Doctoral theses at NTNU, 2022:65

Jan William Johnsen

Interdisciplinary approach to criminal network analysis

Opportunities and challenges

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Jan William Johnsen

Interdisciplinary approach to criminal network analysis

Opportunities and challenges

Thesis for the Degree of Philosophiae Doctor

Gjøvik, August 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

ISBN 978-82-326-5817-6 (printed ver.)
ISBN 978-82-326-6833-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2022:65

Printed by NTNU Grafisk senter

Preamble

This thesis is submitted in partial fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU). This work has been carried out at the Faculty of Information Technology and Electrical Engineering, in the Department of Information Security and Communication Technology at NTNU from 2016 until 2022. This research was carried out under supervision from Prof. Dr. Katrin Franke, Dr. Thomas Walmann, and Dr. Stefan Axelsson.

This work received funding from the Research Council of Norway programme IKTPLUSS, under the R&D project 'Ars Forensica - Computational Forensics for Large-scale Fraud Detection, Crime Investigation & Prevention', grant agreement 248094/O70.

Jan William Johnsen

Acknowledgements

I would like to gratefully acknowledge my supervisors Prof. Dr. Katrin Franke, Dr. Thomas Walmann and Dr. Stefan Axelsson for their fruitful discussions and valuable guidance during these years. Thank you for all your support and important advice regarding my work. I am particularly thankful to Katrin for all the motivation, foresight and practical counsel, which substantially contributed to my professional and personal development. I would also like to thank members of the evaluation committee: Dr. Cristina Alcaraz, Prof. Dr. David S. Doermann and Prof. Dr. Sokratis Katsikas who agreed to review my thesis and provide valuable comments.

I am thankful to the department of information security and communication technology and the digital forensics group at NTNU for aiding me in completing this research. A special thank you to the head of the department Nils Kalstad. Several administrative staff members also played an important role, and I would like to thank Hilde Bakke, Kathrine Huke Markengbakken, Maria Henningsson, Urszula Nowostawska, Katrine Moe and Rachael McCallum who supported me and gave important advice at different stages of my PhD research. Furthermore, I want to thank the senior faculty staff Carl Stuart Leichter, Andrii Shalaginov, Slobodan Petrovic, Geir Olav Dyrkolbotn and Mariusz Nowostawski for enhancing my academic experience with discussions. Finally, a thank you to IT staff Lars Erik Pedersen for administrating and maintaining the servers running my experiments.

I enjoyed working with Dagens Næringsliv (DN) on a real-world case: analysing data manipulation in the Tidal music streaming service. A special thanks to the journalists Markus Tobiassen and Kjetil Sæter and editor Gry Egenes at DN for this unique opportunity and great collaboration.

I am thankful to COINS Research School of Computer and Information Security (COINS) staff and in particular Hanno Langweg for organising seminars, and winter and summer schools during these years. It was a great time for my personal and academic growth.

I am grateful to my colleagues and friends for their ideas and time spent together: Kyle Andrew Porter, Parisa Rezaee Borj, Mazaher Kianpour, Radina R. Stoykova, Sergii Banin, Gaute Wangen, Christoffer Vargtass Hallstensen, Anastasiia Moldavska, Vasilis Gkioulos, Steven Hardey, Grethe Østby, Ambika Shrestha Chitrakar, Merve Baş Seyyar and Edgar Lopez Rojas.

I want to thank the people behind the Ars Forensica project and the many project partners, particularly my fellow PhD candidates: Jens-Petter Skjelvåg Sandvik, Gunnar Alendal, Stig Andersen, Nils Martin Mikael Karresand, Jul Fredrik Kaltenborn and Rune Nordvik.

The delivery of this thesis heralded the end of a monumental effort. I am grateful to my wife, Dan Zhang, for all her love and generous support, whose patience gave me the capacity to pursue my goals. Similarly, I want to thank my parents, sister and family for always giving the greatest support to my endeavours.

Abstract

The Crime as a Service (CaaS) model allows cybercriminals to specialise in certain illicit fields, instead of being jacks-of-all-trades with in-depth computer knowledge. This model facilitates serious cyber-enabled and -dependent crimes, e.g. by exchanging information on abusive tactics and engagement in selling illegal materials, products and services. A minority of proficient cybercriminals drives the CaaS model. This minority group develops advanced hacker tools and supports the underground forums' majority population without the same technical skills. Law enforcement tries to disrupt the CaaS model, but their focus has so far been on taking down famous underground forums. Their approach has been shown to have limited impact on CaaS activities in practice. Investigators must instead target specific actors to have an enduring crippling effect. Consequently, there is a strong need for objective and reliable identification of the most prominent cybercriminals.

Knowing which actor to put investigative efforts into means that investigators must scrutinise large quantities of unstructured data from underground forums. However, it is unfeasible to use contemporary investigative methods to examine unstructured data, and employing expert knowledge in manually analysing large amounts of data is absurd. Yet, a substantial improvement can be achieved by leveraging computational methods.

This thesis aims (i) to provide a scientific basis for identifying and profiling cybercriminals in investigations and (ii) to derive advanced computational methods for the machine processing of unstructured data from underground forums. Our empirical studies work towards inferring actors' proficiency by using an interdisciplinary approach. This approach combines methods from natural language processing and social network analysis. Our approach equips investigators with methods to profile several thousands of underground forum actors and differentiate between novices

and proficient actors. Thus, investigators can efficiently and effectively analyse criminal networks to identify actors to further focus investigative resources.

Our initial systematic studies on network centrality measures found them promising for ranking actors in a way that scientifically captures their relative importance. Still, there are two shortcomings in particular: (i) they appear to favour actors with higher communication frequency than important actors in the CaaS model and (ii) the results lack interpretability as to why the actors are given a distinct centrality score and ranked in a certain order. In fact, centrality measures disregard new and prominent actors with fewer posts, and they may erroneously single out actors as prominent cybercriminals in underground forums.

We provide new insights in order to highly automate the process of inspecting underground forum actors' posts using rigorous preprocessing steps and topic modelling. With our approach, investigators can complement the centrality measures' results and understand individual actors and their role in the CaaS model. As a result, investigators can ascertain the value of a potential high-impact cybercriminal actor quickly.

The main contributions of this thesis are (i) designing an interdisciplinary approach for improving our understanding of underground forum communication, (ii) identifying proficient cybercriminals on both the high-level and individual viewpoint in large-scale datasets, (iii) develop a theoretical foundation for rigorous preprocessing steps for more efficient and effective algorithms, and (iv) support decision making and otherwise help investigators scrutinise large amounts of unstructured data. Advanced computational methods give us insights into underground forums' inner workings, subsequently allowing us to exclude about 90% of novice users and focus our analysis on the more proficient cybercriminals.

Sammendrag

Modellen Datakriminalitet-Som-en-Tjeneste (DST) gjør at nettkriminelle kan spesialisere seg innen visse ulovlige områder, i stedet for å være altmuligmenn med utstrakt datakunnskap. Denne modellen muliggjør ondarta kyber-tilpasset og kyberavhengige forbrytelser, f.eks. ved å dele informasjon om skadelige fremgangsmåter samt salg av ulovlige materialer, produkter og tjenester. Et mindretall med dyktige nettkriminelle driver DST-modellen. Denne minoritetsgruppen utvikler avanserte hackerverktøy og støtter majoritetsbefolkningen i undergrunnsfora som ikke har de samme tekniske ferdighetene. Politiet prøver å stanse DST-modellen, men deres fokus har så langt vært å ta ned kjente undergrunnsfora. Deres tilnærming har vist seg å ha begrenset innvirkning på DST-aktivitetene i praksis. Etterforskere må i stedet fokusere på bestemte aktører for å få en varig lammende effekt. Derfor er det sterkt behov for objektiv og pålitelig identifisering av de mest fremtredende nettkriminelle.

Å vite hvilken aktør man skal etterforske betyr at etterforskere må granske store mengder ustrukturerte data fra undergrunnsfora. Det er imidlertid umulig for moderne etterforskningsmetoder å undersøke ustrukturerte data og svært ressurskrevende å bruke ekspertkunnskap i manuell analyse av store datamengder. Likevel kan en betydelig forbedring oppnås ved å utnytte datamaskinbaserte beregningsmetoder.

Denne oppgaven tar sikte på (i) å gi et vitenskapelig grunnlag for å identifisere og profilere nettkriminelle i etterforskninger og (ii) å utrede avanserte beregningsmetoder for maskinbehandling av ustrukturerte data fra undergrunnsfora. Våre dyptgående studier jobber mot å utrede aktørenes ferdigheter ved å bruke en tverrfaglig tilnærming. Denne tilnærmingen kombinerer metoder fra naturlig språkbehandling samt analyse av sosiale nettverk. Vår tilnærming gir etterforskere metoder for å

profilere flere tusen undergrunnsforumaktører og skille mellom nybegynnere og dyktige aktører. Dermed kan etterforskere effektivt analysere kriminelle nettverk for å identifisere aktører og konsentrere etterforskningsressurser på dem.

Våre innledende systematiske studier av nettverks sentralitetsberegninger fant det lovende for å rangere aktører på en måte som vitenskapelig fanger den relative betydningen av aktørene. Likevel hadde de spesielt to mangler: (i) de ser ut til å favorisere aktører med høyere kommunikasjonsfrekvens enn viktige aktører i DST–modellen, og (ii) resultatet mangler tolkbarhet for hvorfor aktørene får en eksplisitt sentralitetspoeng og så blir rangert i en viss rekkefølge. Sentralitetsberegninger ser faktisk bort fra nye og fremtredende aktører med færre undergrunnsforuminnlegg, og man kan feilaktig anklage aktører som fremtredende nettkriminelle i undergrunnsfora.

Vi gir ny innsikt for å automatisere prosessen med å gjennomgå undergrunnsforumaktørers innlegg ved hjelp av strikte dataforhåndsbehandlingstrinn og emnemodellering. Med vår tilnærming kan etterforskere komplementere resultatet fra nettverkssentralitetsberegninger og forstå individuelle aktører og deres rolle i DST-modellen. Resultatet er at etterforskere raskt kan fastslå verdien av enhver potensiell høyt fremtredende nettkriminell.

Hovedbidragene til denne oppgaven er (i) å utforme en tverrfaglig tilnærming for å forbedre vår forståelse av undergrunnsforumkommunikasjon, (ii) identifisere dyktige nettkriminelle både på overordnet og individuelt nivå i store datasett, (iii) utvikle et teoretisk grunnlag for strikte dataforbehandlingstrinn for mer effektive algoritmer, og (iv) støtte beslutningstaking og ellers hjelpe etterforskere med å granske store mengder ustrukturerte data. Avanserte datamaskinbaserte beregningsmetoder gir oss innsikt i undergrunnsforums indre; som lar oss utelukke om lag 90% nybegynnere og fokusere analysen vår på dyktigere nettkriminelle.

Contents

Tables

Figures

Abbreviations

BERTopic BERTopic.

AI Artificial Intelligence.

BERT Bidirectional Encoder Representations from Transformers.

BOW Bag of Words.

CaaS Crime as a Service.

CCI Cyber Crime Investigations.

CEO Chief Executive Officer.

CoC Chain of Custody.

COINS COINS Research School of Computer and Information Security.

COO Chief Operating Officer.

CSV Comma Separated Values.

CTI Cyber Threat Intelligence.

DB database.

DF Digital Forensics.

DFaaS Digital Forensics as a Service.

DFP Digital Forensics Process.

DFR Digital Forensic Readiness.

digraph directed graph.

DN Dagens Næringsliv.

DST Datakriminalitet-Som-en-Tjeneste.

GEXF Graph Exchange XML Format.

GSDMM Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture.

HTML HyperText Markup Language.

ICT Information and Communications Technology.

ID identification.

IOCTA Internet Organised Crime Threat Assessment.

LDA Latent Dirichlet Allocation.

LSA Latent Semantic Analysis.

ML Machine Learning.

NBL Newcomb-Benford's Law.

NER Named-Entity Recognition.

NLP Natural Language Processing.

NMF Non-Negative Matrix Factorization.

NTNU Norwegian University of Science and Technology.

OCG Organised Criminal Group.

OSINT Open-Source Intelligence.

PCA Principal Component Analysis.

PhD Philosophiae Doctor.

pLSA Probabilistic LSA.

PSCP PuTTy Secure Copy Protocol.

RAT Remote Access Trojan.

RQ Research Question.

SNA Social Network Analysis.

SQL Structured Query Language.

SQLi Structured Query Language injection.

SSH Secure Shell.

TF-IDF Term Frequency - Inverse Document Frequency.

UID Unique Identifier.

URL Uniform Resource Locator.

US United States.

VPN Virtual Private Network.

Økokrim Norwegian National Authority for Investigation and Prosecution of Economic and Environmental Crime.

Glossaries

Crime as a Service is when a professional criminal or group of criminals develop advanced tools and other services which are then for sale or rent to other less experienced criminals.

Big data is data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process.

Commonality is a sharing of features or characteristics in common.

Corpus is a large or complete collection of writings.

Cross-validation is a technique to assess how the results of a statistical analysis or machine learning model generalise to an independent data set. Typically used when there is not enough data available for partitioning them into separate training and test sets..

Cyber-dependent crimes where a digital system is the target as well as the means of attack.

Cyber-enabled crimes where existing crimes are perpetrated through the use of the Internet.

Data wrangling is the process of transforming and mapping data from one raw form into another format, which makes it more appropriate and valuable for ensuing tasks such as analysis and modelling.

Emoji is a pictogram or smiley used in electronic messages and web pages to fill in emotional cues in typed conversation.

Heuristic pertains to a trial-and-error method of problem solving used when an algorithmic approach is impractical.

Jargon is the language, especially the vocabulary, unique to a specific trade, profession or group.

Kingpin is the leader in a corporation, movement, undertaking, etc.

Pendency is the state or time of being pending, undecided or undetermined.

Web crawling is a computer program that automatically searches documents on the web.

Whistle-blower is a person who informs on another or makes public disclosure of corruption or wrongdoing.

Part I

Introductory chapter

Chapter 1

Introduction

1.1 Problem description and motivation

Cybercrime is a growing industry where the returns are great, and the risk is low [12, 119]. Europol's annual Internet Organised Crime Threat Assessment (IOCTA) reports on the dynamic and evolving cybercrime threats [53, 54, 55, 56]. IOCTA reports repeatedly highlight the problems caused by today's cybercriminals and describe how they gather in underground forums which act as marketplaces for illicit materials, products, and services [53, 144]. Underground forums facilitate wide-spread adaption of 'as-a-service' models, to provide a wide range of cybercriminals opportunities to enhance their cyber-enabled and cyber-dependent criminal activities [113, 153, 166]. The Crime as a Service (CaaS) business model allows any buyer to execute cyberattacks without understanding what is involved in its execution at considerably less cost than previously [153]. For example, CaaS grants easy access to advanced phishing and exploit kits, malware development and distribution, hacking expertise, fraud, compromising infrastructure, infecting systems, stealing (credit card) information, money laundering of traditional and electronic currencies, and child sexual exploitation material [87, 113, 166].

The actors in underground forums are roughly divided into two distinct groups [1, 53, 84]: a minority and majority group. Their distinction is based on individuals' technological skills, knowledge, and expertise. Only individuals in the minority group hold high enough technical skills to develop new malware strains or identify zero-day exploits and conduct their own cyberattacks. These individuals and groups also sell entry-level hacker tools and other services to the majority group – without the same technical skills – making their cyberattacks successful [53]. Researchers [39, 180] have estimated that cybercrime was costing the

world over 6 trillion USD annually by 2021, primarily backed by the adoption of CaaS.

Law enforcement naturally thinks about crime and criminal groups in terms of networks, modelled by mathematical graphs. The goal is to map the relations between criminal actors to develop effective ways of solving the crime at hand, preventing future crime and disrupting the criminal network. For example, Organised Criminal Groups (OCGs) in the past had bureaucratic and hierarchical structures [5, 36] with a 'supreme leader at the top of the hierarchy and different levels of subordinates, each with a different role in the hierarchy [102].' Forensic investigators can organise gang members in a hierarchical pyramid and focus their efforts on leaders and other key actors to cause a large and long-term disruption of the criminal network and their activities. A 'key' or 'important' actor is an individual with power in the network who can maintain essential processes or aspects of a criminal organisation such as money laundering. Traditional methods for identifying important criminal actors involved analysing actors' network positions because a structural network position often coincides with particular roles, functions or tasks [107]. The network position indicates, for example, to what extent an actor can exert influence, transfer, or selectively share information with other network actors.

Network centrality measures are the most common structural positions researchers and law enforcement focus on when analysing criminal network data. They use various centrality measures to focus on interesting actors with a high network centrality. 'Having a high network centrality indicates that an individual may be an important actor in the network [107].' In particular, degree and betweenness centrality are featured in related literature as able to identify prominent leaders (e.g. they are well connected) and other key actors in physical criminal networks of particular interest [117]. Thus, law enforcement can use centrality measures to prioritise limited resources on key criminal actors with higher centrality. See Subsection 1.7.3 for details about the various centrality measures used in this study and Section 1.8 for a description of how related work uses network centrality measures to identify important actors.

Forensics is the intersection between application, technology and methodology [65]. Using centrality measures is the same, but the application has changed from physical hierarchical networks to more decentralised and loosely structured networks. Network centrality measures have identified leaders of criminal organisations in previous work. However, they have not been verified to work similarly on the new application of finding key underground forum actors, such as professional cybercriminals within the CaaS business model. It is also important to call attention to the fact that early related work often analysed data from police investigations and is, therefore, analysing small networks. See Table 1.5 for a complete comparison

of network sizes. Today's underground forums' network sizes are vastly different, and this thesis analyses two real-world underground forum datasets with several hundred thousand users on each forum.

Forensic methodologies must be consistent, repeatable, and well-documented to comply with current methodological standards (see Subsection 1.7.1 for a description of the Daubert standard) in digital forensics. The field of digital forensics lacks systematic studies to assess under which conditions methodologies work or do not work. A significant drawback formerly consisted of the lack of scientific work that validated network centrality measures methods in forensics. Using unvalidated methods for a particular application and presenting their result in a court of law is a problem for the right to a fair trial, human rights, and procedural accuracy [171]. Moreover, relying on untrustworthy evidence is a factor in wrongful convictions [170]. For example, assuming key criminal actors also have a high network centrality can cause law enforcement to waste resources on non-important individuals in the best case. In contrast, in the worst case, they may inaccurately prosecute them for being among the leaders of a criminal network.

This thesis re-evaluates existing methods for the new application of identifying important/key cybercriminals in underground forums. It is essential to show the limitations of methodology to forensic investigators and that traditional methods do not work as intended or as described in related work. Important actors no longer hold a specific hierarchical role or position in the criminal network. Instead, important actors are more knowledgeable individuals who are technologically proficient and specialise in providing commodities such as malware and exploit development, reverse engineering, hacking and phishing. Rather than finding structural network positions, this thesis proposes a novel and comprehensive methodology for finding important actors by analysing the technical skills and expertise they exhibit when communicating with other cybercriminals.

1.2 Research scope and aim

Criminology is a branch of sociology and 'the scientific study of crime as a social phenomenon, of criminals, and penal treatment [124].' Criminologists study the personal characteristics motives of people who commit crimes, the effect of crime on individuals and communities/society, and sociological methods for preventing crime. On the other hand, forensic (described in Subsection 1.7.1) scientists 'examine and analyse evidence from crime scenes and elsewhere to develop findings that can assist in the investigation and prosecution of perpetrators of crime or absolve an innocent person from suspicion [179].' Criminologists and forensic experts study very different problems; that is, the first study criminal actors' behaviour and personality types, while the latter study analytical methods and ensure

accurate results. This research only addresses the problem description and the research questions from a forensics perspective.

We consider investigators from law enforcement agencies and investigative journalism to have the highest impact for revealing information relating to proficient CaaS cybercriminals. Investigators from each domain approach this problem under different and possibly complementary conditions. For example, law enforcement agencies typically investigate underground forums after a court order, except, perhaps, when conducting forensic intelligence. In different circumstances, investigative journalists do not require court orders to initiate their inquiries. Thus, a journalistic investigation can break a story and ignite subsequent law enforcement actions [164]. This thesis take on the perspectives from law enforcement and investigative journalism because of this duality when examining criminal networks.

This research's general objective is to: establish effective and efficient computational methods which can aid investigators to find actors or groups of actors belonging to the minority population. We define actors in the minority population as those individuals who have roles that demand some technical skills, such as reverse engineers, malware and exploit developers, providers of malicious services, distributors of stolen material, hackers, fraudsters, and so forth. The aim does not involve finding the role/profile of underground forum users, nor the time when users change roles. Finding or classifying users' roles in underground forums highly depends on the specific application. Thus, the thesis more broadly distinguishes proficient cybercriminals from other non-proficient individuals.

The datasets under investigation are two leaked real-world criminal underground forums. These leaked datasets are two raw database dumps, including data about user profiles, user ranks, private messages, public messages, and other data types. Social Network Analysis (SNA) requires users and a connection between them as variables to construct and analyse graphs, while Natural Language Processing (NLP) requires user-produced messages/posts to analyse their content. Our unique access to complete databases gives us some ground truth, which allows cross-examining the thesis's results against data found there.

OCGs and loosely associated cybercriminals are not the only ones using CaaS. For example, 'nation-state hackers and cybercriminals are increasingly impersonating each other to try and hide their tracks as part of advanced attack techniques [61, 138].' Renting or buying tools/exploits developed by others allows them to learn and imitate attack techniques. Not only do nation-state hackers take advantage of others' work to become more successful in their cyberattacks, but adopting other's tools will also confuse investigators. Investigators can be confused because they see the same attack signatures/patterns multiple cybercriminal groups

use. Therefore, it can be challenging to attribute a computer attack to a specific group of cybercriminals or nation-state hackers. This thesis will purely focus on cybercriminals, as we lack the necessary datasets and intelligence/classified data to concentrate on nation-state hackers.

A few researchers have published material with a similar theme as presented in this thesis, where they combine network and text analysis. However, they acquire different (types of) criminal datasets and use other analytical methods than those presented here. For example, researchers [2, 51, 111, 172] typically have corpora gathered through web crawlers or police reports. They use Named-Entity Recognition (NER) from NLP to extract individuals and connect them in a network, which can be analysed using SNA methods.

This thesis presents research with a novel interdisciplinary approach on two real-world criminal underground forums in Section 1.9. This is the first research with this unique interdisciplinary approach of combining network analysis and topic modelling. The thesis has two novel results. The first uses topic modelling to effectively divide an underground forum's users into minority and majority populations. In contrast, the second discusses how past researchers could have incorrectly used centrality measures to identify key criminal actors. The proposed topic modelling method can remove a significantly large proportion of an underground forum's users, making subsequent law enforcement investigatory actions more efficient because they can be aimed at actors within the minority population. Section 1.10 and Chapter 6 discuss the incorrect usage of centrality measures, which gives forensic scientists a more in-depth understanding of the limitations of these measures.

1.3 Research questions

Studying underground forums to identify professional cybercriminals means finding nuances in the way they communicate, their communication patterns, and the text they produce. Finding the correct information is challenging because it is buried in large amounts of data and not easily obtained using current analytical methods. Two research fields, in particular, stand out using methods with the potential to extract knowledge about individual underground forum users: SNA and NLP. SNA (described in Subsection 1.7.3) investigates social structures through mathematical graphs. In contrast, NLP (described in Subsection 1.7.2) can process and analyse large amounts of textual data.

This thesis combines NLP and SNA in an interdisciplinary approach to understand what people are talking about and with whom they are communicating, giving investigations relevant and reliable information admissible in a court of law. There-

fore, this thesis addresses the following general Research Question (RQ):

- What valuable information can be extracted from the relationship between underground forum communication patterns and post content to identify professional cybercriminals?

This thesis intends to address more specific RQs from forensic and journalist investigators' perspectives:

1. Through which analytical approach can investigators acquire leads on high-impact cybercriminals when the only available data are from underground forums?
2. How can NLP be applied to identify key actors who talk about relevant topics efficiently and effectively?
3. How can we model interactions between actors in the criminal network and identify professional cybercriminals in the model?

Effectiveness refers to the ability to produce the intended or expected result [140], while efficiency refers to acting in the best possible way with the least waste of time and effort [141]. Thus, effective is about doing the right things, while efficient is doing something right. The methods presented in this thesis use NLP to understand how key actors distinguish themselves from other types of cybercriminals.

