Proceedings of

The 33rd Conference of Open Innovations Association FRUCT

Žilina, Slovakia
24-26 May 2023

UNIVERSITY OF ŽILINA

IEEE

IEEE ComSoc IEEE Communications Society

ORACLE

BRAIN:IT

electronics an Open Access Journal by MDPI

inventions an Open Access Journal by MDPI

Proceedings of the 33rd Conference of Open Innovations Association FRUCT

Žilina, Slovakia, 24-26 May 2023

Organized by

FRUCT Association
University of Žilina

Technically sponsored by

The conference patrons:

33rd Conference of Open Innovations Association FRUCT:
Publisher: FRUCT Oy (Finland), 2023. 444 p.

ISBN 978-952-69244-9-6
ISSN 2305-7254 e-ISSN 2343-0737

This proceeding includes the papers of the following topics:

- Artificial Intelligence in Text Analysis and Generation
- Artificial Intelligence, Robotics and Automation
- Coding Theory, DevOps and DevSecOps Technologies
- Emerging Wireless Technologies, 5G and beyond
- Internet of Things: Apps and Enabling Technologies
- Gamification, E-learning and Smart Data in Education
- Commercialization of Technologies and Digital Economy
- Location Based Services: Navigation, Logistics, Tourism
- Wearable Electronics: Novel Architectures and Solutions
- Natural Language Processing and Speech Technologies
- Big Data, Knowledge Management, Data Mining Systems
- Cloud, Fog and Edge Computing and Engineering, HPC
- Predictive Analytics, Probability and Statistics
- Audio Pattern Recognition, Semantic Audio
- Computer Vision, Image & Video Processing
- Crowdsourcing and Collective Intelligence
- Software Design, Innovative Applications
- Blockchain Technology and Applications
- Artificial Intelligence Applications
- Intelligence, Social Mining and Web
- Smart Systems and Embedded Networks
- Networks and Applications
- e-Health and Wellbeing
- Security and Privacy
- Algorithms and Modeling
- Workshop: The DataWorld

The reports were present at the 33rd Conference of Open Innovations Association FRUCT held on 24-26 May 2023 in Žilina, Slovakia

The editor-in-chief: Dr. Sergey Balandin
The associate editors: Michal Kvet and Tatiana Shatalova

ISSN 2305-7254
e-ISSN 2343-0737

ISBN 978-952-69244-9-6

Proceedings

33rd Conference of Open Innovations
Association FRUCT

Žilina, Slovakia
24-26 May 2023

Organization Committee of the 33rd Conference of Open Innovations Association FRUCT

Local Chair: Michal Kvet ***Publishing team leader:*** Tatyana Shatalova

FRUCT President: Sergey Balandin

Program Committee

Albert Abilov
Guntis Arnicans
Ivaylo Atanasov
Serena Baiocco
Sergey Balandin
Ekaterina Balandina
Sergey Bezzateev
Ankur Bist
Iurii Bogoiavlenskii
Ales Bourek
Doina Bucur
Tien-Fu Chen
Vladimir Deart
Mario Döller
Adam Dudáš
Roman Dunaytsev
Jan-Erik Ekberg
Pumudu Fernando
Dieter Fiems
Andrey Fionov
Alexander Geyda
Philip Ginzboorg
Boris Goldstein
Oleg Golovnin
Marco Grossi
Andrei Gurtov
Grigory Kabatiansky
Carlos Kamenski
Alexey Kashevnik
Lazhar Khriji
Vladimir Khryashchev
Athanasios Kiourtis
Olga Kolesnichenko
Mikhail Komarov
Georgy Kopanitsa
Dmitry Korzun
Kirill Krinkin
Kirill Kulakov
Nadezda Kunicina
Andrey Kuzmin
Miroslav Kvassay
Michal Kvet
Marek Kvet
Ksenia Lagutina
Rustam Latypov
Sergey Listopad
Andrei Lobov
Hsi-Pin Ma
Anton Makarov
Anna Maltseva
Oleg Medvedev
Alexandrov Mikhail
Dmitry Mouromtsev
Dmitry Namiot
Anand Nayyar
Victor Netes
Marina Nikitina
Stavros Ntalampiras
Valentin Olenev
Martin Omana
Giuseppe Pace
Michele Pagano
Ilya Paramonov
Kiran Patil
Evelina Pencheva
Maria Elizabeth Pereira
Edison Pignaton de Freitas
Konstantin Platonov
Jari Porras
Jenni Rekola
Joel Rodrigues
Kurt Sandkuhl
Vladimir Sayenko
Anton Shabaev
Manoj Sharma
Tatyana Shatalova
Liudmila Shchegoleva
Tatiana Sherstinova
Nikolay Shilov
Maria Skvortsova
Alexander Smirnov
Manfred Sneps-Sneppe
Sergey Staroletov
William Steingartner
Elena Suvorova
Takeshi Takahashi
Sandeep Tamrakar
Naser Tarhuni
Nikolay Teslya
Timofey Turenko
Frane Urem
Andrey Vasilyev
Vladimir Vinnikov
Fabio Viola
Adeesha Wijayasiri
Lenis Wong
Hao Yu
Michal Zabovsky
Victor Zakharov
Victor Zappi
Mark Zaslavskiy
Yunpeng Zhang
John Zhang

Preface of the 33rd Conference of Open Innovations Association FRUCT

On behalf of the organizing team, I warm welcome you to the 33rd Conference of Open Innovations Association FRUCT. This year, the conference is hosted by the University of Žilina. The FRUCT33 conference embraces a hybrid format, combining onsite participation in Žilina, Slovakia, with online engagement through MS Teams.

