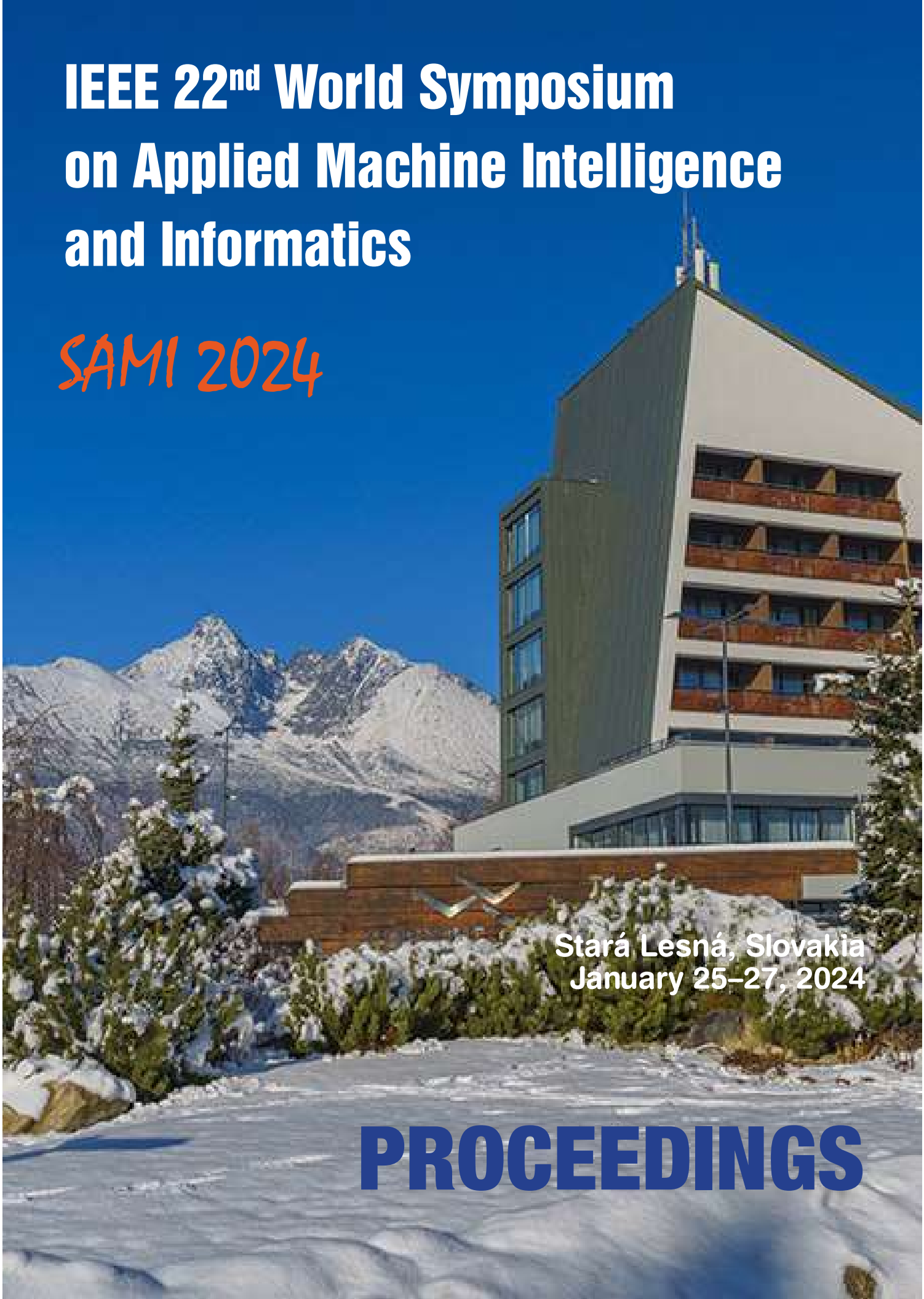
IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics

SAMI 2024



Stará Lesná, Slovakia
January 25–27, 2024

PROCEEDINGS

Organizers and Sponsors

Technical University of Košice, Slovakia
Óbuda University, Budapest, Hungary
University Research and Innovation Center
Antal Bejczy Center for Intelligent Robotics
ELFA Ltd., Košice, Slovakia
Slovak Academy of Sciences
Hungarian Fuzzy Association
SMC TC on Computational Cybernetics
IEEE Computational Intelligence Chapter of Czechoslovakia Section

Sponsors

IEEE Hungary Section
IEEE Joint Chapter of IES and RAS, Hungary
IEEE Control Systems Chapter, Hungary
IEEE SMC Chapter, Hungary

Technical Co-Sponsor

IEEE SMC Society

	Part Number	ISBN
XPLORE COMPLIANT:	CFP2408E-ART	979-8-3503-1720-6
USB:	CFP2408E-USB	979-8-3503-1719-0

Welcome from the Chairs

Computational Intelligence and Intelligent Technologies are very important tools in building intelligent systems with various degree of autonomous behavior. These groups of tools support such features as ability to learn and adaptability of the intelligent systems in various types of environments and situations. The current and future Information Society is expecting to be implemented with the framework of the Ambient Intelligence (AmI) approach into technologies and everyday life. These accomplishments provide the wide range of application potentials for Machine Intelligence tools to support the AmI concept implementation. The number of studies indicates that this approach is inevitable and will play essential and central role in the development of Information Society in close future.

