DEPARTMENT OF COMPUTER SCIENCE

Cybersecurity Research Datasets: A Replication and Extension

Qihang Zhang

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Friday 24th September, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. I have identified all material in this dissertation which is not my own work through appropriate referencing. Except where specifically acknowledged, it is all the work of the Author.

Qihang Zhang, 24th September 2021

Abstract

The 21st century has seen more data-intensive and collaborative scientific research than in the past. Data sharing is a crucial component of the scientific process, because it enables researchers to verify findings and go beyond previously discovered facts. Therefore, data is no longer purely a result of research but becomes the raw material and driving force of scientific activity. For cybersecurity field, the paper “Cybersecurity Research Datasets: Taxonomy and Empirical Analysis” helps to improve our understanding of how cybersecurity research datasets are created and used, and how to encourage greater sharing among researchers.

This paper replicates and extends the above paper as a means of examining the main findings of the original paper and uncovering the latest features and sharing state of cybersecurity datasets. I start with replication with following every step of the original paper. Based on the successful replication, my research is extended in three aspects. Firstly, I added the recent papers and databases, discovering whether data sharing between researchers has improved and how well the taxonomy developed in the original paper fits with recent studies. As well, two models used in the original paper are optimised, including the binary classifier used to distinguish between the inclusion or exclusion of datasets and the regression model, which can represent the characteristics of cybersecurity datasets more accurately and comprehensively. It is determined that papers that make the created datasets publicly available have higher citation rates, but that the proportion of shared datasets is consistently low. A key to breaking this status quo is to focus on incentives for sharing by removing barriers and rewarding publication. Accordingly, I offer suggestions on how to improve data sharing behaviour in cybersecurity in the future, with the expectation that it will contribute to the development of cybersecurity.

Acknowledgements

I would like to thank my supervisor, Dr Matthew Edwards, for his invaluable guidance and feedback throughout the project and the writing of this dissertation. His insights and interaction consistently helped me with the technical and writing issues, which considerably improve the quality of this project. Thanks also to Dr Neill Campbell for taking the time to review my thesis.

In addition, I am grateful to my parent and friends who encouraged me during the entire year of MSc and cheered me on no matter what. Thank you for making this stressful time so much easier. I would also like to thank my fellow students in the Bristol Computer Science Department, with whom the interaction and discussions have eased my anxiety in studying. This thesis may not be successfully completed without all these people and more, and I will be forever grateful to everyone who has helped me.

COVID-19 Statement

This project was largely unaffected by the COVID-19 pandemic, except for loss of time during establishing remote connections. There's unstable connection regarding crawling papers from DBLP and Google scholar due to the remotely work.

Contents

Chapter 1

Introduction

Since the formal advent of the Internet in the late 1980s, many routine, mundane tasks have been simplified by Internet's availability. However, it has become an area of chaos, with many unscrupulous people using it for criminal purposes. Nowadays, cyber-attacks and cybercrime are common, increasing in frequency and severity, which threaten various aspects, including politics, economics and everyone's interests. Therefore, cybersecurity has received increasing attention.

With the arrival of the Big Data era, cybersecurity research and practice are becoming more data-oriented. For example, cybercrime indicators can be used to quantify risk better, which could provide a basis for proactive defence based on previous targets. Researchers have been using cybersecurity datasets as inputs to their work and as outputs of their research for many years, but these datasets are not always shared with the broader research community. In the "Cybersecurity Research Datasets: Taxonomy and Empirical Analysis", Zheng et al. [1] conduct a statistical and regression analysis of top computer security publications from 2012 to 2016 to construct a taxonomy of cybersecurity datasets and examine the use and creation of data. The study on thousands of research papers found that three-quarters of the existing datasets used by researchers in their papers were publicly available, but less than one-fifth of the datasets created by researchers were publicly shared. Despite the increasing focus on cybersecurity research involving data and the exhortations to share datasets publicly, the proportion of publicly shared datasets is consistently low. Their paper also identified that papers making the created datasets publicly available had higher citation rates by using linear regression. Therefore, they argue that incentives for researchers to share datasets with the wider research community are underappreciated, and it is critical to focus on incentives for sharing by removing barriers and rewarding publication. This finding started to shift the argument about data sharing from community service to individual rationality. When individual rationality starts to be taken seriously, more researchers are believed to choose to made datasets public so as to promote the field of cybersecurity.

1.1 Aims and Objectives

Zheng et al. have contributed to improving our understanding of which datasets are created, how they are used, and how to encourage greater sharing among researchers. By replicating the work in the original paper, I aim to confirm the study's key findings and demonstrate that they are unaffected by any methodological threats. Building on this, I added the recent papers and databases, discovering whether data sharing between researchers has improved and how well the taxonomy developed in Zheng's paper fits with recent studies. As well, the models used in the original paper are optimised, including the binary classifier used to distinguish between containing datasets or not and the regression model, which can be used to represent the characteristics of cybersecurity datasets more accurately and comprehensively. Finally, I propose my suggestions for the current state of cybersecurity datasets.

1.2 Structure of Thesis

Chapter 2 reviews the existing literature on cybersecurity datasets, focusing on data sharing and taxonomy. In **Chapter 3**, a replication of the original paper "Cybersecurity Research Datasets: Taxonomy and Empirical Analysis" is undertaken, including the collection of datasets, the construction and analysis of taxonomies, and regression analysis to identify the motivations for making datasets publicly available to others. **Chapter 4** extends the original paper by re-analysing the latest databases and studies with an optimisation approach to identify any improvements in data sharing and the extent to which the taxonomy fits the latest studies. **Chapter 5** summarises the entire paper and makes recommendations in terms of incentives for researchers to share. Finally, **Chapter 6** identifies the weaknesses of the paper and provides an overview of the future.

Chapter 2

Background

2.1 Data-sharing on Cybersecurity

The 21st century has seen more data-intensive and collaborative scientific research than in the past. Data sharing is a crucial component of the scientific process, because it enables researchers to verify findings and go beyond previously discovered facts. Therefore, data is no longer purely a result of research but becomes the raw material and driving force of scientific activity. The sharing and reuse of research data are considered an important driver of scientific innovation and knowledge discovery. However, “scientific data are being lost at a rapid rate and in 20 years, 80% of data will be inaccessible and unusable” [2]. The collection, preservation and sharing of research data have become an urgent and complicated issue today.

Sharing scientific data can definitely increase researchers' productivity and academic influence[3]. This is done by demonstrating the quality of research data and research; enhancing reputation of researchers and their institution; facilitating communication and collaboration with other researchers and attracting more funders and partners; managing data more effectively to ensure long-term preservation and integrity; and saving time and effort for subsequent research.

Unfortunately, since cybersecurity data are not often shared with the wider academic community, reproducing discoveries and generating innovations using current data are difficult or impossible. In Laube and Böhme’s [4] research on the sharing of cybersecurity data, they identified a framework for analysing defenders’ strategies to share cybersecurity information privately or publicly and found that cybersecurity defenders were more reluctant to share information than would be socially expected, as the decision to share were motivated by selfish rather than altruistic reasons, although those who do share cybersecurity information can profit from it. Generally speaking, data sharing behaviour benefits others first before it benefits researchers themselves. So, if researchers make it their code of conduct to benefit others, they are more likely to share research data and are more likely to share data on a pro bono basis.

A number of challenges could contribute to this situation, legal and privacy concerns that hinder sharing are commonly brought up. It is worth noting that the motivation for sharing data is not powerful. When it comes to sharing data, competitive considerations can discourage companies and academics providing security products or services from sharing and negatively impact information-sharing incentives. Evidence presented by Moore and Clayton [5] suggests that rivalry between security services providers can detract from sharing data, thereby undermining overall security. Additionally, they discovered that firms who clean up websites with phishing content are unlikely to exchange data with rivals, which results in unnecessary delays in remediating impacted websites. Furthermore, unlike many forms of data, Zheng et al. [1] indicated that security datasets are often seen as especially sensitive since many cybersecurity data may contain some private and harmful data, which may adversely affect others or assist in criminal activities.

Economic analysis of cybersecurity provides further insight into attackers' and defenders' behaviour. A fundamental hurdle is that sharing can be expensive, while the benefits largely flow to others. Gordon et al. [6] observed that sharing increases the probability that companies would make the best and economic investment on security. Additionally, the data incentive was found to be inadequate, with companies being tempted to free-ride on those who shared their data if no coordination was there. Gal-Or and Ghose [7] argue that sharing information is more beneficial with the high substitutability of products, indicating that such collaborations benefit more in highly competitive industries.

Therefore, an increasing number of researchers have been focusing on how to address the current state of data sharing. Pete and Chua [8] focused on the usability of cybersecurity data. They found that addressing the technical issues, like the setup and download of datasets as well as accessibility of big data technologies, would facilitate cybersecurity dataset adoption in the wider research community. Coull and Kenneally [9] propose a framework for implementing disclosure controls that comprehensively addresses data sharing risks by synergistically considering both policy-level risks and technical-level disclosure issues (e.g., data anonymisation). In terms of incentives, Costello [10] points out the lack of incentives for sharing data, which leads to a low level of motivation for researchers to participate in creating and managing data. For such cases, Wang [11] proposes an incentive mechanism to give proper recognition to those directly involved in the design and production of the data by using data publication, avoiding drowning genuine contributors in a long list of authors. Zheng et al. [1] focused on cybersecurity data sharing among researchers and found that three-quarters of the existing datasets used by researchers in their papers were publicly available, but less than one-fifth of the datasets created by researchers were publicly shared. This demonstrates severe structural imbalances in the supply and demand of cybersecurity research data. In order to overcome these barriers, incentives need to be aligned for producers, seekers and beneficiaries.

2.2 Taxonomy of Cybersecurity Datasets

As the Internet becomes a part of day-to-day activities, cyber-attacks have increased significantly, and attackers have progressively improved and introduced innovative attack methods. According to the 2020 Cyber Security Risk Report [12], cybercrime will cost nearly $6 trillion per year by 2021. To help identify and defend against cyber-attacks, a growing number of researchers are exploring the taxonomies of cybersecurity.

Some databases use protocols or other technical features of the data to categorise their datasets, such as BGP or DNS data [13]. It is an obvious approach to group datasets, but this method presents certain cybersecurity concerns when applying it to datasets. For instance, hijacking reports and route announcements for BGP are quite distinct beyond protocol; nonetheless, they are both regarded as BGP data. Researchers who investigate Internet disruptions or attacks would be interested in BGP hijacking reports, whereas those who study Internet topology would be interested in BGP route announcements.

In addition, researchers have always preferred to focus on cyber-attacks and try to classify them, mainly because most papers and studies are conducted concerning cyber-attacks. A taxonomy of four separate dimensions that include network and computer attacks was developed by Hansman and Hunt [14]. The attack vector, the first dimension, was utilised to categorise the attack. The second dimension classified the attack's target. The third dimension focused on Howard’s taxonomy [15] and was made up of his vulnerability classification number and criteria. The final dimension emphasised the payload or associated effects. Various layers of information were provided inside each dimension to provide characteristics of attack data, thus improving the knowledge of computer and network security. Meyers et al. [16] presented a taxonomy for cyber-attacks comparable to Hansman and Hunt [14], which classified cyber-attacks into nine categories.

As can be seen, most existing data taxonomy methods are based on the attack aspect, but this paper focuses on the larger area of cybersecurity. Of course, there are other researchers who divide the cybersecurity datasets from a broader perspective. For example, researchers from other fields, such as psychology, policy, management, and others, could also see the value of cybersecurity due to its multidisciplinary character, according to Suryotrisongko et al. [17]. So, in order to uncover the entire cybersecurity research field, they examined 99 papers selected from various cybersecurity publications and broke down the research subjects into eight categories, which include not only computer technology security, but also human/social security, system/technology, and so on.

Taken together, in Zheng's paper, the data is divided into four main categories: attacker-related datasets; defender artifacts; end user and organisation characteristics; and macro-level Internet characteristics, which focus on a more comprehensive and relevant dataset describing the context of cybersecurity research to figure out the unique characteristics of cybersecurity research datasets.

Chapter 3

Study Replication

In this chapter, I replicated the analysis of the original paper as closely as possible, including the data collection and processing, taxonomy construction, as well as statistical and regression analysis of datasets. As the original paper did not provide a detailed description of the methodology, I elaborated on the methods used to complete the previously omitted sections.