1.4 Research methodology

1.4.1 Description of datasets and data processing cycle

The nature of the study determines the methods of data collection, the size of the population being studied and the application of appropriate data analytical methods [185]. The data gathering could have been achieved in a few different ways: (i) using a web crawler to collect data from the forum; (ii) ask users to participate in questionnaires; (iii) ask administrators to collaborate with us by delivering a dataset; or (iv) acquire the underground forum dataset by other means. The issues are that web crawlers are limited to accessing certain areas of the forums and easily countered using anti-web crawler techniques. Moreover, it is unlikely that users and administrators would help us in the process of developing methods for identifying proficient cybercriminals, who law enforcement can prioritise their resources. Therefore, collecting data by other means, such as downloading leaked

databases, is the only realistic way to give us an unbiased view of the research problem and accurate results.

Some online underground forums promote the sharing of leaked data, usually acquired through unethical activities such as compromising competing underground forums and stealing their database. The databases of two underground forums Nulled and Cracked [56] were leaked in 2016 and 2019, respectively. The work presented here involves the study of these two real-world underground forums. Nulled and Cracked are both hacker communities that facilitate the brokering of compromised passwords, provide tools and leaks, and generally act as marketplaces for services, products and materials.

The relational databases of Nulled and Cracked contain a lot of data about the forum and its users. Not all of the data and metadata in the databases are relevant for our experiments, nor is it relevant to describe the relationship between database tables (i.e. the database schema). Thus, it is sufficient to say that we are particularly interested in extracting information about users, their profiles and forum posts. Table 1.1 provides a short overview of these two datasets and shows the number of users and posts we collected information about.

Table 1.1: Statistics over datasets

Dataset	Users	Emails/Posts	First post	Last post
Enron	75 416	252 759	Oct 1998	Feb 2004
Nulled	599 085	3 495 596	Nov 2012	May 2016
Cracked	321 444	2 459 543	Mar 2018	Jul 2019

Readers may observe Enron in Table 1.1. We use the public Enron corpus as a benchmark to test our hypotheses against, compare the result between all three datasets, and allow other researchers to reproduce our experimental results using this public dataset. The Enron corpus is a dataset of workplace communication and is studied frequently in many scientific disciplines. Although Enron has been part of the datasets in our studies, this subsection only focuses on the underground forum datasets because Enron is well-known.

The initial assumption is that we downloaded original database dumps from the sharing individual. In other words, the leaked files have not been changed by any individual handling the files between the website compromise and sharing of the files. The database dumps should be identical to the ones of the underground forums at the time of extraction, including any database inconsistencies, discrepancies, missing values, etc. This initial assumption is difficult to verify without accessing and comparing the files with the original underground forum databases.

We inspected relevant database columns, checked our assumptions for the values these columns should contain and found no concerning inconsistencies. We checked for inconsistencies such as: (i) forum post timestamps outside the time range in Table 1.1, (ii) forum users having multiple identification (ID) numbers, (iii) the content/text produced by users. We addressed relevant inconsistencies, such as users having multiple IDs during the experiments. To the best of our knowledge, the datasets are valid and will produce a result that reflects the real world.

Data preprocessing involves extracting information from raw data to produce insightful results. Following a rigorous data processing cycle is crucial for more accurate and reliable results. Figure 1.1 show the data processing cycle we followed in our research. These steps were not necessarily linear as illustrated here, and we had to go back to re-do steps because the automated processing had made incorrect changes.

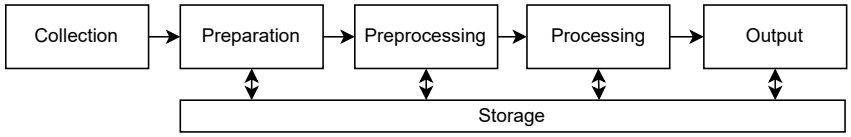


Figure 1.1: Data processing cycle

Collection Before any data can be processed, it must first be collected, which includes extracting or gathering raw data from available sources. It should be collected from accurate and reliable sources. Some underground criminal forums share leaked data with other users. We downloaded two leaked underground forums Structured Query Language (SQL) database dumps from such forums in 2016 and 2019.

Preparation The raw data needs to be prepared and cleaned to remove noise and format it in a way that makes sense for downstream analysis. The SQL is a relational database, which means it structures the raw data in relation to another piece of data in the database. Thus, this step involves creating a dataset using feature selection. This was achieved by manually inspecting database tables to select columns and merge them with columns from other tables. The columns of interest were data related to user and forum post, such as user ID, post ID, post content, etc. Table 1.2 show the values extracted from the underground forum databases.

user ID	username	role ID	role name
post ID	post timestamp	post thread ID	from user ID
to user ID	post content		

Table 1.2: Values extracted from databases

Preprocessing This step involves converting the selected raw data into a readable format, depending on the specific downstream analysis. For example, SNA need a matrix representation which illustrates the users and their relationship with each other, while NLP need a tabular representation of forum posts. Additionally, the forum posts must undergo some text normalisation steps. More specifically, NLP analysis requires text without special characters (newline, tabular, return), HyperText Markup Language (HTML), BBcode, forum-specific text and stopwords. Readers are referred to both Article III, Article IV and Article V for detailed descriptions of text preprocessing and the order in which they were performed.

Processing The datasets are analysed using SNA and NLP algorithms. This involved multiprocessing whenever possible due to a large amount of data.

Output The analysis result is presented in a readable format to the user, such as graphs and files. Interpreting these results are the basis for the conclusions in our work.

Storage This phase is involved in every step of the data processing process. After processing the data successfully, all information was stored for later use to promote a faster, easier means of accessing information. Storing intermediary data allowed it to be directly used as input in the next step of the data processing cycle.

In law enforcement, inaccurate data could mean the wrong person is accused and prosecuted for committing a crime. Maintaining data accuracy and integrity is at the forefront of every step throughout the data processing cycle. Data accuracy implies accurate data that can be used as an actual source of information to make important and informed decisions. We maintain data accuracy by removing as little as possible to avoid changing the original context of the forum posts. Furthermore, we rigorously check the result after each change in the preprocessing of the data and the order in which we make changes.

1.4.2 Description of general research methodology

Performing a literature review before formulating the research questions was necessary to avoid doing research that has already been addressed. Therefore, an extensive literature review played an important role in this research to both understand the research subjects and formulate the specific research questions. The literature review has also been done to present the state-of-the-art and existing solutions related to the research problem and questions. It was here – and in combination with the available dataset – that appropriate methods and theory were identified. Peer-reviewed research articles, books, technical reports, workshops, news articles and so forth have been used for the literature review.

This thesis aims to develop methods which can aid investigators to identify proficient individuals or groups of cybercriminals. Developed methods must follow strict legal standards of admissible evidence, such as the Dubert standard described in Subsection 1.7.1. Moreover, existing methods to identify important criminal actors (e.g. leaders) in related research must be re-evaluated according to this same standard. They must be re-evaluated because the application of these algorithms has changed with the new network structure of underground forums.

Using only quantitative or qualitative methodologies might obtain incomplete and incomprehensible answers to our research questions. This thesis is empirical research utilising mixed-methods research methods to draw from diverse research methodologies and data sources. We use qualitative and quantitative research methods to best address the research problem. The role of individual underground forum members can be inferred using the approach proposed in this thesis.

Qualitative research methods examine various forms of data from different angles to construct a rich, meaningful and complete understanding of the complex real-world phenomenon. Qualitative research methods involve a type of 'ethnography research' and aim to examine an entire group that shares a common culture, such as Internet-based communities [103]. Text analysis is qualitative technique used to identify the specific characteristics in the content produced by proficient and non-proficient cybercriminal actors.

Network centrality is a quantitative measure used to find important/influential actors in a network. However, it is debatable how valid the centrality measures are. The goal of this thesis is not just to perform a traditional SNA but rather to evaluate centrality measures by conducting empirical studies analysing exactly how network centralities correlate with the data. To this end, we use quantitative research methods involving 'correlation research' to statistically investigate the relationship between two or more variables [103]. Article V shows there is a positive correla-

tion between individuals with higher network centrality scores and the number of replies to their forum posts, and thus, demonstrate centrality is sub-optimal when used to identify important cybercriminal actors.

Identifying, developing and implementing appropriate methods and performing experiments with relevant datasets are crucial to provide proof-of-concept of the proposed solutions. The proposed solutions were applied on relevant data and the results were analysed to explain their benefits and limitations of using those methods. The interpretation of the experiment results help deriving the conclusions of this thesis. Figure 1.2 provides a sequential flow diagram showing important elements of the experiments conducted in this thesis. SNA and NLP analysis methods use different feature sets because they require distinct types of data preprocessing and algorithm input. Thus, Figure 1.2 show a branching to either graph or text analysis.

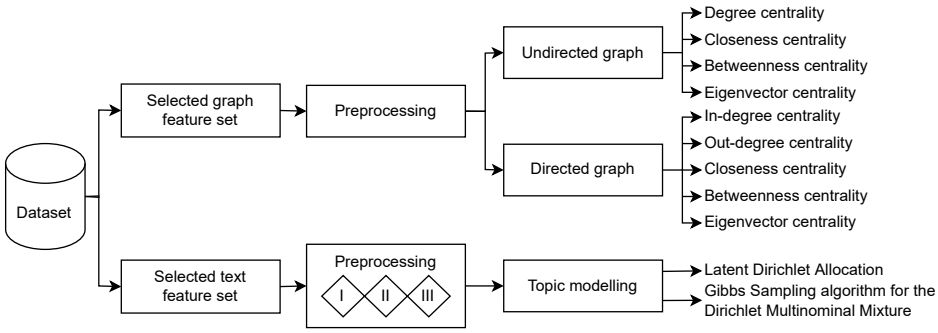


Figure 1.2: Experimental design

The selected features for the graph analysis are users (as vertices) and posts (as edges). The edges are either undirected or directed, and unweighted or weighted, as described in Subsection 1.7.3. Preprosessing graphs involve resolving discrepancies in the user ID or the possibility to have multiple e-mail aliases so that actors are only represented by one vertex in the graphs. Detailed graph construction and preprocessing are described in Article I, Article II, Article IV, and Article V.

The other feature accessible to law enforcement is the post content. The selected features for the text analysis are users and the content of posts. We studied three different ways of structuring the forum posts in Article III to achieve intelligible results from topic modelling algorithms. Detailed text preprocessing steps are described in Article III, Article IV, and Article V.

1.5 Thesis outline

This thesis consists of two parts. Part I contains Chapter 1 and gives an overview of this research project and it contains the following ten sections:

- Section 1.1 begins by describing the problem addressed in this thesis and motivates the approaches applied in this research.
- Section 1.2 and Section 1.3 explain the research scope and define the research questions addressing the problem stated in Section 1.1.
- Section 1.4 show statistics of the databases used in our study, and describes the data processing cycle and general research methodology.
- Section 1.5 describes the thesis outline.
- Section 1.6 provides a list of publications, their relationship to the RQs, and a brief summary of the thesis's contributions.
- Section 1.7 provides theoretical background knowledge from several areas. It describes forensic science from a civil and criminal perspective, essential machine learning theory and practices, social network analysis methods and thoughts surrounding the ethical and legal aspects.
- Section 1.8 synthesises related work which is relevant to this thesis.
- Section 1.9 gives a synopsis of the publications and their main results.
- Section 1.10 summarises the contributions of this thesis and highlights the impact of our real-world case in the context of investigative journalism.
- Section 1.11 provides a recapitulation of the thesis and the most significant findings, including theoretical implications, practical considerations and proposals for future work.

1.6 List of publications and summary of contributions

Part II constitute the main part of the thesis. It includes six research articles in Chapters 2 - 5. The candidate also has one related real-world casework publication of importance in Chapter 8. More specifically:

Research articles

- Chapter 2 contains Article I: **Jan William Johnsen** and Katrin Franke. *"Feasibility study of social network analysis on loosely structured communication networks."* In Procedia Computer Science, volume 108, 2017, pages 2388-2392.
  Summary: Our findings indicate that network centrality measures from SNA have limitations in that they rank actors with less important contributions to the hacker community [92].

- Chapter 3 contains Article II: **Jan William Johnsen** and Katrin Franke. *"Identifying central individuals in organised criminal groups and underground marketplaces."* In International Conference on Computational Science. Springer Cham, 2018. pp. 379-386.
  Summary: Our research demonstrates that centrality measures often rank actors with a natural higher frequency of communication rather than indications of criminal activities [94].

- Chapter 4 contains Article III: **Jan William Johnsen** and Katrin Franke. *"The impact of preprocessing in natural language for open source intelligence and criminal investigation."* In 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019. pp. 4248-4254.
  Summary: Extracting useful knowledge from unstructured data is a challenge. Our research established a way to generate more reliable results from automated processes applicable in forensic contexts [95].

- Chapter 5 contains Article IV: **Jan William Johnsen** and Katrin Franke. *"Identifying proficient cybercriminals through text and network analysis."* In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019. pp. 1-7.
  Summary: Our research establishes an interdisciplinary approach to analyse criminal networks by looking at how they communicate and the content they produce [96].

- Chapter 6 contains Article V: **Jan William Johnsen** and Katrin Franke. *"On the feasibility of social network analysis methods for investigating large-scale criminal networks."* Manuscript submitted for publication in April 2022.
  Summary: Our findings qualitatively demonstrate that network centrality measures identify those actors who receive more replies/attention and discusses how this can negatively affect a forensic investigation. Additionally, it establishes that our proposed method generalises well to other datasets of criminal underground forums [97].

- Chapter 7 contains Article VI: Andrii Shalaginov, **Jan William Johnsen** and Katrin Franke. *"Cyber crime investigations in the era of big data."* In 2017 IEEE Big Data 1st International Workshop on Big Data Analytic for Cyber Crime Investigation and Prevention. IEEE, 2017. pp. 3672-3676.
  Summary: We explain the challenges and implications of big data in digital forensic investigations. We strongly argue the need for new forensic tools and research into computational methods to support forensic investigations [162].

Additional publications

- Chapter 8 contains the casework: **Jan William Johnsen** and Katrin Franke. "Digital Forensics Report for Dagens Næringsliv." In Dagens Næringsliv. 2018. pp. 1-78.
  Summary: We were engaged by the newspaper Dagens Næringsliv to analyse millions of Tidal users' listening habits. We document our approach in a forensic report and indicate that 350 million playbacks were manipulated to artificially increase the listening for two distinct albums [93].

Publications' relation to research questions

Figure 1.3 shows the relationship between the publications and which research questions they attempt to answer. Article I, Article II, Article IV and Article V attempt to answer RQ 1 and RQ 3. While Article III, Article IV and Article V attempt to answer RQ 2. Article VI and Article VII do not directly contribute to answering the research questions in this thesis, but they have other significant real-world contributions for forensic investigators when they work with methods without peer-reviewed documentation. The contributions of these articles are described in Section 1.10.

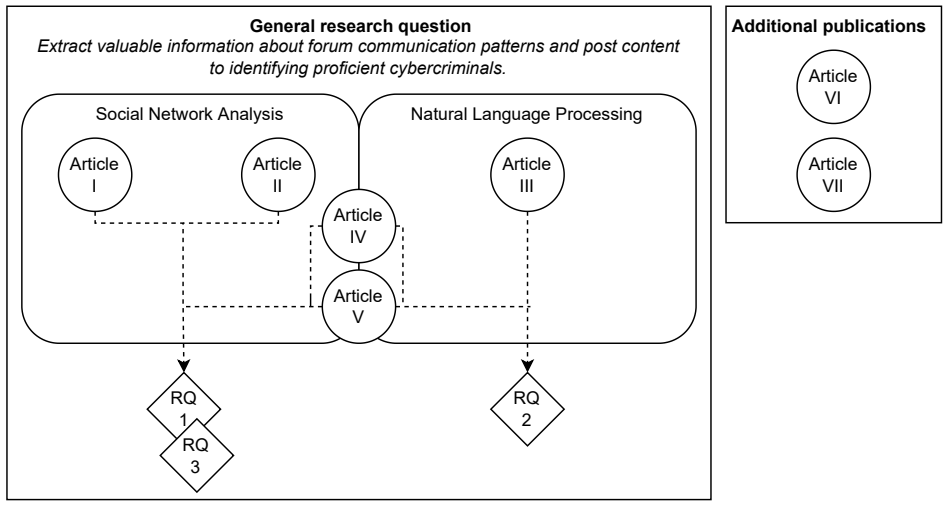


Figure 1.3: Relationships between publications and research questions

Figure 1.3 also illustrate the respective discipline where each publication belong. Article I and Article II only use methods from the discipline of SNA, while Article III only apply methods from the discipline of NLP. The interdisciplinary work resides in the two publications Article IV and Article V, which apply methods from both SNA and NLP.

Main contribution summary

- Establish procedures for validating existing graph-based methods (centrality measures) and applied them to re-evaluate their ability to find key cybercriminal actors in the CaaS business model.

- Demonstrated that current centrality measures are inaccurate when used to identify potential important CaaS cybercriminals. The measures identify cybercriminals who receive more attention and are talkative in underground forums, i.e. administrators, moderators and other talkative users. This has negative consequences for procedural accuracy because lesser criminals can be accused of being important cybercriminal actors, such as leaders.

- Showed that network centrality measures consider and quantify the structural position of vertices in a network. Thus, they can differentiate between central and peripheral actors. Their result is challenging to interpret, but it must be interpreted in the context that the underlying graph is modelling.

- Developed a novel method using topic modelling to accurately identify high

centrality actors.

- Developed a novel comprehensive methodology for identifying professional cybercriminals by combining existing technologies from text analysis. This method identifies the minority group of technically skilled cybercriminals, i.e. important cybercriminal actors.
- Developed a forensically sound and systematic, novel approach to preprocess data for forensic investigations.
- Validated the research presented in this thesis by analysing two large real-world criminal underground forums. The study of these datasets is unique when compared to related work. The two main benefits allow us to analyse complete data and validate results with the ground truth.

1.7 Theoretical background

This section lays out the theoretical foundations for the methods applied in later chapters. The forensic science section sets up (i) the important Daubert standard for evaluating expert witness testimony's admissibility and, thus, evidence reliability; (ii) describing big data's challenge from a digital forensic perspective and the possible computational method solutions to approach this issue; and (iii) discussing forensics from an investigative journalism perspective.

The Machine Learning (ML) section (i) discusses the importance of preprocessing data and evaluating models to test their generalisability; (ii) explains the topic modelling algorithms used in our studies; (iii) describes the ML theorems that motivates our approach; and (iv) discusses the Newcomb-Benford's Law (NBL) whose curve was the idea for detecting suspicious patterns in our real-world casework.

The section continues by explaining SNA and fundamental graph theory, which is necessary to understanding network centrality measures. Finally, the section discusses ethical and legal considerations for the secondary use of potentially illicitly obtained datasets.

1.7.1 Forensic science and forensic intelligence

Forensics applies science and the scientific method to provide facts for cases in the judicial system. It involves examining objects or substances related to a case by following the scientific method, legal standards of admissible evidence and other criminal procedures. Forensic analysts can use almost any scientific discipline to analyse evidence in legal cases, if only they apply the scientific method's principle

to arrive at objective facts. Extraordinary scientific innovations and advances have made forensics an essential part of the judicial system over the decades.

Forensic science is – for the most part – reactive and responds to crimes after they have been committed [6]. A first responder or forensic scientist collects, preserves and analyses evidence from a crime scene. Some forensic scientists collect the evidence themselves, while others may occupy laboratory roles or analyse evidence brought to them by other persons. Forensic scientists may work independently; however, all of them share the responsibility to maintain a Chain of Custody (CoC) to avoid allegations of tampering or misconduct. CoC is the chronological documentation (audit or 'paper trail') showing all the steps from the moment evidence (physical or electronic) was seized, and the custody, control, transfer, analysis and disposition of that evidence [208].

Most countries make a distinction between civil and criminal law. This section will, for simplicity, discuss differences from the perspective of the United States (US). Criminal and civil law differ concerning (i) how cases are initiated, (ii) how cases are decided, (iii) what kinds of punishment or penalty may be imposed, (iv) what standards of proof must be met, and (v) what legal protections are available to the defendant [46]. Both criminal and civil litigation leverage forensic science when arguing their case in a courtroom. The prosecutors of criminal cases must use evidence to prove guilt 'beyond a reasonable doubt'. At the same time, plaintiffs in civil cases only have to offer evidence demonstrating that their claims have a greater than 50% chance of being true, i.e. establishing liability on the preponderance of the evidence. Consequently, the evidential weight is higher in criminal courtrooms.

Traditional methods of forensic investigation view forensic science as a purely reactive resource within a legal framework [199]. However, information (e.g. from terrorist attacks) suggests that perpetrators were often involved in other criminal activities before progressing to serious crimes. Forensic intelligence is the proactive forensic science to handle future crimes before they occur or become a real threat [6, 199]. It involves gathering and analysing data earlier in the crime analysis cycle to detect, reduce, deter, disrupt, and prevent crime. The knowledge forensic intelligence can provide beforehand will help law enforcement deploy their resources appropriately [150], to save time, human resources and other resources [6].

The Daubert standard

The supreme court in the US has established standards to determine scientific evidence's admissibility to address the problem of 'junk science' in their courtrooms [64].

In the beginning, the Frye standard (general acceptance test) allowed expert opinions admitted only 'if the technique is generally accepted as reliable in the relevant scientific community [37]'. However, the Frye standard was strict and inconsistent with the applicable evidentiary rules of Rule 702 [31]. It also has a few issues: (i) difficult to admit novel approaches yet to gain acceptance within the relevant scientific communities, and (ii) admits evidence from accepted approaches in a niche scientific community, though these approaches are generally not accepted by larger scientific communities.

The Daubert standard has superseded the Frye standard and equips judges with an incomplete list of factors to consider when considering the admissibility of expert testimony [31, 64, 98, 208]:

1. Whether the expert's technique or theory can be tested and assessed for reliability.
2. Whether the technique or theory has been subject to peer review and publication.
3. The known or potential rate of error of the technique or theory.
4. The existence and maintenance of standards and controls.
5. Whether the technique or theory has been generally accepted in the scientific community.

These criteria are critical because not all scientific experts are reliable [63], e.g. through honest mistakes, incompetence, self-deception or outright dishonesty [173]. The Daubert standard first qualifies the expert witness to give an opinion on a specific subject. Then it evaluates the expert's methodology (i.e. can it be tested, has it been peer-reviewed, potential error rates, etc.). This evaluation determines whether or not an expert has used a scientifically valid methodology and that it can be adequately applied to the facts at issue [98]. Courtrooms cannot reliably eliminate junk science without the active support of and scrutiny from the scientific community. In this context, this thesis contributes to studies of network centrality measures and topic modelling in a forensic context.

Digital forensics

Digital forensics is a forensic branch encompassing the recovery and investigation of digital material [208]. Like forensic science, digital forensics is also commonly used in criminal and civil cases. There exist numerous models for the digital forensic process [145, 204]. However, we will describe the five-step

process involving: identification, collection, examination, analysis and presentation [145, 208]. Figure 1.4 illustrate this process where steps mainly happen in order; however, it is essential to note that this process involves repetition, both within each step and going back to re-do previous steps.

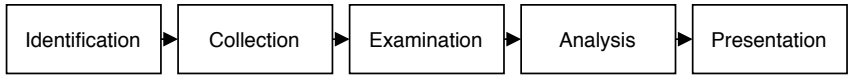


Figure 1.4: Digital forensics process model [208].

1. Identification This phase identifies that an incident/crime has occurred and finds relevant evidence. The initial steps taken in this phase are crucial because this will determine whether the efforts result in admissible evidence in a court of law. For example, law enforcement personnel must acquire relevant search warrants to seize digital devices and data in a criminal case. At the same time, companies can investigate their own equipment without a warrant as long as other laws (e.g. privacy and human rights) are preserved.

2. Collection This phase consists of preserving relevant data by protecting the crime/incident scene and acquire/extract data with forensically sound methods and techniques, e.g. using a write-blocker when acquiring a hard disk image. The CoC begins whenever an investigator takes custody of evidence at a crime/incident scene. The CoC is maintained during the next phases.

3. Examination Data must be examined and prepared for later analysis. This involves examining collected data to assess and extract relevant pieces of data that can be used as potential evidence. Additionally, raw data often requires restructuring, parsing and preprocessing while preserving its integrity.

4. Analysis This phase consists of processing data to address the investigation's objective to determine the facts about an incident, and support or refute hypotheses about the incident. The methods include detecting patterns and relationships between data objects, linking data objects and timeline analysis. Computational power can be used to automate the data mining task, known as computational forensics which are described in a following subsection.

5. Presentation This involves the presentation of findings from the analysis phase in the form of reports. The results of the investigation must be supported by actions taken and accounted for in the CoC. Finally, the results of the investigation are presented to a court of law or other audiences.

