University of BRISTOL

Department of Computer Science

‘CyBOK of Evil’ What does a cyber criminal learn from dark web training material?

Andrew Gordon Thomas

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

September 2020 - CSMSC-20
Word count: 18501

Declaration:

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work.

Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references.

Any views expressed in the dissertation, other than referenced material, are those of the author.

Andrew Gordon Thomas, September 2020

Executive Summary

The anonymity and accessibility of the dark net has enabled a cyber criminal subculture to thrive. The implications of this from a cyber security perspective are significant as such an environment provides protection for criminal guides and exploits to be traded with impunity. Crucially, a cyber criminal no longer has to develop technical competence through a traditional formal curriculum as the knowledge or services can simply be purchased through informal channels on the darknet [72].

This research project analysed the learning material available on dark web markets and forums as a means to improve our understanding of the type of material that cyber criminals are likely to be consuming. To achieve this, the content was mapped to knowledge areas within the Cyber Security Body Of Knowledge (CyBOK) [62] in order to determine the areas with least and most emphasis. The maps were also evaluated against similar maps of formal cyber security qualifications [40] in order to compare their curriculums with the available cyber criminal learning material. A keyword classifier program was constructed and utilised in order to categorise the full data set within the required timescale.

To achieve these aims the project implemented the following:

- Undertook thematic analysis of the source data in order to develop a codebook defining the key categories present within the darknet learning material.
- A keyword classifier program that can automate the assignment of categories to string items based on the prevalence of keywords through which the majority of the source data will be parsed.
- A statistical analysis of the results to create a mapping to CyBOK knowledge areas in order to determine the area(s) with greatest and least emphasis respectively.
- The produced mapping will be compared with the mappings of formal cyber security qualifications [40] to determine any relationship between this and the informal cyber criminal curriculum on the dark web.

The results were then contrasted against prior research into the dark web as well as analysed on a yearly basis to detect any emerging trends. Lastly, the implications of our findings were discussed in the context of the wider literature.

Acknowledgements

I would like to thank my dissertation supervisors for their invaluable guidance throughout the course of this project.

In addition to this, I would like to thank them individually for the following:

Dr Matthew Edwards For sourcing the darknet data sets that were used for analysis.

Dr Joseph Hallett For providing the code used to generate the bar and spider charts and taking the time to help validate the codebook.

Thank you very much for your contribution to this project.

Contents

List of Figures

List of Tables

1 Introduction and Project Aim

1.1 Introduction

The dark web is the hidden internet, a section of the deep web that has been intentionally hidden and is not accessible through standard web browsers. The user base of the dark web has increased significantly since its inception due to increased publicity as well as the accessibility of specialised routing services [32]. One of the most frequently used is Tor. Short for The Onion Router, Tor is a distributed overlay network designed to anonymise TCP-based applications such as web browsing, secure shell, and instant messaging [66]. This allows users to access the internet and the dark web anonymously by routing their web traffic through other users' computers such that the traffic cannot be traced to the original user [32].

Within the last decade a number of dark web markets and forums have emerged where goods and services are traded anonymously [72]. These websites facilitate the trade in a wide range of illegal products and services. These not only include drugs and weapons but also a range of cyber criminal learning material [73].

Dark web marketplaces, such as the infamous Silkroad, operate and appear much like any online marketplace on the open internet. Transactions between buyers and vendors are anonymised through the use of cryptocurrency such as Bitcoin [56] and communications are encrypted using PGP [34]. A significant amount of cyber criminal material is traded on these dark web markets with research conservatively estimating that the overall revenue generated for cyber criminal commodities on dark web markets as being at least US $15M between the years 2011–2017 [75]. This also includes criminal learning material such as guides and tutorials that cyber criminals can use to improve their skills.

These illegal services are also discussed and traded on dark web forums. These forums constitute arenas in which the propagation of hacking techniques as well as discussion on cracking and ethics take place [69]. Much like the markets, these can be of considerable scale with certain forums containing thousands of posts [12]. By analysing the content of the communication between malicious hackers on these forums this study can provide insight into the concerns, motivations, and goals as well as the environment in which they act [69]. Of concern for cyber security professionals is that research suggests many of these guides, if they are meticulously followed, are effective and leave limited interceptive opportunities for law enforcement [73].

In order to formulate comprehensive strategies and policies for governing the internet, it is important to consider insights on its farthest reaches such as the dark web [21]. While previous research has already detected the presence of cyber criminal learning material on the dark web, beginning to map the content of this material can allow us to better understand the motivations of users on the dark web as well as evaluate the appropriateness of cyber security qualifications to counter this material.

Previous research into the content of cyber security qualifications has revealed that the focus of these qualifications differ considerably [40]. While only an initial study, it detected that some qualifications are heavily weighted towards only a few areas within cyber security and some areas are greatly under-represented across all qualifications. This in itself is not necessarily bad, there are a variety of valid reasons why such discrepancies might exist [40]. However, the implications of this study need to be evaluated in further detail to determine if there are any adverse consequences to this disproportionate weighting. Using the same benchmark of CyBOK knowledge areas, the mappings created through this research will be directly contrasted with that of the cyber security curriculums. By contrasting our findings with the curriculums of traditional cyber security qualifications, this study will determine whether any relationship exists between them.

This project also intends to identify any emerging trends on the dark web by analysing the learning material on a yearly basis. By detecting these trends, this can help us understand areas within cyber security that are likely to continue to be targeted by cyber criminals in the future. For example, if the number of items related to money laundering is increasing every year it is indicative that the current state of anti-money laundering (AML) techniques and regulation is not effectively deterring this activity. If such trends are identified, this will also allow cyber security professionals to prepare and enact preventative measures if necessary, in order to deter these trends.

Previous research has stressed the importance of mapping content on the dark web [23]. In this regard, the study will produce a program that can be improved upon or repurposed to provide more detailed and up-to-date mappings of dark web content. For example, integration with an automated web scraper to provide real-time updates on dark web trends, allowing for early threat detection and prevention. Data representations, in the form of spider maps and bar charts, will also be created in order to summarise our findings. This will provide a direct comparison with the representations created from research into cyber security curriculums [40].

1.2 Project aim & objectives

This research project aims to map the learning material available to cyber criminals on the dark web onto CyBOK knowledge areas in order to determine the content of a typical cyber criminal curriculum. By contrasting our mapping with similar research into formal cyber security curriculums [40] as well as prior research into the dark web we can compare the current state of cyber security education to a typical cyber criminal education on the dark web.

More generally, this project also seeks to determine whether there is learning material that is accessible on the dark web which is not adequately covered within the scope of CyBOK or the formal cyber security curriculum.

The project has the following objectives:

1. Analyse the dark web learning material to determine the most frequently occurring topics.

2. Once a general understanding of common themes has been established, a codebook defining these key categories will be created and validated. The aim was to manually categorise at least 3000 dark web market listings that would form the ground truth data set, which would act as the initial training data set for the classifier.

3. Construct and cross-validate a classifier program that can automate the assignment of categories to string items based on the prevalence of certain keywords through which the majority of the source data will be parsed and categorised.

4. Create data representations of our findings, including spider maps, that reflect the variation of cyber criminal learning material available on the dark web according to CyBOK knowledge areas.

5. Contrast our findings with previous research to inform the current state of cyber security knowledge. Specifically the key differences between the learning material accessible on the dark web versus what is being taught via accredited courses will be discussed.

6. These results will also be split into yearly subgroups and the differences analysed to identify any emerging trends.

2 Background and Context

This research project concerns both the dark web as well as its implications for cyber security and in particular, cyber security qualifications. In this chapter, I outline an overview of each of these areas as well as discuss and evaluate the relevant literature to provide an understanding of the context behind our research project.

2.1 The Dark Web

2.1.1 Definition and origins

Before diving into background research, it is important to clarify the terminology used when discussing the dark web as this can often be very murky [36]. Journalism, academic literature, and popular books provide competing and contradictory definitions [36]. Some erroneously conflate it with the deep web [38], defining it as comprising everything not indexed by a search engine. Others still consider it through a moral lens, defining it very broadly as anything bad that happens on the Internet. This definition would also encompass much activity present on the open internet such as trolling or phishing [36]. An example of this is in a study by the University of Arizona which included terrorist activity discovered on the nonweb virtual world Second Life as "dark web" material [18].

For the purposes of this study, the dark web is referring specifically to a section of the deep web that has been intentionally hidden and is not accessible through standard web browsers without the use of specialist routing software such as Tor. This is the technical definition that has a basis in prior research as the "former definition is technically misleading and the latter is subject to contentious debate," [36].

Additionally, the term "dark net", though often used interchangeably with "dark web" [20], refers to the collection of networks and technologies used to share digital content on the dark web [29]. In this context, it is the underlying proxy network, such as Tor [66], that enables the dark web to exist. The difference between the dark web and the darknet can be thought of as analogous to that of the World Wide Web and the Internet [8].

The deep web, comprising content not indexed by a traditional search engine (there are search engines which specialise in indexing dark web content [1]), is substantial in size with recent estimates [67] suggesting it is approximately 400–500 times larger than the surface web, also known as the Internet, that we normally use. Much of this deep web is perfectly benign and simply includes content that is behind paywalls, login screens or within databases [36]. The dark web however constitutes only a tiny fraction of this but it is growing rapidly. An initial 2017 report [57] identified only around 4,000 dark web sites. Two years later another report [26] discovered there were 8,416 active dark web domains, more than doubling in size.

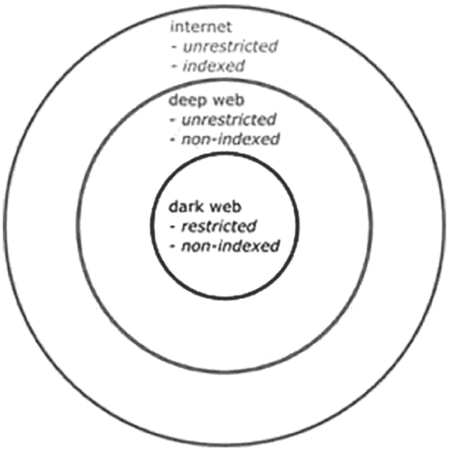


Figure 2.1: Graphical representation of the relationship between the Internet, the deep web and the dark web [29].

Nonetheless, the dark web should not be disregarded from a cyber security perspective. One powerful aspect of the dark web is the level of anonymity it provides for its users. The location of servers hosting dark web sites (also known as hidden services) are not provided to the client users [15]. This makes it essentially impossible for forensic researchers to investigate using traditional means like examining a server's IP address or by checking for registration details. There are two obvious advantages of this for the cyber criminal: the government cannot find out who is running the service, nor can they shut it down [15].

Most sites on the dark web make use of Tor, short for "The Onion Router", open source software that acts as an overlay network providing online protection [66]. Originally developed in the 90s by the US Navy as a means for military personnel to communicate abroad anonymously, Tor utilises powerful encryption and a network of volunteers which make it virtually impossible to find a user's real identity when they access a site using it [47]. Using Tor it is simple to set up a dark web site and neither the Internet Service Provider (ISPs) that route the traffic, nor law enforcement agencies, nor even the developers of the Tor project itself, have visibility into the hosted servers' location or the identity of its operator. The end result is an anonymous network that allows users to communicate while hiding their identities from one another and from third parties [31].

Released to the public in 2002, Tor very quickly attracted thousands who wanted to use it for a variety of purposes, ranging from legitimate to highly illegal, due to its anonymous nature [47]. While there are undoubtedly positive uses for this new technology, such as aiding dissidents in avoiding the control of authoritarian regimes, the ability to traverse the Internet with complete anonymity also nurtures a platform ripe for those who wish to engage in illegal activities [21]. Notably, this makes the dark web particularly useful for cyber criminals, who are constantly trying to hide their tracks [59] and now has the potential to host an increasingly high number of malicious services and activities.

Some research has been conducted into the dark web in order to categorise and analyse its contents. A 2019 study by Faizan & Khan manually assessed and categorised 3,480 English language dark web sites in order to determine the composition of the dark web [30]. The study found the websites were spread across a wide range of categories with no one category comprising more than 10% of websites. Many were Bitcoin related including hidden services for Bitcoin mixing, Bitcoin doubling, Bitcoin wallets, and trading. The largest category, that of services, contains a broad range including e-mail, encryption-decryption, anonymity and privacy tools, escrow and jabber and many others. Other cyber criminal categories with a significant presence include carding (7.8%), ethical hacking (3.2%) and Tor (1.9%).

This study, along with many others [18, 47, 70], demonstrate that the dark web facilitates a wide variety of criminal transactions for a range of malicious actors – from drug and arms dealers to terrorists to hackers for hire. Many studies have sought to map the dark web, typically from a particular perspective or focusing upon a particular crime type such as terrorism [18]. One such study analysed a corpus of over 4000 unique dark web pages to determine the motivations of users on the dark web [27]. It found the dark web was impacted (but not dominated) by illicit commerce and money laundering, but almost completely devoid of violence and extremism. In short, it appears criminality on this'dark web'is based more upon greed and desire, rather than any particular political motivations.

Of some concern for cyber security professionals, not only is the dark web growing but research has indicated that the proportion of the dark web dedicated to illegal content is also increasing steadily. A 2016 study that mapped dark web sites indicated that only 28% hosted illicit material [55]. However, a subsequent 2019 study discovered the number of dark web sites that hosted illicit content had risen by 20% since the previous study [51]. Notably the rate of criminal adoption is far outpacing that of other users, such as privacy activists [48]. According to these latest statistics, this would mean over 400,000 criminals are using the dark web every day. The increase in its user base, particularly the percentage engaging in illicit activity, is indicative that criminals are increasingly becoming more aware of the benefits provided by the dark web.

2.1.2 Hackers on the dark web

Within the criminal sphere of the dark web, the presence of hackers has been well documented [47]. These hackers distribute malware, exchange attack methods, share known vulnerabilities in networks or software and collaborate to breach difficult cyber defenses. On the more severe end, there are tools on the dark web available for download that have been developed to attack industrial control systems and take power grids off-line as well as botnets, such as Storm Bot 2.0 capable of 300 gigabits per second of attack, enough to "knock small countries offline." Often the worst cyber attacks have originated from within the dark web. For example, the Target breach of 2013, one of the largest data breaches in history was caused by malware purchased on the dark web [32]. This research demonstrates first hand the severe threat that can be posed by cyber criminal communities within the dark web.

Dark web hackers are concerned with a wide variety of cyber criminal material. One notable example is the trade in software exploits. The market price for an exploit factors in how widely the target software is used as well as the difficulty involved in cracking the software [53]. In particularly high demand are "zero-day" exploits which are bugs that have not yet been discovered and fixed by software and antivirus companies [47]. Zero-day exploits enable particularly stealthy and sophisticated attacks against specific targets, giving rise to APTs, advanced persistent threats.

These pose a significant threat to both companies and governments [37]. In addition, as the economic damages of a successful APT attack can be very high it is one of the main reasons companies invest in cyber security measures [61]. The widespread accessibility of these exploits that the dark web could enable is especially dangerous as the likelihood of being detected in one of these attacks is effectively nil.

Dark web sites also facilitate the trading of proprietary information, particularly lists of stolen usernames and passwords known colloquially as "combo lists" [9]. These can be for a wide range of websites and many are often of significant scale. Organisations as large and diverse as Marriott International (383 million), Exactis (230 million customers), Quora (100 million), Dixons Carphone (5.9 million) and Cathay Pacific (9.4 million) have all suffered major data breaches [9]. One vendor detected nearly 28 billion credential-stuffing (a cyberattack in which credentials obtained from a data breach on one service are used to attempt to log in) attempts between May and December 2018 alone [9].

One widely reported list, dubbed 'Collection 1-5', featured over 2.2 billion unique usernames and passwords [9]. Cyber criminals only require a small percentage of these credentials (1-2%) in order to work on other accounts and they can generate a decent return on investment. The result is customer identity theft on a massive scale, hitting brand reputation and corporate security risk from spear phishing and business email compromise (BEC) if internal accounts are hijacked [9].

Money laundering, often involving the use of cryptocurrencies, and the trade in credit card information — commonly referred to as carding — also feature prominently [72]. A study even identified numerous dark web websites which specialise solely in these activities [21]. The cryptocurrency Bitcoin in particular can be easily laundered through unregulated exchanges which avoid identity checks. Bitcoin is a decentralised digital currency [56] heavily utilised on the dark web [71]. Though not entirely anonymous, the components of Bitcoin, such as addresses, private and public keys, and transactions, are all read in text strings, such as a public address, that in no way directly link to anyone's personal identity [3]. These characteristics make Bitcoin a particularly appropriate vehicle for money laundering.

Responding to these new developments, police forces devised some methods for tracing the actors involved in Bitcoin transactions by tracing them back to an interaction with a Bitcoin exchange [65]. However, to counter this, cyber criminals are now turning in droves to Bitcoin mixers which aim to disassociate the cryptocurrency from their often-criminal source [74]. These are online third-party services which break down Bitcoins into many different parts and mix those parts with other broken parts from other clients thereby obscuring their origins from law enforcement. Another concerning development is the introduction of high privacy cryptocurrencies such as

Monero. These further obscure the transaction chain making it even more difficult for law enforcement to track [47]. As an anonymous cryptocurrency is essential to dark web trade as it preserves the identity of buyers and vendors. The emergence and widespread adoption of a new highly anonymous cryptocurrency would likely only increase confidence for the users of these markets.

2.1.3 Markets

Much like legitimate services on the open internet, illegal services and items are often discussed, shared or sold on dark web forums and marketplaces. Markets are websites, such as the notorious Silk Road [22], where goods and services are traded anonymously. These range from tame items such as books and clothes, to more illicit goods such as drugs and weapons. These markets appear much like any e-commerce market on the open internet with features including a user interface, high resolution photos and product descriptions [47]. User reviews and ratings help to maintain trust among anonymous users and maintain the overall integrity of the market [9].

These markets not only facilitate the trade in illegal drugs and weapons but also a wide range of cyber crime learning material that cyber criminals can use to improve their skills [48, 55]. These markets are not a niche concern. The combination of accessible anonymity software, the publicity surrounding dark web sites and the use of cryptocurrency to facilitate anonymous transactions has enabled these dark web markets to thrive [71]. Between 2011 and 2013, the Silk Road market alone processed more than 1.2 billion dollars' worth of transactions [47].

These transactions are anonymised through the use of cryptocurrency, typically Bitcoin [56], and communications between buyer and seller are encrypted using encryption software such as PGP [34]. The dark web economy is highly decentralised and adaptable [71]. These marketplaces often disappear, typically due to law enforcement action or exit scams, but it is not long before another replaces it. For example, there have so far been three incarnations of Silk Road, the first modern and most notorious cryptomarket [58].

There has been some research into the variety of content available on these markets. One such study focused on Alphabay, a large market that came to prominence after the fall of Silkroad [13]. Baravalle & Lee found that the majority of listings on the site were for drug products, comprising almost half (45%) of the total market. This was followed by listings relating to fraud (13%), such as fake identification documents, with all other categories representing only a very small portion of the marketplace.

2.1.4 Forums

Another area where cyber criminal knowledge is shared is on dark web 'black hat' hacker forums as well as on forums that accompany the previously discussed cryptomarkets. Web forums are discussion sites which support online conversations. They capture each conversation in a "thread" with a title description and the ensuing postings are usually time-stamped and attributable to a particular online poster [18]. These forums constitute arenas in which the propagation of hacking techniques as well as discussion on cracking and ethics take place. Like the markets, these forums can be of considerable scale with certain forums containing thousands of posts [12]. To demonstrate this, one such dark web forum included in our study, called Crackingfire has been determined to have approximately 14,000 active users [46].

By analysing the content of the communication between malicious hackers we can gain an insight into the concerns, motivations, and goals as well as the environment in which they act [69]. An intimate understanding of these communities will greatly aid proactive cyber security, by allowing cyber security practitioners to better understand their adversaries. For example the concerns, ambitions, and modi operandi of malicious hackers are often showcased in these forums, suggesting that a thorough understanding of how these communities operate will aid in the early detection of cyber attacks [69].

Prior research into dark web forums has focused on terrorist and extremist groups [18]. One such study analysed dark web forum posts in order to seek out users who were starting to show radical tendencies [60]. The aim was to determine whether a relationship was present between the levels of radical tendency on these forums and real life terrorist incidents. In order to achieve this, the researchers utilised a part-of-speech tagger to isolate keywords and nouns which were then entered into a sentiment analysis program to determine the polarity of the post. The semantic analysis program is essentially a classifier, assigning each post with a score that range from -5 (very negative) to 5 (very positive). This approach proved effective, finding sentiment scores over time correlated to real world terrorist events. By categorising a large amount of forum titles using an automated classifier to establish trends on the dark web, this research intends to utilise some of techniques used in this successful study.