When the replication is complete, I first correct some issues encountered in this chapter and then refine the model to obtain more accurate findings. Finally, the characteristics of the cybersecurity dataset are further analysed with updated models and expanded samples by adding the recent data from 2017-2020. Such extensions will be described in Chapter 4.

3.1 The Sample

3.1.1 Data Sources

As described in the original paper, I first selected the four most well-known computer security research conferences: IEEE Symposium on Security and Privacy (S&P), USENIX Security Symposium (USENIX), ACM Conference on Computer and Communications Security (CCS), as well as Network and Distributed System Security Symposium (NDSS). It was then complemented with outlets that regularly publish data-intensive cybersecurity papers: the Workshop on the Economics of Information Security (WEIS), Internet Measurement Conference (IMC), as well as International Conference on Financial Cryptography and Data Security (FC). Finally, the proceedings incorporate workshops associated with top conferences: the AI & Security Workshop at CCS (AISEC), Cyber Security Experimentation and Test (CSET) Workshop at USENIX Security, and the Workshop on Bitcoin and Blockchain Research at FC (BITCOIN).

3.1.2 Data Collection

I started by downloading all the publications from 2012 to 2016 and obtained information on their citations. All papers could be found in DBLP [18], but URLs for a few conferences are now unavailable. For these problematic papers, I used Google Scholar to search them from alternative sources. In addition, as these papers are somewhat out of date, most of the URLs redirect to the latest pages, so we got the latest URLs by a posting request and fetched paper names and download links on them via the Request and BeautifulSoup libraries for Python. The Request library is a module for accessing the web, which can send a request with parameters to a web address and get the required information. And BeautifulSoup can extract data from HTML or XML files, in particular, it can parse the HTML file obtained by Request into a tree structure and then easily get the corresponding attributes of a given tag.

During the download process, there are different downloading logics for different conferences, which are mainly divided into five platforms: IEEE Xplore, ACM Digital Library, NDSS Symposium, USENIX, and Springer. Consequently, targeted download methods are designed for them respectively, and my program will automatically select the method when it recognises the platform. Of these, IEEE and ACM require an authorised login to download, so we manually record cookiess and update them at regular intervals to ensure a stable download process. The purpose of the cookies is to inform the website about the login information when getting a page with Request.

Once papers have been downloaded, I perform two tests on them. The first is to check the size of the downloaded file, a file that is too small usually implies that there was a problem with the download process and that the complete paper was not downloaded locally. The other is to ensure that the file is the thesis itself and not other material such as a PowerPoint presentation. If either of these occurs, I manually verify the download website of these papers and re-download them.

Eventually, we crawled 2,212 papers from their corresponding websites. Of which, most papers were published on CCS with 37.22%. USENIX, NDSS, SP, and IMC followed behind, with similar numbers of papers. The least number of papers were published in workshops. Among them, CSET has no workshop report in 2015 and BITCOIN has only been in operation since 2014.

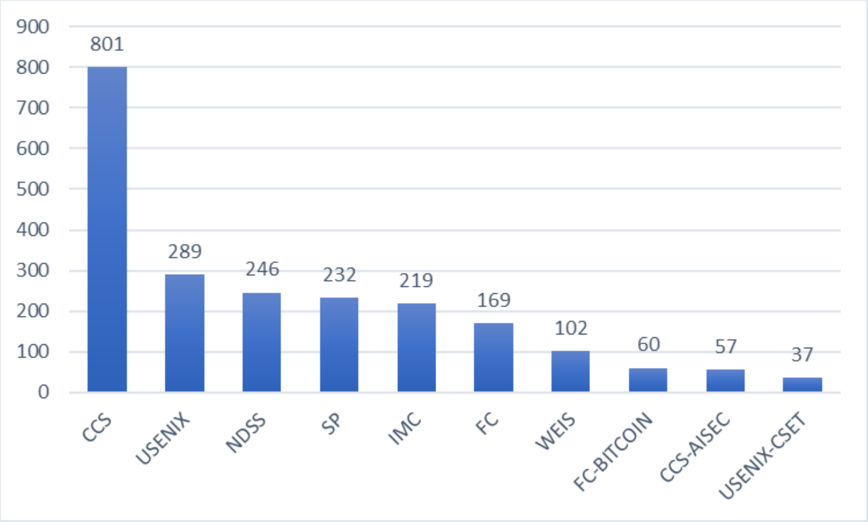


Figure 3.1: Distributions of all crawled papers based on the conference (2012-2016)

In terms of year, there is a general upward trend from 338 papers in 2012 to 526 in 2016. This is due to the increasing popularity of the Internet and the rising importance of cybersecurity. As seen in the graph, the majority of conference papers have an upward trend.

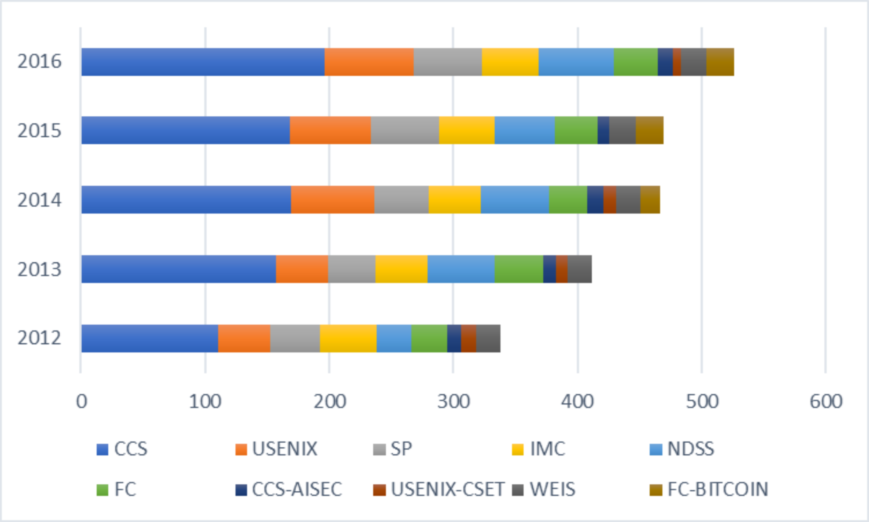


Figure 3.2: Distributions of all crawled papers over years (2012-2016)

Then, a selenium library for Python was used to automatically crawl all papers' citation information from Google Scholar, which is an automated testing tool for the web. This was achieved by running scholar on Firefox using Python as well as by simulating input and click operations through the program to get the HTML information of the page, finally using BeautifulSoup and the re module in Python to find the citations in HTML information. However, as Google Scholar is sensitive to machine crawling and intercepts it, a human-machine test has to be repeated after every few hundred papers are counted.

3.2 Data Processing

For Zheng’s paper, a binary classifier was constructed to distinguish dataset-related papers and non-dataset-related papers. Dataset papers are defined as those with at least one dataset used or created during the research. Non-dataset papers are papers that do not include a dataset as defined above. 957 articles have been categorised by manual and machine in the original paper according to whether they contain data or not. However, upon examination, it is found that some of the papers had duplicates, for example, “Analysis of a ‘/0’? Stealth Scan from a Botnet”, “Analysis of a ‘/0’ Stealth Scan from a Botnet”, “Analysis of a ___/0 __? Stealth Scan from a Botnet” were considered to be three papers in the original classified papers, probably due to a programming error in the download at the time of crawling. The incorrectly formatted name first caused the paper to be unrecognised when cited from Google Scholar, which was shown as having 0 citation in the original dataset, when in fact it had more citations than most similar papers. In addition, a few papers had some problems with classification due to duplicate records, making the dataset in the classified papers also have more repetitions. So, a manual check was carried out to remove all duplicate papers and datasets and correct several classification problems caused by duplication, ultimately resulting in 922 papers and 860 datasets in dataset-related papers. Note that multiple datasets can be used or created for a single dataset-related paper.

3.2.1 Parameters of Dataset Classifier

Benefiting from the classification made in the original paper, I randomly selected 400 classified papers for the construction of the classification model, while ensuring sufficient coverage of all conferences and years from 2012 to 2016. Among them, 220 papers included data and the remaining ones did not.

To construct features, following the process from the original paper, I first extracted a list of words in 'basic form' (i.e. case- and tense-insensitive) for each paper with the textblob (an open-source text processing library of Python) and also used the built-in list of NLTK (a Python library for natural language processing) to filter all stop words, which are mainly function words with no real meaning, including inflectional auxiliaries, adverbs, prepositions, conjunctions, etc. From the final word list of each paper, we removed words with low frequency and all numbers, then built a word vocabulary from all papers and computed a TF-IDF vector for each paper. TF-IDF (Term Frequency-inverse Document Frequency) is a statistical analysis method for keywords, which is used to assess the importance of a word to a document set. And the value of TF-IDF is calculated to classify the dataset's features and filter out useful and essential information. The larger the TF-IDF value is, the more relevant the word is to the text.

Of TF-IDF, TF (Term Frequency) is the frequency with which words occur in a document. For normalisation purposes, it is common practice to take the ratio of the word's frequency in the document set to all words. IDF (Inverse Document Frequency) indicates the importance of a given word. The main idea is that if a feature item appears very frequently in one text and at the same time appears less frequently in other texts, this feature item has good category differentiation and should be given a higher weight.

3.2.2 Selection of Dataset Classifier

Subsequently, I constructed several supervised learning models with the sklearn library of Python used in the original paper, including Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes, C-Support Vector Classification, and Random Forest. For construction, the input is each paper's TF-IDF vector, and the output is whether the paper contains datasets, i.e. 0 and 1. However, after following the previous steps to clean up, the word count per paper was still too high. Since too many parameters would lead to underfitting, I selected the top 400 words of the TF-IDF vector as the input to the model.

For the evaluation, I use confusion matrix and 10-fold cross-validation. The confusion matrix is an essential tool for evaluating the performance of classifiers, as shown in Table 3.1, which shows all possible cases during the confusion matrix dichotomy problem. Each column of the confusion matrix represents the predicted category with the total of each column indicating the number of data predicted to be in that category, while the row represents the real attribution category. FP (False Positive, which is judged to be a negative sample but is in fact a positive sample) and FN (False Negative, which is judged to be a positive sample but is a negative sample) represent Type I and Type II errors. In terms of 10-fold cross-validation, it is done by dividing the dataset into ten equal parts, with nine of them being used as training set and one as validation set in turn. Finally, I took the average performance of the ten times as a result to evaluate.

False Positive Rate = False Positive / (False Positive + True Negative)

False Negative Rate = False Negative / (False Negative + True Positive)

Predicted Actual	TRUE	FALSE
TRUE	TP (True Positive)	FN (False Negative)
FALSE	FP (False Positive)	TN (True Negative)

Table 3.1: Confusion Matrix

The Multinomial Naive Bayes appears to be overfitting, predicting 1 for all inputs, so the table 3.2 does not show the performance. Among the 922 papers, Random Forest had the highest accuracy, 720 ones were correctly classified. It is consistent with the choice of the original paper

Machine Learning Model	Accuracy	False Positive Rate	False Negative Rate
Random Forest	78.12%	24.15%	20.00%
Bernoulli Naive Bayes	67.84%	31.95%	32.34%
Gaussian Naive Bayes	68.19%	28.57%	34.47%
C-Support Vector Classification	75.91%	28.05%	20.85%

Table 3.2: Performance of models

3.3 Characteristics of Datasets

3.3.1 Taxonomy

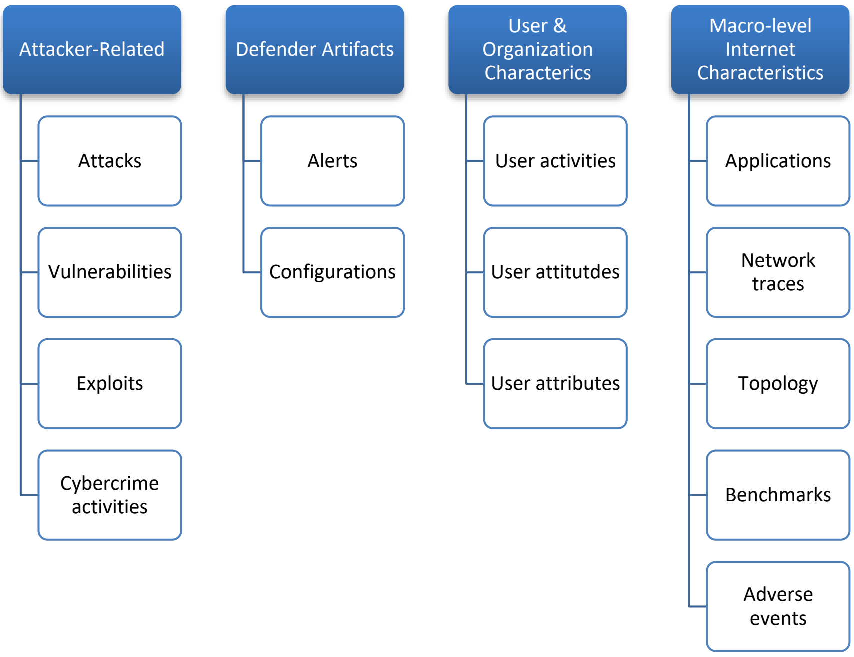


Figure 3.3: Datasets categories

As described in the original paper, cybersecurity research datasets are grouped into four categories: Attack-Related dataset, Defender Artifacts, User and Organisation Characteristics, as well as Macro-level Internet Characteristics, each of which is further divided into different sub-categories.

