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

Andrew Thomas
University of Bristol
Bristol, England
vz19513@bristol.ac.uk

CCS Concepts

• **Security and privacy** → *Malware and its mitigation*; **Human and societal aspects of security and privacy**;

Keywords

dark web, cyber security, education, blackhat, darknet markets, forums

1 Executive Summary

Justification

The anonymity and accessibility of the dark web has allowed a cyber criminal subculture to thrive. The implications of this from a cyber security perspective are significant as it enables criminal guides and exploits to be traded with impunity. A cyber criminal no longer has to develop technical competence through a traditional formal curriculum as the knowledge or services can simply be purchased through informal channels [16].

This research aims to analyse the learning material available on dark web markets and forums as a means to understanding the type of material that cyber criminals are consuming. To achieve this, the content will be mapped to knowledge areas within the Cyber Security Body Of Knowledge (CyBOK) in order to determine the areas with least and most emphasis. The mapping will also be evaluated against similar mapping of formal cyber security qualifications [7] in order to identify mismatches between their curriculums and the available cyber criminal learning material.

Aims and Objectives

This research 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 the formal cyber security curriculum [7] 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.

The core objectives are:

(1) To create a mapping of cyber criminal learning material available on the dark web according to CyBOK knowledge areas.
(2) Identify any trends and contrast our findings with previous research to inform the current state of cyber security knowledge.

Deliverables

- A codebook defining the key categories within the source data, formulated from the manually categorised ground truth data.
- A 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 [7] to determine potential areas of deficit.

Value Added

Previous research has determined that the focus of different cyber security qualifications vary considerably [7]. This research revealed that there were areas of cyber security, such as physical-layer and hardware security, which were heavily under-represented.

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.

Previous research has stressed the importance of the continuous monitoring of the dark web in order to detect and prepare for emerging cyber criminal threats from within darknet communities [5]. The study will address this by producing yearly mappings from the source data in order to identify emerging trends. In addition, the study will produce a program that can be re-purposed to provide more detailed and up-to-date mapping of the darknet. For example, integration with an automated web scraper to provide real-time updates on darknet trends.

2 Introduction

Motivation

Within the last decade a number of dark web markets and forums have emerged where goods and services are traded anonymously [16]. 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. 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.

It is important that the state of cyber security education adapts to reflect this new reality and previous research has already identified the pressing need to start mapping dark net content [3].

Hypothesis

Due to the exploratory nature of the study, there is no fixed hypothesis that will be evaluated. However initial preparatory analysis of the data as well as previous research [16] suggests adversarial behaviours will be the most prominent knowledge area.

Scope

The data analysed in this study has been obtained from three publicly available separate sources:

- 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 88,215 posts selected from relevant categories within the following two datasets:

- The first set of data which covers cryptomarket discussion forums has been sourced from the DNM archives [2].
- The second set of data that covers the hacking discussion forums has been sourced from the University of Arizona's Artificial Intelligence Lab [1].

Preparation commenced in February and the project will conclude by early September. For 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 darknet sites 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 web does not necessarily translate to an actual lack of such.

However, the source data is a substantial subsection which includes the larger and more popular dark web sites that can be reasonably assumed to be broadly representative of the dark web as a whole. In addition, our initial study can serve as a basis for more wider and more far reaching studies in the future. For example, if content from a previously undiscovered dark web community emerges.

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 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.

The study does not test or in any way factor in the efficacy of the learning material. This is 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. However, prior research has already demonstrated the efficacy of dark web learning material. A study by van Hardeveld et al. [17] analysed tutorials concerned with cashing out stolen credit card details 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.

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. To account for this, we will also create mappings by year to detect any emerging trends.

3 Literature Review

The Dark Web

The dark web refers to a section of the deep web that has been intentionally hidden and is not accessible through standard web browsers. The deep web is the part of the World Wide Web whose contents are not indexed by search-engines. This deep web is of substantial size with recent estimates [14] suggesting it is approximately 400–500 times larger than the Surface Web also known as the Internet that we normally use. The dark web constitutes only a tiny fraction of this but it is growing. A 2017 report [9] identified only around 4,000 dark web sites. Two years later another report [6] found there were 8,416 active dark web domains.

