FARSIGHT SECURITY

Farsight Security

Global Internationalized Domain Name Homograph Report

Q2 / 2018

Mike Schiffman

Engineering Team Lead, Farsight Security, Inc.

FARSIGHT SECURITY

Table of Contents

Executive Summary

Internationalized Domain Names (IDNs) enable a multilingual Internet. Using IDN standards and protocols, Internet-users are able to register and use domain names in scripts other than Basic Latin. Yet IDNs are often abused by cybercriminals to conduct malicious activities, such as phishing or malware distribution.

In this new research report, Farsight Security set out to determine the prevalence and distribution of IDN homographs across the Internet. We examined 100M IDN resolutions over a 12-month period with a focus on over 450 top global brands across 11 sectors including finance, retail, and technology.

Almost	Over	Over	Across
100M	12	450	11
IDN resolutions	Month period	Top global brands	Sectors

Among the key findings:

- 100M total IDN resolutions observed; 27M unique fully qualified domain names (FQDNs)
- 8,000 IDN homographs representing or containing a top global brand name
- Unicode “confusables” make up a significant percentage of the characters found in IDNs; 91% of all characters observed in IDN homographs are considered “confusable” -- a “confusable” is a Unicode code point that is often easily confused with other characters, ligatures, and/or digraphs.
- Brands in banking and other related sectors are frequently imitated using IDN homographs with ~750 unique FQDNs observed per month
- 91% of IDN homographs offered some sort of webpage
- We found clear violations of the ICANN Guidelines for the Implementation of Internationalized Domain Names
- 66% of all IDN homograph IP addresses were found to be geolocated in the United States
- 93% of IDN homograph FQDNs had IPv4-based address records

Farsight is committed to making the Internet a safer place for online transactions for all users. If your organization would like your brand included in future Farsight research, please contact us at sales@farsightsecurity.com

Introduction

Just as the Domain Name System (DNS) enables the vast majority of online transactions, Internationalized Domain Names (IDNs) enable an internationalized Internet. The Internet is global by design, but before the advent of IDNs, English was the de facto language of DNS by way of the Basic Latin script. In contrast, given the IDN standards and protocols, Internet users are able to register and use domain names in almost any written language.

The perennial problem with IDNs is that they also enable cybercriminals to register lookalike domain names with ease. The intent here is typically to hoodwink end-users into thinking the bad actor's lookalike site is actually the real thing. These lookalike domain names (known as "homographs") completely sidestep any traditional security controls an organization may have in place, and are, in most cases, indefensible. By leveraging IDN homographs, cybercriminals can easily lure users into going to phishing websites that are pixel-perfect replicas of the brands they're impersonating – often completely undetected by today's defensive solutions.

The fact that some attackers are registering confusing Internet DNS names for the purpose of misleading Internet users is not news. Every user of the Internet learns – often the hard way – that much of the email they receive is forged, and many of the web links they are prompted to click on are malicious. Yet IDN allows forgeries to be nearly undetectable to either the human eye or traditional end-user Internet software such as email clients and web browsers.

Using its products, DNSDB™ and Brand Sentry™, Farsight Security conducted new research to determine the prevalence and distribution of IDN homographs across the Internet. Specifically, we examined 12 months of IDN data to look for trends, signals, and anomalies in what might otherwise seem to be noise. Furthermore, we created a filter in the form of 466 top global brand names, which was used as a sieve to strain out targeted homograph abuse.

In January 2018, Farsight Security published a blog article first announcing its work in IDN homograph research. This report expands on that initial research, both in scope and depth of analysis. We are of the opinion that fostering a continued dialogue about this work is important to the public interest. With continued examination and vigilance exhibited in responses ranging from end-user education to remedial efforts by software vendors, real steps can be taken not only to decrease the frequency of this kind of malfeasance, but also to lessen its ultimate impact.

Our findings confirm that the potential security risks posed by IDN homographs are significant. Any ultimate defense against this variant of Internet forgery will necessarily depend upon Internet governance and security automation. It is to inform the need for such solutions that we offer the findings below.

This report is organized into three sections. First, there is an introduction section containing background material on IDNs and their underpinning foundational technology. The second section contains our key findings and discussion. Finally, we close with an explanation of our methodology for IDN instrumentation.

[1] Even SSL certificate checking is "bypassed" if the phishing site includes its own certificate matching the exact name of the (fake) website.

IDNs, Unicode, Punycode: Background

Internationalized Domain Names in Applications (abbreviated as IDNA, or instances thereof referred to simply as “IDNs”) is a system to represent characters other than those found in the Basic Latin script/ASCII familiar to those who use English.

For example:
Chinese: 百度.中国
Persian: تهران.ایران
Russian: яндекс.рф

This system was implemented to bridge the digital divide between English-speaking and non-English speaking users of the Internet. It enables the registration of domain names utilizing character sets of users’ native languages.

Unicode

Unicode is the standard for digital representation of the characters used in writing all of the world’s languages. It provides a unique number for every character. This numerical value representing a Unicode character (i.e.: U+03B1) is called a code point. The latest version of the Unicode standard contains 136,755 characters covering 139 modern and historic scripts.
Shown below are some Unicode characters and their respective code points (note that for ASCII values, the code points are identical to their common ASCII counterparts’ byte values).

F: U+0046
A: U+0041
R: U+0052
S: U+0053
I: U+0049
G: U+0047
H: U+0048
T: U+0054
♥: U+2665
✪: U+272A

IDNs are bound in form and structure to Unicode. An important distinction to note is that Unicode itself is technically not an actual encoding format; it is just a massive lookup table. Determining how Unicode characters are actually encoded into sequences of bytes is handled by mechanisms like the Unicode Transformation Format (UTF).

Punycode

DNS names are most commonly clamped to the case insensitive Letters, Digits, Hyphen (LDH) namespace which, in and of itself, is completely unsuitable for representing anything other than ASCII names. It certainly cannot directly represent the often-multi-byte Unicode character encodings, at least not without assistance. Rather than attempting to expand the DNS repertoire, core Internet engineers decided to employ an ASCII Compatible Encoding (ACE) scheme to encode the Unicode data. This ACE is Punycode, a lossless method for converting Unicode into LDH ASCII. Punycode encoding is performed on a per-label basis and the "xn--" prefix heralds the start of a so-encoded label. When a DNS label is Punycode encoded, it is known as an "a-label" while its Unicode equivalent is known as a "u-label".

Example Unicode (u-label) to Punycode (a-label) conversion:
αβγδεζηθικλμνξοπρστυφχψω --> xn--mxacdefghijklmnopqr0btuvwxy

IDNs ultimately represent Unicode labels and may appear as such to the end user, but over the wire, they are sent encoded using ASCII-conformant Punycode.

IDN Homographs

It's no secret that different letters or characters can look very much alike. Sometimes this comes about with changes in case or font when rendering text in the same language or script. Perhaps best known is the resemblance in many fonts of "Latin Small Letter l" (U+006c) to "Latin Capital Letter I" (U+0049) or the visual similarity between "Latin Capital Letter O" (U+004f) and "Digit 0" (U+0030) (which gave way to the "slashed zero"). The slashed zero is an instance of homoglyphic (single character) confusion that was resolved earlier than the invention of the printing press. Characters from different alphabets or scripts may also appear indistinguishable from one another to the human eye. Individually, these "confusables" are known as homoglyphs, but in the context of the words that contain them, they constitute homographs. In this document, we refer to them as "homographs," a less popular but more accurate term for our subject of interest.

For example, consider the following domain names:
xn--frsightsecurity-ulm.com --> fаrsightsecurity.com
This is "Cyrillic Small Letter A" (U+0430)

xn--farsghtsecurity-xng.com --> farsɩghtsecurity.com
This is a "Latin Small Letter Iota" (U+0269)

xn--farsghtsecurity-blc.com --> farsıghtsecurity.com
This is "Latin Small Letter Dotless I" (U+0131)

Once processed as Unicode, these multi-block labels look "normal" to the casual observer. While all of the above examples of homographic domains are benign, many others are out there, and they may not be.

