Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 91–98

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

A Bibliometric Analysis of Phishing in the Big Data Era: High Focus on Algorithms and Low Focus on People

Mirjana Pejić-Bach*, Ivan Jajić, Tanja Kamenjarska

Faculty of Economics and Business, University of Zagreb, Square of John F. Kennedy 6, 10 000 Zagreb, Croatia

Abstract

The phishing attacks, based on social engineering to persuade potential victims to provide valuable information, have significantly increased in the pandemic Covid-19 era, characterised by ubiquitous big data technologies. This paper aims to assess the theoretical and empirical research on phishing emails and big data that has been done to identify trends and recommend new areas for research. Using the VOSviewer program, the search results from the Web of Science (WoS) database were extracted. A mapping technique, using VoS Viewer, was used to examine articles on big data and phishing emails. The findings show that most of the field's research is carried out in nations in Asia and the United States of America and that the number of publications in this area is increasing exponentially. However, it is evident that researchers predominately concentrate on technical fields like computer science. Even though they are used in relatively small quantities, machine learning techniques, particularly artificial neural networks, are associated with most of the phishing publications that have been studied. Six clusters correspond to the main phishing domains: Phishing target or victim, Phishing channel, Big data analytics, Big data machine learning, Phishing attacker, and External phishing protection. The results indicate that real-time data collection and the development of effective algorithms are new approaches to combating phishing assaults. However, research outside of the technical domains is scarce.

Keywords: bibliometric analysis; phishing; big data; VOSviewer; text-mining, co-occurrence

* Corresponding author. Tel.: +385-1-238-3464.
E-mail address: mpejic@net.efzg.hr

10.1016/j.procs.2023.01.268

1. Introduction

With the popularisation of the internet and its wide acceptance, the daily routine without access to the internet is unthinkable. At the same time, everyday business operations would be unbearable without this global connectivity [1]. The Eurostat statistics denote that 89% of the European Union (EU) individuals used the internet in the last 3 months, while 90% of them use it daily in 2021 [2]. With such a large proportion of Internet users, a large amount of data is available online and is generated daily. Solutions needed to be made to sufficiently analyse, extract and store those huge amounts of data. The term "Big data" arrived in the 90s, although until the mid-2000s was not so widely used [3]. Big Data denotes datasets that are so large that they are not easy to handle using traditional database management systems [4]. Therefore, by having the ability to understand and extract the most relevant information, businesses can gain valuable insight into the targeted customer base.

Businesses use this big data to gain consumer information, and cyber criminals also developed social-engineering-based attacks, among others, to steal personal information from the victim, called phishing attacks [5]. A phishing attack is generally a form of the message where the attacker uses false information and wants the victim of this email to enter their data or click on a link to a malicious website. There are many forms of phishing except emails, such as Short Message Service (SMS), Instant Messaging (IM) and such [6]. A victim can receive a message requiring information about their bank account, but of course, in a false manner [7]. Therefore, an individual must be careful about which websites they buy, enters information and visit, as all that can affect the phishing attack messages they might receive.

Research in the form of a bibliometric analysis of phishing attacks is still scarce, largely focusing on a specific industry or general cyber security threats. To our best knowledge, there are no studies on the inter-relation between big data and phishing, using the visual representation conducted by VOSviewer, as in similar studies on the security issues [8],[9],[10]. This research focuses on the relationship between big data and phishing, detecting the most important connecting points, such as keywords, citation, and co-authorship by author and country. The papers were extracted from the Web of Science (WoS) database.

The outline of the article is as follows. After the introduction, the methodology is presented. Findings are described in the third chapter of the article, while the conclusion and future implications are discussed lastly.

2. Methodology

To gather relevant research articles for this paper Web of Science database was used. The database was searched in May 2022, using the keywords "big data" and phishing, which resulted in 136 articles, with the first paper emerging in 2010. The search was limited to peer literature written in the English language. Co-occurrence analysis using VosViewer was conducted with the goal of topic extraction using cluster analysis and heat maps. Keywords that occur at least twice in the papers are included. The overall link strength of the co-occurrence (all terms) is 161, and the analysis found 119 links for 31 items. The lin/log modularity normalisation approach is utilised.

