based approach. Among them, only 4.82% were registered for brand protection. Most of the homographic IDNs are yet to deliver useful content, but malicious IDNs which escape all blacklists are discovered.
- From the perspective of adversaries, the choices of available homographic IDNs are substantial.

VII. Semantic Attack

In this section, we present our study on IDNs which impersonate brand domains based on semantic similarity. As described in Section VI, we focus on Type-1 semantic attack, which compounds a brand name with non-English keywords (named Type-1 IDNs afterwards).

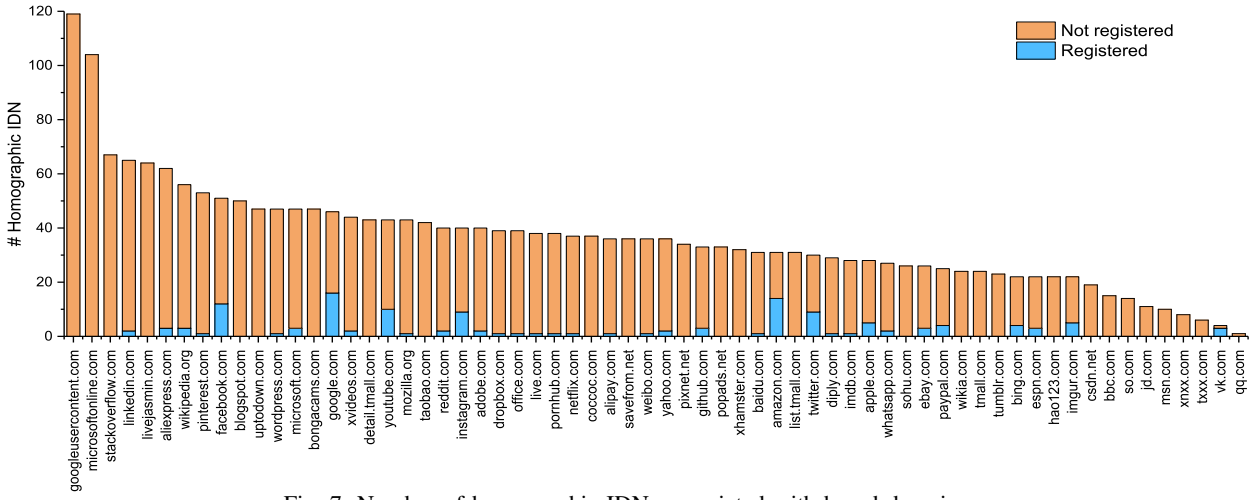


Fig. 7: Number of homographic IDNs associated with brand domains

A. Detecting Type-1 IDNs

To uncover more Type-1 IDNs, instead of manual analysis, we developed an automated approach and compared the entire 1.4 million registered IDNs with brand domains (Alexa Top 1K SLDs). In particular, we first removed the non-ASCII characters from all IDNs, and then computed SSIM Indices on the rendered domain name images. Different from previous experiments, we selected IDNs whose ASCII-only part is **identical** to a brand domain (i.e., SSIM Index equals 1.0). Our assumption is that adding non-English keywords and replacing ASCII characters with homoglyphs at the same time would make the IDN quite distinguishable, reducing their chances of fooling users.

B. Registered Type-1 IDNs

In total, **1,497** IDNs are detected under this category. All blacklisted phishing IDNs (see Table IX) are detected as well. We manually checked their semantic meanings to understand their intentions.

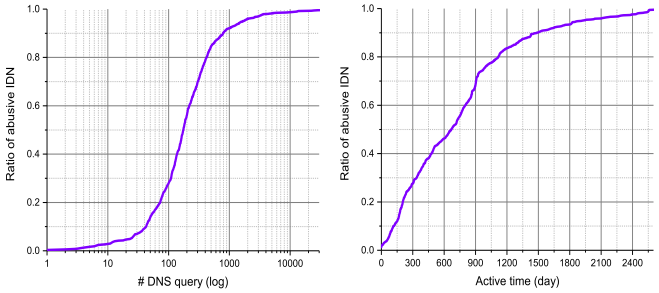
Brand domains. We found that 102 brand domains are targeted by this attack, top 10 listed in Table XIV. Particularly, 36 of the brands are mainly facing Chinese customers. A prominent reason behind these IDNs is to impersonate a brand service. For instance, we observed that every Type-1 IDN related to *58.com* (the biggest website serving classified ads in China) appends a service keyword to "58" (e.g., *58 汽车.com*, meaning 58 automobile). From WHOIS data of all 1,497 domains, we found that only 45 IDNs were registered under email accounts of brand companies, with at least 226 registered using personal email addresses.

