



Analyzing Social Engineering Research through Co-Authorship Networks Using Scopus Database during 1926-2020

Leila Khalili * | Department of Knowledge and Information Science, Azarbaijan Shahid Madani University, Tabriz, Iran

Nayana Darshani Wijayasundara  | Librarian (PhD), University of Sri Jayewardenepura, Nugegoda, Sri Lanka.

Abstract

Hacking the human brain and manipulation of human trust to obtain information and get monetary gains is called social engineering. This study aims to visualize and analyze the co-authorship networks in the Scopus citation database's social engineering research from 1926 to 2020. The present quantitative study used the bibliometric method and social network analysis. The study collected data from the Scopus database. A total number of 1994 records was taken as the sample of the study. Researchers used descriptive and inferential statistics and social network analysis to obtain results; to do this, different software types were used in the study (SPSS, Microsoft Excel, Text Statistics Analyzer, ISI.exe, Pajek and VOSviewer). Findings indicate the top three sources of publishing and the related subject areas. Furthermore, the top three core authors and countries were identified. Also the authors with high centrality measures in co-authorship network were identified. Majority of papers had only one author. The Collaborative Coefficient among researchers was 0.36. Based on the results of Spearman's test, there was a significant association between the number of documents, the number of citations and the rate of total link

* Corresponding Author: l.khalili@azaruniv.ac.ir

How to Cite: Khalili, L., Darshani Wijayasundara, N. (2022). Analyzing Social Engineering Research through Co-Authorship Networks Using Scopus Database during 1926-2020, *International Journal of Digital Content Management (IJDCM)*, 2(4), 15-45.

strength of the countries. Likewise there was a positive and high significant association between degree and closeness centralities. The researchers' frequently used keywords in this area were social engineering, phishing, and information security; in addition, the frequency of keywords was not compatible with Zipf's Law. A small sample of keywords cannot properly follow the Zipf's distribution.

Keywords: Bibliometric, Co-authorship Networks, Centrality Measures, Social Engineering, Zipf's Law.

INTRODUCTION

In Information Security "human element" is regarded as the "weakest link"(Lineberry, 2007). Even the securest technical protection systems can be detoured by attackers; they can divulge a password, influence the user to open a malicious email attachment or visit a prearranged website by manipulating them. Social engineering is the terminology coined for this process of manipulation (Heartfield & Loukas, 2015). The term is borrowed from the 20th century political sciences, where it signified clever methods to solve social problems. The positive connotation was eroded over the years, especially after the Second World War. The term got associated with negative flavor and the stereotypical manipulations of politicians to gain advantage in electoral votes (Duff, 2005). Nowadays still there is some negative aura around the term; however, it is somehow neutralized & is employed in Information System Security to describe cases in which people are defrauded to give away the private information (Hansson, 2006). A variety of items, like revealing passwords and giving access to the internal infrastructure of organization, etc. are involved. The popularity of the term in recent years is indebted to the increase in potential attacks and the disastrous aftermath it entails (Ivaturi & Janczewski, 2011).

Hacking the human brain can be the rudimentary definition of social engineering. No matter how much technology advances, the attacks on security persist due to the difficulty of upgrading or patching human brains (Townsend, 2010). "Social engineering is used by everyday people every day in everyday situations" (Hadnagy, 2010), thus, social engineers analyze the behavioral traits of people to plan for the future. Social engineering, for its own objective utility, calls for methods to control human behavior. The evocation of strong human emotion is a common tactic employed by social engineers. To begin with, the attacker tries to build trust with the victims by weaving a credible story. Basic human instincts such as greed, sympathy, or fear are often invoked in such stories (Townsend, 2010). The attackers gain the trust by convincing the victims to relate to such emotions. Rusch (1999) enumerates two substitute methods for persuading an individual; first 'central route to persuasion' which is strong analytical reasoning, and second 'peripheral route to persuasion' which elicits emotions.

Social engineering is the artful exploitation of people who are in fact the weakest link of information security systems. The attackers deceive the victims into releasing certain information or performing malevolent action. They begin by collecting background information on their would-be targets. Dumpster diving and phone calls are common methods for gathering such information. The increasing use of social networking sites, in turn, leads to a surge in the number of accessible tools and techniques for social engineering (Huber, Kowalski, Nohlberg & Tjoa, 2009).

The psychological aspect of social engineering cannot be ignored, as the attackers use the weaknesses in humans in their cyber-attacks (Montanez, Golob, & Xu, 2020). Cyber-attackers use social engineering techniques to get personal information from people, and they collate information through various sources. These attackers are targeting users, not the systems (Saeed & Shereef, 2020). Social engineers try to persuade people by appealing to their emotions, such as kindness, fear, trust and social obligations (Zulkifli, Zawawi & Rahim, 2020). Besides, some people use ways to build interpersonal relationships, leading to trust and commitment (Gao & Kim, 2007). Along with that, sometimes social engineers may ask someone for bank account details, promising that they will make a bank deposit as a prize for winning a lottery. Social engineering cyber-attacks occur in many forms. Some of them are baiting, pretexting, shoulder surfing (Wang, Sun & Zhu, 2020), phishing, frauds, scams, spear phishing, social media sock puppets (Montanez et al., 2020) and even forensic analysis (Oosterloo, 2020). Greavu-Serban and Serban (2014) have identified five models of how social engineers persuade people. The models are simplicity, interest, incongruity, confidence and empathy.

Nowadays social networking sites (SNSs) such as Facebook are used by attackers to gather primary background information on prospective victims. Furthermore SNSs facilitate the automation of attacks by providing data in machine readable form. Moreover, the automation of attacks is smoothed by SNSs, as they provide data in machine readable format. The automation primarily intends the reduction of human intervention time to a minimum which is the final aim of automated social engineering (ASE). Classic social engineering attacks are expensive due to the fact that building and maintaining rapport with someone to finally exploit the relationship is a time consuming task; accordingly, classical social engineering

attacks are costly. In contrast, automated social engineering bots are cheap and promising since they need little human time resources and can be scaled (Huber et al., 2009).

The nature of the net is itself an important factor in the redefinition of the arena of interaction among the individuals and their inclination to reveal private information. The "Net" generation prefers social networking sites for their communication; for posturing, role playing, or sounding off. However, access to such forums is rather easy, so anyone with an internet can review the users' personal information the posted content (Rosenblum, 2007). The best strategy to resist social engineering is to increase the level of awareness through education. In order to mitigate the effect of such activities, organizations should implement multi layered training to enforce policies like "need-to-know" access (Ivaturi & Janczewski, 2011).

Larabee's suggestion (2006) for classifying these attacks is taxonomy which is based on three extensive criteria; "close access technique", "online social engineering" and "intelligence gathering". Furthermore, Heartfield and Loukas (2015) introduced the taxonomy of semantic attacks; it means deceiving a user and thus manipulating the user-computer interference, aiming to rupture the information security of a computer system. Likewise, Ivaturi and Janczewski (2011) fleshed out various types of social engineering attacks, employing a taxonomy approach. Taxonomy is divided into two major categories; Person-Person and Person via Person which is divided into several sub-categories (Figure 1).