3. Bibliometric analysis

Research on big data and phishing is concentrated in the Computer Science area (91.18%), followed by Engineering (28.68%), Telecommunications (16.91%), Operations Research Management Science (2.94%) and Other research areas that include areas such as Automation Control Systems, Business Economics, Chemistry, Criminology Penology, Education Educational Research, Information Science Library Science, Instruments Instrumentation, Psychology and Science Technology.

The majority of the research are proceeding papers (63.97%) and Articles (32.35%), while the least published types of documents are Review Articles (3.68%), Early Access (1.47%) and Book Chapters (0.74%). Therefore, most of the extracted publications (63.24%) are indexed in Conference Proceedings Citation Index – Science (CPCI-S), followed by Science Citation Index Expanded (SCI-EXPANDED) (22.06%), Emerging Sources Citation Index (ESCI) (12.5%), Social Sciences Citation Index (SSCI) (3.68%) and Other indices which contain the Book Citation Index – Science (BKCI-S) and Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH) (1.48%). The results indicate that majority of the research results in this field are published in conference proceedings, which is confirmed by the analysis of the most frequent publication titles (Table 1), indicating that the IEEE INTERNATIONAL CONFERENCE ON BIG DATA for the years 2013, 2014, 2016, 2017, 2018 and 2020 holds the majority of the publication titles in the field of phishing emails and big data (22.06%). This trend is preceded by the IEEE TRUSTCOM BIGDATASE ISPA with 5.89%, 2018 IEEE INT CONF ON PARALLEL DISTRIBUTED PROCESSING WITH APPLICATIONS UBIQUITOUS COMPUTING COMMUNICATIONS BIG DATA CLOUD COMPUTING SOCIAL COMPUTING NETWORKING SUSTAINABLE COMPUTING COMMUNICATIONS (3.68%), IEEE ACCESS (3.68%), INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY (3.68%) and LECTURE NOTES IN COMPUTER SCIENCE (2.21%). The rest of the publication titles related to phishing emails and big data hold 1.47% or less.

Table 1. Publication Titles (5+ papers).

Publication Titles	Record Count	% of 136
IEEE INTERNATIONAL CONFERENCE ON BIG DATA 2013, 2014, 2016, 2017, 2018, 2020	30	22.06%
IEEE TRUSTCOM BIGDATASE ISPA	8	5.89%
2018 IEEE INT CONF ON PARALLEL DISTRIBUTED PROCESSING WITH APPLICATIONS UBIQUITOUS COMPUTING COMMUNICATIONS BIG DATA CLOUD COMPUTING SOCIAL COMPUTING NETWORKING SUSTAINABLE COMPUTING COMMUNICATIONS	5	3.68%
IEEE ACCESS	5	3.68%
INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY	5	3.68%

Source: Authors' work.

Fig. 1 shows that the research on phishing emails and big data remains under-explored and is yet to gain research attention. It can be seen that most of the research will be published in 2020 (16.91%), with the constant rise in the number of publications. However, the number of publications started to decrease in 2021 and 2022. Even though we do not know if, until the year 2022, the number of papers will increase, we are certain that this number is already in downfall, confirmed from the previous 2021. year. One of the reasons for such a disinterest in the topics mentioned above could be seen in the current increase of other cyber-attack forms due to the instability in the world, especially in Europe, due to the Russian Federation's aggression on Ukraine.