DNS statistics. We queried Farsight Passive DNS using the Type-1 IDNs to assess their active time and query volume. The results are illustrated in Figure 8. Similar to homographic IDNs, Type-1 IDNs are frequently visited, with 735 days of active time and 1,562 queries in average.

IDN usage. Only a few Type-1 IDNs are meaningful to visitors. According to our manual analysis on a sampled set (100 websites), more than 85% are inactive, including unresolvable (55%), error (9%), parked (21%) and empty (2%). The result suggests that most Type-1 IDNs are owned by opportunistic registrants. Nevertheless, we discovered 2 domains involving malware delivery (*xn--bet365-n82p.com* and *xn--bet365-g37i416dc3e.com*, which impersonate *bet365.com*).

TABLE XIV: Top 10 brand domains ordered by Type-1 IDNs

Domain	Alexa	# Type-1 IDN	Rate	Protective Registrations
58.com	861	270	18.04%	1
qq.com	9	139	9.29%	22
go.com	391	114	7.62%	0
china.com	166	84	5.61%	0
bet365.com	332	81	5.41%	5
1688.com	191	74	4.94%	0
amazon.com	11	63	4.21%	2
sex.com	537	39	2.61%	0
google.com	1	34	2.27%	0
as.com	634	33	2.20%	0
Total		**931**	**62.2%**	**30**



(a) Active time (b) Query volume

Fig. 8: ECDF of active time and query volume of semantically abusive IDNs

C. Summary of Findings

By exploiting the semantics of brand domains, attackers can create deceptive IDNs for malicious activities like phishing. Though such attack has never been reported before, our detector has already identified 1,497 IDNs which are likely

involved. While a few of them have been used for malicious activities, most of them are still in "sleep mode".

VIII. Discussion

Recommendations. Our work has identified at least 6,241 malicious IDNs and some of them are in active use and very deceptive for users. We believe to solve the issues around IDN abuse, efforts from all parties in the Internet ecosystem are required. For registries maintaining DNS zones, checking if a domain registration request is intended for malign purposes is necessary. As an example, we found a brand protection system is deployed on three TLDs (e.g., *cn*), by performing resemblance checks on visual appearances, pronunciation and semantics [11]. For registrars selling domains, domain parking should be avoided for abusive IDNs, which could curb attackers' revenues from domain name fraud. For browsers, our analysis on off-the-shelf browsers shows not all of the browser vendors correctly implement the guidance from IETF, and we recommend them to deliver code patches promptly. We also notice that policies based on the diversity of character sets are not enough to prevent IDN abuse. IE 11 seems to address this issue adequately, which prompts an alert when the domain name contains Unicode characters. For end users, we believe education is necessary to let them understand the harm and look carefully for indicators.

IX. Related Work

IDN. Though the IDN implementation has been rolled out for more than a decade, there are only few studies covering this scheme and its security implications. In particular, the homograph attack caused by IDN received most attention from the security community. The first research of this issue was done by Tobias et al. in 2006, who analyzed a small amount of users' network traffic to find IDNs impersonating Alexa top 500 sites, and measured their popularity and intention [25]. Hannay et al. showed homograph attack was gaining traction from the cyber-criminals [20]. Chris et al. looked into different ways in which IDNs are abused, and found they were utilized for malware distribution and botnet communications [37].

We revisit this topic but our study is much more comprehensive in terms of scale, observations and attack vectors identified. By scanning zone files from major TLDs and iTLDs, we discovered over 1.4 million IDNs, which are orders of magnitude more than previous works. We measured the entire IDN ecosystem, including hosting, registration and usage. In addition to homograph attack, our study discovered new semantic attack launched through IDNs.

Domain-squatting. The attacks from IDN aim to confuse web users when recognizing domain names, in hopes of hijacking their web traffic, which can be classified as domain-squatting attacks. Previous studies have revealed different forms of such attacks, like typo-squatting [1, 29, 50]. Recent studies even show that the configuration issues and hardware errors of users' machines could be exploited by attackers to harvest domain requests, which is called bitsquatting [43, 53]. The semantic attack discovered by our research complements the existing works in this area and suggests the attack vectors under this category are not yet exhausted. Regarding the impact of domain-squatting, most of the reputable domains are targeted by this attack vector [1] but the overall negative externalities to the Internet users are still moderate [29].

DNS abuse. DNS has been abused by attackers to cover their infrastructures from a long time ago. They obtain domain names from domain registrars and link them to a broad spectrum of cyber-criminal activities, like blackhat SEO [14], malware [19] and spam [3, 34]. A great amount of effort has been devoted by the research community to detecting such malicious domains, mainly through DNS analysis, URL analysis and code analysis [4–6, 45, 47, 55]. In parallel, many studies focus on understanding attackers' operational models behind domains [18, 22, 38, 48, 49] and how to protect DNS against abuse [7, 13].