Other research has also focused specifically on hacker forums. A study by Almukaynizi et al. used social network analysis alongside machine learning techniques to analyse dark web discussions and predict future cyber threats [10]. The study concluded that there were features which can be computed from hacker social networks that can provide important indicators of future cybersecurity incidents, demonstrating conclusively how analysis of these forums can provide threat intelligence for cyber security professionals.

2.2 Implications for cyber security

As discussed, prior research has already demonstrated that the dark web is a major platform for cyber criminal activity. The question now is: how do cyber security professionals react to mitigate these threats?

The most obvious solution would appear to be to simply take these websites offline. However, this is easier said that done. As discussed previously, the Tor network hides the IP address of hidden services making them difficult to take down via conventional means. This does not make them completely impenetrable however and therefore this has not stopped crime agencies from attempting to seize the most notorious sites involved in criminal activities. A 2017 analysis by Chertoff discussed the effects of law enforcement taking down two different dark web sites: Playpen and Silk Road [20].

In the case of Playpen, a dark web site dedicated to child abuse, the FBI was able to infiltrate the hosting web server. They were able to do this after a source reported that the site was misconfigured and therefore, leaking its real IP address [6]. The FBI then took the unprecedented step of seizing the Playpen server and transferring the site to an FBI server. They then used a hacking tool to identify the IP addresses of users accessing the site. It proved highly effective and resulted in sufficient evidence to bring about 1500 cases against people accessing images of child abuse on Playpen. Although this action was only possible due to the carelessness of the service operator, it was undoubtedly able to prevent continued access to and production of images of child abuse.

The take down of Silk Road, Chertoff argues, was less successful. Although the operator was arrested, after the event there was an explosion in the dark web market for illegal goods. The seizure did little, if anything, to dissuade people from starting new dark web marketplaces. Another study looked at the shutdowns of multiple major dark web marketplaces and concluded that there was no evidence that these large scale exits deterred buyers or sellers from continuing to engage in online drugs sales and purchases, with new platforms rapidly arising to replace those taken down [14].

As the seizure of Silk Road demonstrated, this tactic of merely taking down a site but not pursuing any of its users, is not only resource intensive but has a short-term pay-off and is largely ineffective in the long run as other marketplaces will simply pop up to meet the demand. Chertoff summarises that the seizure of a dark web site is most likely to be effective if it can be safely assumed that every user will be a criminal and therefore serve as evidence for criminal prosecution, which will not be the case for the dark web markets and forums where cyber criminal material is traded and discussed.

In light of this, cyber security researchers have since stressed the importance of monitoring the dark web in order to detect and prepare for emerging cyber criminal threats from within dark web communities [9, 23]. One such study conducted by Ciancaglini et al. [23] proposed a number of methods we should be utilising to monitor the dark web. One such method involves the building of a semantic database which contains important information regarding a hidden site. This information can then be used to help track future illegal activities on the site and associate them with malicious actors. A similar article by Victor Acin also supported a continued monitoring of the dark web. It stressed that any data gathered "must be fresh, targeted, contextualised and actionable" to be of any practical use to cyber security professionals [9].

While our study will not be focused on creating profiles based on individual users, this study will involve creating a semantic database relating to knowledge areas within cyber security that can be used to gauge trends within the cyber crime community more generally. There is also scope for the model to be re-purposed for subsequent research or fitted with a web crawler in order to provide ongoing analysis so these trends can be measured in real-time.

2.3 Cyber security qualifications

Previous research clearly indicates a skills shortage within the cyber security industry [33]. The current and future demand for cyber security skills looks likely to be outstripping supply. As evidence of this, a 2016 report found that almost half of UK organisations feel they lack the in-house skills required to handle the cyber threats they are facing [2]. The UK National Cyber Security Strategy identified multiple issues that need to be addressed in order to close this gap [4]. These range from a lack of young people entering the profession, a lack of exposure to cyber and information security concepts in computing courses as well as the absence of established career and training pathways into the profession [4]. This is a concerning trend when contrasted with the increasing numbers of cyber criminals operating on the dark web as it suggests the profession will be ill-equipped to manage the expected growth in cyber crime.

Currently there are a range of cyber security qualifications available for students looking to enter the industry. Although they have a variety of goals, many of these intend to equip their graduates with the skills required to counter and deter cyber criminals. Just as cybersecurity professionals must hone their skills continually to keep up with a constantly shifting threat landscape, cybersecurity programs need to evolve to ensure they continue to produce knowledgeable and effective graduates who will be desired by employers [43]. It is therefore essential that the state of cyber security education adapts to reflect this new reality.

Contrary to what one might expect, cyber security qualifications are not all alike. Previous research has already determined that the focus of different cyber security qualifications vary considerably [40]. This study by Hallett et al. analysed the curricular frameworks of 4 cyber security qualifications to determine their emphases in terms of cyber security, using CyBOK as a basis for comparison. Not only did the emphasis of each qualification differ, the research also revealed that there were areas of cyber security, such as physical-layer and hardware security, which were heavily under-represented. In addition to this, one curricular framework (IISP) was almost entirely focused on only two of the five knowledge areas.

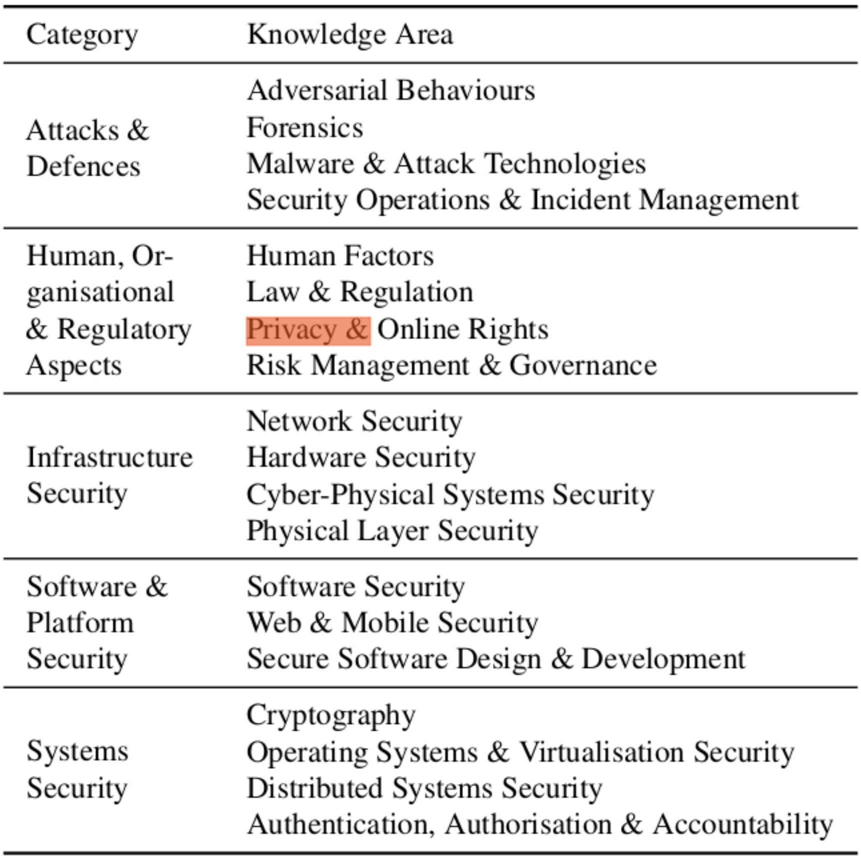
The study is careful to point out that these findings may not necessarily a bad thing. It may be the case that employers prefer candidates with skills in the more heavily weighted categories and the course directors are catering to this demand. It is worth noting that while this study seeks to determine if the weighting of these curriculums are proportionate to cyber criminal threats posed by hackers, cyber security is a broad field and not every aspect revolves around tackling criminality. It also includes areas such as cryptography, one of the more mathematical aspects of cyber security, where much of the focus is on algorithms and mathematical proofs [62].

There are some limitations to this study. It does not account for the time the frameworks allocate to individual topics. It may be the case that a topic is briefly covered in the curriculum framework but covered in great depth when taught or vice versa. It also does not analyse the curriculum of every available cyber security qualification, therefore it is possible these may differ significantly from the findings. Even with these limitations in mind, the study does provide a good general overview of the content present within some of the most prominent cyber security qualifications.

By contrasting the knowledge area mappings created by the Hallett et al. study against the mappings of learning material on the dark web, we can further evaluate the weightings of these respective curriculums. This will allow the study to determine how these curriculums relate to and whether they compliment the material cyber criminals are consuming on the dark web. In addition, this comparison can allow us propose specific changes to to these curriculums in terms of emphasis or content to address these deficits. There are difficulties in determining the exact framework that a self-taught cyber criminal undergoes on the dark web as there is no set curriculum from which every cyber criminal graduates. A individual cyber criminals learning experience will likely be formed by a combination of their unique personal skill set and intentions. With this limitation in mind, by analysing the prevalence of certain subjects available on the dark web we can make reasonable assumptions about the learning areas most in demand within the cyber criminal underworld.

2.3.1 The Cyber Security Body of Knowledge

CyBOK is a broad guide to foundational cyber security knowledge which has been developed through consultation with industry and academia [62]. As the name suggests, it intends to act as a guide to the body of knowledge; the knowledge that it codifies already exists in literature such as textbooks, academic research articles, technical reports, white papers, and standards [62]. It intends to follow on from other such bodies of knowledge in computer science, such as Software Engineering Body of Knowledge (SWEBOK), aimed at identifying and describing the body of knowledge a software engineering professional with an undergraduate degree [16].



Category	Knowledge Area
Attacks & Defences	Adversarial Behaviours Forensics Malware & Attack Technologies Security Operations & Incident Management
Human, Organisational & Regulatory Aspects	Human Factors Law & Regulation Privacy & Online Rights Risk Management & Governance
Infrastructure Security	Network Security Hardware Security Cyber-Physical Systems Security Physical Layer Security
Software & Platform Security	Software Security Web & Mobile Security Secure Software Design & Development
Systems Security	Cryptography Operating Systems & Virtualisation Security Distributed Systems Security Authentication, Authorisation & Accountability

Table 2.1: Overview of the 19 CyBOK knowledge areas and their categories [40].

CyBOK distills cyber security knowledge into 19 top-level knowledge areas and 5 broad categories. Previous research has established the use of CyBOK knowledge areas as a basis for comparison due to its broad foundational scope [40]. By mapping forum titles and market listings onto the CyBOK knowledge areas, we can capture the emphasis each group of dark web sites place on each of the CyBOK knowledge areas and categories. These weightings will be used to compare any differences in emphasis between these groups of websites as well as explore more generally the areas within cybersecurity that users on the dark web are interested in. By understanding the differences in emphasis between these three sets of websites, the study can establish the area(s) which present the greatest cyber security threat.

In addition, by utilising a shared benchmark, this ensures our study can provide a direct comparison with the mappings created by previous research. Using the knowledge areas from CyBOK as a basis for comparison, this study intends to contrast the formal cyber security curriculum with the informal, "underground" curriculum cyber criminals are consuming on the dark web. This comparison will evaluate whether these qualifications are suitably equipping their graduates against the methodology their adversaries are likely to be learning from content on the dark web.

3 Methodology

Our analysis consists of five steps: a) Data Preparation, b) Creation of a Codebook, c) Creation of a Classifier Program, d) Categorisation of Source Data and e) Data Representation. The following sections document the detailed procedure and analysis for each step.

3.1 Data preparation

The data used in this analysis was sourced, manually checked to ensure it contained relevant items and irrelevant fields were removed.

3.1.1 Source data

The source data analysed in this study has been obtained from three publicly available sources which have been detailed below. Through utilising pre-existing collections, this allowed for more time to be focused on data preparation and analysis instead of collection.

- The market data comprises of 12,354 product listings drawn from the TradeRoute, Hansa, Apple and Silk Road 3 cryptomarkets. These listings were sourced from a larger collection held by the University of Arizona. The only categories selected were targeted for the inclusion of training materials: any product categories referring to guides, eBooks or tutorials, and security practices. The listings include the subject, vendor and price, but not the feedback from buyers or the content of the tutorials themselves.

The forum data comprises of 118,899 posts selected from relevant categories within the following two data sets:

- The first set of data which covers cryptomarket discussion forums has been sourced from the DNM archives [17]. This data set contains 99,118 posts drawn from Abraxas Forums, Silk Road 2 Forums, The Hub Forums, Utopia Forums, Evolution Forums and other cryptomarket forums. The posts date from 2013 to 2015.
- The second set of data that covers the hacking discussion forums has been sourced from the University of Arizona's Artificial Intelligence Lab [12]. This

data set comprises 19,781 posts drawn from the CrackingFire and CrackingArena forums. The CrackingFire data set dates from 7th April 2011 to 21st February 2018 and the CrackingArena data set dates from 8th April 2013 to 24th February 2018.

The forum categories were manually checked to ensure they contained relevant cyber criminal learning material and any categories that did not (such as categories for new user introductions or other off topic discussion) were excluded from the study. In addition to this, any fields that were not relevant to the study, such as price or the post body, were also excluded in order to simplify analysis. A short Python script was written in order to extract the topic titles from the dark web forum data (Appendix B).

Scope and limitations of source data

The post content of the dark web forums were not analysed, only the thread titles. The primary reason for this is it would have been too time consuming for the classifier to sort every post. It is therefore possible that content within a thread does not match the title or for one thread to contain many more posts than another thread.

It should also be noted that, for obvious practical reasons, the data sets do not comprise of every single dark web market and forum. There will be other, more restricted dark web sites where cyber criminal knowledge is being shared and there will also be many dark net websites which are no longer accessible (for example due to seizure by law enforcement). Due to its inherently secretive nature, it is important to state that the lack of observable activities on the dark net does not necessarily translate to an actual lack of such.

However, the source data represents a substantial subsection of the larger and more popular dark web sites that can be reasonably assumed to be broadly representative of the dark web as a whole. As discussed earlier (p. 8), dark web forums and markets are already well documented as arenas for the trade and discussion of cyber criminal material. In addition, our initial study can serve as a basis for wider and more far reaching studies in the future. For example, if content from a previously undiscovered dark web community emerges this too can be analysed to determine its contents from a cyber security perspective and assess any potential threats posed.

3.2 Threats to validity

This section will briefly discuss the various threats to the validity of this study, both internal and external, as well as discuss some of the methods that have been implemented to control for these threats.

The internal validity of our results is dependent primarily on an accurate codebook and classifier program. If the categories defined in the codebook are too vague then our results will lack any meaning. To control for this, after the codebook categories were defined a 10% subset of the manually categorised data set was categorised by an external rater using only the codebook as a guide. A Cohen's kappa value was then calculated to check for agreement between the co-raters. If a suitable value was not achieved, the process was repeated until a Cohen's kappa value indicating substantial agreement was achieved. This will ensure the codebook categories

Since many of the items will be automatically assigned a category by the classifier, as opposed to by a human rater, it is imperative that the classifier is functioning as intended. To control for this, the classifier will undergo extensive cross validation to confirm it is assigning categories to items correctly. A Cohen's kappa value was calculated to ensure the classifier able to achieve substantial agreement with a human rater assigning the same items.

Dark web forums and markets contain a large variety of categories, many of which are dedicated to content that is not of interest to this study. If these categories were included in the analysis, this would bias the results as it would reflect content that is not learning material. To prevent a large amount of irrelevant data being analysed, each category were manually checked to ensure they contained relevant cyber criminal learning material. Those that did not, were excluded from the study. This was typically done by checking the title of the category to determine its most likely contents. In some cases where this was not entirely clear, a brief scan of the the topics within this category was then conducted to determine whether any relevant learning material was present.

The analysed data consists only of cryptomarket and forum categories that relate specifically relate to guides and learning material. In addition, it is somewhat dated with the most recent forum titles dating from 2018 and market listings from 2015. Any developments in the dark web content from this point onwards will not be reflected in our findings. We therefore have to be careful when generalising our findings beyond these specific areas and time periods.

3.3 Creation of a Codebook

The creation of a codebook was required in order to firstly, define the most prevalent topics within the ground truth data and secondly, to provide distinct categories for the classifier to sort items into. Furthermore, being able to sort items into categories that are separate to the broader knowledge areas, has the advantage of providing an extra layer of detail that will allow the study to determine which specific areas feature most prominently within the source data.

3.3.1 Conception

The aim of the codebook was to define distinct categories that could then be used to sort the dark web learning material onto CyBOK knowledge areas. Initially the study began by manually categorising market listings directly to CyBOK knowledge areas however this approach was soon abandoned as we were unable to achieve significant agreement between co-raters as many of the items had characteristics which crossed multiple knowledge areas.

Additionally, it became clear the thirteen knowledge areas were not accurately reflecting the nature or the range of content available within the learning material. For example, there were a significant percentage of listings relating to drugs and e-books that we believed could be relevant and should be covered by the codebook. A significant percentage of the items were also being categorised simply as "Adversarial Behaviours" but within that category, the items included a large variety of very different behaviours that were not being represented. Secondly, many items related to content that could be placed in multiple knowledge areas, making it difficult to achieve agreement between co-raters.

In order to create this codebook, a substantial portion (over 2,000 cryptomarket listings) were manually categorised and sorted into the most common categories. Once these categories were defined and agreed, they were then mapped to the most relevant CyBOK knowledge area.

3.3.2 The Codebook

The following are the thirty five categories defined within the codebook as well as a few examples of items that could be placed in each category.

1. **Anonymity – Other**
   Item relating to a guide, software or service designed with the intention of increasing anonymity in some respect but unrelated to any other anonymity subcategory.

   Example items: voice changing software, anonymous email account service or privacy oriented operating systems such as TAILs [28] or Whonix [24].

2. **Anonymity – Tor**
   Item relating to the Tor (The Onion Router) [66].

   Example items: a guide on how to setup the Tor browser or discussion around Tor security features.

3. **Anonymity – VPNs**
   Item relating to virtual private networks.

   Example items: selling of VPN accounts or guides on how to set up a VPN.

4. **Anonymity – Proxies**
   Item relating to proxy servers.

   Example items: a guide on how to set up a proxy server or selling of a subscription to a proxy service.

5. **Carding**
   Item relating to the theft of bank cards (colloquially known as carding) or the use of such cards. This is inclusive of tools or guides that are highly likely to be used to aid these activities such as CVV checkers or information on how a banks credit card system operates.

Example items: credit card dumps, a list of websites vulnerable to credit card fraud or tutorials on how to learn carding.

6. **Cashing Out**
   Item relating to the conversion of currency either between platforms (for example, from credit card to Paypal), from one form of currency to another (for example, Paypal to Bitcoin) or other methods for concealing the origins of money obtained illegally.

   Example items: discussion relating to cryptocurrency mixing services such as Bitcoin fog or relating to the most effective money laundering techniques.

7. **Clearing Criminal History**
   Item relating to the removal or modification of criminal records.

   Example items: services dedicated to the removal of names from criminal registers or the removal of mugshots from police databases.

8. **Counterfeit Currency**
   Item relating to the creation or use of fake currency.

   Example items: a guide on how to create fake US dollars or how to pass fake currency in-store.

9. **Cryptocurrency – General**
   Item relating to cryptocurrencies but otherwise unrelated to both trading, cashing out or fraud. This category is also inclusive of other blockchain technologies.

   Example items: a discussion on the best websites to purchase cryptocurrency, the process of sending cryptocurrency or technical analysis of cryptocurrency.

10. **Cryptocurrency – Trading**
    Item relating to the trading of cryptocurrencies.

    Example items: a discussion of cryptocurrency trading signals or other cryptocurrency trading strategies.

11. **Denial of Service**
    Software, scripts, guides or other methods relating to denial of service attacks.

    Example items: a guide on how to implement a successful denial of service attack or a list of tools that can be used to aid denial of service attacks.

12. **Digital Forensics**
    Item relating to the the recovery and investigation of material found in digital devices. This includes methods of or software designed to circumvent forensic techniques.

    Example items: a guide on how to erase a hard-drive, recover deleted files.

13. **Doxing**
    Item relating to the publishing of private or identifying information about a particular individual on the Internet.

    Example items: a guide on how to effectively dox a targeted individual.

14. **Drugs – Production**
    Item related to the production or refinement of drugs.

    Example items: a guide on how to synthesise MDMA, brew beer or grow cannabis plants.

15. **Drugs – General**
    Item related to to drugs but unrelated to any other sub-category.

    Example items: books on drug culture, discussion around strains of cannabis or a guide on how to use certain drug paraphernalia.

16. **eBooks – Other**
    An electronic version of a printed book (that can be read on a computer or handheld device) which unrelated to any other category. This category is also inclusive of audio books.

    Example items: PDF or EPUB files of popular novels, books on political theory or cookbooks.

17. **eBooks – Technical**
    An electronic version of a printed book (that can be read on a computer or handheld device) which is related to an aspect of computer science. This category is also inclusive of audiobooks.

    Example items: PDF or EPUB files of popular novels, books on political theory or cookbooks.

18. **eWhoring**
    Item relating to the practice of e-whoring (posing as an attractive woman with the intention of extracting money).

