Overview of phishing landscape and homographs in Arabic domain names

Humza Ahmad | Laszlo Erdodi

Department of Informatics, University of Oslo, Oslo, Norway

Abstract

In this paper, we summarize the latest social engineering phishing attack types with the focus on domain name manipulation. Providing a fake domain is a crucial part of phishing attacks that can be carried out with different techniques such as cybersquatting, typosquattings, or homographs. We argue that homographs with special International Domain Names can be very serious threat for many users that can hardly be identified. We present the analysis of phishing attacks with Arabic domains and conclude that because of the linguistic complexity of the Arabic language and the missing support by the industrial tools expose millions of users to sophisticated domain manipulation based phishing attacks. To reduce the future risk of homograph attacks originating from Arabic internationalized domain names (IDNs) we suggest a change to the registration policy. We also present a browser extension to assist against homographs in current Arabic IDN namespace.

KEYWORDS

Arabic language, browser extension, homographs, phishing, social engineering

1 | INTRODUCTION

As a social engineering network attack, phishing aims to steal personal and financial information such as credit card numbers and login credentials with spoofed web pages. According to Federal Bureau of Investigation’s (FBI) internet crime activity report of 2019, cybercrime resulted in a loss exceeding $3.5 billion. Of this about $1.7 billion is attributed to phishing attacks. It is also important to mention that this report consists of the cases that were reported to the FBI in the United States, but the global figures are much higher than this. CheckPhish reported a spike of 225% in phishing sites during COVID-19 pandemic. The Anti Phishing Work Group reported discovery of 165 772 and 146 994 new phishing sites for first and second quarter of 2020.

Study shows that the cause of users falling for such schemes is mainly due to inexperience (no proper training or awareness regarding such attacks), carelessness, temptation of lottery, and so on. Another reason of the success is the continuous improvement of these attacks that means more and more realistic and tricky ways to mislead the users. Defense against phishing schemes can be segmented into two approaches, the user training approach and the automated approach. Even when properly trained, users are susceptible to fall victim to highly engineered phishing attacks as seen in Twitter take over, which was a sophisticated spear phishing attack that targeted twitter employees. In this paper we
focus on domain component of a phishing attack, peculiarly we explore homograph attack vector with internationalized domain names (IDN) of Arabic character set. The paper is organized as follows. Section 2 presents the background of phishing type social engineering attacks, Section 3 discusses the anatomy of these attacks. Sections 4 and 5 focus on the domain manipulation attacks and the defense, in Section 6 our proposed solution is presented to mitigate homograph attacks. Finally the conclusions are summarized in Section 7.

2 | BACKGROUND

Phishing is a social engineering technique where user trust is abused. In Reference 6 the authors give an overview and insight to social entrapment and attack scenarios. A general operating procedure for phishing attack is:

1. Reach the user by means of a communication channel like email or SMS.
2. Persuade the user to click on a link which redirects to an attacker controlled site.
3. Believing the look-alike site to be real, user provides his information. Depending upon the scam this can be personal information like login credentials, social security number, or financial information like credit card details.
4. Using this information, malicious actors can do severe damage to user’s reputation or monetary assets. The attacker can also masquerade as this user to carry out further attacks.

2.1 | Types of phishing

Depending upon the communication channel and target group phishing attacks can be classified into following categories.7-12

- **Vishing** refers to phishing done over phone calls. The word *Vishing* is a combination of voice and phishing. Vishing attack occurs when a malicious actor impersonates a valid person or organization such as a bank, or government agency in order to obtain personal or financial information from the victim over a phone call. Voice over IP (VoIP) technologies make it easy for attacker to spoof and route calls.

- **SmiShing** is a type of phishing attack where the mode of initial entrapment is SMS, rather then by means of email. *SmiShing* is a portmanteau of “SMS” (better known as text messages) and “phishing.” Most of the time people are vigilant when dealing with emails but when using phones they tend to be less wary and believe their phones to be more secure then computers. Attacker can use two approaches to gain private information. They could either make the user download a malware infected app or redirect them to a fake website.