X. Conclusion

To make Internet more accessible to people whose primary languages are not English, IETF initiated the IDN standard and many registrars have opened up the registration for IDNs. Through quantitative analysis, our study shows the volume of IDNs has been steadily growing over years, and now more than 1.4 million IDNs are registered. Despite the increase in volume, their value to Internet users is far under expectation. Through stratified sampling analysis, we found only 19.8% IDNs deliver meaningful content, compared to 33.6% of ASCII domains. Moreover, visits to them are far less frequent than non-IDNs under gTLDs like *com*. What makes IDN more problematic is that new attack vectors have been enabled and abused for cyber-attacks like brand phishing. IDN is known to enable homograph attack and we discovered 1,516 IDNs resembling known brands. At least 100 of them are confirmed malicious. Still, attackers have a large candidate pool of deceptive IDNs, given that 42,671 IDNs can be used for homograph attack and most of them are unregistered. What remains less known is that, IDN can be designed to confuse users by padding keywords or translating English brand names (called semantic attack). We discovered 1,497 IDNs under the first case, and some brands (like *58.com*) are targeted by over 100 IDNs. We believe the development of IDN needs rectification and efforts should be spared by all entities in Internet, including registries, registrars and Internet software.

Acknowledgments

We thank all anonymous reviewers for their helpful suggestions to improve the paper. We also thank Fengpei Li, Jinjin Liang, Jianjun Chen, and Yiming Zhang for their valuable feedback.

This work was supported by the National Natural Science Foundation of China (grant 61772307, 61472215, U1636204), the National Key Basic Research Program (grant 2017YFB0803202) and CERNET Innovation Project NGII20160403.

Any views, opinions, findings, recommendations, or conclusions contained or expressed herein are those of the authors,

and do not necessarily reflect the position, official policies or endorsements, either expressed or implied, of the Government of China or Qihoo 360.

REFERENCES

[1] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *NDSS, 2015*.

[2] H. Alvestrand and C. Karp. Right-to-left scripts for internationalized domain names for applications (idna). Technical report, 2010.

[3] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. *Spamscatter: Characterizing internet scam hosting infrastructure*. PhD thesis, 2007.

[4] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. Building a dynamic reputation system for dns. In *USENIX security, 2010*.

[5] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting malware domains at the upper dns hierarchy. In *USENIX security, 2011*.

[6] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon. From throw-away traffic to bots: Detecting the rise of dga-based malware. In *USENIX security, 2012*.

[7] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. Dns security introduction and requirements. Technical report, 2005.

[8] L. W. G. L. AY Fu, X Deng. The methodology and an application to fight against unicode attacks. In *Proceedings of the second symposium on Usable privacy and security*, 2006.

[9] Chromium. Idn in google chrome. https://www.chromium.org/developers/design-documents/idn-in-google-chrome.

[10] CNNIC. Cnnic participates in icann idn program conference.

[11] CNNIC. Introduction of brand protection services. https://www.cnnic.net.cn/gjymaqzx/gjymaqlm/lmfw/201507/t20150706_52503.htm.

[12] A. Costello. Rfc 3492-punycode: A bootstring encoding of unicode for internationalized domain names in applications (idna). *Network Working Group, IETF*, 2003.

[13] D. Dagon, M. Antonakakis, P. Vixie, T. Jinmei, and W. Lee. Increased dns forgery resistance through 0x20-bit encoding: security via leet queries. In *CCS, 2008*.

[14] K. Du, H. Yang, Z. Li, H.-X. Duan, and K. Zhang. The ever-changing labyrinth: A large-scale analysis of wildcard dns powered blackhat seo. In *USENIX Security, 2016*.

[15] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the https certificate ecosystem. In *IMC, 2013*.

[16] P. Faltstrom, P. Hoffman, and A. Costello. Rfc 3490: Internationalizing domain names in applications (idna). *Network Working Group, IETF*, 2003.

[17] FarSight-Security. Dnsdb data. https://www.farsightsecurity.com/solutions/dnsdb.

[18] M. Felegyhazi, C. Kreibich, and V. Paxson. On the potential of proactive domain blacklisting. *LEET, 2010*.

[19] C. Grier, L. Ballard, J. Caballero, N. Chachra, C. J. Dietrich, K. Levchenko, P. Mavrommatis, D. McCoy, A. Nappa, A. Pitsillidis, et al. Manufacturing compromise: the emergence of exploit-as-a-service. In *CCS, 2012*.