(1) Attacker-Related

Datasets that are determined to be malicious, such as scams or malware, or used by attackers, are labelled as attacker-related datasets, which contains four sub-categories. Attacks are primarily information about deliberate attempts to damage a digital asset by hostile perpetrators. Vulnerabilities are information about an exploitable weakness in a digital asset that attackers can take advantage of. Exploits contain instructions on how to carry out attacks. Finally, Cybercrime activities refer to illicit activities that are apart from Attacks, with information mainly about the operations and infrastructure employed by hostile perpetrators to carry out the attack.

(2) Defender Artifacts

To prevent or avoid attacks, individuals and organisations construct defences such as security configurations or firewalls. Such defender artifacts consist of configurations and alerts. Of which, configurations contain information on the setup and configuration of defence artifacts, while alerts are defender artifacts' output.

(3) User & Organisation Characteristics

Numerous datasets are used to examine the behaviours of individuals or organisations. User activities reveal information about the online actions operated by various kinds of users. User attributes include information relating to users' or organisations' own features. User attitudes are comprised of information on individuals' beliefs or attitudes regarding a topic, which is frequently collected through surveys.

(4) Macro-level Internet Characteristics

Macro-level Internet Characteristics includes datasets that are devoted to the study of network features. Network traces are commonly dumps of network traffic containing both application-level and lower-level information. Topology datasets are often related to the information about connections between Internet components. Applications include information about Internet end products and services. Benchmarks carry information regarding Internet performance measurements. Finally, Adverse events comprise information about events that endangered digital assets, where there is no established harmful intention.

3.3.2 Additional Dataset Characteristics

If a dataset was in existence prior to the research to be conducted, the dataset is labelled as an existing dataset. Otherwise, researchers create the dataset. This comes in two forms of created datasets. Suppose it is produced from other datasets, such as crawling an application list from the google play store and further analysing or classifying them. In that case, the paper creates a derived dataset. If the dataset was completely generated by the researchers and no dataset was used as input, for example, by a questionnaire in the form of asking users about their views on phishing, I consider that the dataset was created primarily by the researchers.

From figure 3.4, we can see that papers without data make up the majority of papers. By observing a large number of papers during the classification above, it can be seen that most of such papers are proposing or improving models as well as performing simple analysis of cybersecurity phenomena, so they do not require the use of datasets. Moveover, compared to papers with existing datasets, there are more created datasets in papers, including primary and derivative datasets. However, the categorisation of 922 papers is not able to fully reflect the situation in the field of cybersecurity and maybe a little biased. This happens because the analysis of cyber security is limited and not all cybersecurity papers are analysed, but we can still get a general trend and comparison.

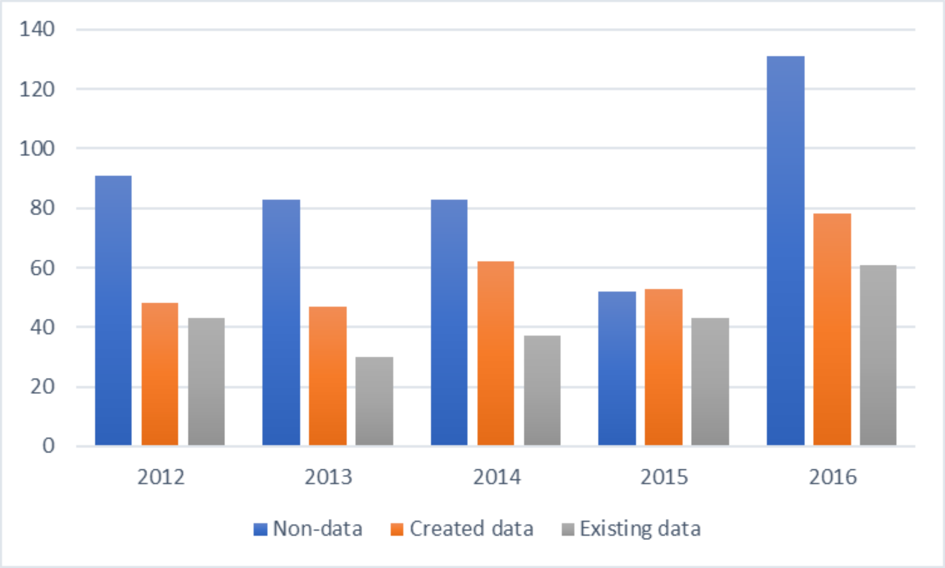


Figure 3.4: Numbers of classified papers (2012-2016)

3.4 Empirical Analysis

This section performs statistical and regression analysis on the dataset identified above. Firstly, section 3.4.1 describes the frequency of different categories of datasets being created, used and made public, as well as the features of citations. Then, in section 3.4.2, regression models are developed to visually analyse the impact across dataset characteristics. It is found that papers that make the created datasets publicly available are rewarded with moderately high citation rates compared to other papers.

3.4.1 Analysis of Datasets

Using the taxonomy described above, we classified the cybersecurity datasets used in the papers into four main categories and 14 sub-categories. The leftmost numerical column in table 3.3 displays the percentage distribution of the datasets in different sub-categories. Macro-level Internet characteristics account for 49.2% of the total dataset, nearly half of all datasets, and 22.4% for User & Organisation Characteristics. Datasets related to attack and defence make up the remainder, at 20.6% and 7.7%, respectively. From a specific sub-category perspective, datasets about Applications occupy the most, nearly a quarter, followed by Attack and User activities, which together account for 48.6%.

The next column in Table 3.3 examines how the datasets are used in each sub-category. It indicates the proportion of datasets in each sub-category that are created via research, rather than reusing existing data. 72.7% of the datasets describing vulnerabilities were created, compared to 30.6% of the attack datasets. This might imply that attack datasets are particularly important inputs for research or that vulnerabilities are more likely to be discovered. Similarly, it is more probable that Network traces, Benchmarks, and Adverse events will be produced than utilised. Among User & Organisation characteristics, User attitudes were created in 90% of cases. As for this type of data, most studies have most likely chosen to collect the dataset through questionnaires or in-person questions. However, the probability of using existing data for user attributes is high, probably due to the greater availability and comprehensiveness of data in this area on the web.

The last column in the table shows the fraction of datasets that are public or not. For four main categories, Defender Artifacts datasets have the highest public rate at 62.5%. In contrast, User & Organisation Characteristics datasets are the most petite public, which may be due to the private nature of user and organisation activities. In terms of sub-categories, datasets of vulnerability, alert, application, user attribute, and Topology are more likely to be public, while Network traces, Benchmarks, and User attributes are less likely. Of these, user attributes are only 10% public, which may be attributed that User attitudes contain a large amount of individual information.

		% Datasets	% Created	% Public
Attacker-Related	Attacks	12.6	30.6	50.9
	Vulnerabilities	4.7	72.5	37.5
	Exploits	2.3	35.0	70.0
	Cybercrime activities	1.0	55.6	44.4
Defender Artifacts	Alerts	2.6	31.8	77.3
	Configurations	5.1	54.5	47.7
User & Organization Characteristics	User activities	11.5	41.4	39.4
	User attitudes	1.2	90.0	10.0
	User attributes	9.7	28.9	62.7
Macro-level Internet Characteristics	Applications	24.5	35.5	62.1
	Network traces	9.5	61.0	22.0
	Topology	9.1	23.1	69.2
	Benchmarks	3.7	81.3	28.1
	Adverse events	2.4	66.7	33.3

Table 3.3: Classification features of datasets (2012-2016)

In general, while cybersecurity researchers often produce and utilise datasets, the researchers who collect them rarely share them. There are many plausible (and less plausible) explanations for this, including worries about privacy, competitive issues, etc. Additionally, preparing and requesting data for sharing can be expensive and time-consuming.

Despite these drawbacks, the publication of datasets can certainly provide benefits to researchers. One possible benefit that is valued highly by researchers is the number of citations received by papers. It can be hypothesised that papers including publicly available datasets will receive more citations, as other academics may use the datasets in later studies.

As a whole, papers without data or using existing datasets receive 22 (median) citations, and papers that generate datasets but do not make public are cited 24 times. In comparison, a median of 37 citations is given to papers that make datasets public. Figure 3.5 shows the breakdown of the median number of citations on the basis of datasets sub-categories and usage. The median citations vary considerably by sub-categories, papers with User Attitudes datasets that were created are cited approximately 4 times, while papers relating to Alerts datasets are cited 45 times. Notably, most sub-categories demonstrated that papers with created datasets tended to be cited more than papers using existing data. This trend was more apparent for the papers relating to Cybercrime activities, Alerts, User attributes, Benchmarks, and Adverse events, indicating that these created data are likely to have a higher value for subsequent papers. Furthermore, for Attacker-related features, the citations for the papers with existing or created datasets were broadly similar. For User & organisation characterises, papers including existing datasets of User attitudes are cited more frequently, which may be related to their low public rate (10%), leading subsequent scholars to prefer using the complete existing data.

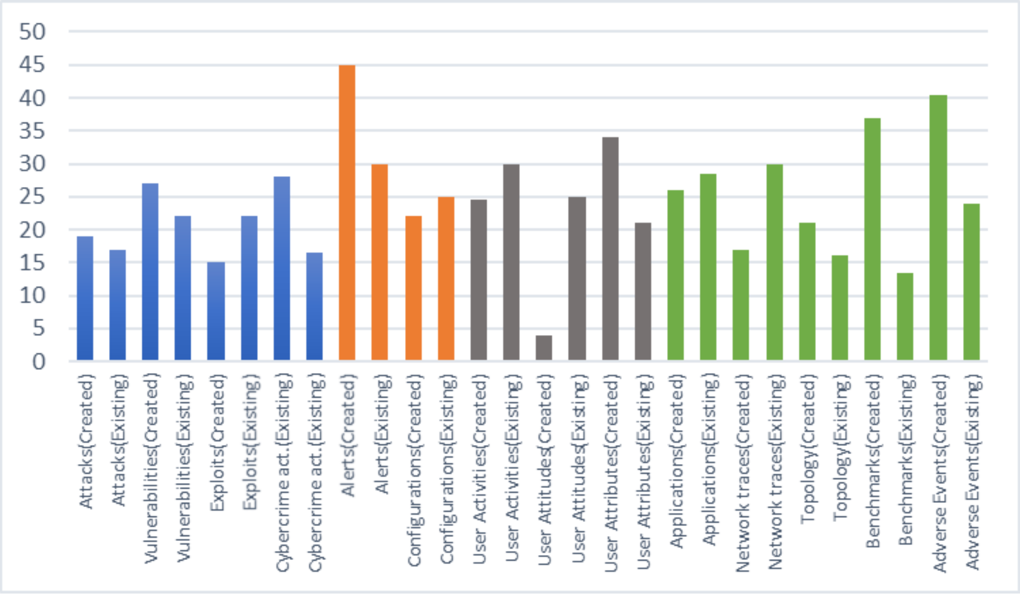


Figure 3.5: Median citations per sub-category (2012-2016)

3.4.2 Regression Analysis

To better understand the impact on each of the above factors, several Multiple Linear regressions are constructed using Least Squares with the number of citations as the dependent variable. And the explanatory variables for the regressions comprise:

(1) Number of years since publication. The number of citations in scientific papers is time-dependent. Generally, the longer a paper has been published, the more quotations it will receive. Price's research [19] states that a paper will normally attract the attention of researchers and begin to increase the number of citations two years after it is published.