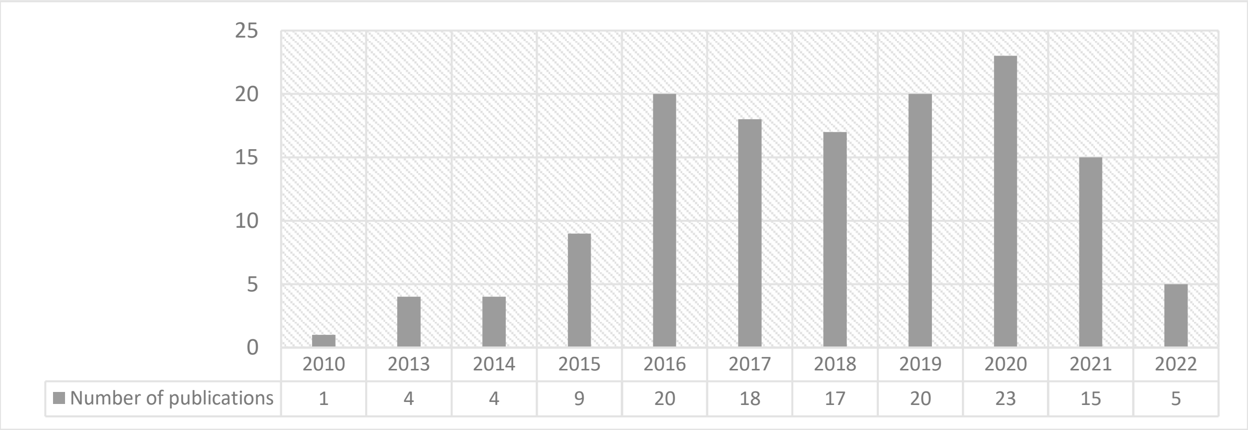


Fig. 1. Publication trend of the research papers on big data and phishing; Source (Authors' work).

The analysis shows in Table 2 that most of the research publications are conducted in Asia (58.84%), with the majority in the People's Republic of China (25%). Furthermore, North and South America hold 26.47% of the records (USA- 23.53%; Canada-1.47% and Brazil-1.47%). The second continent with most of the publications (21.33%) in the field of phishing emails and big data in Europe (England - 5.15%; France-4.41%, Netherlands- 2.94%, Italy-1.47%, Czech Republic-1.47%, Portugal-1.47%, Slovenia-1.47%, Switzerland-1.47% and Denmark-0.74%). The least research attention is observed in Australia (7.35%) and Africa (1.47%). It should be noted that one research can include authors from different continents. Thus, the record count is higher than the extracted publications.

Table 2. Most research by the country

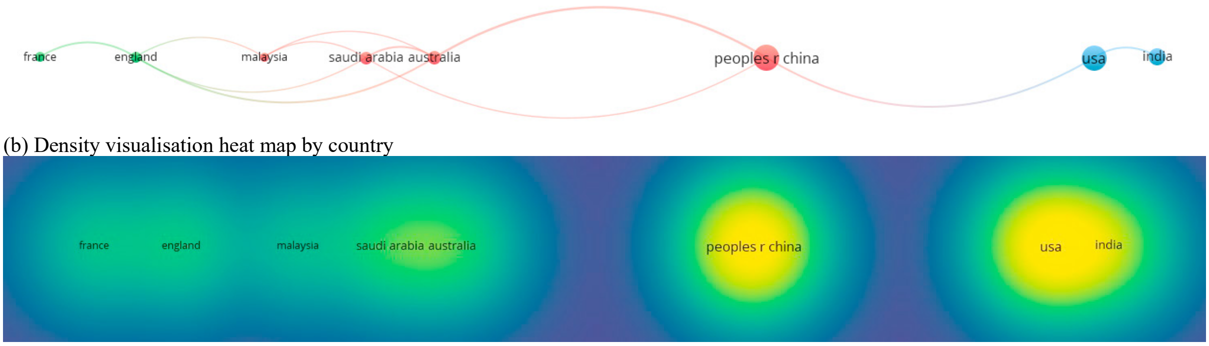
Countries/Regions	Record Count	%
ASIA (PEOPLES R CHINA, INDIA, SAUDI ARABIA, MALAYSIA, IRAQ, SOUTH KOREA, BANGLADESH, INDONESIA, PAKISTAN, TAIWAN, AZERBAIJAN	80	58.84%
NORTH AND SOUTH AMERICA (USA, CANADA, BRAZIL)	36	26.47%
AUSTRALIA	10	7.35%
EUROPE (ENGLAND, FRANCE, NETHERLANDS, ITALY, CZECH REPUBLIC, PORTUGAL, SLOVENIA, SWITZERLAND, DENMARK)	27	21.33%
AFRICA (SOUTH AFRICA)	2	1.47%

