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 budding 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 [48].

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) in order to determine the areas with least and most emphasis. The maps were also evaluated against similar maps of formal cyber security qualifications [28] in order to identify mismatches between their curriculums and the available cyber criminal learning material.

A keyword classifier program was built and utilised in order to categorise the full data set within the required timescale.

To achieve these aims the project implemented the following:

- A codebook defining the key categories present within the darknet learning material.
- A classifier program that can automatically assign categories to 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 [28] to determine potential areas of deficit.

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 datasets 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.

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 [45].

Within the last decade a number of dark web markets and forums have emerged where goods and services are traded anonymously [48]. These websites facilitate the trade of a wide range of illegal products and services. These include drugs and weapons but also a wide range of cyber criminal learning material [CITE].

Dark web marketplaces operate and appear much like any online marketplace on the open internet.

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 2011–2017 [50].

This includes criminal learning material such as guides and tutorials that cyber criminals can use to improve their trade.

Of concern for cyber security professionals, research suggests that many of these guides, if they are meticulously followed are effective, leaving limited interceptive opportunities for law enforcement [49].

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.

While previous research has already demonstrated the presence of cyber criminal learning material, mapping 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.

These illegal services are also discussed on dark web forums and much like the markets, these can be of considerable scale with certain forums containing thousands of posts [8].

Previous research in the content of cyber security qualifications has revealed that the

focus of these qualifications differ considerably [28]. In fact, some of these qualifications are heavily weighted towards only a few areas within cyber security and some areas are greatly under-represented across all qualifications.

Using the same benchmark, the mappings created through this research will be directly contrasted with that of the cyber security curriculums, to determine whether these are effectively covering the material that their cyber criminal adversaries are learning on the dark web.

This project also intends to identify emerging cyber-criminal trends on the darknet by analysing the data on a yearly basis. In addition, the study will produce a program that can be improved upon or re-purposed to provide more detailed and up-to-date mapping of the dark web. For example, integration with an automated web scraper to provide real-time updates on dark web trends.

Data representations, in the form of spider maps and bar charts, will be created in order to summarise our findings and provide a direct comparison with the representations created from research into cyber security curriculums [28].

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 curricula [28] as well as prior research into the dark web we can evaluate the effectiveness of the current state of cyber security education in addressing these threats.

More generally, this project also intends 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.

In order to achieve this aim, the project intends to complete the following objectives:

1. Carry out analysis of the source data to determine the most frequently occurring topics present within the data set. The aim was to manually categorise at least 3000 market listings that would form the ground truth data set. This would act as the initial training data set for the classifier.

2. Once a general understanding of common themes has been established, a codebook defining these key categories will be created and validated, formulated from the manually categorised ground truth data.

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 work published in each area.

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 [26]. Some erroneously conflate it with the deep web, 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.

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," [26].

The term dark net, though often conflated with the dark web, refers to the collection of networks and technologies used to share digital content on the dark web [21]. 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 [cite FOSS].

The deep web, comprising content not indexed by a search engine (excluding the few niche search engines which specialise in indexing the dark web), is substantial in size with recent estimates [46] 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 [26]. The dark web however constitutes only a tiny fraction of this but it is growing rapidly. A 2017 report [37] identified only around 4,000 dark web sites. Two years later another report [19] found there were 8,416 active dark web domains, more than doubling in size.

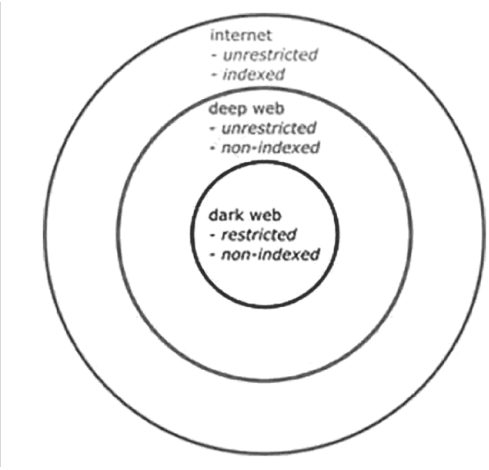

Figure 2.1: Graphical representation of the relationship between the internet, the deep web and the dark web [21].

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 [10]. 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 can not find out who is running the service, nor can they shut it down.

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. 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 [33]. 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 [23].

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 [33]. 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 [14]. Notably, this makes the dark web particularly useful for cyber criminals, who are constantly trying to hide their tracks [39] and now has the potential to host an increasingly high number of malicious services and activities.

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 [33](cite original). Many studies have sought to map the dark web, typically through a particular lens or focusing upon particular crime types such as terrorism [12]. One such study analysed a corpus of over 4000 unique dark web pages to determine the motivations of users on the dark web [20]. 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, 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, research has indicated that the proportion of the dark web dedicated to illegal content is increasing steadily. An initial 2016 study indicated that only 28% of dark web sites hosted illicit material. However, a follow up 2019 study discovered the number of dark web listings that hosted illicit content had risen by 20% since the previous study (cite, p128, maz). Notably the rate of criminal adoption is far outpacing that of other users, such as privacy activists (cite, p130, maz). 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 [33]. 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 [24]. This research demonstrates the severe threat that can be posed by cyber criminal communities within the dark web.