Digital data is volatile and can easily be tampered with or modified. The digital forensics community has, therefore, extensively used the term 'forensically sound' to qualify or justify the use of a particular technology or methodology [120]. The term can sometimes refer to a series of steps to follow or the adoption of several principles [198]. McKemmish [120] proposes four criteria to determine whether or not a digital forensic process may be considered forensically sound:

1. The meaning and interpretation of the digital evidence remain unaffected.
2. Errors has been identified and explained to ensure the reliability of the evidence.
3. The process can be independently examined and verified.
4. An analyst with sufficient and relevant experience has done the analysis.

Cyberattacks on digital devices generate a great deal of data; however, not all data qualify as forensic evidence admissible in a court of law. Forensically sound processes must gather digital evidence which complies with the legal criterion for it to be admissible [32]. The process model must adopt several principles (in harmony with, e.g., the Daubert standard) to ensure that the end-product does not lose its evidentiary weight [120] while conserving critical aspects of digital forensics: reliability, repeatability, and verifiable results [62]. Digital forensic experts must maintain a precise CoC to support litigation and present technical testimonies in terms which laypeople can understand [157].

Investigative journalism

Journalism comes in so many shapes and sizes that it is easy to arrive at distinctions that are not absolute. Therefore, this section will discuss the characteristics of journalism, albeit somewhat simplified. News journalism deals very rapidly with received information defined by tips from the public and authority entities, such as ministries, police, universities and spokesmen [29, 80]. In contrast, investigative journalism is characterised by long in-depth research that may take months or years to unveil matters concealed either deliberately or accidentally by someone in a powerful position [29, 86]. Investigative journalists 'investigate a single topic of interest, such as serious crimes, political corruption or corporate wrongdoing [190].'

'Investigative journalism and detective [(i.e. law enforcement)] work share many similarities, including the need to follow leads, dig up clues, and gather evidence [77].' One main distinction is that law enforcement can take legally permissible actions by arresting suspects and seizing materials; however, they must also

follow legal restrictions, such as acquiring court orders. Investigative journalists can, in comparison, operate relatively more freely yet be within the bounds of the law. There are many examples where investigative journalists have uncovered concealed (either deliberately or accidentally) crimes [86], such as the Watergate scandal [197]. In more recent years, investigative journalists analysed large amounts of leaked data from the Panama Papers [193] and Paradise Papers [194] to unearth decades of financial wrongdoing, which provoked subsequent law enforcement actions. Figure 1.5 illustrates the scale these leaks had, forcing journalists to use technology and big data techniques to expose misconduct and corruption.

'The key skill required in investigative journalism is research [29]' because the work is hypothesis-driven; journalists need to know how to collect, verify and analyse data; put the story together in narrative order, and finally publish and defend the story [86]. Their work often reveals the truth by collecting admissible information and holding it up for the scrutiny of the courts. Collaborative efforts between these domains are needed to combat the extent and severity of the CaaS business model.

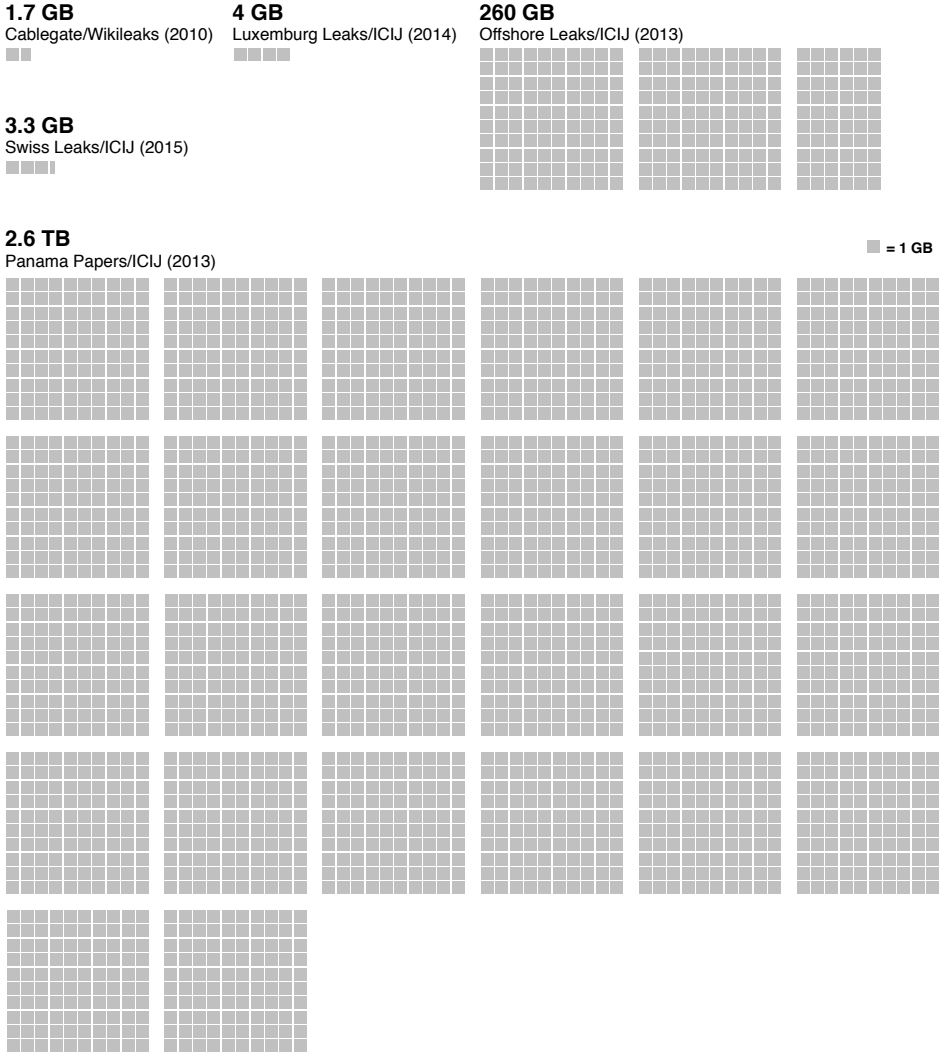


Figure 1.5: Size perspective of investigative journalist cases. Copied from Walmann [183] and ICIJ [88].

Computational forensics

Numerous technical, legal, and resource challenges exist in the digital forensics domain [3, 167, 181]. We refer curious readers to the work being done towards Digital Forensics as a Service (DFaaS) and Hansken [10, 11] in particular, which research addresses challenges in the digital forensic domain. This thesis addresses the most relevant challenge from the increase in data dimensions [148], i.e. big

data: volume, velocity, variety, value and quality [188]. The challenge of big data has already negatively affected the timely process of justice [148] because of the increase in (i) the number of digital devices per case, (ii) the frequency with which digital devices occur in cases and (iii) the amount of data in each case and individual device is growing [148, 162]. For example, already at the time of the Panama Papers (mentioned in the investigative journalism subsection), the Norwegian National Authority for Investigation and Prosecution of Economic and Environmental Crime (Økokrim) conducted investigations with twenty times more material seized [183] (i.e. more than 52 TB from just one criminal case).

Law enforcement increasingly encounters data that cannot be analysed with today's forensic tools, which is seen as the end of the digital forensic 'golden age' [67]. Investigators' new paradigm resolution is to converge data science and traditional digital forensics to tackle the fundamental challenge of analysing the vast amount of data efficiently and effectively [65, 76] while preserving forensic principles to present the results in a court of law. Forensic investigators progressively rely on computational methods and data analysis to support them in their daily casework, e.g. by applying machine learning to triage and analyse forensic disk images and network traffic dumps. Computational forensics is an emerging research domain in forensics concerning a systematic approach for investigating forensic problems using computational methods [65]. It involves interdisciplinary methods such as computer-based processing, analysis, modelling, simulation, visualisation and pattern recognition.

Forensic investigators need to enhance their current toolkit with forensically sound and advanced techniques, methods and algorithms to find tiny pieces of evidence hidden in chaotic environments [33, 208]. The research discipline of computational forensics helps investigators in this process by employing systematic and hypothesis-driven studies to scrutinise forensic problems. Computational forensics provides: (i) better analysis, (ii) improved processing of large-scale data, and (iii) representation of expert knowledge [161], while working towards [65, 208]:

- In-depth understanding of forensic principles.
- Evaluating the basis of particular scientific methods.
- Systematically applying techniques of computer science, applied mathematics and statistics to forensic sciences.

One criticism by the digital investigation community is that (partially) automating some tasks can deteriorate the investigation quality [30]. However, studies in computational forensics carefully control automated solutions to improve the speed and

quality of forensic investigations. The effect is that it allows expert investigators to spend more time on the manual facets of investigations [69, 89, 151, 182].

1.7.2 Machine learning

ML is the study of computer algorithms that improve automatically through experience [100, 125]. ML has gained popularity with the advances in Information and Communications Technology (ICT) and growing volumes and varieties of data. Faster and cheaper hardware combined with the abundance of data also lends support to increasing the capabilities of ML. The advantage of ML – instead of traditional statistics – is the use of inductive learning methods, where the learner discovers rules by observing examples.

More formally, a ML model is the mapping of a function $f : X \rightarrow Y$, where f is a one-to-one or many-to-one function which maps the relationship from elements in the input set X to elements of the output set Y. The function $f(x) = y$ is formally known as the target function and would ideally map every $x \in X$'s full relationship to every $y \in Y$. However, $f(x)$ is unknown to us, and the ML model tries to find a heuristic hypothesis function $h(x)$ instead. This $h(x)$ approximates the unknown target function. ML models 'discover' this function by searching for all possible hypotheses H to find the hypothesis $h(x) \in H$.

Machine learning process model

ML is a process that involves many sequential steps to produce a deployable model. Figure 1.6 illustrates an overview of the steps in the ML process model. Experts and forensic investigators involved in the process may be called upon to testify as expert witnesses in a court of law. As discussed in Subsection 1.7.1, the Daubert standard strictly outlines the admissibility requirements of expert testimony. In court, expert witnesses may be challenged as to their work or qualifications as forensic scientists, e.g., whether they have performed the test correctly, whether the results are interpreted accurately, or whether the underlying science is valid, reliable, and relevant [40, 58]. Therefore, this subsection – and related research articles – focus on data preprocessing and model evaluation, which are the two aspects with the most impact on the final model, and thus, admissibility in a court of law [137].

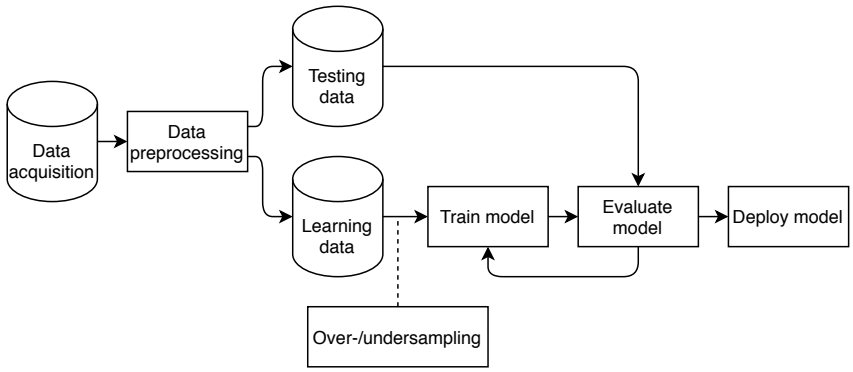


Figure 1.6: Machine learning process model

Acquiring and preprocessing data can be summarised with the common saying that 'data scientists typically spend 80% of their time on data wrangling and 20% on data analysis and modelling.' These are not actual numbers by any means, but this saying highlights the importance data scientists put on processing raw data. What is essential for this thesis is that courts take careful notice of how evidence has been acquired, stored, processed and analysed [126]. This means that a significant focus will be on the documentation (i.e. CoC) of all activities carried out on the data during the entire length of the investigation. Therefore, this thesis and related publications put a significant focus on documenting the data preprocessing steps to be relied upon to assess the forensic soundness of the process. Comprehensive documentation also helps with reliability and repeatability and produces verifiable results for experiments, which are essential in investigations.

Checking and evaluating an ML model is necessary to test the trained model's generalisability [184]. The quickest method of evaluating performance is by creating a train and test split of the dataset. The training set is used to prepare the model, while the testing set is withheld. This allows us to compute the model's performance by comparing the model's predictions with the actual values of the test set. We can extend this type of train and test split using cross-validation, where we do more than one split. Some frequently used cross-validation methods include k-fold and leave-p-out (leave-one-out if $p = 1$), as seen in Figure 1.7.

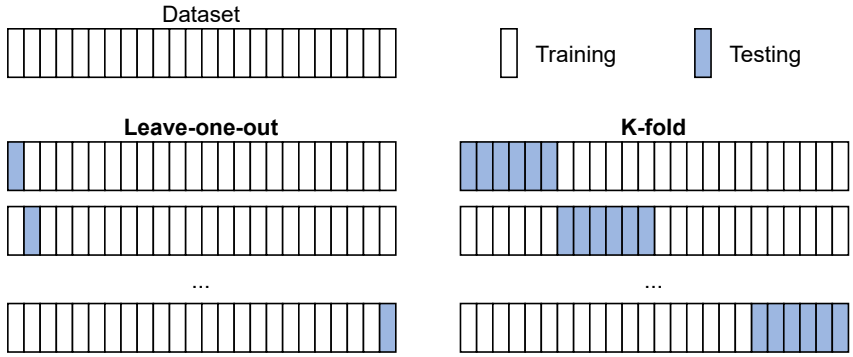


Figure 1.7: Leave-one-out and k-fold cross-validation methods

Natural language processing and topic modelling

Traditional analytical methods have difficulties analysing unstructured data such as text documents and speech recognition [20, 68, 70]. NLP is a subfield in Artificial Intelligence (AI) – with close ties to ML – which give computers the ability to understand, analyse and manipulate human language. In NLP, the ML training algorithms study thousands or millions of text samples – such as words, sentences and paragraphs – to understand the context of human speech, writing and other modes of communication [20]. The strategy is to apply preprocessing steps, such as tokenisation, stop word removal, normalisation and stemming/lemmatisation to reduce unnecessary text noise. The text then goes through an embedding, which converts text into vectors of numbers [72] using methods such as Bag of Words (BOW), Term Frequency - Inverse Document Frequency (TF-IDF) and Word2Vec/Doc2Vec. These embeddings can be used by ML algorithms.

In NLP, a topic model is a statistical model for discovering abstract 'topics' that occur in a corpus [18, 196]. Topics are abstract groups of words that best represent the information in a document collection (i.e. a corpus). More specifically, 'documents can be described by a distribution of topics, and each topic can be described by a distribution of words [177].' For example, 'meow' and 'cat' will appear more often in documents about cats than words like 'poison' and 'radiation', except for documents discussing the thought experiment from Schrödinger. A topic model captures our intuition that certain topic-related words appear more or less frequently in groups.

Some of the most commonly used topic models are [4]: Latent Semantic Analysis (LSA), Probabilistic LSA (pLSA), Latent Dirichlet Allocation (LDA), Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), Gibbs

Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM), and more recent BERTopic [75]. However, many of these topic models have some unfavourable conditions when applying them to different scenarios. For example [4], LSA has topics which are difficult to interpret, pLSA can overfit for large corpora, PCA is expensive to compute for high-dimensional datasets, and NMF sometimes provides semantically incorrect results. Some algorithms become unsuitable because of their technical limitations when analysing our large-scale datasets, while others are less useful for a forensically sound process model.

On the other hand, BERTopic leverages state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) [42] to create dense clusters of words with easily interpretable topics. BERT is pre-trained on both formal and structured text from BookCorpus and the English Wikipedia and another type of text embedding. We tested BERTopic on our corpora in Article V and found it had difficulties with the type of data, which primarily consist of informal and noisy text. This is not an uncommon problem, as classical topic modelling algorithms have poor performance on short, informal, and texts which lack regular patterns [109].

BERT has presented state-of-the-art results in a wide variety of NLP tasks, and it has the potential to improve on the type of task in this thesis. Improvements demand a significant focus on fixing the limitations of BERT (e.g. 512 token limit) and fine-tuning performance with supervised learning approaches, which involves creating labelled datasets, defining a loss function, etc. The LDA and GSDMM methods in this thesis already have good performance through removing a significantly large portion of underground forum users. Fixing limitations and fine-tuning BERT will probably give a minor performance boost, but at a higher cost than the effort is worth. Moreover, fine-tuning BERT is better suited to a large study with a scope in this direction. Although this thesis has a different scope (see Section 1.2), future work can use the methods proposed here to, for example, create partially labelled datasets.

The remainder of this subsection focuses on describing LDA and GSDMM, which are the two algorithms used to produce the results in our research articles. LDA [21] is a generalised version of the older approaches of LSA and pLSA. LDA is a popular topic model because it generalises well. It works by treating every document as a mixture of topics and every topic as a mixture of words. Figure 1.8 graphically illustrate how the LDA algorithm categorises a set of documents into pre-defined k unobservable groups (i.e. topics). The two hyper-parameters α and β regulate two Dirichlet distributions for the document-topic density and topic-word density [21], respectively. Thus, LDA allows for a nuanced way of categorising documents, as each document has some probability of belonging to several topics.

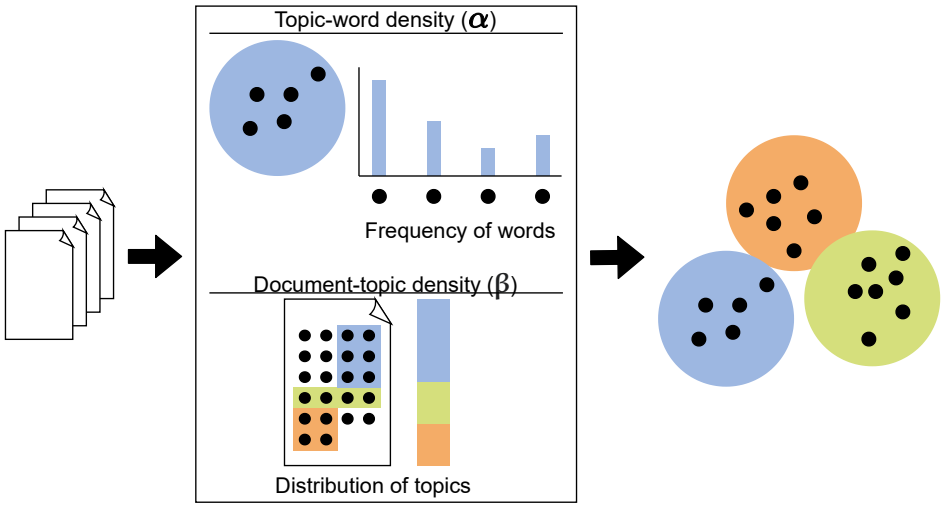


Figure 1.8: Latent Dirichlet Allocation

More formally, LDA is a statistical generative process model, which assumes documents x are generated by latent variables. That is, each document is a mix of topics, and each topic is a mix of words. LDA creates a synthetic approximation $h(x)$ of the underlying complex and unobserved process $f(x)$, which generated the documents. The $h(x)$ is a distribution of k topics over words. Figure 1.9 shows a plate notation of the LDA process model.

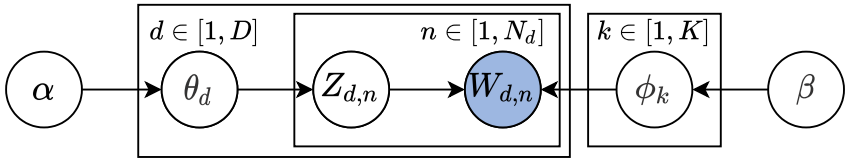


Figure 1.9: Latent Dirichlet Allocation plate notation

The rectangles in Figure 1.9 represent repeated entities, while the shaded and empty circles are observable and latent variables, respectively. The outer plate D represents documents, while the inner plate N represents the repeated word positions in a given document; each position is associated with a choice of topic and word. The plate notation variables are defined as follows:

D	number of documents.
N	number of words in a document (document d has N_d words).
K	number of topics.
α	Dirichlet prior parameter for per-document topic distributions.
β	Dirichlet prior parameter for per-topic word distribution.
θ_d	topic distribution for document d.
ϕ_k	word distribution for topic k.
$Z_{d,n}$	topic for the n-th word in document d.
$W_{d,n}$	specific word.

Although LDA's latent space works for our scenario in Article V, LDA also comes with its drawbacks: poor performance on shorter texts and requires a predefined k number of topics. As seen in Article V, LDA performs well on a specific type of task while having poor performance on a different kind of task. This observation is explained by the 'no free lunch' theorem detailed in Subsection 1.7.2. Analysts can mitigate the drawbacks by choosing the best conditions for when to apply the LDA algorithm versus another algorithm. We apply the GSDMM algorithm [203] for short text classification. GSDMM is similar to LDA with one significant difference: it assumes that a document can only belong to a single topic. Figure 1.10 shows a plate notation of the GSDMM process model. The plate notation variables are similar to LDA notation, with the exception of θ_d because GSDMM does not find the topic distribution for document d. It has been shown that the single topic assumption is more appropriate for short text and that GSDMM, therefore, will outperform LDA on short texts [118].

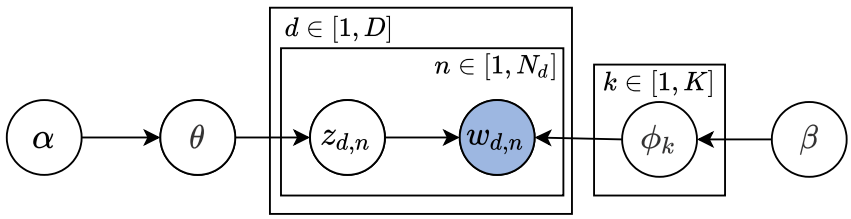


Figure 1.10: GSDMM plate notation [203, 90]

The GSDMM algorithm initialises by randomly assigning each document into k number of topics. For each iteration i, the algorithm consecutively removes each document from its current cluster and assigns it to a cluster according to its probability of belonging to each cluster. The algorithm stops when it reaches the maximum number of iterations.

The goal of any topic model is to find the best fit for the data. A topic model with a

good fit has found more coherent topics, which means that topics contain semantically similar high-scoring words (i.e. those that best represent the topic). The model fit is often measured by perplexity or held-out likelihood. Chang et al. [35] have shown that those measurements and human judgement are often not correlated, so it would be problematic to use them to assess topic relevance. Instead of estimating the relevance, we use perplexity in Article III to qualitatively compare topic models that have been tuned with different combinations of hyper-parameters. The ideas presented in Article IV and Article V do not rely on perplexity but instead assume a human investigator selects appropriate and coherent topics.

Machine learning theorems and principles

A ML model is a simplified representation of reality, which discards unnecessary details and focuses on the aspect we want to understand. This implies that a model cannot learn a simplified representation without making assumptions. This implication is associated with two theorems and one principle, which are essential in the area of ML. They are the 'no free lunch' and 'ugly duckling' theorems and 'Occam's learning' principles.

There are two theorems named 'no free lunch' [130]. The original is by David Wolpert [200] which – in essence – states that we cannot learn from data without making assumptions [73]. Wolpert argues that two predicting algorithms have the same ability to generalise under the assumptions that: (i) the test data and the training data are statistically independent and (ii) the labels have nothing to do with the features [130]. There is another version of the 'no free lunch' theorem by Shalev-Shwartz and Ben-David, which states that no model is always the best [163]. Thus, there are datasets where algorithm A outperforms algorithm B in predicting accuracy on unseen examples. Fundamentally, this means that we need to try several ML algorithms because all algorithms are equally good (and on average equally bad) across all possible learning tasks.

Satosi Watanabe's 'ugly duckling' theorem [186] reflects his recognition that categorising things (as in classification problems) is fundamental to virtually all conceptual processes [9]. Watanabe proved that 'any two objects are equally similar to each other as any other two objects, and are equally as dissimilar to each other as any other pair [186].' Thus, the number of attributes commonalities that two randomly chosen objects (e.g. a duckling and a swan) can have is equal to the number of commonalities that two objects classified as similar may possess. In less formal words, the set of attributes for a duckling contains indefinitely numerous possible shared and unshared attributes with the set of attributes for a swan, making an ugly duckling and a swan just as similar to each other as two swans. The consequence is that we can only differentiate a swan from an ugly duckling with the right set of

attributes.

Lastly, 'Occam's learning' principle states that simple solutions are more likely to be correct than complex ones [38, 192]. This guiding principle is often used to reduce complexity. It can influence the selection or refine models (e.g. pruning for decision trees) to combat overfitting. Another way to reduce complexity is to trim the dimensions of features used by the model, which combat the curse of dimensionality. There are two caveats to this principle that are important to note: (i) there is little empirical evidence that demonstrates that the world is simple and (ii) it is hazardous to reduce complexity at the expense of accuracy. Thus, 'Occam's learning' principle only applies when the predictive power between two models is equally good [59].