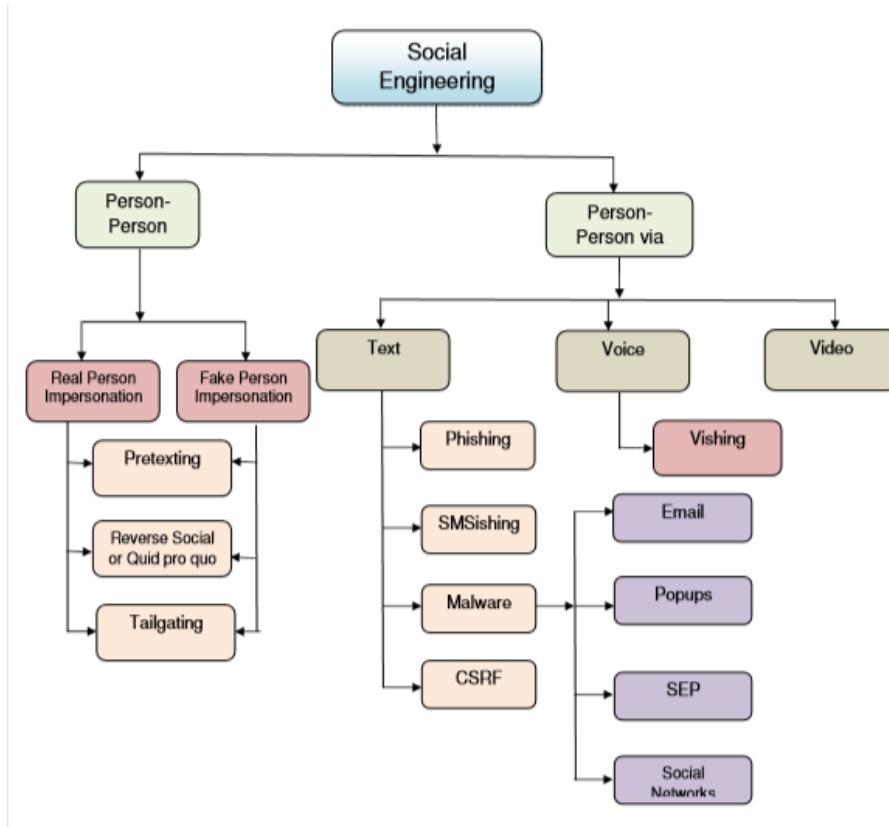


Figure 1. The taxonomy designed by Ivaturi & Janczewski (2011)

These days many large companies throughout the world try working from home, and it has immensely increased due to the COVID-19 pandemic. The International Labor Organization (ILO) estimates that nearly 18% of the global workforce has occupations suitable for working from home. One adverse effect of working from home is the decreased interpersonal relationship. People work in isolation with many technical tools to communicate like Zoom, email, MS Teams, Skype, Facebook (Saeed & Shareef, 2020), Dropbox and many other forms when they perform their work from home. Besides, many of us have transferred our daily banking and shopping activities online due to the pandemic. This transformation has made the ground for social engineering attacks. Due to the COVID-19 pandemic, online teaching and learning have bloomed more than ever before. Online education practice in universities is an instance where social engineering is

related to higher education's digital culture (Priatna, Malyawati, Sugilar & Ramdhani, 2020).

Recently the rate of social engineering attacks has rapidly risen and that, to be expected, has led to the weakening of the cybersecurity chain (Kalnin, Purin & Alksnis, 2017). Social engineering is very rewarding for cybercriminals. According to the report of CyberEdge, 79 percent of the attacks was successful in 2017. There is still a tendency to increase. About 62 percent of the attempts was successful in 2014. It increased to 71 and 76 percent during the next two years. With reference to the State of the Phish report of Wombat Security issued in 2019, about 83 percent of all companies gave an account of phishing attacks the preceding year. Apropos the same report, 49 percent of the attacks was via SMS and voice phishing, and 4 percent through infected thumb drives. Even a considerable number of information security professionals (64 percent) reported being spear phished in 2017. A study carried out in 2018 revealed that 17 percent of people have been victims of social engineering attacks. Companies are no exceptions either referring to Accenture reports during 2016 and 2017, around 69 percent of them experienced the social engineering attacks (Olson, 2019). Additionally, social engineering attacks are among the most perilous threats in the world. According to Cyence, the cyber security analyst company, the USA was the target of most social engineering attacks in 2016 and subsequently tolerated the highest cost; Germany and Japan followed. The approximate cost of the attacks was \$121.22 billion in the USA (Arana, 2017).

As the above mentioned statistics indicate, social engineering affects many peoples and institutions. Then, it is essential to know how the scientific communities deal with this widespread problem; in other words, how much scientific communities have investigated in this significant area. Scientometric and bibliometric studies can be used to integrate scientific outputs of a domain to obtain a better systematic review, statistical analysis, and science visualization. Furthermore, one way that helps researchers achieve research goals in their field is to understand and overview the scientific framework of the field, which is possible by visualizing the scientific map. Researchers obtain different characteristics of that field's publications by analyzing the scientific map of a field. According to Chen (2018) visualization is an effective method to generate a systematic review of the history and the situation of a scientific field. It may give an

insightful understanding of a research topic by identifying landmark studies in the development of the field, critical contributions in the past, and potentially transformative ideas.

Visualizing the scientific map can be presented in the form of co-authorship and co-word networks. Co-authorship networks are a type of network that consists of the author as the network node and lines between the nodes are the co-authors. Co-authorship network is an essential category of social networks and is widely used to determine the structure of scientific collaborations and researchers' position. Looking at scientific societies as networks of collaboration and co-authorship can help researchers understand these societies' behaviors and relationships better; it also helps policymakers in each scientific community to identify and encourage more effective behaviors. Moreover, the frequency of used keywords in the social engineering area may provide a better vision for researchers to design a better taxonomy. According to Lee, Chen and Tsai (2016) in visualizing keyword networks, the emergence of a new cluster indicates the beginning of a trend and the persistent cluster represents a continuation of an existing trend.

Taking into consideration the quantity of studies on social engineering, as compared to the number of individuals and institutions involved worldwide, there is a need for further research from different perspectives. Also, a review of the literature indicates that previous studies have not addressed the bibliometric approach on social engineering. Thus, due to lack of bibliometric research in this area, this research may partly help to reduce the gap in the literature.

Literature Reeviw

Review of the literature for present study is presented in three sections.

Bibliometric Analysis and Co-authorship

The analysis of co-authorship networks by bibliometric method can be applied to track nearly every feature of scientific collaboration networks (Glanzel & Schubert, 2004). Four previous bibliometric studies which are almost from ICT area, are reviewed here.