Source: Author's work

4. Co-occurrence analysis

The co-occurrence analysis was undertaken in the VosViewer program for the keywords and country of authors. First, we looked at co-authorship by the nation for two documents as a minimum number of documents per country (Fig. 2). The number of items detected is seven, with nine linkages totalling 10 link strengths. In addition, three clusters were discovered. The most important countries in this research bibliometric analysis are the People's Republic of China, the United States of America and India. However, research collaborations still emerge the most often between the countries that are related to each other in terms of geographical location or business relationship.

(a) Interconnected countries



(b) Density visualisation heat map by country

Fig. 2. Co-authorship by country (Source: Authors' work).

Second, we have conducted the co-occurrence using keywords. All keywords that appeared at least twice in the research papers were clustered into six logical categories.

Six distinct clusters were discovered. The main research paper keywords and the biggest nodes are grouped in Clusters 3, 4 and 5 with the corresponding supporting keywords. The main research node and keyword "Phishing" is co-occurring with social engineering, personality traits, and behaviour, to list a few. The keyword "Big Data" is co-occurring with the third node ", Machine Learning" and the fourth node "Phishing Detection" in the same cluster. Table 3 represents research paper clusters. Clusters are also presented in Fig. 3. The node's size signifies the keyword frequency: the larger the node, the greater the keyword frequency. The line thickness is determined by the closeness of the connections between the two terms [11]. The clusters are denoted using their specific focus. The clusters that focus more on technical issues, like big data analytics and big data machine learning, are more connected to other clusters and encompass a larger number of keywords and papers.

Table 3 – Research paper clusters

Cluster	Keywords (# of occurrences / Link strength)	Example papers	Cluster focus
C1	Taxonomy, Impact, Performance, Model, Individual Differences	[12] - Taxonomy; [13] – Model, Individual differences; [14] - Individual differences	Phishing target or victim
C2	Data Mining, Algorithm, Phishing Website, Phishing Email	[16] – Phishing Email; [17] – Algorithm, Phishing Website; [18] – Phishing Website	Phishing channel
C3	Malware, Big Data Analytics, Generative Adversarial Networks, Online Security	[19] – Big Data analytics; [20] – Generative Adversarial Networks; [21] – Online Security;	Big data analytics
C4	Big Data, Phishing Detection, Classification, Machine Learning, Dimensionality Reduction, Features, Deep Learning, Ensemble Learning, Protection Scheme	[22] – Classification; [23] - Deep learning; [24] – Ensemble; [25] - Protection scheme, Features	Big data machine learning
C5	Phishing, Social Engineering, Personality Traits, Behavior, Security, Phishing Attack, Autoencoder	[26] – Autoencoder; [27] - Social engineering, phishing attack	Phishing attacker
C6	Cybersecurity, Spark	[28] - Cybersecurity	External phishing protection