[2] Note that, as per section 3 of RFC 1034 (https://tools.ietf.org/html/rfc1034): "the DNS specification attempts to be as general as possible in the rules for constructing domain names". As such, technically, they can consist of octets of any value. However, it is generally "accepted" (and in many places, enforced), that only the LDH syntax is allowed.

Our Approach

The research presented in this report is sourced from our corpus of data-at-rest, DNSDB™. From DNSDB™, we extracted 12 months of IDN data from May 01, 2017 to April 30, 2018 and ran this dataset and a list of top global brands through Brand Sentry™ [3] to obtain a corpus of IDN homographs. These were then post-processed with a variety of bespoke tools to learn additional information including client-facing features of identified web sites.

Brands to Watch

Of key importance to this report is the capacity to monitor broadly and detect with accuracy homograph domains masquerading as popular global organizations. Any such project can only stem from the construction of a list of objectively globally popular "brands". To this end, we created just such a list containing top global brands as defined by respected ranking organizations. Our brand list contains the following "sectors":

- **Banking**
- **Credit and Loans**
- **Insurance**
- **Financial Management**
- **Ecommerce**
- **Clothing Retailers**
- **Jewelry Retailers**
- **Luxury Retailers**
- **Cryptocurrency Exchanges**
- **Computer Security Firms**
- **Select brands from the Alexa Top 50**

Farsight Security compiled this list using data from market intelligence firm "SimilarWeb", web traffic rankings provided by "Alexa Internet", and metrics of cryptocurrency popularity generated by "CoinMarketCap". From each category, the top 50 organizations were chosen. After deduplication and the removal of unwanted sites (mostly "adult-oriented sexual content"-related), the brand list stood at 466 entries.

[3] Farsight Security's Brand Sentry™ is a purpose-built service designed to monitor our Passive DNS channels in real-time for lookalike domains, and issue alerts when it detects something suspicious. For more information, see: https://www.farsightsecurity.com/solutions/incident-response-team/brand-sentry/ value. However, it is generally "accepted" (and in many places, enforced), that only the LDH syntax is allowed.

[4] SimilarWeb, a market intelligence firm: https://www.similarweb.com

[5] Alexa Internet, a web traffic and analytics firm: https://www.alexa.com

[6] CoinMarketCap, a cryptocurrency tracking site: https://www.coinmarketcap.com

[7] Brand list brands were deduplicated as per the following algorithm:

1. If a brand occurred in the Alexa Top 100 and another sector, it was removed from the Alexa sector.
2. If a brand occurred in multiple sectors, it was kept in the first seen sector (sorting alphabetically) and removed from all others.

[8] Farsight Passive DNS collects DNS response data received from caching, recursive DNS servers distributed around the global Internet. This data is aggregated and made available via the Farsight SIE platform after it is imported in an anonymized form into the Farsight DNSDB system. Passive DNS uses observed cache miss traffic collected from these recursive resolvers to build a database detailing relationships between domain names, IP addresses, and name servers.

Key Findings

Below are highlights from our full analysis. Each subsection reports relevant findings among the IDN data corpus as a whole and, where appropriate, exclusively considers matches against the brand list discussed above.

Total Observed IDNs

During the 12-month period, Farsight Security observed nearly 100 million (99,432,594) IDN-based DNS cache misses. The following chart shows the per-month breakdown.

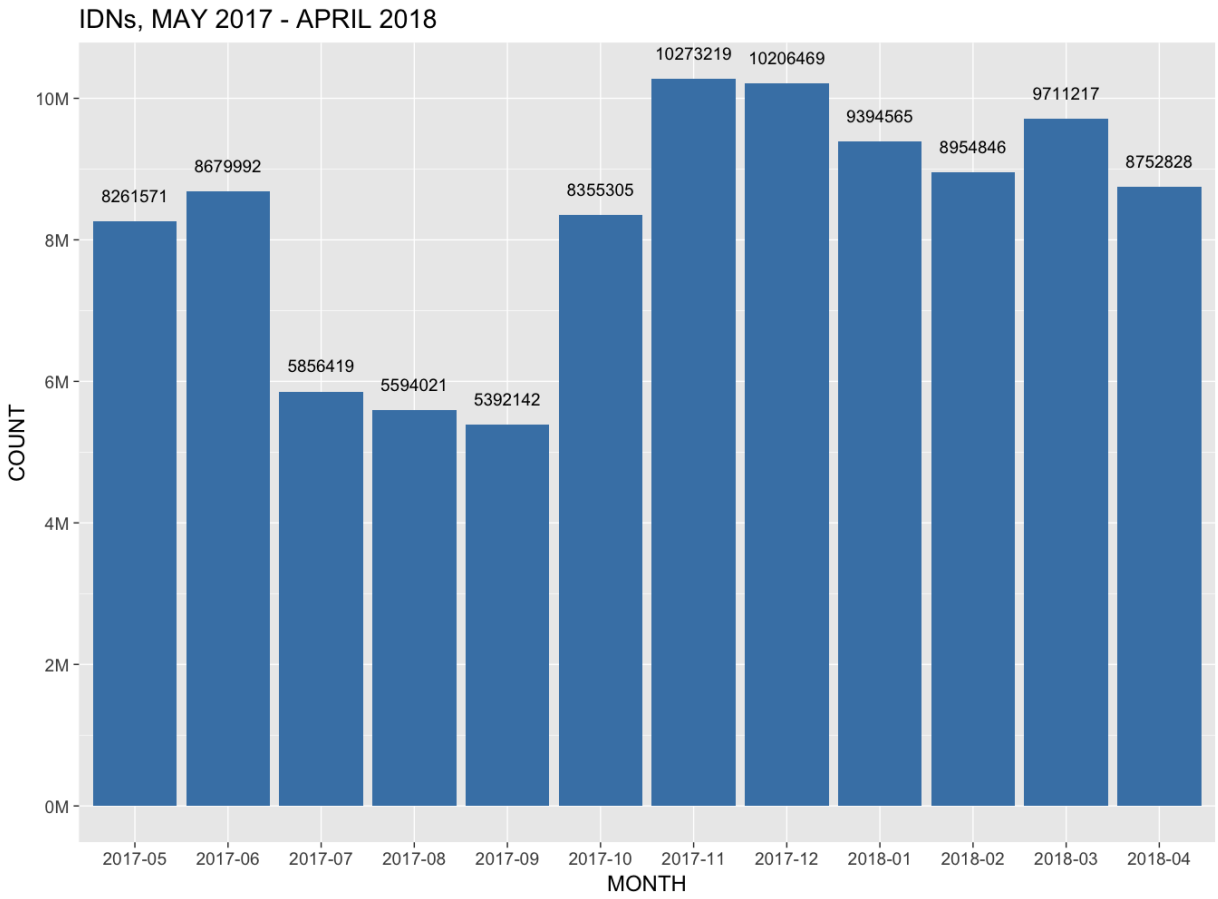

Figure 1 **IDNs Observed Per Month**

There is a noticeable dip during the summer months[9]. It is possible this lull in activity is related to "people doing less Internet" as they take advantage of the warmer weather. We will see this trend repeat elsewhere.

[9] (Summer, of course, for those living in the Northern Hemisphere.)

On FQDN Deduplication

Note that the above observations are deduplicated with respect to each fully qualified domain name (FQDN) on a per-month basis. Multiple observations in the same month for the same FQDN are collapsed into a single entry for all record types. For example, if in May 2017 we observed six queries for the A record for "xn--xample-23a.com" and two queries for its MX record, this would be recorded as a single observation of "xn--xample-23a.com" for May 2017. Note that this deduplication does not hold across successive months. Observations for "xn--xample-23a.com" in May 2017 are not deduplicated with respect to identical observations in June 2017, and so on.

