Computational criminology: at-scale quantitative analysis of the evolution of cybercrime forums

Jack Hughes

Churchill College

This thesis is submitted for the degree of Doctor of Philosophy.
DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60,000 words.

Jack Hughes
July, 2023
Abstract

Computational criminology: at-scale quantitative analysis of the evolution of cybercrime forums

Jack Hughes

Cybercrime forums and marketplaces are used by members to share hacking techniques, general community-building discussions, and trade hacking tools. While there is a large corpus of literature studying these platforms, from a cross-forum ecosystem comparison to smaller qualitative analyses of specific crime types within a single forum, there has been little research into studying these over time. Using the CrimeBB dataset from the Cambridge Cybercrime Centre, this first contribution of the thesis explores the evolution of a large cybercrime forum, from growth to gradual decline from peak activity, with research questions grounded in the digital drift framework from criminological theory. This finds a trend towards financially-driven cybercrime over time, by users and the forum as a whole. The second contribution of the thesis presents a method for detecting trending terms, using a lightweight natural language processing method to handle queries, given the size of the dataset. Evaluation using manual annotations showed more relevant salient terms were detected over TF-IDF. Finally, the third contribution of the thesis applies signalling theory to analyse the usage of argot (jargon and slang) on the forum, finding a negative correlation with reputation usage, and clustering to find a decreasing use of argot over time. Part of this contribution includes a lightweight argot detection pipeline with word embeddings aligned with manual annotations. Overall, the combination of different approaches, including criminological theory driving research directions, natural language processing to analyse forum text data, machine learning for classifications, and data science techniques, all contribute to provide a unique interdisciplinary perspective within the field of cybercrime community research, both drawing insights into these communities and contributing novel tools for measurements of large, noisy text data.
ACKNOWLEDGEMENTS

This work would not have been achieved without the support from many people, who I would like to thank. Firstly, thank you to my supervisor Prof. Alice Hutchings for her guidance, understanding, and support throughout the entirety of the PhD. I would like to thank all of the members of the Cambridge Cybercrime Centre, with which I have had the pleasure to work with. Thank you to those that have provided valuable feedback on this thesis. Thank you to the Engineering and Physical Sciences Research Council (EPSRC) (Doctoral Training Studentship) and the Department of Computer Science and Technology for generously providing the funding to enable my PhD. Finally, thank you to my family and friends who have supported me throughout this journey.
# Contents

1 Introduction ........................................ 15
   1.1 Contributions ..................................... 17

2 Background ........................................... 21
   2.1 Cybercrime Forums and Markets .................... 22
   2.2 Criminological Theories ............................ 24
      2.2.1 Pathways ..................................... 25
      2.2.2 Cold Start Problem ............................ 25
         2.2.2.1 Argot and the Criminal Underworld .......... 26
         2.2.2.2 Argot Detection in Cybercrime Communities .... 26
      2.2.3 drift & Churn ................................. 27
      2.2.4 Reputation and Trust on Underground Forums ...... 28

3 Methods .............................................. 31
   3.1 Data ............................................... 32
   3.2 Ethics ............................................. 33
      3.2.1 Limitations .................................... 34
   3.3 Natural Language Processing ....................... 35
      3.3.1 NLP and Cybercrime Discussion Platforms ....... 35
      3.3.2 Evolving Terminology ......................... 36
      3.3.3 Trending Topics ................................ 36
         3.3.3.1 Trending Topics: Fightin’ Words ............ 38
         3.3.3.2 Use case: Detecting Emerging Threats ....... 38
      3.3.4 Argot Detection ................................ 39
   3.4 Named Entity Recognition .......................... 39
   3.5 Machine Learning for Modelling Trajectories ....... 40
      3.5.1 Modelling Trajectories ......................... 40
      3.5.2 Interpretable and Explainable Models .......... 41

4 Measuring the Evolution of Cybercrime Forums .......... 43
   4.1 Introduction ...................................... 43
CHAPTER 1

INTRODUCTION

Cybercrime forums are defined as online platforms where members can discuss topics including hacking techniques and developing malware, and share tools and guides, often for financial gain [Motoyama et al., 2011]. These platforms provide some structure: threads of conversation are categorised into larger topics, and administrators set and enforce community rules [Afroz et al., 2013]. These forums have seen growing and declining numbers of users, with over a decade’s worth of activity. A large corpus of research has studied these, primarily on the larger more popular forums. Large, public cybercrime forums on the surface web can provide an entrypoint into cybercrime for new members, and contain a large corpus of discussions on cybercrime topics. Insights can be drawn from these to show how individuals start, sustain, and desist from criminal activities, who may benefit from interventions. This thesis presents new approaches for analysing forum datasets, to provide these insights.

Much of the existing literature has taken perspectives from distinct disciplines (such as computer science or criminology), and observed the dataset as a static whole. This thesis both takes an interdisciplinary approach, and uses over a decade of cybercrime forum data history to take a longitudinal approach to analysis. Firstly, the interdisciplinary approach uses theories from criminology to ground research questions into cybercrime forums, which are combined with large scale computational analysis, collating techniques from data science, machine learning, and natural language processing. Criminological theories used include signalling theory, information asymmetry [Gambetta, 2011], digital drift and pathways [Hutchings, 2016]. Digital drift and pathways theories have been used to explain how individuals move into criminal activity, sustain criminal activity, and desistence from offending. Secondly, the longitudinal approach uses data for over a decade of activity, where the turnover of new members brings changes in interests and topics on the forum, reshaping these communities. Measurements show how groups on forums move around topics, and drift off the forum over time. Applying these theories to forum
Signalling theory has been used to explain how criminals in the real world use clues to signal trustworthy aspects to other criminals, such as through tattoos [Gambetta, 2011]. In environments where trust is low, such signals are designed to be easy to present yet hard to mimic by those outside of the group. This work analyses argot, defined as slang and jargon used by groups, which can be used as a signal of trust, to compare this to the use of reputation systems.

The thesis starts with measurements into how forums have evolved over time with the churn of users year-on-year combined with changing interests. This uses drift theory as a framework to understand forum users’ pathways. Later, lightweight natural language processing pipelines are developed to identify trending topics on forum data over time. The final main chapter builds a data pipeline to detect argot using machine learning with language models, to then measure the correlation between argot usage and reputation voting data. This applies signalling theory, finding evidence that forum users are using argot to signal trustworthiness. Evidence of the use of argot to overcome the cold start problem faced by new forum members is also uncovered.

All of the chapters in this thesis use forum datasets from the Cambridge Cybercrime Centre, available for research use under a data sharing agreement. Forum datasets can be obtained through scraping, leaked datasets, or using datasets available publicly or through data sharing agreements. The forum data used for this thesis is from CrimeBB [Pastrana et al., 2018b]. This corpus contains hacking-related forums of different sizes. For each site, there are many boards. Boards are created by administrators, and are useful in providing context and they are typically stable. Boards can be categorised with varying granularity, such as marketplace related activity (selling and buying boards), or hacking techniques of ‘beginner’ and ‘advanced’ levels. Also, ‘Programming’ boards exist, but these do not always give a good idea of what type of tools are being built, e.g. discussions may talk about a given problem to solve, but it may not be clear if this problem is malicious, e.g. part of malware development. Boards can also vary in size: they can be for niche interests, or form a general ‘marketplace’. Boards consist of threads. Threads are created by forum members, and may contain keywords unique to the forum (e.g. proper nouns of products or labels such as ‘[Selling]’). While creating a thread may appear to be a large contribution to the forum (suggesting a new topic to discuss), it may not always be the case. One user may create a thread for a useful hacking tutorial, while another may create an advertisement for a product that generates no replies or responses. Within threads, there are a series of posts. The first post is created by the author of the thread, and provides more detail beyond the title of the thread. Successive posts (replies) can optionally quote other posts. Posts can also contain quotes, citations, links, images, code blocks, emojis, or just be a simple ‘thanks’. This is the structure of typical forums, however some forums may
contain extra data. Reputation votes are used by some forums to provide a mechanism for trust. Members can “vouch” for other members, with a positive or negative quantity and optionally provide a reason. Sometimes reputation systems are limited to just marketplace activity, while other systems may also include general discussions.

Forums provide a platform for those with similar interests to discuss ideas and share tools, and larger popular forums can provide a centralised hub covering a wide range of users, from those learning how to start hacking to members providing support. Studying these forums helps to gain further understanding of the cybercrime landscape, as they contain large amounts of discussions about hacking techniques and current trends. Given the large size of the dataset, and the noisy jargon–filed posts within it, novel approaches are needed for analysis, in which this thesis contributes novel tools to these problems.

1.1 Contributions

During working on this thesis, I have been involved in multiple projects. Projects where I was the main contributor are outlined further in this thesis, and all papers including those co-authored are summarised in this section.

A Social Network Analysis and Comparison of Six Dark Web Forums


This work was carried out in collaboration with co-authors, from computer science (Ildiko Pete), criminology (Yi Ting Chua), and psychology (Maria Bada). It uses social network interaction graphs to analyse six forums available only on the Tor anonymity network. Centrality and structural measurements show a small group of users that form central hubs, and qualitative analysis found these members typically post general content across different topics. In this work, I assisted with the project design, I developed the interaction graph model, and I wrote the code implementation for building the model and running measurements. I also assisted with qualitative analysis of post contents.

Turning up the dial: The evolution of a cybercrime market through set-up, stable, and COVID-19 eras

Anh V. Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. Turning Up the Dial: The Evolution of a Cybercrime Market Through Set-up,
This work was carried out in collaboration with co-authors, and it explores the creation of a contract system on a cybercrime marketplace, designed to establish a new mechanism for trust. The primary focus of the paper highlights three different stages the contract system went through, categorising products traded, measuring transaction values, and the evolution of members’ interests over time. In this work, I assisted with the project design, and contributed to the section of the paper exploring the cold start problem on the marketplace. This uses clustering and qualitative analysis to analyse how members overcome the issue of low trust when starting on a new marketplace, finding that new members are more active in low-level currency exchange while they build reputation and trust.

Detecting trending terms in cybersecurity forum discussions


I was the first author on this paper, and work was carried out in collaboration with co-authors from natural language processing and linguistics (Seth Aycock, Andrew Caines, and Paula Buttery) and criminology (Alice Hutchings). This paper uses a lightweight Bayesian approach to detect trending terms on large corpora containing jargon. Seth Aycock contributed an initial R prototype, which combined the tidylo R library with a small dataset sample from CrimeBB. I converted this prototype and library to Python, to connect with existing tools, efficiently scaled up to the entire CrimeBB dataset, and provided support for Elasticsearch databases. I also carried out further analysis and evaluation work. This paper has been reworked into Chapter 5.

Too Much Data: Opportunities and Challenges of Large Datasets and Cybercrime


This book chapter explores common issues faced and opportunities with cybercrime forum data, combining insights from the three authors’ prior works and projects. I
PostCog: A tool for interdisciplinary research into underground forums at scale


This work was carried out in collaboration with co-authors, and it presents a new web-based user interface for accessing the CrimeBB dataset, to enable access to cybercrime forum data regardless of technical skills. I contributed to this work by providing assistance with development issues, usability testing, and code review.

Argot as a Trust Signal: Slang, Jargon & Reputation on a Large Cybercrime Forum


I was the first author on this paper, and work was carried out in collaboration with co-authors. This explores the use of argot, defined as slang and jargon, on a cybercrime forum, and the correlation of this usage with reputation score data. This work also presents a pipeline for detecting argot, using aligned word embeddings. I worked on building the analysis pipeline, data analysis comparing argot and reputation, clustering of variables over time, and manuscript writing. Chapter 6 is largely based on this paper.

The Art of Cybercrime Community Research

Jack Hughes, Sergio Pastrana, Alice Hutchings, Sadia Afroz, Sagar Samtani, Weifeng Li, and Ericsson Santana Marin. The Art of Cybercrime Community Research. In ACM Computing Surveys, 2024. [Hughes et al., 2024]

I was the first author on this paper, which is currently under submission, and work was carried out in collaboration with co-authors. This survey paper covers over 100 prior works on cybercrime communities, describing the current state of the art, and discussing both inherent challenges in the field and challenges that can be overcome. This work sets out good practices and proposes a list of considerations for future cybercrime community researchers. In this work, I took the lead on organising the paper, including reading papers for the survey and writing the manuscript.
Digital Drift and the Evolution of a Large Cybercrime Forum


I was the first author on this paper, and work was carried out in collaboration my with co-author. This measurement study explores the time-based aspect of a forum dataset, to explore how the evolution of the forum has changed it, by size and by discussion topics. This work is guided by the criminological theory of digital drift. In this work, I designed the measurement study and carried out the data analysis. This paper has been reworked into Chapter 4.
CHAPTER 2

BACKGROUND

This chapter sets out the topics covered in this thesis. The interdisciplinary nature of
the thesis means this chapter covers several topics, including cybercrime forums, data
challenges, criminological theories, and the analytical methods and approaches used to
explore and measure forum datasets.

Cybercrime forums are a platform for cybercriminals to communicate, with structured
threads grouping topics of discussion into single themes unlike their predeceasing IRC
channels. Some actors are starting to move to newer platforms including Discord and
Telegram channels, which are more similar to IRC. These lack the partial structure of forum
conversations, and while they are not ephemeral (unlike IRC), the lack of themed topic
adds considerable challenges to analysis. This thesis focuses in on the more established
cybercrime forums, due to the availability of datasets which cover over a decade of data,
supporting analysis of the evolution (growth and decline) of forums.

The structure of forums assists with measurement studies, however the data still has
limitations. Primarily this is textual data, which needs to be cleaned, normalised, and
tokenised. Datasets are typically scraped, which requires considerable effort and may not
include deleted content. Furthermore, ethical concerns are raised with the use of these
datasets, as they contain information on individuals.

The work in this thesis is grounded in criminological theory to drive analysis approaches
and methods. Theories include signalling theory and digital drift, and the cold start problem
from economics. Signalling theory reasons how individuals can provide signals (e.g. tattoos
or particular slang) to indicate they can be trusted, digital drift reasons that individuals
may drift in and out of crime rather than persist in this behaviour, and the cold start
problem is the challenge faced by new members to a marketplace or trust platform in
transitioning from being a ‘new’ member to an ‘established’ or ‘trusted’ member.

Methods used include long-term measurements of forum data, latent Dirichlet allocation
(LDA) topic modelling from natural language processing (NLP) to explore the changing
discussion topics over time (Chapter 4), term-frequency inverse-document-frequency (TF-IDF) to compare to the log odds method introduced for identifying trending terms (Chapter 5), and a combination of machine learning and word embeddings to identify argot (slang and jargon) used on the forum to measure for correlation with reputation scores (Chapter 6).

2.1 Cybercrime Forums and Markets

Before cybercrime forums, IRC was a popular way for cybercriminals to communicate, and has been well studied, from data collection [Fallmann et al., 2010], to various roles of members [Thomas and Martin, 2006], and text analysis [Franklin et al., 2007]. While IRC is typically ephemeral – older messages are not kept long term – visitors to cybercrime forums can view older posts and threads.

Forum data contains a mix of private and public data, such as public posts within topics and private messages between users. Data can be collected by researchers by scraping public data, or from leaks which may additionally include private data. Motoyama et al. [2011] obtained full dumps of six forums, which included public and private messages, to measure the increase of social degree over time after joining the forum. They identified cross-forum activity by matching member email addresses, exploring the effects of different activities on reputation scores, to give an overview of how forums operate. Working with personal data, particularly private messages not intended to be viewed by others, introduces ethical considerations in research. However, training classifiers on only public data for predicting interactions or classifying marketplace listings can be limited without this data [Portnoff, 2018]. To overcome this challenge, these tasks could be iterated on with domain experts to improve modelling with ground truth [Portnoff, 2018].

Cybercrime forums vary in size and content, ranging from large general-purpose cybercrime forums which have been online for over a decade, to smaller forums focused on specific topics that grow and decline with trends. Afroz et al. [2013] explored the operations of five underground forums. They found sustainable forums have network effects with a moderate increase in new members, do not reduce connectivity over time, have admins that enforce bans, and focus around central key topics.

Researchers may choose to focus on a single forum. Pastrana et al. [2018a] look at one forum with a specific ground-truth group of members: those who have been involved in illicit activity, named “key actors”. They analyse the evolving interests of these users over time, and use an ensemble method to identify similar “key actors”, including k-means clustering, social network analysis, and logistic regression. These models include various features related to forum activity, such as centrality, and reputation measures. The use of
multiple models provides cross-validation of results, improving on previous work [Benjamin and Chen, 2012; Marin et al., 2018] which uses a single metric – member reputation scores – to validate results. Other work has shown reputation can contain intrinsic bias [Zhang and Li, 2013]. They also propose an SVM classifier for identifying posts using NLP techniques, which Caines et al. [2018] later release to produce annotations of post type and intent. Pastrana et al. [2018a] note their approach may be limited, as it focuses on one forum only, and highlight the need for longitudinal analysis methods to identify the changing interests and behaviours of key actors over time.

Pastrana et al. [2018a] does not include private interactions on the forum. While private messages and interactions contain additional data, these are usually obtained when a forum is breached and leaked. Additionally, this raises further ethical questions around the use of private messages [Thomas et al., 2017]. Furthermore, models which rely on private messages will have limited use on other forums where this data is not available.

Cybercrime forums vary in type. Some can be found on the open (‘surface’) web, and others on the ‘dark’ web, which operate as hidden services on anonymity networks. Larger forums cater to English speakers, although other-language forums exist [Afroz et al., 2014; Grisham et al., 2017]. Forums can also specialise on certain topics, e.g. carding (credit card fraud) [Mikhaylov and Frank, 2016; van Hardeveld et al., 2016], or be more general, with a wide range of different boards (high-level categories of topics) available [Pastrana et al., 2018a]. The size and topic of a forum can affect the social network structure of the forum [Pete et al., 2020]. Also, some forums have a reputation system, which can be used as a proxy to trust other members within the forum [Franklin et al., 2007; Pastrana et al., 2018a].

This thesis combines perspectives across many fields, including theory from criminology. The social aspects of cybercriminals has been studied across criminological literature and quantitative social science [Anderson et al., 2021; Collier et al., 2021; Dupont and Lusthaus, 2022; Maimon et al., 2017; Morselli, 2010; Paquet-Clouston et al., 2018], including the application of social network analysis approaches [Décary-Hétu and Dupont, 2012; Ouellet et al., 2019; Perkins et al., 2023], ranging from studying small real-world criminal groups to larger online communities. This has included using ground truth data following arrests, to build up real-world social networks of key hackers and model how a network could be destabilised [Décary-Hétu and Dupont, 2012], and the use of time-series data in longitudinal network analysis for exploring online ecosystems of hacking defacement groups [Perkins et al., 2023].

One approach for observing the evolution of forums is to measure how the social network structure has changed over time. Public data from cybercrime forums typically does not include direct user to user interactions, such as a ‘friend’ mechanism. Instead, the social graph has to be approximated, with some limitations. For example, a user
to user connection could exist by users replying in the same thread, or a user to topic or user to board graph could be built to show interests of the forum. Prior work has used the thread-reply approach to approximate the social graph [Pete et al., 2020], which found some highly connected members who form hubs, posting general discussions and tutorials on broader topics, and most members have a lower number of connections. This approximation is limiting for measurements, as it can cause a bias in results towards those who create threads. Using this approach, the thread creator acts as the ‘hub’ within the social network, inflating their importance within the network without regard to the quality of the thread or their actual influence within the network.