Statistics

There is overlap between machine learning and statistics, but they are not identical. This thesis will not try to develop a clear distinction between these fields. Instead, we will give a general and possibly oversimplified distinction that may not satisfy everyone's viewpoints. In our opinion, statistics is interested in learning something about data to arrive at new scientific insights based on the data. In contrast, ML solves complex computational tasks without understanding the problem well enough to write a program that can perform the job. As long as the prediction works well, any statistical insight into the data is unnecessary.

Statistical methods can be put into multiple sub-groups, such as computational and exploratory statistics. For example, ML is closely related to computational statistics [191], while exploratory statistics uses methods such as NBL to summarise the main data characteristics, often visually. NBL (or the first-digit law, expanded by Benford in 1938 [13]) is an observation about the frequency distribution of leading digits in many real-life sets of numerical data. Intuitively, one might assume that leading digits of numbers would be uniformly distributed so that each digit from one to nine appears 11.1% of the time. The simplified version of NBL is that it is often the case that one occurs more frequently than two, two more frequently than three, etc. More precisely, NBL predicts that the frequency for leading digits using base 10 logarithms will decrease, where one is likely to occur 30.1% of the time, while nine occurs 4.6%.

NBL (sometimes just 'Benford's law') is counter-intuitive but shows that natural data often begin with one, two or three more frequently than seven, eight and nine. Therefore, the most prominent applications of the NBL are the detection of fraud in forensic accounting/auditing, insurance claims, expenses reimbursements, and so forth [34, 121, 133, 134]. Figure 1.11 illustrates the NBL with an example,

where we have a set of real-world number observations and the expected frequency distribution set by NBL. We see that these observations violate this frequency distribution, and, therefore, we can begin to suspect that the observations have been tampered with. No test is foolproof, and not every real-life set of numerical data satisfies the NBL. Consequently, the NBL provides fraud examiners with a different method to test data for indications of fraudulent activity, but further analyses are needed to establish the existence of fraud conclusively.

The casework report in Chapter 8 does not use the NBL to look at the first digits, but it uses the idea that numbers follow a curve. We can observe that playbacks on some regular days follow the expected curve, while some days have irregular playbacks, similar to the observed data in Figure 1.11.

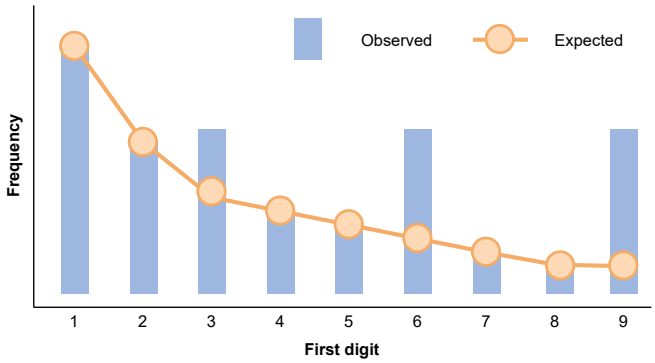


Figure 1.11: Observations violating the Newcomb-Benford frequency distribution

1.7.3 Social network analysis

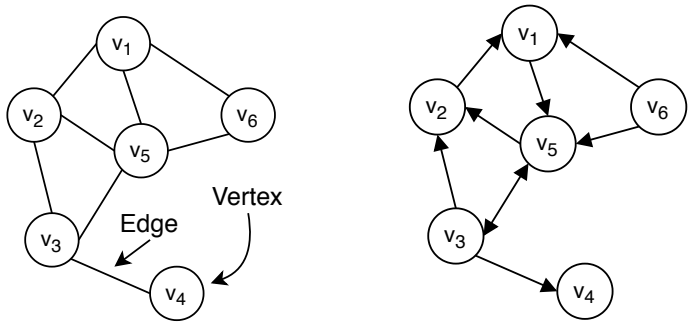
People form many social relationships, and it is normal for a person to be a member of different social networks. A few examples of networks include families, colleagues, friends, even criminals and gangs. SNA is a multidisciplinary research area [147] and contains techniques forensic investigators can use to uncover the social relations and structures formed by these networks. Social networks are formally defined as graphs, with a set of actors and a set of ties. This subsection continues with a detailed and mathematical explanation of graph theory before explaining the essential theory about network centrality.

Graph theory

Graph theory is a mathematical field, and it is the study of mathematical structures known as graphs, which represent pairwise relationships between objects. More formally, a graph $G = (V, E)$ consists of the sets V and E, where V is a non-empty set of vertices and E is a set of edges. The terminology for vertices and

edges varies between different disciplines, e.g. law enforcement agencies call them entities and relationships, while SNA calls them actors and ties. However, this subsection will follow the exact mathematical terminology.

The edges in graphs can have different properties depending on the context they represent. Figure 1.12a illustrates an undirected graph where the edge set contains unordered (v, u) pairs of edge elements. This means a mutual relationship between the elements $(v, u) = (u, v)$. Contrastingly, Figure 1.12b shows a directed graph (digraph) with ordered (v, u) pairs of edge elements. The order is significant as there is an exclusive relationship between the elements $(v, u) \neq (u, v)$.



(a) Graph **(b)** Directed graph

Figure 1.12: Figure with two graphs [91]

Graphs and digraphs are typically represented by either a list or a matrix. In the latter case, graphs are described by square $n \times n$ adjacency matrices, where n is the number of vertices ($|V|$). Each cell in the matrix indicates whether pairs of vertices are connected by an edge or not in the graph. The unordered pairs of elements in E for undirected graphs make their adjacency matrices symmetric, as seen in Table 1.3 for Figure 1.12a. In contrast, the ordered pairs set for digraphs can make them asymmetric, as shown in Table 1.4 for Figure 1.12b.

A binary adjacency matrix A is a zero-one matrix with 1 in its (i, j)th cell when v_i and v_j are adjacent; otherwise, the cell value is 0. Thus, a cell represents the presence or absence of edges between rows and columns. Note that the adjacency matrix cells may contain other numerical values [152], making it a weighted (di)graph. The advantage of a matrix representation is that it is easy to work with because edge checking, adding and removing are done by examining the (i, j)th cell. However, matrices require more memory, which is preferred when working with dense graphs. Adding and removing vertices for dynamic networks makes matrix representation relatively slow.

Table 1.3: Symmetric graph matrix

	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	0	0	1	1
v_2	1	0	1	0	1	0
v_3	0	1	0	1	1	0
v_4	0	0	1	0	0	0
v_5	1	1	1	0	0	1
v_6	1	0	0	0	1	0

Table 1.4: Asymmetric digraph matrix

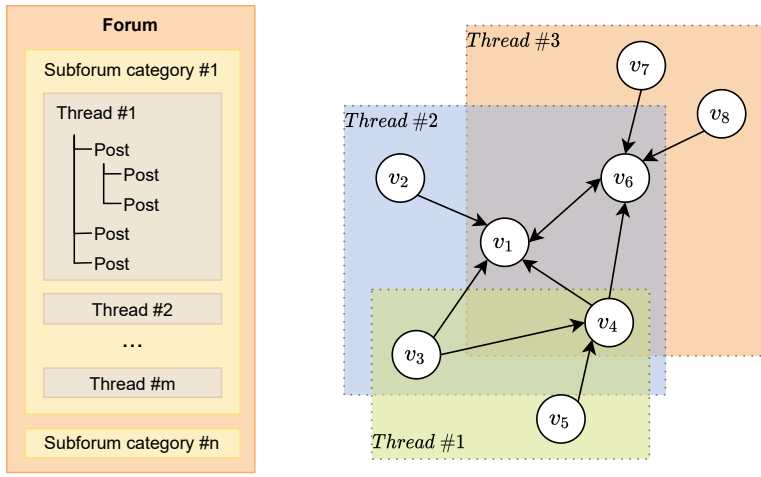
	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	0	0	0	1	0
v_2	1	0	0	0	0	0
v_3	0	1	0	1	1	0
v_4	0	0	0	0	0	0
v_5	0	1	0	0	0	1
v_6	1	0	0	0	0	0

The vertex degree – denoted $deg(v)$ for vertex v – is the number of edges incident (i.e. directly connected) with a vertex [152]. A loop contributes twice to the vertex degree. For example, the degrees in Figure 1.12a are $deg(v_4) = 1$, $deg(v_6) = 2$, $deg(v_1) = deg(v_2) = deg(v_3) = 3$ and $deg(v_5) = 4$. An isolated vertex has degree zero because it is not connected to any other vertices in a graph. On the other hand, v_4 in Figure 1.12a is called a pendant vertex since it relies on the other connected vertex to stay connected to the graph.

Graph construction

Underground forums share a similar hierarchical message structure: a forum consists of several broad and general subforums. Each subforum contains single-topic forum threads, which are composed of an ordered collection of posts [146]. Figure 1.13a illustrates a typical hierarchical message structure. The forum administrators choose the general subforum categories, while individual members contribute with a thread for a specific subforum category. Other forum members can post messages in the forum thread to reply to the thread starter, reply to other post authors, or provide generic comments.

The research articles in this thesis create interaction networks based on the posts in an underground forum. The network's vertices represent forum members who participate in forum discussions, and we often refer to them as actors. A thread starter initiates an interaction. Any actor replying to the thread at any given time will participate in the interaction, regardless of whether their reply directly addresses the thread starter or any other participant in the thread. Figure 1.13b illustrates a small interaction network with three thread starters: v_1, v_4, and v_6. An interaction explicitly manifests itself as a directed or undirected network edge starting from post author v_j and going to thread starter v_i.



(a) Forum hierarchy **(b)** Constructing an interaction network

Figure 1.13: Structural hierarchy and how to build an interaction network

Network centrality measures

SNA breaks down the notion of a network into different levels, focusing on distinct components of a network [147]. The typical levels concentrate on individual actors, pairs of actors, groups of three actors, subgroups or complete networks. This thesis focuses on network centrality measures that analyse a network at the actor level. Centrality measures address the most important or central actor in a network. Thus, the measures identify actors with more opportunities and fewer constraints than other actors. These actors may benefit from exchanges, have more significant influence and be a focus of attention from those in less favoured positions [79, 147]

There are many answers to who is the most important actor, so there are many ways to measure the importance of vertices, as explained in this subsection. Figure 1.14 and Figure 1.15 demonstrate the popular centrality measures: degree (in- and out-degree), betweenness, closeness and eigenvector on an undirected and directed graph. Each measure has its definition of 'importance', and it is crucial to understand the differences for a particular application of them.

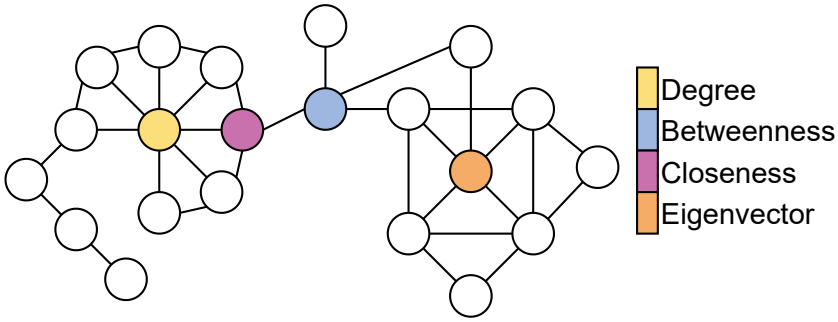


Figure 1.14: Undirected network centrality measures. Adapted from Ortiz-Arroyo [139].

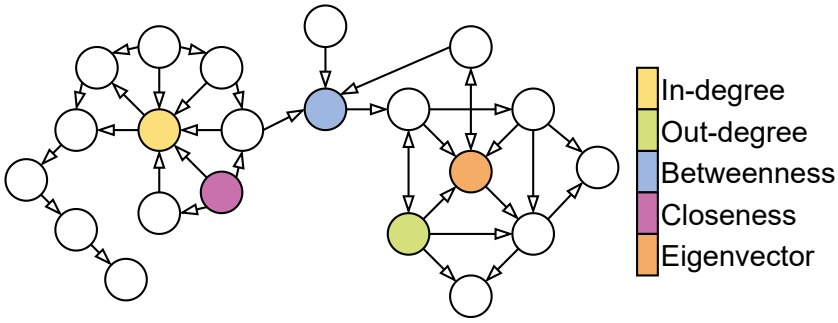


Figure 1.15: Directed network centrality measures

The most straightforward measure is degree centrality, which calculates the number of directly adjacent vertices an actor has in a network. For simplicity, let A be a binary adjacency matrix of size $n * n$, where $A_{ij} = 1$ if i has an edge to j; otherwise $A_{ij} = 0$. Furthermore, let the principal diagonal of A equal 0, i.e. $A_{ii} = 0$. Equation 1.1 [22, 66, 205] shows the formula to calculate degree centrality for actor i in a symmetric matrix.

$$C_D(i) = \sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} A_{ji} \tag{1.1}$$

It is important to note that network centrality measures can only make meaningful comparisons between actors in the same network or networks of the same network size [147]. Thus, we normalise centrality scores by calculating their network proportion to compare actors between networks of different sizes or for networks that may change their size over time [202]. Equation 1.2 [66] give the formula for normalising degree centrality.

$$C_D^{'}(i) = \frac{C_D(i)}{n-1} \tag{1.2}$$

Degree centrality has been split into an in-degree (C_{D^-}) and out-degree (C_{D^+}) centrality measure for digraphs. The in-degree is the number of edges adjacent to other vertices. This measure is often used to measure prestige or popularity [147] because actors seek ties to them. Equations 1.3 and 1.5 [147] provide the formulas for in-degree and out-degree centrality, respectively. These measures are normalised using the Equations 1.4 and 1.6 [147].

$$C_{D^-}(i) = \sum_{j=1}^{n} A_{ij} \tag{1.3}$$

$$C_{D^-}^{'}(i) = \frac{C_{D^-}(i)}{n-1} \tag{1.4}$$

$$C_{D^+}(i) = \sum_{j=1}^{n} A_{ji} \tag{1.5}$$

$$C_{D^+}^{'}(i) = \frac{C_{D^+}(i)}{n-1} \tag{1.6}$$

Degree, in- and out-degree measures only consider the direct ties for each actor and ignore the other actors and edges in the network. Thus, these measures can only be viewed as calculating an actor's level of involvement or activity in the network [147]. On the other hand, other network centrality measures are more robust because they consider the other actors in the network. Betweenness centrality captures the dimension of centrality that it is not important how many people they know, but rather where they are placed within that network [147].

Betweenness calculates how often an actor i sits on the shortest paths (geodesic) between two other actors h and j. There can be multiple geodesics between two actors in a network. The idea is that the i actor is important since this actor can significantly influence the network by choosing to withhold or distort information [147]. The formula is given in Equation 1.7 [22, 66, 205], where δ_{hj} is the total number of shortest paths from vertex h to j and $\delta_{hj(i)}$ is the number of those paths that pass through i. Normalised formulas for betweenness centrality are given for a graph and digraph in Equations 1.8 and 1.9 [185], respectively.

$$C_B(i) = \sum_{h \neq i \neq j} \frac{\delta_{hj(i)}}{\delta_{hj}} \tag{1.7}$$

$$C_B^{'}(i) = \frac{C_B(i)}{(n-1)(n-2)/2} \tag{1.8}$$

$$C_B^{'}(i) = \frac{C_B(i)}{(n-1)(n-2)} \tag{1.9}$$

Closeness centrality captures an actor's potential pendency by measuring the sum

of distances between an actor i and all the other actors. The length of their shortest paths defines the distance. The formula is given in Equation 1.10 [22, 66, 205], where $d(ij)$ is the distance (length of the shortest path) connecting actor i to actor j. The normalised formula for closeness centrality is given in Equation 1.11 [207].

$$C_C(i) = \sum_{j=1}^{n} d(ij) \qquad (1.10) \qquad C'_C(i) = (n-1)C_C(i) \qquad (1.11)$$

Eigenvector centrality expands on the notion of degree centrality by measuring the edges of neighbouring actors. Thus, neighbours highly affect an actor's eigenvector score because their importance is based on their friend's importance [147]. The formula is given in Equation 1.12 [19, 23, 132], where $\lambda \neq 0$ is the largest eigenvalue calculated, $M(i)$ is a set of neighbours to vertex i, j is a neighbouring vertex, and A_{ij} takes a binary value depending on whether or not i and j are neighbours.

$$C_E(i) = \frac{1}{\lambda} \sum_{j \in M(i)} C_E(j) = \frac{1}{\lambda} \sum_{j=1} A_{ij} C_E(j) \qquad (1.12)$$

1.7.4 Ethical and legal deliberation

Data collection is a central part of the scientific process, and scientists follow strict guidelines and regulations to conduct their research ethically [52, 57]. We acquired our datasets in non-traditional ways: (i) from leaked hacker underground forums and (ii) hard disk with data from a whistle-blower. Thus, we want to briefly discuss some ethical and legal considerations because of the secondary use of possible illicitly obtained data. Readers are referred to Thomas et al. [174] and Boustead and Herr [27] for more lengthy discussions and arguments surrounding this issue.

Arfer and Jones studied the 2015 Ashley Madison data leak. They argued that using data that have initially been illicitly collected is itself ethically permissible, stating that: 'we cannot undo the past, but we can make the most of the present by getting what social and scientific value we can out of undesirable events [7].' In contrast, other scientists have denied using leaked datasets when 'the negatives have outweighed the positives, especially when they could gather all or most of the same data in a more legal and accepted manner [27].'

We downloaded two leaked databases of deep web underground forums. They were leaked in 2016 and 2019, and we published our results from both leaks. Researchers generally obtain informed consent before conducting research that involves people [27]; however, it is an impossible task to contact a total of 920 529 forum users to get their permission. To further complicate matters, it is unlikely

that cybercriminals would consent to us conducting our research to prevent their criminal activities. Thus, we must consider that our research poses a minimal additional risk of harm to subjects since we cannot feasibly obtain their consent.

The two main potential harms in our work were the invasion of privacy and revealing information about individuals that might be used against them. We implemented the following safeguards for the first issue: (i) protect integrity and confidentiality to avoid accidental leakage and control sharing; and (ii) not revealing any identities or attempt deanonymisation. However, implementing safeguards for the second main issue would contradict our studies' goals: to investigate and reveal hidden information that could be used against individuals.

We justify using leaked datasets because our research contributions maximise benefits and minimise harms. Thus, the potential benefits of analysing this data outweigh the harms caused by obtaining it [44]. The data from leaked underground forums describes potentially illegal or harmful activity, so the apparent justification is that law enforcement may use our research to prevent or limit such activity in the future. Gathering informed consent from subjects would considerably reduce the available data and risk skewing research results. Furthermore, 'the use of [data with an illicit origin] provide[s] researchers with the opportunity to test their hypotheses against ground truth data [174].'

Investigative journalists sometimes work with whistle-blowers and follow strict procedures to keep their anonymity [29]. Journalists from the Norwegian newspaper Dagens Næringsliv (DN) hired us to analyse data on a hard drive they had acquired. The data lacked any personal or sensitive personal information and primarily contained songs, such as who played which song at what time. Thus, we mainly safeguarded the confidentiality of the data. Furthermore, the journalists did an excellent job of validating and verifying the information on this hard drive until, at one point, we were confident that this data was original and accurate. The journalists also gave the accused party time to respond to our findings of wrongdoings before publishing the story and our casework report found in Chapter 8.

1.8 Related work

This section describes related work which used SNA and NLP to analyse data from criminal networks. Subsection 1.8.1 starts with describing early research on physical criminal networks, which laid the foundation for future understanding and SNA research on other types of networks. The subsection continues covering more related work which analysed key actors in criminal underground forums. Subsection 1.8.2 covers the use of NLP to analyse user's reputations/influential actors, identifying threats and understanding the community.

1.8.1 Social network analysis and criminal network analysis

As the name suggests, SNA predominantly analyses social network structures, such as friendship, acquaintance and business networks. The networks were typically established by interviewing actors [147], which resulted in smaller networks focused on some individuals. It was not until the Enron scandal [189] and the subsequent release of the Enron email dataset [165] that researchers had the opportunity to study a 'large-scale' real-world dataset. The Enron email dataset is well-researched from the perspective of social science. For example, Diesner and Carley [43] studied the Enron network's structural properties before and after the Enron scandal. Their findings of the network's two snapshots at two different times suggest that the organisation became denser, more connected and centralised during the scandal than in normal times. For example, in higher Enron positions, actors were sending more emails during normal times while receiving more emails during the crisis.

Investigators in intelligence and law enforcement have long used networks to model the collective effort in offences such as terrorist attacks, narcotic trafficking and armed robbery. They try to examine the characteristics of individual offenders as well as the criminal organisation. Past research articles since the 1990s have proposed using methods from SNA for analysing criminal activities and understanding criminal network structures [128]. The idea is that the knowledge and insights gained from this analysis may help investigators 'target offenders for removal or select effective strategies to disrupt a criminal organisation [202].'

Social network analysis of physical criminal networks

Past research has largely restricted itself to what is arguably the most common and straightforward network concept: centrality. Centrality is often used to indicate a member's importance within a group or a network [147]. Various centrality measures have been proposed for identifying prominent actors in criminal or covert networks. However, centrality has different meanings and implications when approaching it in a criminal setting [128].

Network centrality measures have been developed for the social sciences, and Sparrow [169] has accentuated that they are designed for small, static networks and with very few types of edges (typically only one as discussed in Subsection 1.7.3). Sparrow focused on exploring physical criminal networks due to the scarcity of ICT and introduced the opportunities for network analytic techniques to the area of criminal intelligence. He asserts some centrality measures have greater relevance in identifying 'network vulnerabilities' than the others because some individuals' arrest would impede continued operation of criminal activities. The idea that the

most valuable targets are central and difficult to replace is substantiated by today's law enforcement [53].

Baker and Faulkner [8] studied the organisation of a conspiracy (a type of white-collar crime) to price-fix switchgear, transformers and steam turbine generators. They explore the extent to which social theories on legal/social networks can be generalised to illegal networks. They 'find that the structure of illegal networks is driven primarily by the need to maximise concealment, rather than the need to maximise efficiency [8],' due to sparse and decentralised networks.

Under other conditions, Krebs [101] mapped the covert network surrounding the nineteen terrorist hijackers responsible for the September 11th attack in 2001. He created this mapping iteratively as data became available from various news sources but encountered similar challenges as Sparrow [101, 169]: (i) incompleteness because investigators will not uncover all vertices and edges, (ii) fuzzy network boundaries from the difficulties of choosing who to include or exclude and (iii) these networks being not static but dynamic groups. Krebs's results indicate that the terrorist network was sparse and many of the hijackers distanced themselves from each other. This could be a conscious strategy to 'minimise the [damage to the covert network] if a cell member is captured or otherwise compromised [101].'

It is easy to identify central actors in smaller networks due to the relatively few criminal actors and connections. Such analysis can often be done by eye; however, SNA is required when criminal networks become more complex and involve large numbers of actors and connections. Morselli [128] investigated the individual positioning within an illegal drug distribution network with the overlap between degree and betweenness centrality using data obtained from an investigative criminal case. He found a strong and significant correlation between degree centrality and being arrested, while betweenness centrality only had a positive relationship. Thus, lower-level members in the criminal organisation were most likely to be arrested during the investigation, suggesting that the visibility that comes with high degree centrality does outweigh any strategic capital that may come with betweenness centrality [127]. Morselli suggests that law enforcement should monitor participants shift towards greater brokerage and evaluate them because this should indicate a more privileged place in the network.

Lu et al. [111] gathered information from news sources about a hacker community and used text mining to extract entities and their relations, i.e. NER. They studied this community's social structure and found that network centrality scores are low, implying that this hacker community was decentralised. In other words, actors have similar centrality scores, and no single actor stands out. Furthermore, they

examined four centrality measures to determine an individual's importance and identify the network leadership.

Décary-Hétu and Dupont [49] analysed data from a police operation against computer hackers running botnets. Similarly to our research, they pointed out that missing data leads to incomplete networks and erroneously give minor actors a much higher 'criminal score', i.e. centrality score. Furthermore, they also suggested mitigating this problem using other data sources, such as emails or chat messages. They expressed that, although centrality measures provide scientific and objective measures of the structure of networks and positions of key actors, they should not be presented as a 'silver bullet' for this type of problem [49].

Hardin et al. [81] examined people's ability to learn from studying social relationships in the Enron email corpus [165] by applying six different centrality measures. They touched on an important point in their article: 'A measure of centrality [...] aims to assign a ranking or magnitude to each [vertex] that captures the relative importance of that [vertex] in the context of the graph's structure [81].' Hardin et al. simply assumed that higher volumes of emails must have more significance than lower ones without giving details for interpreting their results properly. Importantly, their work suggests that centrality measures help judge various employees' functional importance instead of reflecting an organisation's administrative structure.