Bahrami and Rouzbahani (2021) studied the cyber security of smart manufacturing execution systems using bibliometric research. They found that Germany, China, and Italy compared to other

countries had more significant research. Also in a bibliometric study conducted by Mat et al., (2020), "android malware" was the keyword used to collect data from web of science (WoS); it covered the span of time between 2010 to 2019. They provided descriptive statistics of a total number of 1278 papers, based on factors like the year of publication, productivity in continents/ countries, research area, the categories of WoS, authors, institutions, countries; also the highly cited articles were recognized. The highest publications belonged to China, the USA, India, Italy, and South Korea respectively. The top authors came from Italy, Luxembourg, Malaysia, China, and India. Besides, Firdaus et al., (2019) noticed a gap due to lack of bibliometric study on blockchain research; thus they conducted a research in this field. They selected a total number of 1119 articles published from 2013 to 2018 for additional analysis. The finding reveals that the preference for publication in conferences was higher than journals or books among researchers. The USA is at the top of the list of blockchain publication, followed by China and Germany. Apart from Canada, India, and Brazil, research collaborations between countries increase the research publication. In the same vein, Rialti, Marzi, Ciappei and Busso, (2019) aimed to methodize the research on big data and the dynamic capabilities between 2007-2017. The study was carried out on 170 manuscripts collected from WoS. The outcome of the bibliometric analysis was four clusters of papers on the topics and the clarification of the content of each cluster. The distribution of publications over the years and the most influential authors were identified as well.

Glanzel and Schubert (2004) emphasize the complexity of scientific collaboration as an event in research which has become the focus of systematic studies since 1960s. The most palpable and best documented instance of such collaboration is co-authorship. Based on the previous studies, they announced the increase of collaboration among researches. Citing from Schubert and Braun (1990) and Glanzel (2001), they stated the dramatic rise of internationally co-authored papers in the last two decades. Also Persson, Glänzel and Danell (2004) observed that from 1980 to 1998 the number of papers grew somewhat, about 36%; in the same period the number of authors grew about 64%. In conclusion, they noted that the one acceptable interpretation of this growth is rooted in the alteration in the patterns of the documented scientific communication and collaboration.

Moreover, several previous studies (Bharvi, Garg & Bali, 2003; Kronegger, Ferligoj and Doreian, 2011; Henriksen, 2018) have indicated that the tendency for collaboration and co-authorship among researchers has increased.

Centrality Measures

Centrality is considered as one of the most important and frequently used conceptual tools for investigating actor roles in social networks (Ni, Sugimoto & Jiang, 2011). Centrality measures (degree, betweenness, and closeness) are used to find out the patterns of connection and communication. Degree specifies the number of collaborators (Newman, 2001); the authors who have the higher degree are the most active, due to their most ties to other actors in co-authorship network (Wasserman & Faust, 1994, p. 178). Betweenness refers to the total number of shortest paths connecting two nodes and passing through a particular node; this quantity is an indicator of who the most influential people in the network are, the ones who control the flow of information between most others (Newman, 2001). The frequency of lying on the shortest path between two nodes determines the control over the interaction between them; in other words, the more a node lies on the shortest path between two other nodes, the more control it has over the interaction between the two non-adjacent nodes. Closeness centrality of authors indicates how close a node in a network is to other nodes (Wasserman & Faust, 1994, p. 165). In a network closeness centrality designates the extent of influence of a node (Ni, Sugimoto & Jiang, 2011). Valente, Coronges, Lakon and Costenbader (2008) reported strong correlations among the centrality measures, although the quantity of correlations was varied for these indicators. The degree of correlation among degree, betweenness, closeness, and eigenvector shows the distinctness and relation of these measures simultaneously. Based on Spearman test, Meghanathan (2016) discovered a very strong and positive association between the degree and closeness centrality metrics.

Zipf's Law

With reference to Zipf's Law, the most popular word is supposed to be used twice as often as the second most popular, three times as often as the third popular and so forth (Grobman & Cerra, 2016, 187). Put

differently, it is a law for determining the relationship of word frequency and its rank (Sahoo & Bhui, 2018). In Zipf's Law the frequency of words is ranked from the most frequent to the least one; it also reckons the consistency of the values produced by multiplying frequencies and rank numbers (Zipf, 1949).

To find the most frequently used keywords by LIS researchers on their public library research, Sahoo and Bhui (2018) used Zipf's Law. Ciftci et al., (2016) too, had a bibliometric analysis in educational sciences and teacher education. Although they referred to the Zipf's Law word frequency in the title of their article, the data they provided was not compatible with the Law. Robles (2019) intended to measure Zipf's Law's efficiency as a pre-processing phase for classifying websites into four categories; as a small sample cannot follow Zipf's distribution accurately, the sites with less than 300 words were removed and thus the accuracy increased to 93.2%. As the intention of Corral, Boleda and Ferrer-i-Cancho (2015) was finding a very long text by a single author to apply Zipf's Law for word frequency, they searched for the longest literary texts ever written. They used Zipf's Law to analyze numerous long literary texts in four languages, comprising varied levels of morphological complexity and in all cases the frequency of words was compatible with Zipf's Law.

Purpose of the Study

This study aims to visualize and analyze the co-authorship networks in the Scopus database's social engineering research from 1926-2020. In order to address the main objective, the following sub-objectives have been defined:

- To report the descriptive characteristics of documents
- To identify the co-authorship pattern and CC for authors
- To indicate the association between the number of authors per paper and publishing year
- To indicate the visualization of the co-authorship network (authors and countries) and co-word network
- To test the association between keyword occurrence and total link strength (TLS)
- To test the association between centrality measures
- To indicate the compatibility of keywords' frequency based on Zipf's Law

- To indicate top authors based on centrality measures of the co-authorship network (degree, betweenness and closeness)

Methods

The present quantitative study used the bibliometric method and social network analysis. Data were collected from Scopus databases using a query on "Social engineering" from 1926 to 24 August 2020. A total number of 2246 records was retrieved; then, the researchers limited documents to journal papers, conference papers, books, and book chapters in the following stage. The result was 1994 records that were saved as CSV file format to be used in bibliometric software. Data gathering was carried out on 24 August 2020. It should be noted that due to wrong or incomplete recording of the affiliation information in country field, some retrieved words were not names of the country; these keywords were manually omitted from the main file.

In the present study the compatibility of Zipf's Law with keyword frequency was measured. Also the formula created by Ajiferuke, Burell and Tague (1988) was used to obtain the CC. CC is a measure of collaborative strength in a discipline that has the merit of lying between 0 and 1 and tends to zero as single-authored papers dominate. Based on this formula, each paper carries a single "credit" with it, this credit being shared among the authors. Thus if a paper has a single author, the author receives one credit; with two authors, each receives 1/2 credits, and with three authors, each receives 1/3 credits and so forth. Furthermore, the centrality measures (degree, betweenness and closeness) for top ten authors in co-authorship network were computed.