(2) Place of publication. The reputation and popularity of the publication outlet undoubtedly affect the frequency of citations. Didegah et al. [20] concluded that academic papers published in high-impact, high-ranking journals are more likely to receive attention and are more likely to be highly cited. This factor is expressed as a categorical variable in the regression. Where CCS is regarded as the reference, CSET, AISEC and BITCOIN are combined into one workshop, thus we have a total of seven dummy variables.

(3) Creation and disclosure of the dataset. It can be assumed that a paper will have more citations when it creates and makes the dataset publicly available. Accordingly, I constructed a categorical variable on the use of datasets, which contains: non-data (baseline), created not public, created public and existing data.

(4) Category of datasets. For papers that include datasets, it is expected that the type of created data will affect their citation frequency. This is also defined as a categorical variable in the regression, with a baseline of attack sub-categories.

The tool we used to perform linear regression is Eviews, abbreviated as Econometrics Views, which is a professional econometric software with data processing, graphing and statistical analysis functions. With Eviews, it's efficient to obtain the regression results, including regression coefficients, P-values, R^2, etc. For the data input, I first manually converted variables to several dummy variables. In terms of year, I have used the difference between 2017 and the year of publication as the independent variable, as all data are dated to 2016 and earlier.

Table 3.4 illustrates the results of four linear regressions that progressively incorporate the above-mentioned explanatory factors. The data used in the model here differs somewhat from the original article, as the data has been adjusted as described in section 3.2. By comparing the results of the regressions with the original paper, it can be noted that the differences are not significant, but only minor variations in individual variables.

Model (1) discovers that, as expected, years since publication has an effect on citations. Each year after publication results in an additional 17 citations amount. Around 10% of the difference in citations is accounted just by the year after publication.

The addition of place of publication (model 2) accounts for an additional 8.8% of the variation in the number of citations. It can be observed that papers published in FC, WEIS, and workshops (CSET, AISEC and BITCOIN) receive fewer citations than papers in CCS. In contrast, papers published in SP and USENIX are significantly more likely to be cited. For the others, IMC and NDSS are statistically insignificant, indicating that their citations did not differ from CCS.

Model 3 adds a categorical variable on the use of the dataset in the paper. Papers that created datasets and made public are more likely to be cited relative to papers that did not include datasets. In addition, papers that created datasets but not made them publicly available and papers with existing datasets are cited at no different rates than papers without datasets.

Model 4 incorporates sub-categories of datasets. It can be noted that the higher number of observations in Model 4 is due to the analysis unit being the all datasets to compare citations by sub-category. Papers from datasets such as Alerts, Network traces and Topology are cited less frequently compared to Attacks datasets. As can be also seen, Created Not Public is not involved in the regression, because papers without datasets would not be present in the dataset analysis. As a result, Created Not Public is used as a benchmark and the variables for Created public and Existing become dummy variables. Therefore, Created Public statistically significantly implies that papers that create datasets and make public tend to attract more citations than those that do not, while using existing data makes no difference.

	Dependent variable: **CiteNum**			
	(1)	(2)	(3)	(4)
Years Published	16.7811***	16.6385***	16.9247***	17.8736***
FC		-19.9323**	-19.6567**	-16.8016
IMC		-11.0596	-14.8203	-7.5739
NDSS		-0.1162	-1.2638	7.9496
SP		45.9036***	44.9381***	26.1598***
USENIX		15.1954**	13.1560*	0.9834
WEIS		-28.7385**	-29.2826***	-30.0766**
Workshops		-39.0002***	-39.8492***	-32.9881***
Created Not Public			-1.3860	
Created Public			26.3706***	24.6836***
Existing			1.8697	-2.2856
Vulnerabilities				-7.5447
Exploits				-22.5575
Cybercrime activities				-3.3549
Alerts				-27.1974*
Configurations				-11.0836
Applications				0.2218
Network traces				-22.0675**
Topology				-27.5021***
Benchmarks				-26.5794
Adverse Events				-13.8974
User Activities				1.2027
User Attitudes				-8.2003
User Attributes				-1.1190
Constant	-0.3931	0.0129	-1.5552	4.6162
Observations	921	921	921	860
R^2	0.1010	0.1885	0.1953	0.1849
Adjusted R^2	0.1001	0.1815	0.1858	0.1630

Table 3.4: Result of Linear regression (Note: *p<0.1; **p<0.05; ***p<0.01)

3.5 Summary of Replication

Overall, I performed a better replication of Zheng's paper, analysing the cybersecurity dataset from 2012-2016 in a similar methodology. During the replication, it was found that there were issues such as duplication of data in the original paper, but the subsequent analysis of the adjusted data revealed that the analysis was slightly different from the original paper, which showed that the main conclusions of Zheng's article were not impaired.

In terms of statistic analysis，by examining nearly 900+ papers, I applied the taxonomy method presented in the original paper to the dataset, revealing how datasets are being generated, used and shared from the perspective of sub-categories. It is found that Macro-level Internet characteristics account for most of the total dataset at 49.2%, nearly half of all datasets, and 22.4% for User & Organisation Characteristics. Datasets related to Vulnerabilities, Network traces, Benchmarks, and Adverse events are more probable to be produced than utilised. However, when cybersecurity researchers often produce and utilise datasets, the researchers rarely share them.

Regarding regression analysis, it shows papers that created datasets and made public tended to attract more citations than unpublic ones while using existing data makes no difference, which supported the main assumptions and conclusions of the original text. It can also be observed that papers published in FC, WEIS, and workshops (CSET, AISEC and BITCOIN) receive fewer citations than papers in CCS. In contrast, papers published in SP and USENIX are significantly more likely to be cited.

Chapter 4

Improvement and Extension

This part extends the original paper by re-analysing the latest databases and studies with an optimisation approach to identify any improvements in data sharing and the characteristics of recent cybersecurity datasets. There are three directions of improvement: enlarge the datasets by adding recent papers, develop a binary classifier that can better determine the existence of datasets in a paper, and modify the regression model for a better fit.

4.1 Enlarge the Datasets

Our data sources are the same as in Zheng's paper, collecting papers from the four major computer security research conferences, three conferences that regularly publish data-intensive research and three related workshops. Focusing on these conferences, it will make the research work relatable and better characterise cybersecurity data. Compared to the replication part, we collected all papers from 2017 to 2020.

I still used Python to capture the URLs of papers from DBLP, crawled papers from the corresponding URLs, and then automatically searched for the number of citations of each paper from Google Scholar. Once they were downloaded locally, I ran two tests on them. The first was to check the size of the file, as a tiny file size implies that there had been an issue with the download process and that the whole paper failed to be downloaded locally. The other is to ensure that the file is the thesis itself and not some other material, such as a PowerPoint. When either of these occurred, I manually verified the download sites for these papers and re-downloaded them.

Ultimately, I crawled a total of 2345 papers. Similar to the case from 2012-2016, the majority of papers are published in CCS, with USENIX, NDSS and SP following closely behind, and the least number of papers are published in workshops. Specifically, CCS publishes 32.32% of the total papers, down from 37.22% in the years before. At the same time, USENIX has grown rapidly in recent years, publishing an average of around 110 papers per year, twice as many as before. NDSS and SP have also seen a relatively large increase in the number of papers. Notably, AISEC has no workshop papers in 2020 and BITCOIN has papers till 2018.

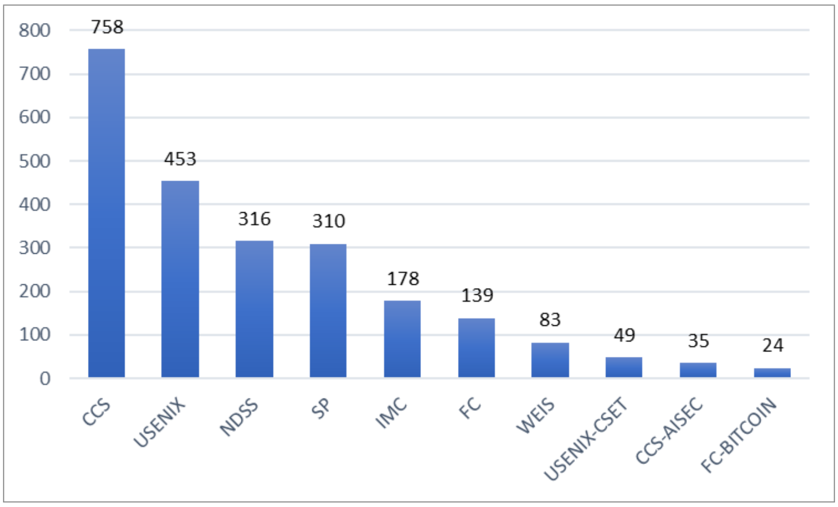


Figure 4.1: Distributions of all papers based on the conference (2017-2020)

In terms of year, there is a general upward trend from 556 papers in 2017 to 610 papers in 2020, the increase of which is not significant. Nevertheless, there has been a significant increase on the publication in recent years compared to the period of 2012-2016. It can also be observed from figure 4.2 that most conference publications are trending upwards, but CCS papers dropped sharply in 2020.

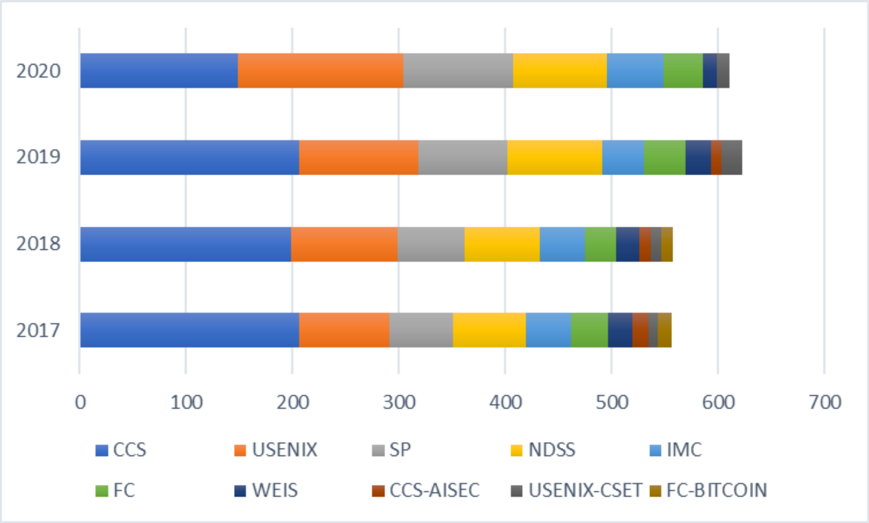


Figure 4.2: Distributions of all papers based on the year (2017-2020)

4.2 Improve the Binary Classifier

4.2.1 Construct an Improved Binary Classifier

To construct a more accurate binary classifier, I supplemented four supervised learning algorithms to find a better classification algorithm: K-Nearest Neighbour, Adaboost, Gradient Boosting, and Hard Voting. ***Nearest Neighbour*** is one of the simplest machine learning algorithms, which determines the class of a sample based on the classification of the nearest sample or samples. ***Adaboost*** is an iterative algorithm based on the principle of training several weak classifiers on the same training set and then pooling these weak classifiers together to produce a more robust classifier. [21] ***Gradient Boosting*** is analogous to Adaboost, but it uses negative gradients to measure the error made by the base learner in the previous round. ***Hard Voting*** is a strategy of Ensemble learning, and therefore is not a stand-alone machine learning algorithm. It is done by building and combining multiple algorithms to form its own model. For example, when several algorithms are used for model building to examine the existence of datasets, the paper will be evaluated as including data if more than half of the algorithms are judged to include data, and vice versa. For my Hard Voting model, I selected three algorithms to be combined in a permutation process to achieve high accuracy, which are Random Forest, Gradient Boosting and Gaussian Naive Bayes.

I evaluated the algorithm using the classified paper that was already identified in the replication stage. It was clear by the accuracy that Hard Voting Classifier was the most accurate, but it still had a false-positive rate of 20.26% and a false-negative rate of 16.81%. Therefore, I used the Hard Voting algorithm to classify crawled papers, of which 1208 predictions included data. A sample of 221 papers was randomly selected for inspection, ensuring that all conferences and years from 2017 to 2020 were adequately covered, with 116 predicted to include datasets versus 105 that did not.