Source: Author's work

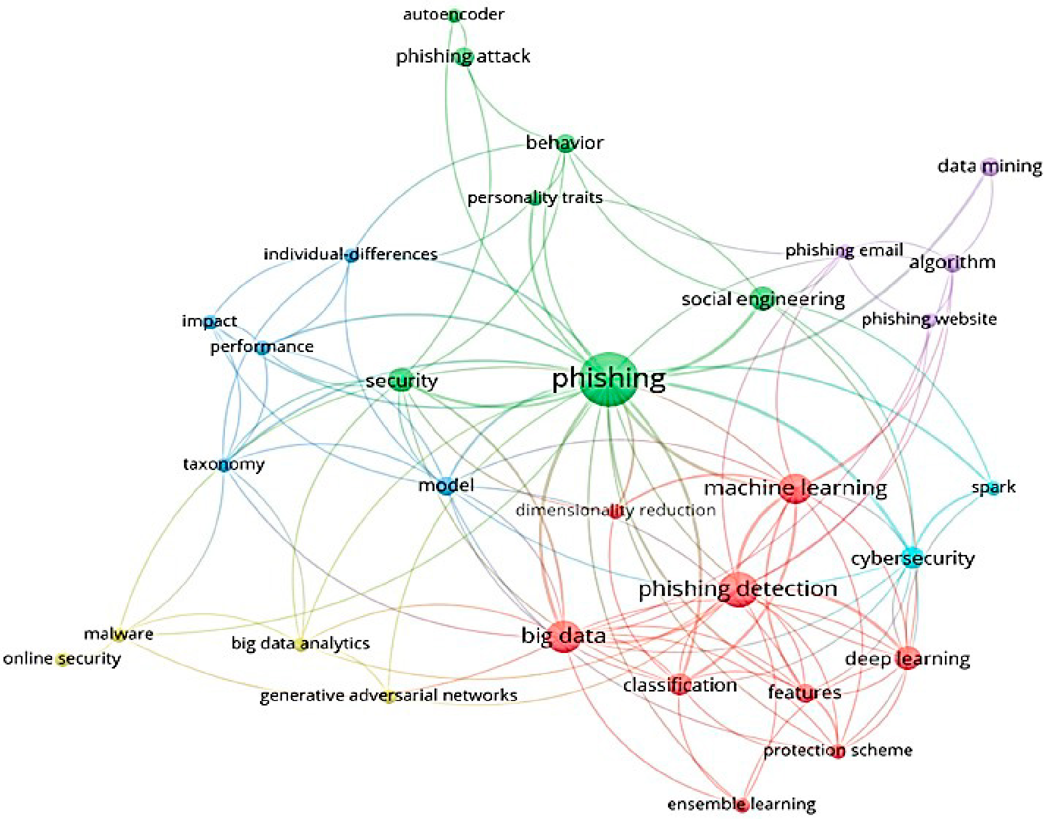


Fig. 3. Co-occurrence of the keywords; Source: Author's work.

Cluster 1 contains the keywords Taxonomy, Impact, Performance, Model, and Individual Differences. The papers in this cluster examine the topic of phishing targets or victims. [11][12] focus on the challenges of securely streaming over the internet with great emphasis on the security approaches and challenges of becoming a potential phishing victim. They find that the best solution for preserving potential phishing victims is seen at the organisational level through implementing technology security governance following strict taxonomy classification. [13] Besides email and spoofed websites, social media is a new market for phishing targets. Users with heuristic processing (thinking what-if) are less prone to a phishing attack. The paper was one of the first to research the alignment between Big Five personalities and the heuristic-systematic information processing model. [14] also used the Big Five personality model to investigate participants' performance in the Phish Derby simulation competition for detecting phishing emails, showing that extraversion and agreeableness are associated with poor phishing email detection and reporting. Furthermore, older participants performed better than younger ones, while more educated participants performed poorer. Although this conclusion is counterintuitive, since it could be expected that a higher educational level is a protection against fraud, other research indicates that this is not the case with phishing attacks [15]. This line of research indicates that technological security needs to be implemented stricter through rules and regulations; however, considering that an individual's personality is the final line of defence against phishing attacks.