Dark web hackers are concerned with a wide range 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 [35]. In particularly high demand are 'zero-day' exploits which are bugs that have not yet been discovered and fixed by software and antivirus companies [33]. Zero-day exploits enable particularly stealthy and sophisticated attacks against specific targets, giving rise to what security researchers term APTs, advanced persistent threats.

These pose a significant threat to both companies and governments [27]. 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 [41].

The widespread accessibility of these exploits, that the dark web could enable, is particularly dangerous as the likelihood of being detected in one of these attacks is effec-

tively nil.

Dark web sites also facilitate the trading of proprietary information, particularly lists of stolen usernames and passwords known colloquially as 'combo lists'. These can be for a wide range of websites and many are often of significant scale. Organisations as large and diverse as Exactis (230 million customers), Dixons Carphone (5.9 million), Marriott International (383 million), Quora (100 million) and Cathay Pacific (9.4 million) have all suffered major data breaches [5]. 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 [5].

One widely reported list, dubbed 'Collection 1-5', featured over 2.2 billion unique usernames and passwords [5]. 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.

Money laundering, often involving the use of cryptocurrencies, and the trade in credit card information — commonly referred to as carding — also feature prominently [48]. A study even identified numerous dark web websites which specialise solely in these activities [14]. Bitcoin in particular can be easily laundered through unregulated exchanges which avoid identity checks.

Bitcoin is a decentralised digital currency heavily utilised on the dark web (cite?). 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 [2]. These characteristics make Bitcoin a particularly appropriate vehicle for money laundering.

Responding to these new developments, police forces around the world have made advances and devised systems of flagging Bitcoin transactions linked to illegal activity. However, to counter this, money launderers are now turning in droves to Bitcoin mixers known as tumblers. 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 [33]. As an anonymous cryptocurrency is essential to dark web trade, the emergence and widespread adoption of a highly anonymous cryptocurrency would 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 [15], which act as anonymous marketplaces selling everything 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 [33]. User reviews and ratings help to maintain trust among anonymous users and the overall integrity of the market [5].

These markets not only facilitate the trade of illegal drugs and weapons but also a wide range of cyber crime learning material that cyber criminals can use to improve their skills. These markets and forums are not a niche concern. Between 2011 and 2013, the Silk Road market alone processed more than 1.2 billion dollars' worth of transactions [33]. The combination of the readily available software, the prominence of dark web sites and the use of cryptocurrency to facilitate anonymous transactions is the perfect storm for cyber crime.

Transactions are anonymised through the use of cryptocurrency, typically Bitcoin, and communications between buyer and seller are encrypted using software such as PGP. The dark web economy is highly decentralised and adaptable. 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 [38].

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 [9]. This study found that the majority of listings on the site were for drug products, almost half (45%) of the total market. This was followed by listings relating to fraud (13%), such as fake ids, with all other categories representing only a 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 supporting online conversations. They capture each conversation in a "thread" and the ensuing postings are usually time-stamped and attributable to a particular online poster [12]. 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 [8]. To demonstrate this, one such dark web forum included in our study, called Crackingfire has been determined to have approximately 14,511 active users [32].

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. 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 [47].

Prior research into dark web forums has focused on terrorist and extremist groups [12]. One study analysed dark web forum posts in order to seek out users who were starting to show radical tendencies [40]. In order to achieve this, the researchers utilised a part-of-speech tagger to isolate keywords and nouns which are 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 [6]. 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 seem 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 article by Chertoff discussed the effects of law enforcement taking down two different dark web sites: Playpen and Silk Road [13].

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 [4].

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, afterwards 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

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 payoff 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 [5, 16]. One such study conducted by Ciancaglini et al. [16] 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 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 [5].

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. 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.

Some research has been conducted into the dark web in order to categorise and analyse its contents. In a study not dissimilar to ours, Faizan & Khan manually assessed and categorised 3,480 English language dark web sites in order to establish [22].

The content was spread across a wide range of categories with no one category comprising more than 10% of websites. Only 38% were for illicit purposes. 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%). The marketplace and forum categories will also likely contain a variety of cyber criminal materials as well.

2.3 Cyber security qualifications

Cyber security qualifications intend to equip their graduates with the skills 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 [31]. It is therefore essential that the state of cyber security education adapts to reflect this new reality.

Previous research has clearly indicates that there is a skills shortage within the cyber security industry [25]. The current and future demand for cyber-security skills looks likely to be outstripping supply. Almost half of UK organisations feel they lack the in-house skills required to handle the cyber threats they are facing [1].

The UK National Cyber Security Strategy identified multiple contributing factors that

need to be addressed to close this gap. 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 [3].

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 [28]. 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 qualification (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 or as this study seeks to determine, they are proportionate to cyber criminal threats posed by hackers.

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 is also does not include every available cyber security qualification. Even with these limitations in mind, the study does provide a good general overview of 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 to ensure they are effectively countering the methodology that cyber criminals are being taught 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 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 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 [42]. It intends to codify 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 [28]. By using a shared benchmark, this allows our study to provide a direct comparison with the mappings created by this study.

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 dark web learning material.

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.

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 [11]. This data set contains 99118 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 [8]. This data set comprises 19781 posts drawn from the CrackingFire and CrackingArena forums. The

CrackingFire forum data set dates from 4/7/2011 – 2/21/2018 and the CrackingArena data set dates from 4/8/2013 – 2/24/2018.