Community changes have also been explored in computational linguistics. Danescu-Niculescu-Mizil et al. [2013] find the evolution of forum jargon changes over time, with new community members using variations of existing terms, such as ‘smell’ in place of ‘aroma’, which takes time to become the norm for older forum members. They use these linguistic changes to predict how long users will stay part of a community. More widely, community changes have been explored in different areas, including changes of user interest over time [Benjamin and Chen, 2012; Bhalerao et al., 2019; Pastrana et al., 2018a; Sundaresan et al., 2016], and changes of forum interests over time [Afroz et al., 2013; Franklin et al., 2007; Garg et al., 2015; Motoyama et al., 2011; Vu et al., 2020]. There is a burgeoning collection of papers on communities changing over time, with a very small subset focusing specifically on cybercrime communities [Garg et al., 2015; Motoyama et al., 2011].

### 2.2 Criminological Theories

The following subsections discuss relevant criminological theories. Pathways [Hutchings, 2016] provides a framework for thinking about individuals entering and leaving criminal activity as a series of events, which can help to identify useful interventions. The cold start problem, which has been explored in reputation systems [Lam et al., 2008; Lika et al., 2014], is where new users face the challenge of having zero reputation in a market while competing with established vendors’ reputation. Pathways theory can help to explore how this problem is overcome to enable trading of hacking tools, in markets where there are established vendors of hacking tools. One approach to overcoming the cold start problem is the use of argot (slang and jargon) which can be used to signal knowledge and trustworthiness on text-based forum platforms. Pathways theory can also apply to the concept of digital drift [Goldsmith and Brewer, 2015; Holt et al., 2019] and churn on forums, in which members drift in and out of forum activity and crime, with some members either bouncing off activity shortly after joining, or later in their activity.
2.2.1 Pathways

Pathways refers to the criminological idea [Hutchings, 2016] that individuals move towards criminal activity through a gradual shift from legal to illegal activities, through a set of activities which these individuals participate in. This idea suggests that crime is an individual journey, with events happening to an individual that develops into a criminal career. Pathways can be useful to telling us how offenders get involved in crime and how they later desist. Online platforms and cybercrime have a different aspect, in which individuals can disconnect from their online identity, providing a mechanism to move in and out of crime more easily than in real-world crime groups. For example, prior research has explored a pathway from gaming into cybercrime through ‘bother’ services (DDoS-as-a-service) [Hutchings and Clayton, 2016], in which gamers use these services to take down and disadvantage opponents on online multiplayer games. These services turn the manual task of creating and managing the infrastructure of an attack into a simple interface for users, lowering the barrier to entry. Intending to cause disruption to computers using denial of service tools is an offense in the UK under the Computer Misuse Act 1990 (amended in 2006), and individuals have been sentenced under the Computer Fraud and Abuse Act in the United States for carrying out DDoS attacks. UK law enforcement has shown interest in disrupting these potential pathways, in which they have carried out intervention activities, such as the NCA running advertisement articles on three UK gaming websites for their ‘Cyber Choices’ campaign [Games Radar, 2019; Kotaku, 2019; PC Gamer, 2019] to inform readers about the illegality of these services.

Related work around pathways includes work by Pastrana et al. [2018a] who explored the evolution of member interest based upon 9 categories: hacking, market, common, gaming, money, tech, coding, web, and graphics. Interests were defined as a weighted sum of threads and posts created by users, and while the results indicated there was a shift to general areas of the forum, further analysis could be applied to these changes. Also, Bhalerao et al. [2018] examined supply chains in cybercrime marketplaces, showing links between users. This used a large set of annotations to label advertisements, and then used a time ordering to construct supply chains. This work also addresses how to work with classifications which have less than 100% accuracy, by taking a conservative approach with modelling to identify and count the supply chains that are highly likely to exist, although this number is small.

2.2.2 Cold Start Problem

Established members on a forum have had time to build a reputation and gain trust among other members. New members start with a blank profile, and have to gain trust and reputation. However, since new members start with zero reputation, it is non-trivial
to gain trust as members will not have any indicators to trust new members by. This is referred to as the cold start problem. While this is an established concern in online business and reputation systems [Lam et al., 2008; Lika et al., 2014], it is especially pertinent in cybercrime communities, where actors are inherently untrustworthy, and doing business with the wrong person can be costly in many ways, such as in undercover stings and general scams.

The only prior work into overcoming the cold start problem within the cybercrime underground is by Vu et al. [2020], in relation to a cybercrime marketplace. My contribution to this paper uses a combination of clustering and regression, to identify the group of members who overcome the cold start problem. This finds the majority of members build their reputation by participating in low value exchange-type transactions for exchanging currency. I expand on this work in this thesis, to explore the use of argot as a text–based mechanism to signal trust and overcome the cold start problem in the online environment.

2.2.2.1 Argot and the Criminal Underworld

The term argot is used to describe a specialised language. Originally the term was used to describe the particular ways in which criminals communicated, a useful way to prevent outsiders from understanding what is being said. The linguist Maurer provides a number of case studies in the 1930s and 1940s into the use of argot by particular types of law breakers and other shady figures. These include pickpockets [Maurer, 1964], prostitutes [Maurer, 1939], professional gamblers [Maurer, 1950], narcotic addicts [Maurer, 1936], moonshiners [Maurer, 1949], forgers [Maurer, 1941], and grifters [Maurer, 1947] or con artists [Maurer, 1940].

This thesis uses argot in its original context, that is, specialised language used within deviant subcommunities. However, the term has become more general over time, and is often used interchangeably with slang and jargon.

2.2.2.2 Argot Detection in Cybercrime Communities

Argot detection typically uses a comparison between a base and target corpus, to identify words which may be out-of-dictionary or are used in a different context. One approach is proposed by Seyler et al. [2021], named DarkJargon, who aim to map ‘dark’ jargon to ‘clean’ jargon to make sense of new slang terms used on cybercrime forums. They compare two methods: KL-divergence and cross-context lexical analysis (CCLA), finding that KL-divergence outperforms CCLA on their simulated dataset. They also use their approach on a real-world corpus, however as they are unable to identify false negatives with their approach, they only validate the output of the top ‘dark’ jargon words. The authors use forum and Reddit datasets, where Reddit is used as a source of ‘clean’ jargon for a control. While they build word vectors for both the forum and Reddit datasets, the
approach in this thesis uses pre-trained word vectors for the ‘clean’ sample, reducing the
time needed to collect and train on non-underground forum datasets. This thesis uses the
‘DarkJargon’ approach for a baseline for evaluation.

A different approach is taken by Yuan et al. [2018], who aim to identify ‘obfuscated’
words. They use an approach based on word2vec [Mikolov et al., 2013], however they
find that comparing word vectors from two separate models does not work. Instead, they
propose a change to word2vec’s skipgram model, which concatenates the word vectors
from one hot encoding. This is equivalent to changing the dictionaries of the two corpora,
prepend “A” to words used in one corpora and “B” to words used in another corpora.
For evaluation, they find the approach has a precision of 0.91, but recall is 0.772. Therefore,
of the predicted words, these are likely to be true positives, but their approach had many
false negatives, which resulted in a lower recall. In addition, the evaluation was limited to
just drug and cybercrime product names, and requires collecting a large ‘clean’ corpus for
model training.

2.2.3 Drift & Churn

This thesis (Chapter 4) explores the churn of users. Churn can be defined as users leaving
the forum and not returning. Users may leave the forum to join other communities
which provide greater interest, or they may instead no longer have interest in cybercrime.
Detecting this movement is out of scope of this thesis, and would require further analysis
methods. Similar concepts to churn exist in criminology, with the ‘drift’ [Matza, 1964] and
‘digital drift’ [Goldsmith and Brewer, 2015; Holt et al., 2019] frameworks, which explain how
drift into and out of criminal pathways can often be ‘accidental’ or ‘unpredictable’ [Holt
et al., 2019]. In drift theory, Matza points out offenders drift in and out of crime, enabled
by a loosening of social control [Matza, 1964]. According to digital drift [Goldsmith and
Brewer, 2015; Holt et al., 2019], an application of drift theory, forums and platforms
provide a mechanism for engaging and disengaging from discussions about hacking and
crime. In this thesis, the focus is more on measuring the drift aspects of the framework,
rather than the social control elements of the theory.

Churn in general social networks is commonly used for prediction tasks that use
feature-based approaches [Karnstedt et al., 2010]. The aim of this task is to identify
which users will churn, and to calculate the churn rate (number of members leaving in a
period divided by active members). For example, if “popular” members churn, then other
members may be likely to churn. Churn analysis has commonly been used in industry
to identify “high-value” users who may leave, so companies can target retention efforts.
However, in industry, such as with telecoms, users enter a service contract with a switching
cost. Forums contain a much lower switching cost, and members are weakly-tied. Within
forums, if the number of users becomes too small, and network effects diminish, this can
cause a decrease in marketplace activity and forum posts. There are differing incentives at play here. Forum administrators will want to retain users so their community continues to flourish, while still actively moderating their platform, removing members who break platform rules. Counter to forum administrators wanting to avoid the death of a forum, interventions by law enforcement aim to disrupt cybercrime economies, with the removal of key actors likely to have a disproportionately large effect on the network.

There is not one single type of churn [Karnstedt et al., 2010]. Churn can be typical (members stop posting), holiday (members stop posting for a period of time but later return), bursty behaviour (members post infrequently, leading to misidentification of churn), and inconsistent behaviour. Churn has been used to explore the relation between a user’s value in a community and the probability of a user churning [Karnstedt et al., 2011]. This used a user-to-user thread reply graph for features, with edges weighted by the number of replies, and forum specific metrics for users. Metrics include the average length of posts started by a user, average length of thread the user participates in, popularity of posts, initialisation of posts, and polarity.

Less work has focused specifically on churn within cybercrime forums. One study used private message interactions from three carding forums [Garg et al., 2015]. They compare the change in topology (structure) of the network between members banned on the forum to the regular churn of users on the forum. While two of the forums did not have significant results due to a low number of bans, one forum found the change in small world structure between the two types was similar, finding that bans did not affect the overall topology. The authors also use the Louvain method [Blondel et al., 2008] to cluster communities in the forums, with latent Dirichlet allocation (LDA) [Blei et al., 2003] topic models to discover topics, to identify specialities. They find most communities sub-specialise in specific crimes, with smaller two-tiered communities of 100-230 members, and larger multi-tiered communities. While the study is useful, it is not clear if this generalises to larger general-purpose hacking and cybercrime forums.

2.2.4 Reputation and Trust on Underground Forums

Communities on underground forums may use a reputation system as a proxy for trust. Dupont et al. [2016] looks specifically at a system on a hacking forum, which uses a weighted approach for feedback. Members of a higher status have greater impact on a user’s score: a new user posting positive feedback awards 1 point, whereas a moderator can award 5 or 10 points. They find that only a small fraction of forum members participate in the reputation system, and beginners are over two times more likely to report positive feedback of members compared to administrators.

Reputation systems can help members to establish trust on forums. Yip et al. [2013] explored trust among cybercriminals on carding forums, finding one key challenge of
needing to determine if another forum member can be a trusted individual, a dishonest trader (‘ripper’, who may provide worthless goods or sell products with backdoors [Dupont et al., 2016]), or a law enforcement associate.

To combat ‘rippers’ on forums, Dupont et al. [2016] note a sanction system used on the forum. Administrators may completely remove all of a user’s positive reputation feedback on the forum, leaving only negative reputation feedback. However, Lusthaus [2012] finds that such sanctions are not as useful, since there is no longer a large cost to switching profile: if a member has negative reputation, they can lose this negative signal by creating a new forum profile. Lusthaus compares this to conventional crime, where individuals have a known identity and need to increase anonymity, whereas in cybercrime, individuals start with no identity.

Work by Holt et al. [2016] looked into the role of trust signals in cybercrime marketplaces for stolen data. They use a zero-inflated Poisson regression model to explore the relationship between marketplace signals and reputation received. One finding showed that having negative feedback correlates with receiving more positive reputation votes. They hypothesise that this can be due to either a seller having a large enough user base, or rippers using positive feedback to obscure negative feedback.

Reputation on forums has also been used as a validation metric [Pastrana et al., 2018a], as a proxy for trust. However, little work has explored the limitations of this metric, as forum members may game this system to appear more trustworthy than they are. Vu et al. [2020] explore how one forum has replaced the reputation system with a contracts system, which aims to provide a public feedback log of self-reported transactions between members.
CHAPTER 3

METHODS

This chapter introduces some of the analytical approaches used in this thesis, including approaches from natural language processing and machine learning. While there are many different techniques that could be used in this research, the methods used were determined by constraints. Firstly, the dataset is over 80 GB in size, and contains mostly unstructured noisy (slang and jargon rich) text. Typical natural language processing and machine learning models are trained on large corpora of websites which include forums such as Reddit, which are useful to provide comparative benchmarks but performance decreases on underground forums. The use of LLMs which are trained on large corpora can be useful to reduce the amount of additional labelled data required, however this depends on if the corpora includes vocabulary of the forum to be analysed. Also, no training data exists for this type of data, and therefore this limits the type of machine learning that can be carried out. Papers that utilise training data involve manually creating training data, which limits which type of model can be used. For example, large language models require large sets of annotated data in the target domain (for example, hacking community-specific jargon and slang) when used for classification tasks, and require access to GPUs. For the work described in this thesis, GPUs were not used in training models, and annotations were manually created by academics familiar with cybercrime topics. In other fields, annotations may be outsourced to services such as Amazon Mechanical Turk, but this raises ethical concerns over sharing cybercrime forum datasets to unknown individuals, and the quality of labels will be of lower quality compared to domain experts. Methods outlined in this section are chosen due to these limitations, in addition to selecting those that can provide useful insights for criminological and social science scholars.
3.1 Data

This thesis uses data from CrimeBB [Pastrana et al., 2018b], available from the Cambridge Cybercrime Centre for academic research use under a data sharing agreement. CrimeBB contains data going back over 20 years from underground hacking forums, with differing languages, sizes, and topics. The majority of this thesis focuses on the HackForums subset of CrimeBB, the largest forum in the dataset. The snapshot of HackForums used for this analysis contains over 680,000 users, over 190 administrator-curated subforums, 42,000,000 posts, and 4,000,000 threads, from 2007 to 2020. This dataset is selected as the timescale of data available allows for longitudinal analysis of the evolution of the forum, and the scale provides a more representative view of the open cybercrime platform ecosystem. However, note that findings in this thesis may not reflect activity patterns found on smaller closed and more profit-oriented cybercrime discussion platforms. This subset also includes reputation scores: positive or negative votes can be sent between users to build a feedback system, to enable trust and trade. While the forum is quite large in size, the majority of members are inactive lurkers, drifting off posting activity. As the dataset only contains posts, not views, it is not possible to tell if these members have completely stopped engaging with the forum. There is a small concentration of highly active users, and a subset of these may be considered ‘key actors’ in the cybercrime literature: members of the forum that could be of interest to law enforcement.

In Chapter 5, the forum Multiplayer Game Hacking (MPGH) from CrimeBB is also used in analysis, to explore trends. This is a smaller forum than HackForums, focusing on sharing gaming-related hacking techniques. The snapshot of MPGH used for analysis contains over 9,700,000 posts, 785,000 threads, and 511,000 users from 2005 to 2020.

Each forum is structured by subforums, which are based on general topics e.g. hacking methods or marketplace, and are defined by the forum administrators. Each subforum contains threads, which are an ordered collection of posts focusing on a defined topic set by the first post in the thread, such as a particular tutorial the author is sharing. Later posts can be providing a reply to the original first post, a reply to a later post by another user, or new information on the topic. While threads are typically focused on a particular topic, longer threads may become off-topic.

As the dataset comes from an automated scraper, there is likely to be occasional data quality issues, and as the data is unprocessed, data cleaning needs to be carried out. For topic modelling work, chunks of the forum post text are removed which are not the main content of posts, including quote, link, and code blocks. These are identified by using regular expressions to identify relevant markup blocks. This approach is specific to the CrimeBB dataset. Then, the dataset is tokenised using nltk’s [Bird et al., 2009] TweetTokenizer, selected as this tokenizer is suited to handling URLs and punctuation
based emoticons. This tokenised form is stored in the database for quick retrieval, as tokenising is a computationally expensive step.

### 3.2 Ethics

All of this work uses data from individuals on forums. This raises ethical questions to limit harm arising from the collection and analysis of such data [Thomas et al., 2017]. This research uses only data collected through the use of web scrapers from publicly available forums, and informed consent was not gained from all members as this would be considered to be spamming. The British Society of Criminology’s Statement of Ethics [British Society of Criminology, 2015] recommends for online research into criminal activity, informed consent may not be required where data is publicly available, and the research outputs focus on collective rather than individual behaviour. This work aims to not offend the users of the forums studied, and findings are reported objectively.

In addition, steps are taken to avoid publishing details that could identify individuals, including usernames and original post contents. Usernames are not included in any research outputs, and post contents is paraphrased to limit searchability. Care is taken to handle any personal data collected securely and sensitively. All research presented in this thesis has been approved by the Department of Computer Science & Technology’s ethics committee.

In the unlikely event that evidence is found of a vulnerable individual at risk of coming to direct harm, or illegal content is found, there is a legal duty to report such content, with guidance provided by the University of Cambridge’s Research Strategy Office. Precautions are taken if links are followed in a post, any pasted links to images in posts are to be treated with caution, and files are not be downloaded.

Data used in this work is stored securely, and access is limited to just the researchers and collaborators involved in these projects. Access is granted for a limited time only, and after work is complete, data is removed from devices where analysis took place.

CrimeBB data may pose minimal psychological risks for researchers, which is mitigated by researchers meeting regularly to provide an overview on progress and offer support to other members. The University of Cambridge offers counselling services, which the researchers who participated in this project can turn to if they require.

The outputs of this work will be of public interest for the research community, and there is little potential harm to the forum members, even though they might be involved in criminal offences, as they will not be identified. When publishing research, it is important to carefully consider whether any of the results from the analysis could be used to cause harm to particular groups, and results published are moderated accordingly.
3.2.1 Limitations

The methods and dataset used in this thesis are subject to limitations. Depending on the forum, members who have deleted their accounts prior to data collection may not have their data available. The HackForums dataset contains posts of members who have deleted their account, thus their post becomes “anonymous”. A placeholder value for the user ID (-1) is used instead, and it is not possible to match posts together by user. Therefore, it is not possible to measure start and end posts for this group of posts without associated user IDs. In this thesis, this group is excluded from analysis to avoid results containing a very active ‘-1’ user, but note results may be limited due to this. With all measurements, it is important to lookout for anomalies in results which could be caused by data quality issues or limitations. The dataset contains 42,165,425 posts, with 65,154 posts (0.15%) having the placeholder user ID (-1).

Thread creation dates are not available, but this can be approximated from the first post creation date in the thread. First posts can be obtained by sorting by post id, to overcome an issue where the dataset had incorrectly parsed AM and PM timestamps, and therefore sorting by date for a thread that contains posts before and after midnight can cause the ordering to be incorrect.

The dataset only contains data for users who have actively posted at least once. It does not include users that register with the forum and read about topics, consuming material but never actually posting publicly. This can result in bias. The dataset has post data for each forum profile, and note that forums users may use more than one profile. Stylometric methods could be used to group these profiles together, however analysis in this thesis assumes each profile is a unique user and this further analysis is left for future work.

In addition, there have been some data quality issues. An issue was found where 24-hour times were parsed as 12-hour times in the dataset. This was detected during annotation work when labelling the first post in each thread would show a post that did not look like it was the first post (i.e. a reply to a previous post). This was confirmed to be the case in the database, as sorting by PostID and Date did not return the same ordering for some threads. This related to an old bug in the code which has since been fixed, highlighting the importance of storing source files so parsing could be re-run.