Researchers used descriptive statistics (frequency and percentages), inferential ones (Spearman correlation) and social network analysis to obtain results. In the present study, different software types were used (SPSS 20 for descriptive and inferential statistics, Microsoft Excel 2010 to draw the graphs, Pajek for centrality measures, Text Statistics Analyzer and ISI.exe to indicate the co-authorship pattern and VOSviewer 1.6.15 for visualization).

Results

In this section, the researchers present the findings of the study based on research sub-objectives.

Descriptive Characteristics of Documents

Published items in the paper format are more than the book; furthermore, most papers are articles published in journals. Authors publish a majority of documents in English; the total number of documents based on language is more than 1994 cases; probably this is due to the publication of few documents in both English and local languages.

For 73 years, from 1926 to 1999, only 172 papers have been published on SE. Steadily during the following years, the number of publications has increased. Authors have published the highest number of papers in 2019. The researchers collected study data on 24 August 2020. However, up to the end of 2020 number of papers in this area grew.

Published documents on the SE area are related to other 25 subject areas. The top subject areas related to SE were computer science, social science, engineering, art and humanities, mathematics, business, management and accounting, and decision sciences. Readers should note that as subject areas overlap, researchers mentioned a document in different subjects.

Authors have used 159 sources to publish 1994 documents in the SE area. Readers should note that the publishing source with four documents and less has not been considered in this study. The sources had been published in a range of 60 documents to 2 documents. The top three source of publication were "Lecture note in Computer Science", "ACM International Conference" and "Advanced in Intelligence Systems & Computing".

Co-authorship Pattern and CC of Authors

The graph (Figure 2) shows the number of papers on the vertical axis and the number of authors on the horizontal axis. A majority of 859 papers had only one author; in total, the rest of the papers (1135 out of 1994) had more than one author; furthermore, one paper had 11 authors, and one paper had ten authors. Additionally, in the present study, the CC among researchers was 0.36, which is a sign of a tendency for single author in this domain.

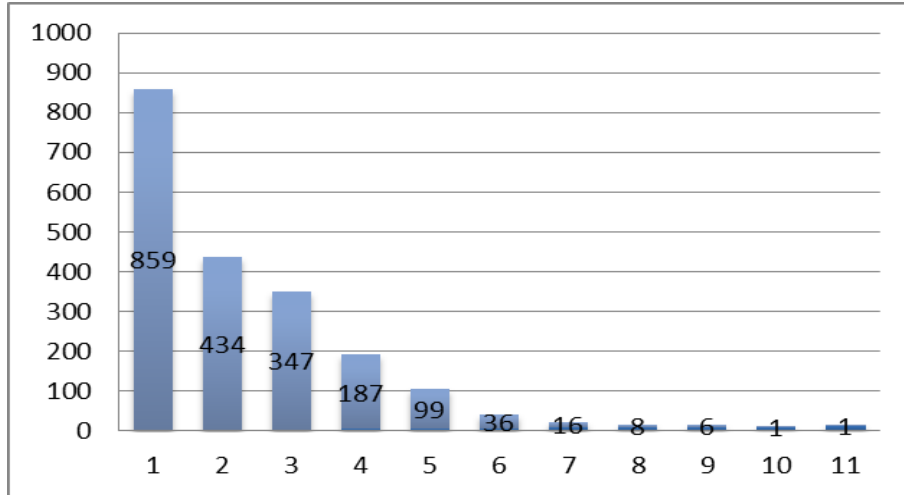


Figure 2. Co-authorship pattern

Association between Number of Authors per Paper and Year

Researchers used Spearman's coefficient to test the correlation between the number of authors per paper and the publication year. The result ($r= 0.336$, $P\text{-value}= 0.000$) indicates that with 99 percent confidence, there was a significant and positive correlation between two indicators. With the rise of the year, the number of authors per paper is increased (Table 1).

Table 1. Spearman correlation between author number and year

		Author number	Year
Author number	Correlation Coefficient	1.000	.336**
	Sig. (2-tailed)	.	.000

** . Correlation is significant at the 0.01 level (2-tailed).

Visualization of Co-authorship

Visualization and analysis of co-authorship network were carried out based on authors and countries using VOSWiever.

Visualization of Co-authorship Based on Authors

Based on findings, 1994 documents on social engineering were written by 3672 authors. The Largest connected co-authorship network for 3672 authors consisted of 126 authors, 14 clusters, 347 links and TLS of 383. The TLS or total link strength shows the strength of the co-authorship links between nodes in the network. In

other words, the TLS shows the total strength of the co-authorship links of a particular researcher with other researchers in the networks (Van Eck & Waltman, 2017).

Due to the high number of authors, VOSviewer, by default, considers the top 1000 authors with the highest TLS; therefore, the TLS of the co-authorship links for each of 3672 authors, calculated by VOSviewer, and the authors with the most remarkable link strength were selected for visualization. Figure 3 indicates the top authors with the most remarkable link strength and the highest number of papers. This network had 1000 authors, 186 clusters, 2443 links and 2942 TLS. The authors in the red areas and large fonts are the core authors.

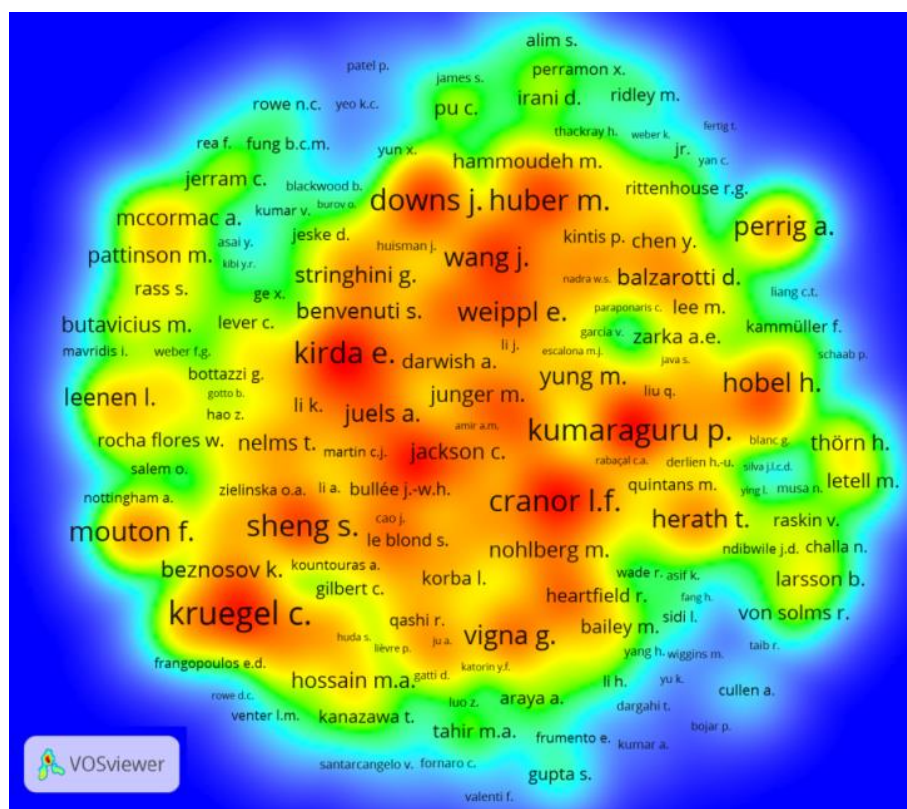


Figure 3. Co-authorship network of 1000 authors with most citations

Visualization of Co-authorship Based on Countries

One hundred countries had participated in writing 1994 documents; this network had 32 clusters, 227 links and its TLS was 357. Figure 4