Building upon a rich legacy of fostering enduring academic and business collaborations, the FRUCT conference has consistently been at the forefront of innovation. The program for this conference comprises 10 sessions, featuring 3 keynote talks, an invited talk, demo and poster section, the 6th DataWorld workshop, and the Oracle day. Spanning three days, the conference program accommodates both onsite and online participants. The first one and a half days of the conference (May 24-25, 2023) are primarily dedicated to attendees present in person. However, the latter half of the second day and the entire third day (May 26, 2023) are reserved for online sessions. Consequently, the conference proceedings are tailored to optimize the experience for both onsite and online participants.

For the onsite portion of the conference, we will adhere to the traditional format of presentations. Furthermore, the onsite sessions will be live streamed via MS Teams, ensuring that remote participants can also benefit from these sessions. As for the online component, all presentations have been pre-recorded by the authors and uploaded to YouTube. The conference program includes links to individual presentations as well as playlists encompassing all talks for each section. To manage your participation effectively, please consult the conference program brochure, which can be downloaded from www.fruct.org/program33.

The online conference sessions consist of two modules. Firstly, we encourage you to watch the pre-recorded presentations on YouTube. The conference program provides playlists for each session along with links to the individual presentations. Secondly, the sessions feature a question-and-answer segment, during which attendees can interact with the authors of the papers presented. These Q&A sessions will take place on MS Teams. We kindly request all paper authors to join their respective Q&A sessions and respond to questions from the conference attendees. The conference program includes designated time slots and corresponding MS Teams links for these sessions. We encourage you to allocate time beforehand to watch the relevant videos. Additionally, we invite you to provide feedback to the conference authors through likes, dislikes, and comments on the YouTube videos. We also encourage you to subscribe to the FRUCT channel for updates and future content.

We are proud to announce that the conference is technically sponsored by IEEE. All conference papers have undergone rigorous peer reviews. Full papers were selected based on stringent criteria, including research quality, paper length, structure, format, and other formal requirements. Each full paper submission was reviewed by at least three expert peers, and acceptance was granted only to those that received positive review comments. Authors were given the opportunity to address all review comments or provide compelling justifications if they chose not to implement specific suggestions. The second volume of the conference proceedings accommodates all other accepted submissions that were not classified as full papers and were not submitted to IEEE Xplore. This partitioning of the proceedings ensures that the highest quality FRUCT publications can undergo proper international indexing and be published in renowned databases such as Web of Science.

We are delighted to present the proceedings of the 33rd Conference of Open Innovations Association FRUCT. With a total of 104 conference submissions, we are proud to announce that 42 papers have been accepted for publication as full papers, resulting in a commendable conference acceptance rate of 40%.

Once again, we extend our warmest welcome to all participants and express our gratitude to the University of Žilina for hosting the FRUCT33 conference. We hope that the ensuing discussions, presentations, and interactions will inspire new avenues of open innovation and contribute to the advancement of research and industry collaboration.

The accelerating pace of innovation and the increasingly shorter lifespan of commercially viable technologies pose unique challenges for the IT and ICT industries. Fierce competition among market players and rapid technological progress fueled by extensive investments in research and development necessitate a proactive response from educational and research institutions worldwide. The FRUCT community strives to foster cooperation and cultural exchange, supporting regional teams in effectively aligning university research and education with industrial challenges. Our primary mission is to strengthen collaboration within the academic community, enhance the visibility of research teams, and facilitate direct personal connections between academic and industrial experts.

The FRUCT conference embodies the principles of continuous development and strategic partnerships between industrial and academic research, which serve as crucial factors for success in the modern innovation ecosystem. Throughout the world, there exist remarkable success stories of such frameworks, which yield significant benefits for all involved parties, fueling their respective research and development endeavors. While fundamental science driven by universities and academic organizations should not be tethered directly to existing industries, industrial research greatly benefits from early access to results and information on emerging trends and weak signals. Likewise, many universities actively engage in applied research, but to maximize their efficiency, they require feedback channels from the industry. Thus, establishing stronger connections between academia and industry is pivotal, especially given the shrinking innovation cycles discussed earlier. An intriguing new trend to address this need involves constructing open innovation frameworks specifically designed to develop strategic partnerships between industrial and academic research, enabling the identification of suitable research partners and facilitating collaborative incubation of new competencies.

The FRUCT association is actively working to involve students and postgraduates in scientific activities at an early stage, fostering joint teams to tackle challenging scientific problems using knowledge-intensive technologies, and elevating the prestige of scientific and research work. Through the development of various processes, FRUCT supports win-win cooperation and the advancement of strategic partnerships between academic and industrial research. These processes serve to overcome barriers to open innovation, demonstrating how businesses can embrace social responsibility and contribute to long-term research and academic collaborations.

The FRUCT conference stands as a significant event celebrating academia-to-industry cooperation. With over 100 participants representing 26 countries, the 33rd FRUCT conference promises to be a vibrant gathering. Additionally, we anticipate that the presentations on YouTube will garner at least tenfold more views, extending the reach and impact of the conference beyond its physical boundaries.