Also, reputation data contained issues that could affect measurement results. On multiple occassions, forum administrators had reset the reputation system, leading to users sending automated bursts of reputation votes to try to re-gain their scores. This can result in non-representative data if reputation votes are summarised without taking this limitation into account.

Forum data contains a partial structure. Board categories, boards, and threads provide ‘ground-truth’ structure to measure, but post content does not. Topic models can provide
an approximation of discussions, but note this is not a perfect representation of content. Parameters are chosen using coherence scores and manually validate detected topics, but the number of topics chosen is selected by the user (i.e. there is not a “perfect” number of topics). A greater number of topics can be chosen to create a granular model, however this can be harder to visualise.

3.3 Natural Language Processing

Natural language processing (NLP) approaches can be used to analyse the text content shared by members. However, many existing NLP tools are trained on large corpora of websites, which introduces issues when using these on smaller forum data containing differing vocabulary not included in the corpora. Posts include slang and jargon, variations on words, and sentences may not be well-formed. In this thesis, NLP is used to understand what topics are discussed, and explore the changing use of slang used on the forum.

Prior literature has explored automated techniques for creating dictionaries [Schwartz and Ungar, 2015] of terms used on social networks, and question-answer forums, a form of information sharing communities, similar to cybercrime forums. With the large amount of data on question-answer forums, Farrell et al. [2001] proposed a technique for summarising threads, to allow returning users to catch-up quickly, by selecting sentences based on a score of salience of words in sentences, instead of selecting whole posts. This could be applied to cybercrime forums to quickly understand current trends of forums based upon summarised posts.

A different approach to make sense of these large datasets can look at the individual interesting users of the forum, by measuring the changing activity of experts. Fu et al. [2016] identifies experts by looking at their changing behaviour over time, using temporal-based features, for later analysis. Pal et al. [2012] explores the set of experts further using clustering to find different types: experts who are consistently active, experts who are only active initially, and experts who are only active later on.

3.3.1 NLP and Cybercrime Discussion Platforms

NLP techniques have been used for a variety of studies of cybercrime discussion platforms and marketplaces, including prediction of private messages using text-based features of public posts [Overdorf et al., 2018], identifying key actors by including stylistic features of posts [Abbasi et al., 2014] with social network analysis features, understanding evolving terminology, and classifying members by post content styles to identify which features have greater importance [Lui and Baldwin, 2010].
3.3.2 Evolving Terminology

Yang et al. [2017] look for new keywords in Chinese-language datasets which are used to describe illicit topics. They use a filter to match similar words, and a search engine to find the meaning of these words, instead of using NLP or ML approaches. Yuan et al. [2018] use a different method to look at new keywords, using word2vec to compare keywords to known terminology on different platforms, including dark, legitimate, and reputable websites, to determine where these terms may be used. Benjamin and Chen [2015] also use word embeddings to look at language used on underground hacking forums, to help understanding by showing related terms for a given term.

3.3.3 Trending Topics

Trending topic tools cover a range of use cases, for looking at both historic datasets and real-time datasets, and typically focus on microblogging platforms, such as Twitter. Common techniques for trending topics detection include TF-IDF (term frequency inverse document frequency) and LDA (latent Dirichlet allocation). Term-frequency inverse-document-frequency (TF-IDF) [Jones, 1972] identifies common terms in a document (e.g. forum thread), but not common across all documents (e.g. the whole forum), which can be used to cluster and classify documents. This technique provides a mechanism for ranking tokens which are “important” to a document. However, forum text is noisy, with varying spelling of words and creative use of punctuation. While TF-IDF is a popular NLP technique, use on forum data would require stemming or lemmatisation, and defining a document either as individual posts, or a thread of posts, for best performance.

TF-IDF assumes each document is based on a single topic, although with forum data, posts and threads may discuss several topics. LDA [Blei et al., 2003] takes a different approach by assuming each document is built from a number of topics, with posts weighted over each topic, by learning a distribution of terms in topics. Similar to TF-IDF, this method also requires finding a suitable tokenisation approach and representation of a document. LDA learns the distribution of terms in topics, which can be used to classify or cluster documents. NLP techniques can either use unigrams (single words), bigrams (pairs of words), or n-grams (n-length list of words). While LDA learns a distribution of terms in topics, this is not as lightweight computationally as TF-IDF.

Burst and dynamic topic models have been used for detecting trending topics, including a burst model proposed by Kleinberg [2003]. Takahashi et al. [2012] and Koike et al. [2013] combine Kleinberg’s burst model with a dynamic topic model. While these approaches measure frequency changes over time to detect bursts, Chapter 5 uses a different approach similar to “two-point trends” discussed by Kleinberg [2016] with “rising” and “falling” words. In addition, Chapter 5 uses a Bayesian approach instead of measuring absolute
However, these techniques have limitations, and improved models have been proposed. Aiello et al. [2013] explores the use of common NLP methods for detecting trending topics on Twitter, related to major events which differ in time scale and topic churn rates. Methods used included document-pivot (one document per topic), which suffered from cluster fragmentation problems and noise, and feature-pivot (e.g. one word per topic) which may capture misleading term correlations. While they found standard NLP techniques are suitable for a small number of focused topics, they suggest using novel techniques for topics evolving in parallel. The best performing methods were n-gram co-occurrence (looking at groups of words that typically appear in the same document) and DF-IDF topic ranking (similar to TF-IDF, although this looks for common topics in a given time period, which are not common in prior time periods). They carried out post-processing to boost the score of proper nouns, as they find these are useful keywords for trending topics.

Follow-up work by Martin et al. [2015] detects bursts of phrases for a topic detection system, using DF-IDF to group co-occurring bursty phrases, followed by topic ranking, using the apriori algorithm [Agrawal and Srikant, 1994] which identifies frequent item sets (n-grams). They also look at windowing, where events which are focused on real-time activity (e.g sports) have a smaller window of activity, with greater topic recall than longer topics (e.g. politics) and discussions continuing after events. Super Tuesday (the Tuesday in which many US states hold their primary elections) performed better with fewer prior Tweets as this was a longer event, than shorter events such as sports games which performed better with a longer window. They also looked at topic recall for different sized windows, and found the apriori algorithm [Agrawal and Srikant, 1994] performed well for ranking.

Previous research has focused on static snapshots of events, whereas Shamma et al. [2011] use temporal analysis, to identify both peaky and persistent topics. They use normalised term frequency, with the number of tweets containing the word, rather than the number of times a word is used, and the peaks look at terms particular to an exact window of time. Persistence looks at peaks of normalised term frequency, assuming these terms have not been used before, and have been used more frequently afterwards.

While much of the literature focuses on detecting English-language trending topics, many cybercrime forums are not English-speaking, which can add complexity into analysis. Also, there are some cases where topic modelling may produce poor quality results, and could be refined with user feedback, which is explored by Hu et al. [2014] with iterating models (hierarchical-LDA trees). Additionally, topic models can assist qualitative analysis. Kigerl [2018] used clustering with a topic model across different carding forums, to group topics found by LDA into categories of discussion topics including types of markets.
3.3.3.1 Trending Topics: Fightin’ Words

As part of work into trending topics on underground HackForums, this thesis (Chapter 5) uses a method based upon work by Monroe et al. [2008] who aimed to identify words used by political parties. This is similar to work on trending topic analysis, although the data in their analysis is labelled by topic (political parties). This uses a model-based approach, modelling terms as a function of political party, to compute the likelihood of terms used by a political party as log likelihoods (“log-odds”). They used an uninformative Dirichlet prior. They find classification methods are not suitable for this task, as they can overfit, and many approaches to feature selection problems do not use probabilistic models. However, they find this type of model does not make the same assumptions as classification techniques, and can be used for analysis techniques in addition to classification. They experiment with other techniques, including odds ratios and TF-IDF and WordScores, but find their Bayesian approach to be superior for their task.

Grimmer and Stewart [2013] survey the usefulness of automatic content analysis, including work by Monroe et al. [2008], noting the importance of continually validating model outputs. Visualisation tools can support validation, such as Kessler [2017], which adapts Monroe et al. [2008] to show which terms are distinct to each corpus. Kessler [2017] uses a scatter plot combined with clustering to offset x and y values of the plot into visually distinct ‘strands’ (visual line clusters of data points), improving readability and reduce noise.

3.3.3.2 Use case: Detecting Emerging Threats

While this work uses historical forum data, techniques have also looked at real-time streaming data, to detect existing and new threats. Large-scale frameworks such as DISCOVER [Sapienza et al., 2018] have been developed, for detecting emerging threats based upon known terms across Tweets and blogs. These systems require annotations for training threat detection models, which can benefit from annotation tools. Behzadan et al. [2018] releases a tool for exploring Twitter data, used to release an annotated dataset of 21,000 Tweets on cyber threats.

Techniques for detection have included using topic detection using LDA on source code and posts [Samtani et al., 2015, 2017], combined with a SVM for classification of threats, trained on known previous threats such as keyloggers and exploits. Once emerging threats have been detected, topic ranking [Bose et al., 2019] can be used to highlight known serious threats. Pre-processing approaches, including PoS (part-of-speech) tagging and sentiment analysis have been used to increase performance in identify threats [Macdonald et al., 2015], although this requires using models which can handle the variations of jargon used on forums. There have been other approaches to look at trends on forums and marketplaces,
including Tavabi et al. [2019] who use a large topic model to map the evolution of different forums as they evolve.

These communities also evolve over time, with changing meanings of words, and an evolving lexicon, which should be taken into account with longitudinal topic modelling. Bhandari and Armstrong [2019] have looked at subforums of Reddit to explore the use of high affinity terms used by communities, looking at how the semantics of these have changed.

### 3.3.4 Argot Detection

This thesis uses a method for argot detection utilising word embeddings. These place words into a space, providing semantic representations learned from examples of usage. While this work uses Euclidean space for embeddings, other approaches have used different types of spaces for the task of hypernym detection [Le et al., 2019]. However, when word embeddings are created for two separate corpora (e.g. cybercrime forum and Reddit posts), comparisons between them are not meaningful. Marchisio et al. [2021] address this problem by comparing Euclidean and graph-based alignment methods, for transforming the word embedding spaces. They find that their performance varies on context. In this thesis, based on their results, the Euclidean approach is used to align the two embedding spaces, using a set of annotations.

Some approaches for argot detection use supervised machine learning models. While these have acceptable performance for test cases, Queiroz et al. [2020] highlights the issue that if models are used over longer periods of time, performance can degrade as lexical changes are introduced to the forum. The authors suggest a relabelling approach could be used, and labelling additional data.

Further methods could be explored to support evolving lexicons. Ryskina et al. [2020] analysed how new words are likely to be formed, finding that both semantic sparsity (surrounded by few words in the embedding space) and the frequency of growth rate are predictive of this. Hamilton et al. [2016] aligned vector spaces across the corpora representing different time periods, to measure which words have changed in meaning over time.

### 3.4 Named Entity Recognition

Chapter 5 uses a Bayesian approach to identifying trending topics, with filtering by noun phrases using a part-of-speech (PoS) tagger. However, an alternative approach may use named entity recognition (NER) to detect trending topics, and later, for extracting events from text. However, Caines et al. [2018] note named entity recognisers are trained on well-formed English text, and their performance is degraded with noisy text.
There has been prior work in using NER on noisy text, including with a shared challenge at W-NUT 2017 [Derczynski et al., 2017]. One approach by Aguilar et al. [2017] used a convolutional neural network with both character-level and word-level features combined with contextual information, input into a bidirectional LSTM, for this task. Jansson and Liu [2017] also used a bidirectional LSTM for word and character embeddings, but combined these with an LDA topic model.

Additionally, contextual data can be used to assist with this task. Xing and Paul [2017] combined word embeddings with Twitter network and geolocation data to improve the accuracy of NER. While CrimeBB does not contain this type of data available about HackForums users, the forum structure provides hierarchy with administrator defined subforums, which could be used as a feature to combine with embeddings.

### 3.5 Machine Learning for Modelling Trajectories

Longitudinal studies measure the change of variables over time, and modelling approaches can be applied to detect common patterns. Techniques include group-based trajectory modelling (GBTM) and k-means longitudinal (KmL).

#### 3.5.1 Modelling Trajectories

When analysing changes of communities in cybercrime forums over time, it may be useful to cluster changes using time-series data. Modelling techniques such as group-based trajectory modelling (GBTM) [Nagin and Odgers, 2010] have been used in criminology to model the development of offenders, by fitting time-series data to curves, and have implementations in common statistical packages. However, model fitting is manual, not scaling easily to large datasets.

An alternative approach by Genolini and Falissard [2010] uses k-means for longitudinal (KmL) data, a popular clustering technique in computer science applied to time-series data. They note while clustering is often used for exploratory analysis, parametric-based models such as GBTM have intrinsic assumptions of what the data looks like to fit the data to curves, and find KmL can be more flexible as the method is not model-based, however it can be harder to test goodness of fit. They compare this method to GBTM, and find they have similar results when trajectories are polynomial curves, and KmL has better results for non-polynomial trajectories. This approach is aimed at analysis in epidemiology, to improve existing clustering techniques.

Clustering is used in this thesis for looking at groups using argot over time (§6.4.2). This work found both GBTM and KmL were not suited for this task. Instead, Gaussian Mixture Modelling was used for analysis, in which the underlying algorithm uses an approach similar to GBTM.
From statistics and signal processing, Farajtabar et al. [2017] look at tracking changes in a social network, to develop their system “COEVOLVE”. This uses a point process model to track changes in a social network, across both network changes and information diffusion, at the same time.

3.5.2 Interpretable and Explainable Models

Using machine learning models to identify trend patterns can provide valuable insights, however models cannot always explain how these patterns are found or classified. Interpretable and explainable models provide a step towards providing understanding of model results. For example, the argot classifier (§6.3) can provide predictions but cannot explain why those predictions are made.

Interpretability and explainable models are hot topics in machine learning, and important in cybercrime research for understanding results of prediction and clustering, instead of producing black-box predictions. These two have slight differences: you can directly inspect the decisions in interpretable models, but cannot with explainable models. Rudin [2019] found explainable machine learning models often approximate, not exactly matching the model. Instead, they suggest using interpretable machine learning for directly inspecting the model. Also, interpretable models can be used to understand the vulnerabilities or limitations of a model which adversaries could exploit [Hutchings et al., 2019].

The use of complex black-box models, such as convolutional neural networks, can be used for better classification with a complex decision boundary [Deliu et al., 2017]. However, it is not trivial to interpret these models. There have been approaches developed to explain predictions of classifiers, such as LIME [Ribeiro et al., 2016] which looks at the decision boundary locally on examples of text classification and image classification. They note this approach is suited for models which are not easily interpretable (e.g. random forests), instead of those that are (e.g. Naive Bayes). While the argot classifier (§6.3) cannot explain why predictions are made, future work could develop interpretable and explainable models to provide researchers with insights to which features define argot.
CHAPTER 4

MEASURING THE EVOLUTION OF CYBERCRIME FORUMS

Cybercrime forum datasets are large and complex. Prior research uses aggregated time series data to create a picture of the whole dataset, or focuses on a smaller sample of cross sectional data, often for a specific subcommunity or crime type. This chapter is based upon the paper by Hughes and Hutchings [2023b], which uses the longitudinal time series aspect of cybercrime forums to measure and observe the evolution of forums at a macro scale. Applying the digital drift theoretical framework, borrowed from criminology, this research uncovers a large amount of churn on the forum, with only a small proportion of users continuing long-term engagement. Measurements show a continual shift in forum activity, with year-based cohorts moving from starting in hacking discussions, towards starting in general discussions, and later towards eWhoring (a fraud type where intimate images are used for financial gain). The group of members who are active on the forum for over 12 months, typically have their last post in the marketplace, while members active for shorter periods of time have their last post in hacking-related boards. Overall, there is an increasing trend towards financially-driven cybercrime, at both the user and forum level. Users post more in financially-related boards over time, and forum activity has trended away from gaming/social activity, trending more towards market-related boards.

4.1 Introduction

Cybercrime forums have been studied in detail, from economic analysis of marketplaces [Afroz et al., 2013; Akyazi et al., 2021; Alldi, 2017; Franklin et al., 2007; Fu et al., 2010; Garg et al., 2015; Kumar and Bhargavi, 2020; Motoyama et al., 2011; Park et al., 2018; Vu et al., 2020; Yue et al., 2019; Zhang and Li, 2013], to identifying individual key actors [Benjamin and Chen, 2012; Pastrana et al., 2018a; Sundaresan et al., 2016], to
exploring certain types of activities in detail [Atondo Siu et al., 2021; Caines et al., 2018; Hutchings and Pastrana, 2019; Pastrana et al., 2019], such as eWhoring [Hutchings and Pastrana, 2019; Pastrana et al., 2019]. While much prior work has focused on specific aspects, such as analysing cross sectional data at one point in time, this research takes a macro view of the forum over time, to explore evolution of the forum, and the changing behaviour of the forum users. While these forums have a large number of users, only a small fraction of these members continue to remain active, while other members “churn” or “drift”. The online aspect of cybercrime forums enables members to disconnect easily from their online profile, and drift in and out of forum activities due to weak ties. In addition, there is an increasing shift towards marketplace related discussions after joining the forums.

Cybercrime forums are not static. They evolve over time: growing from a small set of users to a large ecosystem of many subcommunities, and sometimes reducing back down in size. Datasets constantly evolve, and need to be kept up to date to follow and track changes. Changes on the forums can be observed at three different levels:

**Micro**: At the micro layer, the changing interests and aims of users can be observed over time [Benjamin and Chen, 2012; Bhalerao et al., 2019; Pastrana et al., 2018a; Sundaresan et al., 2016]. These changes can be grouped into pathways on the forum. This can include types of interactions made: moving from social posts to tutorial posts, moving from only posting in existing threads to creating their own, sharing tutorials, participating in wider general social activity, and engaging in marketplace activity. While users may post to the same board about the same topic, there could be change of intention e.g. development of a tool, followed by selling the tool.

**Meso**: At the meso layer, subcommunities of members discuss topics relevant to their interests, such as to discuss a particular crime type [Atondo Siu et al., 2021; Caines et al., 2018; Hutchings and Pastrana, 2019; Pastrana et al., 2019,?]. These subcommunities change over time, growing and shrinking at differing rates. They may use jargon unique to their subcommunity, and contain both members focused solely on a particular type of activity and members spread across different areas of the forums. The definition of these subcommunities depends on what the research task is.

**Macro**: At the macro level are the large scale forum trends over time [Afroz et al., 2013; Franklin et al., 2007; Garg et al., 2015; Motoyama et al., 2011; Vu et al., 2020], including across forums, such as measuring the rising and falling of keywords and boards, across all members. Also, dictionaries or tools for detecting jargon in the forum (collecting spelling variations and matching to general categories) can be used to track changes. Other large scale trends include the churn of users on the forum: those that join and leave in the same year may have different intentions to those that stay for multiple years.

Most macro research of forums uses the data as a “static” dataset to collect aggregate
statistics over the entire forum. While aggregate statistics are useful for measurements, users are not stable data points. Topics of interest and activity levels vary over time, and longitudinal analysis is needed to understand how this engagement fluctuates, which is the main contribution of this chapter. Users who are more active on forums are not necessarily more involved in crime, as general discussion boards are often the most popular type of board posted to, provide a community building mechanism for users, which can help forums to retain users beyond solely technical and marketplace posts.

This chapter uses data from CrimeBB, a dataset available for researcher use from the Cambridge Cybercrime Centre, consisting of over 20 forums and 100 million posts spanning over 20 years. Focusing on HackForums, the largest and longest-running English-language forum, board activity is measured, the churn of users on the forum who drift off the platform, and the shifting interests over time for users with different levels of activity.