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 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.

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 be 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 net 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 (pg numb) earlier, dark web forums and markets are 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 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 on the dark web.

3.2.1 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 I began by manually categorising market listings directly to CyBOK knowledge areas however it soon became clear this was not accurately reflecting the nature or the range of content available. 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. Additionally, 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.

In order to create this codebook, a substantial portion (over 2000 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.

3.2.2 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 anonymity focused operating systems such as TAILs or Whonix.

2. **Anonymity – Tor**
   Item relating to the TOR (The Onion Router) web browser.

   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 by 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, 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.

3.2.3 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 Cohen's kappa calculated to ensure inter-rater reliability. 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 [17] is a popular descriptive statistic for summarizing an agreement table [52]. 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 [34].

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]

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.

3.3 Creation of a Classifier

A classifier 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.

3.3.1 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 would be the most effective means of categorising the remaining source data. In ad-

dition, 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

The classifier works by 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 [43] 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: 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 where necessarily.

This methodology has been used to good effect in numerous similar studies. One such study analysed the messages on the micro blogging website Twitter [30]. 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 [36]. 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.

3.3.2 Training the Classifier

Before the classifier can be used to categorise items, it must be trained on a sorted data set. The classifier will extract and then load the relevant 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.

2. The full ground truth data set is tokenised and the frequency of each token within that corpus is calculated. 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. Significance is also calculated 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. 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 does not need to be trained again unless the ground truth data set changes.

3.3.3 Using the Classifier

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

1. Initially the classifier checks for the presence of certain key phrases and assigns them to the appropriate category. These are typically phrases that include a token or tokens that 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 category.

3. These scores are then summed for each category and returned in a final array and the category with the highest total score is then assigned to the item, provided it is over a threshold level [GET AND EXPLAIN]. If this score is below the threshold value, that is contains no relevant keywords for any category, the item is assigned to 'Other' as the default category.

3.4 Validation of Classifier

Validation techniques were utilised to ensure the constructed classifier was accurately assigning categories to the input items.

3.4.1 K-fold cross validation

The primary method of validation utilised was k-fold cross validation [51]. The k-fold cross validation technique is one of the most used approaches by practitioners for model selection and error estimation of classifiers [7].

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 [44]. 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 [7], 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 as it needs to only run k times which is usually 5 or 10. 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 summary, k-fold cross validation was appropriate for the input data set, practical to implement with the available tools and within the required timescale.

Once the method of cross-validation was decided, 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.

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 would occasionally would result in a very small number of items 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, all k-samples (1 to 10) are 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' [CITE]

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.

3.4.2 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 looking at discrepancies between the total number of items assigned to each category manually and by the classifier, attempting to identify areas that were being over or under assigned. Items in the categories with the largest discrepancy were then manually scanned to identify similarities which could lead to the root cause of this discrepancy if there was any.

The root cause of the discrepancy was typically an over or under weighted keyword and was remedied by either adding relevant entries to the ground truth data or by adding to the list of keywords and phrases that are detected by the classifier.

For example, many of the thread titles related to Bitcoin mixing (also known as tumbling) was erroneously being categorised as 'Cryptocurrency – General'. I therefore added a few relevant phrases in the initial checking stages to ensure these items were categorised correctly as 'Cashing Out'.

4 Results and Analysis

4.1 Results

The remaining uncategorised items were sorted by the classifier and merged with the manually categorised items.

4.1.1 Raw results

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

4.1.2 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. Categories that were not related to cyber security (such as 'Drugs') were not included in this list and were not assigned a knowledge area.

4.2 Data Representation

The spider maps and bar charts were created 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 Mirror, Mirror study. The output PDF files containing the graphs were then trimmed using online tools to remove unnecessary white space.

4.2.1 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 Mirror, Mirror study.