    Example items: autopilot bots or image sets intended to be used for e-whoring purposes.

19. **Fraud**
    Item relating to wrongful or criminal deception intended to result in personal or financial gain. This category is not inclusive of credit card fraud which is covered within the "Carding" category.

    Example items: guides on how to defraud a company or discussion relating to phishing techniques.

20. **Hacking – General**
    Item relating to hacking but unrelated to any other hacking subcategory.

    Example items: discussion relating to general hacking techniques.

21. **Hacking – Website**
    Item relating to the hacking of websites or web applications.

    Example items: a guide on how to hack accounts on a particular website or tools that aid website hacking such as SQL injection tools.

22. **Hacking – Mobile**
    Item relating to hacking specifically related to mobile devices or mobile operating systems such as Android or iOS.

    Example items: a guide on how to jailbreak an iPhone or root an Android smartphone.

23. **Hacking – Wireless Networks**
    Item relating to hacking involving wireless networks.

    Example items: a guide on how to or software intended to be used to crack wireless network passwords.

24. **Hacking – Phreaking**
    Item relating to the hacking of telecommunications systems. This also includes items relating to burner phones.

    Example items: a guide on how to obtain a burner phone or how to obtain free calls.

25. **Hacking – Malware Supply Chain**
    Item relating to malware delivery methods.

    Example items: discussion relating to methods for spreading a virus or the most reliable hosting websites to deliver viruses.

26. **Lock picking**
    Item relating to the practice of picking locks or other methods of bypassing door locks.

    Example items: selling of lock picking tools, guides on how to pick locks or methods for bypassing biometric locks.

27. **Malware Authorship**
    Item relating to the creation of malware.

    Example items: virus building software or guides on how to create an effective virus.

28. **Modifying Credit**
    Item relating to the modification of a individuals credit report or credit score as well as the setting up of credit tradelines.

    Example items: services offering to modify credit scores or secure trade lines.

29. **Resources – Contact Lists**
    Lists of email addresses, phone numbers, usernames & passwords, websites or other contact details and unrelated to any other category.

    Example items: lists of email addresses and password combinations (known colloquially as combo lists) or lists of darknet websites.

30. **Resources – Identity Documents**
    Item relating to the use or creation of identity documents.

    Example items: a guide on how to create fake passports, identity cards or drivers licences.

31. **SEO**
    Item relating to the modification of results from online search engines.

    Example items: a guide on how to increase a websites ranking on Google search or steal a competitors traffic.

32. **PGP/GPG**
    Item relating to the PGP/GPG cryptographic software. This category was extended to include other encryption software such as Veracrypt or Truecrypt.

    Example items: a guide on how to use PGP or discussion around the effectiveness of Veracrypt.

33. **Transportation/Stealth**
    Item relating to the smuggling of goods.

    Example items: a guide on how to conceal drugs to avoid detection by law enforcement or discussion on the effectiveness of a vendors concealment techniques.

34. **Weaponry Explosives**
    Item relating to the manufacture or sale of firearms, explosive devices or chemicals.

    Example items: a guide on how to 3D print a firearm or manufacture a pipe bomb.

35. **Other**
    Item is unrelated to any other category.

Example items: discussion relating to internal matters within the forum such as the nomination of moderators, forum rules or new users introducing themselves.

3.3.3 Validation of the codebook

A randomly selected ten percent sample of the sorted items was then manually categorised by a co-rater, one of my project supervisors Joseph Hallett, using only the codebook as a guide, and a Cohen's kappa value was calculated to ensure inter-rater reliability. A short Python script was written to calculate this value (Appendix A). This process was repeated with the codebook being tweaked each time until an acceptable Cohen's kappa value was achieved.

The Cohen's kappa coefficient [25] is a popular descriptive statistic for summarizing an agreement table [77]. Cohen suggested that the kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as indicating near perfect agreement [52].

The final Cohen's kappa value achieved between co-raters was 0.92 indicating near perfect agreement. This was calculated using a Python script and the sklearn library [Appendix A]. This codebook was then utilised as the basis for a classifier that performs keyword analysis on a string item in order to assign it to the most appropriate category from within the codebook.

3.4 Creation of a Classifier

A classifier program was constructed in order to categorise the remaining uncategorised source data. The following section will outline the steps taken during the design and implementation of the classifier.

3.4.1 Motivation

Manually categorising the remaining items would not have been possible within the required timescales. As a result, it was decided that the creation of a classifier program to automate this process would be the most effective means of categorising the remaining source data. In addition, this also meant the project would output a program that could be re-purposed for future research.

Programming language

The chosen programming language for the classifier was Python as it allowed for the importing of ready made libraries such as pandas for data manipulation and sklearn for calculating Cohen's kappa.

Classifier algorithm

One problem faced by the classifier is that some items are ambiguous even when assigned by human coders. For example, by containing words or a word that could potentially relate to multiple topics.

The classifier intends to control for this by implementing a keyword analysis algorithm to essentially determine the most relevant words in the item and which category they relate to. This will involve breaking each string item into separate tokens before applying an algorithm to each token to determine its relevance to each category. Keyword analysis will be performed using the techniques outlined in the article by Rayson & Garside [63] which has been conclusively shown to discover key items in the corpora which differentiate from one corpus from another. This involves firstly the creation of a word frequency list for each corpus and then the creation of a contingency table, such as the example below.

	Corpus One	**Corpus Two**	**Total**
Freq of word	a	b	a+b
Freq of other words	c-a	d-b	c+d-a-b
Totals	c	d	c+d

Table 3.1: Sample contingency table for word frequencies

In the example, note that the value *c* corresponds to the number of words in corpus one, and *d* corresponds to the number of words in corpus two (N values). The values *a* and *b* are called the observed values (O). The expected values (E) are then calculated according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Once the expected values are calculated, the log likelihood value is calculated according to the following formula:

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right)$$

Using this methodology, a log-likelihood value for each token is calculated via a contingency table. The word frequency list is then sorted by the resulting LL values. This gives the effect of placing the largest LL value at the top of the list representing the word which has the most significant relative frequency difference between the corpora. In this way, we can see the tokens that are most indicative (or characteristic) of one corpus, as compared to any other corpus, at the top of the list.

The Rayson & Garside article concludes that a fully automated approach is not recommended and that further checking of these keywords is required in order to understand the reasoning behind their significance and potential explanations sought for the patterns displayed. Based on this conclusion, manual checks of these log likelihood lists will be performed and amendments will be made to the classifier in instances when it is detected that the log likelihood calculations are causing items to be miscategorised.

This methodology has been used to good effect in numerous similar studies [42, 54]. One of these studies analysed the messages on the micro blogging website Twitter [42]. This study used the log likelihood ratio to analyse these messages to determine the most distinctive words for each day of the week. Another study which implemented this methodology intended to classify blog text according to the mood reported by its author during the writing process [54]. As evidence of its validity, this study found that the classification accuracy was not substantially worse than human performance on the same task.

In conclusion, this method is appropriate for our research firstly, as it is relatively simple to implement with the tools available in the required timescales and secondly, it has a basis in similar research where it has been demonstrated to be effective.

3.4.2 Training the Classifier

Before the classifier can be used to categorise items, it must be trained on a sorted data set in order to weight the keywords for each category.

The training of the classifier comprises of five key steps:

1. The ground truth data set – consisting of all the manually categorised items so far – is extracted into thirty five separate CSV files each containing all the keywords for a particular category (Appendix C).

2. The full ground truth data set is tokenised and the frequency of each token within that corpus is calculated (Appendix D). The text is amended to lowercase and non-alpha numeric characters are removed just prior to this stage and replaced with white space.

3. The next stage is the calculation of the log likelihood values for each token in each corpus (Appendix E, I & J). Significance is also calculated (Appendix F) for each log likelihood value and added to the *sig* column. The thresholds for significance are taken from the aforementioned website.

4. Any non-significant ($p > 0.01$) keywords are now excluded from the model (Appendix F & G). The primary aim of this is to make the next stage faster and to prevent any non-significant keywords from being incorrectly factored in to the calculation of a log likelihood score.

Once the classifier is trained, it is primed and does not need to be trained again unless the ground truth data set is modified.

3.4.3 Using the Classifier

Once the classifier is trained, it can then be used to categorise string items contained in any properly formatted text source.

1. Initially the classifier checks for the presence of certain key phrases and assigns them to the appropriate category (Appendix H). These are typically phrases that include a token or tokens that are at high risk incorrectly categorising the item.

   For example, "no carding" contains the token "carding" which has a very high log likelihood score for the "Carding" category even though the phrase expressly indicates the item should not be categorised as such. Typically the phrase instead indicates that an item is a guide on how to acquire a good or service via fraudulent means and therefore should be categorised as 'Fraud'.

2. For each token, the classifier creates an array containing the log-likelihood score of that token for each codebook category (Appendix J).

3. These scores are then summed for each category and returned in a final array. The category with the highest total score is then assigned to the item, provided it is over a threshold value. The threshold value utilised in this study was 15 indicating a total log likelihood value that was significant ($p > 0.0001$) to at least one category. A lower threshold value ($p > 0.001$) was also tested in the process of building the classifier but this resulted in a high number of items being miscategorised and was therefore increased. If the highest score in the array is below this threshold value, that is contains no relevant keywords for any category, the item is assigned to "Other" as the default category.

3.5 Validation of Classifier

As the bulk of the source data will be automatically assigned by the classifier, it is essential that it assigns categories accurately. Validation techniques were therefore utilised to ensure the constructed classifier was accurately assigning categories to items to a level of at least substantial agreement with a human rater.

3.5.1 K-fold cross validation

The primary method of validation utilised was k-fold cross validation [76]. The k-fold cross validation technique is one of the most used approaches by practitioners for model selection and error estimation of classifiers [11].

In k-fold cross-validation, the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k - 1$ folds are used for learning [64]. Rule-of-thumb methods suggest to fix large values of k (5, 10 or 20), since it is usually preferable to exploit a larger number of patterns for training purposes [11], therefore a k value of 10 was selected for this cross-validation.

This method of cross-validation was selected for a number of reasons. Firstly, the data set was appropriate: both the training set and the validation set are drawn from the same population. There were also advantages over other methods of cross-validation. One such advantage of this method over repeated random sub-sampling is that it ensures all observations are used for both training and validation, and each observation is used for validation exactly once.

Computationally it is not as expensive as leave-one-out cross-validation. K-fold cross-validation needs to only run k times which is usually 5 or 10 whereas leave-one-out cross-validation needs to be be run n times, where n is the number of items in the data set. This would have been extremely time consuming as the data set is large so performing leave-one-out cross-validation would not have been practical.

In addition to this, there are also significant generalisation issues with leave-one-out cross-validation. As 99% of the training data remains the same each time, the model becomes over-fit on the training data and the outcomes produced are more-or-less the same each time. Using k-fold cross validation instead, this provides a better understanding of how effectively the model will categorise items that are not part of the training data.

In summary, k-fold cross validation was appropriate for the input data set, practical to implement with the available tools and within the required timescale. A separate Python program was constructed in order to automate the cross-validation process and integrated many of the functions previously created for the classifier (Appendix J & K). If the total number of items was not an exact multiple of ten then the sample was rounded down to the nearest multiple of ten before being split. This occasionally would result in a very small number of items (less than 10) being excluded from the cross-validation process.

After each k sample has been categorised by the classifier, Cohen's kappa is then calculated for each sample and then averaged to produce the Cohen's kappa value used for validation. This validation process was then repeated, making changes to the model and ground truth data each time (see optimisation), until a Cohen's kappa value above the benchmark was obtained. Once validated, the full ground truth data set was used as the training data for the final classifier.

For reference, the Cohen's kappa value initially achieved between human raters, obtained in the process of validating the codebook (myself and a supervisor) was 0.92. The decided benchmark for the automated model was a Cohen's kappa value of 0.70 which would qualify on the upper end of 'substantial agreement' [52]. After achieving the benchmark Cohen's kappa score via k-fold cross-validation, the classifier is trained on the full ground truth data set. Cohen's kappa is again calculated and evaluated to ensure agreement.

3.5.2 Manual validation checks

While it would have been excessively time consuming to cross-check the automated assignment of every item manually, I did perform some broader manual checks. This involved calculating the discrepancies between the total number of items assigned to each category manually and by the classifier (see Table 4.2). These numbers were then compared in order to identify any categories that were being over or under assigned.

The categories with the largest discrepancies were those that tended to overlap with other categories ("Carding", "Cashing Out") or those in which the items typically contained a lot of variation without any particularly distinct keywords ("eBooks -

Other", "Other"). Items in these categories were then manually scanned to identify similarities which could lead to the root cause of this discrepancy if any was present. This root cause was typically an over or under weighted keyword and was remedied by either adding further relevant items containing this keyword to the ground truth data or by adding to the list of keywords and phrases that are detected by the classifier prior to keyword analysis.

For example, many of the thread titles related to bitcoin mixing (also known as tumbling) were erroneously being categorised as "Cryptocurrency – General". I therefore added a few relevant phrases (such as "bitcoin mixing") in the initial checking stages to ensure these items were categorised correctly as "Cashing Out". Less frequently, this was caused by items in the ground truth data set that had been incorrectly manually categorised. This was typically either due to a word, such as a type of software, that I was not aware of at the time or simply manual error. Once these were detected, they were manually corrected to prevent any such error reoccurring.

4 Results and Analysis

4.1 Results

4.1.1 Cross validation results

The final results of the k-fold cross validation are included in the table below.

K	Cohen's kappa
1	0.76
2	0.67
3	0.72
4	0.74
5	0.71
6	0.76
7	0.72
8	0.72
9	0.71
10	0.75
Mean	0.73

Table 4.1: Cohen's kappa agreement for each k model

The model returned a mean average Cohen's kappa value of 0.73 indicating the classifier was able to achieve substantial agreement with a human rater [52]. For context, the level of agreement between two human raters using the codebook was 0.92 indicating near perfect agreement. So although the classifier was able to achieve substantial agreement, significant improvements would need to be made in order to be as effective as a human rater.

4.1.2 Over or under assigned categories

After k-fold cross validation, the full manually categorised ground truth data set was sorted by the classifier. The table shows the total number of over or under assigned items per category and per data type. For example, the most over assigned category (at this stage) was "Cashing Out" as the classifier assigned 210 more items to this category than the human rater.

Category	Titles	Listings	+/-
Cashing Out	90	120	210
eBooks – Technical	62	120	182
Anonymity – VPN	28	12	40
Cryptocurrency – General	-15	51	36
Anonymity – Tor	21	13	34
Counterfeit Currency	8	19	27
Hacking – General	-1	18	17
Cryptocurrency – Trading	3	13	16
Transportation/Stealth	-14	26	12
eWhoring	0	7	7
Modifying Credit	5	1	6
Doxing	-1	6	5
Hacking – Phreaking	6	-1	5
SEO	0	4	4
Digital Forensics	1	2	3
Fraud	-12	14	2
Drugs – Production	14	-14	0
Resources – Identity Documents	2	-2	0
Denial of Service	0	-1	-1
Hacking – Wireless Networks	-1	-1	-2
Malware Authorship	0	-2	-2
Clearing Criminal History	1	-4	-3
Lockpicking	1	-4	-3
Hacking – Mobile	-1	-4	-5
Hacking – Malware Supply Chain	-1	-5	-6
Resources – Contact Lists	1	-8	-7
Weaponry & Explosives	-2	-6	-8
PGP/GPG	-16	7	-9
Hacking – Website	-3	-16	-19
Drugs – General	-12	-16	-28
Anonymity – Proxies	-27	-11	-38
Anonymity – Other	-26	-34	-60
Carding	-24	-66	-90
Other	-74	-22	-96
eBooks – Other	-4	-216	-220

Table 4.2: Table containing number of items over or under assigned by the classifier

For context, the ground truth data set contained 5,486 items in total. As discussed on page 26, steps were taken to remedy this prior to analysis of the remaining uncategorised items.

4.1.3 Raw results by category

The remaining uncategorised items were sorted by the classifier and merged with the manually categorised items. The table below contains the number of items assigned to each category. The table is broken down between the hacker forum topics, cryptomarket forum topics and the cryptomarket listings.

Category	Hacker	Crypto	Market	Total
Other	588	600	727	**1915**
eBooks – Technical	125	82	1639	**1846**
Cashing Out	72	179	888	**1139**
Carding	78	57	899	**1034**
Fraud	131	47	524	**702**
Anonymity – Proxies	418	47	63	**528**
Hacking – Website	385	6	104	**495**
eBooks – Other	2	4	459	**465**
Drugs – Production	0	23	354	**377**
Hacking – General	173	18	181	**372**
Anonymity – Other	36	119	200	**355**
Resources – Contact Lists	298	12	22	**332**
Transportation/Stealth	20	190	90	**300**
Cryptocurrency – General	25	71	201	**297**
Anonymity – VPN	55	70	161	**286**
Drugs – General	3	38	227	**268**
Anonymity – Tor	15	99	70	**184**
PGP/GPG	17	120	36	**173**
Weaponry & Explosives	0	9	119	**128**
Resources – Identity Documents	45	22	53	**120**
Digital Forensics	3	13	101	**117**
Counterfeit Currency	2	10	93	**105**
Hacking – Wireless Networks	32	14	47	**93**
SEO	19	6	55	**80**
Cryptocurrency – Trading	3	4	69	**76**
Modifying Credit	1	5	68	**74**
Hacking – Phreaking	7	23	22	**52**
Malware Authorship	20	3	28	**51**
eWhoring	11	2	34	**47**
Hacking – Mobile	10	0	30	**40**
Doxing	2	3	24	**29**
Denial of Service	9	2	11	**22**
Hacking – Malware Supply Chain	2	2	17	**21**
Lockpicking	0	1	17	**18**
Clearing Criminal History	0	0	9	**9**

Table 4.3: The total number of items in each category for each data set

4.2 Data representation

4.2.1 Results by category

Bar charts were created to summarise the results by codebook category for each data type and overall.