This chapter explores how different cohorts have varied in scale over time, and changes in different topic interests, to show how users churn on the forum, with year-based cohorts halving in size every 12 months. This chapter takes a longitudinal approach at analysing and describing changes to this dataset over time, which is useful for later research into cybercrime pathways, as the topics of interest and intentions of users evolve.

This chapter uses the digital drift theoretical framework to explore trends in user cohorts, and the forum overall, over time. This theory prompts thoughts on the continual ebbs and flows of user engagement. As new members join and other members leave, it is possible to hypothesise a change in topics on the forum due to the large changing community. For example, a topic may be popular in one year and then decrease as users drift off the platform and new users join. Also, as the community changes, the value of the community to members changes, leading to fluctuations in retention levels. Drift theory helps to guide the exploration of ideas around the changing community and the effect this has on the forum. This chapter addresses the following research questions:

- RQ1: How has the size of the forum, and topics of discussion, changed over time?
- RQ2: What are popular topics among user cohorts over time?
- RQ3: How do retention levels change over time?
- RQ4: With new members joining the forum and old members leaving, what effect does this have on the interests of forum members?

The effect of churning on the forum is studied, and topic models are used to explain how the cohort interests have changed. Over time, the forum has become more financially driven, in terms of the activity on the forum as a whole and as groups of users turn to the economic aspects of cybercrime.

Contributions include:
Categorisation of members to filter active users from inactive users, to separate analysis of users that post a few times and leave from those that remain active on the forum for multiple months.

Measurements of churn on the forum, finding only a small proportion of users continue engagement.

Measurements of starting on the forum, including finding a shift from beginner hacking discussions towards general, then towards eWhoring discussions. This is also shown to occur month-on-month after joining the forum.

Measurements of last posts on the forum show for members posting less than 12 months, they post on hacking boards, whereas members who are more active are typically posting in the marketplace before leaving.

4.2 Background

This chapter explores cybercrime forums, which are discussed in §2.1, and the churn of users. The background of churn is covered in §2.2.3.

4.2.1 Digital Drift

The combination of uniting digital drift theory (§2.2.3) with the practical aspects of social network measurements provides a unique perspective in the field. This uses measurements from standard longitudinal approaches and topic modelling, instead of using social network graph approaches, as the forum dataset does not have ground truth data of social interactions, and approximating these can lead to inaccurate results.

4.3 Methods

4.3.1 Data

Data used in this chapter is from HackForums, a subset of the CrimeBB [Pastrana et al., 2018b] corpus, discussed in §3.1.

4.3.2 Categories of Users By Level of Activity

Different forums have varying numbers of users. However, as these are types of social networks, they tend to have a 'long tail' of activity: few users post the majority of the content. Figure 4.1 breaks this down into categories of activity levels, and Table 4.1 lists
the number of members per category. These categories are selected from initial analysis of the dataset, using measurements of the data, to group members by similar activity levels. The **Unknown** category is for posts which the user has since deleted their account, with the placeholder ID ‘-1’. The second category, **New**, is of users that have less than 12 months of activity on the forum. Those with 12 or more months of activity are broken down into thirds of posting volume. The categorisation is useful to visualise the level of activity of forum members, and to begin to sample from the forum. Sampling is essential, as it is not trivial to get meaningful findings from the forum as a whole due to the size of the dataset, and the heterogeneity of user activity.

**Figure 4.1:** Categories on the forum by level of activity

### 4.3.3 Analytical Approach

This chapter uses the categories of activity levels to explore effects of drifting and churn on the forum, how the interests of users have changed over time, and their first and last

**Table 4.1:** Table of categories on the forum by level of activity

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>New</td>
<td>601,642</td>
</tr>
<tr>
<td>Low</td>
<td>12,789</td>
</tr>
<tr>
<td>Med</td>
<td>12,815</td>
</tr>
<tr>
<td>High</td>
<td>13,211</td>
</tr>
</tbody>
</table>
posts. Latent Dirichlet allocation (LDA) topic models are used to observe the trends of moving between first and last topics on the forum. Coherence is used with the $u_{\text{mass}}$ metric to select the number of topics with the greatest similarity for documents (posts) per topic. Then, these topics are combined for measurement where they have a strong overlap (e.g. “general” topics). This merging is used to group two topics of common forum messages, such as “please” and “thanks”, to simplify visualisations where we wish to focus on specific technical topics discussed.

Drift and churn is measured by looking at activity levels declining over time per type of board as defined by Pastrana et al. [2018a], namely common, hack, tech, coding, games, market, money, web, and graphics. Year-on-year cohorts are analysed, to identify proportions of users that remain active on the forum beyond the year they join in.

The interests of cohort years using topics and boards are measured. As the forum userbase grows, the forum takes on a wider range of interests and the topics included start to change over time. Also, the amount of users replying to a new thread (typically contributing something new or asking a question) are compared to those replying to an existing thread.

4.3.4 Limitations

Dataset limitations are discussed in §3.2.1. In addition, note members who post a single time on the forum and leave their accounts idle will be grouped into New. Later analysis focuses on members with regular posting activity to explore changes over time, and note that the analysis does not explore deeply this group of single post accounts. This chapter uses topic modelling, however note this may not perfectly represent the content discussed on the forum. Manual validation of topics is used to ensure topics are representative, by sampling the dataset and qualitatively checking the results, however there is no “perfect” number of topics.

Measurements show a shift towards marketplace activity from general chatter and hacking discussions. It is important to note that these are found in aggregated data: while there is an overall shift, members individually move into marketplace discussions at different rates and will each be interested in different topics. Overall trends are representative of groups, not individual users.

4.4 Results

Measurements are split into three themes. First, activity over the entire timespan of the forum to explore churn on the forum. Second, exploring how users get started on the forum, including which boards members post to first and continue to engage in for up to
6 months. Third, declining activity levels are explored for the users who had shown the most commitment to the forum (more than 12 months of activity).

### 4.4.1 Activity From Joining to Leaving

It is readily apparent that there is a large amount of drifting and churn occurring on the forum. When users engage with posts on boards for the first time, some users may continue to sustain an interest in these boards, while a significant proportion ‘drift’ away. Along with drifting, the forum also experiences churn. Instead of users changing their interests and moving away from boards, ‘churn’ looks at how long year-based cohorts remain on the forum. Across all activity categories, Figure 4.2 shows that only smaller groups of each year cohort continue to interact with posts and threads on the forum over time. Forum activity peaks in 2012, with a gradual decline thereafter.

![Figure 4.2: Churn of users on the forum](image)

Note that a small cohort of users who joined each year continue to be active in all subsequent years, including those who joined in 2007. This is more visible in Figure 4.3, which visualises the proportion of active users in each year after joining the forum. The year 2009 has the highest retention of users, which decreases year on year.
Further explored are the first and last posts of forum members, starting with the type of first post made, and measuring the change in topics from first to last post. Measuring whether a first post was a new thread or a reply in an existing thread is useful to show the difference between new users who first post to existing threads, contributing to existing discourse, compared to users opening new threads to request or share information and tools. This shows that the number of new threads remains steady over time, but the number of replies increases sharply with an influx of new users between 2010 and 2013. A majority of users reply to existing threads for their first post on the forum.

Over the entire duration of activity on the forum, users may change topics they are interested in. Two latent Dirichlet allocation (LDA) topic models are trained over first and last posts on the forum, using coherence scores to select the number of topics that each have the highest similarity within. Then the topics are validated, to check they represent real concepts. Table 4.2 shows the results of the LDA topic model for first posts, and Table 4.3 shows the results of the LDA topic model for last posts.

Figure 4.4 shows a Sankey diagram of first post topic to last post topic, using these two models. For each post, a topic is selected using the topic of greatest weight from the LDA model. Separate LDA models are trained on the first and last posts, with the Sankey diagram mapping using each user’s first and last post to link between the first topic and last topic. This shows ‘general’ and ‘requests’ are the most popular topics to
Table 4.2: Topic model result for first posts

<table>
<thead>
<tr>
<th>Number</th>
<th>Top 5 Topic Words (with Weights)</th>
<th>Topic Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.045*&quot;website&quot; + 0.027*&quot;page&quot; + 0.025*&quot;site&quot; + 0.024*&quot;script&quot; + 0.017*&quot;id&quot;</td>
<td>Web</td>
</tr>
<tr>
<td>1</td>
<td>0.073*&quot;thanks&quot; + 0.065*&quot;please&quot; + 0.055*&quot;pm&quot; + 0.055*&quot;would&quot; + 0.044*&quot;interested&quot;</td>
<td>General</td>
</tr>
<tr>
<td>2</td>
<td>0.028*&quot;file&quot; + 0.017*&quot;app&quot; + 0.013*&quot;files&quot; + 0.011*&quot;key&quot; + 0.010*&quot;use&quot;</td>
<td>Files</td>
</tr>
<tr>
<td>3</td>
<td>0.031*&quot;phone&quot; + 0.025*&quot;android&quot; + 0.025*&quot;hi&quot; + 0.021*&quot;video&quot; + 0.019*&quot;youtube&quot;</td>
<td>Mobile</td>
</tr>
<tr>
<td>4</td>
<td>0.023*&quot;discord&quot; + 0.020*&quot;account&quot; + 0.018*&quot;buy&quot; + 0.017*&quot;contact&quot; + 0.016*&quot;accounts&quot;</td>
<td>Account (Online)</td>
</tr>
<tr>
<td>5</td>
<td>0.057*&quot;good&quot; + 0.042*&quot;pack&quot; + 0.026*&quot;work&quot; + 0.025*&quot;size&quot; + 0.021*&quot;python&quot;</td>
<td>Pack</td>
</tr>
<tr>
<td>6</td>
<td>0.056*&quot;windows&quot; + 0.043*&quot;system&quot; + 0.033*&quot;software&quot; + 0.029*&quot;program&quot; + 0.016*&quot;apps&quot;</td>
<td>Software</td>
</tr>
<tr>
<td>7</td>
<td>0.091*&quot;code&quot; + 0.013*&quot;number&quot; + 0.011*&quot;return&quot; + 0.011*&quot;string&quot; + 0.010*&quot;bin&quot;</td>
<td>Coding</td>
</tr>
<tr>
<td>8</td>
<td>0.015*&quot;can&quot; + 0.012*&quot;new&quot; + 0.012*&quot;like&quot; + 0.011*&quot;know&quot; + 0.010*&quot;get&quot;</td>
<td>General</td>
</tr>
<tr>
<td>9</td>
<td>0.060*&quot;help&quot; + 0.058*&quot;need&quot; + 0.049*&quot;please&quot; + 0.040*&quot;account&quot; + 0.038*&quot;link&quot;</td>
<td>Request</td>
</tr>
<tr>
<td>10</td>
<td>0.022*&quot;server&quot; + 0.020*&quot;rat&quot; + 0.019*&quot;ip&quot; + 0.019*&quot;use&quot; + 0.014*&quot;help&quot;</td>
<td>RAT</td>
</tr>
</tbody>
</table>

Table 4.3: Topic model result for last posts

<table>
<thead>
<tr>
<th>Number</th>
<th>Top 5 Topic Words (with Weights)</th>
<th>Topic Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.031*&quot;learn&quot; + 0.027*&quot;looking&quot; + 0.022*&quot;hello&quot; + 0.020*&quot;add&quot; + 0.020*&quot;hacking&quot;</td>
<td>Help</td>
</tr>
<tr>
<td>1</td>
<td>0.062*&quot;code&quot; + 0.015*&quot;id&quot; + 0.014*&quot;text&quot; + 0.012*&quot;source&quot; + 0.012*&quot;pi&quot;</td>
<td>Code</td>
</tr>
<tr>
<td>2</td>
<td>0.021*&quot;help&quot; + 0.017*&quot;need&quot; + 0.015*&quot;know&quot; + 0.013*&quot;get&quot; + 0.010*&quot;someone&quot;</td>
<td>Request</td>
</tr>
<tr>
<td>3</td>
<td>0.124*&quot;discord&quot; + 0.058*&quot;service&quot; + 0.050*&quot;telegram&quot; + 0.037*&quot;software&quot; + 0.035*&quot;windows&quot;</td>
<td>Account (Online)</td>
</tr>
<tr>
<td>4</td>
<td>0.018*&quot;good&quot; + 0.016*&quot;like&quot; + 0.012*&quot;time&quot; + 0.011*&quot;people&quot; + 0.010*&quot;one&quot;</td>
<td>General</td>
</tr>
<tr>
<td>5</td>
<td>0.050*&quot;account&quot; + 0.027*&quot;accounts&quot; + 0.025*&quot;phone&quot; + 0.023*&quot;need&quot; + 0.021*&quot;back&quot;</td>
<td>Account (General)</td>
</tr>
<tr>
<td>6</td>
<td>0.082*&quot;please&quot; + 0.071*&quot;thanks&quot; + 0.064*&quot;pm&quot; + 0.055*&quot;send&quot; + 0.037*&quot;would&quot;</td>
<td>Request</td>
</tr>
<tr>
<td>7</td>
<td>0.036*&quot;bot&quot; + 0.030*&quot;contract&quot; + 0.029*&quot;selling&quot; + 0.020*&quot;a&quot; + 0.017*&quot;bitcoin&quot;</td>
<td>Market</td>
</tr>
<tr>
<td>8</td>
<td>0.043*&quot;file&quot; + 0.018*&quot;open&quot; + 0.018*&quot;version&quot; + 0.017*&quot;download&quot; + 0.010*&quot;pc&quot;</td>
<td>Files</td>
</tr>
<tr>
<td>9</td>
<td>0.030*&quot;free&quot; + 0.021*&quot;money&quot; + 0.017*&quot;game&quot; + 0.013*&quot;payment&quot; + 0.011*&quot;video&quot;</td>
<td>Money</td>
</tr>
</tbody>
</table>

post about. Other joining topics include discussions of online accounts, remote access trojans, and packs. Leaving topics include general help, general discussions of accounts, money, and marketplace. This may indicate that members find information they need before leaving, or they stop participating on the forum after posting a marketplace listing. General discussions may include user introductions, with topic words including ‘I’m new’, the other topics provide clues into some of the reasons why users chose to join the forum.

Modelling using only the first and last posts reduces the computation time required to train the model. However, this has limitations. Modelling reduces complexity across the different activities of users, such as one user that only posts twice in a month, and another user that posts a range of content across different topics during their time on the forum, reducing these down to two sets of data points. This approach is chosen to simplify the visualisation, however more time periods can be added with the trade-off of greater computation time and a visualisation with more complexity.
4.4.2 Starting Off

In this subsection, the start of topic and activity pathways that exist on the forum are explored. By breaking the first posts down by board per year, it is possible to observe where these posts concentrate over time. Figure 4.5 shows the first post as a proportion of max in a year (max is 1.0, other boards are a proportion of this). This shows the top 10 boards, rather than all. Using the year-by-year proportional heatmap, new members appear to start posting in Beginner Hacking up to 2012. When the new user counts were at a peak, new members made their first post in Hacking Tools and Programs. Then, following a decline in the number of new users, The Lounge and Introductions were used to make their first post. More recently, since 2019 eWhoring, in which users use ‘packs’ of explicit images to take money from victims, has been a popular board for users making their first post.

While Figure 4.5 shows how users’ first posts change over time, the evolution of cohorts can be observed, with changes in the boards posted to by members joining in various years. Figures 4.6, 4.7, and 4.8 show members posting to beginner hacking for the first time in 2010 and 2015 and 2019 respectively, with a column for each month after the joining month. 2010 is selected as this is when forum has become established, 2015 follows peak activity on the forum, and 2019 is the last full year of data analysed. This shows that the interests of year cohorts changes over time. Colour coding is used to indicate the type of
board: ‘Financial’ for financial-related discussions including market and money making activity, ‘Gaming/Social’ for both gaming and general social chat, and ‘Knowledge’ for exchange of ideas and tools.

The 2010 cohort (Figure 4.6) has most users engaged in knowledge boards around hacking techniques, with a smaller proportion interested in general chat. Over time, there appears to be a shift to the majority of members involved in general chatter and marketplace related boards. The 2015 cohort (Figure 4.7) is immediately more interested in general chatter, and there also are discussions of gaming. Again, there is a shift towards marketplace related discussions beyond the first month.

The 2019 cohort (Figure 4.8) is primarily interested in financial related boards, specifically ‘eWhoring’, and general chatter. Overall, this shows a marked reduction in knowledge exchange boards, and an increase in interest in money making and market-related boards. This includes both at the cohort level (within each figure) and across the forum as a whole. Gaming and social related boards remain important to users throughout, indicating that there is a consistent level of community building and social interaction that is important over time.
Proportion of members active in each board (each month after joining) - starting in 2010 with first post in any board.
| Proportion of Members Active in Each Board (Each Month After Joining): Starting in 2015 with First Post in Any Board |
|---|---|---|---|---|---|---|
| **0 Months After Joining** | **1 Month After Joining** | **2 Months After Joining** | **3 Months After Joining** | **4 Months After Joining** | **5 Months After Joining** | **6 Months After Joining** |
| The Lounge | 17.43% | The Lounge | 5.96% | The Lounge | 3.19% | The Lounge | 2.64% | The Lounge | 2.29% | The Lounge | 1.92% | The Lounge | 1.83% |
| Hacking Tools and Programs | 12.97% | Marketplace Discussions | 1.97% | Premium Sellers Section | 1.7% | Premium Sellers Section | 1.44% | Premium Sellers Section | 1.31% | Premium Sellers Section | 1.3% |
| Beginner Hacking | 11.11% | Premium Sellers Section | 1.97% | Marketplace Discussions | 1.4% | Marketplace Discussions | 1.2% | Marketplace Discussions | 1.09% | Marketplace Discussions | 1.02% |
| Grand Theft Auto | 7.94% | Beginner Hacking | 3.08% | Buyers Bay | 1.42% | Buyers Bay | 1.33% | Buyers Bay | 1.32% | Buyers Bay | 0.99% |
| Marketplace Discussions | 5.8% | Hacking Tools and Programs | 1.65% | Beginner Hacking | 1.26% | Counter Strike | 0.98% | Counter Strike | 0.98% | Secondary Sellers Market | 0.92% |
| Premium Sellers Section | 6.35% | Buyers Bay | 2.36% | Counter Strike | 1.48% | Free Services and Giveaways | 1.09% | Counter Strike | 0.98% | Marketplace Discussions | 0.42% |
| E-Whoring | 6.02% | Free Services and Giveaways | 2.33% | Hacking Tools and Programs | 1.47% | Beginner Hacking | 1.13% | Secondary Sellers Market | 0.99% | Secondary Sellers Market | 0.88% |
| Counter Strike | 5.43% | Counter Strike | 2.23% | Free Services and Giveaways | 1.64% | Hacking Tools and Programs | 1.06% | 0-Whoring | 0.96% | Beginner Hacking | 0.81% |
| Remote Administration Tools | 4.99% | 0-Whoring | 1.27% | Secondary Sellers Market | 1.05% | 0-Whoring | 0.96% | Beginner Hacking | 0.79% | 0-Whoring | 0.78% |
| Botnets, IRC Bots, and Zombies | 4.97% | Remote Administration Tools | 2.03% | Secondary Sellers Market | 1.02% | 0-Whoring | 0.82% | Beginner Hacking | 0.79% | 0-Whoring | 0.7% |

**Figure 4.7:** Members posting to any board for the first time in 2015
## Proportion of Members Active in Each Board (Each Month After Joining) - Starting in 2019 with First Post in Any Board

<table>
<thead>
<tr>
<th>Month After Joining</th>
<th>New</th>
<th>Low</th>
<th>Med</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>56%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.8:** Members posting to any board for the first time in 2019.
4.4.3 Declining Activity

The digital drift theoretical framework hypothesises that members move away from activities, often unpredictably. This framework is used to explore how activity declines over time for different activity categories, comparing the types of boards posted to by those with low, medium, and high post counts. Figure 4.9 shows drifting occurring on various categories within activity categories Low, Med, and High.