When we aggregate observations across the entire 12-month time period and deduplicate, the net result is 26,734,125 unique internationalized FQDNs. This is the number is a reference point we will most often cite when referring back to our "original" corpus.

IDN TLD Distribution

We found 797 unique top-level domains represented among of 26.7 million IDNs[10]. The following chart shows the top 20 IDN-containing top-level domains along with their aggregate counts. Two noteworthy things are readily apparent:

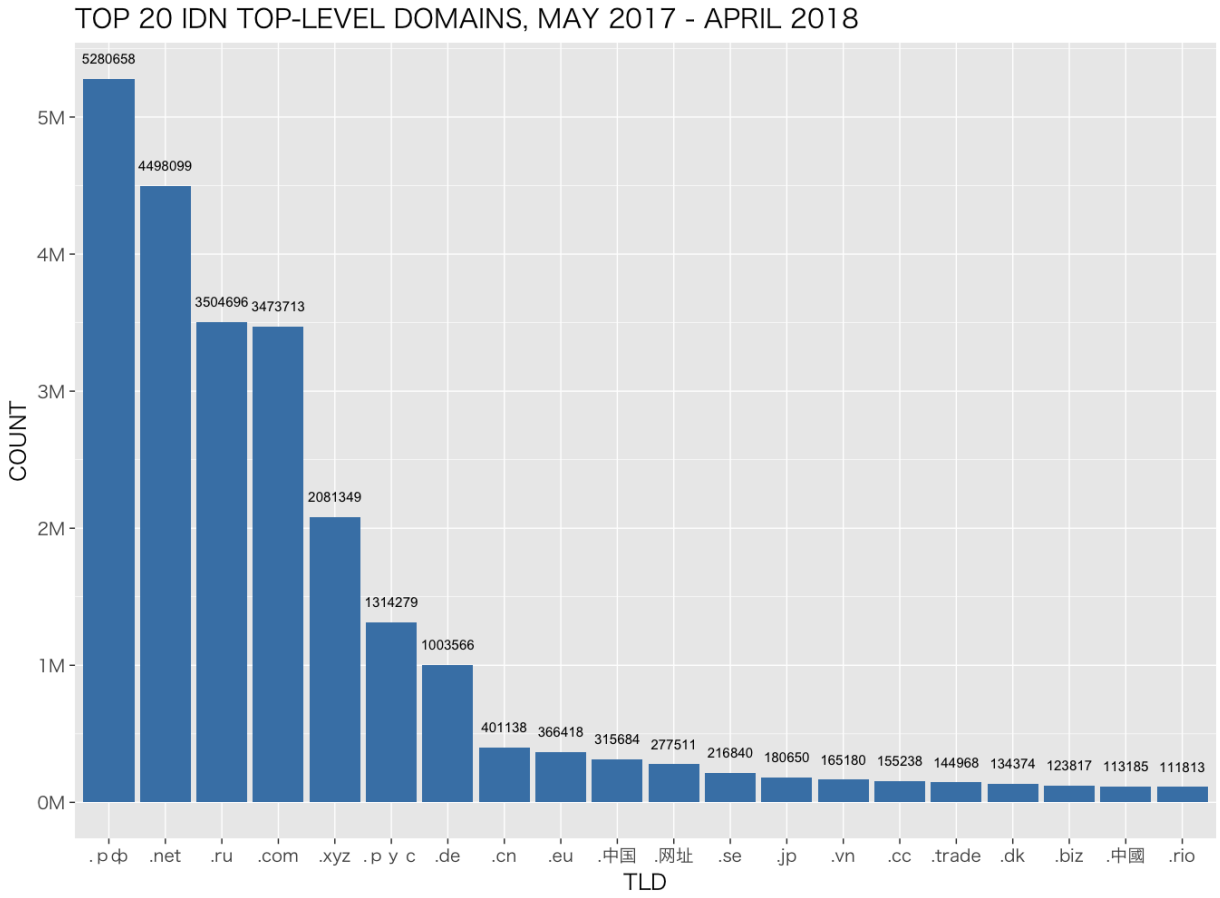

Figure 2 **Top 20 IDN TLD Distribution (May 2017 - April 2018)**

[10] Important to note that while many IDNs do have internationalized suffixes (for example ".pф") it is perfectly legal for an IDN can have a non-internationalized suffix such as ".com" or ".biz".

first, .com is a mere fourth on the list, even losing out to .net. Second, and more interesting is the "market dominance" of Russian-language based TLDs. The most common domain suffix among observed IDNs (.рф) is the Cyrillic country code for the Russian Federation, while that of third rank (.ru) is the ccTLD Latin equivalent for the Russian Federation, and sixth (.рус) is a Cyrillic gTLD that transliterates to "rus" (as in "Russian"). With the additional consideration of six other Russian-language TLDs[11], all told, 10,130,898 IDNs (38% of all observed IDNs for that time period) were registered against a Russian-language TLD. Digging deeper, Farsight Security performed a similar measurement of TLD distribution for all IDNs in DNSDB from June 2010 to December 2017 (in this seven and-a-half year period, a total of 43,400,311 unique IDNs were observed). Again, the top 20 are shown in the chart below.