- **Search Engine Phishing** is a type of phishing where websites are seeded with keywords that a user may search. These sites are indexed by search engines and attacker waits for users to turn up. Malicious actors can also pay the search engine provider to provide link to their website in adverts. Most common examples for this are sites offering free stuff, discount offers, job offers, fake market places.

- **Spear Phishing** unlike phishing in general, where phishing messages are sent to millions of users in hope that some will fall prey to the scam, is a targeted attack. In-depth user profiling is performed to carry out this attack. This information is then used to contrive personal, authentic and more persuasive scenarios to entrap the target user. Spear phishing can be carried out by means of regular email phishing, smishing, and vishing.

- **Whaling** is a special kind of spear phishing attack. It is also known as CEO Fraud. In this kind of attack, malicious actors target the executives of an organization. If the attack is successful the attackers by masquerading as executives can authorize financial transactions. Using the same account the attacker can also get private information of the employees of the organization this can be social security numbers, tax-return information. For the most part the success rate for this type of attack is because CEO's usually do not attend security awareness meetings.

- **Pharming** also known as DNS Spoofing or DNS chache poison is a type of phishing attack where DNS record is poisoned. The domain naming system is used to map human readable name for website to its IP address. By poisoning DNS cache the attacker remaps the IP address for valid domain name to IP of attacker controlled machine. The fake website will appear as identical to the original; thus, entrapping the target user.
- **Clone phishing** attacks take advantage of legitimate messages that the target user may already have received. The attacker creates a duplicate message and replaces valid URLs with malicious ones. Malicious actors utilize the narrative of there being an issue with the attachment or link in previous email to lure the user to click on malicious link or download malware.

- **Watering hole phishing** is a type of phishing attack where attackers try to infect one of the websites frequented by the targeted user. If successful, once targeted user visits the site again malware is installed on their system.

### 2.2 Required expertise

A generic email based phishing attack can be carried out by anyone as the requirement of technical knowledge is not high. It only requires knowledge of domain and website setup. Readily available tools can be easily obtained that simplify the process. Carrying out vishing or smishing attacks mandates more effort as setup and configuration for voip and SMS gateways is not easy. But in this case as well attackers can make use of existing infrastructure of service providers catering to the telecom industry. Apart from the technical know-how the most important factor is social skills.

### 3 ANATOMY OF PHISHING ATTACK

As discussed in Section 2, a phishing attack can be spilt into three segments:

#### 3.1 Social entrapment

In a phishing attack the message part can be regarded as the entry point to secure a base foothold. No matter how good the website and domain setup is if the target does not fall for the story then it is of no use. For a successful entrapment the attacker will fabricate a story to which the target can relate to on a personal level. For example it could be in regards to an upcoming meeting or followup on a recent purchase. The aim of the message is to make the reader perform the required action for the attack to proceed, which in this case is usually clicking on the enclosed uniform resource locator (URL).

#### 3.2 Indistinguishable web page

Landing on an unknown web page is bound to raise flags. To not raise any suspicions it is important to mimic the real website behavior. Usually the aim is to obtain login credentials or billing details like credit card information. To not raise an suspicion, after siphoning information from login page the user is redirected to the actual website. As mentioned in Section 2.2 there are tools and scripts available which make this much easier to setup.

#### 3.3 URL attack vector

What makes a site unique is its URL. Each page and artifact on a website has its own URL. Figure 1, illustrates the anatomy of a URL. As shown, a URL consists of a protocol https://, a fully qualified domain name (FQDN) www.exampleurl.com and file path /info/aboutus.html. FQDN (also known as host name) is mapped to an IP address, which identifies the host machine over the network.

![Uniform resource locator structure](https://www.exampleurl.com/info/aboutus.html)
The host name is further divided into domain levels, starting from top-level domain (TLD) then second-level domain name, third-level domain name, nesting can be done to n-level domain name. The domain name part (usually TLD and second-level domain e.g., google.com) cannot be freely manipulated as domains have to be registered with a domain registrar. But the attacker has free reign over subdomain portion. Consider the below URL https://apple.login.webappACJio22H3.fastrestore.ru

The attacker has added subdomain “apple” to his domain “fastrestore.ru.” Users might not check the full URL and may think that this link points to Apple website. This project will focus study on domain names nested to second level like google.com.