The primary topics of the FRUCT conference are as follows:

- Artificial Intelligence in Text Analysis and Generation
- Artificial Intelligence, Robotics and Automation
- Coding Theory, DevOps and DevSecOps Technologies
- Emerging Wireless Technologies, 5G and beyond
- Internet of Things: Apps and Enabling Technologies
- Gamification, E-learning and Smart Data in Education
- Commercialization of Technologies and Digital Economy
- Location Based Services: Navigation, Logistics, Tourism
- Wearable Electronics: Novel Architectures and Solutions
- Natural Language Processing and Speech Technologies

- Big Data, Knowledge Management, Data Mining Systems
- Cloud, Fog and Edge Computing and Engineering, HPC
- Predictive Analytics, Probability and Statistics
- Audio Pattern Recognition, Semantic Audio
- Computer Vision, Image & Video Processing
- Crowdsourcing and Collective Intelligence
- Software Design, Innovative Applications
- Blockchain Technology and Applications
- Artificial Intelligence Applications
- Intelligence, Social Mining and Web
- Smart Systems and Embedded Networks
- Networks and Applications
- e-Health and Wellbeing
- Security and Privacy
- Algorithms and Modeling
- Workshop: The DataWorld

We extend our gratitude to all the authors, reviewers, and participants who have contributed to the success of this conference. The special words of thanks go to the local organizing team and especially Michal Kvet, who despite all problems and obstacles managed to organize this conference. I wish to thank all people who contributed efforts and a lot of personal time to the organization of the FRUCT conference, and all members of the organizing committee and FRUCT Advisory Board for reviewing the papers and other forms of contribution to the success of the 33rd FRUCT Conference. I hope that the proceedings and the ensuing discussions will inspire fruitful collaborations, foster innovative solutions, and drive further advancements in the field of open innovations.

May 2023

Sergey Balandin
FRUCT President

TABLE OF CONTENTS

Volume 2

Index

Domain Blacklist Efficacy for Phishing Web-page Detection Over an Extended Time Period

Ivan Skula, Michal Kvet
University of Zilina
Zilina, Slovakia
skula@dobraadresa.sk, michal.kvet@fri.uniza.sk

***Abstract*—Phishing domains and web pages are the most common techniques cybercriminals use and a backbone of social engineering techniques causing tremendous losses globally. A domain blacklist is one of the oldest techniques used for phishing detection and has been superseded by more modern and more accurate techniques - in practice and research. Analysis which was conducted using the 10-year phishing data from 2013 to 2022, collected from PhishTank and PhishStats websites, was aimed to calculate and assess the domain blacklist efficacy in capturing phishing web pages during this time period and for the future. The complete process consisted of data collection and consolidation - merging the data from both sources, data cleansing, and blacklist creation, followed by the analysis to calculate and collate the figures and observations. The last step was to review the gathered results and summarize the conclusions. The results show that only a small portion of the phishing domains (≈22%) re-occur and therefore are an eligible target of blacklist detection. Though, this is not a negligible number, especially when between ≈6% and ≈62% of records (from PhishTank) found in the blacklist were previously unclassified. A casual look at more recent trends doesn't provide a lot of supportive arguments in favor of blacklist as a future-proof technique either. However, the increased use of newly registered domains proves that cybercriminals must tap into the pool of new domains as current solutions utilizing blacklists effectively eliminate the re-used domains.**

I. Introduction

Phishing is the number one cybercrime type by number of victims annually [1]. It was continuously at the top for the last four years and has such a great lead over the remaining types that it will retain its position for the foreseeable future. Although the same report recorded last year a slight decrease in the number of victims (from ≈323K to 300K), another report [2] states a 61% increase in phishing attacks in 2022 when compared to 2021. Report's data for Q1-Q3/2022 [3] also indicate a growing trend. Finally, both our datasets - from PhishTank and PhishStats - show a substantial increase in phishing attempt volumes. And so, although the first phishing attacks were observed more than three decades ago [4], no perfect solution has been found yet.

Domain blacklist was among the first detection techniques used against phishing, usually as part of the web browser. This is still true today, as all commonly used modern browsers carry a highly accurate phishing detection functionality [5], [6]. Already in late 2004 there were criminal groups focusing on phishing attacks like the known "Rock Phish" group which employed single-use URLs. This approach caused concern among security professionals as it bypassed the majority of existing anti-phishing solutions which relied on URL lists [7]. Current research in phishing detection leans more towards techniques like predictive analytics and machine learning, which were proven to be highly accurate [8] and unlike domain blacklist can assess also never before seen domains; however, supplementing these techniques with a blacklist to achieve even a marginal gain would practically be translated into significant financial as well as non-financial savings due to number of the impacted victims globally.

A. Blacklist in the phishing research

Phishing techniques are constantly evolving to exploit the existing, newly found, or newly created gaps in the technology (e.g. URL obfuscation techniques to bypass standard rules-based detection systems), utilizing free web hosting, free blog sites, or web-based storage followed by the more recent use of public cloud infrastructure or AI-created targeted phishing messages. Research shows that blacklisting as a primary technique for phishing detection is a thing of the past. Domain blacklist has two inherent characteristics which limit its use or efficacy:

- it can't assess never before seen domain
- it requires another classification technique to support updating the blacklist

The first point directly impacts the efficacy of the domain blacklist. If a ratio of re-occurring phishing domains can be identified it would be possible to formulate the theoretical maximum efficacy of a domain blacklist. If only a fraction of domains are re-occurring, then only this fraction of domains can be fed into the blacklist assessment and be properly evaluated as phishing or not.

The second point is critical for updating a blacklist. Domain blacklist needs to store previously seen phishing domains. There are two main assessment approaches

- human-driven
- automated/machine-driven

The human-driven approach requires a human to decide and flag the domain as phishing or not. The machine-driven approach most commonly leverages classification machine learning algorithm [9], [10].