Through manual examination, I ultimately determined that out of the 116 papers predicted to contain datasets, only 86 actually did. And of the 105 papers with no data predicted by the model, 15 were incorrectly classified. For the papers inspected, the correct classification rate was 79.64% and the false-negative rate was low at 14.85%, which was a fairly positive result overall, although the false-positive rate was high at 25%. A total of 101 dataset-related papers are finally obtained, which contain 143 datasets, as one paper could use or create multiple datasets

Machine Learning Model	Accuracy	False Positive Rate	False Negative Rate
Random Forest	78.12%	24.15%	20.00%
Bernoulli Naive Bayes	67.84%	31.95%	32.34%
Gaussian Naive Bayes	68.19%	28.57%	34.47%
C-Support Vector Classification	75.91%	28.05%	20.85%
K-Nearest Neighbour	66.08%	39.74%	29.15%
Gradient Boosting	79.18%	25.97%	16.60%
Adaboost	72.87%	28.31%	26.17%
Hard Voting	81.64%	20.26%	16.81%

Table 4.1: Performance of models

4.2.2 Characteristics of Classified Datasets

For 221 papers tagged, I further studied the datasets included in them. For the classification of the datasets, I manually grouped them into four main categories and fourteen sub-categories to ensure the accuracy of the categorisation. Then I tracked how they were used by grouping them into existing, derived, and primary datasets. The primary datasets are created by the researchers alone, while the derived datasets are created with at least one existing dataset. There is also a particular focus on whether the dataset is publicly available. I looked for clear statements in the paper that the dataset is freely accessible. Public repositories with some limitations on downloading the datasets (e.g. IMPACT) are included since it is still public. When a dataset is made public, we directly access the link to check whether it is available and valid.

In order to have a more comprehensive analysis as well as to understand the variations in the datasets being created and used over the years, I combined the data just classified with the data from the previous section, a total of 1142 papers. The analyses in the later parts are all based on aggregated data in 2012-2020.

Table 4.2 provides a breakdown of the datasets based on their usage and whether or not the datasets are publicly available. Of the 574 data-related papers, a total of 1003 datasets are created or used. For the datasets created, a division is made between primary datasets and derived datasets. It can be seen that 69% of the existing datasets used by the researchers are publicly available. Unfortunately, researchers who have created datasets are unlikely to be rewarded by making their own datasets public. The proportion of public availability of both types of created data is relatively close, with close to 80% of datasets not being made public. And the number of datasets created and made public is 85 (27+58), representing only 8.5% of all datasets. This apparent discrepancy highlights the fact that researchers have a negative attitude towards making their data publicly available.

	Not Public		Public	
Dataset Type	#	%	#	%
Created Deriv.	99	78.57	27	21.43
Created Prim.	231	79.93	58	20.07
Existing	182	30.95	406	69.05

Table 4.2: Number and percentage of datasets about creation and publication (2012-2020)

Using the taxonomy described previously, I grouped the cybersecurity datasets into four main categories and fourteen sub-categories. The leftmost numeric column in Table 4.3 presents the proportion of datasets in the different sub-categories, the middle column examines the creation of datasets in each sub-category, and the last column shows the percentage of publicly available datasets.

The proportion of data related to Macro-level Internet Characteristics is still the highest at 49%, with Applications-related datasets accounting for over one-quarter. In terms of public rate, the willingness of researchers to share has become lower in recent years, with most sub-categories showing a slight decline in disclosure rates. It is worth pointing out that, in the original statistics, datasets about User Attitudes are created in 90% of cases, while it was made public for only 10%, but this situation has improved in recent years. A rise can be observed in the public rate of User Attitude datasets to 15%, which further leads to an increased likelihood of others using their datasets, thus creating a virtuous circle. Overall, the characteristics of datasets do not change significantly from the previous statistics.

		% Datasets	% Created	% Public
Attacker-Related	Attacks	12.2	30.6	47.1
	Vulnerabilities	5.2	64.7	35.3
	Exploits	2.2	40.9	68.2
	Cybercrime activities	1.6	50.0	43.8
Defender Artifacts	Alerts	3.0	30.0	66.7
	Configurations	4.9	50.0	47.9
User & Organization Characteristics	User activities	10.6	40.0	37.1
	User attitudes	1.9	78.9	15.8
	User attributes	9.4	30.1	57.0
Macro-level Internet Characteristics	Applications	25.3	34.8	57.6
	Network traces	9.6	62.1	22.1
	Topology	8.1	25.0	68.8
	Benchmarks	3.6	80.6	25.0
	Adverse events	2.3	60.9	30.4

Table 4.3: Classification features of datasets (2012-2020)

Figure 4.3 shows how the types of datasets appearing in the papers change over time. The proportion of Attacker Related datasets shows an upward trend, rising from 13% to 44% in 2019, before slipping to 18% in 2020. On the one hand, this may be an error due to the random selection of data. On the other hand, it may be attributed to the fact that papers on attackers are not up to date. For example, AISEC was not released in 2020, datasets of which are more preferred to be attacker-related based on past experience. The Defender Artifacts datasets have also increased in recent years. It is worth noting that datasets related to Macro-level Internet Characteristics increase significantly in 2020. The main growth factor is the Application sub-category, which indicates researchers' preference for datasets with information on Internet end products and services, such as websites, extensions, applications, code, etc.

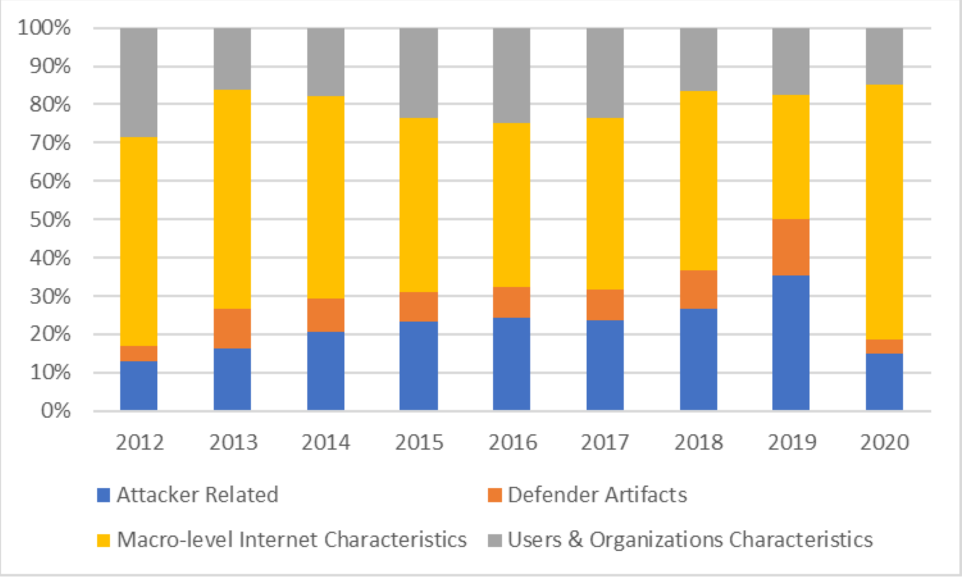


Figure 4.3: Dataset categories (2012-2020)

The public availability of datasets certainly benefits researchers, so it is assumed that articles with publicly available datasets will receive more citations, thereby increasing the impact of authors and papers, which is one possible benefit valued by researchers. Since 2012, papers with no data have the lowest number of citations (median) at 44, while papers with existing datasets and created datasets have been cited 53 and 58, respectively. Specifically, papers that created and made public datasets are cited at a median of 68, more frequently than the non-public ones. The findings support, to some extent, the idea that papers using datasets or publicly available datasets are more likely to be cited and used by other scholars in future research.

Compared to 2012-2016, the passage of time leads to an overall 145% increase in the median frequency of citations, from 24.3 calculated in 2018 to 59.6 calculated this year. Beyond the overall increase in citations over the years, we now focus on citation differences between the various sub-categories. The overall picture of citation frequency across sub-categories is not significantly changed, but a pleasing trend can be identified compared to the replication part. Namely, papers containing created datasets have a faster increase in citations than papers with existing datasets in recent years, and this trend is particularly evident in Defender Artifacts datasets. For example, (1) the citation gap between two types of Vulnerabilities datasets continues to widen, with papers containing created datasets now far outstripping those with existing datasets; (2) created datasets of Configurations received more references than existing datasets for the first time.

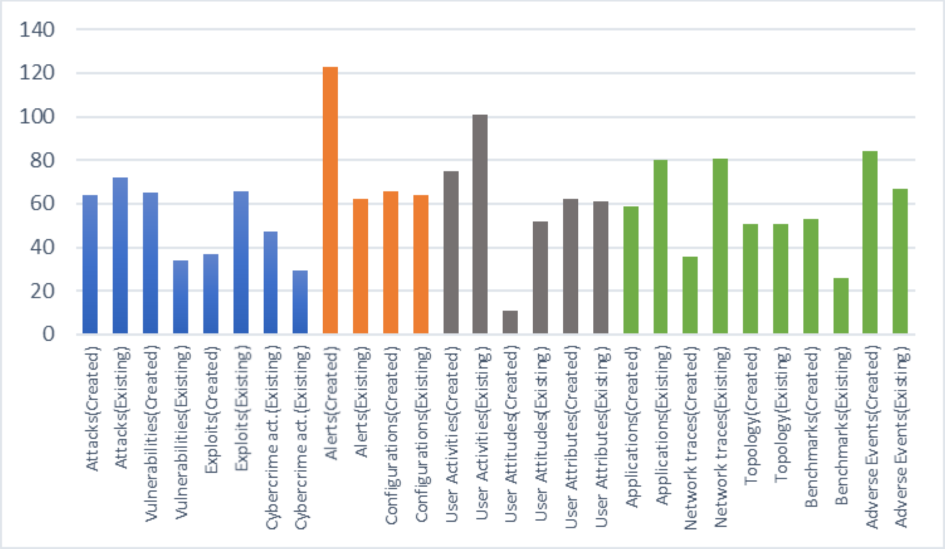


Figure 4.4: Median citations per sub-category (2012-2020)

4.3 Modify the Empirical Analysis

4.3.1 Selection of Regression Model

The original paper used multiple linear regression based on least squares to perform the regression analysis. However, when the dependent variable is not normally distributed, it is not possible to use the usual multiple linear regression model so that a transformation with a linking function is required. Because a fundamental assumption for linear regression models is that the error term (ε) conforms to a normal distribution N (0 , δ^2), so that the dependent variable serves to a normal distribution N (aX+b , δ^2), where the prediction function y = aX + b. It can be obtained from the probability density function of the normal distribution. In short, when the error term is normally distributed, its dependent variable is also necessarily distributed normally. Therefore, before fitting the data with a linear regression model, data is required to conform or approximately conform to a normal distribution, otherwise the result will be incorrect or a poor fit.

Figure 10 shows the distribution of the citation frequency for 1142 papers since 2012, which clearly shows a decreasing trend of citation frequency and does not conform to a normal distribution. The number of papers that receive citations between 0 and 10 is 161, accounting for 14%. In addition, papers with less than 50 citations account for 50%.