4 | DOMAIN NAME MANIPULATION

Below are different methodologies and techniques for manipulating domain names for phishing attacks.

4.1 | Cybersquatting

Also known as Domain Squatting is the act of registering a domain for the purpose of profiting from a trademark belonging to an individual or a company.14 For example if a user owns example.com cybersquatters can register the domain with another TLD like example.no or exemple.org. Cybersquatters can profit off by selling back the domain to trademark owner at an inflated price or by carrying out brand related frauds and scams like phishing.

4.2 | Typosquatting

Also known as URL Hijacking is the process of registering domains with typographical errors.15 Typosquatting is based on the notion that users make typing errors when entering domain names. A classic example for typosquatting is goggle.com for google.com.

4.3 | Homograph domain

These are domains which appear to be visually similar. Since the incorporation of IDN by the Internet Corporation for Assigned Names and Numbers individuals and companies can now register domains in their native languages. IDN scheme allows registration using Unicode characters.16 For example:

- faceb00k.com: the letter “o” is in “facebook” are replaced by the number “zero.”
- facebook.com: the letter “o” is actually the Cyrillic small letter “o” (0x43E), not the Latin “o” (0x6F).

5 | DETECTION TECHNIQUES

Over the years evolutionary research has been done in the field of security. This extensive research has also led to the development of numerous approaches for defense against phishing attacks. Here we look into the techniques used in detection of fake sites; not into message evaluation mechanisms which are utilized for detection of spam, phishing emails.

5.1 | Blacklists

As the name suggests this methodology implores on the concept of gatekeeping. This is the most simple defense mechanism. Once a site/domain is confirmed to be malicious it is added to a blacklist. These lists are maintained and updated on a regular basis. When users try to open a URL in their browser, the browser and security solutions check whether the
domain is blacklisted. If it is blacklisted a warning message is displayed to the user. Blacklists are also used by search
engines and email providers to check for and filter out malicious content. The drawback of blacklisting is that it only
helps defend against domains which are already known. \(^\text{17}\)

### 5.2 Visual similarity

Visual similarity based phishing detection techniques utilize the feature set like text content, text format, HTML tags,
cascading style sheet, image, and so forth, to make the decision. These approaches compare the suspicious website with
the corresponding legitimate website by using various features and if the similarity is greater than the predefined threshold
value then it is declared phishing. \(^\text{4,18,19}\)

### 5.3 Heuristics and machine learning

Heuristics and machine learning based approaches use data analysis techniques to determine whether a site is malicious
or not. This can be the analysis of the network traffic, visual, and textual features extracted from the site or a combination
of both. Some of these features can be,

- **URL based features**
  - URL length
  - Checking for Typosquatting
  - Checking for Cybersquatting
  - Examining URL for use of brand names
  - Subdomain count

- **Domain based features**
  - Domain blacklist status as mentioned in Section 5.1
  - Domain age (days since domain was registered)

- **Page based features**
  - Page rank (Global and Country Based)
  - Estimated number of visits per day, week, month
  - Average page views per visit
  - Average visit duration
  - Site references from social media

- **Content based features**
  - Visual features as mentioned in Section 5.2

Heuristics and machine learning based approaches are highly dependent on the data used for training the classifiers.
ML based phishing URL detectors boast a success rate of more than 90%. An issue with machine learning based phishing
domain detectors is that, these classifiers work very well with long URLs but do not generate good results for only domain
names (second level domain). \(^\text{13,20}\)

### 5.4 Available tools

Majority of the complete protection suites are only commercially available, some also offer community versions but with
limited functionality. Below are some of the well-known phishing detection tools.
**Immuniweb**: As part of their community tools, they provide a domain checking tool. The tool checks for brand infringement, domain squatting, typo squatting, and potential phishing. For detection they use their own proprietary mechanisms in conjunction with PhishTank, Google Safe API, OpenPhish.21

**CheckPhish and Bolster**: Bolster provides AI based security tools for zero-day detection and prevention. CheckPhish is their phishing URL detector available for the community. Like Immuniweb their detection mechanism is proprietary.22,23

**Open Source - GitHub**: Open source phishing detectors are available on Github. Most of the available tools are based on machine learning approach for phishing URL detection. As discussed in Section 5.3, the phishing domain detectors work very well for long URLs but do not generate good results for base domain names.