The essential importance of the Machine Intelligence in this historically challenging effort points out the responsibility of MI community including all fields like Brian-like research and applications, fuzzy logic, neural networks, evolutionary computation, multi-agent systems, artificial life, Expert Systems, Symbolic approaches based on logic reasoning, Knowledge discovery, mining, replication and many other related fields supporting the development and creation of the Intelligent System. The importance embedding these systems in various kinds of technologies should bring profit and different role of mankind in production and in everyday life. We expect to have intelligent technologies, solution and even humanoid robots to help the mankind to improve and keep the ideas of humanity and democracy.

The role of Machine Intelligence Quotient will play an important role in the future to be able to evaluate the degree of the autonomous behavior of the designed system. It is belief that it will be domain oriented problem and should also be important to use this information for decisions made by humans e.g. in evaluation of many information system in commercial tender to pick up the system with the highest MIQ. The usefulness of this parameter will be dependent on many influences including technological, domain oriented and also commercial aspects of the CI application in various systems. The commercial need to have “intelligent” solution and products should increase the interest for MI tools.

This year number of contribution showed up from mechanical Engineering domain, control and also pure computer science. We do believe that this multidisciplinarity will be very useful to emerge more AI applications in Information Society and will help making products and solutions more “intelligent”.

This proceedings is a small contribution of knowledge dissemination and presentation of important problems and advances in Computational intelligence theory and applications. Hungary and Slovakia as members of EU will do their best to contribute to European Research Area and support the development of Computational Intelligence technology for the benefit of the mankind.

Levente Kovács and **Liberios Vokorokos**
General Chairs

Committees

General Chairs

Levente Kovács, Óbuda University, Budapest, Hungary
Liberios Vokorokos, Technical University of Košice, Slovakia

Founding Honorary Chair

Imre J. Rudas, Óbuda University, Budapest, Hungary

Honorary Committee

Stanislav Kmet, Technical University of Košice, Slovakia
Anton Čižmár, Technical University of Košice, Slovakia
Levente Kovács, Óbuda University, Budapest, Hungary
Peter Mésároš, Technical University of Košice, Slovakia

International Scientific Committee

Alin Albu-Schaeffer, German Aerospace Center, Germany
Philip Chen, University of Macau, Macau
Paolo Dario, Scuola Superiore Sant'Anna, Italy
Paolo Fiorini, University of Verona, Italy
Hamido Fujita, Iwate Prefectural University, Japan
Huijun Gao, Harbin Institute of Technology, China
Tamás Haidegger, Óbuda University, Budapest, Hungary
Keith Heipel, University of Waterloo, Canada
Oussama Khatib, Stanford University, USA
Kazuhiro Kosuge, Tohoku University, Japan
Gernot Kronreif, ACMIT Gmbh, Austria
Bernd Liepert, KUKA Roboter AG, Germany
Ren Luo, National Taiwan University, Taiwan
Vincenzo Piuri, Universita' degli Studi di Milano, Italy
Bruno Siciliano, University of Naples, Italy
Peter Sinčák, Technical University of Košice, Slovakia
Masayoshi Tomizuka, University of California, Berkeley, USA
Jacek Zurada, University of Louisville, USA

International Organizing Committee Co-Chairs

Frantisek Babič, Technical University of Košice, Slovakia
Marián Bucko, Elfa, Slovakia
Ladislav Főző, Technical University of Košice, Slovakia
Norbert Ádám, Technical University of Košice, Slovakia

Technical Program Committee Chairs

Szilveszter Kovács, University of Miskolc, Hungary
Rudolf Andoga, Technical University of Košice, Slovakia
Ivana Budinska, Slovak Academy of Science, Slovakia

Technical Program Committee

Norbert Ádám, Technical University of Košice, Slovakia
Rudolf Andoga, Technical University of Košice, Slovakia
František Babič, Technical University of Kosice, Slovakia
Péter Baranyi, Széchenyi István University, Győr, Hungary
Peter Bednár, Technical University of Kosice, Slovakia

Secretary General

Anikó Szakál
Óbuda University, Budapest, Hungary
E-mail: szakal@uni-obuda.hu

Iveta Zamecnikova
Technical University of Košice, Slovakia
E-mail: zamecnikova@elfa.sk

Table of Contents

URL and Domain Obfuscation Techniques - Prevalence and Trends Observed on Phishing Data

1st Ivan Skula
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
skula@dobraadresa.sk

2nd Michal Kvet
Faculty of Management Science and Informatics
University of Zilina
Zilina, Slovakia
michal.kvet@fri.uniza.sk

***Abstract*—For phishing to be successful, it is necessary to instill confidence and appear legitimate in the eyes of the potential victim, especially when mimicking a known brand. To achieve this, attackers employ various obfuscation techniques. Some are aimed to bypass existing technical (software) protections; others are aimed against the targeted victim (person). On the side of prevention, these techniques are seen as a clear sign of phishing, and many detection algorithms use these characteristics to decide whether to show or block the given webpage. Analysis conducted on 15 years of phishing data (2009-2023) collected from PhishTank and PhishStats websites focused on the prevalence and trends of various obfuscation techniques. These figures would allow validation and weighting of the relevancy of these indicators in phishing web page detection throughout the covered period and also provide a future baseline for creating a robust phishing dataset. Analysis steps required collecting and consolidating the phishing URL data. Due to the nature of the phishing data collection and their potential overlap, it was necessary to cleanse and filter out incorrect and duplicate records. The analysis's core part summarizes the selected techniques' prevalence and highlights notable observations. A noteworthy finding is that they occur rarely despite being a powerful indicator of phishing. Any of the techniques reviewed is present in less than ≈3% of the phishing URLs across the entire 15-year period. The most common techniques (in order of prevalence) are the - use of IP addresses, URL shorteners, ports, and Punycode. The remaining ones are extremely rare, with single or maximum double-digit occurrences.**

***Index Terms*—phishing, URL, domain, obfuscation techniques, trends**

I. Introduction

Last year (2022) has been another record year for phishing, with more than 4.7 million attacks recorded. Only in Q4 was this number more than 1.35 million [1]. Despite all the efforts to tackle or at least reduce phishing, the numbers are higher year by year. The overall direct financial losses are significantly lower than other types of online crime, e.g., 52 million USD as opposed to 3.3 billion USD of investment fraud. Yet, phishing is Nr.1 online crime type by the number of victims [2]. Combining that fact with the statement, "Phishing remains a key access vector for most online fraud schemes." [3] explains why accurate phishing detection is critical.