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 are encrypted using strong cryptography. 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. Anonymity is also guaranteed for users as the Tor client software, commonly used to access the dark web, routes Internet traffic through a worldwide volunteer network of servers thereby hiding user's information and eluding monitoring attempts.

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 [3]. Notably, this makes the dark web particularly appropriate for cyber criminals, who are constantly trying to hide their tracks [11] and now has the potential to host an increasingly high number of malicious services and activities.

Markets and Forums

Much like on the open internet, these illegal services and items are shared or sold on darknet forums and marketplaces. Markets are websites, such as the notorious Silk Road [4], which act as anonymous marketplaces selling everything from tame items such as books and clothes, to more illicit goods such as drugs and weapons. Aesthetically, these sites appear like any number of shopping websites, with a short description of the goods, and an accompanying photograph.

Transactions are anonymised through the use of cryptocurrency such as Bitcoin and communications between buyer and seller are encrypted using 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 cryptomarket [10].

Another area where cyber criminal knowledge is shared is on darknet 'black hat' hacker forums. 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 can be of considerable scale with certain forums containing thousands of posts [1].

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 [15].

A multitude of cyber criminal materials are shared or traded on these websites. One such example is the trade in software exploits. These markets serve as platforms for buying and selling zero-day exploits, and an exploit's price factors in how widely the target software is used as well as the difficulty of cracking it [8]. These sites also facilitate trade in proprietary information including passwords for surface Web paid-pornography sites and PayPal passwords [18].

Money laundering, often involving the use of cryptocurrencies, and the trade in credit card information — commonly referred to as carding — also feature prominently [16]. A study even identified numerous darknet websites which specialise solely in these activities [3].

Implications for cyber security

The accessibility of cyber criminal services and material on both dark web markets and forums contributes to the rise in cybercrime by lowering barriers for entry. cyber criminals no longer have to have the technical competence to achieve their ends since the expertise or relevant software can simply be purchased [16]. This market is not insignificant with research conservatively estimating the overall revenue generated for cyber criminal commodities on darknet markets as being at least US $15M between 2011–2017 [16].

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. Indeed, demonstrating that the dark web has turned into a major platform for cyber criminal activities is crucial in order to develop the necessary tools to monitor all parts of the Internet.

A study by Ciancaglini et al. [5] proposed a number of methods we should be utilising to monitor the dark web. One such method involves building a semantic database that contains important information regarding a hidden site which can then be used to help track future illegal activities on the site and associate them with malicious actors.

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.

Prior research into formal cyber security qualifications has determined that different curriculum frameworks have different emphasis [7]. In addition, it was also discovered that the more technical areas of cyber security such as physical security and hardware security, are covered scarcely or not at all.

By contrasting the knowledge area mappings from this study against the mapping 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 learning from the dark web.

4 Proposed Approach

Research Questions

This research intends to answer the following research questions:

- Is there learning material accessible on the dark web that is not adequately covered within the scope of CyBOK or the formal cyber security curriculum?
- What are the key differences between the learning material accessible on the dark web in comparison to what is being taught via accredited courses?

Methodology

The study will be broadly split into three stages:

1. Analysis of the source data

The aim of the stage is to analyse the source data to identify key themes. Firstly, the forum categories and marketplaces will be manually checked to ensure they contain relevant information and any that do not are excluded from the study.

Manual thematic analysis of a broad subsection of the source data will be conducted in order to establish distinct categories that can form the basis for a codebook. These categories will also be assigned to the most appropriate CyBOK knowledge area in order to create the mappings required for final stage of the study. The aim will be to manually categorise at least 3000 market listings to form the ground truth data set.

CyBOK is a broad guide to foundational cyber security knowledge which has been developed through consultation with industry and academia [12]. It codifies 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 [7].

Once an initial codebook is completed, a randomly selected ten percent subsection of the ground truth data will be cross-checked by a co-rater. Cohen’s kappa will be calculated to determine the reliability of the codebook definitions. If this is not achieved initially, amendments will be made to the ground truth data and the codebook until a Cohen’s kappa value greater than 0.70 is achieved.