indicates the co-authorship network based on number of received citations. The largest set of connected countries consists of 83 items and 15 clusters; some of the 100 countries in the network are not connected to each other. Based on the number of documents, authors from the USA, UK and India had the most co-authorship; also based on the number of citation, the USA, UK and Canada, and finally in terms of TLS, US, UK and Germany had the most co-authorship.

In the co-authorship network, the USA, with 511 publications, was present in 13 clusters and was connected to 41 countries; the USA's TLS with other countries in the network was 103, and this core country received 6526 citations. The UK, the second core country in co-authorship networks, had 224 documents and 2178 citations; the UK, with 34 links and 81 TLS, had a presence in 8 clusters of networks. Germany with 75 documents and 19 links co-authored with 5 clusters; TLS of this country with other countries in the network was 36 and received 497 citations. India, the other core country in co-authorship networks, had 119 documents and 871 citations; India, with 15 links and 17 TLS, had the presence in 2 clusters of the networks. With 65 documents and 950 citations, Canada was the third country based on the number of received citations; this country with 10 links and 17 TLS had a presence in 10 clusters of networks. Readers should note that the TLS shows the total strength of the co-authorship links of a specific country with other countries in the networks.

Frequency of Authors' Keywords

Table 3 indicates 126 author keywords that are used five or more than five times. The keywords "social engineering", "phishing", and "information security", respectively with 516, 187 and 85 times frequencies, were the most common used keywords. It should be noted that the frequency of some words involved compounds .(anti-phishing/anti phishing, cybercrime/cybercrimes, cybersecurity/ cyber security, graphical password/graphical passwords, human factor/human factors, internet of things/IoT, phishing attack/phishing attacks, social engineering attack/social engineering attacks, and social network/ social networks)

Table 3. Author Keywords with five times frequency and more

Keyword	N	Keyword	N	Keyword	N	Keyword	N
social engineering	516	internet of things	11	spear phishing	8	social science	6
Phishing	187	Facebook	10	Threats	7	Training	
information security	85	information security awareness		Android		Twitter	
cyber security	75	network security		big data		unidirectional communication	
Security	74	online social networks		computer security	vulnerability analysis		
Malware	41	ransomware		data security	welfare state		
machine learning	36	south Africa		Passwords	Botnet		
social engineering attacks		technology		risk analysis	critical infrastructure		
social networks		web security		risk management	Cyberspace		
human factors	29	advanced persistent threat		9	Attack	6	Democracy
Privacy	26	apt			bidirectional communication		e-commerce
cyber crime	24	culture	Community		Engineering		
Deception		data mining	deep learning		Ethnicity		
security awareness		ethics	Ideology		Firewall		
Anti-phishing		23	internet		indirect communication		Impersonation
							5

Keyword	N	Keyword	N	Keyword	N	Keyword	N
vulnerability/ vulnerabilities	20	intrusion detection	8	information assurance		industrial control systems	
Authentication	19	persuasion		insider threat		Information	
Spam	18	phishing detection		Migration		information security culture	
identity theft	17	risk		Nationalism		Neoliberalism	
Trust	15	cryptography		physical security		neural network	
Education	14	detection		Race		Participation	
social media		development		Research		Personality	
Fraud	13	email		Scada		Religion	
phishing attacks	12	encryption	Science	risk assessment			
social networking sites		hacking	shoulder surfing	Simulation			
Awareness	11	modernity	Singapore	social engineering attack framework			
Classification		penetration testing	social engineering attack detection model	social marketing			
cloud computing		pharming	social networking	Taxonomy			
graphical passwords		psychology	social psychology	Technocracy			

Association Between Keyword Occurrence and TLS

TLS indicates the number of documents in which two keywords occur together. Using Spearman ratio, the association between keyword occurrence and TLS, as presented in table 4, is a positive and significant relationship ($r=.700$). It means the keywords that occur more have higher TLS value.

Table.4 Correlation between keyword occurrence and TLS

Keyword occurrences	TLS	
	Correlation Coefficient	.700**
Sig. (2-tailed)	.000	

** . Correlation is significant at the 0.01 level (2-tailed).

Zipf's Law

Furthermore, compatibility of keywords' frequency with Zipf's Law was analyzed. The figure 6 indicates the rank and frequency of keywords; due to the wide range of keywords with low frequency, only the keywords with fifteen frequencies and more were considered to be checked with Zipf's Law. The most frequent word here "social engineering" should be twice as frequent as the second popular word, three times as frequent as the third popular words and so on. As seen in the figure, data does not conform to Zipf's Law. However with excluding "social engineering" and considering "phishing" as the most frequent keyword, it almost is twice as frequent as "information security".

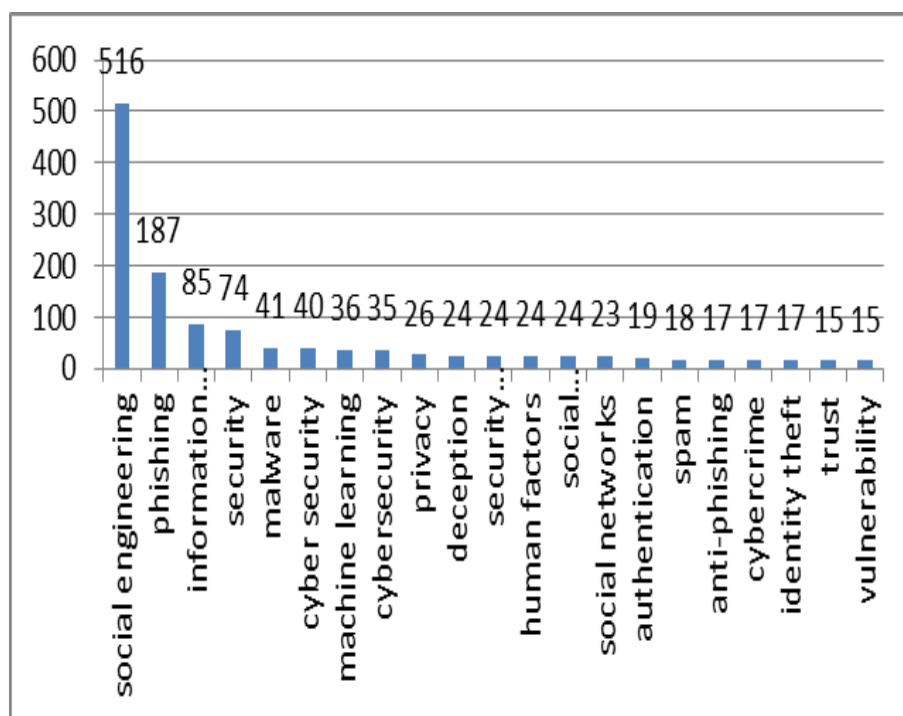


Figure 6. Frequency of keywords based on Zipf's Law

Centrality Measures

Top 10 authors, based on centrality measures (degree, betweenness and closeness), are presented in table 5. As the data of table shows Kumar has the highest number of links with other authors; this author