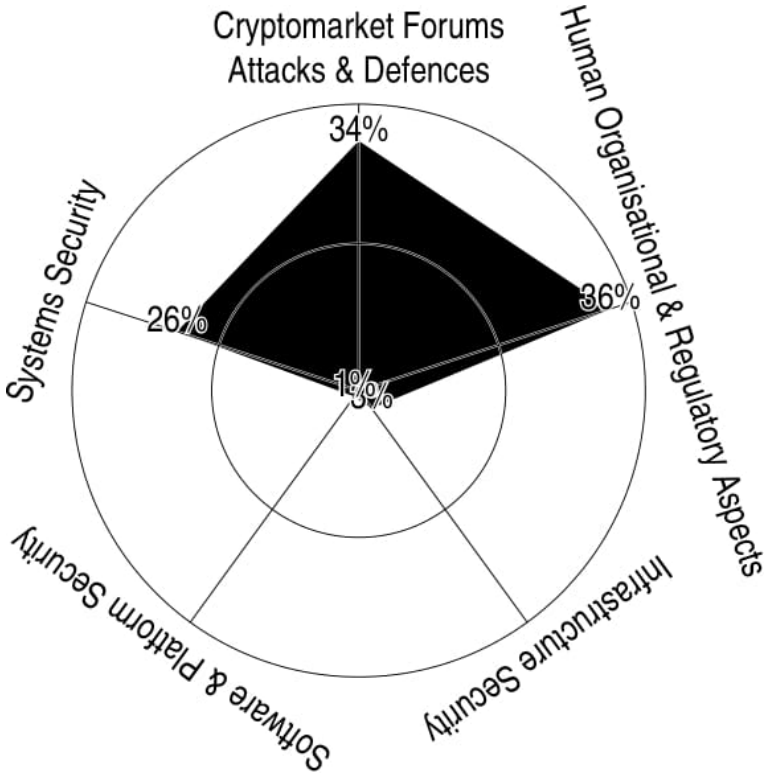

Figure 4.1: Cryptomarket forums

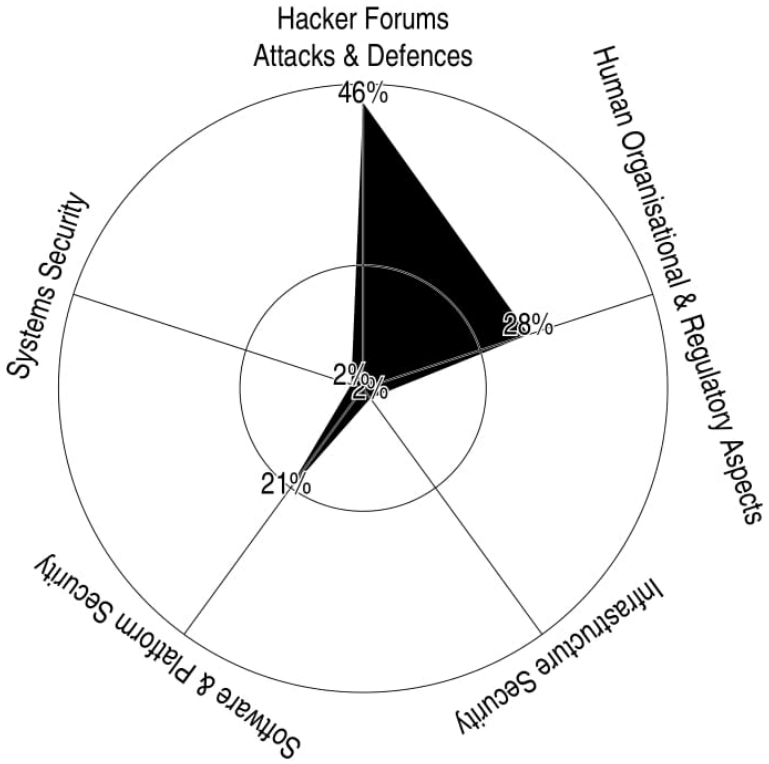

Figure 4.2: Hacker forums

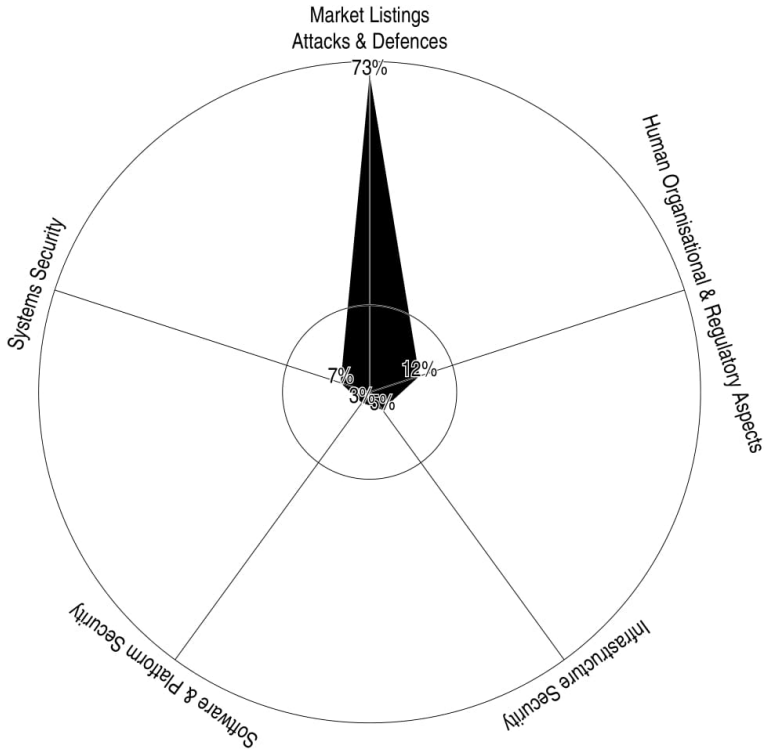

Figure 4.3: Market listings

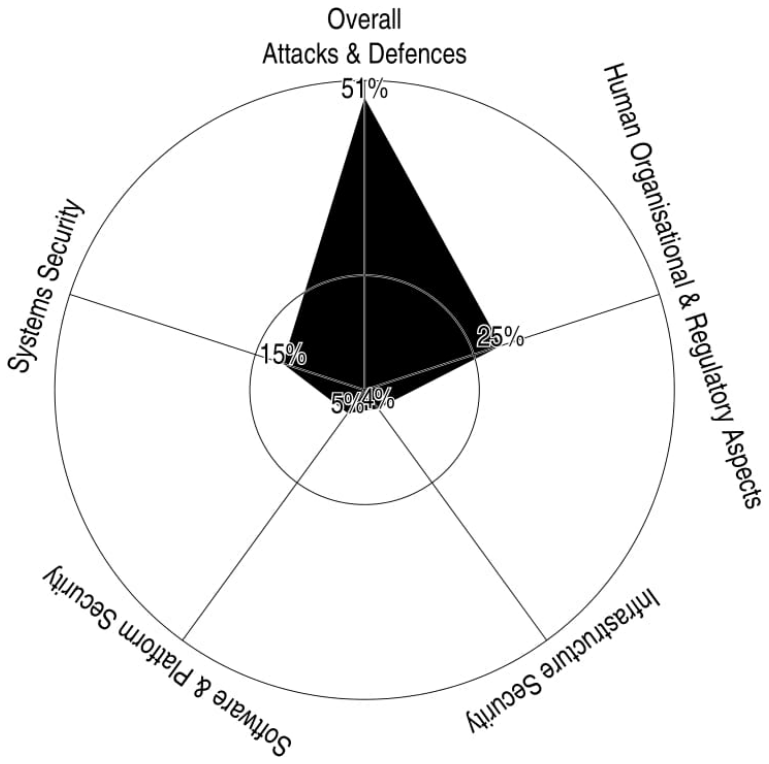

Figure 4.4: Overall

4.2.2 Bar Charts

Bar charts were created to split the results by individual CyBOK knowledge area. It should be noted that the following knowledge areas (LIST) were not covered by the classifier and are therefore set to zero percent.