Phishing spans across a wide array of electronic channels (e-mail, SMS, voice call, web) and employs a multitude of techniques to stay under the radar and bypass not only technical measures (e.g., endpoint detection systems, anti-virus software, firewall) but also convince the user - a potential victim - that they are interacting with a genuine website. The oldest and most common techniques used are **obfuscation techniques**. URL or domain obfuscation - for which various techniques are reviewed in this analysis - focuses on concealing or making it hard to understand or recognize the actual destination URL or domain. They do so by manipulating one or multiple URL components "Fig. 1".

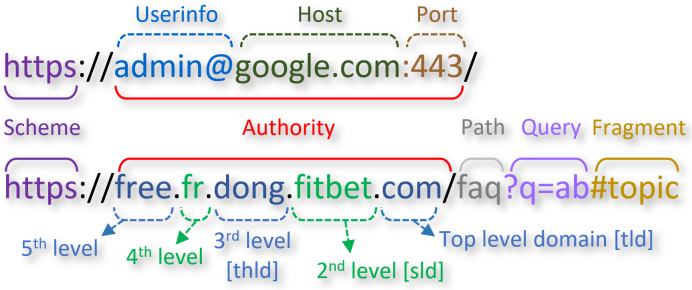


Fig. 1. Components and sub-components of the URI

This paper is structured as follows - section II. explains the purpose of the obfuscation techniques and lists the references of some of the obfuscation techniques described in this paper in the research. Section III. talks about data collection, filtering, and cleansing steps to prepare the dataset for the next step - section IV - which summarizes the most common obfuscation techniques and their collected statistics. Section V. analyzes the overlap between the discussed obfuscation techniques. Section VI. suggest next steps and potential future works linked to this paper and Section VII. provides a summary of the gathered results.

The primary objective of the analysis was to evaluate the prevalence of selected obfuscation techniques among the confirmed phishing web pages through the extended time period and uncover the real-world figures, Year-Over-Year changes, and historical and actual trends. These obfuscation techniques are strong indicators of phishing, as stated in the papers listed in the next section, and quantifying their prevalence among phishing webpages would allow the formulation of their importance or capacity to expose the phishing webpages.

979-8-3503-1720-6/24/$31.00 ©2024 IEEE

II. URL and domain obfuscation techniques

The use of obfuscation techniques on URLs or domains to commit a scam is a form of semantic attack [4]–[6]. Examples of other types of obfuscation techniques that are commonly used but can't be identified from the domain or URL are redirects (deployed on the client side via <meta> tag forcing refresh, javascript, or deployed on the server side); another example is QR codes.

Obfuscation techniques can serve diverse objectives, but the two most important ones are **evading detection** and/or **increasing credibility**, which are often coupled. In some scenarios, obfuscation techniques can improve on both; in others, they might counteract.

Puny code is an example of an obfuscation that positively impacts both objectives. The potential victim sees and might believe to be accessing the genuine domain. The chances of phishing detection are significantly reduced because one of the most common clues - domain or URL perception was passed, and the credibility of the currently visited domain increased.

An example of the opposite scenario, when improving one objective reduces the other, is replacing the domain name with an IP address. Using an IP address could help the attacker bypass the domain watch lists, but it might reduce credibility in the eyes of the victim when the URL is shown with IP in the browser's address bar. To counterbalance this negative impact, the attacker might deploy another technique - secured HTTP (HTTPS) to improve the site's credibility in the eyes of the user.

A. URL obfuscation techniques in the research

Different phishing detection-focused research papers have leveraged indicators of various obfuscation techniques reviewed in this analysis. For example, the at "@" as an indicator of phishing is used in [5], [7]–[9]. IP obfuscation technique as an indicator of phishing is the most referred phishing page feature and is used in [7], [9]–[12]. IP address formatted as a single decimal value was described in [5]. The presence of the port as part of the URL to identify a phishing page is used in [8], [10]. URL shorteners are addressed in [12], [13]. Finally, a combination of the techniques has been mentioned in [5]. As seen from the list of mentioned references, listed obfuscation techniques were commonly used across the defined period as indicators of phishing among the analyzed URLs.

In article [14], authors mentioned the calculated presence of selected descriptive characteristics, some of which are common with those analyzed by us, like - ports, IP presence and IP encoding and URL shorteners, though the provided numbers are difficult to compare due to data cleansing approach(de-duplication) which is not described in detail. Also, their data time window was ranging only from 2016 to 2021.

III. Phishing Domains Data

To analyze data over an extended period, it is prudent to use data from multiple sources. However, there are various phishing datasets [15] available online; most are limited to short periods only. The most comprehensive dataset, amongst the publicly available ones, was an archive of PhishMonger. It contains data spanning three years (2016-2018) and sourced from PhishTank [16], [17].