Another characteristic linked to blacklist, domains and impacts the blacklist efficacy is the assumption that once

the domain is observed as genuine, it will remain safe from phishing Or the opposite - the domain once observed as involved in phishing will always host malicious content. Such a simplified view does not apply to the phishing domains or domains in general; therefore, the efficacy of phishing detection based purely on a blacklist will be less accurate than other more advanced or combined approaches. This is also a reason why ambiguous domains - domains that were flagged as False-Positives(FP) during the analyzed time period - were removed from the dataset from the moment when they were identified as FP and added to the graylist.

Areas of focus: The primary objective of the analysis was to assess the efficacy of the phishing blacklist over an extended period, but other questions were formulated along with this objective.

1) Efficacy of blacklist in fraud detection
   a) If each newly identified phishing domain is added to the blacklist, what portion of phishing attempts could be detected via this blacklist year by year?
   b) Is the YoY efficacy of the blacklist increasing, decreasing, or steady?
2) Prevalence of domain re-use for phishing
   a) What is the ratio of phishing pages hosted on unique domains?
   b) How often is the phishing domain re-occurring?
3) Time periods related to domain re-occurrence
   a) What is the usual period before the domain first re-occurs?
   b) What is the usual time period between the re-occurrence of the domain?

For any of the above questions, we were not able to find any direct or even indirect answers in the published research.

II. Phishing domains data

Data for analysis were obtained from two websites collecting reported phishing pages from diverse sources - **PhishTank**[1] and **PhishStats**[2]. The PhishTank data was collected using a custom build web-scrapping tool. PhistStats data were obtained as a complete historical database, and more recent data were collected through web-scrapping. Volumes of data used in the analysis - all confirmed phishing records as well as genuine ones - by Year and source are visualized in "Fig. 1". Presented data volumes already reflect the data cleansing operations described below on each dataset from both sources individually while considering five levels of the domain name. The overall volume of records (bright-colored bars) from PhishTank was generally higher than from PhishStats. When counting only confirmed phishing records (dark-colored bars), the volume reversed in favor of the PhishStat.

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

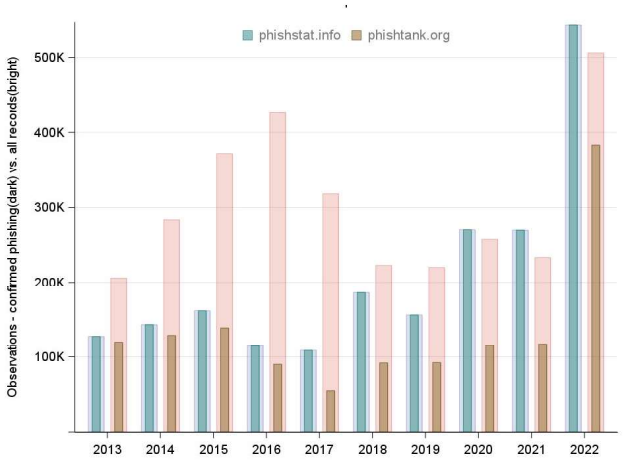

Fig. 1. Data volumes in separate source data sets after data cleansing

A. PhishTank

PhishTank's reported phishing data are manually evaluated by the people registered on the site. The final decision is always based on multiple evaluators' opinions. There are three different states in which the record of the reported phishing can be:

- True Positive (TP) - Phishing,
- False Positive (FP) - Non-Phishing
- Unknown (UNK) - assessment didn't conclude, or a final decision couldn't be achieved

B. PhishStats

All records present in the database were considered confirmed phishing web pages (TP).

III. Dataset preparation

A. Analysis of data overlap between PhishTank and PhishStats

Since the data originate from multiple sources and capture the same event - a phishing attempt - an analysis of data overlap was required. The objective of the analysis was to understand whether there is an overlap between the two datasets and, if yes, then to what extent. This was especially important as there are only a few free sources of phishing data available to the general public in a consolidated manner, like in the case of PhishTank and PhishStats.
The initial analysis considered only the confirmed phishing records as only those were relevant for the blacklist creation and assessment. Each dataset was divided into separate monthly parts and de-duplicated, so each month-part contains the domain only once. The overlap was calculated by comparing the full domain names lists (without the scheme and path parts as depicted in "Fig. 3") within these monthly parts of each data source.
Similar overlap analysis was conducted in [11], though they used PhishTank and OpenPhish as a source of data, and the period was limited to 75 days between March and June 2019.

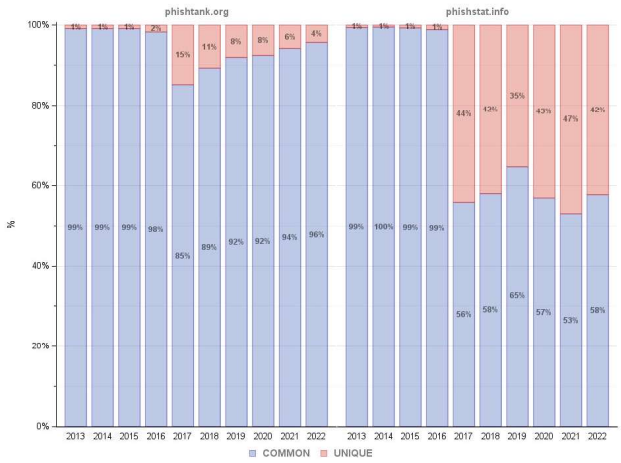

Fig. 2. % share of data overlap between datasets

Also, the focus of their analysis was on the persistence of the blacklisted URLs and didn't cover blacklist efficacy. Also in [12] authors conducted an overlap analysis along with comparative and descriptive analysis of 14 selected blacklists but PhishTank and PhishStats were not considered.