The first subplot highlights the difference between activity categories. This shows the most active users have less variability in the types of boards they post in, while those that post less often are more likely to post in the market and hack categories. The following subplots highlights the declining activity of members over multiple months within each group. The market category contains the highest level of activity over time, and hacking...
boards contain a sharper decline over time compared to these.

While these subplots highlight differences between activity category groups, they assume that all members start on the same month. Next, an alternative approach is taken, shown in Figure 4.10. This shows the board in which users posted their last post, by year, aggregated into a small set of categories. This aggregation adapts Pastrana et al. [2018a] who categorised boards into the following categories: Coding, Common, Games, Graphics, Hack, Market, Money, Tech, and Web. Later boards established since this earlier work are added to the existing categorisation. For activity category New across all years ‘hack’ is the most popular board type, with a gradual rise in popularity of ‘market’ boards in recent years. For activity categories Low Med High ‘common’, ‘hack’ and ‘tech’ were popular leaving board categories, but from 2011 onwards, ‘market’ has become the most popular board type posted to before leaving. Overall, this shows a shift towards posting in the marketplace before leaving, with a more significant shift occurring among members of 12 or more non-contiguous months of activity.

### 4.5 Discussion & Future Work

This work has shown different aspects to forum activities, guiding analysis using the digital drift theoretical framework. There was a significant amount of drift and churn. Only a small proportion of users actively contribute to the forum for over 1 year of activity, with these members contributing to a larger number of unique boards. There also is a skew of members posting replies to other threads rather than creating their own. The majority of members ‘churn’, with less than 100,000 active members remaining on the forum over time. These churning members have all actively posted, indicating that there is something causing a lack of attachment to the forum. Future work should analyse what topics are discussed prior to these quickly–churning groups of users. For example, users may ask a
question and receive no reply, or have their question answered outside of the forum, or find that they do not have technical skills to carry out a hack, or post one marketplace advertisement and then leave.

Observing the churn year-on-year, a large number of users join the forum and leave within a single year, with typically less than half of users remaining. This significant level of churn reflects the idea of drift theory, with groups of users leaving the forum and a smaller group of more persistent users who remain on the forum for longer.

In measuring the interests of members joining the forum, there are movements of those posting in beginner hacking moving to more general and market oriented parts of the forum. Breaking this down further by first post per joining year, there is a shift from beginner hacking to general boards to the eWhoring board. However, the raw number of members joining is significantly higher during earlier years of the forum. These findings support the digital drift theory, showing movement out of hacking-related posts, towards more general chatter, followed by an emergence of interest in a specific activity type.

Grouping boards into categories shows first posts are typically in the hack category, and last posts are either in the hack or market category. In selecting users that are active for more than a year, a last post in the market category is more likely to occur.

Future work should explore trends in other similar forums. This work was limited to a single forum dataset, which may not be fully representative of the full cybercrime platform ecosystem. However, this forum was selected as it one of the largest cybercrime forums.

Future work should further use topic models to explore how the forum has changed over time, both overall and per year cohorts. Training topic models for large number of documents (posts) for different parameters (number of topics) is computationally expensive. However, it would be useful to explore the set of posts made by members when joining and leaving, and additionally sampling at a suitable frequency in between. Also, future work should explore the use of dynamic topic models [Blei and Lafferty, 2006], which are used to model the change of topics over time. This can help to build a set of pathways for groups moving through the forum. These pathways could be used to measure if groups are joining the forum for specific intentions (e.g. hacking), for looking at current popular discussion topics (e.g. eWhoring), and if groups joining for specific intentions drift into the group interested in popular discussion topics.

4.6 Conclusion

This chapter uses the digital drift theoretical framework to present an analysis of the evolution of forums over time. This includes a categorisation of members into groups, to filter out active users from inactive users. First churn is measured on the forum, finding only a small proportion of users continue engagement over multiple years, with
the majority of users leaving within a year. Initial topic modelling of users’ first and last posts highlighted a shift from general welcome messages and hacking-specific topics in their first post, to requests and market-related activity in their last. Exploring the change over time by year-based cohorts, there is a shift across first post topics on the forum. In the early days of the forum with many newcomers, first posts by members were posted in technical hacking boards. In the middle period of the forum, there was a peak in new members joining, and general discussion boards were the most popular type of board to start activity on. More recently, there has been a decline in the number of new users, and a clear shift of new members becoming active for the first time in marketplace related activity. Then, topics of members leaving the forum is measured, finding those with a lower level of activity have their final post on hacking-related boards, whereas members with over 12 months of activity typically place their last post within the marketplace. The findings overall show a shift towards financial related activity, but a reduction in forum activity as a whole, supporting the digital drift theory, and the underlying cybercrime pathways theory. This highlights that forums are not static, and the need to consider analysis using time-based approaches, to account for forum interests changing to new topics. For example, measuring the most popular topic of a forum may return a topic which was popular at the highest active period of the forum, but the forum may have moved to new topics since, such as marketplace related activities.
Cybercrime forum datasets contain a historical collection of post data, with long-term and short-term trends on discussion topics. However, while methods for detecting trends exist, these are limited to structured data, such as hashtags on Twitter. Trends on noisy forum text data could be detected using approaches such as topic models, however these are computationally expensive to train over a large corpus of 42 million posts with an unknown number of topics, resulting in a large parameter search space. This chapter presents a lightweight method for identifying currently trending terms in relation to a known prior of terms, using a weighted log-odds ratio with an informative prior. This method is applied to a dataset of posts from an English-language underground hacking forum, spanning over ten years of activity, with posts containing misspellings, orthographic variation, acronyms, and slang. This statistical approach supports analysis of linguistic change and discussion topics over time, without a requirement to train a topic model for each time interval for analysis. The approach is evaluated by comparing the results to TF-IDF using the discounted cumulative gain metric with human annotations, finding the method outperforms TF-IDF on information retrieval.

This chapter combines two strands of work looking at trending topics, and how they contain slang and jargon which has evolved over time. This chapter is based my published paper which created a tool to look at trending topics on the forum [Hughes et al., 2020]. The tool requires user input in searching for particular date ranges, and therefore more work may be required to show long-term trends automatically.

5.1 Introduction

Underground hacking forums contain a large collection of noisy text data around various topics, with misspellings, changing lexicons, and slang phrases. The evolving domain-specific lexicon includes homonyms, where “rat” may be identified as an animal by
off-the-shelf tools, but is typically defined as a “remote access trojan” in this context, a type of malware used to gain access to a victim’s computer.

This chapter uses post text from HackForums in CrimeBB. The dataset contains over a decade of text data, but detecting trends is non-trivial, due to the informal language used by members, not only technical terms such as “rat”, but also misspellings, slang, orthographic variation, and acronyms. For instance, the following texts demonstrate how posts are structured into threads on given topics, and how users both deliberately and accidentally use noisy language:

User 1:
*Ransomware infects hospitals all over UK: link*

User 2:
*anyone think they made some money from this?*

User 1:
*They might of done but idk they’ll get caught eventually, it’s stupid to commit crimes like this*

User 3:
*Who tf targets hospitals for ransomeware*

User 1:
*I dont believe they actually went for the nhs.. the ransom would be more $$$ lol*

User 4:
*I looked up a few btc addresses and can confirm they made money*

Researchers interested in analysing hot topics on the forum will find it hard to gain a clear perspective due to the noise generate from the volume of data going through the forum every day. Therefore, an overview of trending topics with natural language processing and statistical techniques is useful for identifying what may be of interest to security researchers. This chapter proposes a tool to identify tokens from trending topics, by pre-tokenising post data, followed by adapting a statistical technique for measuring changes, which can be used to scan across the dataset.

1The texts are fabricated so as to preserve user anonymity, but they are based on real ones encountered in the database.
The tool builds upon a weighted log-odds ratio [Monroe et al., 2008] with an informative Bayesian prior [Silge et al., 2020], used to compare differences in two corpora. In this chapter the corpora represents two distinct time periods of interest within the same subforum\(^2\). For known events, one period can be a set of texts preceding the event (the prior) and the other period can be texts following the event (the target). Also, the tool can be used for live listings of trending terms in the present day, by comparing new posts against some fixed prior.

This chapter presents a method which identifies the relative importance of tokens to each time period. The log-odds ratio indicates whether terms are more likely to appear in a given corpus over others. A log-odds score is higher for terms that are both unique and more frequent to a given period. Other NLP methods may require the removal of pre-defined stopwords. However, for the approach presented, as stopwords have a similar distribution across both time periods, they will have a low log-odds score and rank.

The tool looks at “bursty” events: for a token to be trending, the frequency of the token should be significantly different between the prior and target periods, and be more frequent than other terms in the target period. For identifying topics, the method used in this chapter uses a feature-pivot approach (a topic is a cluster of keywords) over a document-pivot approach (a topic is a cluster of documents). The latter may struggle with documents about multiple topics, whereas the former may incorrectly identify correlations between words as topics.

A major challenge in developing the tool is that it is to be used on a large dataset of noisy data, for exploring the evolution of underground hacking forums. Mitigating steps such as storing the pre-tokenised and part-of-speech tagged text are taken, to decrease computation time for longitudinal analysis. While there is focus on a cybercrime context, this type of data has similarities to Twitter data: short posts, and informal language. However, while Twitter data has some minimal inter-tweet connections through hashtags, quoting comments and replies, forums have a rigid discussion-based structure set by the forum administrators.

In this chapter, the following contributions are made:

• Adapting a technique used for capturing the linguistic changes between two corpora, to be used as a trending topics tool for temporal analysis of data.

• Showing the application of this trending topics tool in the context of cybercrime research.

\(^2\)A forum is the whole site, and a subforum or bulletin board is a page on the site, dedicated to a given general topic and created by the administrators. Subforums contain member-created threads consisting of an ordered set of posts typically focused on a single topic.
5.2 Related work

This chapter uses concepts including LDA, TF-IDF, and trending topic models, which are discussed in §3.3.3.

This chapter adapts the Monroe et al. [2008] paper discussed in §3.3.3.1. This adapts the method to time-based analysis, modelling terms as a function of time. It uses an informative Bayes prior, which was used in the R tidylo library by Silge et al. [2020]. While this method was initially used to compare two distinctly different political party news corpora, it is adapted to examine a longitudinal dataset to explore how a particular corpus has changed over time.

5.3 Method

5.3.1 Data

This chapter uses the CrimeBB dataset described in §3.1. The data is “noisy”, containing variations of spelling (e.g., “ransomeware” instead of “ransomware”), orthography (e.g., “NK” and “nk” for North Korea), and length of posts (ranging from short replies “pm me” to longer in-depth tutorials). In addition, due to the size of the dataset, the method used in this chapter requires a lightweight approach in order to measure the evolution of trends and topics over time.

5.3.2 Tokenisation and pre-processing

Forum post text is pre-processed to remove markup blocks and tokenised, described in §3.1. Note this step does not remove a pre-defined list of stop words, however the Bayesian approach will decrease the relevance of a large number of very frequent words which appear equivalently in the prior and target texts.

Following this, PoS-tagging using spaCy [Honnibal and Montani, 2017] is carried out to identify nouns and noun phrases in posts, which results are filtered by. Note that this step is not applied before calculating log-odds, as this would change the distribution of tokens used in a period, affecting the quality of results.

Both the token counts and set of nouns for each post in the forum are stored. These are stored separately for each subforum in HackForums. Note that terms which may vary in their orthographic form are not merged – for instance acronyms or abbreviations with their full forms, spelling errors, and casing differences. It remains a matter for future investigation whether acronyms and abbreviations should always be associated with fully spelled-out forms, or whether they should be kept distinct because they represent different uses of the term. Secondly, future work could utilise a spell-checker to cluster
misspelled words with their intended form, but this will need adaptation to the vocabulary
of the cybersecurity domain. Finally, this approach can capture casing differences (e.g.,
“WannaCry” and “wannacry”, and “NHS” and “nhs”) because all texts are lower-cased
before tokenisation.

5.3.3 Windowing: Prior and Target

The method requires the selection of two time windows: a prior and target period. The
prior period is used to learn a distribution of terms used, as a comparator for the target
period. The size and placement of windows can be varied depending on the desired results:
long-term trend detection would have a longer, and more distant, target window than for
short-term trend detection.

These windows should be selected depending on the dataset used and research questions.
If the prior window and target window overlap the same event, then these terms will
appear in both windows with a similar frequency, and will therefore have low log-odds. If
the prior and target window are too far apart, then the prior may not be representative,
leading to poor quality results. Also, if a topic is re-trending, and the previous trending
period falls in the prior, then this may affect whether a term appears to be trending.

5.3.4 Overview of the log-odds method

The approach in this chapter uses a method implemented in the tidylo R library [Silge
et al., 2020], which this work has re-implemented in Python for compatibility with other
tools. The tidylo R library uses an informative prior Bayesian approach, instead of the
Dirichlet uninformative prior used by Monroe et al. [2008]. A later version of the tidylo
library added support for the uninformative prior. However, this approach continues using
the Bayesian approach as this chapter’s time-based application of the tool is suited to
using an informative prior.

This approach, created for comparing two corpora, is adapted to detect trending tokens.
Instead of selecting corpora by pre-existing classes, prior and target time windows are
chosen, to find terms which are more likely to appear in the prior or target period. Each
period is represented as a “bag-of-words”, for all posts in the selected period.

This Bayesian approach is shown in the following series of equations, based upon the
tidylo implementation.

For the corpus (combined set of posts in both periods) \( y \), define \( y_w \) as the frequency of
token \( w \), and \( y_{wi} \) as the frequency of the token \( w \) in period \( i \). \( n \) is the sum of frequencies
of tokens across all periods, and \( n_i \) is the sum of frequencies of tokens in the period \( i \).

First, calculate \( \omega_{wi} \), the odds of each token appearing in period \( i \), and \( \omega_w \), the odds of
each token appearing in the corpus:
\begin{equation}
\omega_{wi} = \frac{y_{wi}}{n_i - y_{wi}} \tag{5.1}
\end{equation}

\begin{equation}
\omega_{w} = \frac{y_{w}}{n - y_{w}} \tag{5.2}
\end{equation}

Secondly, calculate \( \delta_{wi} \), the log odds ratio to compare the usage of the token \( w \) in period \( i \) to the whole corpus:

\begin{equation}
\delta_{wi} = \log \omega_{wi} - \log \omega_{w} \tag{5.3}
\end{equation}

Thirdly, calculate the variance of the estimate, \( \sigma^2_{wi} \):

\begin{equation}
\sigma^2_{wi} = \frac{1}{y_{wi}} + \frac{1}{y_{w}} \tag{5.4}
\end{equation}

Finally, calculate the log odds score \( \zeta_{wi} \) for each token \( w \) in period \( i \):

\begin{equation}
\zeta_{wi} = \frac{\delta_{wi}}{\sqrt{\sigma^2_{wi}}} \tag{5.5}
\end{equation}

Depending on when the prior and target time windows occur, the tool will either pick up short or long term trending tokens.

## 5.4 Evaluation

The results of the tool are evaluated by carrying out an information retrieval task with human annotators, where the success of the approach is determined by how well they surface relevant trending terms. The log-odds approach is compared with TF-IDF using discounted cumulative gain and the human annotations as a ground-truth ranking of identified terms. This uses both a known cybersecurity event to define the target window, as well as a randomly-selected target window to determine if relevant terms are surfaced without selecting a pre-determined event.

### 5.4.1 Trending Event Selection

Within CrimeBB, the HackForums dataset is used, as this is widely studied in prior cybercrime literature [Bhalerao et al., 2019; Pastrana et al., 2018a,b]. First, for the known event, the spread of WannaCry in the year 2017 is selected. WannaCry is a type of ransomware, which encrypts data until the victim pays a ransom. WannaCry spreads through vulnerable computer systems, instead of directly targeting specific entities, where these systems have not previously updated their systems to patch this issue. One of the largest organisations affected by this attack was the National Health Service (NHS),
the universal public healthcare system in the UK. This event is chosen as it is likely it would have been extensively covered on the forum. Indeed, it was later revealed that the individual who was instrumental in stopping the spread of WannaCry had formerly been an active forum member [Krebs, 2017].

The incident within the NHS began on Friday 12 May 2017 [Smart, 2018], which is selected as the start of the 7 day window for the analysis. The subforum “News and Happenings” is selected with a prior period of 2017-04-12 to 2017-04-18 and a target period of 2017-05-12 to 2017-05-18. The prior contains 404 posts, and the target contains 470 posts.

Secondly, the random subforum is selected, “Monetizing Techniques”, with a random date range for the target (2016-12-23 to 2016-12-29 for the target, and a week in the previous month for the prior: 2016-11-23 to 2016-11-29). The prior contains 195 posts and the target contains 295 posts. This random event and date range is used to avoid bias in the evaluation, where known events are more likely to contain unique keywords.

5.4.2 Log-odds and TF-IDF Results

The approach is compared to TF-IDF for topic ranking, using a similar approach to log-odds. This includes creating two TF-IDF “documents” as the set of posts for a given period (e.g. prior or target), as this is similar to the current method (frequent terms in the period but not frequent across all periods). The same tokenisation and pre-processing approach is used as the log-odds tool, to provide direct comparison. TF-IDF is used as it is a lightweight technique for topic ranking and detection.

For each event and technique, the top 10 tokens are plotted for the prior and target periods. For the “WannaCry” event, Figures 5.1 and 5.2 show the top tokens and scores for the prior and target periods. The results of the log-odds tool for the target period all contain tokens related to the WannaCry ransomware event. While TF-IDF also includes tokens related to the WannaCry ransomware event, it additionally contains terms related to different events (e.g., “notebook”, “pirates”, and “sharing”). Figures 5.3 and 5.4 show the top tokens and scores for the randomly selected event.

5.4.3 Annotation Task

First, a list of ranked terms is generated from both the tool and from TF-IDF, selecting the union of the top 10 terms of each event. For the WannaCry event, these were: amount, computer, cyber, data, hospital, hospitals, malware, microsoft, money, notebook, pirates, ransom, ransomware, security, sharing. For the randomly selected event, these were: affiliate, betting, burst, fiverr, guide, imdb, laptop, mining, movie, movies, network, time, traffic, videos, week.
For each event, three annotators who are subject experts are presented with each post from the prior and target periods with the accompanying tags. The annotators selected the most salient tags for each post, leaving posts not annotated if there were no suitable salient tags. Inter-annotator agreement is measured using multinomial Krippendorff’s alpha with the MASI distance metric of sets [Passonneau, 2006] for comparison, finding an overall agreement of 0.833 between the three annotators. Multinomial Krippendorff’s alpha is used due to multiple annotators each selecting multiple labels for each post. MASI distance is used as the labels for each post form a set, which this distance metric can measure differences between.
5.4.4 Discounted Cumulative Gain

Using annotations combined using majority voting, the ranking of the log-odds tool is compared against TF-IDF, using normalised discounted cumulative gain (nDCG) [Järvelin and Kekäläinen, 2002]. nDCG is a metric used to evaluate the usefulness of a ranking of a list, by measuring the quality (salience) of tokens returned from the tool. nDCG is measured with the annotations of salient tokens, as the metric increases the weight of errors towards the top of the ranked list, compared to other rank correlation measures, such as Kendall’s tau. Additionally, there is no ground truth information on the ordering of all tokens.
For the WannaCry event, the log-odds tool scored 0.979 compared to TF-IDF of 0.877 using the annotated dataset for validation. For the random event, the log-odds tool scored 0.978 compared to TF-IDF of 0.753 using the annotated dataset for validation. For both events, the log-odds tool had a greater discounted cumulative gain score than the TF-IDF approach, finding the ranking of terms provided by the log-odds tool produced more relevant salient terms than the TF-IDF method, for the forum dataset.