This categorised data will be utilised as the ground truth data to which the keyword analysis is then applied in order to identify the most frequently occurring keywords for each category.

2. Automation

The second stage is the creation of a Python program which can automate the categorisation of the remaining source data. The automation is a necessity in order to process the amount of data within the required timescale. Firstly, the program will perform keyword analysis on the ground truth data. This will involve tokenisation in order to calculate word frequency then a further calculation will need to be performed to calculate the log likelihood values for each token. These log likelihood values will determine the most relevant keywords for a particular category.

Secondly, the program will automate the categorisation of the remaining source data based on these calculated log likelihood values. This stage will include any further tweaking of the model. For example, to fix bugs if they are discovered or making changes to the code to improve efficiency. In order to ensure the automation is functioning as intended, output from the automated categorisation will be checked against manual categorisation of the same data and Cohen’s kappa will be calculated until a value greater than 0.70 is achieved.

The third step will involve adding additional functionality to the program. Importantly, this will include functionality to sort the source data by date which will be required in order to create mappings for each year. This will either be achieved by the creation of a function within the Python program or potentially by extracting the source data to a SQL database which will then be queried.

3. Evaluation of results

The final stage will involve evaluation of the results and creation of any graphical representations. Graphical representations will include radar charts that can be used to summarise our findings. In addition, the charts will be used for direct comparison with the charts produced by research into cyber security certifications [7]. The yearly mappings will also be contrasted to determine any general trends. If time permits, more advanced graphical representations such as interactive diagrams may be produced.

Lastly, the results will also be evaluated against the background literature on the dark web and the significance of any new findings will be discussed as well as their implications for cyber security.

Execution

The software used for this task will include: spreadsheet software to store and view the source data, an IDE for writing Python script to read and modify CSV file data and a data analysis library to calculate Cohen's kappa and perform keyword analysis.

Keyword analysis will be performed using the techniques outlined in the article by Rayson & Garside [13]. This involves the creation of a word frequency list for each corpus. For each word, a log-likelihood value is calculated via a contingency table. The list is then sorted by their log-likelihood value with the words at the top deemed to be the most significant keyword for the corpus.

The validity of the program will be tested by conducting a manual analysis of the same source data and checking for alignment. If the Cohen's kappa value for the comparison is not achieved initially, the program will be tweaked until this value is achieved. Execution of the mapping task will involve an understanding of the various forms of cyber crime as well as how they map to the CyBOK knowledge areas.

Evaluation

The evaluative points for this research are firstly, whether we have been able to produce an accurate mapping of our dark web source material onto the CyBOK knowledge areas. Secondly, once the first point has been completed, contrasting our results with prior research to provide meaningful feedback on the current state of the cyber security curriculum.

Furthermore, the model will be quantitatively tested to ensure it is accurately assigning categories. This will be achieved by conducting a manual analysis of a subset of the automatically assigned data. Cohen's kappa will be calculated to ensure significant co-rater reliability and the model will be tweaked if necessary.

The implications of the mapping will be qualitatively measured. Firstly, the mapping will be contrasted with that of the formal cyber security qualifications [7] to determine whether the current cyber security curriculum adequately covers the content available on the dark web.

Secondly, the findings will be contrasted against prior research into the dark web and any further potential implications for cyber security will be discussed. The source data will also be sorted by year and mappings created for the yearly subgroups to determine any emerging trends.

Conclusion

The growing and largely unexplored cyber criminal communities on the dark web poses a significant threat to cyber security. This research intends to discover the most prominent areas of cyber criminal learning material available on the dark web and use our findings to evaluate the current effectiveness of the available cyber security qualifications.

In order to achieve this, the source data will be mapped to CyBOK knowledge areas and this mapping will be used as a basis for comparison against similar mapping of cyber security qualifications in order to assess whether these courses are adequately equipping their graduates to deal with these threats. Mapping will also be created for each year of the data to determine any emerging trends and the implications of these will be evaluated from a cyber security perspective.

The study will utilise publicly available collections of dark web forum posts and market listings. By utilising pre-existing collections, this allows for time to be focused on data preparation and analysis instead of collection. While not fully comprehensive, it is a substantial dataset that includes the more popular darknet websites, which should be generally representative of the darknet as a whole.