Cluster 2 contains the keywords Data Mining, Algorithm, Phishing Website, and Phishing Email. The authors in this cluster explore phishing channels in their research papers from the standpoint of algorithms efficiency. [16] denotes that the reality of phishing emails is still prevalent, and the right solution is to detect it from the big email data. Cybercriminals have successfully camouflaged their attempts; therefore, the existing detection methods are insufficient. The paper used generalisation, word segmentation and word vector generation in their testing phase for implementing their results into a Long Short-Term Memory Network (LSTM) neural network model training phase. Findings show that with 95 per cent accuracy, this model can detect phishing emails. [17] confirms the machine learning approach used by [16] as they outperform any other technique by accuracy. The paper denotes that the machine learning approach can not handle big data sets, but this problem could be solved by combining it with the New instructions (NI) algorithm. [18] proposed a new Phishing Hybrid Feature-Based Classifier (PHFBC) model using machine learning with two Naive Base and Decision Tree algorithms. It used a robust set to validate the model. The findings show that the PHFBC model, through real-time validation and simulations, showed distinctive from the competition in terms of accuracy and minimal misclassification for the Web phishing problem.

Cluster 3 contains the keywords Malware, Big Data Analytics, Generative Adversarial Networks and Online Security. The authors in this cluster explore big data analytics. Web application attacks significantly threaten computer networks and end-users[19], and authors constructed a threat intelligence technique for web attacks where they: Collect web attack data through website crawling; Extract features by using the Association Rule Mining (ARM) algorithm; Simulate the web attack by using the collected data; Propose a new Outlier Gaussian Mixture (OGM) technique for zero-day and known attacks detection using the anomaly detection methodology. The findings show that the proposed scheme outperforms the competition regarding detection and false alarm rates. [20] focuses on phishing URL detection. The authors mention that to oversample the phishing URL dataset, specific criteria need to be set, and the dataset needs to be diverse. They trained text generative adversarial networks (text-GANs) with minority URLs which they later used for oversampling by the set criteria. The findings show significant performance improvements when the oversampling technique has been used. [21] denotes that online phishing websites are on the rise, showing a significant security threat. The authors constructed a PhishMonger platform where they captured live phishing websites in real-time. This platform can be of great use for the cybersecurity and security informatics government bodies and the research community.

Cluster 4 contains Big Data, Phishing Detection, Classification, Machine Learning, Dimensionality Reduction, Features, Deep Learning, Ensemble Learning, and Protection Scheme. The papers in this cluster focus on big data machine learning. [22] confirms that machine learning is the most popular tool for data analysis and results gathering for phishing. They find that covering approach models are more applicable for anti-phishing solutions. [23] proposed a novel voting-based deep learning framework, which showed highly effective system performance. The phishing false alarm was reduced to 75% compared with other models such as LSTM. [24] proposes a weighted consensus clustering for Big data instead of single clustering methods. Euclidean distance metric achieved the highest accuracy, which can be applied to making group decisions on several alternatives. [25] The machine learning approach used different URL parts to find their phishing potential. The authors conclude that the Grey Wolf Optimiser Algorithm significantly outperforms the Firefly Algorithm in detecting phishing based on the URL length, pop-up windows, and HTTP response.

Cluster 5 contains Phishing, Social Engineering, Personality Traits, Behavior, Security, Phishing Attack and Autoencoder. The authors in this cluster research the phishing attacker. [26] propose a new phishing detection approach based on the spam filter to classify the textual content of an email. Autoencoder is used to classify and extract URL features. Both methods should show better anti-phishing solutions. [27] provides an educational standard for security awareness of those users that click on phishing emails.

Cluster 6 contains the keywords Cybersecurity and Spark. The authors in this cluster research external phishing protection. [28] denotes that data-driven analytics is being increasingly used in cybersecurity. The paper's authors use a human-machine collaborative approach to design a semi-supervised solution. This method implemented through PhishMonger's Targeted Brand dataset showed a better phishing detection by 12% from the baseline.

5. Conclusion

Phishing is a persistent and constantly changing threat, meaning cybersecurity professionals must comply with those changes, especially considering the emerging big data technologies. Constant involvement, innovation and breach testing are necessary for any anti-phishing efforts. Phishing is often present on malware websites or through email, although there are other methods such as SMS, instant messages and similar.