Traditional centrality measures do not accommodate the idea of weighted relations, as they were defined solely for binary relations. Schwartz and Rouselle [158] build upon Borgatti's SNA-based key actor approach [24, 25], by incorporating weighted edges to differentiate the strength between actors. They modified centrality algorithms to include attribute and edge weights to help law enforcement and intelligence target actors more appropriately. Furthermore, Memon [123] proposed that generalised network centrality measures should depend on the edges' number and weight. His work suggests that a balance between the number and weight of edges can retrieve more precise knowledge about the actors.

All those studies in previous work are static analyses in which data is collected at a single point and studied. However, this network snapshot can quickly misrepresent a criminal network because criminals establish new relations or break existing ones as their social positions, roles and power change. Xu et al. [202] dynamically analysed a criminal network using descriptive methods from SNA to help detect and describe these changes in criminal organisations. They examined a narcotics network consisting of 924 criminals, where the edges were weighted based on the frequency with which two criminals involved committed crimes together between 1983 and 2002 [201]. A series of yearly networks were generated, and each cent-

rality measure was calculated to capture changes in the network. They found that two leaders alternated in being active as the other one was put in prison. Furthermore, one of the leaders avoided connecting too many other people and still maintained leadership over the network through other persons.

Social network analysis of underground forums

Today's underground forums contain many thousands of active users in contrast to traditional criminal networks. Moreover, the vast majority will be involved in minor deviance levels [143], and only a limited number of cybercriminals are highly skilled users [84]. Therefore, the goal is to identify proficient and high-profile actors who play a significant role in cybercrime activities. There is a broad assumption that cyber criminals with the greatest skill play a central or significant role in underground forums and are looked up to by others. E.g. Holt et al. [84] found that actors with high technical skills are located centrally within communities.

Abbasi et al. [1] proposed a framework using text analytics for key hacker identification. They employed an interaction coherence analysis to extract interactions between users in hacker communities. The result of their study is that certain users use special lexicons, post more embedded code or have a disproportionally higher number of starter threads and forum posts. Moreover, Samtani and Chen [154] utilised SNA to identify key hackers systematically, indicating that many key hackers are the most senior and longest-tenured members of their community.

Hacker forums can be a rich data source to enhance the organisation's Cyber Threat Intelligence (CTI) abilities. Grisham et al. [74] examined hacker forums from the perspective of CTI to identify mobile malware and associated key hackers. The result of their study indicates that many 'key hackers' hold administrative positions in hacker forums. In contrast, Pastrana et al. [143] identified variables relating to forum activity that predict a user's likelihood of becoming an actor of interest to law enforcement. They utilised several approaches to manually identify 113 key actors who had been linked to cybercrime activities and extracted 44 features (including many forum activity variables) and grouped actors based on their activity using k-means clustering. Notably, they say these approaches are not scalable due to the manual effort.

Pete et al. [146] created six networks using posts from exclusively dark web forums and identified actors of importance through network centrality analysis. They analysed data from the CrimeBB dataset [144], which is data scraped from various dark and clear web forums. They have similar findings to our studies, where central actors mostly post in forum threads with broader topics such as general

discussions and tutorials. Furthermore, there are overlaps in the identification of central actors across the centrality measure. Finally, Pete et al. [146] found they could manually inspect users' posts to reveal information about these users, such as the role these users might play in the forum.

Disrupting criminal networks

We only want to briefly discuss disruption strategies, although it is not within this work's scope to provide a detailed analysis of supporting strategies. Studies have shown that criminal organisations must be considered social networks that form responsive and non-hierarchical internal relations rather than organisations with a 'kingpin' [47, 99, 101, 127, 131, 168, 169, 178]. Consequently, law enforcement cannot assume entire criminal organisations will collapse after targeting the 'kingpin' but realise the serious nature of resilience when they try to control organised crime [47]. Instead of using traditional law enforcement strategies, they must look into changing the state of a criminal network by disrupting it in such a way that it cannot efficiently diffuse information, goods and knowledge [47, 48, 110]; e.g. by targeting actors occupying strategic positions (degree and betweenness centrality) or actors which embody special competencies or knowledge in networks.

Duijn et al. [47] studied the dynamic resilience within criminal networks due to disruption to find effective strategies to control criminal networks. They achieved this by simulating and observing the mechanisms of disruption and resilience. They discovered that network efficiency was barely affected even after removing several actors and network efficiency would, instead, increase over time as a direct result of network recovery. However, this decreased the network's internal security because they become more exposed, which offer law enforcement agencies opportunities to target them. Therefore, effective law enforcement disruptions are critical because criminal networks develop capacities to absorb and withstand disruption and change when necessary.

Summary of social network analysis research

Many researchers have considered the structure of underground forum communities and tried to apply network centrality measures for identifying prominent cyber criminals. The level of accuracy of centrality measures depends on the quality of the observed network data [123], which should bring up several issues with the many varying data sources that researchers employ, such as information from police data [49], news sources [111], judges' sentencing comments [28] or web crawling [136, 144, 155]. For example, police data and sentencing comments typically contain data on individuals with a midpoint of an investigation, such as being targeted or prosecuted by the police. This information is most likely biased

towards those individuals because the police have not gathered enough information from other parts of the criminal network. Therefore, network centrality would most likely confirm the bias that these individuals are the 'most central' actors, when they – in reality – are only very visible in an investigation.

Table 1.5 contains a list of previous work that used various resources to construct graphs of criminal networks. These networks are typically constructed from data gathered, in one way or another, by the police. This table compares the network size of our research with 16 other related works. Our research is the earliest work to study networks of such a large scale and complexity in a digital forensic context. The work of Pastrana et al. [143] is the only research that may come close in regard to network size.

The access to leaked and complete databases from hacker communities allows us to avoid most of the challenges faced by Sparrow [169] and Krebs [101], because we can (i) reveal all the vertices and edges to construct a complete network and (ii) consider all users to be within the boundary of the network. Therefore, this unique access to complete databases gave us many opportunities that few other researchers had previously. For example, we look at a much bigger and complete network, which allows us to study network centrality measures in their best-case scenario. Furthermore, it provides some ground truth which we could compare our results against.

The 'no free lunch' theorem (discussed in Subsection 1.7.2) states that there are datasets where algorithm A outperforms algorithm B. Thus, we are not justified in believing that algorithms – designed for small social networks – can generalise to larger unstructured underground networks. Furthermore, it is unknown which centrality measure is better at identifying prominent cybercriminals. Consequently, there is insufficient support for network centrality measures' applicability according to the Daubert standard (discussed in Subsection 1.7.1), potentially rendering their results inadmissible in a court of law. It is important to establish these facts to correctly interpret the centrality results and scores, instead of assuming they assign a 'criminal score' [49] to actors.

Table 1.5: Comparing network sizes from previous work [97] (sorted by year)

Research article	Nodes	Edges
Baker and Faulkner [8] (1993)	78	
Krebs [101] (2002)	19	
Xu and Chen [201] (2003)	164 – 744	
Xu et al. [202] (2004)	924	
Diesner and Carley [43] (2005)	227	
Lu et al. [111] (2010)	23	
Morselli [128] (2010)	174	
Memon [123] (2012)	62	153
Holt et al. [84] (2012)	336	
Décary-Hétu and Dupont [49] (2012)	771	
Abbasi et al. [1] (2014)	4 576	
Hardin et al. [81] (2015)	156	
Samtani and Chen [154] (2016)	6 796	
Grisham et al. [74] (2017)	100	562
Johnsen and Franke [92] (2017)	599 086	371 002
	599 085	2 672 147
Pastrana et al. [143] (2018)	572 000	
Johnsen and Franke [94] (2018)	75 416	319 935
	33 647	98 253
	299 105	2 705 578
Pete et al. [146] (2020)	22 – 16 401	57 – 624 926
Johnsen and Franke [96] (2020)	94 832	490 268
	62 933	794 868
Johnsen and Franke [97] (2022)	21 432	64 938
	299 701	2 741 464
	185 806	1 794 947

1.8.2 Natural language processing on underground forums

Network centrality measures provide a scientific measurement of key actors in related research; however, their results should be interpreted in the context of the network's structure [81], and they should not be presented as a 'silver bullet' for this type of problem [49]. Thus, researchers should use other data sources, e.g. forum features and posts, to interpret and enhance centrality results to better understand the key actors. The use of NLP to analyse underground forums is a recurring technique in this regard. This research is largely focused on three areas: (i) identifying the most influential community members, (ii) understanding the community

structure and social relationships and (iii) identifying threats found in the content and other content-related features.

Identifying the most influential community members

Early work for using other data sources was based on a user's reputation to identify the most prominent cybercriminals. However, recently it has been suggested that most forums lack the ability of peer-assigned reputation scores [115]. Moreover, Motoyam et al. [129] highlighted the challenge of gathering data from underground forums, because of the combination of public, restricted and private sections. Thus, it might be difficult for investigators to acquire the necessary data to interpret centrality results. Their research also indicates that users' reputations come from being publicly active. This implicitly means that underground forum users post valuable data in public and semi-public sections, which investigators can benefit from.

Benjamin and Chen [14] analysed the relationship between hacker behaviour and reputation by exploring how hackers become key actors. The idea is that more trusted users – i.e. those who have good reputations – can often cooperate or receive help from others [160]. Their results suggest that reputable and trustworthy hackers contribute more to content diversity or novelty of information than simple forum involvement or tenure. In contrast to Motoyam et al. [129], reputable users are involved in the forum other than just being publicly active.

Marin et al. [115] recently suggested that a user's reputation score is a strong indicator for identifying key hackers [206]. Thus, their research concentrates on developing an approach for identifying users with higher reputation scores based on 25 forum features. These features are based on forum metrics, such as forum activity, indicated expertise, structural network positions and influence. They evaluate their method against several dark web forums with a peer-assigned reputation score, which is thought to mirror how other forum members evaluate the usefulness of the user's contributions. They suggest that their model can be generalised to other forums that lack a user reputation system or have a deficient one.

Abbasi et al. [1] proposed an automated framework for identifying and characterising expert hackers, which leverages text analysis, feature extraction, and clustering to characterise hacker specialities. Their result was four clusters which identified black market activists, founding members, technical enthusiasts and average users. Abbasi et al. [1] assume that the former three clusters constitute the key actors in hacker communities. Their result indicates that founding members had extensive interactions among themselves and played a vital role in bridging technical enthusiasts and black market activists.

Understanding the community structure and social relationships

The main engine of underground forum marketplaces is trust, where sellers and buyers can be banned for not complying with transaction rules. Related work has examined products and services found in underground forum marketplaces using NLP. Marin et al. [114] scraped and examined seventeen underground forum marketplaces. They examined these marketplaces to understand the product categories using a combination of manual labelling and clustering. Vendors typically advertise and sell products, materials and services related to malicious hacking, drugs, pornography, weapons and software. Additionally, Marin et al. [114] note that many items are cross-posted in multiple marketplaces, sometimes under the same vendor pseudonym. This suggests that large parts of the CaaS business model can be disrupted if the right individuals are targeted.

Huang and Chen [85] proposed a topic-based SNA and clustering approach to identifying the key actors and their roles in the CaaS chain. They used topic modelling to define the possible roles (keyword-based) in the CaaS chain and validate their results. They find that using their method can help identify key actors and their roles, influence levels, and social relationships. More interestingly, they find that many key actors across forums are actually the same individual using various pseudonyms, and they tend to have 'poor' reputation, suggesting that they were reported as deceivers or other complaints about their integrity.

Pastrana et al. [143] mentioned the problems NLP techniques could encounter from the underground forum data. The problems include technical jargon and non-standard expressions, such as non-native English speakers and short texts. They combined the results of a logistic regression model with k-means clustering and SNA to predict the likelihood a user would become an actor of interest to law enforcement. Results were verified using topic analysis to confirm whether these users are engaged in a cybercrime-related activity where the key hackers talk in similar, hacking-related terms.

Identifying threats in the community content

The research group surrounding Prof. Dr. Hsinchun Chen has been particularly active in this problem area, with research from Benjamin, Li and Samtani. Benjamin have several articles focused on examining hacker language and keyword-based identification of hacker assets. In particular, Benjamin et al. [17] proposed a framework which performs comprehensive keyword weighting and search for identifying potential hacker threats, such as attack vectors, software vulnerabilities, financial fraud and other threats. Although their framework is an automated and scalable approach, it does need to be constantly adjusted to reflect changes in

underground forums over time, such as adding or removing keywords. This approach is not particularly generalisable to other types of underground forums and it can be vulnerable to deliberate changes in keywords to avoid detection.

Furthermore, Benjamin and Chen [15] applied NLP to analysing text to understand hacking-specific terms, concepts, tools and other unfamiliarities. More specifically, they used lexical semantics as a way to learn about 'hacker language'. They suggest this work will help security researchers and practitioners learn the latest trends within hacker communities in a temporal analysis, to find emerging hacker terms and threats. Moreover, Benjamin and Chen [16] were motivated to explore computational techniques that support automated categorisation of multilingual underground forum participants into varying groups. They used paragraph vectors to generate fixed-length vector representations of messages posted by users and clustered them into language groups. Their results suggest that paragraph vectors outperform traditional n-gram frequency approaches. Thus, they could categorise participants into different geopolitical origins based on the language usage.

In contrast, Li focused on the carding business, which encompasses stealing, reselling and using large volumes of payment information to commit fraud [105]. Li and Chen [104] presented a deep learning-based framework for identifying top malware and carding sellers. They retrieved relevant threads in an iterative fashion, using a snowball sampling technique with seeding keywords. Their framework identified top sellers with higher seller feedback scores, which is based on the overall customer feedback for the provided products or services. Li et al. [106] extend their idea of analysing customer reviews, by presenting a text mining system for identifying high-quality carding services. Existing methods are unable to adequately estimate the service quality, so they developed a method for evaluating customer reviews which is domain-specific.

Li et al. [105] developed a text mining system for identifying and profiling key sellers. Their system identifies sellers using sentiment analysis of customer reviews and profiles them using topic modelling of advertisements. They base their system on two aspects: (i) sentiment analysis to evaluate seller quality based on customer reviews and (ii) topic modelling to profile sellers based on their advertisements. Thus, their system interprets positive, negative and neutral emotions in the text in consumer reviews to determine the seller's quality.

Samtani et al. [155] focused on the assets which may be used in a cyberattack. Their study aims to understand the functions and characteristics of assets by applying classification and topic modelling techniques. These assets commonly involve attachments (books, videos, pictures, executables, tools and various other programs), source code and tutorials. They developed a framework which de-

termines the purpose of attachments, source code and tutorials. LDA is used to understand the topic characteristics of hacker assets.

Samtani et al. [156] extended their previous work by leveraging a novel CTI framework which automates web crawling, data analysis and text mining to analyse vast numbers of malicious hacker tools. Thus, their framework can find openly available malicious assets to provide intelligence on tools that hackers have developed but not yet used for cyberattacks. In addition to finding many openly available malicious assets, they found that about 10% of forum threads can be considered malicious, while the remaining threads were benign.

It took one month after software producers announced a vulnerability until cybercriminals were selling an exploit in the darknet market [136]. By comparison, it took six months until a security firm identified a malware that exploited this vulnerability. Thus, the development and exploitation of ICT is swift. Nunes et al. [136] proposed a system for CTI to identify emerging cyberthreats, which include information on newly developed malware and exploits that have not yet been deployed in cyberattacks. Their system considers the binary classification of identifying relevant products and topics relating to malicious hacking. They mentioned the challenge that misspellings, word variations and unnecessary text can produce noise for the classifier.

Summary of natural language processing research

Users' reputation scores generally indicate their activeness and quality content posting in underground forums; however, the process in which actors gain or lose reputation is subjective. Thus, the problem with equating a user's reputation to prominence is that it is based on something arbitrary. For example, administrators assign themselves reputation points arbitrarily [14] or non-technical members accumulate reputation over time by providing less significant CaaS-related materials such as tutorials/guides and copies of malware. Basing the assumption that users' reputations can say something about their proficiency is further complicated by the indication that key actors tend to have a 'poor' reputation [85], most forums lack reputation scores [115] and prominent cybercriminals with a few high-impact posts on the forum do not have to be the most active members.

Keyword-based examinations come with other drawbacks because hacker language contains technical jargon and abbreviations which are not immediately apparent to investigators. For example, keywords require expert and domain knowledge to define and to maintain/extend afterwards. Moreover, keywords may not be generalisable to other forums of similar and dissimilar type and cybercriminals can easily change their vocabulary to avoid law enforcement operatives.

Understandably, related works typically gather data using web crawling [136, 144, 155] because it allows them to collect data from underground forums. Moreover, this mimics the approach real investigators would take to analyse a forum and they will encounter many of the same problems. Firstly, they can only accumulate data with the privilege their crawler is assigned on the forum, which could make some semi-private forum sections unavailable. Secondly, they will undoubtedly come across anti-crawling techniques designed to prevent data collection. We do not encounter similar issues because of our access to the complete database.

1.9 Article summaries and main results

This section summarises the work done for this thesis, and it is primarily arranged in chronological order of publication. The six research articles that constitute this thesis's main contribution are found in Chapters 2 - 7. Furthermore, we composed a real-world casework in Chapter 8 which was the main inspiration of this thesis's dual-perspective from law enforcement and investigative journalism.

Feasibility studies

Network centrality measures have currently been applied on small and structured social networks – such as friend networks from Facebook and Twitter – to identify prominent actors. More notably, the network size of related work and forensic investigations has been similarly small because they rely on first- or second-hand information, generally obtained from police investigation reports. However, today's underground forum size is considerably larger (several hundred thousand individual actors), and each actor is loosely connected with other actors. Thus, network centrality measures must tackle a significantly different problem than the one they were originally designed for. Simply applying centrality measures to a larger problem goes against the intuition given by (i) the 'no free lunch' theorem and (ii) the Daubert standard. The first teaches us that we cannot routinely apply an algorithm to a different dataset and expect the same result. At the same time, the latter requires that we test and understand the scientific procedure in the context of the new problem for it to be admissible in a court of law.

Initially, two feasibility studies were conducted to assess the practicality of applying network centrality measures to larger and loosely structured underground forum datasets. Article I is a novel large-scale study that evaluates centrality measures on unstructured networks from a forensic context. It is also the first study that used leaked information from underground forums to construct an interaction network instead of information provided by other sources, e.g. law enforcement reports or investigations, news articles, web crawling, etc.

Article I After acquiring an underground forum dataset in 2016, we conducted

the first feasibility study in Article I [92] on four undirected centrality measures: degree, closeness, betweenness and eigenvector. We studied how these centrality measures would rank actors in private and public communication forms. We put great effort into manually inspecting messages to understand the top-ranking actors. The top-ranking actors were administrators, moderators and other active users on the forum, while a few others contributed cracked software and leaked user credentials. Interestingly, users ranked higher in eigenvector were selling services of converting or trading between currencies.

There are two issues related to our findings: (i) manually inspecting posts is a time-consuming process for investigators, and (ii) centrality measures identify actors who set up and maintain underground forums. It is an impractical and resource-intensive process to inspect forum posts manually, and investigators may quickly get side-tracked or lose focus. We largely automate this process in Article IV, by utilising topic modelling to better understand the postings from a group of individual actors. The more significant issue, however, is that network centrality measures appear to identify active forum users rather than valuable actors within the CaaS business model. Article V clearly shows the relation between centrality and forum activity. Consequently, law enforcement may waste time and resources investigating lower-skilled cybercriminals who start forum threads more frequently and with minimal impact on the CaaS business model.

Article II The network structure and connectivity between a traditional organised criminal group and an underground forum are very different. According to the 'no free lunch' theorem, good results using network centrality on one type of network structure do not automatically transfer to networks with entirely different forms. Article II [94] compares the identified actors from a traditional hierarchical structure found in organised criminal groups with today's more loosely structured underground networks. Additionally, this article includes edge direction as a natural extension of the first study to investigate whether the results would improve. A traditional and hierarchical organised criminal group is represented with the Enron corporation dataset, while the underground network was the same from Article I. This study aims to target prominent individuals, e.g. actors with higher positions in the organisational hierarchy and important CaaS actors.

The network centrality measures again identify actors with more frequent communication as more central or important, even when considering the edge direction. For example, the top-scoring actor in the hierarchical organisation was responsible for the daily operation, which entails frequent communication between other actors. However, the centrality measures also ranked actors (individuals with high-ranking executive positions) in the middle of the Enron fraud scandal higher. A notable observation is that eigenvector identified actors with lower hierarchical

positions due to how this centrality measure works.

Network centrality measures only find important actors – those who provide cracked software or leak credentials – in the CaaS business model if they are publicly active. However, publicly active cybercriminals do not have to be professional and important cybercriminals. Being active underground forum users is undoubtedly not a criterion to justify focusing law enforcement efforts and resources on them. Moreover, centrality measures are also 'fooled' by mass communications, e.g., automatic 'welcome' messages to new forum members or 'thank you' letters for donations and support. This means that centrality measures give actors with more frequent and diverse communication a higher score, making them appear more important than they are. Consequently, law enforcement using centrality measures can inadvertently accuse lesser criminals of being among the criminal organisations' leaders and falsely use this as reliable evidence in an investigation or court of law.

Underground forum corpus preprocessing

The re-occurring challenge in research Article I and Article II has been the significant effort put into manually inspecting messages to understand the top-ranking actors. This process is slow, resource-demanding and potentially prone to errors. Thus, the process of understanding users' forum posts needs to be primarily automated so law enforcement can quickly understand actors' abilities and importance in the underground forum. Network centrality measures can benefit from this automation and make them more viable for investigators because they provide more information than simply a centrality score. The following research article studies the steps to incorporate NLP to understand human-produced text in underground forums.

Article III Text needs preprocessing before machines can analyse the data; however, previous work was lacking in this regard from two perspectives. Researchers typically use the bare minimum of necessary methods to preprocess and convert text. Furthermore, they lack explanations for the preprocessing methods and detailed construction of the algorithms' input. Thus, given the Daubert standard, their research was neither reproducible nor was their methodology admissible in a court of law in the worst-case scenario.

Article III [95] addresses this gap in the literature. This study proposes a series of rigorous text preprocessing steps on an underground criminal forum. These steps involve standard preprocessing steps found in related literature, such as converting text to lowercase, word normalisation and stop word removal. Additionally, Article III includes other text preprocessing steps to address our dataset's particular

nature, e.g. by removing HTML tags (including attributes), HTML entities, symbols, emojis, and a large amount of email and password dumps. The series of text preprocessing steps are further refined and expanded upon in Article IV.

The foundation for further research is established in Article III by identifying the various document construction approaches for the LDA algorithm. It identifies three contrasting construction approaches and the scenarios when investigators can use them. One approach was not particularly useful for investigators because it produced too many documents – which increased computational cost and time – without improving the results. On the other hand, the two other approaches were appropriate (i) when investigators need an overview of the whole forum and its users and (ii) understanding individual forum users. They are used in Article IV.

A new interdisciplinary methodology

It is critically important to target and disrupt highly proficient cybercriminals to affect the CaaS business model. However, it is a non-trivial task to identify proficient cybercriminals and key actors that would significantly disrupt the criminal network's operations. Network centrality measures are currently the methods literature and law enforcement use to identify actors of 'importance' in criminal networks. Article I and Article II show that network centrality measures rank, e.g. administrators and moderators higher, and thus, consider them to be important actors. Article V further solidifies this observation by showing that centrality measures consider authors of popular forum threads as being important.

Administrating a platform where cybercriminals can connect is a technically minor task compared to the skills necessary to develop hacker tools or find exploits. Similarly, publishing popular forum threads is unreliable for identifying proficient cybercriminals. Article IV and Article V propose a novel approach that uses topic modelling algorithms to remove uninteresting people from the larger underground population. The proposed approach is a systematic and iterative process that removes only uninteresting actors – from a law enforcement perspective. Future research can benefit immensely from this work because it can remove uninteresting actors and extract target-specific features from more proficient cybercriminals. The extracted features then allow for the development of methods that support law enforcement with faster and accurate targeting of CaaS actors that have a more significant disruption on criminal activity.

Article IV A challenge of analysing criminal underground forums is the user-generated content, which has many unique characteristics and frequent use of informal language, e.g. short and incomplete sentences; text that is noisy, sparse and ambiguous; regular use of exaggerations and abbreviations; and repetitions

of words and characters (where about $3/20$ of the data contained repeats). We overcame this challenge by following a series of rigorous preprocessing steps established in Article III, while Article IV [96] improves data quality by normalising further. We achieved this by reducing repeating words and characters to their intended or base form while avoiding erroneous changes to other parts of the text.