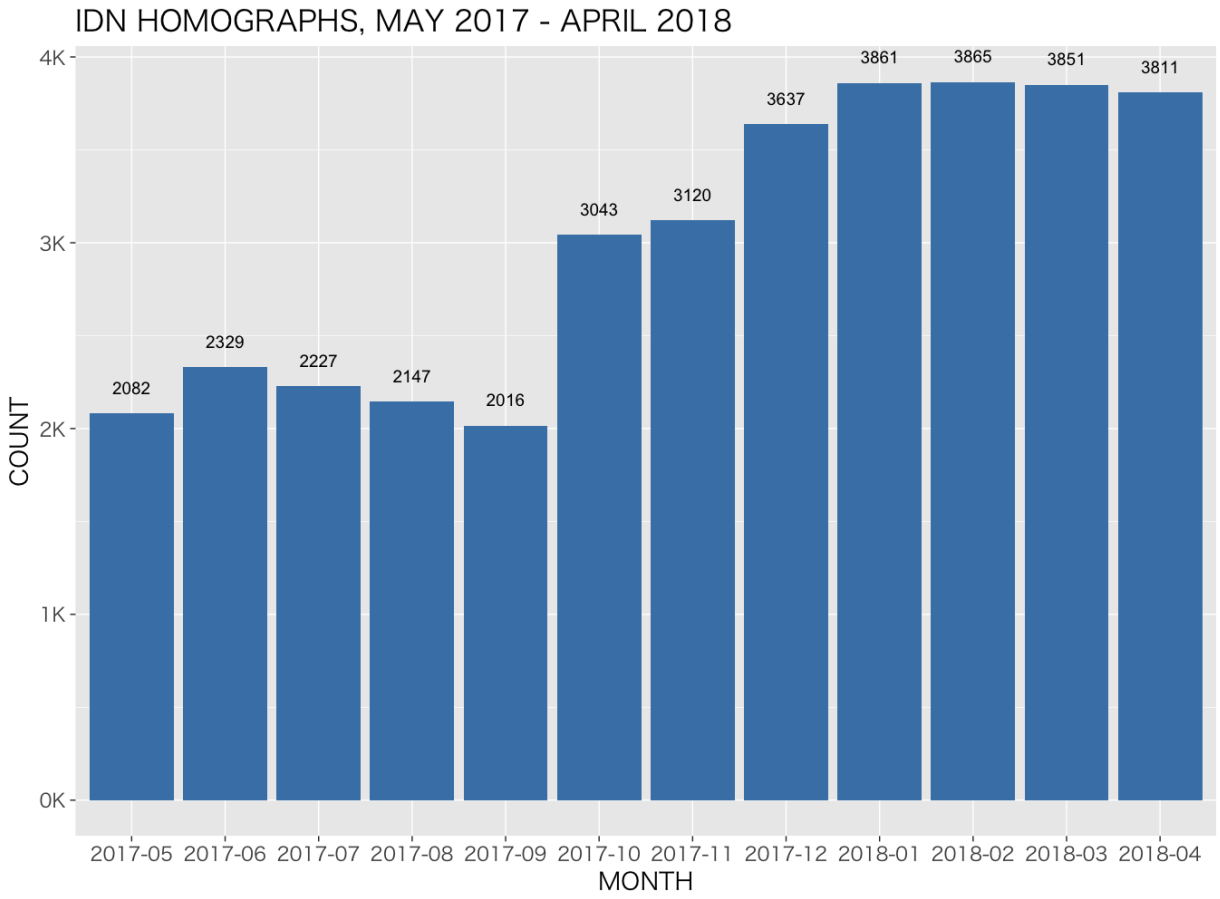

Figure 3 **Top 20 IDN TLD Distribution (June 2010 - December 2017)**

An interesting trend emerges: from June 2010 to December 2017, Russian-language TLDs accounted for only 20% of the World's IDN TLDs. However, from May 2017 to April 2018, the relative frequency nearly doubled to 38%. This is shown in waffle charts below.

[11] The other six Russian-language TLDs (sourced from https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains#Cyrillic_script) occupied spots further down in the TLD list. They are listed below:

- .дети ("children")
- .католик ("catholic")
- .ком ("com")
- .онлайн ("online")
- .орг ("org")
- .сайт ("site")

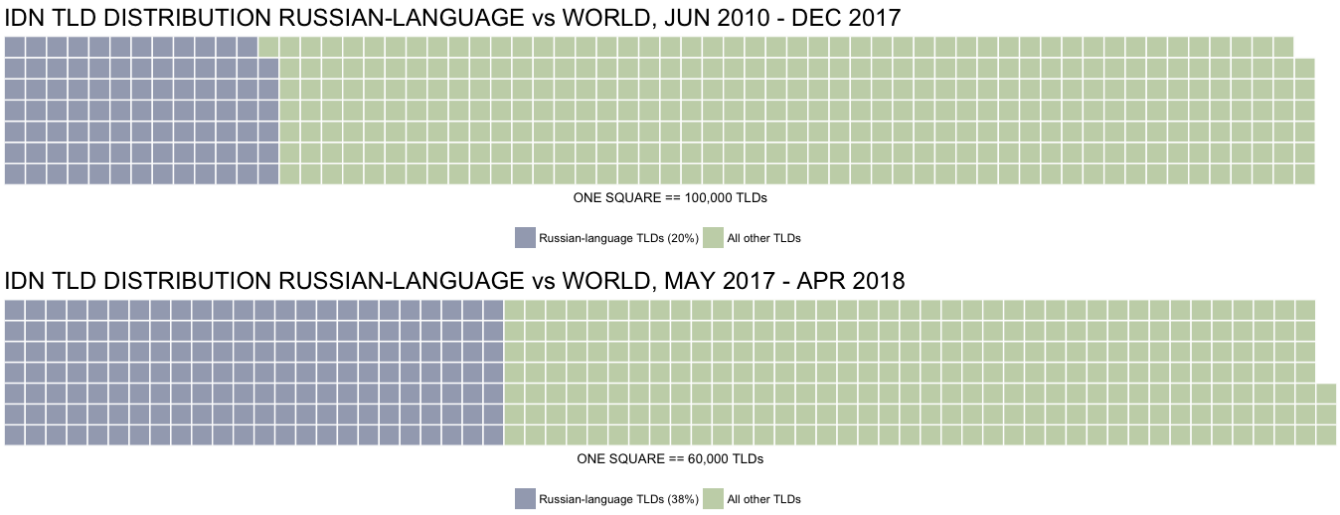

Figure 4 **Russian-Language IDN TLD Distribution**

Based on the above observations, there is a demonstrable surge in IDN use by Russian-language users. Farsight Security did not look further into this trend.

Homograph Prevalence

When we ran the corpus of 26.7 million IDNs through Brand Sentry™, we counted 35,989 homographic IDN domains that resembled one of the brands from our list as being the subject of answered DNS queries.
The following chart shows that per-month breakdown.

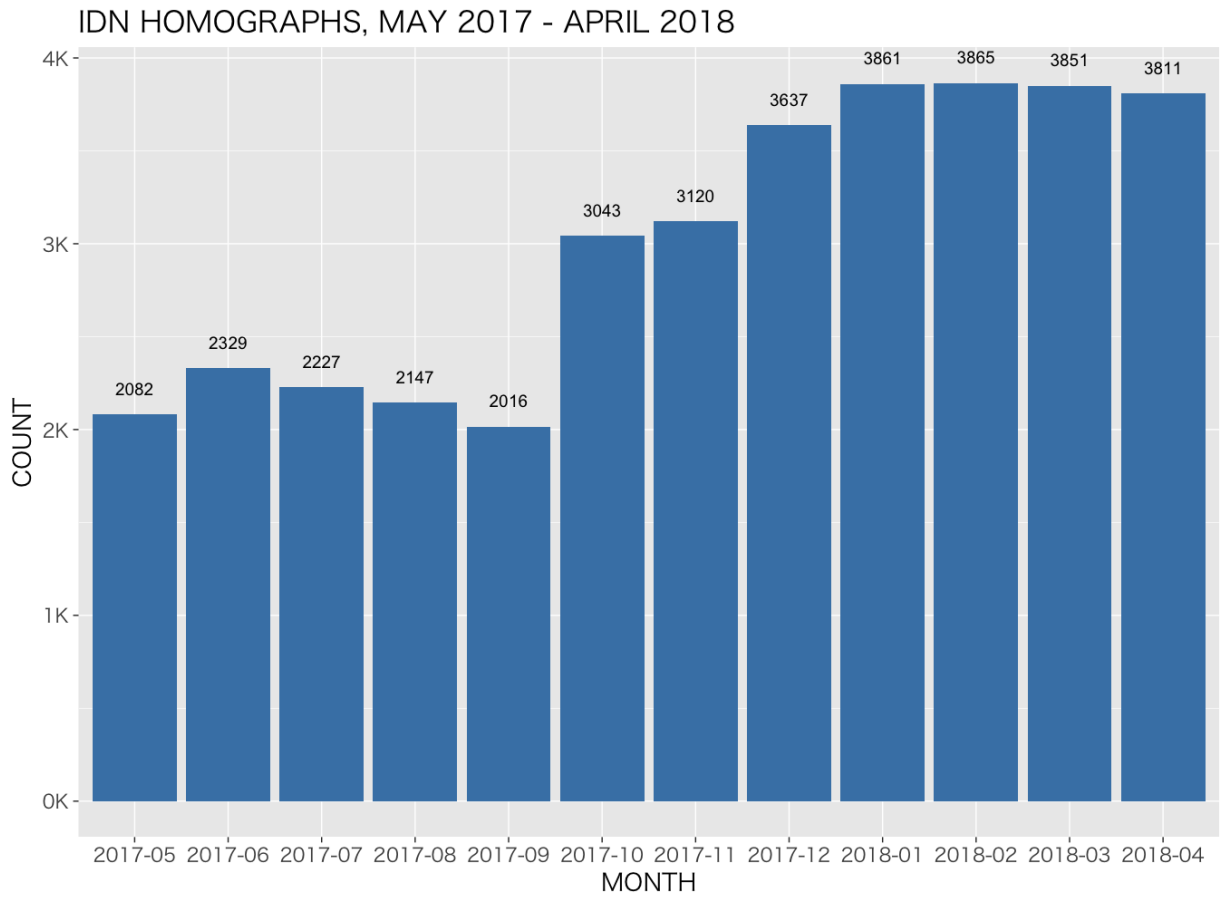

Figure 5 **IDN Homographs**

Here again we notice a summer traffic dip with a ramp-up later in the year.