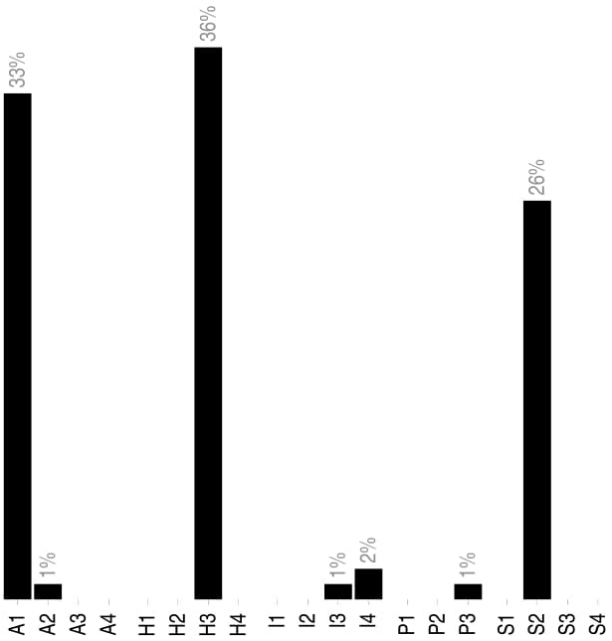

Figure 4.5: Cryptomarket forums

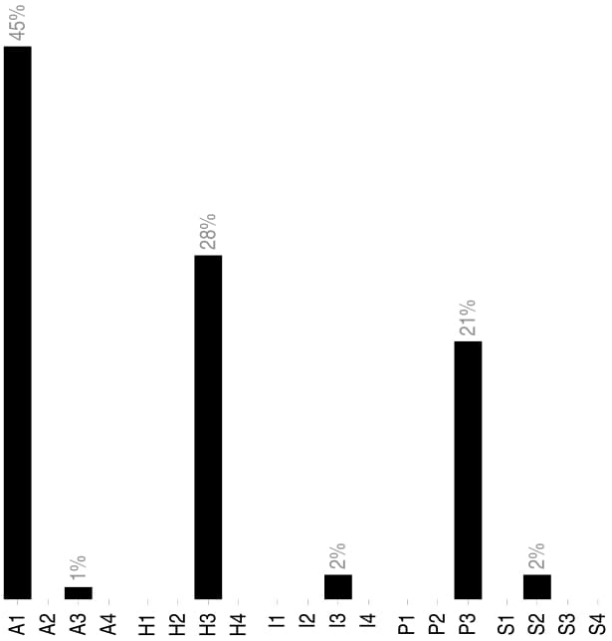

Figure 4.6: Hacker forums

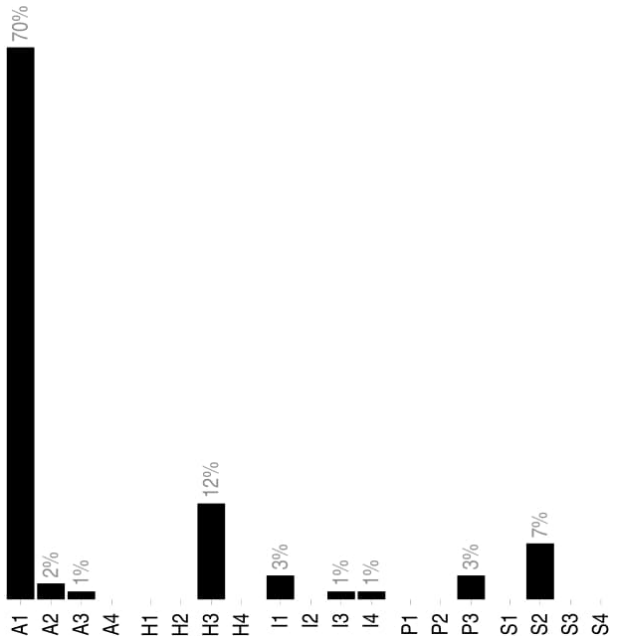

Figure 4.7: Market listings

A1	Adversarial Behaviours	H1	Human Factors	I1	Cyber-Physical Systems Security
A2	Forensics	H2	Law & Regulation	I2	Hardware Security
A3	Malware & Attack Technologies	H3	Privacy & Online Rights	I3	Network Security
A4	Security Operations & Incident Management	H4	Risk Management & Governance	I4	Physical Layer Security