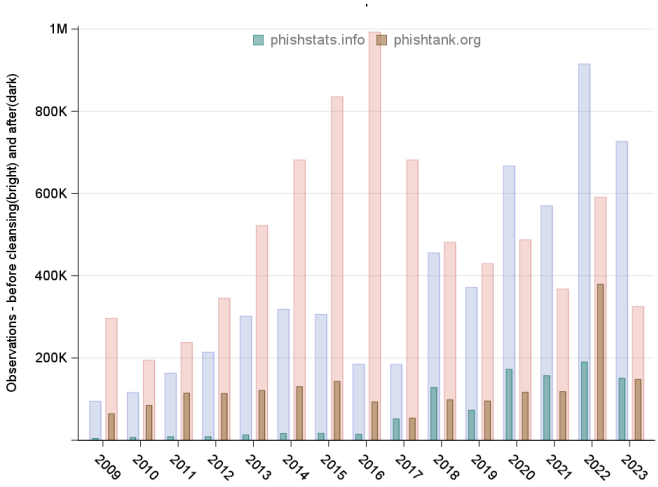


Fig. 2. Volumes in the combined dataset before and after data cleansing

Free or open available dataset of phishing data for multi-year periods is practically non-existent. The only way to collect such data was to acquire them from the owner (**PhishStats**[1]) or scrape them from available online sources(**PhishTank**[2]).

A. Data preparation and cleansing

As a result of the way the above-described sources collect the phishing records, the collected data had to be reviewed and cleansed. This process entailed:

- **Merge the data.** Both datasets - PhishTank and PhishStats were combined into one common data table with a common structure (unified column names, data types, and column lengths).
- **Filter the desired period.** Only data for the period between 1st‘ of January 2009 and 30st of September 2023 were selected. For the data from PhishTank, the 30th of September 2023 was decided as a cut-off date. For PhishStats data, the cut-off date is 9th of August 2023 due to technical issues on the PhishStats website and the inability to collect data beyond this date through the provided API. This consolidated dataset had ≈13M records, of which ≈7.4M from PhishTank and ≈5.6M from PhishStats ("Fig. 2", semi-transparent columns).
- **Remove duplicate records.** When deployed by the attacker, phishing attacks can be spotted and experienced by many potential victims, some of whom can report the phishing webpage to PhishTank or PhishStats. This process results in reporting the same domains multiple times and results in duplicates. To eliminate these duplicates, all the domains that have the same components of domain up to 5th-level subdomain (5 levels of domain granularity

[1]phishstats.info
[2]phishtank.org

were selected based on previous analysis of distribution and share of different levels of domain granularity within the collected phishing data [18]) and were reported within the 24h window from their first occurrence were removed as duplicates. This step removed a significant portion of the records, and the resulting dataset shrank to ≈5.2M records - ≈4.2M from PhishTank and ≈1M from PhishStats.

As seen on the "Fig. 2", many PhishStats records for 2009-2016 have been removed (dark blue column size vs. semi-transparent blue column size). This is because PhishStats sourced its phishing records almost exclusively from PhishTank during this period, and only in 2017 did it source additional data from other sources. Those new sources do not overlap with PhishTank's data. Data overlap analysis between PhishTank and PhishStats is detailed in [18].

- **Select only confirmed phishing records.** The last step meant keeping all data from PhishStats (as it publishes only confirmed phishing records) and only a subset of data from PhishTank. Original unfiltered data for the given 15-year period shrank further to ≈2.9M as ≈2.3M records from PhishTank were removed.

The final dataset of confirmed phishing URLs contained ≈**2.9M** records out of which ≈**1M** were from **PhishStats** and ≈**1.9M** from **PhishTank**.

IV. Obfuscation Techniques and Trends

The obfuscation techniques covered in this paper are all linked to the webpage URL. The occurrence of each obfuscation technique is represented using Year-over-Year volume statistics. The statistics were initially gathered only for filtered confirmed phishing data (as described in the previous section). Figures are visible in every summary table under the Source data = "Phishing" columns.

Later, we decided to add statistics of occurrences among the unconfirmed phishing data (data removed in the last step of the cleansing process). These data are visible in each summary table within the columns with grey-highlighted background color and under the heading of Source data = "Unconfirmed". These figures were added since some obfuscation techniques have had significant occurrences in data that were filtered out. The common premise regarding these obfuscation techniques is that they indicate phishing with very high accuracy; therefore, we believe that most of these records are confirmed phishing. Including these numbers would provide a more comprehensive picture of the prevalence of those obfuscation techniques.

A. Obfuscation using the at "@" sign

At sign "@" has a specific purpose in the URI as part of the authority component "Fig. 1". Part preceding the at "@" sign is a user information sub-component, which is used only rarely (due to security reasons - passing cleartext credentials) [4]. Nevertheless, using this sub-component can help the attacker to deceive the potential victim. An example of such an attack is

	Source data		
	Phishing		Unconfirmed
	"@" missing	"@" present	"@" present
Year			
2009	69 094	18	.
2010	91 601	1	2
2011	123 183	.	.
2012	122 698	4	5
2013	134 029	.	.
2014	146 783	.	1
2015	159 809	.	1
2016	107 925	2	2
2017	105 918	.	.
2018	226 983	.	1
2019	168 569	1	.
2020	289 270	2	10
2021	275 717	1	7
2022	569 637	2	2
2023	298 773	2	1

Fig. 3. Occurences of "@" sign

http://dhl.cz:0@www.dongfengcidef.cl, which tries to evoke the visiting **dhl.cz** domain, while in reality, the browser will navigate to a webpage hosted on **dongfengcidef.cl** domain. Reviewing the figures, the prevalence of this obfuscation technique is very rare, with almost only single-digit occurrences within the confirmed and unconfirmed phishing data. There is also no visible trend from the gathered data ("Fig. 3").