[20] P. Hannay and G. Baatard. The 2011 idn homograph attack mitigation survey. In *SAM, 2012*.

[21] P. Hannay and C. Bolan. Assessment of internationalised domain name homograph attack mitigation. In *Australian Information Security Management Conference*, page 13, 2009.

[22] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck. Understanding the domain registration behavior of spammers. In *IMC, 2013*.

[23] A. Hern. Unicode trick lets hackers hide phishing urls. https://www.theguardian.com/technology/2017/apr/19/phishing-url-trick-hackers.

[24] P. Hoffman and M. Blanchet. Rfc 3491:nameprep: A stringprep profile for internationalized domain names (idn). *Network Working Group, IETF*, 2003.

[25] T. Holgers, D. E. Watson, and S. D. Gribble. Cutting through the confusion: A measurement study of homograph attacks. In *USENIX ATC, 2006*.

[26] ICANN. Centralized zone data service. https://czds.icann.org/en/.

[27] ICANN. Internationalized internationalized domain names in poland. https://www.icann.org/en/system/files/files/bartosiewicz-idn-kl-21jul04-en.pdf.

[28] ICANN. Report of the internationalized domain names working group — responses to survey c.

[29] M. T. Khan, X. Huo, Z. Li, and C. Kanich. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting. In *Security and Privacy, 2015*.

[30] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. R. Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. In *CCS, 2017*.

[31] J. C. Klensin. Internationalized domain names for applications (idna): Definitions and document framework. 2010.

[32] J. C. Klensin. Internationalized domain names in applications (idna): Protocol. 2010.

[33] J. C. Klensin. Internationalized domain names in applications (idna): Protocol. 2010.

[34] M. Konte, N. Feamster, and J. Jung. Dynamics of online scam hosting infrastructure. In *PAM, 2009*.

[35] V. Krammer. Phishing defense against idn address spoofing attacks. In *International Conference on Privacy, Security and Trust, 2006*.

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

[37] C. Larsen. Bad guys using internationalized domain names. https://www.symantec.com/connect/blogs/bad-guys-using-internationalized-domain-names-idns.

[38] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. Domain-z: 28 registrations later measuring the exploitation of residual trust in domains. In *Security and Privacy, 2016*.

[39] J. Liang, J. Jiang, H. Duan, K. Li, T. Wan, and J. Wu. When https meets cdn: A case of authentication in delegated service. In *Security and privacy, 2014*.

[40] M. Lui and T. Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*.

[41] M. Lui and T. Baldwin. Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011.

[42] Mozilla. Idn display algorithm. https://wiki.mozilla.org/IDN_Display_Algorithm#Algorithm.

[43] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen. Bitsquatting: Exploiting bit-flips for fun, or profit? In *WWW, 2013*.

[44] PIR. Zone file access for .org. https://pir.org/resources/file-zone-access/.

[45] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta. Phishnet: predictive blacklisting to detect phishing attacks. In *INFOCOM, 2010*.

[46] Qihoo. Passive dns system. http://www.passivedns.cn.

[47] B. Rahbarinia, R. Perdisci, and M. Antonakakis. Segugio: Efficient behavior-based tracking of malware-control domains in large isp networks. In *DSN, 2015*.

[48] A. Ramachandran, N. Feamster, D. Dagon, et al. Revealing botnet membership using dnsbl counter-intelligence. *SRUTI, 2006*.

[49] K. Sato, K. Ishibashi, T. Toyono, H. Hasegawa, and H. Yoshino. Extending black domain name list by using co-occurrence relation between dns queries. *IEICE transactions on communications, 2012*.

[50] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich. The long" taile" of typosquatting domain names. In *USENIX Security, 2014*.

[51] Versign. How to register internationalized domain names. https://www.verisign.com/en_US/channel-resources/domain-registry-products/idn/index.xhtml.

[52] Versign. Top-level domain zone file information. https://www.verisign.com/en_US/channel-resources/domain-registry-products/zone-file/index.xhtml.

[53] T. Vissers, T. Barron, T. Van Goethem, W. Joosen, and N. Nikiforakis. The wolf of name street: Hijacking domains through their nameservers.

[54] T. Vissers, W. Joosen, and N. Nikiforakis. Parking sensors: Analyzing and detecting parked domains. In *NDSS, 2015*.

[55] S. Yadav, A. K. K. Reddy, A. Reddy, and S. Ranjan. Detecting algorithmically generated malicious domain names. In *IMC, 2010*.

[56] H. R. S. Z. Wang, A. C. Bovik and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. In *IEEE Transactions on Image Processing, 2004*.

[57] A. Zhou Wang; Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. In *Signal Processing Magazine*. IEEE, 2009.