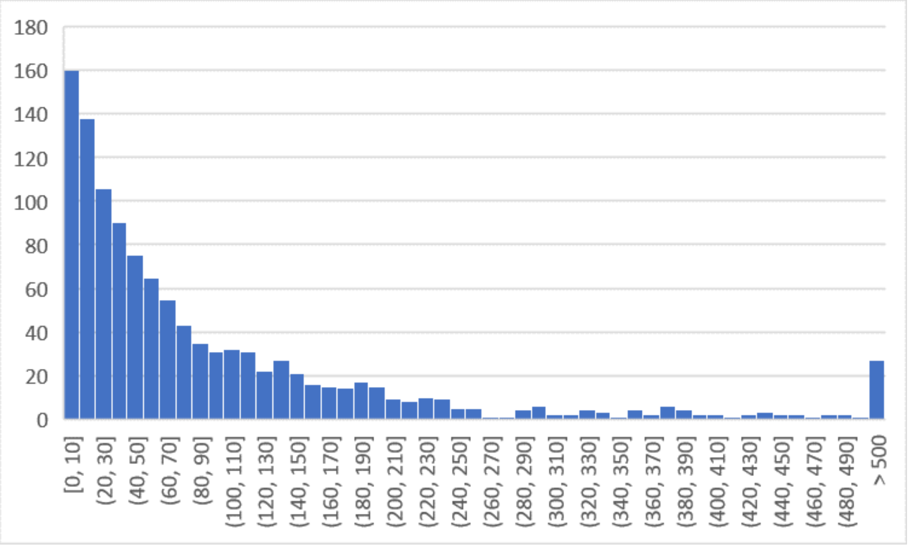


Figure 4.5: Distribution of citations (2012-2020)

With regard to count data (citations), Poisson regression and Negative Binomial regression are commonly used, but Poisson regression requires the data to satisfy equal dispersion (mean and variance are equal). Once the variance deviates from the mean, there is a large bias in fitting when using Poisson regression models. By calculating the citations, the variance of the citations was much greater than the mean, which would lead to a small standard error in the estimation of the model parameter if the Poisson regression is still adhered to. Therefore, it is more scientific to use Negative Binomial regression, which is a promotion of the Poisson regression and can be a good solution to the problem of excessive deviations.

Then I proceed to compare the performance of the Negative Binomial with the previous multiple linear regression using AIC (Akaike information criterion) and BIC (Bayesian Information Criterion), both of which are measures of fitness, taking into account both complexity and accuracy. Complexity corresponds to the “number of parameters” and “amount of training data”, where a larger value indicates an increase in the complexity of the model and a tendency to overfit. Accuracy is the capability of the model to describe the data. The smaller the AIC and BIC values, the better the model. Taking model 4 in Table 4.4 as an example, the AIC and BIC using Negative Binomial regression are 6254 and 6363, respectively, compared to a larger AIC and BIC using multiple linear regression at 7350 and 7454. This also indicates that using Negative Binomial regression leads to better regression results than the model used in Zheng's paper.

In terms of regression tools, we used statsmodels library in Python for regression analysis, as the Eviews used previously is only capable of simple linear regression. It is essential to mention the Alpha variable appearing in the model, which is the default dispersion coefficient output by the Negative Binomial regression and is used to test for over-dispersion. If the Alpha coefficient is significantly non-zero (corresponding to a p-value less than 0.05), then it is reasonable to use Negative Binomial regression, and vice versa, it is probably better to use Poisson regression. It is clear from Tables 4.4 and 4.5 that each Alpha is significantly non-zero, which again validates the correctness of using Negative Binomial regression. In addition, as the negative binomial regression model cannot calculate R^2, we use Log-Likelihood to judge the model's fitness, where a larger value means a better fit.

4.3.2 Statistical Analysis of Regression

For the Negative Binomial regression, I used the four explanatory variables used in the previous multiple linear regression, including the number of years since publication, the location of publication, the usage of datasets and the category of the datasets, to discover how well these four variables fitted and how they differed from previous findings with the recent data and model. For years published, I used the difference between 2021 and the publication year of the paper as the explanatory variable.

Table 4.4 presents the results of four regressions that progressively include four explanatory variables. The regressions only cover papers with created datasets to test the hypothesis that the publication of datasets would receive higher citations.

Model (1) illustrates that the coefficient of years published is 0.27 and reaches 1% significance level, implying that the publication year of papers will positively affect the frequency of citations. When analysing the effect of the explanatory variables on the dependent variable, the Negative Binomial regression needs the OR value (odds ratio) to evaluate. It is calculated by dividing the odds of the event occurring in the experimental group by the odds occurring in the control group, which can also be calculated by exp(coefficient). The OR for years published is 1.31 ($e^{0.2735}$), representing a 31% increase in citations for each additional year of publication, with all else being equal.

Model (2) adds the categorical variable of place for publication. Papers published at FC, IMC, WEIS, and workshops have a lower likelihood of being cited than papers from CCS. In the case of FC, papers published at FC conferences were 0.53 ($e^{-0.6424}$) times the probability of being cited than CCS. Alternatively, papers published at CCS were 1.9 ($e^{0.6424}$) times the probability of being cited than FC. The probabilities of being cited for papers published at IMC, WEIS and workshops are respectively 0.67 times, 0.19 times and 0.25 times that of CCS. The smaller the regression coefficient, the less likelihood of being cited compared to papers from CCS. The citations of the others (NDSS, SP and USENIX) do not differ from CCS. The result is somewhat different from the previous one, including USENIX and SP changing from being significant to insignificant, and IMC being the opposite. However, if just focusing on their regression coefficients, the trends remain consistent with previous results.

Model (3) incorporates a Boolean variable to determine whether datasets created are publicly available, defined as 1 and 0 for papers with and without public datasets, respectively. The coefficient is positive and statistically significant, which can be explained by the 25%

($e^{0.2227}$-1) boost in citation rates for papers with publicly available datasets compared to non-public datasets.

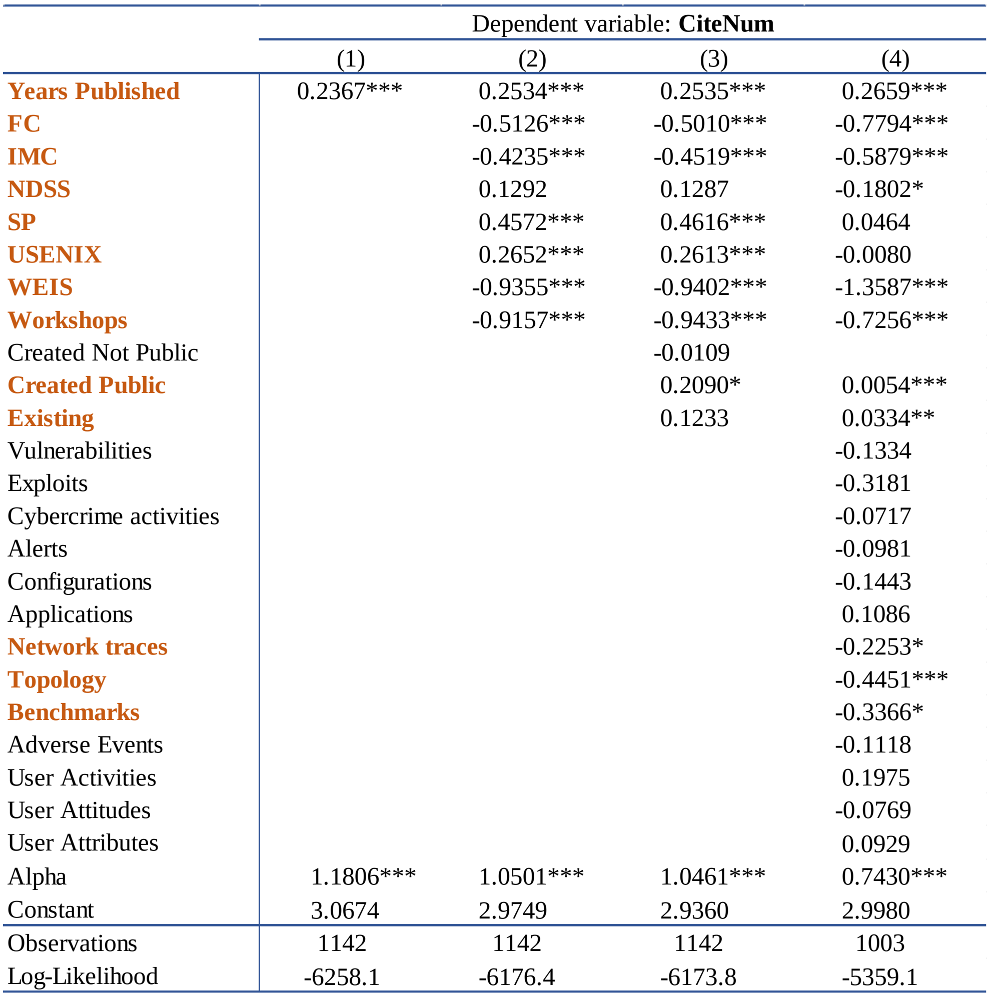
Model (4) adds sub-categories of datasets. There are more observations in model (4) since the analysis unit is all the datasets included in the papers. It can be observed that only Topology and Benchmarks are significant, so that papers containing them are cited less frequently relative to the Attack-related dataset.

	Dependent variable: **CiteNum**			
	(1)	(2)	(3)	(4)
Years Published	0.2735***	0.2887***	0.2924***	0.2741***
FC		-0.6424***	-0.6679***	-0.8675***
IMC		-0.4079**	-0.4431***	-0.5871***
NDSS		-0.2526	-0.2664	-0.3949***
SP		0.3022	0.2572	0.0065
USENIX		0.2172	0.1829	-0.0198
WEIS		-1.6442***	-1.6432***	-1.7281***
Workshops		-1.3686***	-1.3875***	-1.1339***
Created Public			0.2227*	0.2775***
Vulnerabilities				-0.1697
Exploits				-0.0685
Cybercrime activities				0.1582
Alerts				-0.0023
Configurations				-0.1039
Applications				-0.0066
Network traces				-0.2056
Topology				-0.3742*
Benchmarks				-0.4442**
Adverse Events				-0.2167
User Activities				0.2285
User Attitudes				-0.3284
User Attributes				0.2572
Alpha	1.0296***	0.8551***	0.8478***	0.7098***
Constant	2.8318	2.8669	2.8115	3.0408
Observations	331	331	331	574
Log-Likelihood	-3171.3	-1794.7	-1793.0	-3103.0

Table 4.4: Negative Binomial regression tables for papers that create datasets (2012-2020) (Note: *p<0.1; **p<0.05; ***p<0.01)

Table 4.5 lists four more regressions taking into account all papers, not just those with created datasets. The usage of datasets, one explanatory variable, shifted from Boolean variables to categorical variables because the non-datasets and existing-datasets papers were incorporated, where non-datasets served as a baseline for categorical variables. The results for most variables are consistent with the regressions presented in Table 4.4. SP and USENIX gained significance, indicating that papers published in them are more likely to be cited and more popular than CCS. Then, papers containing the Network trace dataset are less likely to be cited than attack-related ones.

Concerning the use and publication of datasets, papers that create datasets and make them public are much more citable than papers that do not have datasets. In addition, papers that created datasets without making them publicly available and papers that used existing datasets were not cited differently from papers without datasets. For model (4), it can be concluded that, compared to non-public datasets, papers with datasets created and public and papers using existing datasets have higher citation rates relative to non-public datasets.



	Dependent variable: **CiteNum**			
	(1)	(2)	(3)	(4)
Years Published	0.2367***	0.2534***	0.2535***	0.2659***
FC		-0.5126***	-0.5010***	-0.7794***
IMC		-0.4235***	-0.4519***	-0.5879***
NDSS		0.1292	0.1287	-0.1802*
SP		0.4572***	0.4616***	0.0464
USENIX		0.2652***	0.2613***	-0.0080
WEIS		-0.9355***	-0.9402***	-1.3587***
Workshops		-0.9157***	-0.9433***	-0.7256***
Created Not Public			-0.0109	
Created Public			0.2090*	0.0054***
Existing			0.1233	0.0334**
Vulnerabilities				-0.1334
Exploits				-0.3181
Cybercrime activities				-0.0717
Alerts				-0.0981
Configurations				-0.1443
Applications				0.1086
Network traces				-0.2253*
Topology				-0.4451***
Benchmarks				-0.3366*
Adverse Events				-0.1118
User Activities				0.1975
User Attitudes				-0.0769
User Attributes				0.0929
Alpha	1.1806***	1.0501***	1.0461***	0.7430***
Constant	3.0674	2.9749	2.9360	2.9980
Observations	1142	1142	1142	1003
Log-Likelihood	-6258.1	-6176.4	-6173.8	-5359.1

Table 4.5: Negative Binomial regression tables for all papers (2012-2020) (Note: *p<0.1; **p<0.05; ***p<0.01)

4.4 Summary of Extension

This section has extended the original paper by re-analysing the latest databases and studies with an optimisation approach to identify any improvements in data sharing and the characteristics of cybersecurity datasets. Specifically, there are three aspects that have been improved.