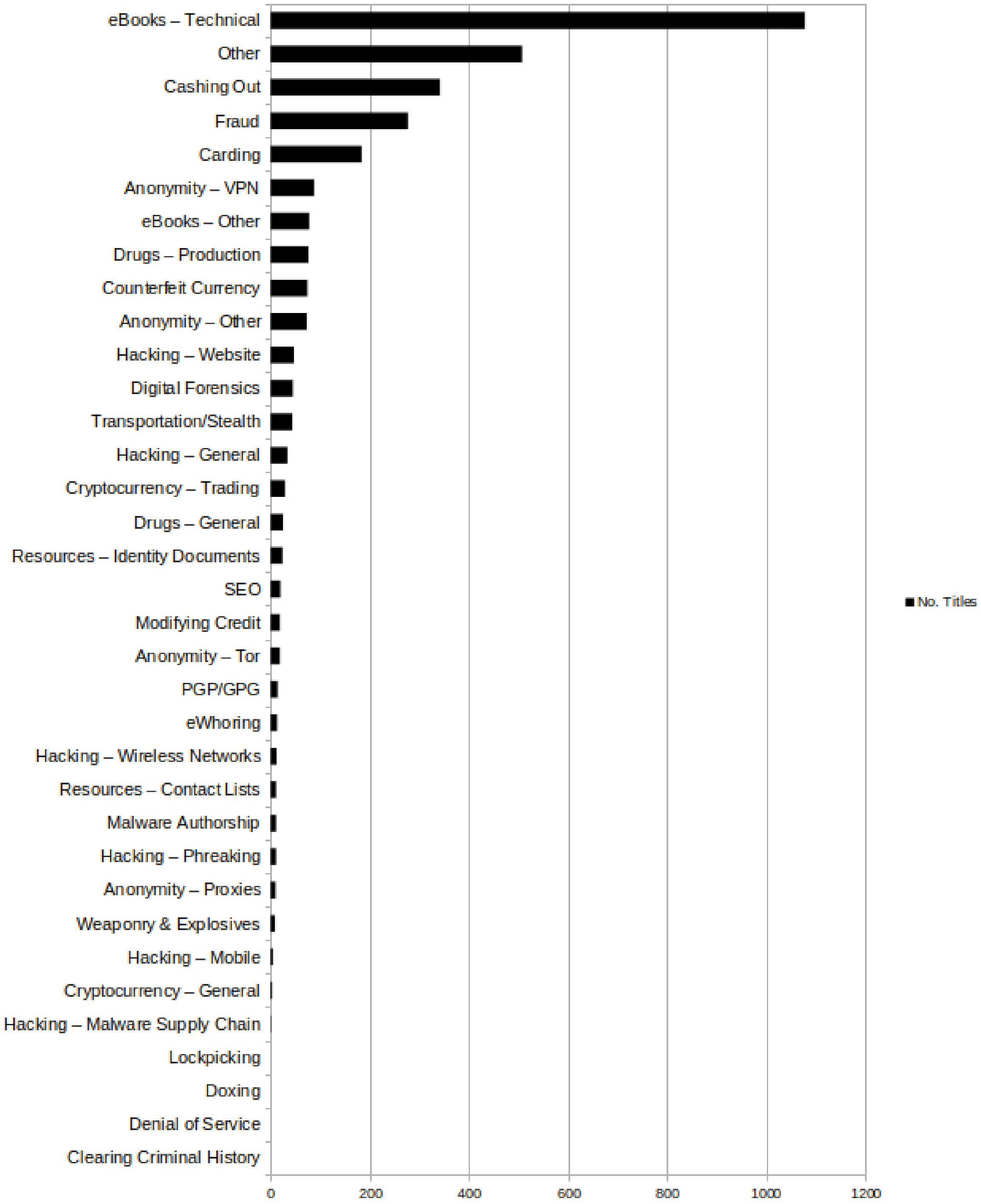


Figure 4.1: Bar chart summarising market listings by category

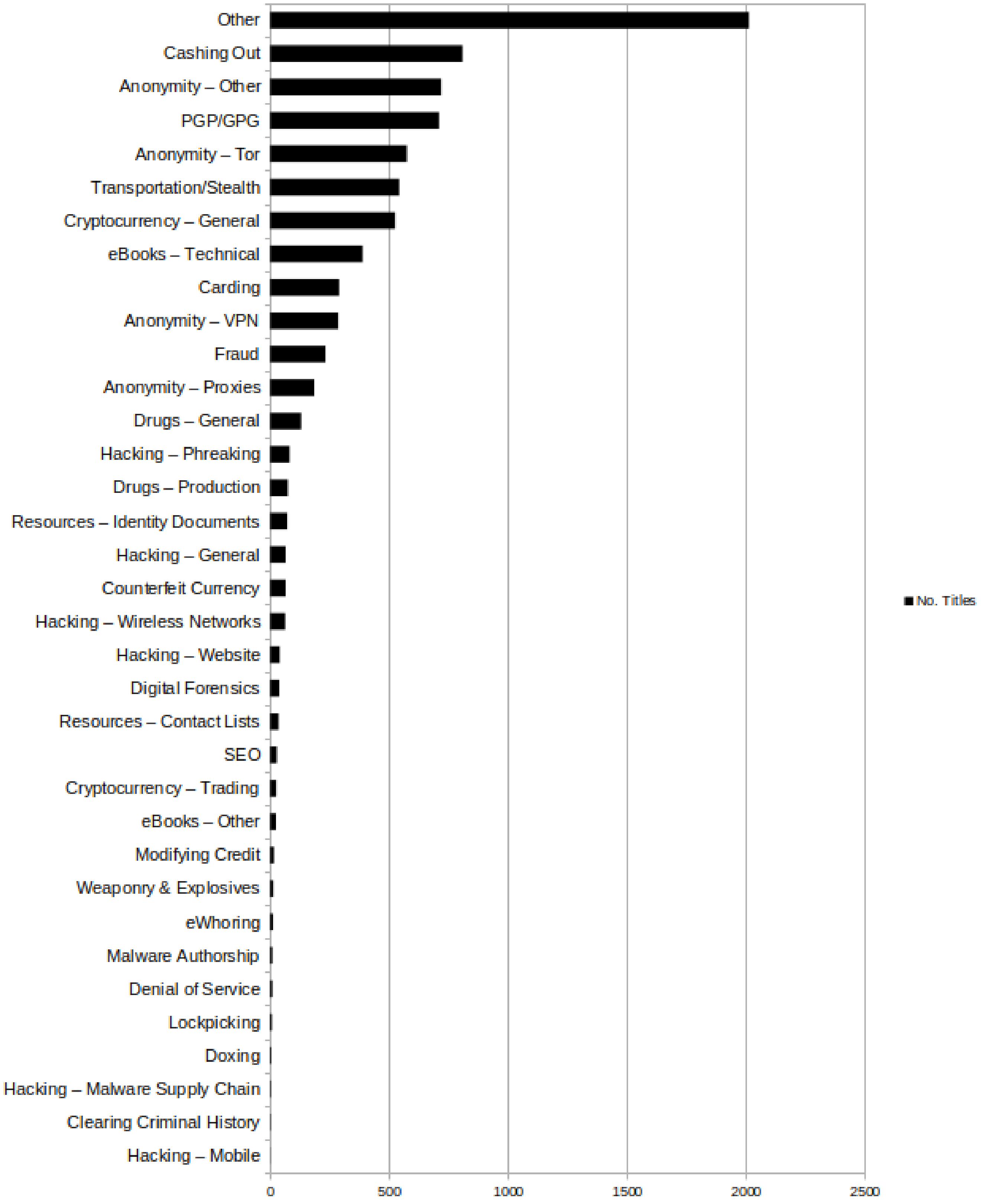


Figure 4.2: Bar chart summarising cryptomarket forum titles by category

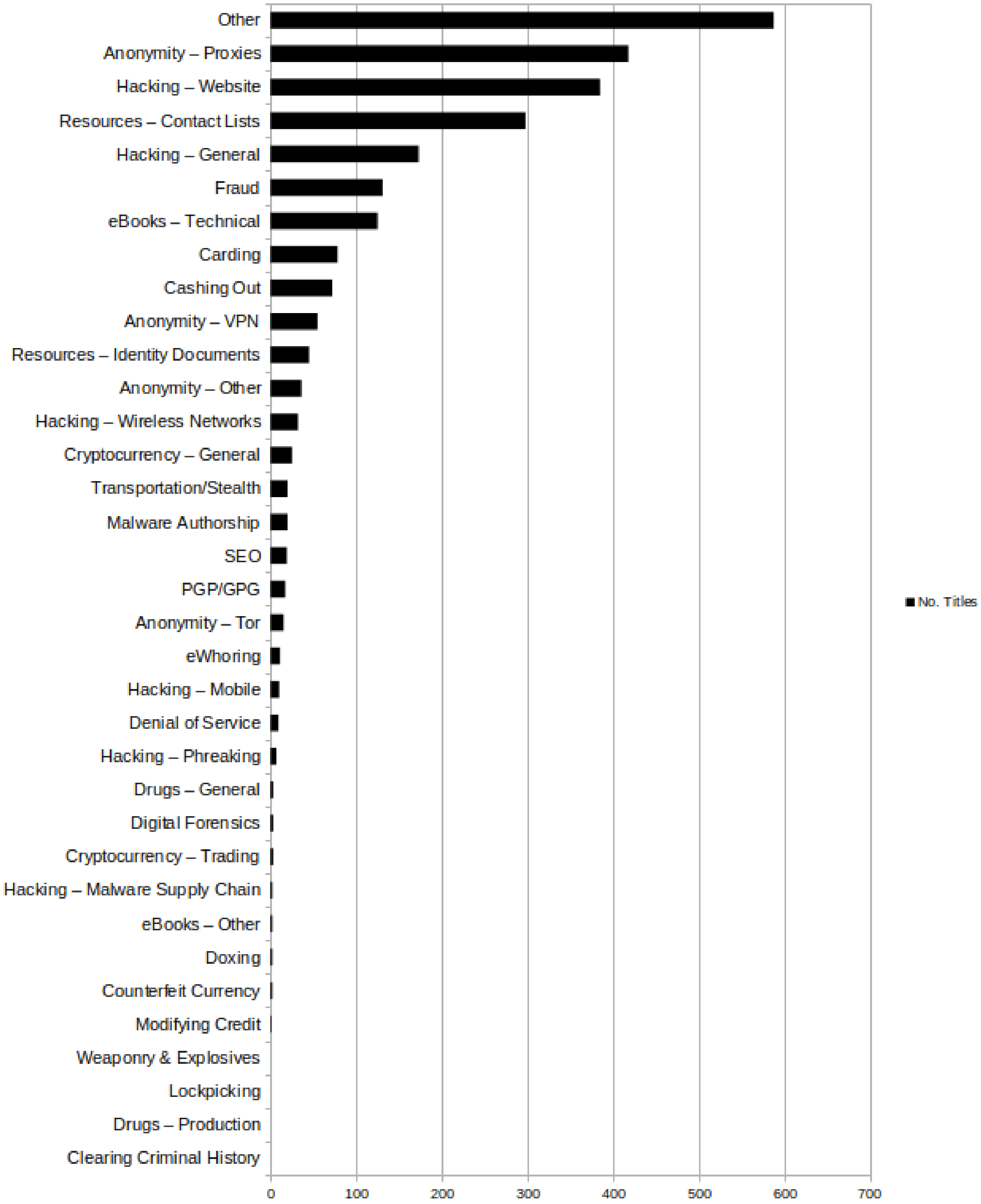


Figure 4.3: Bar chart summarising hacker forum titles by category

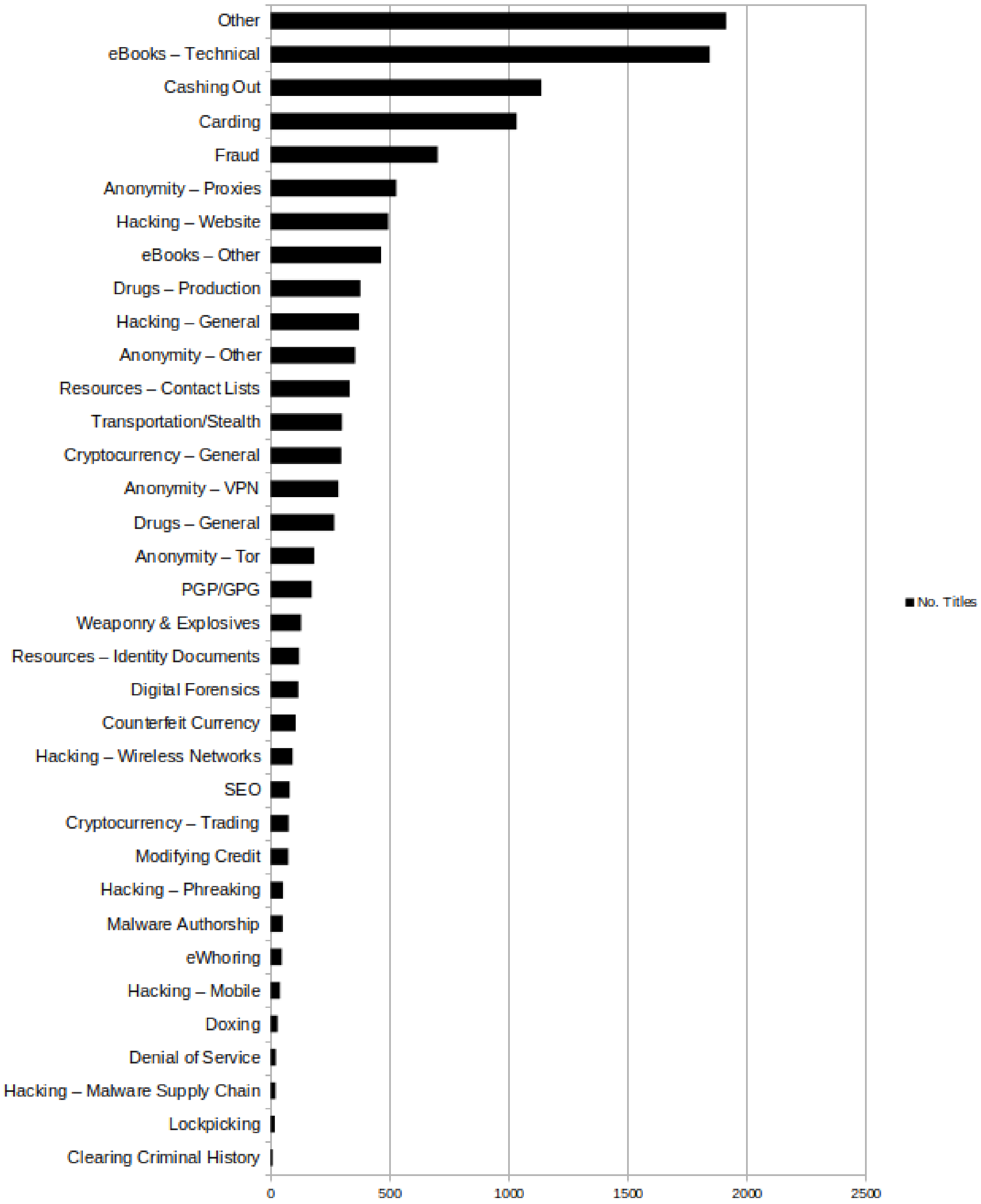


Figure 4.4: Bar chart summarising the total number of items by category

4.2.2 Linking with CyBOK knowledge areas

Each category was assigned to the most relevant CyBOK knowledge area (see Appendix A). This was done with the supervision of my project supervisors and any contentious categories were discussed. It was found some categories overlapped many knowledge areas. For example, “Anonymity - VPNs” contained elements found both in "Privacy & Online Rights" and within "Network Security". Categories that were not related to cyber security (such as "Drugs" and "Other") were not included in this list and were not assigned a knowledge area. The final assignment of codebook categories to knowledge areas and their broader CyBOK category is detailed in the table below.

CyBOK – Category	**CyBOK – Knowledge Area**	**Codebook – Category**
Attacks & Defences	Adversarial Behaviour	Carding
Attacks & Defences	Adversarial Behaviour	Cashing Out
Attacks & Defences	Adversarial Behaviour	Clearing Criminal History
Attacks & Defences	Adversarial Behaviour	Counterfeit Currency
Attacks & Defences	Adversarial Behaviour	Denial of Service
Attacks & Defences	Adversarial Behaviour	Doxing
Attacks & Defences	Adversarial Behaviour	eWhoring
Attacks & Defences	Adversarial Behaviour	Fraud
Attacks & Defences	Adversarial Behaviour	Hacking – General
Attacks & Defences	Adversarial Behaviour	Hacking – Malware Supply Chain
Attacks & Defences	Adversarial Behaviour	Modifying Credit
Attacks & Defences	Adversarial Behaviour	Resources – Contact Lists
Attacks & Defences	Adversarial Behaviour	Resources – Identity Documents
Attacks & Defences	Adversarial Behaviour	SEO
Systems Security	Cryptography	Cryptocurrency – General
Systems Security	Cryptography	Cryptocurrency – Trading
Systems Security	Cryptography	PGP/GPG
Attacks & Defences	Forensics	Digital Forensics
Attacks & Defences	Malware & Attack Technologies	Malware Authorship
Infrastructure Security	Network Security	Hacking – Wireless Networks
Infrastructure Security	Physical Layer Security	Hacking – Phreaking
Infrastructure Security	Physical Security	Lockpicking
Infrastructure Security	Physical Security	Weaponry & Explosives
Human, Organisational & Regulatory Aspects	Privacy & Online Rights	Anonymity – Other
Human, Organisational & Regulatory Aspects	Privacy & Online Rights	Anonymity – Proxies
Human, Organisational & Regulatory Aspects	Privacy & Online Rights	Anonymity – Tor
Human, Organisational & Regulatory Aspects	Privacy & Online Rights	Anonymity – VPN
Software & Platform Security	Web and Mobile Security	Hacking – Mobile
Software & Platform Security	Web and Mobile Security	Hacking – Website
N/A	N/A	Drugs – General
N/A	N/A	Drugs – Production
N/A	N/A	eBooks – Other
N/A	N/A	eBooks – Technical
N/A	N/A	Other
N/A	N/A	Transportation/Stealth

Table 4.4: Codebook categories and their respective assignment to CyBOK knowledge areas

4.2.3 Spider maps

Spider maps were created to summarise the results for each broader CyBOK knowledge area. These were created in order to provide a direct comparison to the spider maps created in the Hallett et al. study [40].

The spider maps were produced using Postscript code written by one of my supervisors (Joseph Hallett) with the inputs and settings customised. This was appropriate as it was the same code used to generate the graphical representations for the Hallett et al. study. The output PDF files containing the graphs were then trimmed using online tools to remove unnecessary white space and converted into image files.

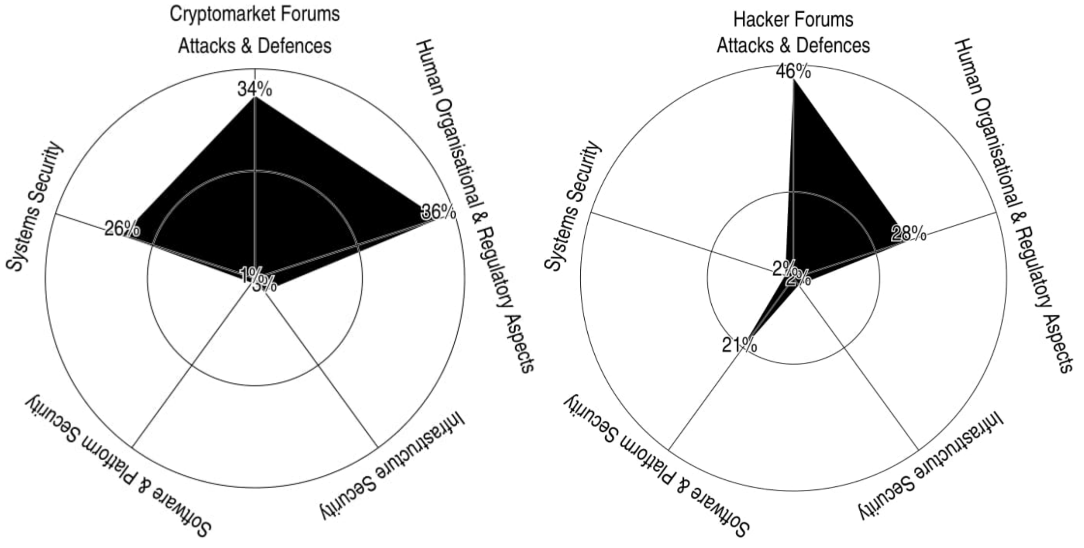


Figure 4.5: Knowledge area spider charts for the cryptomarket forum titles and hacker forum titles

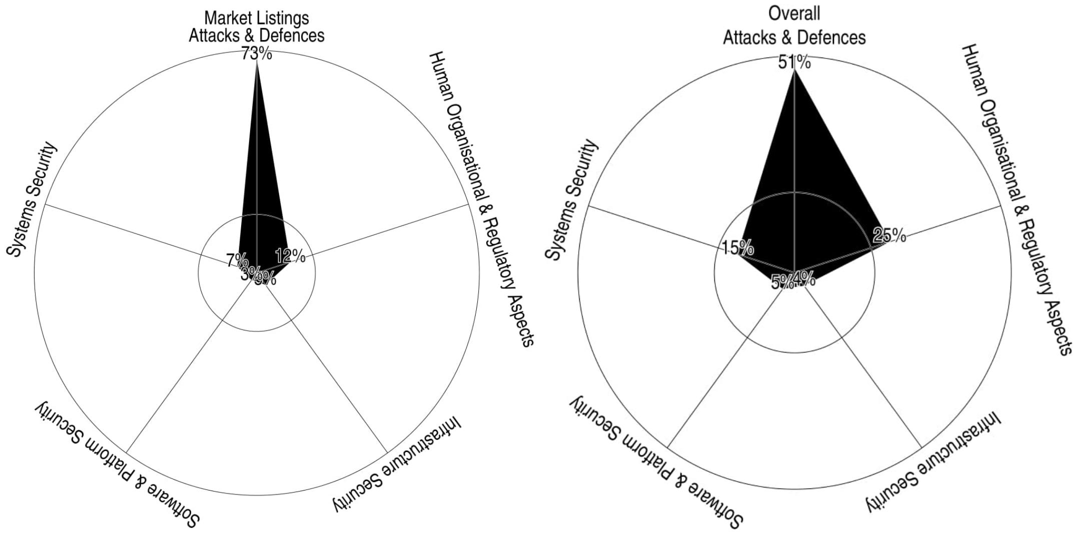


Figure 4.6: Knowledge area spider charts for market listings and all data sets combined

4.2.4 Bar charts

Bar charts were created to summarise the results by individual CyBOK knowledge area.