This paper presents the bibliometric analysis of big data and phishing. The papers were extracted from the Web of Science database and were analysed using the bibliometric analysis and co-occurrence using VosViewer. Results indicated that the first paper emerged in 2010, with a slight decrease in the number of papers in 2021, probably due to other emerging security threats, such as cyber-attacks. Most of the papers are published in conference proceedings, specifying that the research in this area needs rapid publication to catch the ever-changing nature of phishing attacks. Most papers are published in the Republic of China and the United States of America, followed by the European countries.

The research papers are structured based on the co-occurring keywords in six clusters related to the main phases of phishing detection and prevention: Phishing target or victim, Phishing channel, Big data analytics, Big data machine learning, Phishing attacker, and External phishing protection. Most research papers focus on machine and deep learning algorithms and their efficiency. New ways of tackling phishing attacks are seen in real-time data building and efficient algorithm creation based on big data technologies. Usage of machine learning with specific conditions in training and testing phases produces very efficient models for anti-phishing methods. This increased focus on machine learning in the context of big data analytics indicates that any effort to decrease the number of phishing attacks is to build updated daily and always-changing cybersecurity, which can be ensured only by utilising big data analytics. However, other areas of phishing prevention and detection did not produce the same number of papers, indicating that the management and organisational issues of phishing in the big data era are under-represented. This result is confirmed by the fact that most papers are published in technical areas, such as Computer science.

Acknowledgements

This research has been fully funded by the PhisHRban project (KK.01.2.1.02.0267).

References

[1] Mack, E. A., Dutton, W. H., Rikard, R., & Yankelevich, A. (2019) Mapping and measuring the information society: A social science perspective on the opportunities, problems, and prospects of broadband Internet data in the United States. *The Information Society*, **35**(**2**), 57–68. https://doi.org/10.1080/01972243.2019.1574526

[2] Eurostat. (2022) Individuals Internet Use [Data File]. Retrieved from https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_ifp_iu/default/table?lang=en

[3] Elgendy, N., & Elragal, A. (2014) Big Data Analytics: A Literature Review Paper. Advances in Data Mining. *Applications and Theoretical Aspects*, 214–227. https://doi.org/10.1007/978-3-319-08976-8_16

[4] Al-Sai, Z. A., Abdullah, R., & Husin, M. H. (2019) Big Data Impacts and Challenges: A Review. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. https://doi.org/10.1109/jeeit.2019.8717484

[5] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021) Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, **3**. https://doi.org/10.3389/fcomp.2021.563060

[6] Gupta, P., Srinivasan, B., Balasubramaniyan, V., and Ahamad, M. (2015) "Phoneypot: data-driven understanding of telephony threats," in *Proceedings 2015 network and distributed system security symposium*, (Reston, VA: Internet Society), 8–11. doi:10.14722/ndss.2015.23176

[7] Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2020) A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, **76**(**1**), 139–154. https://doi.org/10.1007/s11235-020-00733-2

[8] Ho, H. T. N., & Luong, H. T. (2022) Research trends in cybercrime victimisation during 2010–2020: a bibliometric analysis. *SN Social Sciences*, **2**(**1**). https://doi.org/10.1007/s43545-021-00305-4

[9] Arora, P., & Jain, A. (2021) Cyber Security Threats And Their Solutions Through Deep Learning: A Bibliometric Analysis. 2021 3rd *International Conference on Advances in Computing, Communication Control and Networking (ICAC3N).* https://doi.org/10.1109/icac3n53548.2021.9725480

[10] Cheng, P., Tang, H., Dong, Y., Liu, K., Jiang, P., & Liu, Y. (2021) Knowledge Mapping of Research on Land Use Change and Food Security: A Visual Analysis Using CiteSpace and VOSviewer. *International Journal of Environmental Research and Public Health*, **18**(**24**), 13065. https://doi.org/10.3390/ijerph182413065

[11] Shonhe, L (2020). Continuous Professional Development (CPD) of librarians: A bibliometric analysis of research productivity viewed through WoS. The Journal of Academic Librarianship 46, 102-106

[12] Kumari, A., Tanwar, S., Tyagi, S., & Kumar, N. (2019) Verification and validation techniques for streaming big data analytics in internet of things environment. IET Networks, **8**(**3**), 155-163.