B. Obfuscation via HTML entities

HTML entities are easy to identify as they always begin with an ampersand "&" and end with a semicolon ";". There are two types:

- **Named HTML entities**, are most commonly used to display characters with special meaning in HTML like less-than sign "<" written as "<" used for the HTML tag opening or greater-than sign ">" written as ">" used for closing the HTML tag.
- **Numeric HTML entities**, which are used to express any character using the hexadecimal ("&#xHH;") or decimal format ("&#DD;"). For example character **"@"** can be expressed as **"@"** or **"@"**.

From the gathered statistics ("Fig. 4"), it is clear that HTML entities are also used sparsely, with very few occurrences among the confirmed phishing URLs and only 2-digit numbers within the unconfirmed phishing records. The search considered only those present as part of the domain (Authority), not within the path, query, or fragment ("Fig. 1"). YoY trends show that the numbers slowly increased from 15 (Unconfirmed phishing) in 2009 to almost 90 in 2018 and 80 in 2020. Since then, the figures have decreased to 35 in 2022 and even less in 2023.

C. Obfuscation by specifying port details

To make malicious URLs more convincing, attackers can use obfuscation techniques by explicitly mentioning the port

	Source data		
	Phishing		Unconfirmed
	"&" missing	"&" present	"&" present
Year			
2009	69 112	.	15
2010	91 602	.	11
2011	123 183	.	13
2012	122 702	.	17
2013	134 027	2	29
2014	146 782	1	24
2015	159 808	1	40
2016	107 927	.	21
2017	105 918	.	65
2018	226 976	7	88
2019	168 570	.	63
2020	289 272	.	76
2021	275 716	2	33
2022	569 638	1	35
2023	298 775	.	14

Fig. 4. Occurences of HTML entities

	Source data		
	Phishing		Unconfirmed
	":" missing	":" present	":" present
Year			
2009	68 700	412	539
2010	91 313	289	341
2011	122 619	564	388
2012	122 308	394	436
2013	133 737	292	267
2014	146 373	410	625
2015	159 611	198	410
2016	107 845	82	465
2017	105 855	63	338
2018	226 757	226	226
2019	167 393	1 177	774
2020	287 895	1 377	551
2021	275 470	248	379
2022	567 960	1 679	327
2023	298 466	309	325

Fig. 5. Occurences of the port presence

number right after the colon character ":" placed at the end of the host component (e.g., **http://google.com:80**) (“Fig. 1”). Another intent might be to make the URL look more complex and focus the user’s attention on the port part of the domain while ignoring the preceding domain part, which points to a malicious site. The last use case is targeting a firewall, which might be configured to filter out traffic passing through specific ports. Attackers can leverage non-standard ports to bypass such firewall rules. In some cases, the port colon was present, and the actual port number was omitted. There were only single-digit occurrences each year for such cases.

The figures of port occurrence are higher than those of previous obfuscation techniques, ranging from less than a hundred to almost 1700 in 2022. In general, 2022 stands out with nearly double the volume of records compared to 2021. Still, the volume of records with port details is almost 6x higher than in 2023. Significantly higher numbers are visible in recent history, specifically in 2019 and 2020. The number of port details among the "Unconfirmed" phishing records is spread around ≈400 but with no visible continuous trend “Fig. 5”.

Records with specified port numbers were further analyzed and grouped by

- **Port classification** - ports were grouped based on the usual purpose of the service assigned for a given port number [19]. The most common ports were identified among the groups listed in “Fig. 6”, and as would be expected, the majority of the ports were linked to common ports for HTTP/HTTPS (80, 81, 443, 8080, 8081, 8443, 8090, 8000).
- **Port ranges** - ports were grouped into three defined ranges: 1. Well-known ports, 2. Registered ports, and 3. Unassigned ports “Fig. 7”. Distribution was mainly between the first two groups due to the prevalence of ports linked to HTTP/HTTPS.

	Source data			
	Phishing		Unconfirmed	
	N	%	N	%
Port classification:				
1. Http/Https	2 551	34.7%	2 698	50.2%
2. Remote access/control	140	1.9%	145	2.7%
3. Kerberos	90	1.2%	54	1.0%
4. Web - cPanel	25	0.3%	32	0.6%
5. Trojan/Virus	975	13.3%	968	18.0%
6. Other	3 571	48.6%	1 474	27.4%

Fig. 6. Distribution of port classes

D. *Use of Punycode to mimic genuine domains*

Punycode is an encoding of a non-ASCII Unicode string into an ASCII string. It was defined in 2003 in RFC3492 [20]. The presence of Punycode can be identified through **"xn--" prefix** within the string. Intended regular use of the Punycode allows users to type a domain name into the browser’s address bar in their language-specific character set like Chinese, Cyrillic, and others. A Unicode string is translated using the Punycode encoding algorithm within the browser into an ASCII-compatible string, which is then sent to DNS to return the IP address of the requested domain. Punycode can be highly efficient for homograph attacks or brand spoofing by replacing certain ASCII characters in the domain with a non-ASCII Unicode character, which

	Source data			
	Phishing		Unconfirmed	
Port ranges:				
1. Well known ports (0-1023)	2 826	38.4%	3 008	56.0%
2. Registered port (1024-49151)	3 871	52.7%	2 329	43.4%
3. Unassigned ports (49152-65535)	655	8.9%	34	0.6%

Fig. 7. Distribution of port ranges

	Source data		
	Phishing		Unconfirmed
	"xn--" missing	"xn--" present	"xn--" present
Year			
2009	69 093	19	16
2010	91 591	11	12
2011	123 156	27	17
2012	122 647	55	31
2013	133 916	113	108
2014	146 663	120	324
2015	159 558	251	277
2016	107 831	96	310
2017	105 824	94	365
2018	226 453	530	348
2019	168 255	315	262
2020	288 690	582	445
2021	275 367	351	277
2022	569 190	449	376
2023	298 150	625	295