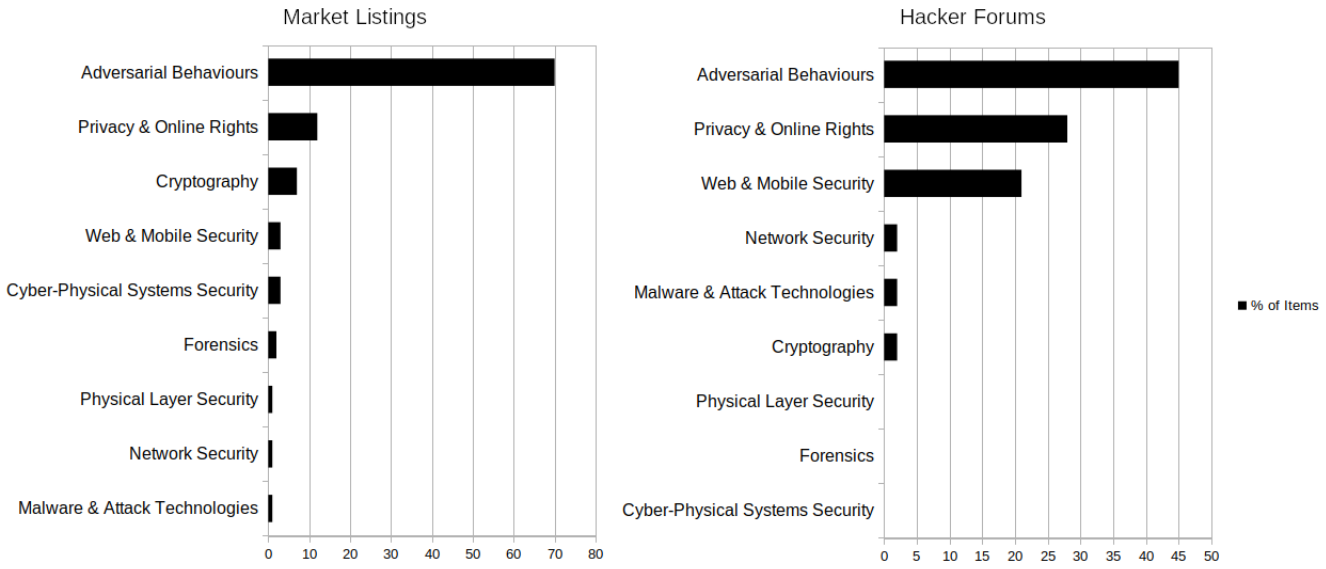


Figure 4.7: Percentage of items by knowledge area for market listings and hacker forums

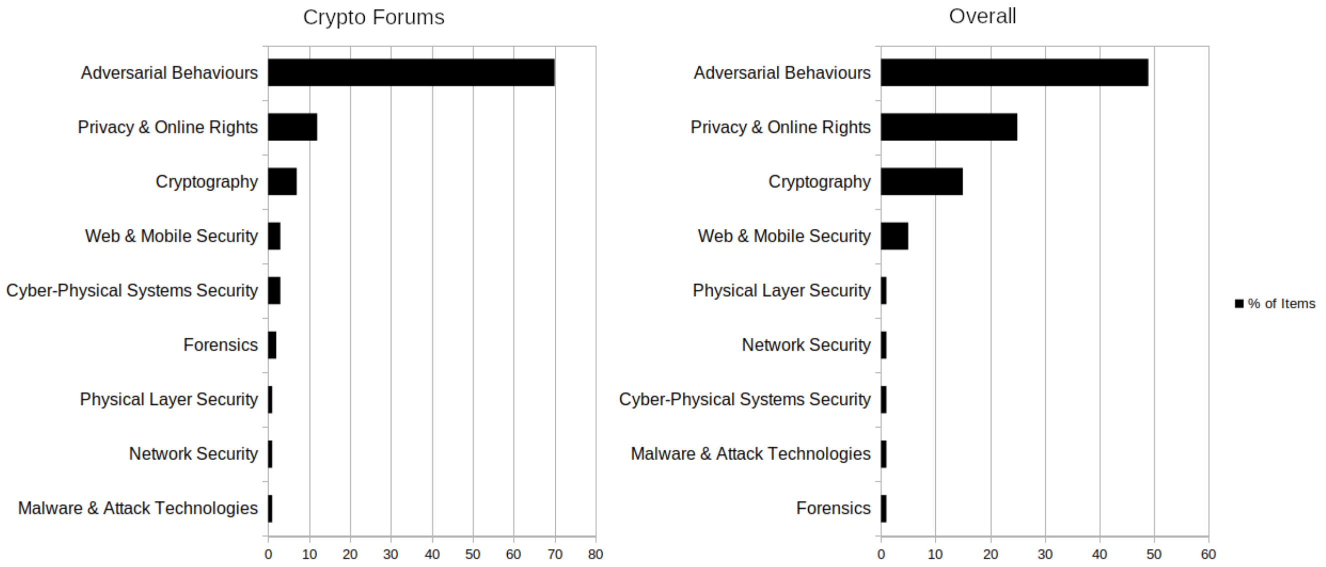


Figure 4.8: Percentage of items by knowledge area for cryptomarket forums and overall

4.3 Discussion

In this section, the implications of our results will be discussed. Firstly, these findings will be discussed in the context of the related literature on the dark web. Secondly, the findings will compared against research into cyber security curriculums in order to evaluate the current state of cyber security qualifications. Lastly, I will discuss suggestions for both policy and cyber security education based on our findings from this study.

4.3.1 Research questions

Most prominent knowledge area

While the study did not test a fixed hypothesis, it was speculated in the research proposal that "Adversarial Behaviour" would be the most prominent knowledge area and this did prove to be correct. This knowledge area alone accounted for almost 50% of cyber security related items. The basis of this prediction was a study that indicated a significant percentage of illicit content on the dark web was dedicated to credit card fraud and money laundering [72]. Our findings concurred with this as categories relating to carding and money laundering were some of the most prominent, particularly within the market listings data set.

Scope of CyBOK

In the research proposal we set out to determine whether there was cyber security related learning material that was not adequately covered within the scope of CyBOK. To answer this question, our study was unable to find any material that was of cyber security relevance which could not be adequately mapped within the scope of CyBOK.

However, this answer should not be considered fully comprehensive as firstly, this study did not cover every dark web site and secondly, it is possible such material was present but to such a minor extent that it was not captured by the study. It is possible there is content within the wider dark web that is not yet covered by CyBOK but further research would be required to determine this.

4.3.2 Against previous research

The findings of our study, with some exceptions, broadly supported the conclusions derived from previous dark web research.

User motivation

Users access the dark web for a variety of different reasons however, previous research has indicated they are typically far more driven by lust and greed than they are by other motives, such as politics [27]. The findings of our study strongly support this conclusion, particularly the greed motive, with many of the most prominent categories relating to fraud and money laundering. There were almost no instances of items related explicitly to terrorism or political extremism, though a very small percentage (1%) of guides were dedicated to "Weapons & Explosives" which could be purposed to meet these ends. There was also almost no items specifically relating to hacktivism - the use of computer-based techniques such as hacking as a form of civil disobedience to promote a political agenda or social change.

Other findings from this study were not supported. For example, the study detected a small percentage (5%) of websites dedicated to providing education and training relating to child exploitation, particularly in terms of grooming vulnerable children [27]. The results of our study did not find any items at all which fit this description, indicating that users looking for these guides are confined to niche websites. However, this may instead be a reflection of internal website policy rather than user motivation as this content is explicitly banned from many cryptomarkets and hacker forums [71].

Political extremists and child predators, whose presence has been documented on the dark web [18, 27], it would seem gravitate to the few sites that cater to their interests rather than across the broader dark web. Though it should be noted that some of the data sets utilised in our study were sourced from websites that are explicitly targeted at cyber criminals, such as the hacker forum data set, so it is difficult to generalise our findings. These findings may only represent these specific areas on the dark web concerned with learning material rather than the dark web at large.

Our study found that the codebook categories with the highest prominence were those concerned with methods for obtaining money, laundering money or preserving anonymity. They seem to indicate an informal cyber criminal journey which users embark on. Firstly users are introduced to common methods for obtaining money illegally (“Carding”, (“Fraud”), secondly methods for laundering this fraudulently obtained money (“Cashing Out”) are provided and lastly, techniques to stay anonymous while engaging in these activities are discussed (“Proxies”, “Tor”, “VPN”, “PGP”).

A suitable conclusion that can be derived from these findings in the context of previous research may be that cyber criminals specifically are primarily driven by financial motivations as opposed to political or emotional motives such as revenge. Expected behaviour for these actors on the dark web can therefore be considered largely through the prism of financial incentives.

Drugs

The existence of the trade in illicit substances, particularly within cryptomarkets, has been well documented on the dark web [14, 50]. A recent study by Martin et al. analysed content on a major cryptomarket (Agora) and found that drug related items accounted for nearly 80% of the total listings [50]. The study concluded that the total market for drugs within Agora “was massive, in all ways – number of products on sale, number of sellers operating in the country and total size of the market” [50].

While our study was focused on learning material and did not encompass either the wider cryptomarket or the dark web at large, our findings continue to support the conclusion that a significant proportion of traffic on these markets is dedicated to drug related activity. Our study found many instances of guides which relate to drugs. These constituted about 7.8% of the market listings, 3.21% of the cryptomarket topics but less than 1% of the hacker topics. These mostly comprised of guides on how to manufacture a certain drug or relating to either drug culture or paraphernalia. The “Transportation & Stealth” category (2% of total items) was also heavily related to drugs as it primarily featured guides on how to ship or conceal drugs without arousing the suspicion of law enforcement.

From a broader criminal perspective, the prevalence of drug related guides within the cryptomarket listings and forums is interesting as it suggests dark web markets are not only being used to trade illicit drugs themselves but also the methodology for drug production. This could potentially exacerbate the drug problem as guides for difficult to manufacture drugs (such as LSD) become more accessible. A potentially valuable study from a drug policy perspective may be to investigate the efficacy of these guides to determine whether an individual without a background in chemistry can use them to synthesise illegal drugs.

Software exploits

Studies have documented the trade in software exploits within the dark web [53] as well as the potential for these exploits to pose a serious threat to organisations and governments [37]. Our findings do not support the idea that this is particularly widespread, at least not within dark web sites we analysed that were focused on cyber criminal learning material. There were only a relatively small percentage of items relating to software exploits (which would fall within the "Hacking – General" category) within the cryptomarket listings or forums. However, this category, alongside website hacking, does feature more prominently within the hacker forum data set so the findings do support the idea that this market exists but suggests it is rather niche.

To determine any trends, the study analysed the number of items within these categories on an annual basis as shown below.

Category	**2011**		**2012**		**2013**		**2014**		**2015**		**2016**		**2017**	
Hacking – General	3	7%	33	7%	28	5%	22	5%	13	9%	21	9%	52	8%
Hacking – Website	21	51%	94	21%	115	20%	50	12%	15	10%	30	13%	53	8%

Table 4.5: Table contains number of items in categories covering exploits within the hacking forum data set as well as their percentage of the total for each year

Category	**2011**		**2012**		**2013**		**2014**		**2015**		**2016**		**2017**	
Hacking (General & Website)	24	N/A	127	429.17%	143	12.60%	72	-49.65%	28	-61.11%	51	82.14%	105	105.88%
Total	41	N/A	448	992.68%	578	29.02%	401	-30.62%	147	-63.34%	224	52.38%	632	182.14%

Table 4.6: Table contains the total number of items both overall and in the hacking categories (General & Website). The percentages represent the total percentage change in the number of items from the previous year.

One reassuring aspect of the above table from a cyber security perspective is it suggests any demand for learning material concerning software exploits, relative to other areas, is not growing. If this were to be the case it would be expected that the percentage of items relating to software exploits would be rising consistently. Excluding 2011, the percentage of items dedicated to this material fluctuates slightly but always remains between 15% to 28% with no clear increasing or decreasing trend. Additionally, the overall number of titles also remained relatively stable and does not appear to be rising in line with the general year-on-year increase in hacker forum titles. For example, from 2016 to 2017 the total number of titles rose 182% whereas the hacking categories only rose 105%.

From comparing our findings with previous research, it would seem most 'average' users of the dark web are more interested in the areas of cyber crime, such as carding, that require less technical learning and therefore yield a faster payoff [72]. This indicates there is certainly a demand for these software exploits but this is largely restricted to a minority of users contained within the hacker forums. This research did not encompass every dark web site however, so there may be other dark web sites or communities where they are also shared.

Malware Authorship & Malware Supply Chain

The term malware can be difficult to define but can be broadly thought of as software that harmfully attacks other software where to harmfully attack can be observed to mean to cause the actual behaviour to differ from the intended behaviour [45]. For clarity, it should be noted that exploits, which were discussed earlier, do not fall under the category of malware. An exploit is not malware itself, but rather it is a method used by cyber criminals to deliver malware [7].

Previous research has identified a trade in malware on both dark web cryptomarkets [49] and on hacker forums [39], though neither study investigates its prevalence relative to other content. Interestingly, our findings discovered only a small percentage of dark web learning material that was concerned with malware creation or methods for their delivery. "Malware Authorship" comprised only 0.42% of the total items and "Malware Supply Chain", which concerns methods for malware delivery, even less (0.17%). This was the case even within the hacker forums, where one might expect the presence of such topics. These findings would suggest dark web cyber criminals are generally not interested in learning the technical aspects of malware creation or the means for deploying such malware.

4.3.3 Differences between hacker forums, cryptomarket forums and cryptomarket listings

There were very significant differences between the results produced for each of the three sets of websites. For example, the categories that dominated each data set differed significantly, suggesting each set of websites attracts its own separate set user base with differing intentions.

The market listings were generally very homogeneous and consisted primarily of technical e-books (21%) followed by guides on carding (12%) and cashing out (12%). As the technical e-book category was largely academic e-books, one potential reason for its prevalence is the expensive cost of purchasing these through legitimate channels [41]. This could be driving students to darknet markets in order to acquire these materials at a fraction of the price.

Both of the forum data sets displayed more variety. The cryptomarket forums were primarily focused on anonymity (15%), stealth (10%), cashing out (9%) and PGP/GPG (6%). The hacker forum data set were focused on hacking (23%), proxies (16%) and contact lists (11%) which comprised almost entirely of combo lists. From these findings it would seem the hacker forums were the group of websites where cyber criminals were most likely to congregate due to high percentage of content dedicated to hacking topics and combo lists.

4.3.4 Comparison over time

The forum data sets were split by year and analysed with the intention of detecting any emerging trends. The market listing data is not dated thereby making analysis over time for this data set impossible, though previous research in this area has been discussed when relevant.

Number of items

Previous research has indicated that the number of dark web services is growing rapidly [51, 55]. By analysing the total number of titles for each year, the study can ascertain whether this trend also applies to learning material on the dark web. The number of titles within the cryptomarket data set increased steadily year on year. There were only 41 titles in 2012 compared to 4517 in 2014, an increase of over 10,000% within two years. The hacker forum data set also showed sustained growth though not to the same extent. The number of title increased from 41 in 2011 to 632 in 2017, an increase of approximately 1400%. Though this study was unable to analyse the market listings, previous research has determined that darknet markets have also grown considerably since their inception and will likely continue to do so in future [20].

The steady increase in topics over time within the forums support the conclusion from prior research that the demand for illicit services on the darknet is increasing. If this trend continues as expected, the wider implications of having a large and unaccountable market for cyber criminal material needs to be considered and anticipated in terms of cyber security education and wider policy.

As Tor is a common tool used to access these websites, the impact of a significant increase in Tor traffic also needs to be considered. Interestingly, the security of a single Tor user is a direct function of the number of overall users. The more users on a Tor network, the stronger the network becomes [47]. In short, the emergence of a very large darknet would be extremely difficult to hack and de-anonymise, thereby increasing the anonymity for its cyber criminal user base and further limiting options for law enforcement.

Knowledge area emphasis

The annual results data was mapped to CyBOK knowledge areas to determine the respective emphasis for each year. The aim was to identify any emerging trends, particularly whether one particular knowledge area was becoming more prevalent, and discuss any implications this may have for cyber security. This was conducted firstly for the cryptomarket forums (see Table below) and then the hacker forums (Table 4.8).

Knowledge Area	2012	2013	2014	2015*
Attacks & Defences	0.357	0.239	0.357	0.395
Systems Security	0.286	0.397	0.225	0.202
Infrastructure Security	0.036	0.013	0.036	0.049
Human, Organisational & Regulatory Aspects	0.321	0.348	0.371	0.348
Software & Platform Security	0.000	0.003	0.011	0.006

*Until July

Table 4.7: Emphasis of cryptomarket forum data set on an annual basis

Analysis of the cryptomarket forum topics suggests the emphasis of each knowledge area appears to remain relatively stable year on year. As it has remained stable for four consecutive years it suggests the cryptomarket forum has continued to draw a similar user base with broadly similar motivations.

One caveat to this conclusion is that the data set does not stretch beyond July 2015. Further research involving more recent data sets would be required to determine whether this trend continues beyond this date.

Knowledge Area	2011	2012	2013	2014	2015	2016	2017	2018*
Attacks & Defences	0.18	0.34	0.36	0.34	0.40	0.53	0.68	0.44
Systems Security	0.00	0.00	0.01	0.04	0.06	0.04	0.02	0.03
Infrastructure Security	0.00	0.03	0.02	0.02	0.03	0.04	0.01	0.00
Human, Organisational & Regulatory Aspects	0.07	0.29	0.28	0.41	0.31	0.21	0.19	0.47
Software & Platform Security	0.75	0.34	0.33	0.18	0.20	0.18	0.10	0.06

*only contains data for January and February

Table 4.8: Emphasis of hacker forum data set on an annual basis

The percentage of titles within the "Attacks & Defences" knowledge areas increased almost year-on-year, rising from 18% in 2011 to 68% in 2017. This did dip the following year but this data was only inclusive of Janurary and February. The consistent growth in this area over many years suggests this trend will likely continue in future and that cyber security curriculums need to be well prepared to counter that demand. However the "Software & Platform Security" knowledge areas (consisting only of the web and mobile hacking codebook categories) showed the exact inverse trend, descending from 75% in 2011 until 10% in 2017.

One potential reason for this trend could be greater awareness of and greater accessibility to these darknet forums. An increase of financially motivated, less technically minded users could explain the increase in fraud related categories (such as "Carding" and "Fraud") over more technically minded categories (such as "Hacking" and "Malware Authorship").

4.3.5 Comparison against research into cyber security curriculums

The results of our study were contrasted against prior research into cyber security qualifications to provide insight into their effectiveness in countering the cyber criminal methodology being compiled on the dark web. To accomplish this our study compared our weightings with the weightings of these curriculums to determine if any relationship between them was present and in particular, whether there were knowledge areas in the learning material that were not appropriately covered within these curriculums.

Adversarial Behaviour

Comparing the spider charts, one clear difference is the knowledge area "Adversarial Behaviour" which features prominently in the darknet learning material, in contrast, is severely under represented in every cyber security accreditation. The implication of this is that cyber security graduates are not being effectively trained in this knowledge area, are unable to produce effective preventative measures thereby enabling these guides to continue to be effective and popular.

This discrepancy suggests it may be worth increasing the presence of this knowledge area within these curriculums in order to reflect the disproportionate demand from cyber criminals on the darknet. In addition, the mappings over time suggest that cyber criminal interest in this knowledge area will only increase over time.

It is worth stating that much of the content assigned to this knowledge area was concentrated within the "Carding" and "Cashing Out" codebook categories. The increasingly high levels of credit card fraud and money laundering related learning materials suggest that any measures that are being taught to curb these are ineffective. In addition, it suggests the introduction of Bitcoin and Bitcoin mixers have made money laundering more accessible, easier to conduct and harder to detect. Further research would need to be conducted in order to identify the specific areas of weakness within these fraud detection and anti-money laundering systems.

Privacy & Online Rights

The second most prevalent knowledge area was "Privacy & Online Rights". This knowledge area featured prominently within every group of websites, particularly the hacker forums where it accounted for almost 30% of content. These findings would suggest a secondary focus of users on the dark web is on defensive tools to stay anonymous while engaging in illicit activity.

In contrast this knowledge area does not feature heavily within the cyber security curriculums, ranging from 0-11% of content in each curriculum. Instead the focus is focus on methods for designing secure systems, as evidenced by the prominence of knowledge areas such as "Risk Management and Governance" and "Secure Software Design Development". This discrepancy I believe is justifiable due to competing goals and motivations of these two different groups of actors. Generally, it is far more valuable for a cyber security professional to be able to design secure systems than it is for them to operate privacy focused software such as Whonix or Tor for example. While there is no doubt a need for cyber security qualifications to teach a basic understanding of these privacy technologies, they are typically of more interest to those engaging in hacking activities as opposed to those intending to trace them.

There are some caveats to our findings in this area. As discussed on p. 34, there were issues with assigning some of the codebook categories directly to knowledge areas as these would often contain characteristic of multiple knowledge areas. This was particularly the case with those that related to “Privacy & Online Rights” knowledge area. Therefore it should be stated that many of the individual codebook categories which comprise this knowledge area also share overlaps with other knowledge areas. For example, the VPNs category will likely contain content that could be included in “Network Security” and the Tor category will likely contain content that could relate to “Web & Mobile Security”. Further research may be required to tease apart exactly how much content in these codebook categories relate to each CyBOK knowledge area.

Low prevalence of infrastructure security

A commonality between all the cyber security curricula and darknet learning material is the low prevalence of infrastructure security. This is particularly evident in the hacker forum data set where there were almost no instances of titles relating to infrastructure security.

This is not to say we should be remiss about infrastructure security. In fact one of the most highly sought after skills for cyber security recruiters is the ability to root hardware [40]. There may be a need for a niche qualification where this is the focus but the findings from this research does not support the notion that infrastructure security should be given equal weighting with other knowledge areas on a general cyber security qualification.