[13] Frauenstein, E. D., & Flowerday, S. (2020) Susceptibility to phishing on social network sites: A personality information processing model. Computers & security, **94**, 101862.

[14] Canham, M., Posey, M. C., & Constantino, M. (2022) Phish Derby: Shoring the Human Shield Through Gamified Phishing Attacks. Frontiers in Education, **6**

[15] Caldwell, T. (2016) Making security awareness training work. Computer Fraud & Security, (**6**), 8-14.

[16] Li, Q., Cheng, M., Wang, J., & Sun, B. (2022) LSTM Based Phishing Detection for Big Email Data. *IEEE Transactions on Big Data*, **8**(**1**), 278–288. https://doi.org/10.1109/tbdata.2020.2978915

[17] Akinyelu, A. A. (2019) Machine Learning and Nature Inspired Based Phishing Detection: A Literature Survey. *International Journal on Artificial Intelligence Tools*, **28**(**05**), 1930002. https://doi.org/10.1142/s0218213019300023

[18] Zuhair, H., & Selamat, A. (2018) Phishing Hybrid Feature-Based Classifier by Using Recursive Features Subset Selection and Machine Learning Algorithms. *Advances in Intelligent Systems and Computing*, 267–277. https://doi.org/10.1007/978-3-319-99007-1_26

[19] Moustafa, N., Misra, G., & Slay, J. (2021) Generalized Outlier Gaussian Mixture Technique Based on Automated Association Features for Simulating and Detecting Web Application Attacks. *IEEE Transactions on Sustainable Computing*, **6**(**2**), 245–256. https://doi.org/10.1109/tsusc.2018.2808430

[20] Anand, A., Gorde, K., Antony Moniz, J. R., Park, N., Chakraborty, T., & Chu, B. T. (2018) Phishing URL Detection with Oversampling based on Text Generative Adversarial Networks. *2018 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata.2018.8622547

[21] Dobolyi, D. G., & Abbasi, A. (2016) PhishMonger: A free and open source public archive of real-world phishing websites. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*. https://doi.org/10.1109/isi.2016.7745439

[22] Abdelhamid, N., Thabtah, F., & Abdel-jaber, H. (2017) Phishing detection: A recent intelligent machine learning comparison based on models content and features. *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. https://doi.org/10.1109/isi.2017.8004877

[23] Haghighat, M. H., & Li, J. (2021) Intrusion detection system using voting-based neural network. *Tsinghua Science and Technology*, **26**(**4**), 484–495. https://doi.org/10.26599/tst.2020.9010022

[24] Alguliyev, R. M., Aliguliyev, R. M., & Sukhostat, L. V. (2020) Weighted consensus clustering and its application to Big data. *Expert Systems with Applications*, **150**, 113294. https://doi.org/10.1016/j.eswa.2020.113294

[25] Anupam, S., & Kar, A. K. (2020) Phishing website detection using support vector machines and nature-inspired optimisation algorithms. Telecommunication Systems, **76**(**1**), 17–32. https://doi.org/10.1007/s11235-020-00739-w

[26] Douzi, S., Amar, M., & el Ouahidi, B. (2017) Advanced Phishing Filter Using Autoencoder and Denoising Autoencoder. *Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017*. https://doi.org/10.1145/3175684.3175690

[27] Carella, A., Kotsoev, M., & Truta, T. M. (2017) Impact of security awareness training on phishing click-through rates. *2017 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata.2017.8258485

[28] das Bhattacharjee, S., Talukder, A., Al-Shaer, E., & Doshi, P. (2017) Prioritised active learning for malicious URL detection using weighted text-based features. *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. https://doi.org/10.1109/isi.2017.8004883