Fig. 8. Occurences of Punycode

	Source data					
	Phishing			Unconfirmed		
	Decimal	Hexadecimal	Octal	Decimal	Hexadecimal	Octal
Year						
2009	3 986	87	16	4 716	188	23
2010	4 481	53	4	2 246	31	4
2011	5 270	27	9	1 658	11	5
2012	5 966	6	1	2 756	9	7
2013	4 468	1	.	3 374	9	3
2014	3 882	2	.	4 654	10	1
2015	3 075	.	.	3 459	4	4
2016	1 459	.	.	6 092	6	13
2017	1 124	.	.	4 806	5	2
2018	2 547	.	.	3 229	3	.
2019	2 807	1	.	2 077	4	1
2020	5 838	.	.	2 751	14	6
2021	2 785	2	.	920	1	1
2022	7 027	.	1	894	.	3
2023	2 446	1	.	1 034	.	.

Fig. 9. Share of various numerical representations of IP

looks identical or very similar to actual ASCII characters. For example, URL **http://account.xn–googe-wsa.com/** which is presented as **http://account.googìe.com**, another example **http://app.xn–sshi-08a.tk/** is shown as **app.sụshi.tk**. The examples show that these character replacements are hard to spot, especially if the characters are carefully selected.
Though the Punycode was defined already in 2003, in the selected period (2009-2023), we observed very low occurrences in 2009 and 2010 (20 and 11, respectively) with a visible growth till 2018, since when the figures stabilized in the range of 300-500 cases annually among the confirmed phishing records "Fig. 8". Though the numbers are not as high as for the obfuscation using the ports, they are not negligible either, with overall ≈800-900 records each year since 2018 (confirmed and unconfirmed phishing records added together).

E. Obfuscation through IP address

Substituting the domain name with an IP address in the URL of a phishing web page is the most prevalent technique of URL obfuscation. The most common objective of such substitution is hiding the actual domain name - which might expose the phishing nature of the webpage to the potential victim. IP addresses can be represented in various notations:

- **IPv4 notation** - the most commonly used and known xxx.xxx.xxx.xxx where xxx is a number between 0-255, e.g., http://211.72.122.11/secured/index.htm
- **Single value notation** - IP is represented as a single value ranging from 0 to 2^{32}, e.g., http://1077629123/phpma/config/ (in IPv4: 64.59.80.195)
- **Hybrid notation** - IP is represented as a variation of the above two techniques, e.g., http://0x4a.0x361142/~cgipecom/www.irs.gov (which can be represented as http://74.3543362/ by converting hexadecimal values into decimal and which further translates into 74.54.17.66 in IPv4)

IP written in the above notations can also represent the numerical value in different formats. The most common are:

- **Decimal** - IPv4 notation example: http://66.147.240.156/~frpaypal/, single value notation example: http://1075516530:82/index.php and hybrid notation example: http://203.10654640:8080/.https/www.wellsfargo.com
- **Hexadecimal** - can be identified through specific prefix "**0x**". IPv4 notation example: http://0xd8.0xb6.0x6c.0x58/signin/, single value notation example: http://0xd2bb6e92/.b.php and hybrid notation example: http://0xa8.0xbb5ce5/vsp/form.html
- **Octal** - can be identified through a leading zero character "**0**". IPv4 notation example: http://0106.0125.0326.0102/www.poste.it/login.html, single value notation example: http://033113520761/start.jsp.htm and hybrid notation example: http://0125.027135477/aw/
- **Combined** - combines the above numerical formats, e.g., http://0x6b.026.0320.189/, which combines hexadecimal with two octal and one decimal formats within the IPv4 notation.

"Fig. 9" shows how decimal format is prevalent compared to hexadecimal or octal. Another notable observation is that many records with IP obfuscation are among the unconfirmed phishing data though the distribution between the numbers in confirmed phishing and unconfirmed phishing records in recent periods is shifting towards confirmed phishing records. The YoY numbers among phishing records don't show any visible trend, but when combined with numbers of unconfirmed phishing data, the average volume revolves around ≈7K records, with a visible decrease in 2021 and 2023. Looking at the % share of combined records with IP obfuscation each year, we see a steady decline from almost **12.2%** in 2009 to **1.2%** in 2023.

	Source data				
	Phishing				Unconfirmed
	missing		present		present
	N	%	N	%	N
Year					
2009	68 851	99.6%	261	0.4%	256
2010	90 953	99.3%	649	0.7%	309
2011	122 698	99.6%	485	0.4%	222
2012	122 051	99.5%	651	0.5%	265
2013	133 349	99.5%	680	0.5%	431
2014	145 951	99.4%	832	0.6%	979
2015	158 983	99.5%	826	0.5%	1 464
2016	107 506	99.6%	421	0.4%	1 830
2017	105 462	99.6%	456	0.4%	1 885
2018	225 714	99.4%	1 269	0.6%	932
2019	167 827	99.6%	743	0.4%	1 310
2020	288 325	99.7%	947	0.3%	1 063
2021	274 552	99.6%	1 166	0.4%	801
2022	568 493	99.8%	1 146	0.2%	920
2023	297 917	99.7%	858	0.3%	565