A lot of derivationally related words (e.g. 'thaaaanks', 'tyty', etc.) in a corpus not only create unnecessary variations of words with the same meaning (e.g. 'thanks'), but this will also make any NLP model more complex. Article IV proposes the following process to normalise derivationally related words to a common base form: (i) begin by finding all patterns with repeating words and characters (criteria: two or more identical characters or series of characters), (ii) reduce and merge patterns into their shortest form possible, (iii) a domain expert can look over the shortest form list to suggest base form words, and (iv) replace repeating words/characters with the base form. The effectiveness of this approach is high because over 70% of the repeating patterns can be replaced by only suggesting base words for the first thousand repeating patterns. This resulted in an additional normalisation of around 17% of all words in the dataset if we replaced all repeated words/characters.

Article IV also proposes an interdisciplinary approach for analysing criminal networks by combining complementary methods from NLP and SNA in two novel ways. The first approach uses the LDA topic modelling algorithm to identify and remove users with low technical skills from an underground forum population. The second approach combines topic modelling and centrality measures to create a sorted list of central actors and then infer their role in the underground forum.

Low-skilled actors would be uninteresting to focus any law enforcement resources on during an investigation. Article IV uses the LDA algorithm to create few and generic topics and assigns users to particular topics. A human analyst selects the topic(s) with the most coherent appreciation words and removes users predominantly writing posts on this topic. The proposed approach removed 79.6% of low-skilled users from an underground forum's population, which allows law enforcement to focus their limited resources more efficiently on a smaller group with proficient actors.

Some network centrality measures can take a long time to calculate massive graphs. In particular, closeness and betweenness centrality have a time complexity of $O(VE)$ and $O(VE + V^2)$ [122], respectively. As described in the previous paragraph, this approach that removes users will significantly reduce the computational cost and time for subsequent analyses. Article IV is an example of analysis time reduction, where we use centrality measures to create a sorted list of potentially interesting actors on time. Then we use topic modelling – with a document construc-

tion from Article III – to understand the user-produced content of central actors. This allowed us to gain an insight into each actor's potential role in the underground forum, where we identified actors with functions such as administrators and reverse engineers. This approach is much more effective and efficient than spending resources on manually inspecting forum posts, which allows investigators to focus their priorities on more prominent cybercriminals more swiftly.

Article V This research differentiates itself from the two earliest studies, which looked at the feasibility of using network centrality measures to study large-scale and loosely connected criminal networks. Article V [97] quantitatively demonstrates a strong relationship between the rank centrality measures assigned to actors and the number of replies those actors receive (when the graph models an interaction network). This result shows that the notion of a 'key actor' must be interpreted within the graph's context, and 'key' only refers to a vertex's position within the graph structure [97]. For example, actively communicating actors are more important in an interaction network and are, therefore, more central in the graph structure. Consequently, network centrality cannot reliably be used as a forensic technique to identify criminal actors who hold key positions in criminal networks.

The article investigates the generalisability of the method presented in Article IV on another dataset from a real-world criminal underground forum. It also extends the method by identifying which combination of topic modelling algorithms can remove users further and the point at which it stops removing users. After two iterations, the method removes a significantly large portion of uninteresting actors from the underground forum datasets. The first iteration removes the majority of users who only post appreciation messages. A total of 77.39% and 79.23% users were identified and removed using the LDA algorithm. On the other hand, the second iteration used the GSDMM algorithm to remove users who primarily post appreciation posts, which resulted in an additional 69.03% and 45.95% reduction. The final result after two iterations is a reduction of 93.00% and 88.77% of the users in the underground forum. The remaining users have a higher potential of being among the minority population.

The two algorithms LDA and GSDMM both use a latent space to uncover topics in a corpus; where GSDMM is a variation of the LDA algorithm. Article V examines the use of BERTopic – which is based on state-of-the-art transformers (e.g. BERT) – for the second iteration. The performance was low when compared to GSDMM, as only 2.6% or 4.5% of additional users could be removed (depending on the document-construction method). Although a transformer-based topic model had poor performance in this instance, it can be improved using supervised learning approaches (e.g. using labelled datasets).

Big data challenges for forensics

Article VI Today's digital forensic investigators are increasingly confronted by enormous amount of data seized in cybercriminal investigations. The digital crime scene is no longer restricted to the immediate area, but rather, it spans many distinct systems, with multiple victims and crossing jurisdictions. Current investigative methods and tools further challenge investigators' capabilities to analyse large and diverse data sources. The reason is that they are preprogrammed with human knowledge and expert opinions, which are designed only to handle minor and structured problems. The rigidity of these tools makes investigators incapable of handling larger quantities and diversity in data.

Explicitly programming methods and tools with human knowledge is time-consuming, but these implementations must also be correct and reliable. However, inflexible preprogramming makes it difficult for tools to accommodate and effectively analyse ever-larger amounts of data. Article VI [162] discusses the vital need for researching new techniques and processing methods for digital forensic investigations, in areas such as data gathering, preprocessing, cleaning, reduction, analysis, interpretation and visualisation. In particular, it suggests possible ways to approach these challenges using computational methods. This thesis demonstrates how computational methods can improve the analysis of vast amounts of data in a forensic investigation. Computational methods can also support forensic investigators in their daily casework by providing a scientific basis and representing human expert knowledge and reasoning.

Finding fraud in an enormous dataset

Casework We were approached by the Norwegian newspaper DN to assist in analysing billions of playbacks from Tidal. DN suspected Tidal had manipulated data, and we were tasked to identify how the manipulation (if any) was done and to what extent. We received 74.1 GB of log files, with 1 590 422 377 log entries. The logs covered 65 days for two distinct periods, and the first thing we did was explore and understand this dataset. The two periods were from 2016.01.21 to 2016.03.03 (43 days) and 2016.04.18 to 2016.05.09 (22 days).

Our casework report [93] demonstrates how to properly document a hypothesis-driven and design of purpose-built analyses to uncover abnormal playbacks in a large amount of data. The report details nine different statistical analysis approaches, where each approach describes: (i) the purpose of the analysis; (ii) the hypothesis being investigated; (iii) the expected result; and (iv) interpretations of both unexpected and expected results. These statistical analyses involve descriptive methods, timeline analysis and other advanced methods for finding suspicious

log entries, such as logical impossibilities (e.g. playing the same or different songs simultaneously) and other repeating log entries. The report describes how we unveiled millions of Tidal users' listening habits and identified over 350 million fraudulent playbacks. The casework report documents reliable, repeatable and verifiable information which can be presented as admissible evidence according to the Daubert standard in a court of law.

1.10 Summary of contributions

Easterby-Smith, Thorpe, and Jackson [50] categorise novel contributions into three branches: (1) *Discovery*, where a new idea or explanation emerges. (2) *Invention*, where a new technique to deal with a kind of problem is developed. (3) *Reflection*, where existing theories, techniques, or group of ideas are re-examined. All three branches are considered to be equally worthwhile as research novelty [112, 142, 159].

> 'Contributing to knowledge means creating new knowledge based on the available knowledge by doing extensive and innovative research [60].'

Forensics is the challenging intersection of application, technology and method [65]. The technology and methodology might have been sufficient when the application in related work was to small-sized criminal networks with only a few actors. However, analysing criminal networks with different characteristics (e.g. network type and size) changes the application, which requires re-validation of the methodology and technology. The Daubert standard (discussed in Subsection 1.7.1) already points to this by saying we need the known error rates, including knowing under what conditions a particular method works or does not. This thesis reviews the existing network analysis methodology used to find key criminal actors in criminal investigations to work towards procedural accuracy and provide a fair trial.

This section presents new knowledge regarding reflection-type research by scientifically establishing the context and improving our understanding of using network centrality measures in forensic investigations. We find the existing methodology is insufficient to ensure accuracy and explainability in investigations and, therefore, we extend the current methods, which falls under the invention-type of research. This methodology contributes to safeguarding procedural accuracy, the right to a fair trial and human rights.

The development of advanced computational methods is fundamental for analysing and identifying prominent cybercriminals in the CaaS business model. However, this development must emanate from systematic data analysis workflows and

methodologies to make them appropriate in forensic contexts and admissible in a court of law. So far, there is no validation done on the network analysis methodologies used in forensic tools such as IBM i2 Analyst's Notebook and Maltego. This thesis is a significant step toward establishing procedures for validating network analysis methodologies. The studies described in this thesis solve the increased need for reliable methods to identify proficient cybercriminals in underground forums by providing a methodological foundation for investigations. This section outlines our knowledge contributions within the criminal network analysis field by first answering the specific RQs. We explain the real-world casework and its contributions before answering the general RQ, which lead to this work.

Research question 1

> RQ 1: Through which analytical approach can investigators acquire leads on high-impact cybercriminals when the only available data are from underground forums?

Related work limitations Researchers [14, 74, 143] see the positive correlation between forum posting behaviour and reputation, which has developed the understanding that longer-tenured users (with more frequent posting) belong to the group of proficient cybercriminals. Indeed, senior users have spent time contributing with great posts and, thus, building a reputation among their peers. However, there are three caveats when only considering reputation or seniority.

Firstly, administrators can quickly increase or decrease forum reputation for members they like or dislike, so reputation is unreliable and not a true reflection of a user's prestige. Secondly, forum users try to publish as many forum threads as they believe will be helpful for the community and increase their reputation. This results in threads with varying degrees of quality, many of which are of no relevance to investigations. Assuming that a user with a high reputation also produces forum threads of high quality implicitly give the same intrinsic value to all of their posts regardless of their content. Thus, a forum thread about a tutorial/guide will be equally valued as a thread about a new exploit or malware/hacker tool. The latter is more interesting from a law enforcement and security perspective, while other posts can be more similar to noise. Thirdly, professional cybercriminals who are new to the forum will be ignored by reputation-based analysis because they have not had the time to increase their reputation by posting many forum threads.

Very few underground forums provide a reputation-based system [115]. Thus, related literature uses other methods – such as network centrality measures – to find senior users and important criminal actors. The network centrality measures would superficially tick some approval boxes in the Daubert standard because they

are well-established, peer-reviewed, and used to identify key actors. However, the current known error rates were established in a different application area, with minor test data from prior research (see Table 1.5). Before the research collected in this thesis, there was a lack of a scientific basis to use network centrality measures to analyse more extensive criminal underground networks.

An interdisciplinary approach Article I, Article II and Article V validate network centrality measures' reliability and reproducibility [171], i.e. ensuring they follow scientific methodology and produce accurate results. These research articles also contribute to our current knowledge of network centrality measures, in particular when the underlying graph is an interaction network. The articles unequivocally show that centrality measures mainly identify actors who receive more attention and the negative consequences when applied in a forensic context. The consequences negatively affect procedural accuracy because investigators can accuse lesser criminals of being key/central criminal actors, such as leaders. Moreover, the worst-case scenario violates a criminal's right to a fair trial because the algorithms are not used in the correct application. The contribution of these articles is discussed further in RQ 3.

The shortcoming of using variables such as seniority, tenure, and reputation is that they are biased towards a particular type of underground forum user. Those users are frequent posters, which also means they post more noise. Instead of using imprecise centrality measures or looking at forum posting behaviour to find professional cybercriminals, we propose in Article IV and Article V the use of NLP topic modelling to semi-automate the evaluation of post content on a large scale. The topic modelling algorithms LDA and GSDMM are used to effectively categorise six million posts to both individualise actors and filter out actors who mainly write appreciation posts to exclude them from future analyses.

This thesis's interdisciplinary research combines approaches from the scientific disciplines of SNA and NLP, so they are better suited to the problem at hand. Network centrality measures' main advantage is differentiating between central and peripheral actors. However, investigators cannot evaluate a criminal based on a centrality score, making it a limiting factor when using centrality measures. Article IV demonstrates how investigators can use centrality measures to create a sorted list over central actors and then use the LDA algorithm to understand each actor better. LDA generates topic clusters that human investigators can interpret and infer actors' roles in the forum, such as administrator, exploit developer, reverse engineer, etc. Therefore, investigators can use their resources more effectively by selectively targeting those actors that would be of more interest to them and cause more extensive disruption to the criminal underground forum.

Another advantage of clustering users' posts into topics is filtering out low-skilled actors from the underground population. We can exclude 79.6% of users in Article IV from further research by knowing the content they produce. Article V extends this methodology further to remove 93.00% and 88.77% from two independent criminal underground forums. We can exclude these users because they only/-mainly post appreciation messages, which is noise when the goal is to identify professional cybercriminals. Article V has an additional and subtle contribution, as it is also the first study to delineate the minority and majority population quantitatively. Further analyses can additionally improve this outline by analysing the remaining actors that have not been excluded in our studies.

Previous work sometimes relied on forum reputation to distinguish between proficient and inept users. The limitations of such an approach are that new proficient users are ignored, and it gives all forum posts the same intrinsic value. The proposed approach in Article IV and Article V will not have the same limitations because it only removes users who have either only or mainly written appreciation posts. Thus, new proficient underground forum users – with fewer posts – are left in the minority group as long as they produce some other type of text other than appreciation posts, which is very likely. Furthermore, clustering posts into topics allows investigators to selectively keep posts with more value to an investigation, such as new malware/threats, exploits and vulnerabilities, etc. Finally, very few underground forums have reputation-based systems [115], which make this approach very useful because it does not assume the existence of reputations to work.

Algorithm reliability for forensics The acquisition of leaked datasets of criminal underground forums from May 2016 and August 2019 benefited our research. The use of complete underground forum datasets not only put us in a better position to evaluate and validate the use of centrality measures in the forensic field, but it is also unique from a related work's perspective. No other related work had the same opportunity to study entire criminal networks and check their findings against the ground truth.

As previously mentioned in this thesis, researchers often analyse networks from police investigations that are relatively small (see Table 1.5). Thus, other researchers have been unable to uncover and describe limitations of network centrality measures because they analysed incomplete criminal networks. Forensic investigators look to related work to see how they should use the methods in commercially available forensic tools to find key actors in criminal networks. We find that investigators may draw false conclusions if they believe network centrality measures will unequivocally identify the most important or key actors in a network.

We validate how centrality measures would hold up when analysing different net-

work organisational structures (i.e. a traditional hierarchical organisation and a loosely connected network) in Article I and Article II. Our research articles (in addition to Article V) reveal how centrality measure's interpretation of 'important' must be in the context of the analysed graph; i.e. centrality measures quantify actors' positions within the *network structure*. For example, centrality measures identify actors who receive more replies to forum threads because our graph is modelled after the forum communication. Thus, centrality measures do not automatically identify important actors in the CaaS business model but rather actors with more communications.

- This thesis combines NLP and SNA into an interdisciplinary approach that allow investigators to understand different aspects of underground forum communication from both a high-level and individual perspective.
- Centrality measures provide law enforcement with a list that prioritises and focuses their investigation on central cybercriminals.
- Topic modelling algorithms complement centrality measures by inferring users' role in underground forums by inspecting the content they communicate to distinguish administrators from, e.g. malware developers.

Research question 2

> RQ 2: How can NLP be applied to identify key actors who talk about relevant topics efficiently and effectively?

Digital forensic investigators stand at a paradigm shift, from a copy and process everything to a more refined and targeted approach. It is ineffective and unfeasible to manually inspect hundreds of thousands of forum users and posts, so investigators must seek to automate their procedures. However, this automation must be well-documented to meet the legal requirements related to accuracy, reliability, explainability, and so forth [171]. Our research gives investigators an in-depth understanding of how to use the complementary methods of NLP and SNA to quickly ascertain the value of actors who can be potential targets during an investigation. This allows investigators to spend time and resources on those individuals with the highest probability of disrupting the CaaS business model.

'Data analysis workflows in many scientific domains have become increasingly complex and flexible [78].' Botvinik-Nezer et al. [26] looked into the complex workflow of data (pre)processing and found that researchers come to different conclusions because they are crunching data differently. Undocumented or different

workflows are a huge problem for procedural accuracy, where courtrooms rely on forensically sound and validated methods to ensure a fair trial. This highlights the need for validating and documenting forensic science, methodologies and workflows. Our studies are a significant contribution in this regard. In particular, our research documents everything in detail to ensure transparency, reproducibly and reliability and evidentiary admissibility in a court of law. Proper data preprocessing will also bring other benefits such as better performance, faster processing times, error and noise reduction, and improved accuracy.

Forensic text analysis A vital goal in forensics is to keep the evidence in its original form. However, this is not always possible, e.g. in the case of DNA evidence, where investigators must destroy a sample to extract meaningful information, which destroys the ability for re-analysis. Similarly, data would be 'destroyed' because it must be preprocessed to create a topic model (i.e. data is no longer in its original form). Article III sets out to explore and document the text preprocessing steps to normalise a large corpus and prepare the data for topic modelling in a forensically sound way. This article demonstrates how to preprocess text from a forensic perspective, which involves the minimum number of preprocessing steps necessary to produce a suitable topic model while avoiding changing the original text's meaning.

Under the Daubert standard, the proponent of a technique must demonstrate that it rests on adequate validation. Analysts can combine a collection of texts in many different ways to produce a document which is analysed by the LDA algorithm. Each combination can produce a different result and have a different interpretation. Related work lacks exploring the various ways to combine or structure text into documents and how the combinations affect the interpretation of the result. It is essential to understand the effect of document construction and algorithms to fulfil the criteria in the Daubert standard. Article III identifies three document construction approaches, where only two of them are effective and efficient. The two document construction approaches found in Article III are applied in future research Article IV and Article V, where we demonstrate how investigators can map topics and cybercriminals to determine which individuals are of interest in an investigation.

Text analysis challenges The informal and spoken language frequently written by underground forum users are significant challenges for the NLP field. For example, underground forum text comprises many repetitions, incomplete sentences, incorrect sentence structures, slang expressions, domain-specific language, repeating words/characters to add emphasis, and short messages. Furthermore, many users are non-native English speakers, resulting in misspelt words, alternative and incorrect use of words, etc. These variations in text add randomness and

intensify the problem that algorithms must learn from, making data preprocessing a necessary step. Related work typically does not address any of the challenges from the informal text. They mostly use the bare minimum to preprocess the text, such as converting to lowercase, removing special characters/symbols, lemmatisation/stemming and removing stop words. Importantly, they lack detailed documentation of their preprocessing steps, which considerably reduces their approach's testability and, thus, admissibility in a court of law.

Our research in Article IV documents the text preprocessing steps in sequence and detail for reproducibility. This article includes standard text preprocessing steps and other preprocessing to account for the specific nature of our data. Importantly, this article proposes an approach to tackle the challenge of repeating words/characters. This approach (i) begins by finding all patterns with repeating words and characters, (ii) reduces and merges patterns into their shortest form possible, and (iii) a domain expert can look over the shortest form list to suggest words that would replace all the repeating words/characters. The effectiveness of this approach is high because over 70% of the repeating patterns can be replaced by only suggesting new words for the first thousand repeating patterns. This approach further normalises around 17% of all words in the dataset if we replaced all of the first thousand repeated patterns.

- Law enforcement must follow and document a series of rigorous preprocessing steps to reduce the noise in a text (to make ML algorithms more efficient) while keeping the text close to its original form. This documentation contributes to a fair trial because the investigative process is transparent and allows third parties to scrutinise the process and reproduce the result.
- NLP can efficiently provide an overview of a large underground forum corpus and identify relevant topics for an investigation.
- Constructing the appropriate input (i.e. documents) for NLP algorithms help examiners to identify key actors effectively.

Research question 3

> RQ 3: How can we model interactions between actors in the criminal network and identify professional cybercriminals in the model?

Modelling criminal communication The goal with modelling the forum interactions is to identify interesting actors who law enforcement can target during criminal investigations. Procedural accuracy also relies on correctly identifying

actors. Research Article I and Article II demonstrates how graphs can model the public communication between forum users and, thus, the interactions between them. Article I is the first study of its kind on a large-scale criminal network (see Table 1.5 in Subsection 1.8.1). Moreover, these research articles show how SNA's network centrality measures identify central actors in two distinct criminal network structures: a traditional hierarchical network and a loosely connected network. Article I and Article II are the first studies to indicate that high-ranking individuals (identified by centrality measures) have fewer essential contributions to the community, e.g. by moderating the forum. This finding is further solidified in Article V, which quantitatively shows that thread starters who receive more replies (i.e. they start popular threads) are those actors whom centrality measures identify as more important.

Although network centrality measures can cut through noisy data, they reveal users who are administrators, moderators or other talkative criminal actors. Removing these actors would have a low impact on any illegal operations, compared to removing key CaaS actors such as reverse engineers, malware/exploit developers, etc. Thus, the main benefit of centrality measures is that examiners receive a sorted list with actors to investigate. The sorted list is an effective way for investigators to prioritise efforts on central actors (i.e. central in the graph structure), but using centrality measures alone is insufficient to identify more prominent cybercriminals, as demonstrated in Article I, Article II, and Article V. However, combining multiple disciplines as in Article IV and Article V gives law enforcement a more detailed knowledge of which actors to prioritise their resources on.

- Graphs can model the interactions between actors in a criminal network, while centrality measures identify actors of potential interest in networks.
- Centrality measures provide law enforcement with the ability to prioritise targets, but they are limited because they cannot explain why individuals receive their centrality scores. Topic modelling algorithms can fill this knowledge gap and explain to investigators what type of CaaS-role actors occupy.

Real-world casework contributions

Similarly to forensic investigators, investigative journalists also discover the truth by collecting admissible information and holding it up to the court's scrutiny. Investigative journalists' work 'is distinct from apparently similar work done by police, lawyers, auditors, and regulatory bodies in that it is not limited as to target, not legally founded and closely connected to publicity [41].' In other words, investigative journalists use the same forensically sound methods as law enforcement, but the same legal constraints do not bind them.

Reliable one-time methods ICT is in constant development, and forensic investigators cannot always rely on off-the-shelf algorithms or tools to solve every problem. Additionally, there will not always be peer-reviewed documentation to support a methodology or approach to a problem. Investigators must be agile and adapt to the scenario they are analysing while always using a methodology that they can defend in a court of law. Thus, ensuring validity, reliability, transparency and reproducibility comes up when dealing with one-time and purpose-built methods. In these scenarios, investigators must have a systematic approach to understanding the data and developed methods and thorough documentation to support it in a court of law.

Hiding manipulated data Besides our work on analysing criminal networks using NLP and SNA, we have another significant contribution in a study related to forensics. Our casework report demonstrates how to systematically approach a problem related to forensics and document the sequential steps to arrive at a conclusion. This report recorded unveiling millions of Tidal users' listening habits and identifying over 350 million fraudulent playbacks. Our report received worldwide attention after it was first published together with the news article 'Strømmekuppet' [176, 187]. The report has since helped the investigative journalists win the SKUP award [83] and Impala award [82] in 2019.

Thorough documentation is vital so that the investigation process is comprehensive and can endure scrutiny by the court and other third parties. Our casework report documents all the purpose-built methods we applied and link it to the evidence we gathered to ensure its results' credibility and accuracy. Furthermore, it considers and rules out underlying factors that may have affected the data. This report provides quality assurance in the investigation, making it admissible in a court of law according to the Daubert standard.

Prepare for unforeseen consequences Our casework report and the related news article had severe ramifications for Tidal because the Norwegian collection society TONO and three other entities filed an official police complaint against Tidal shortly after their release [45, 108, 116]. In January 2019, Økokrim announced that it had started investigating this as a criminal case [209]. However, Økokrim investigators were interrupted in their search because Tidal claimed their search violated another country's territory [175]. After all, the data was stored on servers outside Norway's borders [135]. Økokrim's right to access the servers was handled by the court system when Tidal finally appealed to the supreme court.

This casework had inadvertently led to answering a fundamental question about what the Norwegian police can collect during a criminal investigation. Law enforcement is dependent on collecting evidence to prove a defendant's guilt beyond

a reasonable doubt. In a worst-case scenario, Økokrim would have to throw away large amounts of evidence in many criminal cases if Tidal had won in the supreme court. This outcome would have made it very difficult for law enforcement to investigate crimes because criminals would easily avoid prosecutions by storing their incriminating data on servers outside Norwegian police jurisdiction. Fortunately, the court ruled in Økokrim's favour [71, 149] and said they were entitled to seize documents stored in other countries.

General research question

We started this section by answering the specific RQs and explained our other significant contributions in real-world casework. We finish this section by explaining how our research has contributed to the general RQ:

> What valuable information can be extracted from the relationship between underground forum communication patterns and post content to identify professional cybercriminals?

Only a minority of technically skilled individuals drives the CaaS business model, facilitating a wide range of cybercriminal activities. Law enforcement's challenge is distinguishing a few proficient cybercriminals from the thousands, other users with varying skills. We expect to find an approach to differentiate them in how they communicate, both in the content produced and with whom they communicate.