is among top 10 based on closeness centrality. Based on betweenness and closeness Li is in the first rank; this author acts as a hub in network and links two sections of it. It should be noted that authors with degree 17 were 17 cases and in the table only names of nine of them have been mentioned. Six top authors, based on betweenness and closeness, are in bold color in the table.

Table 5. Top Central Authors

Author	Degree	Author	Betweenness	Author	Closeness
Kumar V.	25	Li B.	.004403	Li B.	.032045
Asai Y.	17	Wang Y.	.004271	Lu L.	.031000
Bhardwaj	17	Lu L.	.003787	Perdisci R.	.030340
Bhattach A.	17	Kirda E.	.003406	Wang Y.	.028905
Brahmach AS.	17	Perdisci R.	.002952	Li K.	.028238
Ghosh S.	17	Antonaka KM	.002421	Neasbitt C.	.028238
Jain A.	17	Chen Y.	.001547	Singh K.	.028145
Kitano H.	17	Zhang Y.	.001498	Kirda E.	.026654
Matsuoka Y.	17	Liu L.	.001424	Kumar V.	.026000
Mondal A.	17	Huang H.	.001111	Antonaka KM.	.025694

In order to find out whether there is any association between centrality measures, Spearman test was used. The results indicate a positive and significant association between three indicators. The relationship between closeness and degree ($r = .871$) was higher than the associations of other indicators (Table 6).

Table 6. Spearman's rho for centrality measures

Betweenness & closeness	Correlation Coefficient	.254**
	Sig. (2-tailed)	.000
Closeness & degree	Correlation Coefficient	.871**
	Sig. (2-tailed)	.000
Betweenness & degree	Correlation Coefficient	.259**
	Sig. (2-tailed)	.000

** . Correlation is significant at the 0.01 level (2-tailed).

Discussion

In the 21st century knowledge is a crucial factor in societies. At present, many of us are using social networking sites to maintain social relationships and exchange data. Knowingly or unknowingly, we exchange our personal and private information through this social networking. Attackers target this human weakness. Manipulation of human trust to obtain information and get monetary gains or other benefits through that information is called social engineering. Due to the worldwide spread of social engineering, there is a necessity to know about the efforts in scientific communities in this area. The present study aims to analyze social engineering publications indexed in Scopus from 1926 to August 2020.

Out of four formats (journal paper, conference paper, book and book chapter) studied in the present study, most published documents were papers, especially journal papers. Nonetheless, Firdaus et al., (2019) in their bibliometric study pertaining to blockchain research, found that researchers were most interested to publish their work in conference rather than journal or in book form. It seems that the conferences are the first channel for the researchers to disseminate their new ideas. Authors have published many documents in English; as the English language has the most scientific audience, the result is rational. For 73 years, from 1926 to 1999, only 172 papers were published on social engineering. Steadily during the subsequent years, the number of publications has increased. In line with this finding, Persson et al., (2004) reported a growth in the number of papers over the years.

The top three sources for publishing social engineering outlets were from computer areas. The top three subject areas for publishing social engineering documents were computer science, social science, and engineering. Subject relevance of social engineering documents with 25 areas indicates that researchers from different domains were interested in researching this area.

Majority of papers (859) had only one author. However, 1135 out of 1994 documents were written by more than one author; many researchers tended to collaborate with other researchers. Based on the present study's finding, one of the main co-authorship patterns in this area seems to be one author. As mentioned above CC is a measure of collaborative strength in a discipline and ranges between 0 and 1. In the present study, the quantity of the CC was also a sign of a tendency

toward single author in this domain. Researchers of the present study tested an assumption to know whether the authorship pattern has changed during recent years. Spearman test indicates a significant and positive correlation between the number of authors per paper and the publication year. It means the number of authors per paper during recent years has increased; in other words, it seems researchers in the early years of the emergence of the idea of social engineering had primarily published single-author papers. In recent years, the publishing pattern has changed to more than one author. In line with the findings of this study, previous studies (Bharvi et al., 2003; Glanzel & Schubert, 2004; Persson et al., 2004; Kronegger et al., 2011; Henriksen, 2018) have also reported an increased tendency for co-authorship among researchers. As Persson et al., (2004) concluded, the interpretation of this tendency probably is the change in the patterns of scientific communication and collaboration in the last two decades.

Based on the study's findings, 1994 documents on social engineering were written by 3672 authors and in 100 countries. The top three researchers, based on the number of documents, were Mouton F., Venter H.S., and Algarni A. and Korda E. together in the third rank. The top three researchers based on the number of citations were Kruegel C., Kumaragruru P. and Kirda E. with 324, 275, and 274 citations. The mentioned researchers are top and core nodes of co-authorship networks in the social engineering area. Without these core nodes, the co-authorship network will disintegrate.

The top three countries based on the number of citation were respectively the USA, UK and Canada; also the top three countries based on the number of documents were the USA, UK and India respectively; moreover, the USA, UK and Germany were the top three countries based on TLS. It signifies these leading countries have a crucial role in co-authorship networks of the social engineering area. The study concludes that without these elite researchers and countries, the co-authorship networks will disintegrate. In line with these findings, Firdaus et al., (2019) in a bibliometric study, indicated that the most active country in blockchain publication was the USA. Also the finding of a bibliometric study on android malware research by Mat et al., (2020) revealed that the USA and India had the highest publication respectively. Furthermore, in the present study the top 10 highly cited countries were the USA, UK, Canada, Australia, India,

South Africa, Austria, Germany, Netherlands and France. Also, Bahrami & Rouzbahani (2021) in their bibliometric study indicated that Germany, China, and Italy were the leading and foremost countries in having done significant research into cyber security of smart manufacturing. It should be noted that, based on country ranking in terms of citation on Scimago from 1996 to 2020, all of the mentioned countries except South Africa, Austria and Italy were among highly cited countries; however, Italy, based on 2020 report of this site, was in the fifth rank (Scimago, 2020); it signifies that some countries are in the vanguard in most scientific fields.