Fig. 10. Occurences of URL shorteners

	Source data			
	Phishing		Unconfirmed	
	N	%	N	%
Top 10 shorteners:				
bit.ly	1 889	16.6%	2 047	15.5%
tinyurl.com	1 865	16.4%	1 633	12.3%
bit.do	775	6.8%	961	7.3%
is.gd	709	6.2%	704	5.3%
cutt.ly	656	5.8%	423	3.2%
t.co	563	4.9%	1 060	8.0%
goo.gl	561	4.9%	1 142	8.6%
ow.ly	524	4.6%	1 220	9.2%
x.co	495	4.3%	758	5.7%
tiny.cc	477	4.2%	355	2.7%
Others	2 876	25.3%	2 929	22.1%

Fig. 11. Top 10 URL shorteners used

F. Prevalence of URL shorteners

URL shorteners were designed for convenience to simplify the sharing of longer URLs, but malicious actors started exploiting them to obfuscate phishing URLs. URL shorteners substitute a URL with a short hash code right after the link to the shortener's primary domain, e.g., **http://bit.ly/13mod8** or **http://tinyurl.com/ykplrqz**. There are hundreds of URL shorteners today (in our analysis, we identified more than 250). The number of occurrences among the confirmed phishing records revolves around ≈700 records, but high numbers are also occurring among the unconfirmed phishing records "Fig. 10". Combined numbers (confirmed and unconfirmed phishing records) gradually grew from 2009, reaching a peak in 2017 (≈**2.2%** of yearly records volume) and since then slowly descended to ≈**0.5% share in 2023**. "Fig. 11" lists the top 10 URL shorteners among the confirmed phishing data (the first three places are the same among unconfirmed phishing records).

	Source data			
	Phishing		Unconfirmed	
	Share %		Share %	
	http	https	http	https
Year				
2009	99.9%	0.1%	99.6%	0.4%
2010	99.8%	0.2%	99.1%	0.9%
2011	99.6%	0.4%	98.7%	1.3%
2012	99.4%	0.6%	98.5%	1.5%
2013	99.5%	0.5%	98.5%	1.5%
2014	99.3%	0.7%	98.1%	1.9%
2015	99.1%	0.9%	97.7%	2.3%
2016	98.2%	1.8%	96.0%	4.0%
2017	87.6%	12.4%	82.9%	17.1%
2018	71.7%	28.3%	66.9%	33.1%
2019	54.2%	45.8%	45.4%	54.6%
2020	47.9%	52.1%	39.3%	60.7%
2021	41.2%	58.8%	30.7%	69.3%
2022	59.3%	40.7%	27.1%	72.9%
2023	28.1%	71.9%	20.3%	79.7%

Fig. 12. Occurences of HTTP vs. HTTPS

G. Employing HTTPS to appear legitimate

The idea behind using HTTPS on phishing sites is to make it appear more legitimate in the eyes of the potential victim. By configuring the HTTPS on the server side, the visitor's communication between the local device (PC, mobile, etc.) and the server becomes encrypted instead of only HTTP cleartext communication, which can be eavesdropped on. Practically, HTTPS has no relevance regarding the potential phishing purpose of the hosted site or provides no risk mitigation in this regard.

As per "Fig. 12" the shift towards HTTPS is obvious and confirms what was presented by APWG in a report from Q2/2021 [21] - from Q3/2020 onwards more than **80%** of phishing pages were already set up with HTTPS. Our numbers show slightly lower figures - the most recent data in **2023 at ≈72%** among confirmed and ≈**78%** among unconfirmed phishing records.

V. Overlap of Obfuscation Techniques

Some of the listed URL obfuscation techniques work or impact different URL parts ("Fig. 1"), so multiple techniques can be deployed within the same phishing URL. A good example of combined obfuscation techniques and showcasing the benefits it provides is an URL **http://www.microsoft.com@2398855780** which will load the Google search page as the decimal part (2398855780) stands for IPv4: 142.251.162.100 which is one of Google's public IP addresses, although it might appear that the URL is pointing towards microsoft.com website.

"Fig. 13" depicts grouping all records into sets where each set employs the same combination of obfuscation techniques. Each group is also assigned a volume of records to show each combination's prevalence. The most common overlapping techniques are URL shorteners in combination with HTTPS,

#	"%2E"	"@"	"&"	":"	"xn--"	IP	SHORT	HTTPS	Obs.	%
1	○	○	○	○	○	○	○	○	1,903,386	65.8606%
2	○	○	○	○	○	○	○	●	910,150	31.4928%
3	○	○	○	○	○	○	●	○	7,406	0.2563%
4	○	○	○	○	○	○	●	●	3,984	0.1379%
5	○	○	○	○	○	●	○	○	50,214	1.7375%
6	○	○	○	○	○	●	○	●	3,466	0.1199%
7	○	○	○	○	●	○	○	○	1,999	0.0692%
8	○	○	○	○	●	○	○	●	1,620	0.0561%
9	○	○	○	●	○	○	○	○	2,846	0.0985%
10	○	○	○	●	○	○	○	●	1,161	0.0402%
11	○	○	○	●	○	●	○	○	3,674	0.1271%
12	○	○	○	●	○	●	○	●	14	0.0005%
13	○	○	○	●	●	○	○	○	2	0.0001%
14	○	○	○	●	●	○	○	●	17	0.0006%
15	○	○	●	○	○	○	○	○	8	0.0003%
16	○	○	●	○	○	○	○	●	6	0.0002%
17	○	●	○	○	○	○	○	○	26	0.0009%
18	○	●	○	○	○	○	○	●	2	0.0001%
19	○	●	○	●	○	○	○	○	2	0.0001%
20	○	●	○	●	○	○	○	●	3	0.0001%
21	●	○	○	○	○	○	○	○	31	0.0011%
22	●	○	○	○	○	○	○	●	3	0.0001%
23	●	○	○	○	○	●	○	○	1	0.0000%
24	●	○	○	●	○	○	○	●	1	0.0000%
									2,890,022	100%

Fig. 13. Obfuscation techniques overlap and volumes matrix

but it is not a combination of obfuscation but rather a characteristic of URL shorteners domains, which all are configured to use HTTPS. The actual most common combination of techniques is HTTPS and IP address. The most common combination without HTTPS is the use of an IP address along with a port. The remaining combinations are extremely rare (single-digit occurrences).