**TypoSquatting Tools**: presents a performance analysis of most popular market tools for typosquatting. These tools use fuzzing techniques to generate an exhaustive list of malicious domain names, and then check if these domain names are in use or not. Among them Dnstwist and Typofinder generate the highest number of typosquatted domain names.

**TypoWriter**: The Paper *TypoWriter: A Tool to Prevent Typosquatting*, presents a tool for defensive preventive registration of domain names using recurrent neural network (RNN). They analyzed the domain logs for .bg TLD. Using deterministic rules pertaining to edit distance, time difference, IP address, they prepared a dataset of domain pairs in the form of valid domain and it’s typo variant. N-grams are constructed for these pairs which are later used to train RNN model. As compared to typosquatting tools mentioned in Section 5.4, TypoWriter aims to generate the least number of most probable domain names, rather than arbitrarily selecting domains for preventive registration.

**Homograph Domains; Defense and Active Attack Vector**

The most easiest and straightforward method of defense against homograph attacks is to prohibit cross script domain name registration. Cross script domain registration implies that users are able to register domains by mixing characters from different languages for example being able to register a domain name by mixing characters from Latin and Cyrillic. This defense mechanism is implemented at the domain registrar end. The registration rules for a TLD are defined by it’s authoritative registry. For example the authoritative registry for .com TLD “Verisign” does not allow commingling of certain language scripts to prevent homograph attacks.25

Punycode encoding is the representation of Unicode in ASCII. This encoding mechanism is used for storing and resolving internationalized domain names in the DNS protocol space. Punycode domains begin with the characters “xn-.” In order to help users identify if the site they are visiting has a multi-script IDN browsers display the punycode for that IDN in the address bar but if the domain name has characters belonging to only one character set like Latin or Arabic then punycode representation is not displayed. Table 1 shows homograph IDN example for domain google.com. Visually both domain names are identical but domain with punycode “xn-gogle-kye.com” generated by amalgamation of Latin and Cyrillic characters.

Though the above methodology provides a defense mechanism against multi-script homograph attacks, attack vector for homograph attacks within the same language namespace still exists. As shown in the Table 2, both domains are registered using Arabic script, visually the domain names appear to be same but as highlighted the last character is different. The character (06C1) is typed using Urdu keyboard whereas the character (0647) is typed using Arabic keyboard.

Unlike the homograph domain name showed in Table 1, where it is not possible to register xn-gogle-kye.com due to registration rules, both domain listed in Table 2, are valid and registered. This capacity to register domain with identical characters in Arabic character set provides a versatile attack space for threat actors to act upon.

To the extent of our knowledge the tools available for phishing detection fail to provide support for Arabic based domain names. This is not the first time that issue for Arabic homograph is highlighted. Since acceptance of IDNs threat factors from multilingual domains have been raised,26 but most of the focus has been on languages which appear to be visually similar to Latin. We believe there can be two contributing factors to this.

- Industrial tools do not provide support for Arabic domains as the stakeholders involved do not perceive it as a tangible threat. This can be based on the fact that majority of the web content is dominated by English language and content in Arabic is less than 5%.27 Though no major attacks have been seen, without proper scientific study; risks from Arabic
| Table 1 | Internationalized domain name homograph cross script (English and Cyrillic) |
|---------|---------------------------------------------------------------|
| **Domain name** | **Unicode representation** | **Punycode representation** |
| google.com | g (0067)  o (006F)  o (006F)  g (0067)  l (006C)  e (0064) | google.com |
| google.com | g (0067)  o (006F)  o (043E)  g (0067)  l (006C)  e (0064) | xn-gogle-kye.com |

| Table 2 | Internationalized domain name homograph single script (Arabic) |
|---------|---------------------------------------------------------------|
| **Domain name** | **Unicode representation** | **Punycode representation** |
| همزة and .com | ء (062D)  ر (0645)  ج (0632)  ت (06C1) | xn–sgbk0dv2a.com |
| همزة and .com | ء (062D)  ر (0645)  ج (0632)  ت (0647) | xn–sgbk0dj.com |

| Table 3 | On combination visually different characters become similar |
|---------|---------------------------------------------------------------|
| **Single Character** | **Combined** |
| ئ + 1 | ئ |
| ل + 1 | ئ |

Based IDNs cannot be written off as close to 24.1% of world population can interpret this language. With the ease of technological access in these demographics it is only a matter of time before localized IDN attacks start surfacing.