5.4.5 Measuring Timing of Latent Dirichlet Allocation

Computation time is measured and compared to training of a Latent Dirichlet Allocation (LDA) model. The WannaCry event subset of posts is used for this task, which contains 2,544 posts. Running the log-odds method on this dataset 100 times results in an average time of 0.0464 seconds, using an M2 chip laptop with 24GB RAM. An LDA model is trained across the dataset 10 times, with a parameter search space of 10 to 50 topics, taking on average 242.263 seconds. This highlights the lightweight approach of the log-odds method, which does not require model training time or parameter search.

5.5 Discussion

Detecting trending topics on noisy social media data is not a new problem for information retrieval and NLP. However, the application of an existing statistical method onto a longitudinal dataset provides a novel lightweight approach to detecting trending terms, which returns terms of more relevance than TF-IDF, and remains computationally less expensive than topic modelling such as LDA.

This work provided an initial step towards detecting temporal linguistic changes over time, by pre-processing text data, followed by using a Bayesian approach with a moving prior and target window depending on whether a user is observing short or long term trends. While the method does not identify the relevant windows itself, the tool can be combined with trending topic detection techniques to identify lexically distinct events, where some terms may re-trend.

Having shown that the statistical model is strong, and using a Bayesian approach can support new and evolving slang in the dataset without fine tuning a language model, there are still ways to further improve the NLP of cybersecurity forum texts. For instance, pre-processing can be improved in order to better deal with noisy texts: this includes the detection of misspellings, orthographic variation, acronyms and abbreviation, and deliberate obfuscation such as leetspeak. In addition, the detection of multiword expressions and named entity recognition techniques for noisy language can be incorporated, since both are likely to be of interest to researchers analysing language use in cybersecurity forums.
Future work should aim to increase understanding of the evolution of forums, changing language over time, and the changing topics of discussion by forum members. This includes automatically detecting and extracting events in the CrimeBB dataset. Although there has been an analysis of forum data, the tool can be used to explore trends in other corpora. Future work could also use this approach to analyse how spam emails have changed following the COVID-19 pandemic.

5.6 Conclusion

This chapter addresses the problem of detecting trends in cybercrime forum data, where text contains slang and jargon and has little structure. Existing techniques for detecting trends include the use of topic models, but these are computationally expensive to train, and require considerable manual effort to validate results and change hyperparameters (such as the number of topics). This chapter presented a new use-case for the log-odds tool introduced by Monroe et al. [2008] and previously implemented in the tidylo R library by Silge et al. [2020]. This chapter includes work on re-implementing the tool in Python for compatibility with existing Python analysis libraries, and the tool is used for detecting trending terms in longitudinal historical noisy text data of an underground hacking forum. The tool can be used for both detecting short term and long term trends depending on the time windowing and separation of windows selected. This approach produced more relevant salient terms over TF-IDF, using annotations.
Argot as a Trust Signal: Slang, Jargon & Reputation on a Large Cybercrime Forum

Signalling theory is applied to a cybercrime forum to explore how argot (slang and jargon) is used to signal trust in untrustworthy environments. This chapter contributes two findings: a new methodology for detecting argot, and findings from analysing relationships between detected argot and reputation systems. This chapter introduces an argot detection tool [Hughes and Hutchings, 2023a], using word embeddings from forum and non-forum datasets, which are aligned using training annotations. Compared with prior work, this approach improves performance, with an increase in the F1 and accuracy scores. Using the detected argot to create per-user variables, this finds a negative correlation between the use of argot and reputation votes. The trajectories of groups of forum members are explored to observe how the use of argot and user reputation in the forum varies over time. These findings indicate forum users are using argot to overcome the cold start problem, a conundrum faced by new users to social networks with ranking systems and marketplaces with feedback systems. A significant group of long-standing users is characterised by high levels of argot in their early forum postings. This decreases once reputation metrics increase. This particular trajectory group are amongst the highest-rated long-term members.

6.1 Introduction

Signalling theory [Gambetta, 2011] has been used to explain how criminals in the real world use subtle clues to signal trustworthy aspects to others in an environment of low trust. One example of signalling in the underground is the use of tattoos. These can be used to demonstrate toughness and resilience to pain. Some tattoos signal the gang membership
of the bearer and status within groups. Tattoos are difficult to remove, making them a permanent indicator of commitment [Gambetta, 2011]. Other groups use signals that are not so obvious, only caring to display group membership to others within that group, without bringing attention to themselves from outsiders. An example of a signal that relies on argot (slang and jargon) is Polari, a lexicon used by gay men during the time homosexuality was criminalised in the UK [Baker, 2003].

Signalling trustworthiness is particularly important for criminals, where they risk interacting with undercover police, and in environments where scruples are generally low. The idea of signalling theory is that signals of trustworthiness are cheap to emit—in this context, by authentic criminals (the way someone dresses, talks, or the tattoos they display on their bodies)—but expensive to mimic by those who are not genuine.

Signalling theory is particularly important when it comes to understanding how cybercrime forums operate [Bakken, 2021; Dupont et al., 2016; Holt et al., 2016; Lusthaus, 2012, 2018; Yip et al., 2013]. In these online spaces, you find people who need to interact with each other, to buy products or exchange services [Lusthaus, 2012]. Anonymity often means they do not know the identity of their contact, and there is no threat of violence that might otherwise deter others from cheating. Nonetheless, the cost of interacting with the wrong person can be painful, including losing their money or their liberty. However, physical signals are almost completely absent. The main medium by which users interact with each other is by text. Therefore, text-based signals can be important for communicating trustworthiness.

To avoid becoming ‘lemon markets’ resigned to the control of rippers who price out genuine sellers [Herley and Florêncio, 2010], forums introduce reputation systems as a signal of trustworthiness. However, sybil attacks could be used to disrupt (‘lemonise’) the market, with false accounts promoting distrust within reputation systems [Franklin et al., 2007]. Reputation systems provide a way for members to assign positive or negative votes to each other, and reputation metrics are displayed on members’ profiles to inform other members. Reputation systems are not failsafe, in that they may be gamed to falsely gain a higher reputation score (or to attack a competitor). Another problem with reputation votes is how reputation is initially gained by those who are new and untrusted, when they require others to trust them to gain reputation votes. In economics, this conundrum is known as the cold start problem.

In this research into the cybercrime underground, the use of slang and jargon by members can signal a level of knowledge and trustworthiness to other members, and results show the use of argot is negatively correlated with reputation metrics. However, when exploring the trajectories of actors who have been active on the forum for a year or more, there are more nuanced results. A significant group of users has a decreasing use of argot, while their reputation increases. These users are amongst the most active and highly
reputedly. Initially high levels of argot being used by this group are measured. However, this rapidly decreases, just months into their forum activity, coinciding with a growth in reputation. This could be due to the group of users using argot to overcome the cold start problem. Once they become highly reputed, reputation metrics take over as a trust signal, and their use of argot diminishes.

In this work, the use of the word ‘argot’ refers to the slang and jargon used by a particular group – in this case, cybercrime forum members. By ‘slang’ this means colloquial language either of new words or current words used in a different sense, and by ‘jargon’ this means words used by a specific group that are difficult for others to understand, such as technical terminology.

Let’s consider the term ‘rat’. The majority of people may immediately think of a rodent, typically bigger than a mouse, common in highly populated areas. However, within criminal communities, the term rat would be used in relation to someone untrustworthy who is likely to betray others, a police informant. Within cybercrime communities, however, there is an entirely different use of the term. Here, a rat is used to describe a commonly traded type of malware; a remote access trojan (or toolkit). This is why existing text mining and analysis tools trained on general text, i.e. not including argot from cybercrime communities, may not perform well within such a domain-specific environment, with cybercrime argot being unique to the specific underground community. Approaches that use contextual embeddings are useful in such cases where terms are used with different meanings, provided the embeddings are trained on the target text.

Another term, ‘leech’, is used on the forum. Typically, this can refer either to the type of parasitic worm or to a person that sponges off others. Within technology, the second sense of the word is commonly used by people who carry out torrenting, to describe users that receive torrents without contributing back to the peer-to-peer network, used to shame the users. This term is also in use on the forum in a similar way, to refer to members that use the forum to solely ask questions without contributing back to other members in any way.

Terms used on the forum may also be entirely new. For example, ‘eWhoring’ is discussed on the forum, and is used to describe a type of fraud, in which stolen or shared sexualised images are used to trick victims into believing they have paid for a virtual sexual encounter [Hutchings and Pastrana, 2019].

Despite argot being an important signal within criminal communities, this work is one of the first to test whether it is an indicator of trustworthiness within the cybercrime underground. One reason that argot is an under-explored trust signal is it cannot easily be measured. Given the specialised nature of argot, it differs across communities, and is often implicit. This chapter aims to detect argot and explore its use as a signal of trustworthiness, with the following research questions:
• How can natural language processing (NLP) techniques be used to efficiently detect argot usage on forums?
• Is there a relationship between the use of argot and reputation?
• How does the level of argot and reputation used by members vary over time?

6.2 Background & Related Work

This chapter covers topics previously discussed in the background sections on argot (2.2.2.1), argot detection in cybercrime communities (2.2.2.2), the cold start problem (2.2.2), and reputation systems (2.2.4).

6.3 Method for Argot Detection

The following section describes the dataset, the annotation process, the baseline comparison method, and the method used for argot detection. An overview of this is shown in Figure 6.1.

6.3.1 Data

This chapter uses data from the HackForums subset of the CrimeBB dataset, described in §3.1. A dataset of reputation votes from HackForums is used in this work. Reputation votes are either positive or negative values between -10 and +10 sent from a user and ‘received’ by another (however, the receiving user does not have a choice to reject these). Figure 6.2 shows the number of each quantity amount sent on the forum. This can include a value of zero, used by members to provide feedback on other members without affecting reputation scores. The value amount a user can send depends on their ranking on the forum [Dupont et al., 2016]. The reputation votes sent were manually analysed for this chapter. It was found this changed in 2017, as scores were reset by the forum administrators due to misuse. Misuse included members sending negative reputation even though they had not traded with them, and members sending positive reputation to their friends regardless of their actions. Following this reset, members used automated scripts to quickly send other members reputation votes to try to recreate pre-reset scores. Therefore, only reputation data prior to 2017 is used in this analysis. Following this, the forum introduced a new contract system [Vu et al., 2020] as a new mechanism for trust. Contracts enable members to have a list of transactions that have taken place, for other members to observe, however there is no guarantee that these transactions reflect real-world transactions. Later in this chapter, a second analysis is run against confirmed
Figure 6.1: Comparison of this approach to the baseline method. Data from the annotation stage is used in different parts of methods for the DarkJargon baseline and this approach: training sets 1 and 2, validation set 3 and testing set 4. More details of the split are in §6.3.4.
received contracts (contracts which both parties have marked as being fulfilled), between
the period 2018-06-11 and 2020-06-11. The contract system provides a transparent log of exchanges between users, and during this time, the reputation system was no longer in use.

While reputation scores can be problematic, they are the strongest signal as to trustworthiness available on cybercrime forums. This has been omitted for the period after 2017, where there were varying volumes in reputation, and spikes in activity, indicating reputation gaming. Before this, the volume of trust reputation voting was relatively stable. Trust is trust within members, and reputation is the main explicit method by which users vouch for others.

### 6.3.2 Tokenisation

The contents of each post are extracted, and blocks including *URLs* and *images* are removed. Then, NLTK’s [Bird et al., 2009] TweetTokeniser is used to tokenise the text, and join tokens (a meaningful group of characters, i.e. words) with whitespace. This is saved as a new column in a database table. This method enables the use of *ts_vector* in SQL statements using ‘simple’ (whitespace) mode to extract and summarise tokens. If Elasticsearch is instead used, a query can be run to obtain tokens from the text.
6.3.3 Annotations

A list of training words is constructed based upon usage in HackForums. First, words are selected that occur in at least 1,000 posts, to remove low-frequency words likely to be misspellings. Then, NLTK’s list of English stopwords (e.g., “the”, “and”, “in”) are used to remove the highest frequency words in order until only 10 stopwords remain in the set. This is used to filter the most frequent words in the dataset, to include both stopwords and other common words. 500 words are sampled for annotation by academics with knowledge of cybercrime communities, to indicate if a given word is argot. This is chosen to provide a suitable sample size for the task, which could be annotated by annotators in a reasonable timeframe. Note for other alignment papers there is often an existing resource such as a dictionary. But in this case there is no such dictionary, meaning that this work has an overhead for creating annotations, and hence the seed vocabulary used is smaller than related work.

Fleiss’ Kappa is used for an indication of inter-rater reliability, and the annotations scored 0.59, which has moderate agreement [Landis and Koch, 1977]. The task of labelling argot is inherently difficult, with differing annotations between annotators. Therefore, majority voting among the three annotators is used to create the annotated training set.

6.3.4 Annotation Split

Annotations are split into two training sets 1 2 with 55% of words, one validation set 3 with 22.5% of words, and one testing set 4 with 22.5% of words, to ensure there are enough samples for each part of the pipeline. A further split separates the training set into two training subsets: the first 1 contains 70% of the negative (non-argot) annotations for alignment of word vector spaces. The second training subset 2 contains the remaining 30% of negative annotations combined with the positive argot annotations, for training the support vector machine (SVM) classifier.

For the baseline approach, the two training subsets are combined 1 + 2 into a single training set, as there is only one training step.

6.3.5 Baseline Comparison

This approach is compared to DarkJargon [Seyler et al., 2021]. The authors use two methods to identify hypernyms (more general related words) of slang: KL-divergence and Cross-context Lexical Analysis (CCLA), finding that KL-divergence outperforms CCLA. This chapter uses the KL-divergence method to build a baseline to compare the approach to. KL-divergence is used to measure the divergence of word co-occurrence between the HackForums corpus and baseline Reddit corpus, to provide a proxy for measuring context words are used in.
First, co-occurrence matrices are built for HackForums and Reddit data from Pushshift [Baumgartner et al., 2020] (sampling data from January 2020, due to the large size of the Pushshift dataset). This uses a window size of 10 (21 items in a window), as this allows for the comparison of results to the DarkJargon approach. Note that both co-occurrence matrices are limited to just those that appear in a dictionary of HackForums words only, to focus only on HackForums data, and reduce the overall size of the matrices. Then Laplace smoothing is used, selecting an alpha of 0.1 to maximise the mean reciprocal rank (MRR) score, which is used to evaluate the order of results.

The authors’ primary goal is to identify relevant hypernyms (general related words), rather than argot. To extend the task to identify argot, an additional step is added to the method: a threshold is placed on the KL-divergence metric to select words used in different contexts between HackForums and Reddit.

To achieve this, the combined training set 1 + 2 and validation set 3 is used for tuning. The diagonal of the KL-divergence matrix from DarkJargon is taken, to get the KL-divergence metric between the same word across the two corpora only. The KL-divergence of HackForums words is sorted in descending order, to incrementally lower the threshold of KL-divergence for predicting argot. For each increment, the F1 score is calculated. This approach is selected as it is more likely that argot will be used in different contexts in HackForums compared to Reddit, due to the varying conversation topics and jargon across different platforms. Therefore, initial steps of decreasing threshold will increase F1 score, up to a given point, and then decrease as more common words are within the lowered threshold. The threshold which has the greatest F1 score is selected: KL-divergence $\geq 0.91517$.

Using the testing dataset 4 with the DarkJargon approach, the accuracy score is 0.670 and the F1 score is 0.654.

### 6.3.6 Log-Odds Approach

#### 6.3.6.1 FastText models

Two FastText [Bojanowski et al., 2016; Joulin et al., 2016] models are used for creating embeddings. The first is a model pre-trained on CommonCrawl data, for comparison. For the second model, this is trained over the dataset of tokenised posts. Parameters used are: vector size=300, window=5, min count=100, continuous bag of words (CBOW), epochs=5. These are selected as they have the same parameters for the pre-trained CommonCrawl FastText model [Mikolov et al., 2017].

A FastText model is used as this enables the training of a model on a large corpus with low computational resources. Alternatively, modern NLP approaches such as BERT [Devlin et al., 2019] could be used instead, provided later annotations of argot are for tokens in
context. In addition, BERT embeddings would need to be generated for every word in context for prediction, which would be computationally expensive (requiring GPU, which was not available for this project due to data restrictions and a lack of GPU resources) compared to this lightweight approach. To test the feasibility of training BERT on CPU, a sample is taken of HackForums posts to estimate the total time required. This uses a learning rate of 0.00005 over 10 epochs. For a sample of 100 posts this takes 302 seconds, and for a sample of 1,000 posts this takes 2972 seconds, both training for 0.05 seconds per post. For the entire HackForums dataset of 42 million posts, this would take 208 weeks to run. This used an M2 chip laptop with 24GB RAM.

6.3.6.2 Create embeddings

Using the FastText models trained on CommonCrawl and HackForums data, word embeddings are obtained for all tokens in the HackForums dataset which have over 100 occurrences across all posts (including multiple usage in a post). Tokens are thresholded as computing statistics for all tokens can be computationally expensive, and those with low frequencies could be spelling mistakes of common words.

6.3.6.3 Align embeddings

Word vectors are aligned using open-source code (licensed under BSD 3-Clause from Babylon Health) [Smith et al., 2017]. This code uses a Procrustes method for aligning two fastText models trained on different languages, using a known dictionary of words to transform the vector space of one model onto another, such that non-dictionary words can also be translated using the alignment. In this pipeline step, this code is adapted to “translate” between two corpora: Reddit and forum data. The “dictionary” used to align words consists of known non-argot words. Specifically, the training data of only negative (non-argot) tokens set aside for alignment of the word vector spaces.

6.3.6.4 Feature collection

Definitions are obtained for tokens from the Urban Dictionary API\(^1\), to create features for the number of definitions a word has, and the number of votes the top definition has.

The features used are: cosine similarity of aligned vectors, distance of aligned vectors, HF word vectors (aligned), CommonCrawl word vectors (aligned), number of votes for the top definition in Urban Dictionary (or zero if not present), and number of definitions in Urban Dictionary (or zero if not present).

\(^1\)https://api.urbandictionary.com/
6.3.6.5 Predicting argot

The cosine similarity of words is calculated across the word vector spaces. A Gradient Boosting classifier [Pedregosa et al., 2011] is trained using features of the second training set and validation set. Examples of predicted argot terms include: adfly, ewhore, installer, rootkit, ub3rs. Note the prediction step predicts some false positives, which includes: again, nobody, and yourself.

6.3.7 Comparison

On the testing set this method has an accuracy score of 0.723 and an F1 score of 0.703. This outperforms the DarkJargon baseline with an accuracy score of 0.670 and and F1 score of 0.654. Note that the detection of argot is a non-trivial task, requiring contextual information and domain knowledge to correctly identify words, including where a word may have multiple senses (usage).

6.4 Measuring Argot and Reputation

First, the hypothesis that there is a relationship between the use of argot and reputation votes is tested. Then, further exploratory analyses are conducted to understand this relationship in greater detail, exploring how it varies over time.