S1	Authentication, Authorisation & Accountability	P1	Secure Software Design & Development
S2	Cryptography	P2	Software Security
S3	Distributed Systems Security	P3	Web & Mobile Security
S4	Operating Systems & Virtualisation Security		

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. The findings will also be used to evaluate the state of current cyber security qualifications.

4.3.1 Against Previous Research

The findings of this study broadly support many of the conclusions derived from previous dark web research.

While the study did not test a fixed hypothesis, it was speculated in the research proposal that based on previous research 'Adversarial Behaviour' would be the most prominent knowledge area and this did prove to be correct. The basis of this prediction was a study that indicated a significant percentage of illicit content on the dark web was related to credit card fraud and money laundering [48].

The sale of both drugs [CITE] and firearms [18] on the dark web have been widely

documented and our findings continue to support this conclusion. Our study found many instances of these particularly within the market listings data.

Software exploits

Studies have documented the trade in software exploits within the dark web as well as the potential for these exploits to pose a serious threat to organisations and governments [CITE].

Our findings do not support the idea that this is particularly widespread. 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, I looked at the number of items within these categories on a yearly basis.

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 4.2: 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

One reassuring aspect of the above table from a cyber security perspective it does not suggest the demand for software exploits is growing. If this was the case, it would be expected that the number of items relating to software exploits would be rising consistently. 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 are less technical and therefore yield a faster payoff. There is certainly a demand for these software exploits but this is largely restricted to a minority of users within the hacker forums, though there may be other dark web sites or communities, not included within this study, where they are shared.

User motivation

Previous research has indicated that users of the dark web are far more driven by lust and greed than they are by other motives, such as politics [20]. The findings of this study strongly support this conclusion, particularly the greed motive, with many of the most prominent categories referring to fraud and money laundering. There were almost no instances of items related to terrorism or extremism. 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 [CITE].

Some other findings from this study were not supported by our results. For example, it detected a small percentage (5%) of websites dedicated to providing education

and training for child exploitation, particularly in terms of grooming vulnerable children. The results of this study did not find any items at all which fit this description, indicating that users looking for these guides are confined to niche websites.

Political extremists and child predators, whose presence has been documented on the dark web [12, 20], 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 completely.

The categories with the highest prominence are concerned with methods of 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 cleaning that fraudulently obtained money ('Cashing Out') are provided and lastly, techniques to stay anonymous while performing 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, rather than political or other emotional motives such as revenge.

Scope of CyBOK

In the research proposal I set out, as a research question, 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 cannot 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 caught by the study. It is possible there is content within the wider dark web that is not yet covered by CyBOK.

Drugs

The existence of the trade in illicit substances, particularly within cryptomarkets, has long been documented on the dark web [CITE]

From a more broader criminal perspective, the prevalence of drug related guides within the crypto-markets listings and forums is interesting as it suggests darknet markets are not only being used to trade illicit drugs (a practice which is well documented [CITE]) but also methodology for production. This will likely exacerbate the drug problem as guides for difficult to manufacture drugs (such as LSD) become more widely available.

4.3.2 Differences between hacker forums, cryptomarket forums and cryptomarket listings

As alluded to in earlier sections, there were very significant differences in the results for each of these data types. The categories that dominated each data set differed significantly, suggesting each set of websites attracts its own separate set userbase with differing intentions.

The market listings were generally very homogeneous and consisted largely of technical e-books following by carding and cashing out. A potential reason for the prevalence of e-books is the typically expensive cost of purchasing these legitimately [29] which is driving students to darknet markets in order to acquire these materials at a fraction of the price.

The forum titles displayed more variety. The cryptomarket forums were primarily focused on Transportation/Stealth, Cashing Out, PGP/GPG and Anonymity.

The hacker forum data set is focused on primarily on proxies and hacking

4.3.3 Comparison over time

The data sets were split by year and analysed with the intention of detecting any emerging trends.

Listings

The market listing data is not dated thereby making analysis over time for this data set impossible.

However previous research into market listings has determined that darknet markets have grown considerably since their inception and will likely continue to do so.

Cryptomarket Forums

Knowledge Area	**2012**	**2013**	**2014**	**2015***
Attacks & Defences	10	250	1013	381
Systems Security	8	416	639	195
Infrastructure Security	1	14	102	47
Human, Organisational & Regulatory Aspects	9	364	1053	335
Software & Platform Security	0	3	31	6
Totals	**28**	**1047**	**2838**	**964**

Table 4.3: Fig X. Number of titles on the cryptomarket forums for each year within each knowledge area
*until July

The table shows a significant increase in titles year on year, supporting previous research suggesting that the user base of the darknet will continue to grow in size.