Prior research has established that cyber security qualifications do not place emphasis on all areas of cyber security equally [7]. Taking this into account, this study aims to determine if the content of these qualifications are reflective of the cyber criminal learning material available on the darknet.

The differences between the two sets of mappings will be analysed in detail to provide insight into the effectiveness of formal cyber security qualifications to counter cyber criminal methodology being compiled on the darknet. In addition, the study will attempt to identify emerging trends on the darknet by producing mappings of the source data on a yearly basis.

References

[1] Azsecure. [n.d.]. Other Forums: AZSecure-data.org. https://www.azsecure-data.org/other-forums.html

[2] 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. 2015. Dark Net Market archives, 2011-2015. https://www.gwern.net/DNM-archives. https://www.gwern.net/DNM-archives Accessed: 2020-05-25.

[3] Michael Chertoff and Tobby Simon. 2015. *The Impact of the Dark Web on Internet Governance and Cyber Security.* Technical Report.

[4] Nicolas Christin. 2013. Traveling the silk road. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*. Association for Computing Machinery (ACM), New York, New York, USA, 213–224. https://doi.org/10.1145/2488388.2488408

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

[6] Cyberscoop. 2019. *How many dark web marketplaces actually exist? About 100.* https://www.cyberscoop.com/dark-web-marketplaces-research-recorded-future/

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

[8] Charlie Miller. 2007. *The Legitimate Vulnerability Market Inside the Secretive World of 0-day Exploit Sales.* Technical Report. www.securityevaluators.com

[9] OnionScan. 2017. *OnionScan Report: Freedom Hosting II, A New Map and a New Direction.* https://mascherari.press/onionscan-report-fhii-a-new-map-and-the-future/

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

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

[12] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, Andrew Martin, Makayla Lewis, and Claudia Peersman. [n.d.]. *Scoping the Cyber Security Body of Knowledge.* Technical Report.

[13] Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora -*, Vol. 9. Association for Computational Linguistics (ACL), Morristown, NJ, USA, 1–6. https://doi.org/10.3115/1117729.1117730

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

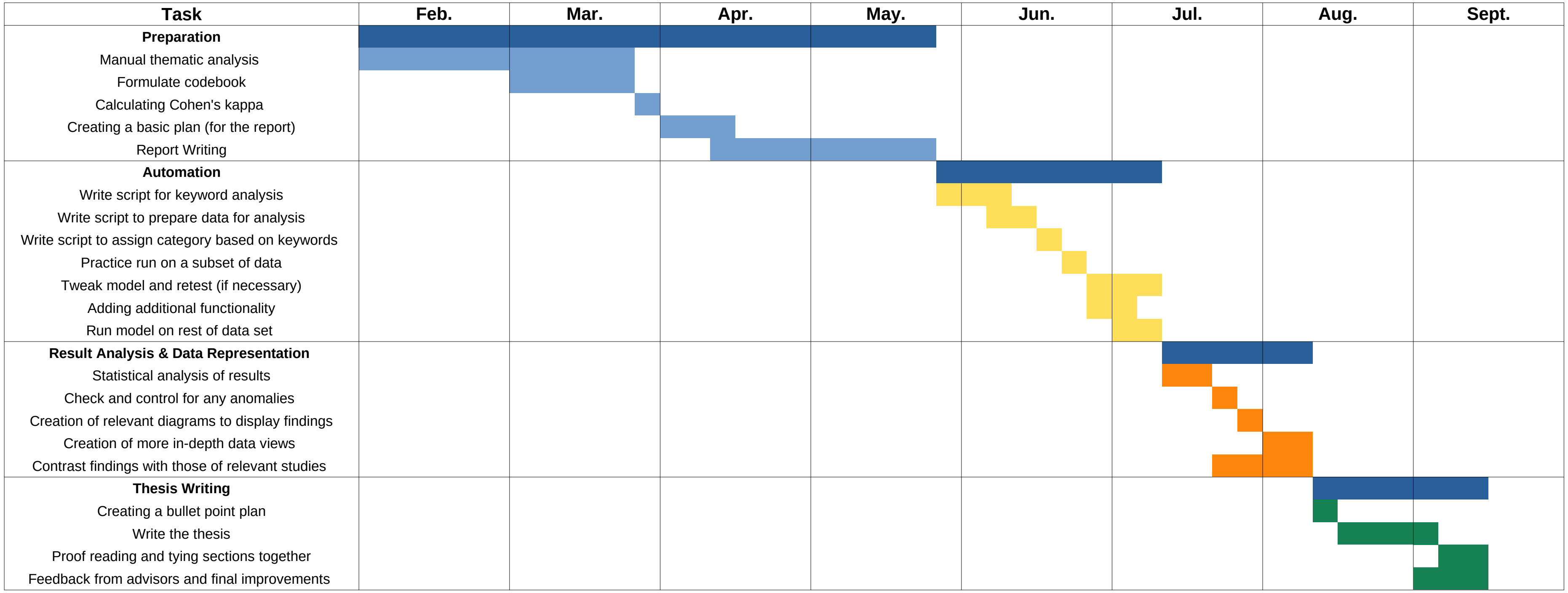
[15] Jana Shakarian, Andrew T. Gunn, and Paulo Shakarian. 2016. Exploring malicious hacker forums. In *Cyber Deception: Building the Scientific Foundation.* Springer International Publishing, 259–282. https://doi.org/10.1007/978-3-319-32699-3_11