As above, when we pool observations from the entire 12-month time period and deduplicate them, we come away with 8,021 unique IDN homographs.

IDN Homographs By TLD

Looking at the just the top-level domains of the IDN homographs, we see a majority originate from the .com TLD, with a total of 4,339 observations or 54% of the total IDN homograph space. This is what we expect; most top global brands are registered in the .com-space so anything attempting to mimic one of these brands would camouflage itself best by also residing in the same TLD. Even the secondly most heavily trafficked IDN homograph TLD, .ru, is only seen 1,397 times – or just under 17% of the total space.

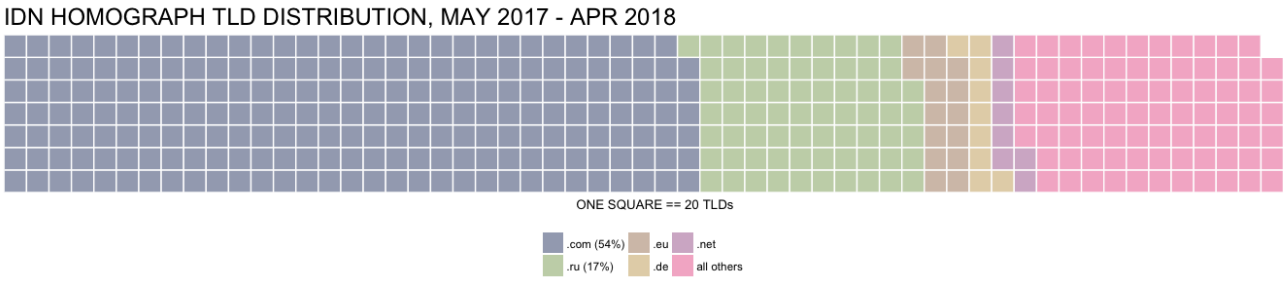

Figure 6 **IDN Homograph TLD Distribution**

Unicode Confusables

The Unicode Consortium[12] maintains a freely available list of so-called "confusables"[13]. This is a list of Unicode code points that are often easily confused with other characters, ligatures, and/or digraphs. Some Basic Latin examples are shown in the table below (displayed using a fixed width font that, with careful inspection, should more clearly reveal slight differences):

[12] The Unicode Consortium is a non-profit corporation devoted to developing, maintaining, and promoting software internationalization standards and data, particularly the Unicode Standard, which specifies the representation of text in all modern software products and standards. Reference: http://unicode.org/consortium/consort.html

[13] "Version 11.0.0 of this file, the one we consulted, is available here: http://www.unicode.org/Public/security/11.0.0/confusables.txt"

Confusable	Confused With
CYRILLIC SMALL LETTER A а	LATIN SMALL LETTER A a
CYRILLIC SMALL LETTER O о	LATIN SMALL LETTER O o
CYRILLIC SMALL LETTER ER р	LATIN SMALL LETTER P p
CYRILLIC SMALL LETTER IE е	LATIN SMALL LETTER E e
LATIN SMALL LETTER DOTLESS I ı	LATIN SMALL LETTER I i
LATIN SMALL LIGATURE OE œ	LATIN SMALL LETTER O/LATIN SMALL LETTER E oe

Table n **Sample Unicode Confusables**

Unicode confusables make up a significant percentage of all characters found in IDNs. Indeed, of the 6,294 entries in the current version of Unicode Confusables, file 1,474 code points, or 24%, are found to “look like” a Basic Latin character or a ligature containing one.

As such, it is important to investigate the distribution and frequency of confusables in our data set.

Confusables Frequency in IDN Homographs

As a baseline, we count the total number of code points observed in all effective second-level domains of the u-label corpus containing 8,021 IDN homographs. All other labels, as well as label separators are considered chaff and ignored[14] as show in Figure 7 below.

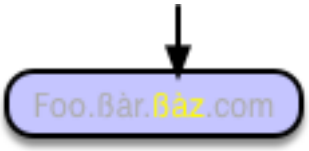

Figure 7 **Label Selection for Code Point Analysis**

[14] We’re only interested in the second level domain because this is the portion that is formally registered, the level of the DNS naming hierarchy at which most brand names are organized independently, and the most plausible entry point for successful homograph abuse.

After discarding additional labels and all suffixes, we are left with 60,080 total code points. We will use this number as a dividend against the number of observed confusables. Further counting reveals 54,687 code points that are confusables, resulting in a finding that 91% of the total homograph code points are considered "confusing"[15]. This is shown in the figure below.

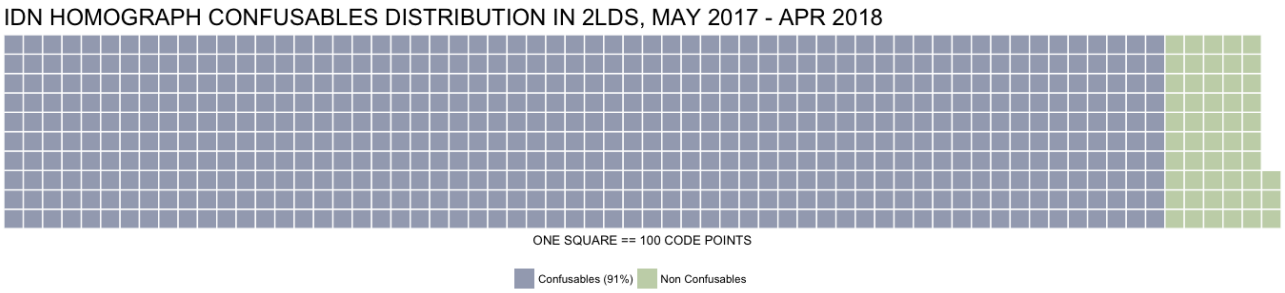

Figure 8 **IDN Homograph Confusables**

IDN Homograph Intra-label Mixed Script Prevalence

According to version 3.0 of the ICANN's Guidelines for the Implementation of Internationalized Domain Names, code points in a single label should belong to the same script[16]. This rule was intended to be enforced at the registry/registrar-level and ostensibly serves as a frontline defense against the registration of homographs.

In our analysis, Farsight Security observed clear violations of this rule. Using the same second-level domain homograph corpus from the previous section[17] we checked for code point script homogeneity using the Unicode "Scripts[18]" file as the single canonical source of determining validity. After removing "Common", "Inherited", "Unknown", and "Unassigned" properties[19], we were left with 157 labels that contained mixed scripts (this represents ~1% of the IDN homograph corpus).

Every instance of intra-label script commingling contained at least one Latin character with Greek and Cyrillic being popular mix-in choices for the violators[20]. This is what we'd expect to see for impersonators of common "Western" brands (Greek and Cyrillic contain several characters easily confused with common English letters). This is shown in the figure below.

[15] Either a "source" or "target" confusable.

[16] While there are corner-case exceptions to this rule as per the current ICANN IDN Guidelines: https://www.icann.org/resources/pages/idn-guidelines-2011-09-02-en, none of the scripts in the IDN homograph corpus Farsight Security analyzed met the criteria to permit mixing.

[17] Since domain owners and/or name server operators are free to create subdomains and hostnames with labels using mixed scripts, we are only interested in script commingling at the second-level domain level.

[18] Version 11.0.0 of this file, the one we consulted, is available here: http://www.unicode.org/Public/11.0.0/ucd/Scripts.txt

[19] Script Properties are enumerated values, mainly used to identify primary script associations. More information is here: http://unicode.org/reports/tr24/#Script

[20] The one surprise here was the discovery of a homograph using code points from Vai, an obscure Liberian script.