6.4.1 Relationship Between Argot and Reputation

An argot score is created for each user. This uses the approach outlined in §6.3 to create a list of argot. Using this list, a materialised PostgreSQL view is created in the database. A summarised count of argot words is used for each member.

When testing the hypothesis that there is a relationship between the use of argot and reputation votes, the overall number of words posted by a user is controlled for. This control variable is used as the more words a user posts, the greater the amount of argot they are likely to use, independent of their reputation.

Reputation votes are votes sent between a pair of users, with a given sender and recipient, and a positive or negative value. These are reputation votes received prior to 2017 (§6.3.1). The sum quantity of votes is used to create a variable of reputation score.

6.4.1.1 Analysis with Reputation

The following variables for analysis (per user) are used: number of argot words used, number of words posted by a user, number of posts, and reputation score.
The first analysis step tests if there is a correlation between average argot proportion per post used and reputation score. This uses cross-sectional data variables. This takes the sum of reputation votes received prior to 2017-01-01, the count of argot used prior to 2017-01-01, and the number of words and posts posted prior to 2017-01-01. The number of words are controlled for by dividing the count of argot words used by the number of words posted, then again by the number of posts, to create a variable for the average argot proportion across posts.

Kendall’s tau-b correlation is used to carry out the statistical test, as it is a non-parametric measure of rank correlation, and can support ties in the data which Spearman’s rank is unable to do.

Kendall’s tau-b correlation was computed to assess the relationship between the reputation score and the average argot proportion across posts, for members (n=29,141) that have a reputation score and average argot proportion across posts > 0, shown in Figure 6.3. There was a significant negative correlation between the two variables, $\tau_b = -0.628$, $p < 0.001$.

### 6.4.1.2 Analysis with Contracts

The second analysis step tests if there is a correlation between average argot proportion per post used and contracts received. Contracts data is used instead of reputation data for this period, as the reputation system is no longer used. This takes the sum of contracts...
received, the count of argot used, and the number of words and posts posted during the period 2018-06-11 to 2020-06-01. This period is selected as the start of the contracts system (reputation system has been reset, more information in §6.3.1).

Then, the relationship between the count of contracts received by members is assessed, and their average argot proportion across posts during this period, shown in Figure 6.4. Only members that have both a positive number of contracts and average argot proportion across posts were included (n=6,159). There was a significant weak negative correlation between the two variables, \( \tau_b = -0.392, p < 0.001 \).

### 6.4.2 Argot and Reputation Over Time

This section explores the relationship over time between argot and reputation. This uses clustering for analysis. Initially, this work used Group-Based Trajectory Modelling (GBTM) [Nagin and Odgers, 2010], however the proc traj library does not handle large datasets and this could not get the package’s statistical models to fit to the dataset. Then, the use of k-means longitudinal (R library kml) [Genolini and Falissard, 2010] was explored. This uses the Lloyd’s algorithm in k-means, which is run multiple times to avoid finding a local optimum, and additionally can fill in missing data values. This work also explored the use of scikit-learn’s [Pedregosa et al., 2011] k-means implementation in Python, as the dataset does not have missing values. However, k-means does not work well with outliers, and expects variables to be normalised, which would become an issue when working with

![Figure 6.4: Average Argot Proportion Across Posts and Contracts Received](image)
multi-trajectory clustering.

Therefore, this approach uses Gaussian Mixture Models [Pedregosa et al., 2011] to identify trajectory groups in the variables. Gaussian Mixture Models can be a more general version of k-means, and are equivalent to using GBTM with a normal distribution. Note that the second stage of GBTM – fitting a polynomial curve to trajectories – is not included here, as (1) polynomials did not fit the trajectory means well, and (2) the dataset uses data for consecutive months, and therefore does not need to estimate values for missing months.

There is a large variation of number of months active for each member, during their first two years. Given the variation across this data, with a large number of members active for only one month and a small number of members active for several months, four are selected groups: members with 1 (inclusive) to 6 (exclusive) months of activity (n=469,714), members with 6 (inclusive) to 12 (exclusive) months (n=37,003), members with 12 (inclusive) to 24 (exclusive) (n=24,541), and members active for every month of their first two years (24 months activity) (n=2,092).

Within these groups, two additional steps are run to normalise the data. First, due to some members not being active for all of the months, linear interpolation is used with exponential weighted mean to fill missing data points. Second, there is a difference of volume of activity within these groups, leading clustering algorithms to cluster on volume. However, the intention is to explore how members have changed over time within their own activity, not how this contrasts to other members. Therefore, the month with the maximum Argot Per Post for each member is identified, and then divide each month by this. This gives features that are proportions for each member, rather than raw data.

These proportions are clustered within each activity level group, shown in Figure 6.5, using the Gaussian Mixture Model approach. Note that in each group, the number of months active is not always consecutive. For example, a member active in months 0 and 12 will belong in the 1 to 6 months group, and missing months 1-11 and 13 onwards will use linear interpolation to fill the inactive months. For members active between 1 and 6 months, the largest cluster contains a high continuous level of argot used by members. The second largest cluster has the greatest argot usage for the first six months, before reaching a steady level. The fourth largest cluster contains members who have posted no argot.

For members active between 6 and 12 months, the four clusters have smaller differences between them over time. The largest cluster gradually increases and decreases in use of argot during the first year, before continuing at a steady level.

Members active between 12 and 24 months have greater diversity in trends over time. The largest cluster also shows a similar pattern of increasing and decreasing during the first year, before continuing at a steady level. The second largest cluster continually increases
over time for the two years. The smallest cluster, containing 2,780 members, shows an interesting pattern where argot per post starts high, then continuously decreases over the two years.

The final group contains highly committed members: those who are active for every month of the two year period, which does not require interpolation to fill missing values. The largest cluster contains a consistent level of argot per post used by members over time. However, the middle cluster contains members with a decreasing level of argot over time.

6.4.3 Cold Start Problem

Next, the cold start problem is explored, in which new members try to overcome the issue of having zero reputation among existing members with established reputation. This can involve members engaging positively with other members, to gain reputation votes, or manipulation to quickly build reputation regardless of their activity. This section analyses the cold start problem for two groups (12 to 24 months and 24 months), as these both contain two declining clusters of Argot Per Post over time, to explore which group variables correlate with this pattern. The two groups with less than 12 months of posting do not contain a declining cluster, therefore these groups are not analysed.

Mean variables for members active between 12 and 24 months are shown in Figure 6.6. For the smallest cluster, which shows a continual decrease of Argot Per Post over time, the argot per post per month sharply decreases within the first five months, while the reputation score increases.

Mean variables for members active for 24 months is shown in Figure 6.7. There is a similar pattern to the previous group, in which the middle cluster contains members with a decreasing level of argot over time. A significant decrease is shown in argot per post at the same time as a significant increase in reputation score. In addition, this group is characterised by a greater number of posts per month and argot per month than other clusters.

Across these two declining groups, they both use a high level of argot in their forum postings, and this decreases once reputation metrics increase. These two groups have thus used argot as a way to overcome the cold start problem.

Note in these figures, there is no clustering or correlating between these variables, which could introduce multicollinearity issues. The purpose of these figures is to highlight the variables for each detected cluster, in order to explore why argot is decreasing in one group while rising in another.
Figure 6.5: Clusters of Argot Per Post of Members, Over Different Activity Levels
Figure 6.6: Mean of Variables for Members Active between 12–23 Months
Figure 6.7: Mean of Variables for Members Active for 24 Months

Mean variables for each cluster group in: Months Active == 24

Cumulative Reputation

Argot Per Post Per Month

Argot Per Month

Posts Per Month
6.5 Discussion & Future Work

First, this chapter presented a method for detecting argot, which outperformed the baseline approach. While this approach improved the accuracy and F1 scores on a strong baseline, future work in argot detection could focus on using more advanced NLP models to further improve metrics. For example, BERT could be used to identify argot. However, the annotations were per-word as this approach obtains a single word vector per-word, whereas BERT would need to train on argot words in context. This would require considerable more annotation time, and further, fine-tuning a BERT model to the task requires powerful GPUs, which were not available. Finally, in this approach, it was possible to identify argot words to provide counts of these words per-user for analysis. With BERT or similar pre-trained language models, argot usage would have to be predicted for each sample in the dataset, instead of a straightforward database query. This would add significant overhead to analysis.

The relationship between both argot and reputation scores was explored, and argot and received contracts, finding both have negative correlations. By analysing trajectories of groups of users over time, there were two groups of members who have decreasing argot usage while their reputation increases. This can indicate signalling between forum members, where argot is initially used as a signal of trust by displaying knowledge of words used within this community. Later, as reputation increases, this becomes the main trust signal, and usage of argot decreases, thereby overcoming the cold start problem. The cold start problem applies where members aim on increasing their reputation score, in order to improve trustworthiness for trading, before reducing argot usage once trust is established. There are also users who wish to continue using the community’s lexicon in order to fit in, requiring an ongoing level of cognitive activity.

Future work in analysing the relationship between argot and reputation could explore if this pattern exists across other cybercrime forums, and use advanced modelling techniques to cluster multivariate datasets (such as changing argot level and reputation over time together). Also, later work could look at whether reputation or argot comes first – do members increase their argot usage as reputation increases, or use forum-specific argot to build their profile? This could analyse groups where members have increasing reputation first before using argot.

6.5.1 Limitations

This work has a few limitations. Firstly, as the dataset uses scraped data based on real-world interactions, it will not be “perfect”. The reputation dataset was analysed before working on it, to identify major outliers in the dataset that could affect the result. This included the reset of the reputation system in 2017 (§6.3.1), leading to members
immediately sending a significantly greater level of reputation votes to each other, in order to try to recreate their previous score. Changes such as these can affect overall analysis, and therefore this chapter excludes this period, and include a second period using data from received contracts.

Also, note that this approach uses a bag-of-words approach. A list of argot is identified, to measure usage. However, these argot words do not contain contextual information. Future work should use more advanced NLP models which take context into account, to improve prediction accuracy and measurements.

6.6 Conclusion

In this chapter, a method for efficiently detecting argot usage on forums was presented, a cross-correlation analysis between argot and reputation was carried out, and the cold start problem was explored with the reputation system. The argot detection method combines pre-trained word embeddings with forum-specific embeddings, using an alignment approach with a set of annotations. Argot usage can be obtained from the dataset using simple database queries, requiring little computational time. This method was used on a subset of CrimeBB to explore the usage of argot over time. Kendall’s tau-b correlation both finding a significant result with between average argot proportion across posts and reputation votes, and between average argot proportion across posts and received contracts. Finally, clustering examined how argot and reputation usage varies over time among forum members, and how members overcome the cold start problem. This found that as time passes and reputation increases, argot usage decreases.
CHAPTER 7

CONCLUSION

This thesis has presented different perspectives of cybercrime forums, from the macro evolution of a forum, to viewing the meso trends on the forum, to the micro usage of argot by users. This has used theory from criminology to ground research questions into cybercrime forums, including digital drift and signalling. Analytical approaches have used various techniques across data science, machine learning, and natural language processing.

The thesis starts by measuring the evolution of a cybercrime forum, grounded by using digital drift theory from criminology. This includes categorisation of members to separate analysis of those who are active for a long period of time from members that join to post once. This work observed a significant amount of drift and churn, with a small proportion of users continuing to actively engage with over a year of activity, and less than 100,000 active members remain on the forum over time. The scale of churn on the forum gives rise to year cohorts, in which significantly more users join the forum in a year than continue on. Exploring the different interests of these year cohorts shows a shift in the early stages of the forum from beginner hacking to general boards, followed by later cohorts having a primary interest in money-making related boards (e.g. eWhoring). These interests evolve over time for users, with those active for more than a year typically shifting from primarily posting in hacking-related boards towards market-related boards. Overall, there is a shift towards financial related activity, but forum activity as a whole has reduced.

The meso level of trends explores what keywords are discussed on the forum, using a lightweight statistical approach. This adapts the log-odds tool introduced by Monroe et al. [2008] and implemented in the tidylo R library by Silge et al. [2020], providing a Python implementation that works on pre-tokenised text for efficient trend detection. Long and short term trends can be detected depending on the time window selected. Using annotations of salient terms during both discussion of WannaCry, and a randomly chosen duration, the approach produced more relevant salient terms over TF-IDF.

The micro level explored the use of argot by members, tested for cross-correlation
between argot usage and reputation scores, and explored the cold start problem with the changing reputation system to contracts. First, this introduced a method for detecting argot, which used combined pre-trained and forum-specific word embeddings with an annotated alignment set, outperforming the baseline on accuracy and F1 scores. Then, the relationship between argot and reputation scores, and argot and received contracts, was tested. Kendall’s tau-b correlation found both to have a negative correlation, with a significant result. Further analysis explored the trajectories of groups over time, finding two groups had decreasing argot usage while their reputation increased. This can indicate signalling of members, where as reputation increases, it can become the main trust signal, thereby overcoming the cold start problem, and use of argot then decreases.

In addition to the contributions to knowledge of cybercrime communities, this thesis develops multiple methods which contributes to measurement approaches of cybercrime communities. The applications of these tools are applicable in more general measurements of social networks, such as Twitter and Facebook, however cybercrime forums introduce the additional challenge of containing messy non-structured public-only data which may not be complete. Using machine learning and natural language processing tools on big datasets in criminological research can provide insights into communities, however most are designed to work with formal English text (such as newspaper articles) and additional work is needed to adapt these methods to noisy text environments (high levels of slang and jargon).

Macro analysis measurements of forums are used to summarise overall forum activity. However, forums contain a large number of users that post a few times, and a small number of users that post the majority of posts. Also, forums are constantly evolving and inherently non-static, and observing data as a static dataset can produce misleading results. This thesis contributes a categorisation of forum member activity levels to avoid bias in macro analysis of forums, and contributes a time-based perspect of forum measurements aims to provide new insights into forum data.

The evolutionary nature of forum data leads to rise and decline of discussion trends. However, given the size of the dataset used in this thesis, detection is non-trivial, and off-the-shelf techniques perform poorly with the high level of slang and jargon used in forum texts. A computationally lightweight approach is needed to support historical trend detection, which supports slang and jargon usage. This thesis contributes a new method for trending term detection, based upon a technique used in political science.

Argot, consisting of slang and jargon unique to a community, is frequently used in forum text data. Detecting the use of argot terms, which can be used to signal knowledge and trustworthiness, is non-trivial. This thesis contributes a computationally lightweight method used for argot term detection, using word embedding models from natural language processing, applied to forum data to identify argot terms.
7.1 Future work

This thesis explores different perspectives of cybercrime forums from the macro to the meso. Future work can address limitations of the dataset (e.g. analysing other forums of different sizes and topics), limitations of methods (e.g. using BERT for argot detection over word embeddings), and limitations of resources (e.g. time needed to annotate larger training sets). One dataset limitation is the high use of slang and jargon, with a lack of normalised text, and the need to carry out pre-processing of the dataset to improve the quality. Future work could include the detection of spelling mistakes, acronyms and abbreviations, and detection of deliberate obfuscation. Also, techniques such as named entity recognition for noisy language could be used to identify key terms and trends. Furthermore, there is a shift of forum members moving to other platforms, such as Telegram and Discord. This introduces further challenges for analysis, as this does not contain as much structure as forums. For example, chat topics may overlap in the same time period, and ‘re-threading’ these to separate out topics is non-trivial.

Trending topic detection used a Bayesian approach with a moving prior and target window, with different parameters selected for short or longer term trends. Future work could adapt this tool to search for window sizes, and identify where terms may re-trend in the same window.

The argot detection method introduced in this thesis improved the accuracy and F1 scores over a strong baseline. Future work could adapt this method to use large language models, such as BERT, to predict on terms in context. However, this is resource intensive: each term would need a collection of annotations in different contexts, BERT fine-tuning requires GPU access, and prediction would need to be run on every term. Future work in argot and reputation analysis could expand this to other cybercrime forums, and use clustering on multivariate datasets.

Future research questions to explore are:

- As forum members shift to other platforms, how can existing analysis methods be adapted to less structured information found in chat discussion platforms?

- How do results in this thesis, on large popular cybercrime discussion platforms, compare to platforms using different languages and of different sizes?

- How can large language models be used to improve argot classification, using contextual information?

- Can large language models be used to provide definitions of detected argot?

A key actors analysis study on web forums, using text analytics and forum involvement to analyse those who may be expert key hackers. Textual analysis observes terms used by members, to extract features which are combined with other features, including quotations and likelihood of conversation. Expert identification uses X-means, as the authors do not know how many clusters are in the dataset. Visualisations show degree centrality was an important characteristic of expert hackers. Also, other users were less likely to be involved in the core of the interaction network. Their case study also showed founding members to play an important role, in both bridging different types of members, and also having a high amount of interactions taking place between these founding members.


Cultural analysis of cybercrime economies, to identify what makes forums successful and sustainable. This includes showing network effects, cheap community moderation, and enforcing bans on members that break rules. Measurements observe the changing social network analysis metrics over time.

An approach for identifying multiple accounts of the same user using stylometry. Uses dumps as ground truth for location (in determining same person), and are also able to see private message, which are more free-form that general posts. This work uses the Stanford log-linear PoS tagger and principle component analysis (PCA), but they faced issues with PoS tagging. They find reasons for multiple accounts includes banning, sockpuppet (multiple accounts to raise demand and create competition), accounts for sale, branding (different types of products), cross-forum accounts, group accounts. They note there are some ways to evade this approach, including anonymity.


Apriori algorithm used for identifying frequent sets


Uses a convolutional neural network with both character-level and word-level features to carry out named entity recognition


Trending topic detection on Twitter, comparing six topic detection methods on three datasets related to major events, looking at events which differ in time scale and topic churn rates. Work looks at volume of activity over time, sampling procedures, and pre-processing, which all affect quality of topics. They propose using n-gram co-occurrence and DF-IDFt topic ranking. Preprocessing looks at tokenization (Twokenizer), stemming (Porter stemming algo), and aggregation (putting documents together as longer documents may cause issues with results). Authors mention that not much work has taken place on exploring these.

U. Akyazi, M. van Eeten, and C. H. Gañán. Measuring cybercrime as a service (caas)

Longitudinal analysis of the cybercrime-as-a-service (CaaS) ecosystem, discussing the types of CaaS offered on cybercrime markets, and finds that there has been no significant change in offerings despite law enforcement interventions.


Using cybercrime market data to examine the cybercrime economy, finding exploits are priced similarly to bug-bounty programs, and the market is continuing to grow


Economic analysis of cybercrime, suggesting the cybercrime economy is more like tech entrepreneurship than traditional crime gangs, due to the reliance on infrastructure, economies of scale, and the lack of a traditional hierarchy.


Building a crime type classifier for an underground marketplace, to explore relationships between crime types and currency exchange. Includes additional qualitative analysis to provide insights to qualitative results.


Paper exploring the history of the lexicon (Polari) used by gay men


Criminological paper discussing the application of signalling theory to drug dealers


Paper introducing the Pushshift Reddit dataset, which contains a large corpus of Reddit data available to researchers


Using Twitter stream data to detect cyber-related threats, using a convolutional neural network. Also, the authors release an annotated dataset of 21,000 Tweets.


Key actor detection study to identify useful variables for this task, using regression to measure the relationship between posting activity and reputation scores. Finds those who contributed most have the greatest reputation score, however the time spent on the forum and quality of posts were not correlated with reputation.


Uses word embeddings to help understand language used on underground hacking forums by showing related terms. This uses recurrent neural network language models, trained on less than 10,000 posts across two forums.


Detecting supply chains using supervised NLP and graph-traversal. This identifies a product category for a given post, identifying replies, and building
an interaction graph for subsequent selling. This approach was only able to identify a handful of chains despite the large size of the forums.