- The second contributing factor that we perceive is the linguistic complexity of Arabic script. For detection, generation of homographs in scripts like Latin, similar characters are assorted based on deterministic rules of substitution, addition and subtraction. But in case of Arabic character set we believe this is not enough as the visual appearance of words can change depending upon the combination of characters used. Unlike in languages like Latin, Cyrillic where characters and words have a fixed visual representation; in Arabic some characters on their own do not have much of visual similarity but when used in combination with other characters they appear to be identical. As show in Table 3.

### 6.1 AI based detection of Arabic sites

In recent years use of machine learning to detect phishing attacks has proven to be quite useful. As suggested in Reference 29 classifiers can be trained on a number of features extracted from website URL and HTML. Though AI seems to be the way forward, the pre-requisite of any machine learning solution is the availability of sufficient amount of quality training.
data. For the issue of Arabic phishing scams in terms of AI based solution, the training set needs to consist of Arabic websites, whose URL and HTML contains content in Arabic scripture. Unfortunately at the time of writing there is a lack of Arabic websites specially in terms of phishing samples. As discussed above, websites with content in Arabic are just about 5% of the whole world wide web, this makes it very difficult to collect and generate a good quality data set to train a machine learning model for detection of Arabic phishing sites. If at a future time, sufficient data samples are successfully collected then in combination with the “Plug n Play” mechanisms presented in Reference 30 and feature selection techniques, an AI based solution can be realized to tackle phishing attacks involving Arabic websites.

6.2 Proposed solution

To mitigate the future risk of homograph domains from Arabic, Urdu, and Farsi lexical we suggest a change in the policy pertaining to registration of Arabic IDNs with respect to homographs. As mentioned above the current policy does not allow registration of domains by mingling characters from languages which contain similar characters. The Arabic character set by itself contains multiple characters and symbols which are visually identical. This is to support other language locales like Urdu, Farsi, Kurdish which have many common characters with Arabic. We suggest that the identical homograph pairs be blacklisted and only single character from the pair be allowed in domain registration. Considering the low statistics of domain registration using Arabic character set these policy changes will not have any major implications on user experience but will greatly curb the risk of future homograph phishing scams based on Arabic IDNs.

For the already existing IDNs we have created a browser extension specifically targeting IDNs based on Arabic lexical. Existing browser extensions which target IDNs only highlight the fact that user has opened an IDN website or it will straight out block the user from visiting an IDN page even if it is a valid one. These extensions rely on whitelisting to manage IDNs. Only examining the punycode is of not much help to the untrained eye as the combination of characters which form the punycode do not make much sense.

This extension examines the Arabic domain name and prompts user with similar existing IDNs. This not only helps with alerting the user to a possible phishing site but also if the user has entered the wrong domain due to different keyboard locale setting. Based on homograph character mappings that we have identified in the tables below; homograph domains are created and then their existence is verified by means of DNS lookup using Google DNS service. The below Figures 2 and 3 shows the output of browser plugin when the user opens the site فيلم جزاءه.com in the browser. The plugin is only activated when user navigates to an Arabic domain. The code for the browser plugin can be accessed at this GitHub repository. It should also be noted that at the moment only homographs of the first order are generated. This implies

\[\text{FIGURE 2} \quad \text{Arabic internationalized domain name in browser}\]

\[\text{FIGURE 3} \quad \text{Output of browser extension}\]
that for each permutation only a single character is replaced whereas the rest are the original characters from the input domain name. This is further illustrated in the Figure 4.