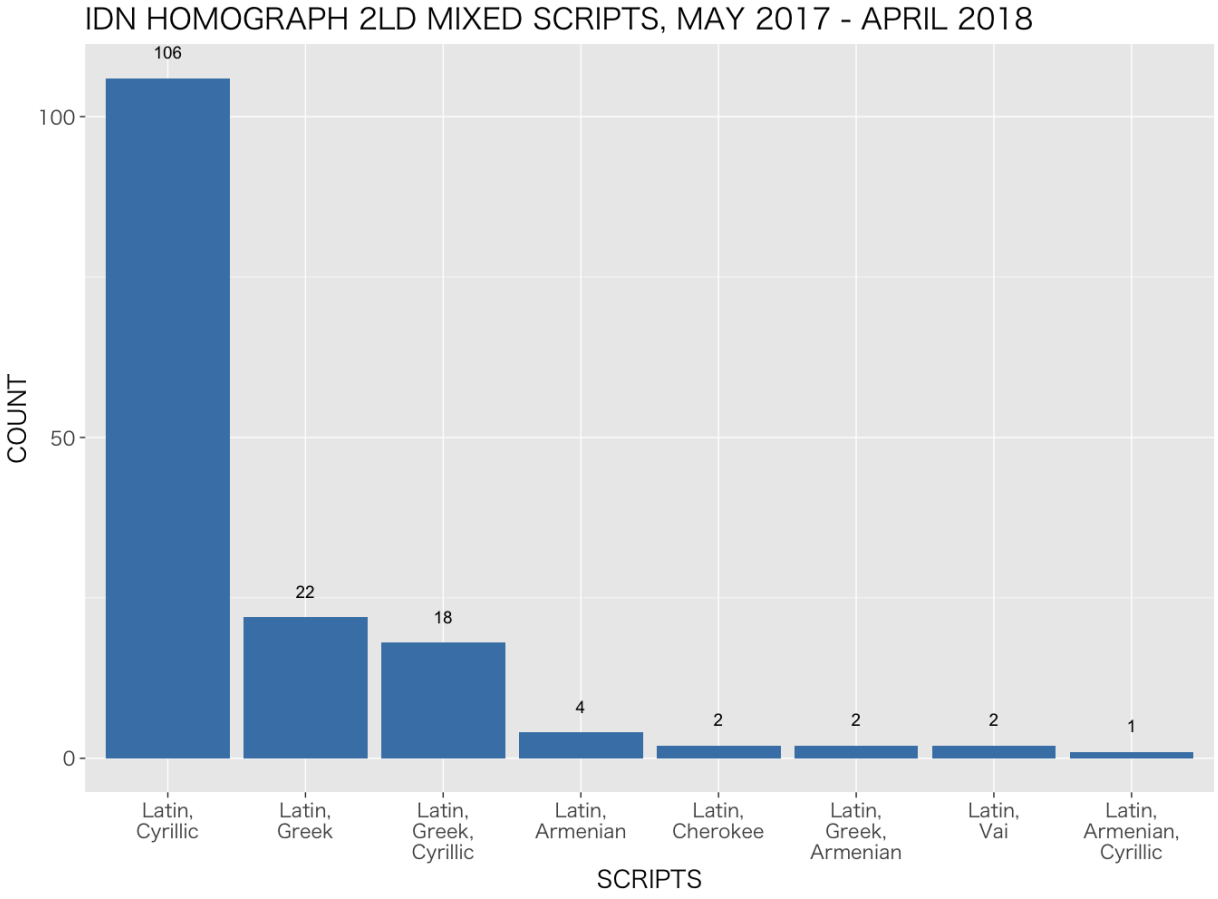

Figure 9 **IDN Homograph Intra-label Mixed Script Breakdown**

Intra-label mixed script homographs are especially troubling for two main reasons:

1. Their possibility allows for much more sophisticated and difficult-to-spot homographs than any single script does.
2. They greatly increase the potential homographic namespace for a brand, making defensive registrations much harder.

The small number of such domains seen limits our concern, but it is still troubling that there are registrars out there allowing such flagrant rule violations.

Farsight Security attempted to perform additional reconnaissance on a handful of these IDNs by looking them up at the registry level. Ironically, we were unable to obtain any results, presumably because these IDNs are technically malformed and failed input validation. It seems that once one of these domains gets registered, it comes with some implicit if unexpected built-in protections.

IDN Homograph IP Geolocation

For an idea of where IDN homographs are hosted, we queried DNSDB™ looking for relevant A and AAAA records[21] for these domains. In total, 43,136 IP addresses were returned, and each one was queried in the MaxMind IP Geolocation database[22]. We find that 66% of all IDN homograph IP addresses were found to be geolocated somewhere in the United States. Farsight Security did not investigate this pattern deeply, but one possible reason for this locational skew could be cheap US-based hosting providers or CDN content distribution with a bias towards the US.

The geographic distributions are shown in the figure below.

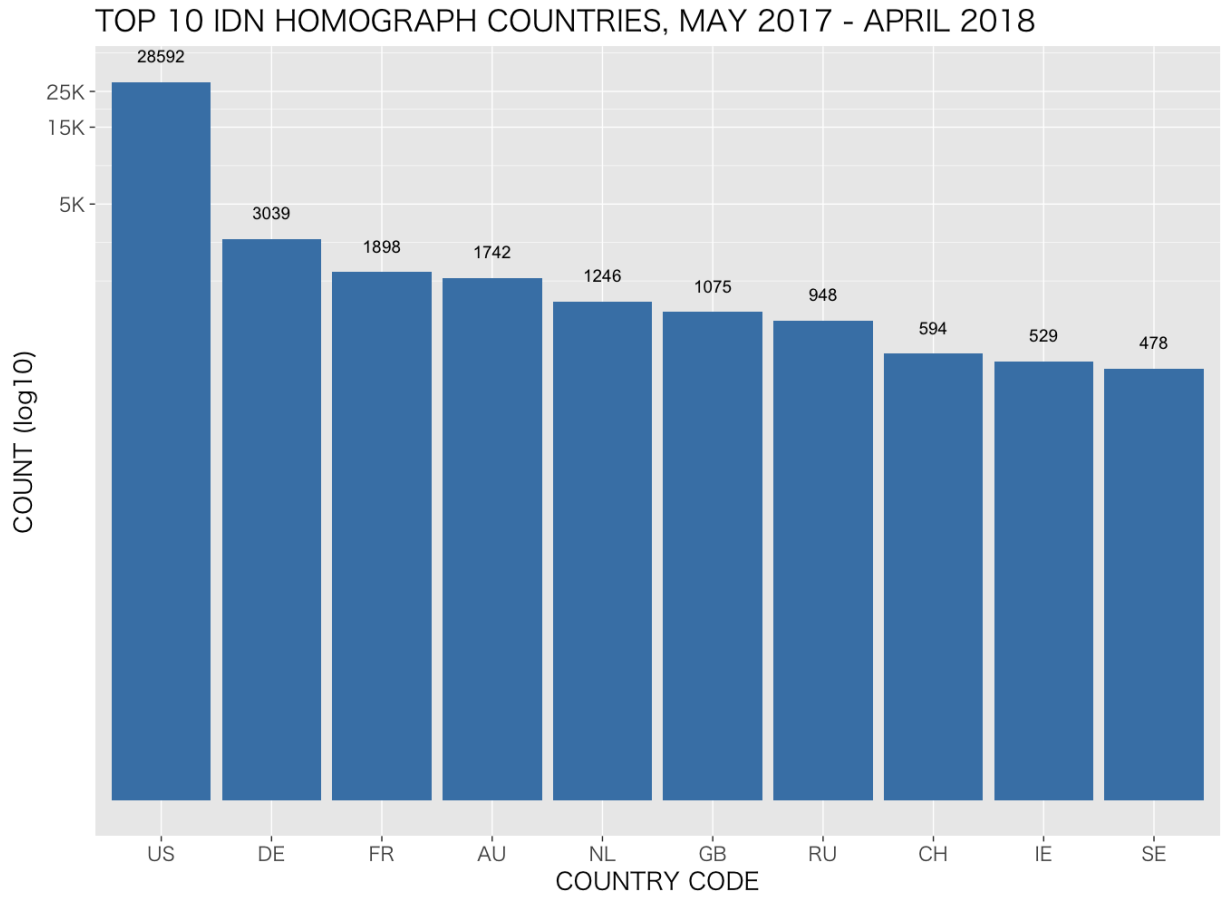

Figure 10 **IDN Homograph IP Geolocation by Country**

Looking at IP protocol version distribution for IDN homographs, we find they show an overwhelming IPv4 association. In fact, 93 of IDN homograph FQDNs had IPv4-based address records as shown in the chart below. While Farsight Security did not scrutinize this trend, it is entirely possible this percentage reflects an assumption that legacy systems and software are both more vulnerable to homograph attacks, and less likely to take advantage of IPv6 connectivity.