The results ("Fig. 2") from the perspective of PhishTank data showed that almost all records from PhishTank data are present also in the PhishStats dataset with a visible drop (gap increased to ≈15% from previous ≈1%), which happened in 2017 and lasted till 2022 (while slowly closing down to ≈4% in 2022). Further checking the recorded date and time of the overlapped records showed that both datasets had the same date and time, meaning that PhishStats was possibly loading Phishtank data into its database.

From the PhishStats perspective, the data show that in the early years (2013 - 2016) PhishStats data were almost identical to Phistank's confirmed phishing data, and only starting from 2017 some additional sources were added. As the analysis was done with only confirmed phishing records in Phishtanks dataset, we performed an expanded analysis to confirm whether these extra data are also not sourced from PhishTank(as the FP or UNK records). In this expanded analysis, all records for PhishTank and PhishStats were considered. Compared to the initial analysis, the volume of additional records in PhishStats was lower than the numbers(≈40%) shown in the initial analysis (see top right part of the "Fig. 2"). Still, the analysis confirmed that additional sources of phishing incidents were added in this period (2017-2022), providing data, that were not present in the PhishTank dataset.

This level of overlap - especially in the early years of 2013-2016 but also later - would practically duplicate all PhishTank records and finally skew the results. We performed further data cleansing and filtering (described in the next section), which addressed this overlap.

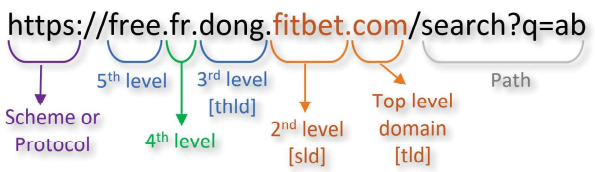

Fig. 3. URL and Domain name components

B. Data transformations and pruning

The next step in the data preparation phase was data profiling. This exercise provided insights on certain data groups which needed to be removed as they were irrelevant to the analysis or might skew the results. Ultimately after all the pruning steps the dataset was left with only ≈36% of the original data.

1) Domains with invalid top and second level domains: This filter kept only domains with a valid predefined Top Level Domain (tld) and non-missing 2nd Level Domain (sld). No condition was applied for 3rd (thld), 4th and 5th domain level names ("Fig. 3"). The most crucial part of the domain name is the second-level domain and top-level domain - in our example "fitbet.com". This is what domain registrars are allowing companies and individuals to purchase and register with them and use them for their intended purposes. Whoever registers a given domain, can create further subdomains within this purchased domain. This step decreased the size of the overall dataset by 0.69%.

2) Obfuscated domain via IP: URLs with the IP address in decimal or hex format were removed. This step shrank the dataset by 2.21%.

3) Obfuscated domain via URLs shorteners: records that were using URL shorteners (e.g. bit.ly, goo.gl, tinyurl.com, ow.ly, and others) were removed. Overall 1.11% of records were removed by this step.

4) Ambiguous domains from phishing perspective: using PhishTank's data, we analyzed the records, which were resolved as non-phishing(FP). All domains of these records were added to the Graylist with the date when they were identified as FP and then removed from the analysis. Further analysis of these records showed that many domains were hosted on free web-hosting domains (000webhostapp.com, weebly.com, duckdns.org, etc.) or free blogging sites (blogspot.com, medium.com). Because these domains have been reviewed and classified as non-phishing sites, they can't be flagged as TP or FP without an inaccuracy resulting from such a generalization. This operation decreased the size of the dataset by another 3.36%.

5) Domains with extremely high occurrence rate: Significant volume of domains have appeared in the dataset once(≈35%) or two times(≈35%). Less than 1% of domains have appeared in the dataset more than 26 times. Some domains that appeared thousands of times were flagged as outliers. In the final dataset, we removed all domains which appeared in the dataset more than 58 times (occurrence rate

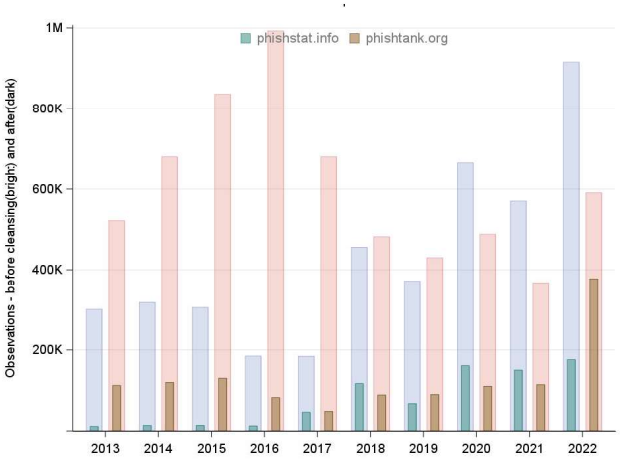

Fig. 4. Data volumes in the combined dataset before and after data cleansing

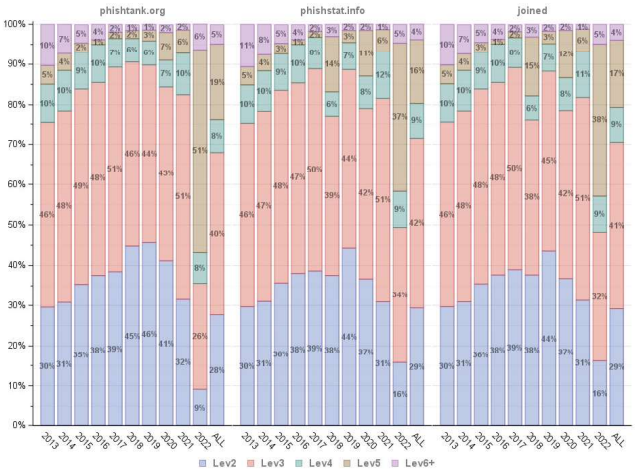

Fig. 5. Ratio of different levels of domain names

outside of 99.73% interval) which constituted 12.35% of records.