Our research in Article I, Article II and Article V systematically addressed the scientific basis of applying network centrality measures to underground forum datasets. We found a previously unidentified weakness where centrality measures select individuals with higher communication frequency as more 'important'. Thus, forensic investigators cannot use network centrality measures to identify proficient cybercriminals reliably, unless they are also very publicly active and receive the most attention from their peers. However, centrality measures' benefit is their ability to create sorted lists of central actors investigators can consider, according to a scientific metric. We continued to investigate how we could join this knowledge from communication patterns with text content.

Our research focuses mainly on preserving the critical aspects of reliability, repeatability, and verifiable results of scientific methodologies according to the Daubert criteria. Therefore, transparent data preprocessing is the centre of attention in all of our research articles. Article III and Article IV are notable contributions for how researchers and investigators can further improve their text preprocessing in a forensically sound way, without too many changes to the original data. Our research culminates in Article IV and Article V.

Article IV demonstrates an interdisciplinary approach to investigating underground forums. It starts by creating a sorted list of central actors using network centrality measures before making several topic models from the post content of each actor. Investigators can infer the user's role in the underground forum by inspecting the keywords from the topic models, which allow an investigator to distinguish, e.g. administrators from reverse engineers (and other types of actors). The proposed method presented in this thesis will allow a more targeted approach, so law enforcement can effectively and efficiently focus their resources and efforts on more technically skilled actors such as reverse engineers instead of arresting forum administrators. This targeted approach can break the current modus operandi where law enforcement shut down underground forums, and the remaining members move to alternative marketplaces. Law enforcement can, instead, focus on CaaS providers who would have a more significant effect of disrupting the increasing cybercriminal activities all over the world.

Article V demonstrates an approach where researchers and investigators can divide all underground forum users into two distinct groups: the minority and the majority population, as discussed in Section 1.1. This approach uses two topic modelling algorithms to categorise user-produced text into a few broad topics; then, a human analyst can select the topics that best indicate the majority population. The specifically selected topics must contain semantically coherent appreciation words such as 'thank', 'you', 'great', 'work', etc. Article V's result shows that 93% and 88.77% of the users in an underground forum can be considered as being part of the majority population. At the same time, the remaining 7% and 11.23% are more likely to belong in the minority population. The significant bulk exclusion of users – from the majority population – not only allows future analysis to be more effective, but it also allows researchers to concentrate their efforts on finding specific features for professional cybercriminals or separate users into specific CaaS groups/roles.

There is significant potential when investigators can use automated or computational methods to reduce the amount of relevant data from a criminal investigation. Our method presented in this thesis reduced the number of actors by around 90%. Consider the example from Subsection 1.7.1 where Økokrim have criminal cases with over 53 TB of data. Although this issue is not directly analogous to the problem addressed in this thesis, the data reduction by 90% (i.e. 5.3 TB) is a huge time saver and will allow investigators to focus on relevant data for the criminal investigation.

1.11 General considerations

1.11.1 Theoretical implications

The digital era produces too much data that can overwhelm even the most well-equipped and well-resourced investigator. Furthermore, the reality of big data goes beyond examining criminal networks and analysing thousands of underground forum users. Investigators from every forensic domain such as property, inchoate, statutory, financial, cyber and terrorism encounter large amounts of structured and unstructured data during their workflow. Human investigators will inevitably employ off-the-shelf algorithms or computational methods to speed up examinations when they find it impossible to keep up with their work.

Our research has shown that off-the-shelf algorithms and methods do not necessarily produce desirable results on the acquired data, even though state-of-the-art literature has indicated they would work in this way. Thus, investigators cannot blindly trust or depend on off-the-shelf algorithms and computational methods; because they may not have been shown to work on their type of data or dataset, so they can, therefore, produce inaccurate or undesirable results. Moreover, algorithms are not the only thing that matter, but every step in the workflow. Each step must be extensively tested, from the raw data, preprocessing and algorithms to tools and methods.

1.11.2 Practical recommendations

Data is the most crucial in the workflow process; however, it may be challenging to acquire. This is particularly true of law enforcement data, which they cannot share due to its sensitive nature and various laws. We recommend that investigators and researchers look for other and similar data sources. For example, we acquired leaked hacker datasets similar to dark web underground forums to conduct our research. Nevertheless, investigators must examine the possible limitations of using datasets from other sources and clarify its impact on their analysis.

The acquired data need to be preprocessed before being analysed. There are many best practices for data preprocessing; however, the essential part is to know and understand the data. A deeper understanding of the data allows investigators to think outside the box and find potential combinations or alternative processing methods. Never underestimate data preprocessing as there are no one-size-fits-all solutions, so every data preprocessing approach must be tested and verified.

Analysts must help algorithms at every step to achieve their goals. Sometimes, however, existing algorithms do not adequately work in certain situations, and they need to be developed on the fly, as demonstrated in our casework. Algorithms

must be well-documented in these scenarios, and the documentation must go beyond just the algorithm. It must also involve the surrounding area, particularly the parameters, input data, and interpretation. Thorough documentation keeps an accurate account of all activities, which will support their case in a court of law.

Finally, law enforcement analysts must be adequately trained to handle big data and computational methods because there is no magical 'push here' button, making machines or algorithms do everything. If law enforcement analysts cannot acquire data proficiency, they should bring data scientists onto their team to ensure they follow proper data analytical procedures.

1.11.3 Recommendations for future work

We identified several directions that can be investigated further.

Dismantling criminal networks

The goal of dismantling criminal networks is to stop their illegal activities, but history has shown that law enforcement operations have a reduced impact before criminal activity resumes as normal. There are certainly some actors that keep the network together more than others. Therefore, it would be advantageous to examine which actors hold the underground forum or CaaS business model together; to focus efforts on criminals with the most impact and disruption when taken down. We suggest a way for researchers to utilise methods similar to (or found in) percolation theory [195] to examine the behaviour of a criminal network when vertices or edges are removed, thereby studying how the criminal network may respond when removing certain actors or relationships.

Countermeasures against identifying key actors

Although this thesis has not directly modelled (a seemingly unlimited number of) defensive strategies or active countermeasures cybercriminals can employ to prevent being identified using our proposed methods, we still consider that proficient criminals try to hide, e.g. by using sparse communication. However, cybercriminals' goal is to profit from their activities, which must involve some risk by being publicly active to get customers for their services or products. Their countermeasures could involve changing the vocabulary or using code words similar to drug traders, which could affect e.g. topic modelling. Code words are associated with specific contexts, so topic modelling algorithms (such as LDA) which cluster topic-related words can pick up those words and still put them into the appropriate clusters.

Graph construction and social network analysis

A limitation of related work is that they assume all forum posts offer the same type of ‘value’ when constructing graphs. For example, they increment graph weights based on the number of posts without considering posts’ contents, as we did in our research. However, it is imprecise to give the same value to releasing a new exploit and a general question for the community. Investigators need to look into approaches to improve the modelling of forum communication through graphs, such as addressing this issue by dynamically assigning a proper value to posts. Edges’ weights most likely must also be normalised to account for the different communication frequencies between actors. An approach such as this will quickly allow investigators to disregard actors with low-value connections since they contribute very little to the network (in the form of the posts they produce). Furthermore, adjusting edge weights will make SNA and network centrality measures more precise when identifying more important actors.

1.12 Bibliography

[1] Ahmed Abbasi, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. Descriptive Analytics: Examining Expert Hackers in Web Forums. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 56–63, The Hague, Netherlands, 2014. IEEE. http://ieeexplore.ieee.org/document/6975554/.

[2] Salim Afra and Reda Alhajj. Integrated framework for criminal network extraction from Web. *Journal of Information Science*, 47(2):206–226, April 2021.

[3] M. Al Fahdi, N.L. Clarke, and S.M. Furnell. Challenges to digital forensics: A survey of researchers & practitioners attitudes and opinions. In *2013 Information Security for South Africa*, pages 1–8, Johannesburg, South Africa, August 2013. IEEE. http://ieeexplore.ieee.org/document/6641058/.

[4] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, 3:42, July 2020.

[5] J.S. Albanese. *Organized Crime in Our Times*. Elsevier Science, 2010. https://books.google.com/books?id=IawY6wzNh_sC.

[6] Joseph Almog. Forensics as a proactive science. *Science & Justice*, 54(5):325–326, September 2014.

[7] Kodi B. Arfer and Jason J. Jones. American Political-Party Affiliation as a Predictor of Usage of an Adultery Website. *Archives of Sexual Behavior*, 48(3):715–723, April 2019. http://link.springer.com/10.1007/s10508-018-1244-1.

[8] Wayne E. Baker and Robert R. Faulkner. The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review*, 58(6):837, December 1993. http://www.jstor.org/stable/2095954?origin=crossref.

[9] Steven James Bartlett. The Species Problem and its Logic: Inescapable Ambiguity and Framework-Relativity. *SSRN Electronic Journal*, 2015. https://www.ssrn.com/abstract=3073801.

[10] H. M. A. van Beek, J. van den Bos, A Boztas, E. J. van Eijk, R Schramp, and M Ugen. Digital forensics as a service: Stepping up the game. *Forensic Science International: Digital Investigation*, 35:301021, 2020. Publisher: Elsevier.

[11] H. M. A. van Beek, E. J. van Eijk, R.B. van Baar, M. Ugen, J.N.C. Bodde, and A.J. Siemelink. Digital forensics as a service: Game on. *Digital Investigation*, 15:20–38, 2015. http://linkinghub.elsevier.com/retrieve/pii/S1742287615000857.

[12] Jacopo Bellasio, Erik Silfversten, Eireann Leverett, Anna Knack, Fiona Quimbre, Emma Louise Blondes, Marina Favaro, and Giacomo Persi Paoli. The Future of Cybercrime in Light of Technology Developments. Technical report, February 2020.

[13] Frank Benford. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4):23, March 1938.

[14] Victor Benjamin and Hsinchun Chen. Securing cyberspace: Identifying key actors in hacker communities. In *2012 IEEE International Conference on Intelligence and Security Informatics*, pages 24–29, Arlington, VA, 2012. IEEE. http://ieeexplore.ieee.org/document/6283296/.

[15] Victor Benjamin and Hsinchun Chen. Developing understanding of hacker language through the use of lexical semantics. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 79–84, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165943/.

[16] Victor Benjamin and Hsinchun Chen. Identifying language groups within multilingual cybercriminal forums. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 205–207, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745471/.

[17] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 85–90, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165944/.

[18] T. Beysolow. *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing*. Apress, 2018. https://books.google.com/books?id=FDBuDwAAQBAJ.

[19] Anand Bihari and Manoj Kumar Pandia. Eigenvector centrality and its application in research professionals' relationship network. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pages 510–514, Greater Noida, India, February 2015. IEEE. http://ieeexplore.ieee.org/document/7154915/.

[20] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009. https://books.google.com/books?id=KGIbfiiP1i4C.

[21] David M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. page 30, 2003.

[22] John M. Bolland. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks*, 10(3):233 – 253, 1988. http://www.sciencedirect.com/science/article/pii/0378873388900147.

[23] Phillip Bonacich. Power and Centrality: A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182, 1987. http://www.jstor.org/stable/2780000.

[24] Stephen P Borgatti. Identifying sets of key players in a network. In *IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No. 03CH37502)*, pages 127–131. IEEE, 2003.

[25] Stephen P Borgatti. *The key player problem*. National Academy of Sciences Press, 2003.

[26] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, and others. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, pages 1–7, 2020. Publisher: Nature Publishing Group.

[27] Anne E. Boustead and Trey Herr. Analyzing the Ethical Implications of Research Using Leaked Data. *PS: Political Science & Politics*, 53(3):505–509, July 2020. https://www.cambridge.org/core/product/identifier/S1049096520000323/type/journal_article.

[28] David A. Bright, Caitlin E. Hughes, and Jenny Chalmers. Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate. *Crime, Law and Social Change*, 57(2):151–176, March 2012. http://link.springer.com/10.1007/s10611-011-9336-z.

[29] Hugo de Burgh. Investigative Journalism. In *21st Century Communication: A Reference Handbook 21st century communication: A reference handbook*, pages 609–617. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 2009. http://sk.sagepub.com/reference/communication/n67.xml.

[30] Lori Cameron. Digital Forensics: 6 Security Challenges | IEEE Computer Society, 2020. https://www.computer.org/publications/tech-news/research/digital-forensics-security-challenges-cybercrime.

[31] Anjelica Cappellino. Daubert vs. Frye: Navigating the Standards of Admissibility for Expert Testimony, July 2018. https://www.expertinstitute.com/resources/insights/daubert-vs-frye-navigating-the-standards-of-admissibility-for-expert-testimony/.

[32] Aparicio Carranza and Casimer DeCusatis. Software Validation and Daubert Standard Compliance of an Open Digital Forensics Model. *Journal of Machine Intelligence and Data Science*, 2021.

[33] Luca Caviglione, Steffen Wendzel, and Wojciech Mazurczyk. The Future of Digital Forensics: Challenges and the Road Ahead. *IEEE Security & Privacy*, 15(6):12–17, November 2017. http://ieeexplore.ieee.org/document/8123473/.

[34] Andrea Cerioli, Lucio Barabesi, Andrea Cerasa, Mario Menegatti, and Domenico Perrotta. Newcomb–Benford law and the detection of frauds in international trade. *Proceedings of the National Academy of Sciences*, 116(1):106–115, January 2019. http://www.pnas.org/lookup/doi/10.1073/pnas.1806617115.

[35] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. page 10, 2009.

[36] Kim-Kwang Raymond Choo. Organised crime groups in cyberspace: a typology. *Trends in Organized Crime*, 11(3):270–295, September 2008. http://link.springer.com/10.1007/s12117-008-9038-9.

[37] Wikipedia contributors. *Frye standard — Wikipedia, The Free Encyclopedia*. September 2020. https://en.wikipedia.org/w/index.php?title=Frye_standard&oldid=977954236.

[38] James A Crowder and John N Carbone. Occam learning through pattern discovery: Computational mechanics in AI systems. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2011.

[39] Cybersecurity Ventures. Cybercrime Report 2017. Technical report, 2017. https://1c7fab3im83f5gqiow2qqs2k-wpengine.netdna-ssl.com/2015-wp/wp-content/uploads/2017/10/2017-Cybercrime-Report.pdf.

[40] William Daubert. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993), 1993.

[41] H. de Burgh and P. Bradshaw. *Investigative Journalism: Context and Practice*. Investigative Journalism: Context and Practice. Routledge, 2000.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[43] Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, 2005.

[44] David M. Douglas. Should researchers use data from security breaches? *Communications of the ACM*, 62(12):22–24, November 2019. https://dl.acm.org/doi/10.1145/3368091.

[45] Øystein Tronsli Drabløs, Bjørn Eckblad, Marcus Husby, and Agnete Klevstrand. Musikeropprør og ny anmeldelse mot Tidal, May 2018. https://www.dn.no/musikk/tidal/stromming/teknologi/musikeroppror-og-ny-anmeldelse-mot-tidal/2-1-346984.

[46] Brian Duignan. What Is the Difference Between Criminal Law and Civil Law?, 2020. https://www.britannica.com/story/what-is-the-difference-between-criminal-law-and-civil-law.

[47] Paul A C Duijn, Victor Kashirin, and Peter M A Sloot. The Relative Ineffectiveness of Criminal Network Disruption. *Scientific Reports*, 4:15, February 2014.

[48] Scott W. Duxbury and Dana L. Haynie. Criminal network security: An agent-based approach to evaluating network resilience. *Criminology*, 57(2):314–342, May 2019. https://onlinelibrary.wiley.com/doi/abs/10.1111/1745-9125.12203.

[49] David Décary-Hétu and Benoit Dupont. The social network of hackers. *Global Crime*, 13(3):160–175, 2012. http://www.tandfonline.com/doi/abs/10.1080/17440572.2012.702523.

[50] Mark Easterby-Smith, R Thorpe, and P Jackson. *Management Research*. SAGE Publications Ltd, London, 2008.

[51] Nael T. Elyezjy and Alaa El-Halees. Investigating Crimes using Text Mining and Network Analysis. *International Journal of Computer Applications*, 126(8):19–25, September 2015.

[52] European Commission. Ethics and data protection, November 2018.

[53] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2014. Technical report, 2014. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf.

[54] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2016. Technical report, 2016. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web_2016.pdf.

[55] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2018. Technical report, 2018.

[56] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2020. Technical report, 2020.

[57] Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, and Lorie Donelle. Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. *PLOS ONE*, 15(8):e0237875, August 2020. https://dx.plos.org/10.1371/journal.pone.0237875.

[58] Margaret G Farrell. Daubert v. merrell dow pharmaceuticals, inc.: epistemilogy and legal process. *Cardozo L. Rev.*, 15:2183, 1993. Publisher: HeinOnline.

[59] Sydney Firmin. Simple is Best: Occam's Razor in Data Science, February 2019. https://community.alteryx.com/t5/Data-Science/Simple-is-Best-Occam-s-Razor-in-Data-Science/ba-p/355159.

[60] Alexis Flynn. What is contribution to the body of knowledge? – Greedhead.net, March 2021.

[61] Tom Foremski. Report: Nation state hackers and cyber criminals are spoofing each other | ZDNet, April 2019. https://www.zdnet.com/article/optiv-report-nation-state-hackers-and-cyber-criminals-are-spoofing-each-other/, urldate = 2020-12-17.

[62] Forensic Focus. Digital Forensics is not just HOW but WHY, July 2012.

[63] Lisa R Fournier. The daubert guidelines: usefulness, utilization, and suggestions for improving quality control. *Journal of Applied Research in Memory and Cognition*, 5(3):308–313, 2016. Publisher: Elsevier.

[64] Henry F Fradella, Lauren O'Neill, and Adam Fogarty. The Impact of Daubert on Forensic Science. *Pepp. L. Rev.*, 31(2):41, 2004.

[65] Katrin Franke and Sargur N. Srihari. Computational Forensics: An Overview. In *Proceedings of the 2Nd International Workshop on Computational Forensics*, IWCF '08, pages 1–10, Berlin, Heidelberg, 2008. Springer-Verlag. http://dx.doi.org/10.1007/978-3-540-85303-9_1.

[66] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, January 1978. https://linkinghub.elsevier.com/retrieve/pii/0378873378900217.

[67] Simson L. Garfinkel. Digital forensics research: The next 10 years. *Digital Investigation*, 7:S64–S73, 2010. http://linkinghub.elsevier.com/retrieve/pii/S1742287610000368.

[68] Adam Geitgey. Natural Language Processing is Fun!, September 2020. https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e.

[69] Zeno Geradts. Digital, big data and computational forensics. *Forensic Sciences Research*, 3(3):179–182, July 2018. https://www.tandfonline.com/doi/full/10.1080/20961790.2018.1500078.

[70] Paramita Ghosh. The Fundamentals of Natural Language Processing and Natural Language Generation, August 2018. https://www.dataversity.net/fundamentals-natural-language-processing-natural-language-generation/.

[71] Henrik Giæver. Nye rettsdokumenter: Økokrim har gitt Tidal status som mistenkt for grovt databedrageri, June 2020. https://www.dn.no/etterbors/nye-rettsdokumenter-okokrim-har-gitt-tidal-status-som-mistenkt-for-grovt-databedrageri/2-1-822585.

[72] Y. Goldberg and G. Hirst. *Neural Network Methods in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2017. https://books.google.com/books?id=Za2zDgAAQBAJ.

[73] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[74] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 13–18, Beijing, China, July 2017. IEEE. http://ieeexplore.ieee.org/document/8004867/.

[75] Maarten Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics., 2020. Version Number: v0.7.0.

[76] Alessandro Guarino. Digital Forensics as a Big Data Challenge. In Helmut Reimer, Norbert Pohlmann, and Wolfgang Schneider, editors, *ISSE 2013 Securing Electronic Business Processes*, pages 197–203. Springer Fachmedien Wiesbaden, Wiesbaden, 2013. http://link.springer.com/10.1007/978-3-658-03371-2_17.

[77] Danielle Haas. Teaching Note: Reporters or cops? CONTACTO and the search for Paul Schaefer, March 2010.

[78] Adam Hadhazy. Complex data workflows contribute to reproducibility crisis, May 2020. https://news.stanford.edu/2020/05/20/complex-data-workflows-contribute-reproducibility-crisis/.

[79] R.A. Hanneman and M. Riddle. *Introduction to Social Network Methods*. University of California, 2005. https://books.google.com/books?id=wAHaygAACAAJ.

[80] Tony Harcup and Deirdre O'Neill. What Is News? Galtung and Ruge revisited. *Journalism Studies*, 2(2):261–280, January 2001. http://www.tandfonline.com/doi/abs/10.1080/14616700118449.

[81] J. S. Hardin, Ghassan Sarkis, and P. C. Urc. Network analysis with the enron email corpus. *Journal of Statistics Education*, 23(2), 2015.

[82] Birk Tjeldflaat Helle. Dagens Næringsliv får internasjonal pris for Tidal-saken, February 2019. https://www.dn.no/medier/tidal/dagens-naringsliv-far-internasjonal-pris-for-tidal-saken/2-1-554289.

[83] Birk Tjeldflaat Helle. Dagens Næringsliv vant Skup-prisen for Tidal-avsløringene, March 2019. https://www.dn.no/medier/skup/tidal/dagens-naringsliv-vant-skup-prisen-for-tidal-avsloringene/2-1-577409.

[84] Thomas J. Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology*, 6(1):891, 2012.

[85] Shin-Ying Huang and Hsinchun Chen. Exploring the online underground marketplaces through topic-based social network and clustering. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 145–150, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745458/.

[86] Mark Lee Hunter, Nils Hanson, Rana Sabbagh, Luuk Sengers, Drew Sullivan, and Pia Thordsen. *Story-Based Inquiry: A manual for investigative journalists*. UNESCO, 2011.

[87] Thomas S. Hyslip. Cybercrime-as-a-Service Operations. In Thomas J. Holt and Adam M. Bossler, editors, *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, pages 815–846. Springer International Publishing, Cham, 2020. http://link.springer.com/10.1007/978-3-319-78440-3_36.

[88] International Consortium of Investigative Journalists. Investigations Archives, 2020. http://www.icij.org/category/investigations/.

[89] Joshua I James and Pavel Gladyshev. Challenges with automation in digital forensic investigations. *arXiv preprint arXiv:1303.4498*, 2013.

[90] Qiang Jipeng, Qian Zhenyu, Li Yun, Yuan Yunhao, and Wu Xindong. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *arXiv:1904.07695 [cs]*, April 2019. arXiv: 1904.07695.

[91] Jan William Johnsen. *Algorithms and Methods for Organised Cybercrime Analysis*. PhD thesis, 2016.

[92] Jan William Johnsen and Katrin Franke. Feasibility Study of Social Network Analysis on Loosely Structured Communication Networks. *Procedia Computer Science*, 108:2388–2392, 2017. http://linkinghub.elsevier.com/retrieve/pii/S1877050917307561.

[93] Jan William Johnsen and Katrin Franke. Digital Forensics Report for Dagens Næringsliv. *Dagens Næringsliv*, page 78, April 2018. https://www.dn.no/staticprojects/special/2018/05/09/0600/dokumentar/strommekuppet/data/documentation/NTNU-rapport_til_publisering.pdf.

[94] Jan William Johnsen and Katrin Franke. Identifying Central Individuals in Organised Criminal Groups and Underground Marketplaces. In *Computational Science – ICCS 2018*, volume 10862, pages 379–386. Springer International Publishing, Cham, 2018. http://link.springer.com/10.1007/978-3-319-93713-7_31.

[95] Jan William Johnsen and Katrin Franke. The impact of preprocessing in natural language for open source intelligence and criminal investigation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4248–4254, Los Angeles, CA, USA, December 2019. IEEE. https://ieeexplore.ieee.org/document/9006006/.

[96] Jan William Johnsen and Katrin Franke. Identifying Proficient Cybercriminals Through Text and Network Analysis. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–7. IEEE, 2020. https://ieeexplore.ieee.org/abstract/document/9280523.

[97] Jan William Johnsen and Katrin Franke. On the feasibility of social network analysis methods for investigating large-scale criminal networks. page 31, 2022.

[98] Terence D. Kadlec and Bruce A. Barnes. Why do Daubert and Frye standards matter to expert witnesses? Whitepaper, Envista Forensics, July 2019.

[99] Peter Klerks. The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections*, 24(3):53–65, 2001.

[100] I. Kononenko and M. Kukar. *Machine Learning and Data Mining*. Elsevier Science, 2007. https://books.google.com/books?id=NUikAgAAQBAJ.

[101] Valdis Krebs. Uncloaking Terrorist Networks. volume 7, page 4. First Monday, 2002. http://journals.uic.edu/ojs/index.php/fm/article/view/941.

[102] Vy Le. Organised Crime Typologies: Structure, Activities and Conditions. *International Journal of Criminology and Sociology*, 1:121–131, 2012.