Furthermore, the findings regarding core authors and core countries are compatible with Structural Hole theory. The theory developed by Ronald Stuart Burt in 1992 and indicates that nodes occupying the bridging positions between different groups have advantages since they control the key information diffusion paths (Lin et al., 2021) in the co-authorship network.

Based on the results of Spearman test, the study concludes that with the increasing number of documents per country, the number of citations and the rate of TLS have increased. This situation means that the leading countries, based on the number of published documents, citations and TLS, are in a good position in co-authorship network.

In the co-authorship network, centrality measures (degree, betweenness and closeness) are used to understand the patterns of connection and communication between authors. Based on findings, in terms of degree centrality, Kumar had the most ties to other authors in the co-authorship network. Based on betweenness and closeness, Li was the top key author; this author acts as a hub in network and links two sections of the network. Li was the most influential author in the network, the one who controls the flow of information between most others. Furthermore, although the relationship between three centrality indicators was positive and significant, the relationship between closeness and degree was higher than associations of other indicators. This means the authors who are close to all the other authors in a network and authors who are on shortest paths between pairs of authors have stronger relationship. This finding agrees with the finding of Meghanathan (2016) that showed based on Spearman correlation, there was a very strong and positive association between the degree and closeness centralities. Also Valente et al. (2008) has reported strong correlations among the centrality measures.

The researchers' most common keywords were social engineering, phishing, and information security. The frequent keywords in social engineering research may help experts design a better taxonomy in this area. Based on findings, the keywords frequency was not compatible with Zipf's Law. In line with this finding, previous bibliometric studies (Sahoo & Bhui; Ciftci et al., 2016) were not fitted with Zipf's Law. While Corral et al., (2015) found that the frequency of keywords in a very long text matched with Zipf's Law. Likewise, Robles (2019), in categorizing websites keywords, after removing the low frequency keywords got better fitness with Zipf's Law.

Conclusion

There was an increase in the number of publications during the years. The co-authorship pattern gradually has changed from single author to multi-author over the years. Despite the role of researchers from 100 countries to publish in social engineering area, the researchers from Germany, Canada, India, UK and the US were more productive, especially the US and the UK are more influential in terms of number of citations and documents. These leading and central countries have critical roles in information flow on the s co-authorship networks of social engineering area. Furthermore, it seems that a small sample of keywords will not properly follow the Zipf's distribution.

Readers should take into account that in the present study the main query was replied using only "social engineering" term in database and the other equivalent and synonym words were not included.

Reference

- Ajiferuke, I., Burell, Q. & Tague, J. (1988). Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, 14 (5), 421-433. <https://doi.org/10.1007/BF02017100>
- Arana, M. (2017). How much does a cyberattack cost companies? *Open Data Security*, retrieved from:<https://opendatasecurity.co.uk/how-much-does-a-cyberattack-cost-companies/>
- Bharvi, D., Garg, K., & Bali, A. (2003). Scientometrics of the International Journal Scientometrics. *Scientometrics*, 56(1), 81–93.doi.org/10.1023/A:1021950607895.
- Bahrami A.H. & Rouzbahani H.M. (2021) Cyber Security of Smart Manufacturing Execution Systems: A Bibliometric Analysis. In: Karimipour H., Derakhshan F. (eds) *AI-Enabled Threat Detection and Security Analysis for Industrial IoT*(pp. 105-119). Springer, Cham. https://doi.org/10.1007/978-3-030-76613-9_6
- Chen, C. (2018). Visualizing and exploring scientific literature with Citespace: An introduction. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 369-370). <https://doi.org/10.1145/3176349.3176897>
- Ciftci, S.K., Danisman, S., Yalcin, M., Tosuntas, S.B., Ay, Y., Sölpük, N. & Karadag, E. (2016). Map of Scientific Publication in the Field of Educational Sciences and Teacher Education in Turkey: A Bibliometric Study. *Educational Sciences: Theory and Practice*, 16(4), 1097-1123.
- Corral, A., Boleda, G., & Ferrer-i-Cancho, R. (2015). Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, 10(7), e0129031. <https://doi.org/10.1371/journal.pone.0129031>
- Duff, A. S. (2005). Social Engineering in the Information Age. *The Information Society: An International Journal*, 21(1), 67 - 71. <https://doi.org/10.1080/01972240590895937>
- Firdaus, A., Ab Razak, M. F., Feizollah, A., Hashem, I. A. T., Hazim, M., & Anuar, N. B. (2019). The rise of “blockchain”: bibliometric analysis of blockchain study. *Scientometrics*, 120(3), 1289-1331. <https://doi.org/10.1007/s11192-019-03170-4>
- Gao, W. & Kim, J. (2007). Robbing the cradle is like taking candy from a baby. In *Proceedings of the Annual Conference of the Security Policy Institute (GCSPi)* (Vol. 4, pp. 23-37).
- Glanzel, W. & Schubert, A. (2004). Analysing scientific networks through co-authorship. Moed, H. F., Glanzel, W. and Schmoch, U. (Eds.), *Handbook of quantitative science and technology research*, Springer, Dordrecht, 257-276. https://doi.org/10.1007/1-4020-2755-9_12

- Greavu-Serban, V. & Serban, B. (2014). Social engineering a general approach. *Informatica Economica*, 18(2), 5-14. <https://doi.org/10.12948/issn14531305/18.2.2014.01>
- Grobman, S. & Cerra, A. (2016). Cybersecurity's Second Wind. In *The Second Economy*, Apress, Berkeley, CA., pp.175-189. https://doi.org/10.1007/978-1-4842-2229-4_10
- Hadnagy, C. (2011). *Social Engineering: The Art of Human Hacking*, Indianapolis: Wiley.
- Hansson, S. O. (2006). A note on social engineering and the public perception of technology. *Technology in Society*, 28(3), 389-392. <https://doi.org/10.1016/j.techsoc.2006.06.006>
- Heartfield, R. & Loukas, G. (2015). A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks, *ACM Computing Surveys (CSUR)*, 48(3), 1-39. <https://doi.org/10.1145/2835375>
- Henriksen, D. (2018). What factors are associated with increasing co-authorship in the social sciences? A case study of Danish Economics and Political Science. *Scientometrics*, 114(3), 1395-1421. <https://doi.org/10.1007/s11192-017-2635-0>
- Huber, M., Kowalski, S., Nohlberg, M., & Tjoa, S. (2009). Towards automating social engineering using social networking sites. In *International Conference on Computational Science and Engineering* (Vol. 3, pp. 117-124). IEEE, retrieved from: https://dlwqtxts1xzle7.cloudfront.net/51931971/Towards_Automating_Social_Engineering_Us20170225-6273-1di1smr.pdf. (6 December 2020). <https://doi.org/10.1109/CSE.2009.205>
- Ivaturi, K. & Janczewski, L. (2011). A taxonomy for social engineering attacks. In *International Conference on Information Resources Management*, pp. 1-12. Centre for Information Technology, Organizations, and People, retrieved from: <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1015&context=confirm2011>. (27 October 2019).
- Kalnin, R., Purin, J. & Alksnis, G. (2017). Security evaluation of wireless network access points. *Applied Computer Systems*, 21(1), 38-45. <https://doi.org/10.1515/acss-2017-0005>
- Kronegger, L., Ferligoj, A., & Doreian, P. (2011). On the dynamics of national scientific systems. *Quality & Quantity*, 45(5), 989-1015. doi: 10.1007/s11135-011-9484-3.
- Larabee, L. (2006). *Development of methodical social engineering taxonomy project*. Naval Postgraduates School Monterey CA., retrieved from: <http://handle.dtic.mil/100.2/ADA457544>. (27 October 2019).