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

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

[18] Ken Westin. 2014. Stolen Credit Cards and the Black Market: How the Deep Web Underground Economy Works . https://www.linkedin.com/pulse/20140822165720-1094155-stolen-credit-cards-and-the-black-market-how-the-digital-underground-works

Gant Chart - CyBOK of Evil



Task	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sept.
Preparation	■	■	■	■				
Manual thematic analysis	■							
Formulate codebook		■						
Calculating Cohen's kappa		■						
Creating a basic plan (for the report)			■					
Report Writing			■	■				
Automation				■	■	■		
Write script for keyword analysis				■	■			
Write script to prepare data for analysis					■			
Write script to assign category based on keywords					■			
Practice run on a subset of data					■			
Tweak model and retest (if necessary)					■	■		
Adding additional functionality					■	■		
Run model on rest of data set						■		
Result Analysis & Data Representation						■	■	
Statistical analysis of results						■		
Check and control for any anomalies						■		
Creation of relevant diagrams to display findings						■		
Creation of more in-depth data views							■	
Contrast findings with those of relevant studies						■	■	
Thesis Writing							■	■
Creating a bullet point plan							■	
Write the thesis							■	■
Proof reading and tying sections together								■
Feedback from advisors and final improvements								■

Risk	Severity (1-5)	Likelihood (1-5)	Exposure (1-25)	Mitigation	Contingency Plan
Risk that keyword analysis does not produce distinct keywords for a category.	2	3	8	The ground truth data on which we are basing the keyword analysis should contain enough entries for each category to produce distinct keywords.	If this occurs, we can manually identify more entries to add to that particular category until significant keywords are determined.
Risk that the source data used for analysis is not representive of the broader dark web.	4	1	6	There are only a relatively small number of darknet sites and the source data has been taken from a wide variety of popular examples.	New dark web sources, if discovered and are deemed highly relevant, could be mined and analysed alongside current source data.
Ambiguity as to how certain titles should map to CyBok knowledge areas.	1	5	5	Any ambiguous titles will be discussed with supervisors to deem a best-fitting category.	These categories and the reasoning for their respective mapping within CyBok will be explained within the study.
Identifying a substantial and relevant category that is not included in the codebook at any stage after its creation.	2	2	4	A substantial amount of the data (over 3000 listings) will be checked manually prior to keyword analysis.	The codebook can be amended to include the newly identified category and then re-checked for inter-rater reliability.
Unforeseen hardware issues such as laptop crashing. (Unable to access university resources due to Covid-19)	4	2	8	There will be backups of key data on both external storage devices and on cloud storage facilities.	If this occurs I will raise with my project supervisors and seek out potential alternative hardware. I will also request an extension to the deadline if necessary.