Below in the Appendix section (Tables A1–A18), we have identified homograph pairs from Arabic script for substitution. Unlike other scripts like Latin, characters in Arabic script are not only similar but also very identical. For identification of homograph characters Unicode range of 0600-06FF (256 characters) was analyzed as dot com TLD only supports Arabic characters in the range of 0600-06FF.33

7 | CONCLUSION

Phishing is not something new, it has been around for the past 26-years and it is not going away anytime soon. Defense against cyber attacks is a continuous effort of user education and enchantment of automated detection systems. To mitigate the risk of single script IDN attacks like Arabic homographs we suggest a change in registration policy. In addition to disallowing registration of domain names by co-mingling of characters from different scripts (like English + Cyrillic), it should also disallow registering domain names created by mixing characters between different Input Locales (keyboard input profiles). In implementation terms with regards to Arabic and its sub languages (like Urdu and Persian) this means that from the unicode profile of Arabic language the identical homograph pairs be blacklisted; allowing only a single element (character) from the set to be used. As described in Section 6, constraint on general script level does not work for languages like Arabic, Urdu, Persian as they use characters from Arabic script. For the existing IDNs using Arabic character set the provided browser plugin helps raise user awareness by prompting the user with active domain names similar to the one which is currently open.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Humza Ahmad
https://orcid.org/0000-0002-3905-823X
REFERENCES

1. FBI. 2019 Internet Crime Report. 2019. https://pdf.ic3.gov/2019_IC3Report.pdf
2. Alafgani J. Phishing Sites Spike 235% During the COVID-19 Pandemic. https://checkphish.ai/blog/covid-19-phishing-sites-spike-235-percent-as-19-pandemic-creates-larger-remote-workforce/?utm_source=checkphish&amp;utm_medium=et
3. APWG. Phishing Activity Trends Report [2nd Quarter 2020]. 2020. https://docs.apwg.org/reports/apwg_trends_report_q2_2020.pdf
4. Zhou Y, Zhang Y, Xiao J, Wang Y, Lin W. Visual Similarity Based Anti-phishing with the Combination of Local and Global Features. 2014: 189–196.
5. Inc. T. An update on our security incident. https://blog.twitter.com/en_us/topics/company/2020/an-update-on-our-security-incident.html
6. Yasin A, Fatima R, Liu L, Yasin A, Wang J. Contemplating social engineering studies and attack scenarios: a review study. Secur Privacy. 2019;2(4):73. https://doi.org/10.1002/spy2.73.
7. Bisson D. 6 Common Phishing Attacks and How to Protect Against Them. https://www.tripwire.com/state-of-security/security-awareness/6-common-phishing-attacks-and-how-to-protect-against-them/
8. Cook S. What is Vishing (with example scams) and how to recognize and avoid it. https://www.comparitech.com/blog/information-security/what-is-vishing-how-to-avoid/
9. Crane C. 10 Types of Phishing Attacks and Phishing Scams. https://www.thesslstore.com/blog/10-types-of-phishing-attacks-and-phishing-scams/
10. Kaspersky. What is Vishing. https://www.kaspersky.com/resource-center/definitions/vishing