First, I updated the cybersecurity-related papers up to 2020, which expanded the datasets for research. After that, I constructed a new Binary Classifier using the Hard Voting algorithm, which had a high prediction accuracy of over 80%. Although it had somewhat higher false positive and false negative rates, it still performed better than the Random Forest used in the original paper. For the classified datasets, the number of created datasets chosen for disclosure was small, accounting for only 8.5% of all datasets, indicating that researchers continue to show a negative attitude towards disclosing the created data. Moreover, the willingness of researchers to share has become lower in recent years, with most sub-categories showing a slight decline in public rates.

Finally, I revised the regression model and used Negative Binomial regression instead to obtain a better fit. The regression results found that for years of publication, citations increased by 31% for each additional year of publication. For the effect of disclosure on citations, there was a 25% increase in citations per year for papers with publicly available datasets compared to non-public datasets. As can be found, both statistical and regression analyses of citations clarifies that publicly created datasets tend to be more favoured by future researchers.

Chapter 5

Results and Discussion

5.1 Summary of Research

This paper replicates and extends the paper “Cybersecurity Research Datasets: Taxonomy and Empirical Analysis”, which covers the collection of datasets, the construction and analysis of taxonomies, and the empirical analysis. The replication examines the main findings of the original paper and explores its analytical methods, while the extension further investigates the characteristics of cybersecurity datasets with improved tools and data.

In the replication stage, I used the same methodology as the original paper and elaborated various procedures to make the study more complete. During the statistical analysis of the original datasets, it was found that the data in the original thesis had issues with data duplication and others. By analysing the adjusted data, I found that the results were nearly identical to the original paper, indicating that the main findings of Zheng's paper were not impaired.

In the extension stage, I downloaded over 2,000 recent cybersecurity-related papers and randomly examined some papers to categorise their datasets, to reveal what types of data are being created, used and shared with the wider community. By studying more than 1,100 papers from 2012-2020, it is found that most researchers prefer to create or use datasets related to macro-level Internet characteristics. For sub-categories, there is a high proportion of datasets related to Applications, Attacks and User activities, which together account for 48% of all datasets. Notably, most sub-categories have seen a slight decline on disclosure in recent years. In addition, the public proportion of primarily created and derived created data is deficient, with close to 80% of datasets not being made public, and these public datasets account for only 8.5% of all datasets. This trend highlights the fact that researchers still have a negative attitude towards making their data publicly available. Subsequently, I constructed a new Binary Classifier using the Hard Voting algorithm, which had a high prediction accuracy of over 80% and performed better than Random Forest in all respects. Finally, I chose Negative Binomial regression instead to analyse the various features of the cybersecurity dataset. It has a higher AIC and BIC than the multiple linear regression used in the original paper. The regression reveals that each additional year of publication increases the citation rate by 31%. Furthermore, most coefficients for publication location are statistically significant, which indicates that publication location does influence the visibility and frequency of papers being used, with papers published in FC, IMC, WEIS, and workshops being less likely to be cited than CCS. It is also notable that papers containing publicly available datasets had a 25% increase in citations compared to non-public datasets. This finding demonstrates that sharing datasets has a positive effect on the number of citations and subsequent papers. Therefore, greater attention is required to the sharing behaviour of researchers if the cybersecurity field needs to develop. Finally, the classification of datasets had a partial effect on the number of citations, where papers containing Attacks datasets were more likely to be cited than papers containing Topology, Benchmark, or Network trace datasets, possibly implying that Attacks datasets are often important inputs to research.

5.2 Suggestions

In light of the above, we need to focus more on the sharing behaviour of researchers to facilitate the development of the cybersecurity field. Some studies demonstrate that the sharing of research datasets fundamentally depends on the interest, either in terms of the researcher's needs or external incentives [25]. The interests of researchers involve both group and individual interests. From the perspective of group interests, there is a requirement to expand the channels for sharing scientific data, standardise the process of sharing scientific data, reduce the difficulty of sharing scientific data, and reduce the time and effort of sharing scientific data. From the perspective of individual interests, it is necessary to adopt appropriate incentives to encourage researchers in sharing scientific data. Therefore, focusing on incentives to share by removing barriers and rewarding publication is critical. In this regard, I propose two recommendations on encouraging researchers to share, namely to alleviate researchers' concerns through data publication and establish appropriate systems of incentives for various aspects related to sharing behaviour from the researcher's perspective.

5.2.1 Establishing Sound Data Publishing Measures

By establishing sound data publishing measures in the field of cybersecurity, the problem of data intellectual property rights can be addressed, the quality of scientific data can be safeguarded, and the value of data reuse can be increased, thus serving as an incentive for researchers to share their datasets.

In the background of data openness, data publishing is becoming an essential way to promote the storage, publicity, and sharing of research datasets, which has recently received attention from global researchers. Data publishing refers that the data is a major research output that undergoes peer review and public release like a paper with standard and permanent data citation information in order to be referenced in other research papers formally. Such publication has been well established in the life sciences, where GenBank, a subject-based data knowledge base, and Dryad, a multidisciplinary data knowledge base, have achieved a significant impact. There are also public data repositories in cybersecurity, such as the IMPACT [22] project supported by the US Department of Homeland Security, known as The Information Marketplace for Policy and Analysis of Cyber-risk & Trust, which aims to coordinate and develop real-world data and information-sharing capabilities by fuelling the global cyber-risk research community. However, because of its high requirements for datasets, it is not suitable for a large amount of sharing between researchers and is somewhat different from the form of GenBank and Dryad. It can be argued that data publishing can be useful for researchers, the public, the data itself, and the development of scientific careers. Specifically, there are the following points:

(1) Data publishing, as a new data-sharing mechanism, can address the issue of intellectual property rights of data, and protect the academic credibility and rights of datasets creators, thus increasing the motivation of researchers to share data.

(2) Data publishing facilitates public users to discover, access and utilise data. The peer-review process serves as a form of data quality assurance for data users, so data publishing can also highlight the transparency of the scientific research process and improve the sharing environment [23].

(3) Data publishing ensures that datasets can be uploaded to trusted repositories and stored properly. Furthermore, data publications become more valuable when associated with other related resources in the same repository to form a complete value chain [24].

As can be seen, data publication not only has the function of data sharing, but also greatly solves the problem of data intellectual property rights, safeguards the quality of scientific data and increases the value of data reuse. This dramatically alleviates researchers' concerns about the risks of sharing, thereby increasing researchers' willingness to share data. Data publishing, as a standardised way of open data sharing, is bound to be a major trend in the future, and it is hoped to be better applied in the field of cybersecurity.

5.2.2 Establish an Incentive System

The sharing behaviour of researchers is the result of the coordination of external triggers such as institutional, technological, personal, and social factors along with six needs: material needs, data protection needs, social needs, personal fulfilment needs, data accessibility needs and self-worth needs [25]. Among them, material needs, social needs, personal fulfilment needs, data accessibility needs, and self-worth needs promote scientific data sharing. In contrast, data protection needs lead to a conflict with the others and impede the sharing of research datasets.

To satisfy their needs, an incentive system based on the various needs of researchers is required to address conflicts and facilitate sharing of datasets. Specifically:

(1) Improve laws and regulations to secure datasets

In response to the data protection needs of researchers, governments and organisations are required to establish a sound environment for sharing and enact appropriate laws, regulations or other measures to secure datasets. For cybersecurity datasets, legal and privacy issues are often actively stated as a barrier to sharing. The Department of Homeland Security's PREDICT project and its successor, IMPACT, have made a start in this area. Overall, the intellectual property rights, datasets security and datasets privacy of providers in the process of sharing should be protected in the form of laws and regulations to reduce the risks of datasets sharing.

(2) Foster a favourable environment for sharing

On the one hand, the social needs of researchers should be satisfied by organising group activities such as workshops during the process of data sharing. Since, in the act of sharing scientific data, researchers are willing to share data for the sake of maintaining interpersonal relationships [25]. On the other hand, the concept of open-sharing in datasets and its benefits to research should be promoted. Through research institutions, journal publishers and other parties, it is possible to create an interpersonal environment that is fair and supportive, open and lively, which is conducive to establishing an environment of datasets sharing.

(3) Enhance the sense of achievement for researchers

It is recommended that the behaviour of scientific data sharing should be incorporated into the system of career promotion and job evaluation to meet the needs of researchers for personal achievement and self-worth. According to Herzberg [26], the main motivating factors for employees are achievement, recognition of achievement, the work itself, responsibility, promotion, and money. As professional promotion and job evaluation are the focus of the vast majority of researchers, the inclusion of the act of sharing datasets can help satisfy a sense of personal fulfilment and make datasets publicly available.

Chapter 6

Conclusion

6.1 Research Contribution

This paper replicates and extends “Cybersecurity Research Datasets: Taxonomy and Empirical Analysis”. The main findings and methods of the original paper are examined in this thesis, and further improvements are proposed to display the latest features of cybersecurity datasets. It is determined that papers that make the created datasets publicly available have higher citation rates, but that the proportion of shared datasets is consistently low. A key to breaking this status quo is to focus on incentives for sharing by removing barriers and rewarding publication. Accordingly, I offer suggestions on how to improve data sharing behaviour in cybersecurity in the future, with the expectation that it will contribute to the development of cybersecurity.

6.2 Limitation and Future Work

From the regression analysis, it can be concluded that publicly available datasets and the sub-category of datasets do have a meaningful impact on citation rates. Still, there is a lot of unexplained variation in citation rates outside. This can be expected since the choice of explanatory variables in this paper is based solely on the objectives and expected findings and does not take into account other characteristics that may influence the citations, such as economic costs, altruistic factors, reputation expectations, etc. However, with the addition of more explanatory variables, there would definitely be better fitting results as well as a more comprehensive analysis of the cybersecurity dataset could be made. Thus, as a next step, we can consider including the above factors in the scope of the study to further enrich the research content.

Additionally, it is worth noting that citing a paper and using a dataset in research are not identical. It is hoped that future work will determine whether the rise in paper citations reflects the authors' reuse of public datasets in their own research. Unfortunately, it is currently difficult to automatically determine whether datasets are being used directly, as the cybersecurity research community has not yet established norms for citing datasets rather than papers. Therefore, measures such as data publication need to be improved as soon as possible, just like GenBank in life science fields where datasets can be cited directly. I believe this will be possible in the cybersecurity field soon with the increasing emphasis on it.

Bibliography

[1] Zheng M, Robbins H, Chai Z, et al. Cybersecurity research datasets: taxonomy and empirical analysis[C]. 11th Workshop on Cyber Security Experimentation and Test. 2018. Available: https://www.usenix.org/conference/cset18/presentation/zheng

[2] Gibney E, Van Noorden R. Scientists losing data at a rapid rate[J]. Nature News, 2013. Available: https://www.nature.com/articles/nature.2013.14416

[3] Mulligan A, Mabe M. The effect of the internet on researcher motivations, behaviour and attitudes[J]. Journal of Documentation, 2011. Available: https://www.emerald.com/insight/content/doi/10.1108/00220411111109485/full/html

[4] Laube S, Böhme R. Strategic aspects of cyber risk information sharing[J]. ACM Computing Surveys (CSUR), 2017, 50(5): 1-36. Available: https://dl.acm.org/doi/abs/10.1145/3124398

[5] Moore T, Clayton R. The consequence of non-cooperation in the fight against phishing[C]. 2008 eCrime Researchers Summit. IEEE, 2008: 1-14. Available: https://ieeexplore.ieee.org/abstract/document/4696968/

[6] Gordon L A, Loeb M P, Lucyshyn W. Sharing information on computer systems security: An economic analysis[J]. Journal of Accounting and Public Policy, 2003, 22(6): 461-485. Available: https://www.sciencedirect.com/science/article/pii/S0278425403000632

[7] Gal-Or E, Ghose A. The economic incentives for sharing security information[J]. Information Systems Research, 2005, 16(2): 186-208. Available: https://pubsonline.informs.org/doi/abs/10.1287/isre.1050.0053

[8] Pete I, Chua Y T. An assessment of the usability of cybercrime datasets[C]. 12th Workshop on Cyber Security Experimentation and Test (CSET 19). 2019. Available: https://www.usenix.org/system/files/cset19-paper_pete_0.pdf

[9] Coull S E, Kenneally E. Toward a comprehensive disclosure control framework for shared data[C]. 2013 IEEE International Conference on Technologies for Homeland Security (HST). IEEE, 2013: 93-98. Available: https://ieeexplore.ieee.org/abstract/document/6698982/

[10] Costello M J. Motivating online publication of data[J]. BioScience, 2009, 59(5): 418-427. Available: https://academic.oup.com/bioscience/article/59/5/418/297578?login=true

[11] Wang D D, Data Papers: Independent Publishing and Sharing Mode of Dataset[J]. Information and Documentation Services, 2015: 95-98.