- Lee, Y.C., Chen, C. & Tsai, X.T. (2016). Visualizing the knowledge domain of nanoparticle drug delivery technologies: a scientometric review. *Applied Sciences*, 6(1), 11. <https://doi.org/10.3390/app6010011>
- Lineberry, S. (2007). The human element: The weakest link in information security. *Journal of Accountancy*, 204(5)44.
- Lin, Z., Zhang, Y., Gong, Q., Chen, Y., Oksanen, A., & Ding, A. Y. (2021). Structural Hole Theory in Social Network Analysis: A Review. *IEEE Transactions on Computational Social Systems*, doi: 10.1109/TCSS.2021.3070321.
- Mat, S.R.T., Ab Razak, M.F., Kahar, M.N.M., Arif, J.M., Mohamad, S. & Firdaus, A. (2021). Towards a systematic description of the field using bibliometric analysis: malware evolution. *Scientometrics*, 126(3), 2013-2055. <https://doi.org/10.1007/s11192-020-03834-6>
- Meghanathan, N. (2016). A comprehensive analysis of the correlation between maximal clique size and centrality metrics for complex network graphs. *Egyptian Informatics Journal*, retrieved from: <https://www.sciencedirect.com/science/article/pii/S1110866516300305>
- Montanez, R., Golob, E., & Xu, S. (2020). Human Cognition Through the Lens of Social Engineering Cyberattacks. *Frontiers in psychology*, 11, 1755. <https://doi.org/10.3389/fpsyg.2020.01755>
- Newman, M.E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1), 016132. <https://doi.org/10.1103/PhysRevE.64.016132>
- Ni, C., Sugimoto, C. & Jiang, J. (2011). Degree, Closeness, and Betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In *Proceedings of ISSI*, pp. 1-13.
- Olson, C. L. (2019). Social Engineering Attacks by the Numbers: Prevalence, Costs, & Impact. retrieved from: <https://datafloq.com/read/social-engineering-attacks-numbers-cost/6068>. (4 December 2020).
- Oosterloo, B. (2020). *Managing social engineering risk: making social engineering transparent*. Master thesis, University of Twente, Netherlands, retrieved from: https://essay.utwente.nl/59233/1/scriptie_B_Oosterloo.pdf. (23 March 2021).
- Persson, O., Glänzel, W. & Danell, R., (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative

- indicators in evaluative studies. *Scientometrics*, 60(3), 421-432. <https://doi.org/10.1023/B:SCIE.0000034384.35498.7d>
- Priatna, T., Malyawati, D.S., Sugilar, H. & Ramdhani, M.A., (2020). Social Engineering to Establish Digital Culture in Higher Education. *Advances in Science, Technology and Engineering Systems Journal*, 5(6), 1474-1479. <https://doi.org/10.25046/aj0506177>
- Rialti, R., Marzi, G., Ciappei, C. & Busso, D. (2019). Big data and dynamic capabilities: a bibliometric analysis and systematic literature review. *Management Decision*, 57 (8), 2052-2068. <https://doi.org/10.1108/MD-07-2018-0821>
- Robles, A. (2019). Classifying Websites Using Word Vectors and Other Techniques: An Application of Zipf's Law. Phd, California State University, Long Beach.
- Rosenblum, D. (2007). What anyone can know: The privacy risks of social networking sites. *IEEE Security & Privacy*, 5(3) 40-49. <https://doi.org/10.1109/MSP.2007.75>
- Rusch, J. (1999). The "social engineering" of Internet fraud. Paper presented at the 1999 Internet Society's INET'99 conference, San Jose CA, retrieved from: <https://vetu.pw/z4.pdf>, (1 July 2021).
- Saeed, R.A. & Shareef, S.M. (2020). Implementation of Artificial Intelligence to predict threats in social media based on user's behavior. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(5),6931-6938. Retrieved from: <http://www.warse.org/IJATCSE/static/pdf/file/ijatcse10952020.pdf>, (19 March 2021). <https://doi.org/10.30534/ijatcse/2020/10952020>
- Sahoo, S. & Bhui, T. (2018). Trend of Public library research in India: a bibliometric study. *Library Philosophy & Practice*, retrieved from: <https://core.ac.uk/download/pdf/189479386.pdf>, (18 June 2021).
- Townsend, K. (2010). The art of social engineering. *Infosecurity*, 7(4), 32-35. [https://doi.org/10.1016/S1754-4548\(10\)70068-1](https://doi.org/10.1016/S1754-4548(10)70068-1)
- Valente, T.W., Coronges, K., Lakon, C. & Costenbader, E. (2008). How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1)16.
- Van Eck, N.J. & Waltman, L. (23 October, 2017). VOSviewer Manual (Manual for VOSviewer version 1.6. 5). CWTS, Leiden. Retrieved from: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.6.pdf. (10 October 2017).
- Wang, Z, Sun, L. & Zhu, H. (2020). Defining Social Engineering in Cybersecurity. *IEEE Access*, DOI: 10.1109/ACCESS.2020.2992807, retrieved from: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9087851>

- (Accessed 24 March 2021).
<https://doi.org/10.1109/ACCESS.2020.2992807>
- Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*, Cambridge university press, New York.
<https://doi.org/10.1017/CBO9780511815478>
- Zulkiffli, S.N.H., Zawawi, M.N.A. & Rahim, F.A. (2020). Passive and Active Reconnaissance: A Social Engineering Case Study. In *8th International Conference on Information Technology and Multimedia (ICIMU)* (pp. 138-143). IEEE.
<https://doi.org/10.1109/ICIMU49871.2020.9243402>

How to Cite: Khalili, L., Darshani Wijayasundara, N. (2022). Analyzing Social Engineering Research through Co-Authorship Networks Using Scopus Database during 1926-2020, *International Journal of Digital Content Management (IJDCM)*, 2(4), 15-45.

DOI: 10.22054/DCM.2022.14016



International Journal of Digital Content Management (IJDCM) is licensed under a Creative Commons Attribution 4.0 International License.