[21] Sometimes, some DNS names lack associated IP address records (instead providing only record types like CNAME, TXT, etc). Other times, there will be many address records associated with a given DNS name.

[22] The free version of MaxMind was used, available here: http://www.maxmind.com

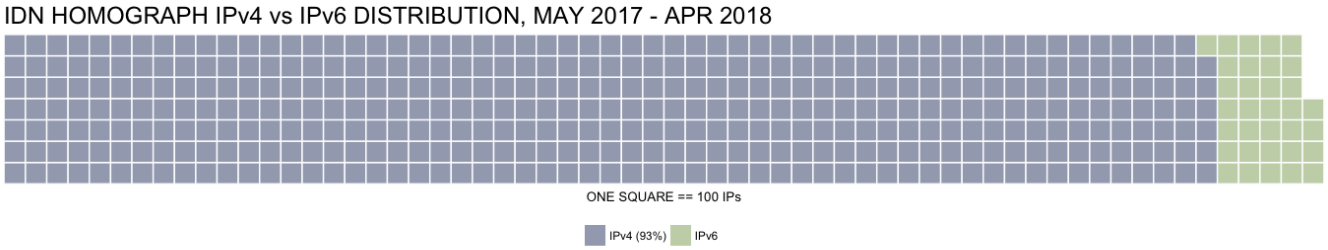

Figure 11 **IDN Homographs IPv4 / IPv6 Usage**

IDN Homograph Query Frequency by Sector

Next, Farsight Security looked at monthly sector-specific query rates for IDN homographs. Delineating a monthly snapshot of each sector type helps more finely illustrate the distribution of potential attacks (graphed with logarithmic y-axes because of the disproportionately higher response rate of IDN homographs in the Alexa Top 100 sector).

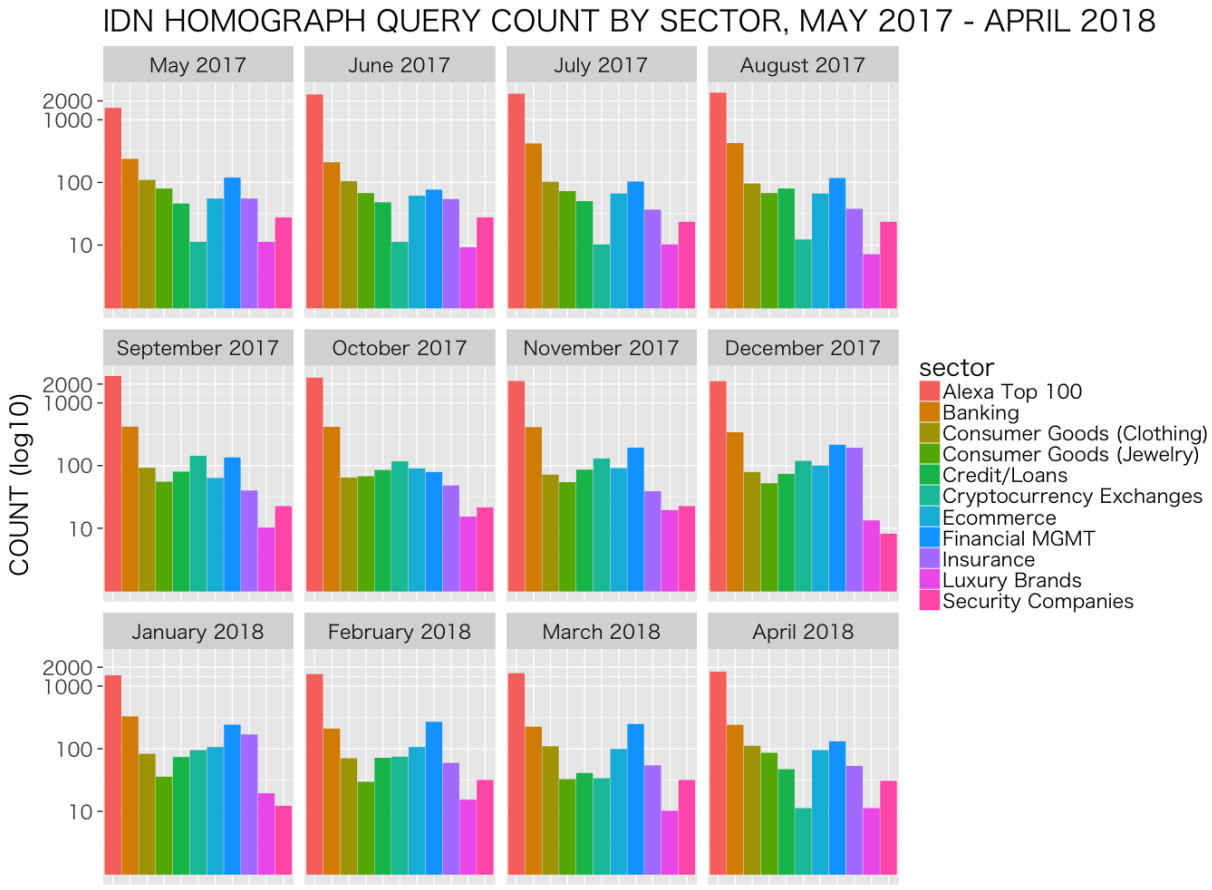

Figure 12 **IDN Homograph Query Frequency by Sector/Month**

The group of homographs encroaching upon the Alexa Top 100 unsurprisingly have the highest query count because they imitate some of the most popular sites on the Internet, representing brands such as Google, YouTube, Facebook, and so on. Naturally, it would seem that high traffic sites beget high traffic imposters. Perhaps attractive for serious phishing campaigns, we did notice many of these were parked, or even "cash parked[23]."

[23] As per: https://www.godaddy.com/domains/cashparking

We can see that brands in banking and other "money" related sectors are the most frequently imitated. All told, Farsight Security observed approximately 750 of these per month. This is no surprise -- phishers, spammers, squatters, and other questionable parties often target brands with maximum potential financial payoff. Businesses in these sectors provide a rich attack surface for many different types of Internet-based malfeasance.

IDN Homograph Websites

On June 1, 2018, we fed each of the 8,021 IDN homographs into the IDN Checker to find out which were listening on TCP ports 80 and/or 443. Our primary goal was to determine if any websites had been stood up, what their purposes might be[24], and to gather any other interesting ancillary details (such as SSL certificate information).
We found:

- **Total sites listening on TCP/80 (the normal web port): 7,332 (91%)**
- **Total sites listening on TCP/443 (the normal secure web port): 6,262 (78%)**
- **Total sites listening on both: 6,146 (76%)**
- **Total sites presenting SSL/TLS certificates: 1,608 (20%)**
- **Total sites presenting expired SSL/TLS certificates: 216 (2%)**

Next, we'll look a bit more closely at the certificate-related statistics.

IDN Homograph Website Certificate Breakdown

Farsight Security found 1,608 sites presenting SSL certificates to web browsers. The breakdown is shown the chart below.