Hacker Forums

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

Table 4.4: Fig X. Emphasis of hacker forum data set on yearly basis
*only contains data for January and February

The percentage of titles within the Attacks & Defences knowledge area increased almost year-on-year, rising from 18% in 2011 to 68% in 2017. The consistent growth in this area over many years suggests this trend will likely continue in future and that cyber security curricula need to be well prepared to counter that demand. However the 'Software Platform Security' category showed the exact inverse trend, starting at 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 less technically minded users could explain the increase in get-rich-quick-type categories (such as Carding and Fraud) over more technically minded categories (such as Hacking).

Number of titles

The steady increase in topics over time within the forums support previous the conclusion from prior research that the demand for illicit services on the darknet will only increase, certainly in the short-term. However, we have yet to see law enforcement develop a consistent and pragmatic method of compromising these websites – if this does happen, it will be interesting to see whether the darknet will adapt to counter these measures or collapse completely.

If this trend continues as expected, the implications of having a large and unaccountable market for cyber criminal material need to be considered and anticipated.

As Tor is an essential tool 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 [33]. The emergence of a very large darknet would be extremely difficult to hack and de-anonymise.

4.3.4 Comparison against research into cyber security curricula

The results of our study were contrasted against prior research into cyber security qualifications to provide insight into their effectiveness in countering the cyber crim-

inal methodology being compiled on the dark web. Our study can add a further perspective by determining whether the weightings of these curriculums are proportionate to the available darknet material.

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 [28]. 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.

Adversarial Behaviour

A 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.

With this in mind, 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 this knowledge area will only increase over time.

It is worth mentioning that much of the content within the Adversarial Behaviour knowledge area is concentrated within the activities of carding and cashing out.

The increasingly high levels of carding 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.

Taking this into account as well as the inability of most generalist curriculums to adequately cover this area, it may be worth creating a specific qualification for those working within in the banking sector.

The areas concerned with banking related crime are both somewhat isolated from the rest of cyber crime and in very high demand. In addition, the introduction of cryptocurrencies have made methods for money laundering increasingly sophisticated.

Privacy and Online Rights

The 'Privacy Online Rights' knowledge area features somewhat prominently within the learning material, particularly the crypto-market forums where there is a lot of discussion relating to proxies, VPNs, the TOR browser and other methods for preserving anonymity.

The darknet focus is primarily on defensive tools to stay anonymous while engaging in illicit activity.

Cyber-security curriculums however focus on 'Risk Management and Governance' which refer to security management systems.

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 data net items.

This would strongly suggest users on the hacker and cryptomarket forums are generally disinterested in legitimate software development or at least, discussing it on the dark net. However, there are a large amount of technical e-books on the market listings, some of which relate to software development.

This knowledge area is far more prevalent within the cyber-security curriculums. This reflects 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.

Similarities between Hacker set and IISP

The mappings of the hacker forum data set and the IISP qualification shared many similarities.

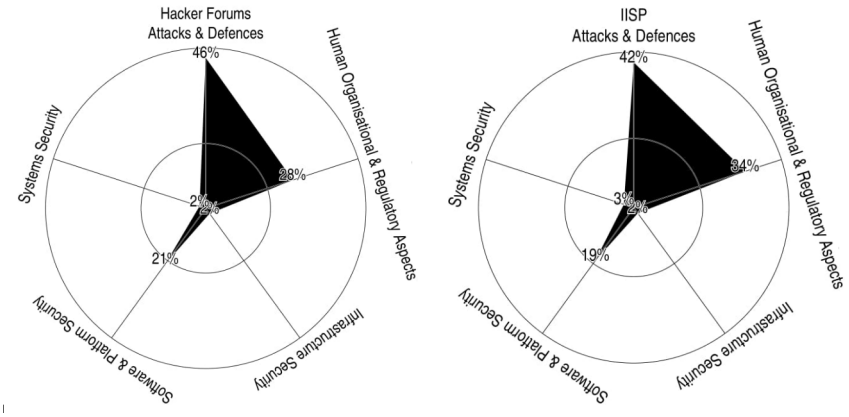

Figure 4.8: Hacker forums comparison with IISP

As shown above, both mappings display a high prevalence towards both the Attacks & Defences and Human Organisational & Regulatory Aspects knowledge areas. At a glance, this alignment suggests this particular curriculum is effectively covering the material available on the hacker forums.

However when analysed in more depth, it should be noted that the focus differs considerably within these broader knowledge areas. 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 knowledge areas the focus is primarily on Risk Management Governance whereas the hacker forums is entirely focused on Privacy Online Rights. Similarly, within 'Attacks and Defences' IISP focuses primarily on 'Security Operations & Incident Management' whereas the hacker forum emphasis is on 'Adversarial Behaviour'.