6) Same-attack duplicates: due to the way the phishing attacks are captured, it is common for the same phishing domain to be reported by multiple sources shortly after the phishing attack is initiated. This results in multiple duplicate records in the dataset recorded within a short time window (minutes and hours). Such records would skew the results of the analysis. Therefore a de-duplication (deletion) of the records of the same domain(meaning the domain with the identical five levels of domain names, as described in the section below) that appeared within 24 hours from the first occurrence was performed. Another supportive reason for this filtering is the data overlap between the data from PhishTank and PhishStats, as described in the section above. Removing the same-domain records within 24h eliminated duplicate records captured in both datasets. Only duplicate records within 24h window were removed, and the remaining ones were left though some of those appearing within the next few days could have referred to the same phishing attack too. As anticipated (“Fig. 2” for the scale of overlap), this step had the biggest impact on the final size of the dataset - 45.44% records were removed. Volumes of data in the joined table before and after cleansing can be seen in “Fig. 4”.

C. Domain granularity level

As hinted in the above analysis, it was critical to decide on the most appropriate level of granularity for domain names. This decision impacted not only the blacklist creation and consecutive analysis but also some of the analyses described above (e.g. overlap analysis and analysis of ambiguous domains). We performed variants of these with diverse domain levels (2-levels, 3-levels, and finally, 5-levels).

As can be seen in “Fig. 3”, domains can have multiple subdomains, each separated by a dot (”.”), but an overall length can’t exceed 253 characters (transmitted as a 255-octet packet) [13]. In our example - https://free.fr.dong.fitbet.com - the domain consists of five levels. Starting from right to left - the top-level domain ”com” (tld), followed by the ”fitbet” as a second-level domain (sld), then ”dong” as a third-level domain (thld), followed by ”fr” as a fourth-level domain and concluding with ”free” as a fifth-level domain.

To arrive at the most appropriate level, we calculated the prevalence of different levels of domain names in the underlying data (considering only confirmed phishing domains). In the analysis, we calculated the % share of each level of domain granularity to understand how common each level is across the analyzed period. The analysis was conducted on both datasets separately and on the joint dataset.

Based on the data - though there are slight differences between the two datasets - initially, most domains were within 2 and 3 levels (joint share even increased from ≈76% in 2013 to ≈89% in 2017). Since 2019 the share of 2-level domains started to decrease, and we observe an increase in domains of level 4 and more. A significant shift is visible in the last year (2022), where the sudden increase in domains with five levels and more is unlike any YoY change seen before (“Fig. 5”). After further review of the data, we identified a sharp increase in the number of different subdomains registered to the same domain (sld.tld) in this period. The average number of different subdomains linked to the same domain was constantly below 2 throughout the whole period except the last year where it almost doubled (an average of 3.7 different sub-domains linked to the same 2-level domain as opposed to less than 2 in the previous years). These results correspond with the findings of other researchers [2], who also observed a significant increase(+83%) in newly registered domains in 2022 compared to 2021.

As per the data, storing the domain name with a maximum of 3 levels (e.g. www.google.com) would provide only ≈70% accuracy (joined dataset for the whole 10-year period). More

recent data (from 2019 to 2021) show an increase in share for domains with 4 and 5 levels. The final decision was, therefore, to proceed with the domain names of phishing URLs with five levels of accuracy to keep the accuracy above the 90% mark.

IV. Blacklist creation

Consolidation of the data from 2 sources into a single dataset and cleansing using the five levels of the domain name concluded the data preparation phase. For the next step, the process of assessing the new records and building the blacklist had to be designed to reflect the real-world process. This meant that the blacklist had to be built chronologically as data appeared in the real world (chronologically, based on their recorded date and time in the dataset from oldest to newest). Only confirmed phishing records (TP) were added to the blacklist. Blacklist was created to record the domain and date when it was classified as confirmed phishing (TP). Graylist was created to record the domain name and the date when the domain was classified as non-phishing (FP).
Every record within the dataset went first through the **Assessment** step - checking whether the domain existed in Graylist or Blacklist. If the domain was found in Graylist, this record was flagged as ambiguous(UNK) and removed from the final dataset. If the domain was found on Blacklist, the record was flagged as confirmed phishing(TP) and remained in the final dataset. If the record wasn't found in any of the lists it continued with the **List update** step where the classification from the original source was used to update the Graylist (this applied only to PhishTank records which were classified as non-phishing, as FP), Blacklist (this applied to records from both data sources in case the record was classified as confirmed phishing, as TP) or remained classified as UNK. Only records that were updating the Blacklist would be left in the final dataset, those classified as FP or UNK were removed from the final dataset. See the process flow diagram in ("Fig. 6"). The end result of passing all the records through this workflow was:

- a Blacklist with all TP records,
- a Graylist with all FP records and
- and a dataset with occurrences of confirmed phishing records

V. Findings

A. Efficacy of blacklist in fraud detection

1) What portion of phishing attempts could be detected via blacklist year by year?: Consolidated figures on the complete 10-year data show that ≈17% of recorded attacks could be detected via Blacklist, which means that this domain has been seen before and classified as a confirmed fraud. The remaining ≈73% will be skipped due to the record not being seen before. From year-over-year statistics, the efficacy ranges from ≈6% up to ≈25% in a given year.