Software and Platform Security

Knowledge areas that relate entirely to the development of software, such as “Secure Software Lifecycle”, are virtually non-existent within the cyber criminal learning material. This would suggest users on the hacker and cryptomarket forums are generally disinterested in legitimate software development or at least, discussing it on the dark net.

Contrast this as well with the high percentage of categories dedicated to hacking methods, particularly within the hacker forums and it is clear these users are far more interested with compromising secure systems than they are with developing them. This finding may not be particularly surprising but it is useful to confirm with evidence.

This knowledge area is far more prevalent within the cyber security curriculums, reflecting the fact these cyber security professionals are far more likely to be involved in the development of secure software programs than their adversaries on the darknet. This discrepancy it would seem can be thought of as complimentary, with cyber security professionals learning to develop secure systems that will not be compromised by their counterparts on the dark web.

Similarities between hacker forums and IISP

The mappings of the hacker forum data set and the IISP qualification shared many similarities, as demonstrated in the figure below.

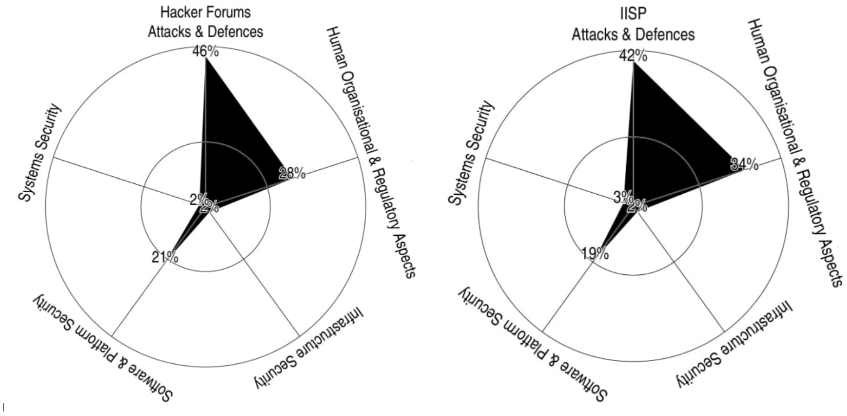


Figure 4.9: Hacker forums comparison with IISP

Both mappings display a high prevalence towards both the "Attacks & Defences" and "Human Organisational & Regulatory Aspects" CyBOK categories. At a glance, this alignment suggests the curriculum is effectively covering the same material available on the hacker forums. However when analysed in more detail, it becomes clear that the focus differs considerably within these broad categories. Firstly, the spread is far broader across the knowledge areas within the IISP qualification whereas the hacker forum set is concentrated almost entirely within three knowledge areas: Adversarial Behaviours (45%), Privacy & Online Rights (28%) and Web & Mobile Security (21%).

Secondly, within the "Human & Organisational" category the focus is primarily on the "Risk Management & Governance" knowledge area whereas the hacker forums is entirely focused on "Privacy & Online Rights". Similarly, within "Attacks and Defences" the IISP curriculum focuses primarily on "Security Operations & Incident Management" whereas the hacker forum emphasis is on "Adversarial Behaviour".

In line with our previous comparisons, these differences can be explained to a great extent in terms of proactive versus reactive intentions. For instance, "Adversarial Behaviours" to an extent features methodology for attacking or compromising secure systems while "Security Operations & Incident Management" comprises the methodology for securing those systems. As this qualification is disproportionately focused in these knowledge areas it would suggest this qualification is geared towards students interested in defensive roles that revolve around the managing risk. In contrast, the typical knowledge areas required to become effective penetration testers, cryptographers or hardware engineers are only very lightly covered in the IISP qualification.

4.3.6 Proposals

This section will outline a few proposals, both for cyber security education and for wider policy, based on our findings within the context of the relevant literature.

For cyber security education

Cyber security is an expanding field [35] with many different specialist roles. These specialisms each require a different skill set. An individual who wants to become a cryptographer will require a significantly different skill set than a penetration tester for example. Despite this most qualifications are marketed as generalist, such as those mapped in the Hallett et al study [40], even though their actual content differs significantly. This study suggests these courses need to be more open about the biases between these different frameworks so students can make informed choices about the courses they take [40].

Our study detected that while the "Adversarial Behaviour" knowledge area was very prominent within the dark web learning material, it did not feature heavily in the cyber security curriculums. This was particularly evident for the codebook categories that related to financial crime, particularly credit card fraud and money laundering. These were the two most prevalent categories that contained cyber criminal material and accounted for approximately 18% of the overall learning material. Additionally, the analysis by year suggests that the demand for such guides has been growing and will likely continue to grow. Other developments in this area are also concerning such as the introduction of privacy-focused cryptocurrencies like Monero which will likely make money laundering even more difficult to detect.

As the number of cyber criminals engaging in such activity increases, it is likely that there will be increased demand for cyber security professionals who specialise in these areas. However the trends detected by this study suggest the failure of generalist curriculums to effectively cover this material. A potential resolution may be the creation a specialist qualification to cater to those working within in the financial sector. These graduates would then be trained in the most cutting edge fraud detection and anti-money laundering techniques to meet the demand from financial services companies.

Lastly, previous research into cyber security qualifications has determined they often rely heavily on ineffective techniques such as multiple choice examinations [44]. Therefore, if any of these proposed changes are to be actioned, they should not only be remedied in terms of emphasis but also using one of the proposed cost-effective methodology identified in the study, such as via oral examination.

For policy

Cyber security education should not be considered in a vacuum. There are many wider contributing factors to the skills gap within cyber security other than cyber security qualifications that also need to be addressed [4]. For example, amending the curriculums of these qualifications will not address the skills gap if there is still a lack of young people willing to enter the profession.

This 2016 government report suggested part of the skill gap was due to insufficient exposure to cyber and information security concepts in computing courses [4].With this in mind, incorporating modules on cyber security within university computer science courses may yield a benefit. This would serve the dual purpose of providing graduates with some components of a cyber security education as well as introduce students to the idea of a career in cyber security at an early stage.

Policy also needs to be considered through the perspective of financial regulation. As our findings demonstrate, a significant percentage of dark web learning material is dedicated to financial crime, particularly credit card fraud and money laundering. Our mappings by year also indicate demand for such material is likely to continue growing in future. Of most concern, research has suggested that if these guides are followed meticulously there are limited interceptive opportunities for law enforcement [73]. Banks and financial regulators will need to consider the impact of a increasing amounts of credit card fraud and money laundering that will be enabled by propagation of guides on the dark web.

5 Evaluation

5.1 Aims and Objectives

This section is intended to evaluate to what extent the project achieved the stated aims and objectives set out within the initial research proposal.

Aim: To create a mapping of cyber criminal learning material available on the dark web according to CyBOK knowledge areas.

A substantial data set of darknet learning material were categorised both manually and through use of a classifier. These categories were then assigned to the most relevant CyBOK knowledge area, the total number of items for each knowledge area was summed and the results analysed. Data representations, in the form of spider charts and bar charts, were created to summarise the results of our analysis thereby achieving this aim.

Aim: To identify any trends and contrast our findings with previous research to inform the current state of cyber security knowledge.

The spider charts that we created were contrasted with those produced through similar analysis of cyber security curriculums. The differences and similarities between these mappings were discussed and on this basis potential improvements for these curriculums were proposed.

The results were also compared between each of the three groups of websites. The significant differences between our findings in this respect indicated that each group of websites attracted its own distinct user base with differing motivations. In particular, the hacker forums was identified as being most likely to attract a cyber criminal user base.

Additionally, the forum titles were broken down into yearly sub-groups and analysed to determine any trends. From this it was determined that the steady increase in topics over time support previous the conclusion from prior research that the demand for illicit services on the darknet will see continued growth.

5.2 Challenges

5.2.1 Codebook

Bias towards market listings

Initially the codebook was formulated only from analysis of the cryptomarket listings. This bias caused issues when analysing the forum title data set as the data differed in many ways. While there were no significantly different categories to emerge (if this had occurred, this would certainly have been factored in through the creation of new categories), there were some differences that had to be accommodated.

For example, a significant number of thread titles were dedicated to the encryption software Veracrypt and Truecrypt. This was relevant to the study but was not being adequately covered by the codebook. To remedy this, the "PGP/GPG" category was broadened to include these items and these thread titles were added to the ground truth data so that the classifier would categorise these accurately. There was a similar issue with items relating to software exploits, which were not present in the market listings but somewhat prominent in the hacker forum data set. As there was no specific category for software exploits, these items were categorised as "Hacking – General" alongside other items, such as hacking tools or general discussion relating to hacking. In retrospect, the study would have benefited from a more detailed codebook formulated from a mix of all three data sets. This would in turn, have made the results clearer and easier to contrast against prior research.

Items relating to multiple categories

Even though the classifier can only assign each item to only one category, many items contained relevant keywords for multiple categories. This was particularly evident for items in both the "Carding" and "Cashing Out" categories where guides often combined features from both categories.

The classifier design accommodated for this by assigning the item to the category which had the highest total log-likelihood value. It was decided this was the most reasonable method to use as allowing an item to be assigned to multiple categories would have significantly complicated the analysis. However, a disadvantage of this methodology is that any items which do span multiple categories will only be assigned to one category. This will mean that some categories, such as "Carding" and "Cashing Out", may be far more prevalent than the results show.

Mapping to CyBOK knowledge areas

The process of mapping each category in the codebook to a CyBOK knowledge area was conducted in a meeting with my project supervisors. Most categories were easy to place but some were more contentious.

In retrospect, there was scope for more detailed categories. Some categories within the codebook were extremely broad and encompassed many items such as the "eBooks

– Technical” category which could include both a guide on how to use Adobe Photoshop and a guide on how to code in Python. Splitting up this category may have provided more insight into the specific areas of computer science that dark web users were interested in.

5.2.2 Classifier

While the design of the model was generally effective and therefore achieved our primary purpose, there were are a few limitations which hinder its overall effectiveness as well as its practicality for use in further studies.

Training data bias towards market listings

Initially the ground truth data set which was used to train the classifier consisted entirely of market listings as opposed to forum thread titles. As a direct result of this and the fact the market data was more formal and homogenised than the thread titles, the classifier was far more effective at categorising the market listings than the forum titles. This was remedied by manually categorising a substantial subsection of the forum titles (1000 titles) and merging these items with the ground truth data set.

Uncommon misspellings

The classifier was ill-equipped to categorise items which contained uncommon misspellings of meaningful words. Such misspellings were fairly common within both of the forum title data sets but much less so within the market listings. While a human manually categorising the item would likely spot the error, the classifier was unable to do so.

The crux of the issue is that these uncommon misspellings, which are not present within the training data, will result in tokens that the classifier assigns a low log likelihood score to and therefore items containing these misspellings have a highly increased chance of being incorrectly categorised.

A potential method that could have been implemented to mitigate this may be to run a spell-check function to correct these errors in the data preparation stage prior to keyword analysis. There are two pitfalls to this solution. Firstly, some colloquialisms used on the darknet will not be recognised as words by most spell checkers. Secondly, some words may be so severely misspelled they cannot be amended by the spell-checker function.

5.2.3 Rare phrases

There was a similar issue for meaningful keywords or key phrases that are not present or rarely occur within the ground truth data. As it is not present within the ground truth data set, the classifier is unable to assign it an accurate log likelihood value for the relevant category and it is erroneously categorised as ’Other’.

Such a case was typically an unusual name for either a drug, software or an explosive. One such example that I spotted and manually re-categorised was "Homemade Semtex - C4's Ugly Sister" which included the relevant keywords "Semtex" and "C4" that were not present with the ground truth data. The classifier does attempt to control for this to an extent by performing checks for certain key phrases prior to the tokenisation stage. These checks are very limited however and were generally based on common phrases I had realised were not being detected by the classifier.

One potential resolution to this issue would be for the classifier to check each token against separate lists of relevant software, drugs and explosives, assigning it to the relevant category if detected. However, the extremely large number of potentially relevant phrases makes controlling for this in totality almost impossible.

5.2.4 Efficiency

A more practical limitation of the classifier is its processing speed. It can take the classifier several hours to train and then classify a list of 5,000 items, each consisting of a string containing approximately five words.

Even with the large data sets, this did not prove to be particularly problematic for my project as I only had to run the classifier a relatively few number of times. The most time consuming stage was performing the k cross-validation which required the classifier to be trained and re-run ten times. This did take many hours but was only required to be done a few times before the desired Cohen's kappa result was achieved. However the classifier code may need to be rendered more efficient if used to categorise significantly larger data sets. This could involve an in-depth analysis of the code to ensure its efficiency or potentially converting the program into a programming language with a faster execution time such as C.

Another aspect that could be hindering the programs efficiency is the storage format containing the training data which pandas must write to and access repeatedly. While CSV files are an appropriate file format for use with the pandas library, a study has shown that other file formats, particularly feather and parquet, have a far shorter load and save time [78]. Amending the storage format to one of these could significantly reduce the length of time it takes the program to run.

5.2.5 Classifier validation methodology

During the write up of the proposal and initial stages of the project, a formal method for validating the classifier had not yet been thoroughly considered. After the creation of the classifier this was discussed with my project supervisors and I conducted research online. For a variety of reasons outlined in the implementation (p.) k-fold cross validation was decided to be the most appropriate methodology.

In order to action this, it required the writing of a Python program which would automate the cross-validation process. The writing of such a program had not been factored in during the planning stage but it proved relatively simple to write and did not set the intended schedule of the project back.

6 Conclusion and Future Work

6.1 Conclusion

This project analysed the broad range of cyber criminal learning material available on the dark web forums and marketplaces. To achieve this aim a codebook was created and a substantial data set was categorised both manually and via an automated classifier program. The results were then mapped to CyBOK knowledge areas in order to determine the areas, from a cyber security perspective, with least and most emphasis.

The study determined that cyber criminal learning material on the dark web is disproportionately focused on the "Adversarial Behaviour" knowledge area, in particular the activities of carding (8%), fraud (5%) and money laundering (9%). Additionally by mapping this content on an annual basis, it suggests this knowledge area will likely continue to grow in future as the percentage of content dedicated to credit card fraud and money laundering is increasing.

Our mappings were compared with the mappings produced by prior research into cyber security curriculums [40]. The content of the curriculums was generally more evenly spread across the knowledge areas whereas the cyber criminal learning material was focused almost exclusively in three ("Adversarial Behaviour", "Privacy & Online Rights" and "Cryptography"). Notably, the "Adversarial Behaviour" knowledge area did not feature prominently within any of the curriculums. These differences between two sets of spider maps could be explained to a great extent in terms of proactive versus reactive intentions. For example, the emphasis on breaking into secure systems versus designing secure systems. It was also reflective of the broader career paths available in cyber security.

Even taking these factors into account, the low levels of content relating to "Adversarial Behaviour" combined with the low levels of students willing to enter the profession [4] is concerning. It suggests graduates from these qualifications may not be gaining the relevant knowledge to tackle the high levels of credit card fraud and money laundering. This research discusses the potential introduction of more specialised qualifications that target these particular areas of cyber crime.

6.2 Suggestions for Future Work

6.2.1 Contemporary data sets

Much of the source data is also somewhat historic in nature and therefore to some extent, our results may not reflect the current state of the content available on the dark web. The most recent items in our study date from July 2018. Any trends detected by the project may since have been halted. It would interesting to replicate our study with a more contemporary data set to determine whether these trends continue beyond the data sets analysed in this study.

6.2.2 Efficacy of content

This research does not test or in any way factor in the efficacy of the learning material due to the volume of the data being so large that to analyse this as well would be impractical within the timescale available. In addition, in the majority of cases the full content of the learning material is not readily accessible. An interesting follow up study would be to investigate the efficacy of the learning material in each category.

Due to its anonymous and often criminal nature, many advertised dark web services are not as advertised. While reputation systems guard against this to some extent, there is little accountability either for new vendors or if a vendor who has already established a reputation of being honest decides to turn to scamming. In some cases, known colloquially as 'exit scams', whole cryptomarkets have disappeared with the site administrators stealing any cryptocurrency (often running into millions of pounds) held on their system [14].

This would also provide further insight into the effectiveness of current cyber security qualifications. For example, it may be the case that the areas with a low level of prevalence on the darknet are in fact, the most effective.

There is some precedence for this as prior research has already demonstrated the efficacy of dark web learning material related to carding and cashing out. This study conducted by van Hardeveld et al. [73] analysed tutorials and determined that if these guides are followed meticulously they were effective and interceptive opportunities for law enforcement were limited. Therefore, we can be reasonably confident that the content advertised does contain valid cyber criminal learning material. A new study could expand on this conclusion to determine whether this held true across other categories.

6.2.3 Classifier optimisation

There were many further improvements I would have liked to have made to the classifier had the project not been time constrained. As shown on page[], even though the classifier was able to achieve substantial agreement with a human rater, it was not able to achieve the accuracy of another human rater.

For example, I would like to have experimented with different keyword analysis algorithms such as log-likelihood ratio to determine whether these would have produced increased validity over the log likelihood values.

Machine learning techniques could also be implemented to allow for the continuous training and improvement of the classifier. This would likely involve some form of 'active learning' techniques whereby the classifier assigns labels to previously unseen items [68]. These items are then manually checked by an expert, amended if necessary and any corrected items are then fed back in to the ground truth data set. This process is repeated continually until desired agreement is achieved. While this is time consuming, it should result in better agreement outcomes for the classifier.

A more complex and long-term goal would be to integrate the classifier with web crawler to produce a continuous mapping of learning material on the darknet. This would produce a far more detailed and current mapping. It would also allow for the more accurate tracking of trends within the darknet.

6.2.4 Advanced data representation

This research study generated a large amount of output data which span three levels of detail. Firstly, the individual codebook category then the CyBOK knowledge area and lastly, the broader CyBOK knowledge area. Much of this data was summarised in static spider and bar charts which appropriately summarised the findings from this study. However, there may have been more effective and engaging methods to display this data.

One such method would involve the production of interactive graphs. This could be done by uploading the results to a relational database to which custom queries could then be applied using a database language such as SQL. The user interface could be provided by either a Python or a Java graphical library such as JavaFX. The resulting application would allow a user to generate interactive data representations based on their own specific criteria.

7 Glossary

Carding Colloquial term for credit card fraud.

Dark Web A section of the deep web that has been intentionally hidden and is not accessible through standard web browsers without the use of specialist routing software such as Tor [36].

Darknet The collection of networks and technologies used to share digital content on the dark web [29].

Denial of Service (DOS) When legitimate users are denied access to computer services (or resources), usually by overloading the service with requests [5].

Digital Forensics The process of identifying and reconstructing the relevant sequence of events that have led to the currently observable state of a target IT system or (digital) artifacts.

Doxing An attack where the victim's private information is publicly released online.

Exploit Software or data that takes advantage of a vulnerability in a system to cause unintended consequences [5].

Hacker In mainstream use as being someone with some computer skills who uses them to break into computers, systems and networks.[5].

PGP (Pretty Good Privacy) An encryption program that enables users to encrypt files and messages [34].

Phishing A fraud that lures users into giving away access credentials to online services to a criminal.

Tor (The Onion Router) A distributed overlay network designed to anonymise TCP-based applications like web browsing, secure shell, and instant messaging [66].

Trolling Behaviour within a community that falls outside acceptable bounds defined by that community [19].

Bibliography

[1] 8 best dark web search engines for 2020. URL: https://www.hackread.com/8-best-dark-web-search-engines-for-2020/.

[2] Data Health Check 2016 | Survey Results. URL: https://datahealthcheck.databarracks.com/2016/#intro-section-2.

[3] Is Bitcoin Anonymous? URL: https://bitcoinmagazine.com/what-is-bitcoin/is-bitcoin-anonymous.

[4] National Cyber Security Strategy 2016 to 2021 - GOV.UK. URL: https://www.gov.uk/government/publications/national-cyber-security-strategy-2016-to-2021.

[5] NCSC glossary - NCSC.GOV.UK. URL: https://www.ncsc.gov.uk/information/ncsc-glossary.

[6] Playpen: The Story of the FBI's Unprecedented and Illegal Hacking Operation | Electronic Frontier Foundation. URL: https://www.eff.org/deeplinks/2016/09/playpen-story-fbis-unprecedented-and-illegal-hacking-operation.

[7] What Is an Exploit? - Cisco. URL: https://www.cisco.com/c/en/us/products/security/advanced-malware-protection/what-is-exploit.html.

[8] What Is The Difference Between Deep Web, Darknet, And Dark Web? URL: https://fossbytes.com/difference-deep-web-darknet-dark-web/.

[9] Victor Acin. Making sense of the dark web. *Computer Fraud and Security*, 2019(7):17–19, jul 2019. doi:10.1016/S1361-3723(19)30075-2.