This work uses product and reply classifiers to build an interaction graph of forum posts, to detect supply chains using a modified breadth-first search. They identified 352 supply chains, noting this a conservative estimate, and more work is needed to test the generalisability of the approach due to the time required for annotating the dataset.


Using Reddit subforum data to explore the use of terms that have strong connections to communities (high affinity), and look at how the semantics of these have changed


Natural Language Toolkit (NLTK) Python library for analysing text


Dynamic topic models, which are used to explore the changes of topics in a corpus over time


A common topic modelling approach, Latent Dirichlet Allocation (LDA) models each document as a distribution over a set of topics. Selecting one topic can identify relevant documents, and selecting a document can identify salient topics.

Louvain method for community detection


FastText paper on training word embeddings


Using Twitter data to explore emerging topics with TF-IDF, which assumes each Tweet focuses on a single topic. They provide a method for ranking events based on an importance score, and the method can be tuned by the user to identify relevant cybersecurity threat events.


Statement of Ethics from the British Society of Criminology. This is used to inform research ethics for this thesis, in which is it non-trivial to gain informed consent from all forum members.


Annotations of cybercrime forum posts for post type, author intent, and addressee. A combined approach of using rule-based and machine learning techniques are used for classification.

Paper suggests maintaining cybercrime infrastructure can be tedious work, which can lead to burnout, and provides three case studies with interviews and qualitative analysis of discussions.


Study looking at the evolution of forum jargon, with new community members introducing new variations of jargon, which took time for older members to use.


Collecting cyber threat intelligence (CTI) from hacking forums. Relevant threat information is extracted using SVMs and CNNs, which the authors compare and find they perform roughly the same. They suggest boundaries in feature space are similar, and CNN is likely to perform better in a different, more complex example.


Summary of a shared challenge at WNUT 2017, focusing on named entity recognition on noisy text


BERT paper, modern large language model used as the baseline for many text classification tasks, following fine tuning for specific domains.

Exploring the dispute system in a cybercrime forum, finding a level of disputes despite the forum using vetting, opened by buyers rather than vendors, and less than a quarter of disputes get resolved.


Analysis of a feedback system on a cybercrime community, which uses a weighted approach. Members of a higher status have greater impact on a user’s score: a new user posting positive feedback awards 1 point, whereas a moderator can award 5 or 10 points. They find that only a small fraction of forum members participate in the reputation system, and beginners are over two times more likely to report positive feedback of members compared to administrators. To combat ‘rippers’ on forums, they note a sanction system used on the forum. Administrators may completely remove all of a user’s positive reputation feedback on the forum, leaving only negative reputation feedback.


Criminology paper using social network analysis (SNA) on police operation in Canada against hackers running botnets. Uses the Keyplayer software, which is used by social scientists for SNA. Authors note prior studies have only focused on simple metrics including centrality and betweenness. They first measure centrality and power, then network fragmentation, to determine if removing a key actor has a big impact on network structure. Importantly, the authors note an issue with SNA, where missing data can produce errors which are misleading. For example, a missing piece of information connecting key actors may cause secondary actors to appear more interesting than they actually are.

Initial work looking at underground economies on IRC and forum marketplaces. Main contributions focus on data collection, introducing a system for automatically identifying and monitoring a large number of these platforms.


A mathematical paper proposing approaches to model both network changes and information diffusion (spread) at the same time, using point processes. They use Twitter data to test their approach.


An approach for summarising topics, by selecting a group of salient sentences to help users catch up quickly. More recent papers are likely to use large language models over this approach.


This work measures underground IRC channels (such as data theft and credit card fraud), to report on the types of products advertised on these channels, and suggests countermeasures including sybil and slander attacks.


Using a longitudinal approach to predict experts by their changing features over time. Focuses on a question and answer forum, with features including timing of answers.

A system for crawling dark web forums, building site maps and spidering to build a complete scrape of sites with little intervention


Book about signalling theory from criminology, which explains how criminals use subtle cues to signal trustworthy features where in a low trust context.


Analysis of three underground forums, to determine how these are organised, with comparisons between traditional crime organisation and cybercrime. The networkx library is used for centrality measures, including degree, in-degree, out-degree, closeness, betweenness, and eigenvector centrality. This finds a two-tier hierarchy in smaller forums, and more layers in larger. The types of forums analysed are different (including by language and topic), so findings may not be generalisable outside of the analysed communities.


A new implementation of k-means for longitudinal data, for varying starting parameters. This is compared to Proc Traj (group-based trajectory modelling, which is model-based). Overall work finds close results when trajectories are polynomial curves, and KmL shows better results for non-polynomial trajectories. The authors note the importance of this, as clustering is often used as an exploratory approach, yet parametric-based models have assumptions of what the data looks like (such as GBTM). Their work uses the Calinski
and Harabaz criterion, instead of the user specifying this. They note that as their work is not model-based, it can be more flexible but also hard to test goodness of fit. This work also only looks at longitudinal single variables. The work aims at the field of epidemiology, in order to improve existing clustering techniques.


Introduces the digital drift framework in criminology, which is an application of drift theory used to explain how underground platforms provide a mechanism for engaging and disengaging with hacking and crime.


Overview paper of political science techniques for analysis political texts


Uses recurrent neural networks to identify mobile malware attachments on posts in hacker forums, then uses social network analysis to identify key hackers.


An approach for aligning word embeddings across two corpora of different time periods, which can be used to measure how words have changed meaning over time by measuring shift.

Discussion of rippers on underground markets (dishonest traders), and how underground markets display characteristics of "lemon markets" from economics.


Analysis of the role of trust signals in cybercrime marketplaces for stolen data, using models to explore the relationship between marketplace signals and reputation received. One finding was negative feedback correlates with receiving more positive reputation votes, which they hypothesise can be due to either a seller having a large enough user base, or rippers using positive feedback to obscure negative feedback.


Criminological theory of digital drift, explaining how drift into and out of criminal pathways can often be accidental or unpredictable


spaCy Python Library


This work looks at techniques for a user where topic models can receive their feedback and improve on this, build new tree-based topic models, and iterate. This is aimed at social scientists who would like to iterate on these models, but have limited machine learning knowledge.


Explores the use of argot, defined as slang and jargon, on a cybercrime forum, and the correlation of this usage with reputation score data. This also presents a pipeline for detecting argot, using aligned word embeddings.
Guided by the criminological theory of digital drift, this measurement study explores the time-based aspect of a forum dataset, to explore how the evolution of the forum has changed it, by size and by discussion topics.

Presents a trending terms tool, using a lightweight Bayesian method identifies key terms based upon user-selected time windows.

Book chapter exploring common issues faced and opportunities with cybercrime forum data, combining insights from the three authors’ prior works and projects.

Survey paper covering over 100 prior works on cybercrime communities, describing the current state of the art, and discussing both inherent challenges in the field and challenges that can be overcome. This paper sets out good practices and proposes a list of considerations for future cybercrime community researchers.

Explores two trajectories (pathways) of cybercrime offenders into hacking and fraud, using criminological theory and interviews to discuss steps from initiation to desistance.

Criminological approach applying theories of neutralisation and rational choice to study operators of DDoS-as-a-service (bother) operators, including quotes from interviews of operators.


Background study on eWhoring, in which offenders fraudently simulate intimate connections with victims for financial gain. Explores market dynamics, including the saturation of free collections of images leading to paid collections becoming more in demand.


Looking at how big data approaches are used within corporations for detecting cybercrime (e.g. automated anti virus) and how approaches can be targeted adversarially.


An approach for named entity recognition in social media texts using a bidirectional LSTM for word and character embeddings, combined with an LDA topic model. This work uses noisy social media texts for testing.


Paper introducing term-frequency inverse-document-frequency (TF-IDF), a commonly used technique in NLP to identify common terms in documents that are not common across the whole corpus.

FastText paper on text classification


Evaluation metric for quantifying the usefulness of a ranking, weighted towards the top of the ranking. Can be used for identifying salient terms, to decrease the effect of terms lower on the ranking.


Work exploring churn on social networks, showing that there is not one single type of churn: typical, holiday, bursty, and inconsistent are all variations of churn.


Measurement study of the relation between a user’s value in a community and the probability of the user churning.


Tool for looking at a comparison of terms used across two corpora. Also uses a technique to spread these into strands to improve labelling in a noisy plot.


Topic modelling task for carding forum members, to profile cybercriminals by topics discussed, including identity fraud, crimeware, free content, and private message requests.

Introducing a burst model to identify trends in streams of data (e.g. email and news articles). Uses an infinite-state automaton model for this approach, to build up intervals of time where terms appear in a burst.


This work discusses two-point trends on information streams, which have “rising” and “falling” words.


Combines a burst model with a topic model, for time series modelling of news and Twitter data discussing topics in parallel.


Sponsored editorial on the legal risks of DDoS attacks and using booters


Article revealing that the individual who helped to stop the spread of WannaCry ransomware had formerly been an active hacking forum member.


Suggestions for designing research studies on cybercrime forums

Research exploring the cold start problem in recommendation systems, and how this can be overcome with limited data by identifying similar established users.


A categorisation to help interpret annotator agreement scores


Hypernym detection paper (is-a relationship between words), using a different hyperbolic feature space


Uses the GroupLens dataset to analyse how users overcome the cold start problem a recommendation system.


Classification of user forum types on a Linux question and answer forum, for clarity, proficiency, positivity, and effort shown in posts created by users. Compares different feature sets from bag of words to those used in prior literature.


Lusthaus explores trust in cybercrime, finding sanctions are not as useful online, since there is no longer a large cost to switching profile: if a member
has negative reputation, they can lose this negative signal by creating a new forum profile. Lusthaus compares this to conventional crime, where individuals have a known identity and need to increase anonymity, whereas in cybercrime, individuals start with no identity.


Criminological book on how cybercrime has become industrialised, and how an economy has developed despite having a low trust environment


The authors create an automated analysis tool, using a parts of speech tagger (PoS tagger) and a sentiment analysis tool, to identify cyber threats on a forum.


Aims to correlate hacker social media use with reports of web defacement, finding an increase in usage of social media correlates with web defacements on weekdays


Addresses the problem where two word embedding spaces are not directly comparable. Compares Euclidean and graph-based alignment methods for transforming the word embedding spaces, and they find performance varies on context.

E. Marin, J. Shakarian, and P. Shakarian. Mining key-hackers on darkweb forums. In
Proposing the use of reputation scores to validate results from key actor
detection, and combine content, social network, and seniority features for
classification

C. Martin, D. Corney, and A. Goker. Mining Newsworthy Topics from Social Media. In
Advances in Social Media Analysis, volume 602 of Studies in Computational Intelligence,
978-3-319-18458-6. URL http://link.springer.com/10.1007/978-3-319-18458-6_2.

An approach for detecting bursts of phrases for a real-time topic detection
system. Uses a variant of TF-IDF to group bursty phrases appearing in the
same messages to identify emerging topics. The authors explored the effects
of windowing, where Super Tuesday performed better with fewer prior Tweets
as this was a longer event, than others which performed better with a longer
window. They used the Apriori algorithm for ranking, which performed well
for their task.

Criminological drift theory where offenders drift in and out of crime, supported
by lower social control

D. W. Maurer. The argot of the underworld narcotic addict. American Speech, 11(2):
116–127, 1936. Publisher: JSTOR.
Exploring the argot of narcotic addicts

Exploring the argot of prostitutes

Publisher: JSTOR.
Exploring the argot of con artists

D. W. Maurer. The argot of forgery. American Speech, 16(4):243–250, 1941. Publisher:
JSTOR.
Exploring the argot of forgers

Exploring the argot of grifters


Exploring the argot of moonshiners


Exploring the argot of professional gamblers


Exploring the argot of pickpockets


An analysis of two carding forums, using keyword analysis to qualitatively explore discussions


Word2Vec paper, used to build word embeddings on textual datasets


FastText paper which pre-trained models on Wikipedia, News, and Common Crawl corpora

Work by political scientists to identify words used by political parties, similar to work on trending topic analysis, although the data used is labelled. They find that classification methods are not suitable for this task, as they can overfit, and find that many approaches to feature selection problems do not use probabilistic models. However, they find this type of model does not make the same assumptions as classification techniques, and can be used for analysis techniques in addition to classification. They experiment with other techniques, including odds ratios and TF-IDF and WordScores, but find their Bayesian approach to be superior for their task.


Criminological study of a drug distribution network, using social network analysis techniques to identify vulnerable parts of the network.


Study looking at ground truth with six SQL dumps of forums. This includes looking at the increase in social degree over time, and the authors measure movement across forums by matching email address. They measure the level of reputation needed before activity begins, and provide a general overview of how these forums operate socially.


Updated work on Group-Based Trajectory Modelling (GBTM), used for modelling groups over time. Can be equivalent to other clustering results, however each group is represented by a mathematical distribution rather than based on determining means for each time step.

Criminological paper exploring the factors that sustains criminal groups over time


The paper presents a method for predicting private interactions on forum data, where only public data exists. They use a classifier trained on data where private data exists, and find natural language based features contribute more to predictions than metadata features. However, the accuracy of the method is not high, and the automated labelling approach focuses on posts where the author did not create any other posts in the surrounding 12 hours.


Explores the set of experts on a question and answer forum using clustering to find different types: experts who are consistently active, experts who are only active initially, and experts who are only active later on.


Criminological article using a theory of behaviour systems to explore how social media fraud operates


Analysis users across forums, to measure how often users post and classify them into types, such as provider and consumer.

Evaluation distance metric for comparing annotator agreement across sets, used by multinomial Krippendorff’s alpha.


This work looks at identifying key actors on an underground forum, using three different techniques: k-means clustering, social network analysis, and logistic regression. They carry out this work to select potential key actors based upon a set of known key actors, from media sources and manual analysis, and validate the predicted key actors by comparing terms used. Clustering and logistic regression use features including those relating to forum activity, network centrality measures, and reputation measures. They also propose an SVM classifier for identifying posts using NLP techniques, and begin to look at the evolution of key actors, based upon their posting activity in various subforums. The authors note that their approach may be limited, as it focuses on one forum only, and does not include private interactions on the forum.


Introduces CrimeBB, a dataset of cybercrime forums, and discusses how scraping works, and the ethics of this. Explores two case studies, one on currency exchanges and one on cybercrime pathways.


Measurement study on eWhoring, in which offenders fraudently simulate intimate connections with victims for financial gain. Presents a pipeline for data collection while considering significant ethical issues, and explores the pathways of offenders.

Sponsored editorial on the legal risks of DDoS attacks and using booters


scikit-learn Python library for common machine learning approaches


Criminological longitudinal study which uses social network analysis to measure affiliations of hackers that deface websites


Uses social network interaction graphs to analyse six dark web forums, with centrality and structural measurements. Findings show a small group of users form central hubs, and qualitative analysis finds these members typically post general content across different topics.


Presents a new web-based user interface for accessing the CrimeBB dataset, to enable access to cybercrime forum data regardless of technical skills. Also discusses the motivation behind this project, and carries out a usability study.


Technical Report, with sections on training classifiers to predict interactions and classifying marketplace listings. While ground truth data is not always available, the author instead suggests models can be iterated on with domain experts.

Exploring concept drift in models for hacking forums, where changing lexicon can cause a decrease in accuracy over time. The authors propose relabelling to address this, however this can be expensive.


Introduces LIME, to explain the prediction of any classifier by building a model around the prediction. Attempts to solve the issue of whether a prediction or model can be trusted, by looking at decisions made locally. They use an examples of text classification and image classification to show how their algorithm works, and test by both statistics and human subjects. They note this is for models which are not easily interpretable (e.g. random forests), instead of those that are (e.g. Naive Bayes). SP-LIME is also introduced, which is a method for selecting representative predictions for explaining a model from a global view. This work is useful in providing explanations of models regardless of the type of model used, supporting comparative work.


Paper proposing that black box machine learning models should be replaced with interpretable models, and that explainable models often approximate explanations rather than directly explain the model.


Analyses how new words can be formed, finding two predictive features:
semantic sparsity, where words are surrounded by few words in the embedding space, and growth rate


An approach for analysing assets on a marketplace, using LDA topic models and classification of source code. Training data is created using snowballing, and the dataset includes source code files and attachments. Classification of code is used to determine the language used, and topic models are used to determine the purpose of the code.


Building a scraping tool to collect information on cyber threat intelligence (CTI), and to identify exploits shared and discussed


Collecting data from social media, blogs and dark web forums to identify threats, such as data breaches. Includes case studies of detected threats


Provides an overview of social science approaches for social media analysis, including creation of dictionaries. This focuses on mainstream social media platforms rather than underground platforms.

‘DarkJargon’ paper used as a baseline in this thesis. This work maps ‘dark’
jargon to ‘clean’ jargon to make sense of new slang terms used on under-
ground forums. They compare KL-divergence and cross-context lexical anal-
ysis (CCLA), finding that KL-divergence outperforms CCLA on a simulated
dataset.

D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling
on Computer supported cooperative work - CSCW ’11, page 355, Hangzhou, China,

Introduces two methods for temporal topics in a stream: peaky topics, and
persistent conversations. Peaky identifies terms particular to exact window
of time. Persistent identifies terms at the peak of normalised term frequency
(number of tweets containing the word rather than the number of times a
word is used), which assumes terms are not used before and more frequently
afterwards.


 tidylo R library, for calculating log odds ratios across two corpora

W. Smart. Lessons learned review of the WannaCry Ransomware Cyber
Attack. Technical report, Department of Health and Social Care,

NHS report on the WannaCry incident, including lessons learnt on prepared-
ness and response

vectors, orthogonal transformations and the inverted softmax. In 5th International
Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017,

Approach using the Procrustes method to align fastText models across different
languages, and provides open-source code under the BSD 3-Clause License.
Code is used in this thesis to align two fastText models used for argot detection.

This work uses instant messenger (Skype) contact details listed on marketplaces, combined with a protocol flaw of Skype to uncover a user’s IP address. They use this personal data to longitudinally measure network behaviour, storing country information and deleting IP addresses immediately after using the protocol flaw.


Combines Kleinberg’s burst model with a dynamic topic model, to identify bursty topics in a corpus.


Analysis of 80 dark web forums, using topic modelling. Highlights similarities across forums, and state sequence diagram can be used to find anomalies over time.


Sets out ethical questions over the use of datasets from illicit origin, and applies these to existing papers, finding inconsistencies in how ethics are approached.

This article in USENIX’s Login explores the roles of members in the cybercrime economy, including those involved in infrastructure, trading, data thieves, and advertisers.


Crime script analysis study on how individuals get involved in credit card fraud through carding forums


Exploring the effect of moving from a simple vote-based reputation system on a hacking forum, to a contract system with a public log of transactions. Multiple distinct eras are defined, and new users are found to overcome the cold start problem by trading in currency exchange to build trustworthiness.


Combines Twitter network and geolocation data (metadata) into word embeddings to improve an existing embedding model


Detecting jargon used on underground platforms in Chinese, using a filter for similar words and a search engine for finding new words.

Explores trust among cybercriminals on carding forums, finding a key challenge of members needing to determine if another forum member can be a trusted individual. They note carding forums provide some organisational structure, which supports the introduction of a market.


Looking at identifying terms in posts using a neural language model, for classifying if terms are from the dark web based upon where they are found.


Authors find a correlation between posts on a hacking forum and DDoS attacks, and hypothesise a direct link from an increase in posts leading to a decrease in attacks, however there is little evidence in the paper to suggest this correlation is causal.


Work analysing reputation data, which shows that reputation can contain intrinsic bias, such as for those users that extend their interests across many topics and actively post in these.