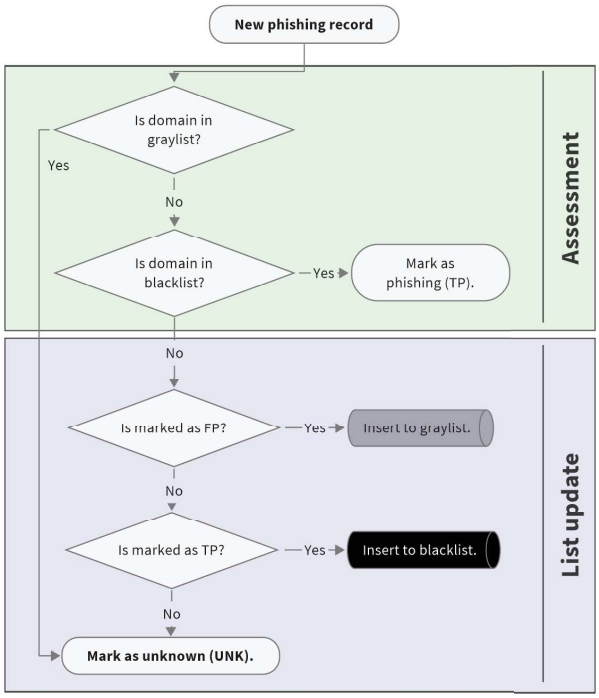

Fig. 6. Flow of phishing record assessment

	Year										
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	All
CATCH	14.8%	17.7%	20.1%	16.4%	16.4%	24.3%	18.4%	24.5%	15.6%	6.5%	16.7%
MISS	85.2%	82.3%	79.9%	83.6%	83.6%	75.7%	81.6%	75.5%	84.4%	93.5%	83.3%

Fig. 7. YoY efficacy of the blacklist-based detection

2) Is the YoY efficacy of the blacklist increasing, decreasing, or steady?: Year-over-year view shows initially between the years 2013 and 2015 an increasing efficacy from almost 15% up to 20%. In the more recent period - from 2020 till 2022 - we see a sharp decrease from almost 25% down to 6.5% ("Fig. 7").

When analyzing the records that were successfully caught by the blacklist we were additionally interested in the ratio - which of these records were classified by reviewers in the PhishTank as confirmed phishing (TP) and therefore would be detected even without blacklist and which were left unclassified (UNK). In the initial period (2013-2016) we see an increasing share of records (from almost 31% to almost 62%) that were not classified by reviewers. These records constituted a clear and tangible contribution of the blacklist to identifying phishing attacks. This trend though reverts to a quick decline ending at only ≈7% share of the unclassified records ("Fig. 8").

B. Prevalence of domain re-use for phishing

1) What is the ratio of phishing pages hosted on unique domains?: Considering all the data across 10 years pe-

	Year										
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	All
TP	69.1%	64.3%	58.8%	38.3%	52.7%	84.0%	85.7%	92.8%	93.5%	93.1%	73.6%
UNK	30.9%	35.7%	41.2%	61.7%	47.3%	16.0%	14.3%	7.2%	6.5%	6.9%	26.4%

Fig. 8. Classification of records captured by Blacklist - % YoY share

	Year										
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	All
RE-OCC	21.5%	27.5%	35.5%	38.4%	37.4%	33.5%	32.1%	37.8%	14.1%	6.1%	21.8%
UNIQUE	78.5%	72.5%	64.5%	61.6%	62.6%	66.5%	67.9%	62.2%	85.9%	93.9%	78.2%

Fig. 9. YoY % share of re-occurring vs. unique domains

riod, ≈78% of confirmed phishing attacks were hosted on the unique domains, and less than 22% were hosted on re-occurring domains. In the year-over-year view, only re-occurrences that happened within 365 days have been considered (to account for the situation where the domain first observed in 2013 had nine years to re-occur, while the domain from 2022 had less than a 1-year window to re-occur). The reason why 365 days window was selected was the ease of comparison on a YoY basis. Also, due to figures from analysis of days between re-occurrences, where more than 95% of domains re-occurred within 365 days. In the final figures, a gradual increase in the share of unique domains is notable ("Fig. 9").

2) How often is the phishing domain re-occurring?: From domains that re-occurred, the vast majority did so only once (almost 68%), some were re-used twice (less than 17%), and only some were re-used three(6%), four(3%) or five times(less than 2%), see "Fig. 10".

C. Time periods related to domain re-occurrence

1) What is the usual period before the domain first re-occurs?: The average time for the first re-occurrence is 51.8 days, and for those domains that re-occurred, 90% did so within 96 days from their first occurrence, 95% within 305 days, and 99% needed 903 days.

2) What is the usual time period between the re-occurrence of the domain?: The average time re-occurrence, irrespective of whether first or second, etc., was 56.7 days. A slightly higher number than the first re-occurrence. 90% of the re-

Re-occurence	
1	67.8%
2	16.8%
3	6.5%
4	3.3%
5	1.9%
6	1.2%
7	0.7%
8	0.5%
9+	1.3%

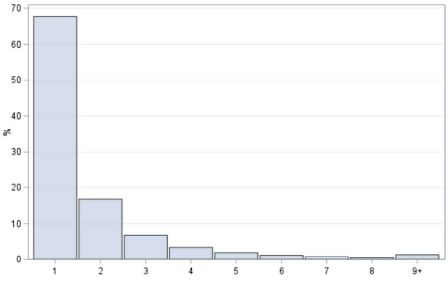

Fig. 10. Frequency of the domain re-occurence

Re-occurrence	Average	365D Threshold	90% Threshold	99% Threshold
1st time re-occurence	51.8 days	96.5%	96 days	903 days
Any re-occurence	56.7 days	96.1%	117 days	977 days

Fig. 11. Days between domain re-occurrence

appearances happened within 117 days, 95% within 320 days and 99% within 977 days ("Fig. 11").