[10] Mohammed Almukaynizi, Alexander Grimm, Eric Nunes, Jana Shakarian, and Paulo Shakarian. Predicting Cyber Threats through Hacker Social Networks in Darkweb and Deepweb Forums. In *Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas on - CSS 2017*, pages 1–7, New York, New York, USA, oct 2017. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3145574.3145590, doi:10.1145/3145574.3145590.

[11] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella. The 'K' in K-fold Cross Validation. In *ESANN 2012*, pages 441–446, 2012. URL: http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[12] Azsecure. Other Forums: AZSecure-data.org. URL: https://www.azsecure-data.org/other-forums.html.

[13] Andres Baravalle and Sin Wee Lee. Dark web markets: Turning the lights on AlphaBay. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11234 LNCS, pages 502–514. Springer Verlag, nov 2018. URL: https://link.springer.com/chapter/10.1007/978-3-030-02925-8_35, doi:10.1007/978-3-030-02925-8_35.

[14] V Bhaskar, Robin Linacre, and Stephen Machin. Dark web: The economics of online drugs markets. *LSE Business Review Blog*, 2017.

[15] Ravishankar Borgaonkar. Tor and onion routing: Protecting your privacy. *From End-to-End to Trust-to-Trust*, page 30, 2008.

[16] Pierre Bourque, Richard E Fairley, et al. *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0*. IEEE Computer Society Press, 2014.

[17] Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. Dark net market archives, 2011-2015. https://www.gwern.net/DNM-archives, July 2015. Accessed: 2020-05-25. URL: https://www.gwern.net/DNM-archives.

[18] Hsinchun Chen. Dark Web: Exploring and Mining the Dark Side of the Web. pages 1–2. Institute of Electrical and Electronics Engineers (IEEE), nov 2011. doi:10.1109/eisic.2011.78.

[19] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/2998181.2998213.

[20] Michael Chertoff. A public policy perspective of the Dark Web. *Journal of Cyber Policy*, 2(1):26–38, jan 2017. URL: https://www.tandfonline.com/doi/abs/10.1080/23738871.2017.1298643, doi:10.1080/23738871.2017.1298643.

[21] Michael Chertoff and Tobby Simon. The Impact of the Dark Web on Internet Governance and Cyber Security. Technical report, feb 2015.

[22] Nicolas Christin. Traveling the silk road. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 213–224, New York, New York, USA, 2013. Association for Computing Machinery (ACM). URL: http://dl.acm.org/citation.cfm?doid=2488388.2488408, doi:10.1145/2488388.2488408.

[23] Vincenzo Ciancaglini, Marco Balduzzi, Robert Mcardle, and Martin Rösler. Below the Surface: Exploring the Deep Web. Technical report, 2015.

[24] Mestre Nuno Ricardo Mateus Coelho. *Paranoid Operating System Methodology for Anonymous & Secure Web Browsing*. PhD thesis, Instituto Superior de Engenharia do Porto, 2020.

[25] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, apr 1960. URL: `http://journals.sagepub.com/doi/10.1177/001316446002000104`, `doi:10.1177/001316446002000104`.

[26] Cyberscoop. How many dark web marketplaces actually exist? about 100., 2019. URL: `https://www.cyberscoop.com/dark-web-marketplaces-research-recorded-future/`.

[27] Janis Dalins, Campbell Wilson, and Mark Carman. Criminal motivation on the dark web: A categorisation model for law enforcement. *Digital Investigation*, 24:62–71, mar 2018. `doi:10.1016/j.diin.2017.12.003`.

[28] Maurice Dawson and Jose Antonio Cárdenas-Haro. Tails linux operating system: Remaining anonymous with the assistance of an incognito system in times of high surveillance. *International Journal of Hyperconnectivity and the Internet of Things (IJHIoT)*, 1(1):47–55, 2017.

[29] Rafiqul Islam Erdal Ozkaya. *Inside the Dark Web*. Routledge, 2019. URL: `https://books.google.co.uk/books?hl=en&lr=&id=GCGeDwAAQBAJ&oi=fnd&pg=PT15&dq=origins+of+the+%22dark+web%22&ots=DlXbAcoEXL&sig=1HmV5an8ocp6Lvcbfqhjw8ZldmQ&redir_esc=y#v=onepage&q=originsofthe%22darkweb%22&f=false`.

[30] Mohd Faizan and Raees Ahmad Khan. Exploring and analyzing the dark Web: A new alchemy. *First Monday*, apr 2019. URL: `https://journals.uic.edu/ojs/index.php/fm/article/view/9473/7794https://journals.uic.edu/ojs/index.php/fm/article/view/9473`, `doi:10.5210/fm.v24i5.9473`.

[31] Joan Feigenbaum, Aaron Johnson, and Paul Syverson. A model of onion routing with provable anonymity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4886 LNCS, pages 57–71. Springer Verlag, 2007. URL: `https://link.springer.com/chapter/10.1007/978-3-540-77366-5_9`, `doi:10.1007/978-3-540-77366-5_9`.

[32] Kristin Finklea. Dark Web. Technical report, 2017. URL: `www.crs.gov`.

[33] Steven Furnell, Pete Fischer, and Amanda Finch. Can't get the staff? The growing need for cyber-security skills. *Computer Fraud and Security*, 2017(2):5–10, feb 2017. `doi:10.1016/S1361-3723(17)30013-1`.

[34] Simson Garfinkel. *PGP: pretty good privacy*. " O'Reilly Media, Inc.", 1995.

[35] Daniel Garrie and Michael Mann. Cyber-security insurance: navigating the landscape of a growing field. *J. Marshall J. Info. Tech. & Privacy L.*, 31:i, 2014.

[36] Robert Gehl. *Weaving the Dark Web: Legitimacy on Freenet, Tor, and I2P*, volume 3. The MIT Press, 2018.

[37] Ibrahim Ghafir and Vaclav Prenosil. Advanced persistent threat attack detection: An overview. *International Journal Of Advances In Computer Networks And Its Security*, 12 2014.

[38] Iain Gillespie. Cyber cops probe the deep web, 2013. URL: `https://www.smh.com.au/technology/cyber-cops-probe-the-deep-web-20131023-2vzqp.html`.

[39] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 13–18. IEEE, 2017.

[40] Joseph Hallett, Robert Larson, and Awais Rashid. Mirror, Mirror, On the Wall: What are we Teaching Them All? Characterising the Focus of Cybersecurity Curricular Frameworks. Technical report, 2018.

[41] III Henry L. Roediger. Why Are Textbooks So Expensive? *APS Observer*, 18(1), jun 2005. URL: `https://www.psychologicalscience.org/observer/why-are-textbooks-so-expensive`.

[42] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2007.

[43] K J Knapp, C Maurer, and M Plachkinova. Maintaining a Cybersecurity Curriculum: Professional Certifications as Valuable Guidance. *Journal of Information Systems Education*, 28(2):101–114, dec 2017.

[44] William Knowles, Jose M Such, Antonios Gouglidis, Gaurav Misra, and Awais Rashid. All that glitters is not gold: on the effectiveness of cyber security qualifications. *IEEE Computer*, 50(12):60–71, 2017.

[45] Simon Kramer and Julian C Bradfield. A general definition of malware. *Journal in computer virology*, 6(2):105–114, 2010.

[46] Andrei Lima Queiroz, Susan Mckeever, and Brian Keegan. Detecting Hacker Threats: Performance of Word and Sentence Embedding Models in Identifying Hacker Communications. Technical report, 2019. URL: `http://tiny.cc/8ws67y`.

[47] Katarzyna Maniszewska and Paulina Piasecka. *SECURITY AND SOCIETY IN THE INFORMATION AGE. VOL. 2*. 2020. URL: `http://www.civitas.edu.pl`, `doi:10.6084/m9.figshare.11555511`.

[48] Marc Goodman. *Future Crimes: Everything Is Connected, Everyone Is Vulnerable and What We Can Do about It*. Doubleday Books, 2016. URL: `https://www.amazon.co.uk/Future-Crimes-Everything-Connected-Vulnerable/dp/0385539002`.

[49] Ericsson Marin, Mohammed Almukaynizi, Eric Nunes, and Paulo Shakarian. Community finding of malware and exploit vendors on darkweb marketplaces. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 81–84. IEEE, 2018.

[50] James Martin, Rasmus Munksgaard, Ross Coomber, Jakob Demant, and Monica J Barratt. Selling drugs on darkweb cryptomarkets: differentiated pathways, risks and rewards. *The British Journal of Criminology*, 60(3):559–578, 2020.

[51] Mike McGuire. *Into The Web of Profit*. Bromium, 2019.

[52] Mary L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012. URL: /pmc/articles/PMC3900052/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/, doi:10.11613/bm.2012.031.

[53] Charlie Miller. The Legitimate Vulnerability Market Inside the Secretive World of 0-day Exploit Sales. Technical report, 2007. URL: www.securityevaluators.com.

[54] Gilad Mishne et al. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, pages 321–327, 2005.

[55] Daniel Moore and Thomas Rid. Cryptopolitik and the Darknet. *Survival*, 58(1):7–38, jan 2016. URL: http://www.tandfonline.com/doi/full/10.1080/00396338.2016.1142085, doi:10.1080/00396338.2016.1142085.

[56] Satoshi Nakamoto et al. Bitcoin: A peer-to-peer electronic cash system.(2008), 2008.

[57] OnionScan. Onionscan report: Freedom hosting ii, a new map and a new direction., 2017. URL: https://mascherari.press/onionscan-report-fhii-a-new-map-and-the-future/.

[58] Charlie Osborne. Silk road dark web marketplace just does not want to die, 2016. URL: https://www.zdnet.com/article/silk-road-dark-web-marketplace-just-does-not-want-to-die/.

[59] Paganini. The good and the bad of the Deep Web - Security AffairsSecurity Affairs, 2012. URL: https://securityaffairs.co/wordpress/8719/deep-web/the-good-and-the-bad-of-the-deep-web.html.

[60] Andrew J. Park, Brian Beck, Darrick Fletche, Patrick Lam, and Herbert H. Tsang. Temporal analysis of radical dark web forum users. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, pages 880–883. Institute of Electrical and Electronics Engineers Inc., nov 2016. doi:10.1109/ASONAM.2016.7752341.

[61] Terry R. Rakes, Jason K. Deane, and Loren Paul Rees. IT security planning under uncertainty for high-impact events. *Omega*, 40(1):79–88, jan 2012. doi:10.1016/j.omega.2011.03.008.

[62] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, Andrew Martin, Makayla Lewis, and Claudia Peersman. Scoping the cyber security body of knowledge. *IEEE Security & Privacy*, 16(3):96–102, 2018.

[63] Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora -*, volume 9, pages 1–6, Morristown, NJ, USA, 2000. Association for Computational Linguistics (ACL). URL: `http://portal.acm.org/citation.cfm?doid=1117729.1117730`, `doi:10.3115/1117729.1117730`.

[64] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. `doi:10.1007/978-0-387-39940-9_565`.

[65] Perri Reynolds and Angela SM Irwin. Tracking digital footprints: anonymity within the bitcoin system. *Journal of Money Laundering Control*, 2017.

[66] Paul Syverson Roger Dingledine, Nick Mathewson. Tor: The Second-Generation Onion Router. In *Proceedings of the 13th USENIX Security Symposium*, 2004.

[67] Dakota S. Rudesill, James Caverlee, and Daniel Sui. The Deep Web and the Darknet: A Look Inside the Internet's Massive Black Box. *SSRN Electronic Journal*, oct 2015. `doi:10.2139/ssrn.2676615`.

[68] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[69] Jana Shakarian, Andrew T. Gunn, and Paulo Shakarian. Exploring malicious hacker forums. In *Cyber Deception: Building the Scientific Foundation*, pages 259–282. Springer International Publishing, jan 2016. `doi:10.1007/978-3-319-32699-3_11`.

[70] Peter W Singer and Allan Friedman. *Cybersecurity: What everyone needs to know*. oup usa, 2014.

[71] Dominik Stroukal et al. Bitcoin and other cryptocurrency as an instrument of crime in cyberspace. In *Proceedings of Business and Management Conferences*, number 4407036. International Institute of Social and Economic Sciences, 2016.

[72] Samaneh Tajalizadehkhoob, Bram Klievink, Ugur Akyazi, and Nicolas Christin. Plug and Prey ? Measuring the Commoditization of Cybercrime via Online Anonymous Markets Rolf van Wegberg and Samaneh Tajalizadehkhoob , Delft University of Technology ;. (August), 2018.

[73] Gert Jan Van Hardeveld, Craig Webber, and Kieron O'Hara. Discovering credit card fraud methods in online tutorials. *OnSt 2016 - 1st International Workshop on Online Safety, Trust and Fraud Prevention*, 2016. `doi:10.1145/2915368.2915369`.

[74] Rolf Van Wegberg, Jan-Jaap Oerlemans, and Oskar van Deventer. Bitcoin money laundering: mixed results? *Journal of Financial Crime*, 2018.

[75] Rolf van Wegberg, Samaneh Tajalizadehkhoob, Kyle Soska, Ugur Akyazi, Carlos Hernandez Ganan, Bram Klievink, Nicolas Christin, and Michel van Eeten. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1009–1026, Baltimore, MD, August 2018. USENIX Association. URL: `https://www.usenix.org/conference/usenixsecurity18/presentation/van-wegberg`.

[76] V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. URL: `https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034`.

[77] Matthijs J. Warrens. Cohen's kappa is a weighted average. *Statistical Methodology*, 8(6):473–484, nov 2011. `doi:10.1016/j.stamet.2011.06.002`.

[78] Ilia Zaitsev. The Best Format to Save Pandas Data - Towards Data Science. URL: `https://towardsdatascience.com/the-best-format-to-save-pandas-data-414dca023e0d`.

8 Appendix

Appendix A: cohens-calculator.py

```
from sklearn.metrics import cohen_kappa_score
import pandas as pd

andy = pd.read_csv(r"data.csv", usecols=[5])
joe = pd.read_csv(r"data.csv", usecols=[6])

print(cohen_kappa_score(andy, joe))
```

Appendix B: titles-only.py

```
import pandas as pd

# Extracts only titles and thread-id and returns this as a CSV file.

# Only id & subject
raw = pd.read_csv(r"data/source/hacker/security_tutorial_posts_hackers.
   ↪ csv", usecols=[2, 4])
raw[’subject’].replace({’Topic: ’: ’’}, inplace=True, regex=True)
raw[’subject’].replace({’Re: ’: ’’}, inplace=True, regex=True)

# Removes duplicates
raw = raw.drop_duplicates()

# Used for searching and assigning keywords
# raw2 = raw[raw[’subject’].str.contains("bitcoin")]
# raw2.insert(2, ’cat’, ’Cryptography - General’)

# View options
pd.set_option(’display.expand_frame_repr’, False)
pd.set_option(’display.max_rows’, raw.shape[0]+1)
pd.set_option(’display.max_colwidth’, -1)

print(raw)
raw.to_csv("data/full-titles-only-hackers.csv", index=False, header=True)
```

Appendix C: addFileToGT.py

```
import pandas as pd
```

```
# 1. Clears ground truth datafiles (data/gt/(0-34).csv)
# 2. Populates these files with content from the target file.
# Item requires column header: 'Product Name'
# Cat requires column header: 'New Code'

def addToGT(inputFile):
    print("Ripping file to ground truth data...")

    # File to rip data from
    raw = pd.read_csv(inputFile, usecols=[0, 1])
    raw = raw.dropna()

    cats = ["Anonymity [U+FFFD] Other", "Anonymity [U+FFFD] Tor", "
       ↪ Anonymity [U+FFFD] VPN", "Anonymity [U+FFFD] Proxies",
           "Carding", "Cashing Out", "Clearing Criminal History", "
              ↪ Counterfeit Currency", "Cryptocurrency [U+FFFD] General
              ↪ ",
           "Cryptocurrency [U+FFFD] Trading", "Denial of Service", "
              ↪ Digital Forensics", "Doxing", "Drugs [U+FFFD] General",
           "Drugs [U+FFFD] Production", "eBooks [U+FFFD] Other", "eBooks
              ↪ [U+FFFD] Technical", "eWhoring", "Fraud", "Hacking
              ↪ [U+FFFD] General",
           "Hacking [U+FFFD] Malware Supply Chain", "Hacking [U+FFFD]
              ↪ Mobile", "Hacking [U+FFFD] Phreaking", "Hacking [U+FFFD]
              ↪  Website",
           "Hacking [U+FFFD] Wireless Networks", "Lockpicking", "Malware
              ↪ Authorship", "Modifying Credit", "Other",
           "PGP/GPG", "Resources [U+FFFD] Contact Lists", "Resources
              ↪ [U+FFFD] Identity Documents", "SEO", "Transportation/
              ↪ Stealth",
           "Weaponry & Explosives"]

    catNumb = 0

    for cat in cats:
        print(cat)

        outputFile = "data/gt/" + str(catNumb) + "-2.csv"
        output = pd.read_csv(outputFile, usecols=[0, 1])

        # Required to clear the file
        output = output.drop(columns="Product Name")
        output = output.drop(columns="New Code")

        # Add new columns (if required)
        output.insert(0, "Product Name", 0)
        output.insert(1, "New Code", 0)

        newData = raw[raw['New Code'].str.match(cat, case=False)]

        # Merging dataframes
        df_row = pd.concat([newData, output])

        # # View options
        pd.set_option('display.expand_frame_repr', False)
        pd.set_option('display.max_rows', raw.shape[0] + 1)
        pd.set_option('display.max_colwidth', None)

        df_row = df_row.drop_duplicates()
        df_row.to_csv(outputFile, index=False, header=True)

        print("Done...")
        catNumb = catNumb + 1

    print("Completed.")


setPath = "data/gt/full.csv"
addToGT(setPath)
```

Appendix D: keywordAnalysis.py

```
import pandas as pd
import nltk
import re

# Required inputs: 35x separate ground truth data .CSV files in data/gt/
# 1. Function counts the frequency of each word in each file.
# 2. Function outputs 35x separate .CSV files containing this information
    ↪ in data/raw-keywords/

nltk.download('punkt')


def getRawKeywords():

    for corpusId in range(35):
        inputFilename = "data/gt/" + str(corpusId) + "-2.csv"
        raw = pd.read_csv(inputFilename, usecols=[0])

        # Forces lowercase
        raw["Product Name"] = raw["Product Name"].str.lower()

        inputString = raw[raw.columns[0]].to_string()
        amendedString = re.sub(r'\W+', ' ', inputString.lower())
        # This must be a string
        nltk_tokens = nltk.word_tokenize(amendedString)

        # Numbers removed
        for token in nltk_tokens:
            if token.isnumeric():
                nltk_tokens.remove(token)

        # Creates wordlist with no duplicates
        nodupes = set(nltk_tokens)
        finalist = []

        # Counts matches - note: certain words manually discounted.
        for item in sorted(nodupes):
            count = 0

            for token in nltk_tokens:
                if (token == item) & (token != "...") & (token != "'s") &
                    ↪ (token != "how") & (token != "to") & \
                        (token != "make") & (token != "and") & (token != "
                            ↪ the") & (token != "of"):
                    count = count + 1

            inputString = count, item
            finalist.append(inputString)

        column_names = ["freq", "word"]
        df = pd.DataFrame(columns=column_names)

        # Sorting
        # This also removes any one letter words
        for string in sorted(finalist):
            if len(string[1]) > 1:
                modDfObj = df.append({'freq': string[0], 'word': string
                    ↪ [1]}, ignore_index=True)
                df = modDfObj

        outputFilename = "data/raw-keywords/" + str(corpusId) + "-
            ↪ keywords2.csv"
        df.to_csv(outputFilename, index=False, header=True)

    print("Completed.")


getRawKeywords()
```