11. Kaspersky. What is Smishing and How to Defend Against it?. https://www.kaspersky.com/resource-center/threats/what-is-smishing-and-how-to-defend-against-it
12. Sander R. Ten types of phishing attacks and phishing scams. https://www.itproportal.com/features/ten-types-of-phishing-attacks-and-phishing-scams/
13. Büber E. Phishing URL Detection with ML. https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5
14. Mathew AR, Hajj AA, Ambusaidi M. Cybersquatting. Paper presented at: 2010 International Conference on Educational and Information Dent Tech. 2010. https://doi.org/10.1109/ICETT.2010.5607800
15. Chen G, Johnson MF, Marupally PR, Singireddy NK, Yin X, Paruchuri V. Combating Typo-Squatting for Safer Browsing. Paper presented at: 2009 International Conference on Advanced Information Networking and Applications Workshops. 2009. https://doi.org/10.1109/WAINA.2009.98
16. Quinkert G, Lauinger T, Robertson W, Kirda E, Holz T. It’s Not what It Looks Like: Measuring Attacks and Defensive Registrations of Homograph Domains. Paper presented at: 2019 IEEE Conference on Communications and Network Security (CNS). 2019. https://doi.org/10.1109/CNS.2019.8802671
17. Zuraiq AA, Alkasasbeh M. Review: Phishing Detection Approaches. 2019: 1–6.
18. Haruta S, Yamazaki F, Asahina H, Sassa I. A Novel Visual Similarity-based Phishing Detection Scheme using Hue Information with Auto Updating Database. 2019. pp. 280–285.
19. Jain AK, Gupta BB. Phishing Detection: Analysis of Visual Similarity Based Approaches. 2017
20. Pettry Z. AI: Deep Learning for Phishing URL Detection. https://www.zpettry.com.ai/+cybersecurity/ai-url-phishing-detection/
21. Web I. ImmuniWeb Domian Security Test. https://www.immuniweb.com/radar/#about
22. Bolster. AI-powered Zero Day Phishing & Counterfeit Website Detection. https://bolster.ai/technology/
23. Bolster. Real-Time Phishing Detection. https://checkphish.ai
24. Ahmad I, Parvez MA, Iqbal A. TypoWriter: A Tool to Prevent Typosquatting. In. 1. 2019. pp. 423–432.
25. Verisign. IDN Registration Rules. https://www.verisign.com/en_US/channel-resources/domain-registry-products/idn/idn-policy/registration-rules/index.xhtml
26. Al Helou J, Tilley S. Multilingual web sites: Internationalized Domain Name homograph attacks. 2010. pp. 89–92.
27. Report IW. DNs and Linguistic Diversity. https://idnworldreport.eu/charts/ids-and-linguistic-diversity/ids-and-linguistic-diversity/
28. Lipka M, Hackett C. Why Muslims are the world’s fastest-growing religious group. https://www.pewresearch.org/fact-tank/2017/04/06/why-muslims-are-the-worlds-fastest-growing-religious-group/
29. Shahrviri V, Darabi MM, Izadi M. Phishing Detection Using Machine Learning Techniques. 2020.
30. Ridhawi IA, Otoum S, Aloeqlay M, Boukerche A, Generalizing AI. Challenges and opportunities for plug and play AI solutions. IEEE Network. 2021;35(1):372-379. https://doi.org/10.1109/MNET.011.2000371.
31. Tharani JS, Arachchilage NAG. Understanding phishers’ strategies of mimicking uniform resource locators to leverage phishing attacks: a machine learning approach. Secur Privacy. 2020;3(5):e120. https://doi.org/10.1002/spy2.120.
32. Ahmad H. Arabic-Homograph-Domains. 2021. https://github.com/Darthcedeus/Arabic-Homograph-Domains
33. Verisign. Registration Rules Arabic. https://www.verisign.com/assets/idn/idn-arabic.html