VI. Future work

A. Temporary blacklist and retention period

Blacklist described in this analysis was built to record and maintain all historically identified phishing domains. Capturing all domains in the database, considering the further need to increase the number of domain levels to be considered and a visible increase in newly registered domains YoY, might prove to be inefficient or even unsustainable in certain use cases. Therefore analysis of defining a retention period after which the record is removed from the blacklist might help keep the blacklist within a manageable size.

B. Complementing the detection with whitelist

Though adding the whitelist into the Assessment process wouldn't improve the efficacy of the blacklist detection, it could improve the overall accuracy of phishing/non-phishing classification by reducing the number of UNK records.

C. Combining the blacklist with analytics algorithms

As described in the process workflow, utilizing the domain blacklist requires a complementary method of classification of the phishing attempts which were not found in the blacklist. For this step, selected algorithms of predictive analytics could be very efficient.

VII. Conclusions

Phishing data from multiple sources and covering a 10-year time window were meant to provide a balanced and sufficient foundation for our analysis.

The main question behind the analysis was about the efficacy of a blacklist. Considering that financial losses linked to phishing are growing year by year [1], even marginal improvements in detection have a meaningful impact (we shouldn't also ignore the non-financial impact of phishing). The results prove that domain blacklist is a relevant detection technique with a capacity to detect between 15% and 20% of the reported incidents throughout the 10-year time window, with a single exception being last year, during which the efficacy dropped to less than 7%. This decrease is attributed to the noticeable increase in the number of newly created subdomain variants for a single domain (sld.tld) which can be observed in the statistics of unique domains occurrence "Fig. 9".

The above detection figures tightly correlate with the ratio of domain re-occurrence or domain re-use, representing the ultimate ceiling of the blacklist's efficacy as the blacklist is only applicable for re-occurring domains. The data showed ("Fig. 9") a steady and continuous increase in the % share of

unique domains and a shrinking % share of the domains being re-used.

The last notable observation is from the analysis of the % share of re-occurring domains(only a small part, around 22% of all domains). This analysis also provided the optimal time-window period for comparison as more than 95% of re-occurring domains do so within one year (365 days) from the previous occurrence, although the average time between any two occurrences is ≈57 days (“Fig. 11”).

Though we observed the reduced efficacy of the blacklist due to the spike in using new domains, this should not be a signal to stop using the blacklists or consider them obsolete. Quite the opposite. We are reading these results as a confirmation of how efficient the blacklists in the current solutions are. The fact that threat actors are keener to register and use new domains than re-use the existing ones is presumably a result of the current phishing detection techniques, which are using the Blacklists efficiently, resulting in the blocking of the re-used phishing domains.

ACKNOWLEDGMENT

This publication was realized with the support of Operational Program Integrated Infrastructure 2014 - 2020 of the project: Intelligent operating and processing systems for UAVs, code ITMS 313011V422, co-financed by the European Regional Development Fund.

EUROPEAN UNION
European Regional Development Fund
OP Integrated Infrastructure 2014 – 2020

MINISTRY OF TRANSPORT AND CONSTRUCTION OF THE SLOVAK REPUBLIC

It was partially supported by the Erasmus+ project: Project number: 022-1-SK01-KA220-HED-000089149, Project title: Including EVERyone in GREEN Data Analysis (EVERGREEN).

Co-funded by the Erasmus+ Programme of the European Union

EverGreen DATA Analytics

REFERENCES

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

[2] G. Aaron, L. Chapin, D. Piscitello, and D. C. Strutt. Phishing Landscape 2022 - An Annual Study of the Scope and Distribution of Phishing. Interisle Consulting Group. [Online]. Available: https://interisle.net/PhishingLandscape2022.pdf

[3] Phishing Activity Trends Report 3rd Quarter 2022. Anti-Phishing Working Group. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q3_2022.pdf

[4] M. Rader and S. Rahman, “Exploring historical and emerging phishing techniques and mitigating the associated security risks,” *International Journal of Network Security & Its Applications*, vol. 5, 11 2015.

[5] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” 01 2009.

[6] T. Skybakmoen and V. Phatak. Comparative Test Report - Q2 2021 Web Browser vs. Phishing . CyberRatings.org. [Online]. Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWLycn

[7] R. Howard, *Cyber Fraud: Tactics, Techniques and Procedures*. Auerbach Publications, April 2009.

[8] A. Jain and B. B. Gupta, “A survey of phishing attack techniques, defence mechanisms and open research challenges,” *Enterprise Information Systems*, vol. 16, pp. 1–39, 03 2021.

[9] J. Ma, L. Saul, S. Savage, and G. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious urls,” 06 2009, pp. 1245–1254.

[10] R. Rao and A. Pais, *An Enhanced Blacklist Method to Detect Phishing Websites*, 01 2017, pp. 323–333.

[11] S. Bell and P. Komisarczuk, “An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank,” 02 2020, pp. 1–11.

[12] T. Phuong Thao, T. Makanju, J. Urakawa, A. Yamada, K. Murakami, and A. Kubota, “Large-scale analysis of domain blacklists,” 01 2020.

[13] P. Mockapetris, “DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION,” Internet Requests for Comments, RFC Editor, RFC 1035, November 1987. [Online]. Available: https://www.rfc-editor.org/rfc/rfc1035.txt