[12] 2020 Cyber Security Risk Report. https://www.aon.com/report-cyber-risk-data-breach-impact-to-organization-ip-mergers-acquisitions-security-threats/index.html

[13] D. of Homeland Security. Information marketplace for policy and analysis of cyber-risk

and trust. https://www.impactcybertrust.org.

[14] Hansman S, Hunt R. A taxonomy of network and computer attacks[J]. Computers & Security, 2005, 24(1): 31-43. Available: https://www.sciencedirect.com/science/article/pii/S0167404804001804

[15] Howard J D. An analysis of security incidents on the internet 1989-1995[M]. Carnegie Mellon University, 1997. Available: https://www.proquest.com/openview/26b4425b41777ee9b6cac10b78da998a/1?pq-origsite=gscholar&cbl=18750&diss=y

[16] Meyers C A, Powers S S, Faissol D M. Taxonomies of cyber adversaries and attacks: a survey of incidents and approaches[R]. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2009. Available: https://www.osti.gov/biblio/967712

[17] Suryotrisongko H, Musashi Y. Review of Cybersecurity Research Topics, Taxonomy and Challenges: Interdisciplinary Perspective[C]. 2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA). IEEE, 2019: 162-167.

[18] DBLP: computer science bibliography. https://dblp.org/

[19] Price D J S. Networks of scientific papers[M]. Princeton University Press, 2011. Available: https://www.degruyter.com/document/doi/10.1515/9781400841356.149/html

[20] Didegah F, Thelwall M. Which factors help authors produce the highest impact research? Collaboration, journal and document properties[J]. Journal of informetrics, 2013, 7(4): 861-873. Available: https://www.sciencedirect.com/science/article/abs/pii/S1751157713000709

[21] Guo Z, Chen J, Yang M. Research on Small Sample Target Detection Technology in Natural Scenes[C]. Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence. 2019: 231-235. Available: https://dl.acm.org/doi/10.1145/3330482.3330509

[22] The Information Marketplace for Policy and Analysis of Cyber-risk & Trust. https://www.impactcybertrust.org/

[23] Murphy F. Transforming Research Communication:Knowledge Management and Data Publishing[EB]. 2017. Available: https://ams.confex.com/ams/93Annual/webprogram/Paper217112.html.

[24] Borgman C L. Data, disciplines, and scholarly publishing[J]. Learned publishing, 2008, 21(1): 29-38. Available: https://onlinelibrary.wiley.com/doi/abs/10.1087/095315108X254476

[25] Hu Y Y, Ding F F, Guo H. Research on Data Sharing Behavior and Countermeasures of Scientific Researchers[J]. Journal of Academic Library and Information Science, 2019, 37(6): 24-29.

[26] Herzberg F. Motivation to work[M]. Routledge, 2017.

Appendix

Code

(This section is just a brief code.)

1. Download the paper

```python
def add_excel(papername,pdf_url,have_datasets):
    wb = load_workbook('C:\\Users\\Desktop\\paper.xlsx')
    ws = wb.active
    ws.append([papername,pdf_url,have_datasets])
    wb.save('C:\\Users\\Desktop\\paper.xlsx')

def get_paper(url, folder, filename):
    if not os.path.exists(folder):
        os.mkdir(folder)
    path= folder + '/' + filename.replace('\n','').strip()
    if not os.path.exists(path):
        requests.adapters.DEFAULT_RETRIES = 5
        s = requests.session()
        s.keep_alive = False
        r = s.get(url=url, headers=headers1, timeout=50)
        r.encoding = r.apparent_encoding
        print(r.status_code)
        with open(path, 'wb') as f:
            f.write(r.content)
        size=os.path.getsize(path)/1024
        while(size<50):
            print('try again!')
            requests.adapters.DEFAULT_RETRIES = 5
            r = requests.get(url=url, headers=headers)
            r.encoding = r.apparent_encoding
            with open(path, 'wb') as f:
                f.write(r.content)
        print(f"{filename} download successfully")
    else:
        print(f"{filename} already exists")

def getHTMLText(url):
    try:
        r= requests.get(url, headers=headers1, timeout=30)
        r.encoding= r.apparent_encoding
        return r.text
    except:
        return "getHTMLText error!"

def get_paper_name(html,s):
    soup= BeautifulSoup(html, 'html.parser')
    title=''
    if('ieeexplore.ieee.org' in s):
        for content in soup.find('title'):
            title=str(content)
    else:
        if(soup.find('h1')==None):
            return None
        for content in soup.find('h1'):
            title=str(content)
    intab = "?/|\.><:*\""
    for s in intab:
        if s in title:
            title = title.replace(s, ' ')
    return title.replace('   IEEE Conference Publication   IEEE Xplore','').strip()

def get_pdf_url(html,s):
    soup= BeautifulSoup(html, 'html.parser')
    for link in soup.find_all('a'):
        url= link.get('href')
        if ('https://dl.acm.org' in s): #CSS
            return s.replace('doi/','doi/pdf/')
        if ('link.springer.com' in s):
            aa=s.replace('chapter/','content/pdf/').replace('1007/','1007%2F')
            return str(str(aa)+".pdf")
        if (url!=None) and (url[0:36]=='https://www.usenix.org/system/files/'): #USENIX
            return url
        if ('ieeexplore.ieee.org/document' in s): #IEEE
            return s[:-1].replace("document/", "stamp/stamp.jsp?tp=&arnumber=")
        if (url!=None) and (url[-9:]=='paper.pdf'): #NDSS
            return link.get('href')
    return None

def download_one_paper(url):
    print(url)
    html= getHTMLText(url)
    papername= get_paper_name(html,url)
    if(papername==None):
        print('no name')
        return
    pdf_url= get_pdf_url(html,url)
    if(pdf_url==None):
        print('no pdf_url')
        return
    print('pdf_url: '+pdf_url)
    add_excel(papername,pdf_url,0)

with open('list.txt', 'r') as f1:
    list = f1.readlines()
for line in list:
    line=line[:-1]
    download_one_paper(line)
```

2. Crawling the citation

```
def get_citNum(name):
    citeNum=''
    driver = webdriver.Firefox(executable_path='D:\\geckodriver.exe')
    driver.implicitly_wait(30)
    driver.get(scholar_url)
    driver.find_element_by_id("gs_hdr_tsi").send_keys(name)
    driver.find_element_by_id("gs_hdr_tsb").click()
    url=driver.current_url
    res = requests.get(url=url,headers=headers)
    soup = BeautifulSoup(res.content,'html.parser')
```

```python
    soup2=soup.find_all('div', {"class": "gs_fl"})
    for i in soup2:
        info=str(i)
        loc=info.find('Citation')
        if(loc!=-1):
            info=re.match(r'(.*)Citation: (.*)</a> <a href(.*)',info)
            citeNum=info.group(2)
            print("citeNum= "+str(citeNum))
            break
    driver.quit()
    return citeNum

wb = load_workbook('C:\\Users\\Desktop\\CyberPapers.xlsx')
ws = wb.active
global citation
for i in range(1,2346):
    name=ws['A'+str(i)].value
    citation=get_citNum(name)
    ws['F'+str(i)].value=citation
    print(f"Paper {i-1}:{citation}")
    if(i%5==0):
        wb.save('C:\\Users\\Desktop\\CyberPapers.xlsx')
        wb = load_workbook('C:\\Users\\Desktop\\CyberPapers.xlsx')
        ws = wb.active
wb.save('C:\\Users\\Desktop\\CyberPapers.xlsx')
```

3. Data process

```python
def writefile(total,filename):
    output = open(filename,'a', encoding='utf-8')
    for i in range(len(total)):
        for j in range(len(total[i])):
            output.write(str(total[i][j]))
            output.write(',')
        output.write('\n')
    output.close()

def readfile(filename):
    file = open(filename, "r", encoding='utf-8')
    lines = file.readlines()
    total = []
    line = []
    for i in lines:
        line = list(i.split(','))[:-1]
        total.append(line)
    file.close()
    return total

def delete_words(paperplace, storeplace):
    n=0
    total=[]
    lines = open(paperplace,'r', encoding='utf-8').readlines()
    for line in lines:
        n=n+1
        if(n%10==0):
            writefile(total,storeplace)
            print('10 is saved!')
            total=[]
        paragraph=TextBlob(line.lower())
        word_list=paragraph.words
        for i in range(len(word_list)):
            word_list[i]=word_list[i].lemmatize("v")
        filtered_words = [word for word in word_list if word not in words and "cid" not in word and not word.replace('.','').replace('-','').isdigit() and not len(word)<=2]
        total.append(filtered_words)
        print(f'{paperplace}--{n}')
    writefile(total,storeplace)

#delete stop words
words = stopwords.words('english')
global total, n
for w in ['!',',','.','?','-s','-ly','</s>','s','(',')','"']:
    words.append(w)
for w in range(97,123):
    words.append(chr(w))
for w in range(65,91):
    words.append(chr(w))
paperfrom=['extension_string1-400.txt','extension_string401-800.txt','extension_string801-1200.txt','extension_string1201-1600.txt','extension_string1601-2000.txt','extension_string2001-2346.txt']
for j in range(6):
    delete_words(paperfrom[j],'extension_after_all.txt')

#cleanup the words
new_total=[]
new_words=[]
total=readfile('extension_after_all.txt')
print(len(total))
for i in range(len(total)):
    all_words=total[i]
    new_words=[]
    for j in range(len(all_words)):
        if(all_words.count(all_words[j])>3):
            if(re.match(r'^[A-Za-z]+$',all_words[j])!=None):
                new_words.append(all_words[j])
    new_total.append(new_words)
writefile(new_total,'new_extension_total.txt')

#Calculate TF-IDF for every words
total=readfile('new_extension_total.txt')
print(len(total))
all_words=list(set(chain.from_iterable(total)))
print(len(all_words))
tfidf=[]
aa=[]
n=0
corpus=TextCollection(total)
for i in range(len(all_words)):
    tfidf.append(corpus.tf_idf(all_words[i],corpus))
aa=dict(zip(all_words,tfidf))
after=sorted(aa.items(), key=lambda x: x[1], reverse=True)[:1000]
save_obj(after,'tfidf_extension')

#Calculate TF-IDF for every papers
total=readfile('new_extension_total.txt')
aa=load_obj('tfidf_extension')
tfidf=dict(aa)
tfidf_total=[]
for i in range(len(total)):
    tfidf=dict(aa)
    for key in tfidf:
        if(key not in total[i]):
            tfidf[key]=0
    aaa=list(tfidf.values())
    tfidf_total.append(aaa)
    n=n+1
    print(f'{n} finish!')
writefile(tfidf_total,'extension_tfidf.txt')
```

4. Machine learning model

```
def get_y(List):
    y=[]
    wb = load_workbook('C:\\Users\\Desktop\\total_paper.xlsx')
    ws=wb['Sheet1']
    for i in List:
        y.append(ws['C'+str(i)].value)
    return y

classified_x=[]
for i in classified_list:
    classified_x.append(tfidf_total[i-1][:400])
for i in range(len(classified_x)):
    for j in range(len(classified_x[i])):
        classified_x[i][j]=(float)(classified_x[i][j])
feature=classified_x
target=get_y(classified_list)

vc=VotingClassifier(estimators=[('rfc',RandomForestClassifier(i)),('gnb',GaussianNB()),('gbc',GradientBoostingClassifier(n_estimators=61))],voting='hard')
vc.fit(feature,target)
predict_results=vc.predict(feature)
conf_matrix = confusion_matrix(target, predict_results1,labels=[1,0])
TP=conf_matrix[0,0]
FN=conf_matrix[0,1]
FP=conf_matrix[1,0]
TN=conf_matrix[1,1]
print(f'Accuracy={(TP+TN)/(TP+FP+FN+TN)},  FPR={FP/(FP+TN)},  FNR={FN/(FN+TP)}')

wb = load_workbook('C:\\Users\\Desktop\\paper.xlsx')
ws=wb.active
for i in range(len(predict_results)):
    ws['C'+str(i+2)].value=predict_results[i]
wb.save('C:\\Users\\Desktop\\paper.xlsx')
```