VI. Future work

This analysis focused on the prevalence of different obfuscation techniques as leveraged by threat actors by analyzing historical phishing URLs. A study or survey examining users' susceptibility to these techniques could be a baseline for comparing the techniques used and their efficiency when employed.

Another branch of future research could focus on non-URL related obfuscation techniques used for phishing attacks like:

- use of QR codes
- Page redirects
- Phishing page content obfuscation

Conclusions

If we ignore HTTPS as a relevant indicator of phishing (as most of today's websites use secured connections), then less than 3% of all confirmed phishing pages across 2009-2023 employ at least one of the obfuscation techniques described. This number might appear low or even negligible, but the factor that makes these techniques interesting is users' susceptibility to them.

Reviewing the four most common techniques - IP address, URL shorteners, port, and Punycode - all show up as most relevant precisely because of their capability to increase the chances for the malicious actor. And increasing chances, even by a small margin, might make a big difference in the overall efficacy of the phishing campaign. From the gathered statistics and ordering the techniques by their occurrence, we can indirectly assume that the order also represents their efficacy when employed by threat actors.

Acknowledgment

It was supported by the Erasmus+ project: Project number: 2022-1-SK01-KA220-HED-000089149, Project title: Including EVERyone in GREEN Data Analysis (EVERGREEN) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Slovak Academic Association for International Cooperation (SAAIC). Neither the European Union nor SAAIC can be held responsible for them.

References

[1] Phishing activity trends report 4th quarter 2022. Anti-Phishing Working Group. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2022.pdf

[2] Internet Crime Report 2022. FBI's Internet Crime Complaint Center. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

[3] Internet Organised Crime Threat Assessment (IOCTA) 2023, Publications Office of the European Union, Luxembourg. Europol (2023). [Online]. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/IOCTA%202023%20-%20EN_0.pdf

[4] T. Berners-Lee, R. Fielding T., and L. Masinter M, "Uniform Resource Identifier (URI): Generic Syntax," Internet Requests for Comments, RFC Editor, RFC 3986, January 2005. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3986.txt

[5] R. Siedzik. (2001, April) Semantic Attacks – What's in a URL? SANS Institute. [Online]. Available: https://www.giac.org/paper/gsec/650/semantic-attacks-url/101497

[6] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Computing Surveys*, vol. 48, pp. 1–39, 02 2016.

[7] Y. Zhang, J. Hong, and L. Cranor, "Cantina: A content-based approach to detecting phishing web sites," in *Proceedings of the 16th international conference on World Wide Web*, 05 2007, pp. 639–648.

[8] R. Gowtham and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," *Computers & Security*, vol. 40, p. 23–37, 02 2014.

[9] R. Verma and K. Dyer, "On the character of phishing urls: Accurate and robust statistical learning classifiers," *CODASPY 2015 - Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, pp. 111–121, 03 2015.

[10] A. Le, A. Markopoulou, and M. Faloutsos, "Phishdef: Url names say it all," *Proceedings - IEEE INFOCOM*, 09 2010.

[11] S. Garera, N. Provos, M. Chew, and A. Rubin, "A framework for detection and measurement of phishing attacks," *WORM'07 - Proceedings of the 2007 ACM Workshop on Recurring Malcode*, 11 2007.

[12] S. Marchal, J. Francois, R. State, and T. Engel, "Phishstorm: Detecting phishing with streaming analytics," *IEEE Transactions on Network and Service Management*, vol. 11, pp. 458–471, 12 2014.

[13] D. Sahoo, C. Liu, and S. Hoi, "Malicious url detection using machine learning: A survey," *arXiv preprint arXiv:1701.07179*, 01 2017.

[14] J. Barros, C. Silva, L. Teixeira, B. Fernandes, J. Oliveira, E. Feitosa, W. Dos Santos, H. Arcoverde, and V. Garcia, "Piracema: a phishing snapshot database for building dataset features," *Scientific Reports*, vol. 12, 09 2022.

[15] K. L. Chiew, E. Chang, C. C. L. Tan, J. Abdullah, and K. Yong, "Building standard offline anti-phishing dataset for benchmarking," *International Journal of Engineering & Technology*, vol. 7, no. 4.31, pp. 7–14, 12 2018.
[16] D. Dobolyi and A. Abbasi., "Phishmonger live phishing website collection for azphish web, university of arizona artificial intelligence lab," 9 2016. [Online]. Available: https://www.azsecure-data.org/
[17] D. Dobolyi and A. Abbasi, "Phishmonger: A free and open source public archive of real-world phishing websites," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, 09 2016, pp. 31–36.
[18] I. Skula and M. Kvet, "Domain blacklist efficacy for phishing web-page detection over an extended time period," in *Proceeding of the 33rd Conference Of FRUCT Association*, 05 2023, pp. 257–263.
[19] Ports Database. SpeedGuide.net. [Online]. Available: https://www.speedguide.net
[20] A. M. Costello, "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)," Internet Requests for Comments, RFC Editor, RFC 3492, March 2003. [Online]. Available: https://www.rfc-editor.org/rfc/rfc3492.txt
[21] Phishing activity trends report 2nd quarter 2021. Anti-Phishing Working Group. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2021.pdf