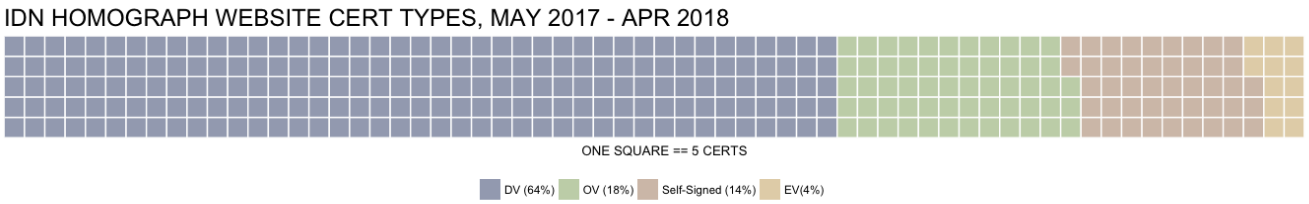

Figure 13 **IDN Homograph Website Certificate Types**

Certificate Authorities (CAs) work hard to ensure that they only issue certificates to the party that actually controls a given domain. Ownership/control of a domain is often demonstrated by having the requestor reply to a challenge email, or by having the requester create a specified web page or specified domain name[25]. With 1,026 (63%) instances, the most popular type of certificate observed was "Domain Validated" (DV). This is considered the lowest level of validation provided by a commercial Certificate Authority (CA) and is normally the cheapest option. These types of certificates have been found on phishing websites[26].

[24] A future version of this paper will include discussion of Farsight Security's automated IDN homograph image analysis and classification.

[25] For example, see https://support.comodo.com/index.php?/Knowledgebase/Article/View/791/0/alternative-methods-of-domain-control-validation-dcv

[26] Indeed in the original blog article "Touched by an IDN" (https://www.farsightsecurity.com/2018/01/17/mschiffm-touched_by_an_idn/) Farsight found a live phishing site presenting a DV certificate.

With 293 occurrences (18%), next most commonly observed was the "Organization Validated" (OV) type of certificate. Certificates at this level have a much more rigorous set of criteria that must be fulfilled before a cert may be awarded (organization authentication, locality presence, telephone verification, etc). [27] This is a stumbling block for most phishers so most of the websites presenting these certificates likely exist as legitimate defensive registrations.

Next, with 226 occurrences (14%) was self-signed certificates. These certificates are, by definition, not signed by a trusted third party, but most typically by the website's operator[28]. They offer nothing in the way of trust and are traditionally associated with man-in-the-middle attack scenarios.

Finally, Farsight Security did observe 63 (3%) "Extended Validation" (EV) certificates. These certificates offer the highest level of trust and require the most extensive vetting of applicants. None of the IDN homograph websites presenting EV certificates appeared fraudulent but instead appeared to be defensive registrations by the respective brand owners. In fact, almost all of these sites belonged to a single financial services brand that obtained 60 different defensive IDN homograph registrations. Each one of these permutations swapped out one or more different Basic Latin characters for one or more various Extended and Supplemental Latin character homoglyphs. Each of these IDN websites presented the same confidence-inspiring EV certificate.

[27] For example, see https://support.comodo.com/index.php?/Knowledgebase/Article/View/253/0/what-is-required-for-validation

[28] DNS-based Authentication of Named Entities (DANE) can be used here as an alternative to the traditional CA-based trust model.

How Can I Protect Myself?

As with many threats targeting users on the Internet, there is no silver bullet to help protect yourself. Vigilance is key, and all the rules for spotting traditional phishing sites still apply to IDN phishing sites as well.

Some web browsers support add-ons or extensions geared toward flagging or outright blocking IDNs. If the majority of your web browsing keeps you within the realm of traditional LDH ASCII domain names, this may be an acceptable security mechanism for you.

The majority of phishing attempts still reach users via email. Regardless of the apparent sender, be extremely suspicious of any emails that include:

Distressing or enticing statements to provoke an immediate reaction, or statements that threaten consequences if you fail to respond.

Account login links, especially when combined with requests or demands to update or confirm your information.

Of course, many legitimate emails contain links to additional information. Instead of clicking on these links, try copying and pasting them into your browser. This can limit the exposure to embedded links with malicious URLs.

When using a web browser, enable phishing filters or the safe browsing feature if available, and keep an eye on the browser address bar:

Any site that requests you enter a password nowadays should utilize encryption. This means the URL should begin with “https://” instead of “http://”, and most browsers will display some type of green padlock symbol or green highlighting of the address bar.

If the “s” at the end of “https://” is missing, or the address bar shows some type of red or orange warning, do not enter your password; further investigation is needed.

Be cautious if the address changes unexpectedly or if after clicking on a link, you are taken to an unfamiliar address.

Be familiar with how your browser handles IDNs. Chrome has an official page (http://dev.chromium.org/developers/design-documents/idn-in-google-chrome) with links to related information for other popular browsers at the bottom of that page.

Finally, for all of the websites that support it, make sure you enable two-factor authentication (2FA). If your credentials are phished, having 2FA enabled can provide an extra layer of security that can both alert you to a compromise of your credentials and prevent an attacker from logging into your account. Note that even with 2FA enabled your cellphone may become the weakest link in the security chain (https://www.nytimes.com/2017/08/21/business/dealbook/phone-hack-bitcoin-virtual-currency.html). If your phone is used as a back-up device for resetting passwords, make sure you protect your cellular account with a strong PIN-code (and hope the customer service agents are well trained enough to enforce its use for sensitive customer requests).

How Can I Protect My Organization?

If you operate a popular website that allows users to interact with one another, log in, purchase and/or download things, chances are your brand (and therefore your users!) will be on some target list for phishers and other Internet-based criminals.

You will want to pay attention to the IDN space, and potentially try to proactively defensively register some of the IDN homographs that could be used to impersonate your brand or subscribe to a service that allows you to monitor recent IDN homograph registration and use in an attempt to impersonate your brand.

What Can I Do as a Registry or Registrar?

If you are a registry operator, you can help. First and foremost, you can follow the Guidelines for the Implementation of Internationalized Domain Names and enforce that your registrars do as well. Looking at Version 4.0 of these guidelines[29], following the guidelines set forth in section 2.5, "Similarity and Confusability of Labels" would go a long way towards helping to stem the flow of new IDN homograph registrations.

Additionally, you can follow the lead of EURid[30], the registry manager for .eu and .ею, and offer "homoglyph bundling". This is a service where domain names that might look confusingly similar are prevented from being registered. Quoting EURid:

> *"Homoglyph bundling is when you register an IDN and the registration system automatically bundles all the homoglyphs of that name (if there are any). This means that several domain names are bundled at one time, and none of the other domain names in that bundle can be registered[31]".*

While this does enable domain name squatters to effectively nab an entire "family" of domain name-based brands, in theory, it makes it easier for the rightful brand owner to dispute[32] fraudulent registrations since there is only a single infringing party. Regardless of this possible caveat, we find the technique of homoglyph bundling to be an exemplary service and would like to see it offered elsewhere.

[29] A link to the current version of these guidelines, 3.0, is provided above. A draft of version of 4.0 may be found here: https://www.icann.org/en/system/files/files/idn-guidelines-10may18-en.pdf.

[30] https://eurid.eu/en/

[31] https://eurid.eu/en/register-a-eu-domain/domain-names-with-special-characters-idns

[32] https://eurid.eu/en/register-a-eu-domain/domain-name-disputes/

For More Information

Farsight Security is the world's largest provider of historical and real-time passive DNS data. We enable security teams to qualify, enrich and correlate all sources of threat data and ultimately save time when it is most critical - during an attack or investigation. Our solutions provide enterprise, government and security industry personnel and platforms with unmatched global visibility, context and response. Farsight Security is headquartered in San Mateo, California, USA. Learn more about how we can empower your threat platform and security team with Farsight Security passive DNS solutions at www.farsightsecurity.com or follow us on Twitter: @FarsightSecInc.