How to cite this article: Ahmad H, Erdodi L. Overview of phishing landscape and homographs in Arabic domain names. Security and Privacy. 2021;4:e159. https://doi.org/10.1002/spy2.159
APPENDIX A. ARABIC HOMOGRAPH CHARACTERS

The Tables A1 to A18 contain homograph characters. Each table contains characters which appear to be visually similar not only when written as single characters but also when written in combination with other characters.

### TABLE A1 Homograph for character “YE”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ی | U+0649 ی | U+0620 ی | U+0678 ی |
| ی | U+0661 | U+0672 ی | U+0622 |
| ی | U+06F1 | U+0623 ی |

### TABLE A2 Homograph for character “ALF”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ١ | U+0661 ١ | U+0672 ١ | U+0622 |
| ١ | U+06F1 | U+0623 ١ |
| ١ | U+0673 ١ | U+0622 ١ |
| ١ | U+0625 ١ | U+0671 ١ |

### TABLE A3 Homograph for character “AEEN”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
|  추진 | U+060F  추진 | U+063A  추진 |
|  ع | U+0639 ع | U+06A0 ع |
|  ع | U+06FC ع |

### TABLE A4 Homograph for characters “BAY,” “PAY,” “TAY”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
|  ب | U+0628 ب | U+067D ب |
|  ب | U+067B ب | U+062B ب |
|  ت | U+0680 ت | U+067F ت |
|  ت | U+067E ت | U+067C ت |
|  ت | U+067A ت | U+062A ت |
**TABLE A5** Homograph for character “WAO”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ﷯         | U+0624  | ﷯         | U+06C7  |
| ﷷ         | U+0676  | ﷴ         | U+0677  |
| ﷳ         | U+0648  | ﷴ         | U+06C9  |
| ﷲ         | U+06C5  | ﷶ         | U+06C6  |
| ﷰ         | U+06C4  | ﷲ         | U+06C8  |
| ﷱ         | U+06CF  | ﷳ         | U+06CA  |
| ﷲ         | U+06CB  |

**TABLE A6** Homograph for characters “DAAL,” “ZAAL”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ﷮         | U+062F  | ﷪         | U+0690  |
| ﷫         | U+068A  | ﷪         | U+068F  |
| ﷬         | U+0689  | ﷪         | U+068E  |
| ﷫         | U+068D  | ﷪         | U+068C  |
| ﷬         | U+068B  | ﷪         | U+068E  |
| ﷬         | U+0688  | ﷪         | U+0630  |
| ﷫         | U+0691  |

**TABLE A7** Homograph for characters “REE,” “ZEE”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ﷮         | U+0631  | ﷮         | U+0632  |
| ﷫         | U+0693  | ﷬         | U+0697  |
| ﷫         | U+0695  | ﷬         | U+0698  |
| ﷬         | U+0694  | ﷮         | U+0699  |
| ﷫         | U+0696  | ﷮         | U+06EF  |

**TABLE A8** Homograph for character “JEEM”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ﷬         | U+062E  | ﷬         | U+062C  |
| ﷫         | U+0682  | ﷬         | U+0683  |
| ﷬         | U+0681  | ﷬         | U+0684  |
| ﷬         | U+062D  | ﷬         | U+06BF  |
| ﷫         | U+0685  | ﷬         | U+0686  |
| ﷬         | U+0687  |
### TABLE A9  Homograph for character “HEEH”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| 🇧 🇫 | U+0647  | ق | U+06C0 |
| 🇧 🇫 | U+06C1  | ق | U+06C2 |
| 🇧 🇫 | U+06D5  | ق | U+0629 |
| 🇧 🇫 | U+0665  | ق | U+06C3 |

### TABLE A10  Homograph for character “TWEEN”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ط | U+0637  | ط | U+0638 |
| ط | U+069F  | |

### TABLE A11  Homograph for characters “SWAAD,” “ZWAAD”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ض | U+0636  | ض | U+0635 |
| ض | U+069E  | ح | U+069D |
| ح | U+06FB  | |

### TABLE A12  Homograph for characters “SEEN,” “SHEEN”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| س | U+0633  | ش | U+069C |
| ش | U+069B  | ش | U+06FA |
| ش | U+069A  | ش | U+0634 |

### TABLE A13  Homograph for character “NOON”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ن | U+0646  | ن | U+06BB |
| ن | U+06B9  | ن | U+06BA |
| ن | U+06BC  | ن | U+06BD |

### TABLE A14  Homograph for character “LAAM”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ل | U+0644  | ل | U+06B5 |
| ل | U+06B7  | ل | U+06B8 |
| ل | U+06B6  | |
| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ك | U+0643 | ك | U+063C |
| ل | U+06AE | ل | U+06AB |
| ل | U+06AD | ل | U+06A9 |
| أ | U+06AC | أ | U+063B |

**TABLE A16** Homograph for characters “FEH”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| ف | U+0641 | ق | U+066F |
| ق | U+06A6 | ق | U+06A7 |
| ق | U+06A4 | ق | U+0642 |
| ق | U+06A1 | ف | U+06A8 |
| ف | U+06A5 | ف | U+06A3 |
| ف | U+06A2 |

**TABLE A17** Homograph for character “GAAF”

| Character | Unicode | Character | Unicode |
|-----------|---------|-----------|---------|
| گ | U+06AF | گ | U+06B3 |
| گ | U+06B0 | گ | U+06B2 |
| گ | U+06B4 | گ | U+06B1 |

**TABLE A18** Homograph for character “BARI YEE”

| Character | Unicode |
|-----------|---------|
| ن | U+06D2 |
| ن | U+06D3 |