Appendix E: addLLValues.py

```
import pandas as pd
import math
from signifance import *
from countwords import *
from sumAllWords import *

# Calculates log likelihood scores and significance of those scores.
# Inputs: data/raw-keywords/
# Outputs: data/sig-keywords/


def calculateLL(a, b):
    # a Frequency of word in corpus one.
    # b Frequency of word in corpus two.
    # e Frequency of word in corpus three.

    # c number of words in corpus one.
    # d number of words in corpus two.
    # f number of words in corpus three.

    totalFreqs = sum(a)
    totalWords = sum(b)

    # c * totalFreqs

    step1 = [None] * 35
    i = 0

    while i < 35:
        step1[i] = b[i] * totalFreqs
        i = i + 1

    # i = c * totalFreqs
    # g = d * totalFreqs
    # h = f * totalFreqs

    step2 = [None] * 35
    i = 0

    while i < 35:
        step2[i] = step1[i] / totalWords
        i = i + 1

    # e1 = (i / totalWords)
    # e2 = (g / totalWords)
    # e9 = (h / totalWords)
    step3 = [None] * 35
    i = 0

    while i < 35:
        if a[i] == 0:
            step3[i] = 0
        else:
            step3[i] = math.log(a[i] / step2[i])
        i = i + 1

    # e3 = math.log(a / e1)
    # e4 = math.log(b / e2)
    # e8 = math.log(e / e9)

    step4 = [None] * 35
    i = 0

    while i < 35:
        step4[i] = a[i] * step3[i]
        i = i + 1

    # e5 = (a * e3)
    # e6 = (b * e4)
    # e7 = (e * e8)

    return 2 * (sum(step4))


def simpleCalculateLL(a, b, c, d):
    # a Frequency of word in corpus one.
    # b Frequency of word in corpus two.
    # c number of words in corpus one.
    # d number of words in corpus two.

    e = c * (a + b)
    f = (c + d)

    g = d * (a + b)
    h = (c + d)

    e1 = (e / f)
    e2 = (g / h)

    if (e1 == 0) | (a == 0):
        e3 = 0
    else:
        e3 = math.log(a / e1)
    if (e2 == 0) | (b == 0):
        e4 = 0
    else:
        e4 = math.log(b / e2)

    e5 = (a * e3)
    e6 = (b * e4)

    return 2 * (e5 + e6)


# Frequency of the word in corpus
# wordFreq = [None] * 35

wordFreq = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
   ↪ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

# Total words in that corpus
totalFreq = sumAllWords()

result = simpleCalculateLL(1, 1, 100, 100)

var = {wordFreq[0]: "Anonymity [U+FFFD] Other", wordFreq[1]: "Anonymity
   ↪ [U+FFFD] Tor", wordFreq[2]: "Anonymity [U+FFFD] VPN",
      wordFreq[3]: "Anonymity [U+FFFD] Proxies", wordFreq[4]: "Carding",
         ↪ wordFreq[5]: "Cashing Out",
      wordFreq[6]: "Clearing Criminal History", wordFreq[7]: "Counterfeit
         ↪  Currency", wordFreq[8]: "Cryptocurrency [U+FFFD] "




      wordFreq[9]: "Cryptocurrency [U+FFFD] Trading", wordFreq[10]: "
         ↪ Denial of Service", wordFreq[11]: "Digital Forensics",
      wordFreq[12]: "Doxing", wordFreq[13]: "Drugs [U+FFFD] General",
         ↪ wordFreq[14]: "Drugs [U+FFFD] Production",
      wordFreq[15]: "eBooks [U+FFFD] Other", wordFreq[16]: "eBooks
         ↪ [U+FFFD] Technical", wordFreq[17]: "eWhoring",
      wordFreq[18]: "Fraud", wordFreq[19]: "Hacking [U+FFFD] General",
         ↪ wordFreq[20]: "Hacking [U+FFFD] Malware Supply Chain",
      wordFreq[21]: "Hacking [U+FFFD] Mobile", wordFreq[22]: "Hacking
         ↪ [U+FFFD] Phreaking", wordFreq[23]: "Hacking [U+FFFD] Website
         ↪ ",
      wordFreq[24]: "Hacking [U+FFFD] Wireless Networks", wordFreq[25]: "
         ↪ Lockpicking", wordFreq[26]: "Malware Authorship",
      wordFreq[27]: "Modifying Credit", wordFreq[28]: "Other", wordFreq
         ↪ [29]: "PGP/GPG",
      wordFreq[30]: "Resources [U+FFFD] Contact Lists", wordFreq[31]: "
```

```
           ↪ Resources [U+FFFD] Identity Documents", wordFreq[32]: "SEO",
       wordFreq[33]: "Transportation/Stealth", wordFreq[34]: "Weaponry &
           ↪ Explosives"}

# The higher the G2 value, the more significant is the difference between
    ↪  two frequency scores.
# For these tables, a G2 of 3.8 or higher is significant at the level of
    ↪ p < 0.05 and a G2 of 6.6 or
# higher is significant at p < 0.01.



def addLLvalues():
    counter = 0
    while counter < 35:
        corpusId = counter

        # Enter raw keywords
        fileName = "data/raw-keywords/" + str(corpusId) + "-keywords2.csv"
        print("Calculating LL values for... ", fileName)
        raw = pd.read_csv(fileName, usecols=[0, 1])

        # Create new columns if needed
        raw["exfreq"] = 0
        raw["LL"] = 0
        raw["sig"] = 0

        # # View options
        pd.set_option('display.expand_frame_repr', False)
        pd.set_option('display.max_rows', raw.shape[0] + 1)
        pd.set_option('display.max_colwidth', None)

        sumExcept = 0
        i = 0

        while i < 35:
            if i != corpusId:
                sumExcept = sumExcept + totalFreq[i]
            i = i + 1

        for i in range(len(raw)):
            searchTerm = raw["word"][i]
            exfreq = countWords(searchTerm, corpusId)
            result = simpleCalculateLL(raw["freq"][i], exfreq, totalFreq[
                ↪ corpusId], sumExcept)

            raw["LL"][i] = result
            raw["exfreq"][i] = exfreq
            raw["sig"][i] = checkSignifance(int(result))

        sortedDF = raw.sort_values(by=[’LL’])

        outputFilename = "data/sig-keywords/" + str(corpusId) + "-
            ↪ keywords2.csv"
        sortedDF.to_csv(outputFilename, index=False, header=True)
        counter = counter + 1


addLLvalues()
```

Appendix F: significance.py

```

def checkSignifance(value):
    if value == 0:
        return "Not significant"
    if value < 3.84:
        return "Not significant"
    if (value >= 3.84) & (value < 6.63):
        return 0.05
    if (value >= 6.63) & (value < 10.83):
        return 0.01
    if (value >= 10.83) & (value < 15.13):
        return 0.001
    if value >= 15.13:
        return 0.0001
```

Appendix G: onlySigKeywords.py

```
import pandas as pd


# Trims away any keywords that are not significant (>0.01)
# Intended to make sorting faster and more accurate.
# Inputs: data/sig-keywords/
# Outputs: data/use-keywords/
def getUsableKeywords():
    for corpusId in range(35):
        fileName = "data/sig-keywords/" + str(corpusId) + "-keywords2.csv"
        raw = pd.read_csv(fileName)

        # Convert column to string to prevent errors
        raw[’sig’] = raw[’sig’].apply(str)

        new = raw[raw[’sig’].str.contains(’0.01|0.001|0.0001’, case=False)
            ↪ ]

        # View options
        pd.set_option('display.expand_frame_repr', False)
        pd.set_option('display.max_rows', raw.shape[0] + 1)
        pd.set_option('display.max_colwidth', None)
        print(new)

        outputFile = "data/use-keywords/" + str(corpusId) + "-keywords2.
            ↪ csv"
        new.to_csv(outputFile, index=False, header=True)


getUsableKeywords()
```

Appendix H: catKeyword.py

```
import pandas as pd
import nltk
import re
from getLLscore import tokeniseString, getLabel



# Uses the data from prior keyword analysis to sort items.
# Calculates a total LLscore ('LLSum') for each entry and assigns a
    ↪ category ('AutoCat') based on this.

def phraseCheck(inputString):
    if inputString.lower().find("no carding|not carding") > -1:
        return "Fraud"
    elif inputString.lower().find("disappear and live free") > -1:
        return "Anonymity [U+FFFD] Other"
    elif inputString.lower().find("anarchist cookbook") > -1:
        return "Weaponry & Explosives"
    elif inputString.lower().find("robert greene|chomsky|karl marx|
        ↪ aleister crowley|1984") > -1:
        return "eBooks [U+FFFD] Other"
    elif inputString.lower().find("e whore") > -1:
        return "eWhoring"
    elif inputString.lower().find("clear your criminal") > -1:
        return "Clearing Criminal History"
    elif inputString.lower().find("gpg4win") > -1:
        return "PGP/GPG"
    elif inputString.lower().find("wifi hacking") > -1:
        return"Hacking [U+FFFD] Wireless Networks"
    elif inputString.lower().find("hacking wireless") > -1:
        return "Hacking [U+FFFD] Wireless Networks"
    elif inputString.lower().find("how to create a virus") > -1:
        return "Malware Authorship"
    elif inputString.lower().find("seo") > -1:
        return "SEO"
    elif inputString.lower().find("dynamite") > -1:
        return "Weaponry & Explosives"
    elif inputString.lower().find("thermite") > -1:
        return "Weaponry & Explosives"
    elif inputString.lower().find("molotov cocktail") > -1:
        return "Weaponry & Explosives"
    elif inputString.lower().find("havij") > -1:
        return "Hacking [U+FFFD] Website"
    elif inputString.lower().find("psilocybe cubensis") > -1:
        return "Drugs [U+FFFD] Production"
    elif inputString.lower().find("nitrous oxide") > -1:
        return "Drugs [U+FFFD] General"
    elif inputString.lower().find("meth manufactur") > -1:
        return "Drugs [U+FFFD] Production"
    elif inputString.lower().find("lsd manufactur") > -1:
        return "Drugs [U+FFFD] Production"
    elif inputString.lower().find("brewing") > -1:
        return "Drugs [U+FFFD] Production"
    elif inputString.lower().find("hydroponics") > -1:
        return "Drugs [U+FFFD] Production"
    elif inputString.lower().find("shotgun") > -1:
        return "Weaponry & Explosives"
    elif inputString.lower().find("c++") > -1:
        return "eBooks [U+FFFD] Technical"
    elif inputString.lower().find("cardable website") > -1:
        return "Carding"
    elif inputString.lower().find("skimmer") > -1:
        return "Carding"
    elif inputString.lower().find("bitcoin mixer") > -1:
        return "Cashing Out"
    elif inputString.lower().find("bitcoin tumbler") > -1:
        return "Cashing Out"
    elif inputString.lower().find("money laundering") > -1:
        return "Cashing Out"
    elif inputString.lower().find("bitcoin blender") > -1:
        return "Cashing Out"
    elif inputString.lower().find("bitcoinfog|bitfog") > -1:
        return "Cashing Out"
    elif inputString.lower().find("p o box") > -1:
        return "Transportation/Stealth"
    elif inputString.lower().find("counterfeit") > -1:
        return "Counterfeit Currency"
    elif inputString.lower().find("how to be invisible") > -1:
        return "Anonymity [U+FFFD] Other"
    else:
        return 0


def sortData(inputFile):
    raw = pd.read_csv(inputFile, usecols=[0, 1])

    # Convert column to string to prevent errors
    raw['Product Name'] = raw['Product Name'].apply(str)

    # Add new columns (if required)
    raw.insert(2, "LLSum", 0)
    raw.insert(3, "AutoCat", 0)

    lines = int(raw.shape[0])

    for i in range(len(raw)):
        temp = phraseCheck(raw["Product Name"][i])

        if temp != 0:
            raw["AutoCat"][i] = temp
        else:
            searchString = raw["Product Name"][i].lower()
            amendedString = re.sub(r'\W+', ' ', searchString.lower())
            array = tokeniseString(amendedString)

            # Progress bar
            print(i, "--", i / lines, "%")

            raw["AutoCat"][i] = getLabel(array)
            raw["LLSum"][i] = max(array)

    raw.to_csv(inputFile, index=False, header=True)


sortData("data/input.csv")
```

Appendix I: countwords.py

```
import pandas as pd

# Counts the freq of a word in every corpus except the target corpus



def countWords(word, corpusId):
    i = 0
    count = 0
    while i < 34:
        if i != corpusId:
            fileName = "data/raw-keywords/" + str(i) + "-keywords2.csv"
            raw = pd.read_csv(fileName, usecols=[0, 1])

            for row in range(len(raw)):
                keyword = str(raw["word"][row])
                string = keyword

                if word == string:
                    count = count + raw["freq"][row]

        i = i + 1

    # Debug printing
    # print(corpusId, word, "Exc: ", count)
    return count
```

Appendix J: k-analysis.py

```
import pandas as pd
from addFileToGT import addToGT
from keywordAnalysis import getRawKeywords
from addLLValues import addLLvalues
from catKeyword import sortData
from onlySigKeywords import getUsableKeywords
from sklearn.metrics import cohen_kappa_score
from splitFile import shuffleAndSplit


def createKSets():
    countdown = 10

    while countdown > 0:
        templateLink = "data/eval/template.csv"
        combined = pd.read_csv(templateLink, usecols=[0, 1])

        for i in range(10):
            if i != countdown - 1:
                inputFile = "data/eval/k" + str(i + 1) + ".csv"
                nextFrame = pd.read_csv(inputFile, usecols=[0, 1])
                combined = pd.concat([combined, nextFrame])

        outputFile = "data/eval/kset" + str(countdown) + ".csv"
        combined.to_csv(outputFile, index=False, header=True)
        countdown = countdown - 1


# Check K sets hold at least one of each category.
def validateKSets():
    cats = ["Anonymity [U+FFFD] Other", "Anonymity [U+FFFD] Tor", "
        ↪ Anonymity [U+FFFD] VPN", "Anonymity [U+FFFD] Proxies",
            "Carding", "Cashing Out", "Clearing Criminal History", "
                ↪ Counterfeit Currency", "Cryptocurrency [U+FFFD] General
                ↪ ",
            "Cryptocurrency [U+FFFD] Trading", "Denial of Service", "
```

```
              ↪ Digital Forensics", "Doxing", "Drugs [U+FFFD] General",
            "Drugs [U+FFFD] Production", "eBooks [U+FFFD] Other", "eBooks
              ↪ [U+FFFD] Technical", "eWhoring", "Fraud", "Hacking
              ↪ [U+FFFD] General",
            "Hacking [U+FFFD] Malware Supply Chain", "Hacking [U+FFFD]
              ↪ Mobile", "Hacking [U+FFFD] Phreaking", "Hacking [U+FFFD]
              ↪  Website",
            "Hacking [U+FFFD] Wireless Networks", "Lockpicking", "Malware
              ↪ Authorship", "Modifying Credit", "Other",
            "PGP/GPG", "Resources [U+FFFD] Contact Lists", "Resources
              ↪ [U+FFFD] Identity Documents", "SEO", "Transportation/
              ↪ Stealth",
            "Weaponry & Explosives"]

    for i in range(10):
        inputFile = "data/eval/kset" + str(i+1) + ".csv"
        raw = pd.read_csv(inputFile, usecols=[1])

        for cat in cats:
            count = 0

            for k in range(len(raw)):
                if raw["New Code"][k] == cat:
                    # print(raw["New Code"][k], cat)
                    count = count + 1

            if count == 0:
                return False

    return True


def addFileToModel(outputFile):
    addToGT(outputFile)
    print("Getting raw keywords for...", outputFile)
    getRawKeywords()
    print("Getting LL values for...", outputFile)
    addLLvalues()
    getUsableKeywords()


def calcCK():
    avg = 0
    for i in range(10):
        samplePath = "data/eval/k" + str(i + 1) + ".csv"
        andy = pd.read_csv(samplePath, usecols=[1])
        auto = pd.read_csv(samplePath, usecols=[3])
        avg = avg + cohen_kappa_score(andy, auto)
        print("CK score for: ", i + 1, " - ", cohen_kappa_score(andy, auto
            ↪ ))

    print("\nOverall: ", avg / 10)


# 1. Shuffles the main dataset then splits it into x10 subsets (labelled
    ↪ k1 to k10)
shuffleAndSplit()

# 2. Creates 10x subsets (comprised of every set except 1)
# Required for next step.
createKSets()

# 3. Subsets are validated to ensure each has at least one item of every
    ↪ category.
# Implemented this as without it, the model can't deal with it.
# This is only really and issue with small datasets, not the full dataset
    ↪ .
check = 0

while check == 0:
    if validateKSets():
        print("Ksets validated.")
        check = check + 1
    else:
        shuffleAndSplit()
        createKSets()
        print("Re-validating...")

# 4. Gets keywords, LL weightings and then sorts the relevant k sample.
# This is repeated ten times.
for i in range(10):
    setPath = "data/eval/kset" + str(i + 1) + ".csv"
    print("Adding ", setPath, "to ground truth data...")
    addFileToModel(setPath)

    kSampleToSort = "data/eval/k" + str(i + 1) + ".csv"
    print("Sorting ", kSampleToSort, "...")
    sortData(kSampleToSort)

# 5. Final step: Calculates & prints cohen's kappa statistics
calcCK()
```

Appendix I: splitFile.py

```
import pandas as pd

# Creates 10 equally sized k samples from the main sorted dataset for
```

```
    ↪ crossvalidation
# Input: Full set of SORTED gt data - location should not change (data/gt
    ↪ /full.csv)


def shuffleAndSplit():
    raw = pd.read_csv(r"data/gt/full.csv", usecols=[0, 1])

    # Duplicates removed
    raw.drop_duplicates()
    raw['New Code'] = raw['New Code'].str.strip()

    # Shuffles set
    shuffled = raw.sample(frac=1)

    # Gets number of rows
    numOfRows = shuffled.shape[0]
    print('Number of Rows in dataframe : ', numOfRows)

    start = 0
    subSection = int(numOfRows / 10)
    counter = subSection
    fileId = 1

    while counter <= numOfRows:
        print(start, counter, subSection)
        print(shuffled[start:counter])

        outputFile = "data/eval/k" + str(fileId) + ".csv"
        fileId = fileId + 1
        shuffled[start:counter].to_csv(outputFile, index=False, header=
            ↪ True)

        start = start + subSection
        counter = counter + subSection
```

Appendix J: getLLscore.py

```
import pandas as pd
import nltk

nltk.download('punkt')


def getTotalLLScore(searchTerm):
    values = [0] * 35

    for i in range(35):
        values[i] = getLLscore(searchTerm, i)
    return values


def getLLscore(searchTerm, corpusId):
    fileName = "data/use-keywords/" + str(corpusId) + "-keywords2.csv"
    raw = pd.read_csv(fileName, usecols=[0, 1, 2, 3, 4])

    for i in range(len(raw)):

        if raw["word"][i] == searchTerm:
            value = raw["LL"][i]
            return value

    return 0


# Helper function - checks if not significant keywords were detected.
# Without this function to check, it kept auto-assigning these to cat34.
def allValuesZero(sumAll):
    count = 0
    for i in range(35):
        if sumAll[0] == 0:
            count = count + 1
    if count == 34:
        return True
    else:
        return False


def getLabel(sumAll):
    # Enter threshold value here. 15 = p < 0.0001
    # Any LLsum below this will be set to Other.
    threshold = 15
    if max(sumAll) < threshold:
        return "Other"
    elif allValuesZero(sumAll):
        return "Other"
    else:
        var = {sumAll[0]: "Anonymity [U+FFFD] Other", sumAll[1]: "
            ↪ Anonymity [U+FFFD] Tor", sumAll[2]: "Anonymity [U+FFFD] VPN
            ↪ ",
              sumAll[3]: "Anonymity [U+FFFD] Proxies", sumAll[4]: "
                 ↪ Carding", sumAll[5]: "Cashing Out",
              sumAll[6]: "Clearing Criminal History", sumAll[7]: "
                 ↪ Counterfeit Currency", sumAll[8]: "Cryptocurrency
                 ↪ [U+FFFD] "
```

```
               sumAll[9]: "Cryptocurrency [U+FFFD] Trading", sumAll[10]: "
                  ↪ Denial of Service", sumAll[11]: "Digital Forensics",
               sumAll[12]: "Doxing", sumAll[13]: "Drugs [U+FFFD] General",
                  ↪  sumAll[14]: "Drugs [U+FFFD] Production",
               sumAll[15]: "eBooks [U+FFFD] Other", sumAll[16]: "eBooks
                  ↪ [U+FFFD] Technical", sumAll[17]: "eWhoring",
               sumAll[18]: "Fraud", sumAll[19]: "Hacking [U+FFFD] General
                  ↪ ", sumAll[20]: "Hacking [U+FFFD] Malware Supply Chain
                  ↪ ",
               sumAll[21]: "Hacking [U+FFFD] Mobile", sumAll[22]: "Hacking
                  ↪  [U+FFFD] Phreaking", sumAll[23]: "Hacking [U+FFFD]
                  ↪ Website",
               sumAll[24]: "Hacking [U+FFFD] Wireless Networks", sumAll
                  ↪ [25]: "Lockpicking", sumAll[26]: "Malware Authorship
                  ↪ ",
               sumAll[27]: "Modifying Credit", sumAll[28]: "Other", sumAll
                  ↪ [29]: "PGP/GPG",
               sumAll[30]: "Resources [U+FFFD] Contact Lists", sumAll[31]:
                  ↪  "Resources [U+FFFD] Identity Documents", sumAll[32]:
                  ↪  "SEO",
               sumAll[33]: "Transportation/Stealth", sumAll[34]: "Weaponry
                  ↪  & Explosives"}

    return var.get(max(var))


def tokeniseString(string):
    nltk_tokens = nltk.word_tokenize(string)
    sumAll = [0] * 35

    for token in nltk_tokens:
        result = getTotalLLScore(token)
        print(token, result)

        for i in range(35):
            sumAll[i] = sumAll[i] + result[i]

            # Debug printing
            # print(sumAll)
            # print(token, max(sumAll), getLabel(sumAll))
    return sumAll
```