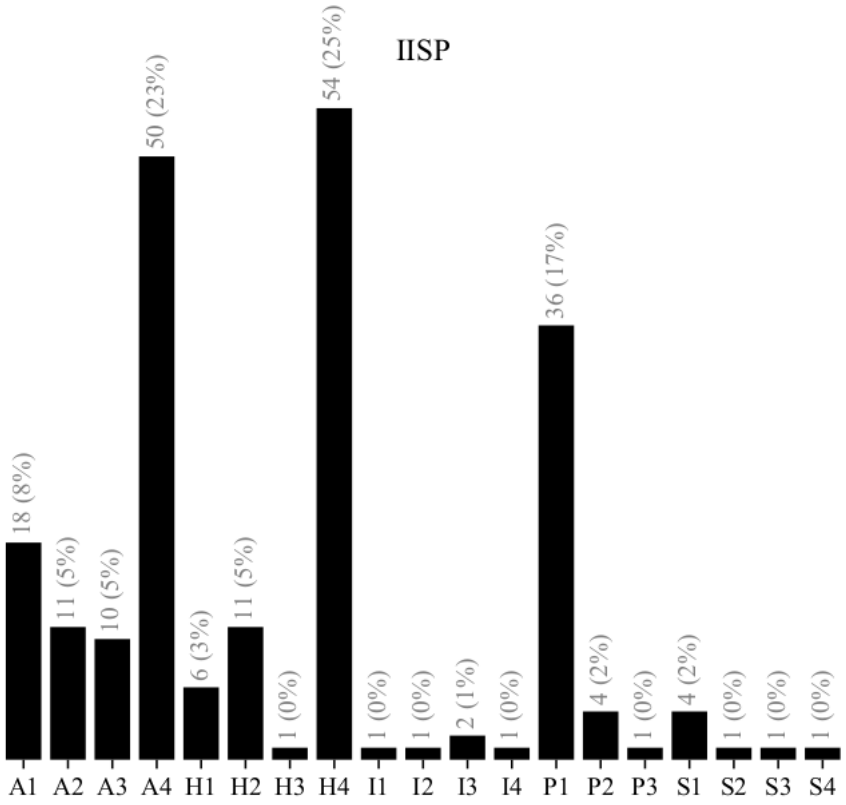

Figure 4.9: Breakdown of IISP curriculum by knowledge area [28]

These differences can be explained to a great extent in terms of proactive versus reactive intentions. For instance, 'Adversarial Behaviours' largely features methods used by those attacking systems while 'Security Operations & Incident Management' comprises the methodology for securing those systems. It is perhaps unsurprising that users on the darknet are far more interested in breaking into systems rather than defending them.

Similarly, the users on hacker forums are primarily interested in tools related to 'target hardening' capabilities, making their activities more difficult to track (i.e. proxies, VPNs, Tor etc.). While there is no doubt a need for cyber security qualifications to teach a basic understanding of these technologies, they are typically of more interest to those engaging in hacking activities as opposed to those intending to trace them.

Delivery methodology

Previous research (All-That-Glitters) into cyber security qualifications has determined they often rely heavily on ineffective techniques such as multiple choice examinations.

Therefore, once these deficits are identified, 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.

These suggestions 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 [3]. 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.

A 2016 report suggested part of the skill gap was due to insufficient exposure to cyber and information security concepts in computing courses [3]. 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.

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 display the results of this 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 were contrasted with those created through analysis of cyber security qualifications. The differences and similarities between these mappings were discussed and on this basis potential improvements for these curricula were proposed.

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

The codebook was formulated only from analysis of the crypto-market listings which caused minor problems at the later stage of analysing the forum thread titles. 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 were often combined.

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 'Technical E-Books' which could include both a guide on how to use 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.

Learning Python

It was decided that Python would be the most appropriate programming language to code the classifier. Prior to this project, I had only used Python sparingly and at a fairly basic level. Thankfully learning Python did not prove too difficult or time consuming as it had many parallels with Java and C, of which I had far more familiarity. In addition, many of the libraries were generally straight forward to use and my project supervisors were able to provide support where necessary.

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 thread 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 example that I spotted and manually re-categorised was "Homemade Semtex - C4's Ugly Sister".

The classifier does attempt to control for this to an extent by performing some checks for certain key phrases prior to the tokenisation stage (further detail on PAGE numb). 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 [53]. 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 [page numb]

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.

5.2.6 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.

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.

Much of this data was summarised in static graphs which appropriately summarised much

6 Conclusion and Future Work

chapter Your text here

6.1 Conclusion

This project looked at the range of cyber criminal learning material available on the darknet.

In order to analyse these data sets, a keyword classifier was built that could assign string items to a CyBOK knowledge area. To ensure it was accurately assigning categories, the classifier was validated using k-fold cross validation before being used

The mappings were contrasted with those of common cyber-security qualifications.

These findings support the introduction of more specialised qualifications, particularly targeting the banking sector and infrastructure security.

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.

An interesting follow-up study would be to replicate this 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.

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. [49] 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 time permitted. For example, I would have experimented with a different algorithms such as log-likelihood ratio to determine whether these would have produced increased validity over the log likelihood values.

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.

Machine learning techniques could also be implemented to allow for the continuous training and improvement of the classifier.

Bibliography

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

[18] Christopher Copeland, Mikaela Wallin, and Thomas J. Holt. Assessing the Practices and Products of Darkweb Firearm Vendors. *Deviant Behavior*, 41(8):949–968, aug 2019. URL: https://www.tandfonline.com/doi/abs/10.1080/01639625.2019.1596465, doi:10.1080/01639625.2019.1596465.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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/.

[38] 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/.

[39] 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.

[40] 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.

[41] 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.

[42] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, Andrew Martin, Makayla Lewis, and Claudia Peersman. Scoping the Cyber Security Body of Knowledge. Technical report.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

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