[103] Paul D Leedy and Jeanne Ellis Ormrod. *Practical research: Planning and design*. ERIC, 2019.

[104] Weifeng Li and Hsinchun Chen. Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 64–67, The Hague, Netherlands, 2014. IEEE. http://ieeexplore.ieee.org/document/6975555/.

[105] Weifeng Li, Hsinchun Chen, and Jay F. Nunamaker. Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System. *Journal of Management Information Systems*, 33(4):1059–1086, 2016. https://www.tandfonline.com/doi/full/10.1080/07421222.2016.1267528.

[106] Weifeng Li, Junming Yin, and Hsinchun Chen. Identifying High Quality Carding Services in Underground Economy using Nonparametric Supervised Topic Model. page 10, 2016.

[107] Matthew J Lindquist and Yves Zenou. Crime and networks: ten policy lessons. *Oxford Review of Economic Policy*, 35(4):746–771, December 2019.

[108] Tord Litleskare. Nok en anmeldelse mot Tidal, May 2018. https://gaffa.no/musikk/nyheter/128955/nok-en-anmeldelse-mot-tidal/.

[109] Chi-Yu Liu, Zheng Liu, Tao Li, and Bin Xia. Topic Modeling for Noisy Short Texts with Multiple Relations. In *SEKE*, pages 610–609, 2018.

[110] Xiaodong Liu, Eleonora Patacchini, Yves Zenou, and Lung-Fei Lee. Criminal Networks: Who is the Key Player? *SSRN Electronic Journal*, 2012. http://www.ssrn.com/abstract=2089267.

[111] Yong Lu, Xin Luo, Michael Polgar, and Yuanyuan Cao. Social Network Analysis of a Criminal Hacker Community. *Journal of Computer Information Systems*, page 12, 2010.

[112] Priti Ranjan Majhi. *Introduction to Research Methodology (Theory and Project Report)*. Himalaya Publishing House Pvt. Ltd, 2019.

[113] Derek Manky. Cybercrime as a service: a very modern business. *Computer Fraud & Security*, 2013(6):9–13, June 2013. https://linkinghub.elsevier.com/retrieve/pii/S1361372313700538.

[114] Ericsson Marin, Ahmad Diab, and Paulo Shakarian. Product offerings in malicious hacker markets. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 187–189, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745465/.

[115] Ericsson Marin, Jana Shakarian, and Paulo Shakarian. Mining Key-Hackers on Darkweb Forums. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 73–80, South Padre Island, TX, 2018. IEEE. https://ieeexplore.ieee.org/document/8367642/.

[116] Willy Martinsen. TONO anmelder Tidal til Økokrim, May 2018. https://www.tono.no/tono-anmelder-tidal-okokrim/.

[117] Giovanni Mastrobuoni and Eleonora Patacchini. Organized Crime Networks: an Application of Network Analysis Techniques to the American Mafia. *Review of Network Economics*, 11(3), January 2012.

[118] Jocelyn Mazarura and Alta de Waal. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6, Stellenbosch, South Africa, November 2016. IEEE.

[119] McAfee. Net Losses: Estimating the Global Cost of Cybercrime. Technical report, 2014. https://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/attachments/140609_McAfee_PDF.pdf.

[120] Rodney McKemmish. When is Digital Evidence Forensically Sound? In Indrajit Ray and Sujeet Shenoi, editors, *Advances in Digital Forensics IV*, volume 285, pages 3–15. Springer US, Boston, MA, 2008. ISSN: 1571-5736 Series Title: IFIP — The International Federation for Information Processing.

[121] Walter R Mebane. Election Forensics: Vote Counts and Benford's Law. *Summer Meeting of the Political Methodology Society*, page 51, July 2006.

[122] Natarajan Meghanathan. Centrality Metrics, January 2016.

[123] Bisharat Rasool Memon. Identifying Important Nodes in Weighted Covert Networks Using Generalized Centrality Measures. In *2012 European Intelligence and Security Informatics Conference*, pages 131–140, Odense, Denmark, August 2012. IEEE. http://ieeexplore.ieee.org/document/6298823/.

[124] Merriam-Webster.com Dictionary. Definition of criminology, November 2021.

[125] Tom M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, 1997.

[126] Reza Montasari. A standardised data acquisition process model for digital forensic investigations. *International Journal of Information and Computer Security*, 9(3):229–249, 2017.

[127] Carlo Morselli. *Inside Criminal Networks*, volume 8 of *Studies of Organized Crime*. Springer New York, New York, NY, 2009. http://link.springer.com/10.1007/978-0-387-09526-4.

[128] Carlo Morselli. Assessing Vulnerable and Strategic Positions in a Criminal Network. *Journal of Contemporary Criminal Justice*, 26(4):382–392, November 2010. http://journals.sagepub.com/doi/10.1177/1043986210377105.

[129] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11*, page 71, Berlin, Germany, 2011. ACM Press. http://dl.acm.org/citation.cfm?doid=2068816.2068824.

[130] Andreas Mueller. Don't cite the No Free Lunch Theorem, July 2019. https://peekaboo-vision.blogspot.com/2019/07/dont-cite-no-free-lunch-theorem.html.

[131] Mangai Natarajan. Understanding the Structure of a Large Heroin Distribution Network: A Quantitative Analysis of Qualitative Data. *J Quant Criminol*, page 22, 2006.

[132] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford ; New York, 2010. OCLC: ocn456837194.

[133] Mark J Nigrini and Linda J Mittermaier. The use of Benford's law as an aid in analytical procedures. *Auditing*, 16(2):52, 1997. Publisher: American Accounting Association.

[134] M.J. Nigrini and J.T. Wells. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley Corporate F&A. Wiley, 2012. https://books.google.com/books?id=Bh5Vr_I1NZoC.

[135] Norsk Telegrambyrå. Økokrim stoppet midt i Tidal-ransaking, February 2019. https://e24.no/i/P3b1X7.

[136] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 7–12, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745435/.

[137] Patrick W Nutter. Machine Learning Evidence: Admissibility and Weight. 21:40, 2019.

[138] Optiv, Optiv's Global Threat Intelligence Center, IntSights, and Carbon Black. 2019 Cyber Threat Intelligence Estimate, 2019. https://www.dhs.gov/sites/default/files/publications/ia/ia_geopolitical-impact-cyber-threats-nation-state-actors.pdf.

[139] Daniel Ortiz-Arroyo. Discovering Sets of Key Players in Social Networks. In Ajith Abraham, Aboul-Ella Hassanien, and Vaclav Sná¿el, editors, *Computational Social Network Analysis: Trends, Tools and Research Advances*, pages 27–47. Springer London, London, 2010. https://doi.org/10.1007/978-1-84882-229-0_2.

[140] Oxford Online Dictionary. Definition: effective, 2020. https://www.oxfordlearnersdictionaries.com/definition/english/effective.

[141] Oxford Online Dictionary. Definition: efficiency, 2020. https://www.oxfordlearnersdictionaries.com/definition/english/efficiency.

[142] Roshan Panditharathna. How can you verify if your research work is novel or not?, January 2020.

[143] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. In Michael Bailey, Thorsten Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis, editors, *Research in Attacks, Intrusions, and Defenses*, pages 207–227, Cham, 2018. Springer International Publishing.

[144] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 1845–1854, Lyon, France, 2018. ACM Press. http://dl.acm.org/citation.cfm?doid=3178876.3186178.

[145] Priya S. Patil and A. S. Kapse. Survey on Different Phases of Digital Forensics Investigation Models. *International Journal of Innovative Research in Computer and Communication Engineering*, 03(03):1529–1534, April 2015. http://ijircce.com/upload/2015/march/18_Survey.pdf.

[146] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. A Social Network Analysis and Comparison of Six Dark Web Forums. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 484–493, Genoa, Italy, September 2020. IEEE. https://ieeexplore.ieee.org/document/9229679/.

[147] C. Prell. *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2012. https://books.google.com/books?id=p4iTo566nAMC.

[148] Darren Quick and Kim-Kwang Raymond Choo. Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digital Investigation*, 11(4):273–294, 2014. http://linkinghub.elsevier.com/retrieve/pii/S1742287614001066.

[149] Rett24. Økokrim får ta beslag i Tidal-dokumenter, June 2020. https://rett24.no/articles/okokrim-far-ta-beslag-i-tidal-dokumenter.

[150] Olivier Ribaux, Simon J. Walsh, and Pierre Margot. The contribution of forensic science to crime analysis and investigation: Forensic intelligence. *Forensic Science International*, 156(2-3):171–181, January 2006.

[151] Golden G. Richard and Vassil Roussev. Next-generation digital forensics. *Communications of the ACM*, 49(2):76–80, February 2006. https://dl.acm.org/doi/10.1145/1113034.1113074.

[152] Kenneth H Rosen and Kamala Krithivasan. *Discrete mathematics and its applications: with combinatorics and graph theory*. Tata McGraw-Hill Education, 2012.

[153] Raj Samani and Francois Paget. Cybercrime exposed: Cybercrime-as-a-service. *McAfee White Paper*, 16, 2013. https://scadahacker.com/library/Documents/Threat_Intelligence/McAfee%20-%20Cybercrime%20Exposed%20-%20Cybercrime%20as%20a%20Service.pdf.

[154] Sagar Samtani and Hsinchun Chen. Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 319–321, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745500/.

[155] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. Exploring hacker assets in underground forums. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 31–36, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165935/.

[156] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker. Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, 34(4):1023–1053, 2017. https://www.tandfonline.com/doi/full/10.1080/07421222.2017.1394049.

[157] S.C. Schroeder. How to be a Digital Forensic Expert Witness. In *First International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'05)*, pages 69–88, Taipei, Taiwan, 2005. IEEE Comput. Soc.

[158] Daniel M Schwartz and Tony DA Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009. Publisher: Springer.

[159] William Sebunje. Research Techniques, September 2015.

[160] Dirk Semmann, Hans-Jürgen Krambeck, and Manfred Milinski. Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology*, 57(6):611–616, 2005. Publisher: Springer.

[161] Andrii Shalaginov. Advancing Neuro-Fuzzy Algorithm for Automated Classification in Largescale Forensic and Cybercrime Investigations. page 380, 2018. http://hdl.handle.net/11250/2491724.

[162] Andrii Shalaginov, Jan William Johnsen, and Katrin Franke. Cyber crime investigations in the era of big data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3672–3676, Boston, MA, December 2017. IEEE. http://ieeexplore.ieee.org/document/8258362/.

[163] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[164] Roy Shapira. Law as Source: How Investigative Journalists Can Superpower Their Reporting, September 2019. https://gijn.org/2019/09/04/law-as-source-how-investigative-journalists-can-superpower-their-reporting/.

[165] Jitesh Shetty and Jafar Adibi. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4, 2004. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.9477&rep=rep1&type=pdf.

[166] Aditya K. Sood and Richard J. Enbody. Crimeware-as-a-service—A survey of commoditized crimeware in the underground market. *International Journal of Critical Infrastructure Protection*, 6(1):28–38, 2013. https://linkinghub.elsevier.com/retrieve/pii/S1874548213000036.

[167] Eugene Spafford. Some Challenges in Digital Forensics. In Martin S. Olivier and Sujeet Shenoi, editors, *Advances in Digital Forensics II*, pages 3–9, Boston, MA, 2006. Springer US.

[168] Toine Spapens. Macro Networks, Collectives, and Business Processes: An Integrated Approach to Organized Crime. *European Journal of Crime, Criminal Law and Criminal Justice*, 18(2):185–215, 2010. https://brill.com/view/journals/eccl/18/2/article-p185_4.xml.

[169] Malcolm K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991. http://www.sciencedirect.com/science/article/pii/037887339190008H.

[170] Radina Stoykova. Digital evidence: Unaddressed threats to fairness and the presumption of innocence. *Computer Law & Security Review*, 42:105575, September 2021.

[171] Radina Stoykova and Katrin Franke. Standard Representation for Digital Forensic Processing. In *2020 13th International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE)*, pages 46–56, New York, NY, USA, May 2020. IEEE.

[172] Saatviga Sudhahar, Gianluca De Fazio, Roberto Franzosi, and Nello Cristianini. Network analysis of narrative content in large corpora. *Natural Language Engineering*, 21(1):81–112, January 2015.

[173] Haak Susan. The Expert Witness: Lessons from the US Experience. (28):32, 2015.

[174] Daniel R. Thomas, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R. Beresford. Ethical issues in research using datasets of illicit origin. In *Proceedings of the 2017 Internet Measurement Conference*, pages 445–462, London United Kingdom, November 2017. ACM. https://dl.acm.org/doi/10.1145/3131365.3131389.

[175] Markus Tobiassen and Bjørn Eckblad. Økokrim ble stoppet midt i ransakingen av Tidal, February 2019. https://www.dn.no/musikk/musikk/okokrim/usa/okokrim-ble-stoppet-midt-i-ransakingen-av-tidal/2-1-541426.

[176] Markus Tobiassen and Kjetil Sæter. Tidals lyttertall er manipulert, May 2018. https://www.dn.no/staticprojects/special/2018/05/09/0600/dokumentar/strommekuppet/.

[177] Chris Tufts. *The Little Book of LDA*. 2020. https://ldabook.com/.

[178] United Nations Office on Drugs and Crime. The Globalization of Crime: A Transnational Organized Crime Threat Assessment. *United Nations publication*, 2010.

[179] U.S. Department of Justice. Forensic Science, January 2021.

[180] Cybersecurity Ventures. Official Annual Cybercrime Report 2019. Technical report, 2019.

[181] Eva A. Vincze. Challenges in digital forensics. *Police Practice and Research*, 17(2):183–194, March 2016. http://www.tandfonline.com/doi/full/10.1080/15614263.2015.1128163.

[182] Alta Waal, Jacobus Venter, and Etienne Barnard. Applying Topic Modeling to Forensic Data. In Indrajit Ray and Sujeet Shenoi, editors, *Advances in Digital Forensics IV*, volume 285, pages 115–126, Boston, MA, 2008. Springer US. http://link.springer.com/10.1007/978-0-387-84927-0_10.

[183] Thomas Walmann. Ars forensica presentation, 2017.

[184] Haiying Wang and Huiru Zheng. Model Validation, Machine Learning. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 1406–1407. Springer New York, New York, NY, 2013. https://doi.org/10.1007/978-1-4419-9863-7_233.

[185] S. Wasserman, K. Faust, Cambridge University Press, M. Granovetter, University of Cambridge, and D. Iacobucci. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. https://books.google.com/books?id=CAm2DpIqRUIC.

[186] Satosi Watanabe. *Knowing and Guessing a Quantitative Study of Inference and Information*. 1969.

[187] Wikipedia. *Tidal-saken — Wikipedia,*. 2019. https://no.wikipedia.org/w/index.php?title=Tidal-saken&oldid=19404461.

[188] Wikipedia contributors. *Big data — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Big_data&oldid=986328266.

[189] Wikipedia contributors. *Enron scandal — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Enron_scandal&oldid=991255605.

[190] Wikipedia contributors. *Investigative journalism — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Investigative_journalism&oldid=987127981.

[191] Wikipedia contributors. *Machine learning — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=994108422.

[192] Wikipedia contributors. *Occam learning — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Occam_learning&oldid=965655834.

[193] Wikipedia contributors. *Panama Papers — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Panama_Papers&oldid=991770999.

[194] Wikipedia contributors. *Paradise Papers — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Paradise_Papers&oldid=991822599.

[195] Wikipedia contributors. *Percolation theory — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Percolation_theory&oldid=994457835.

[196] Wikipedia contributors. *Topic model — Wikipedia, The Free Encyclopedia*. 2020. https://en.wikipedia.org/w/index.php?title=Topic_model&oldid=988077829.

[197] Wikipedia contributors. Watergate scandal — Wikipedia, The Free Encyclopedia, 2021.

[198] Janet Williams. Good Practice Guide for Computer-Based Electronic Evidence, 2012.

[199] Andrew Williamson. Forensic Intelligence. In *Modern Police Leadership*, pages 245–259. Springer, 2021.

[200] David H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, October 1996. https://www.mitpressjournals.org/doi/abs/10.1162/neco.1996.8.7.1341.

[201] Jennifer Xu and Hsinchun Chen. Untangling Criminal Networks: A Case Study. In Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, editors, *Intelligence and Security Informatics*, pages 232–248, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[202] Jennifer Xu, Byron Marshall, Siddharth Kaza, and Hsinchun Chen. Analyzing and Visualizing Criminal Network Dynamics: A Case Study. In Hsinchun Chen, Reagan Moore, Daniel D. Zeng, and John Leavitt, editors, *Intelligence and Security Informatics*, pages 359–377, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[203] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, New York New York USA, August 2014. ACM.

[204] Yunus Yusoff, Roslan Ismail, and Zainuddin Hassan. Common Phases of Computer Forensics Investigation Models. *International Journal of Computer Science and Information Technology*, 3(3):17–31, June 2011. http://www.airccse.org/journal/jcsit/0611csit02.pdf.

[205] Junlong Zhang and Yu Luo. Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network. In *Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*, Bankog, Thailand, 2017. Atlantis Press. http://www.atlantis-press.com/php/paper-details.php?id=25874733.

[206] Xiong Zhang, Alex Tsang, Wei T Yue, and Michael Chau. The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers*, 17(6):1239–1251, 2015. Publisher: Springer.

[207] Leonid Zhukov. Network Analysis. Lecture 5. Centrality measures., February 2015. https://www.youtube.com/watch?v=DfV-pjRTlLg.

[208] André Årnes, Stefan Axelsson, Petter Christian Bjelland, Ausra Dilijonaite, Anders Orsten Flaglien, Katrin Franke, Jeff Hamm, Jens-Petter Sandvik, and Inger Marie Sunde. *Digital Forensics*. Wiley, 2017. https://books.google.com/books?id=Fk0nDwAAQBAJ.

[209] ØKOKRIM. ØKOKRIM investigating Tidal case, January 2019. https://www.okokrim.no/oekokrim-investigating-tidal-case.6181312-411472.html.

Part II

Publications

Chapter 2

Article I - Feasibility study of social network analysis on loosely structured communication networks

Jan William Johnsen and Katrin Franke. In Procedia Computer Science, volume 108, 2017, pages 2388-2392.

Abstract

Organised criminal groups are moving more of their activities from traditionally physical crime into the cyber domain; where they form online communities that are used as marketplaces for illegal materials, products and services. The trading of illicit goods drives an underground economy by providing services that facilitate almost any type of cybercrime. The challenge for law enforcement agencies is to know which individuals to focus their efforts on, in order to effectively disrupting the services provided by cybercriminals. This paper presents our study to assess graph-based centrality measures' performance for identifying important individuals within a criminal network. These measures have previously been used on small and structured general social networks. In this study, we are testing the measures on a new dataset that is larger, loosely structured and resembles a network within cybercriminal forums. Our result shows that well-established measures have weaknesses when applied to this challenging dataset.

2.1 Introduction

Law enforcement agencies report that cybercrime activity is growing and become more aggressive and technically proficient [3, 7] – although the majority of cybercriminals in online marketplaces have relatively low technical skills and capabilities. This suggests that a minority of cybercriminals use marketplaces to sell easy access to sophisticated tools and expertise through a business model called Crime as a Service (CaaS) [3], which allow lesser skilled cybercriminals to have more impact and success in their cyber attacks. A focus on identifying and disrupting criminals in the smaller and more technical skilled group will have a larger impact on stopping illegal activities in underground marketplaces. Because their skills and expertise are difficult to replace by the larger group, with lower technical skills.

Social Network Analysis (SNA) methods have been proposed [9] for the application of identifying central individuals within criminal networks. More specifically, centrality measures are used to determine central individuals by analysing their position in a network [8], represented by a graph as defined in Section 2.2. In previous research, centrality measures have been used to analyse relational structures in organisations [2, 4, 5] and terrorist groups [6]. The network size in these studies is between 30 and 150 individuals. Centrality measures have shown promising results to find central individuals in small and organised networks – although the networks have been incomplete or are just a sample from the total population.

However, real-world datasets are neither small nor organised, and they often require data preprocessing before they can be analysed. Although centrality measures have performed good on networks of smaller sizes by finding interesting individuals, this does not mean they will also perform good on larger and more loosely structured [1] networks. This paper is guided by the research question: *How can graph-based methods be applied to identify important individuals within a real-world online communication network?* Our research question seeks to determine the feasibility of centrality measures in applying it to the area of civil and criminal investigations.

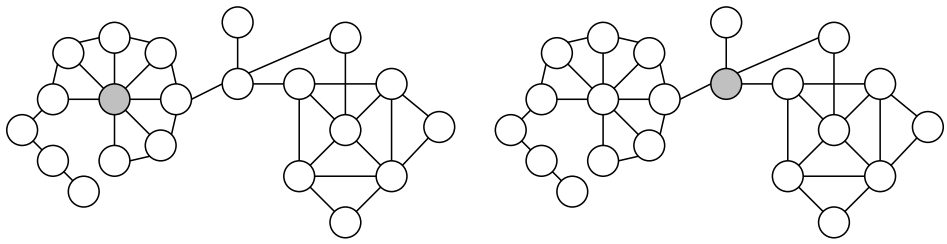
2.2 Methodology

We extracted information to represent the communication within Nulled.IO as graphs: users and the messages between them, represented as vertices and edges, respectively. It has not been pre-filtered and is used in its original form (detailed in Section 2.3) except for separating public and private messages; which results in two graphs with public communication between 26.11.2012 - 06.05.2016 and private communication between 14.01.2015 - 06.05.2016.

The reason for this division is twofold: (i) communication patterns are likely to be different between them, and (ii) civil investigators only have access to public communication in their investigation, whereas criminal investigators will have access to both.

The four centrality measures under evaluation are *degree*, *betweenness*, *closeness* and *eigenvector*. They differ in the interpretation of *important*; thus, different individuals will be ranked as more important in the same network; illustrated in Figures 2.1a - 2.2b.

An (undirected) graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, is represented in terms of the binary adjacency matrix A. Degree centrality is the most basic measure as it only counts directly adjacent vertices. For a vertex $v \in V$, it is defined by $C_D(v) = \sum_{u=1}^{n} A_{v,u}$, where $n = |V|$. The centrality measures discussed in this paper do not consider the diagonal elements in A [8], where $v = u$ because the relationship to oneself is not important.

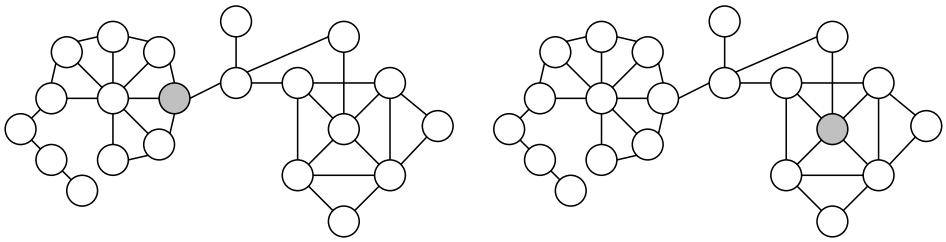


(a) Largest degree centrality **(b)** Largest betweenness centrality

Figure 2.1: Degree and betweenness centrality

Betweenness centrality looks at how often a vertex sits in the *geodesic* (shortest path) between two other vertices. A vertex is considered more important because it can act as a *broker* – i.e. arrange or negotiate plans and deals – and have more influence on the network by choosing to withhold or distort information [8]. Figure 2.1b highlights the vertex in the network with the highest betweenness centrality score because it sits in between two large subgraphs and one vertex. Betweenness centrality for a vertex v is defined by $C_B(v) = \sum \frac{\partial_{u,v,w}}{\partial_{u,w}}$, where $\partial_{u,w}$ is the total number of shortest paths between vertex u and w, and $\partial_{u,v,w}$ is the number of those paths that pass through v, and $u \neq v \neq w$.

Closeness centrality looks at the distance between one vertex and all the other vertices. A vertex is considered more important if it has a short distance to other vertices. In other words: the sum of distances to other vertices is low. Figure 2.2a highlights the vertex in the network with the best closeness centrality score because



(a) Largest closeness centrality **(b)** Largest eigenvector centrality

Figure 2.2: Closeness and eigenvector centrality

it has the shortest distance to all the other vertices. Closeness centrality for a vertex v is defined in $C_C(v) = [\sum_{u=1}^{n} d(v,u)]^{-1}$, where $d(v,u)$ is the distance (length of the shortest path) connecting v to u.

Eigenvector centrality expands on the idea of degree centrality, as it considers the edges to adjacent vertices. A vertex's score is not dependent on how many vertices it is connected to, but on many its adjacent vertices are connected to. This means that a vertex is important only if its neighbours are important – if they also have a higher degree centrality in the network. Figure 2.2b highlights the vertex in the network with the highest eigenvector centrality score. Eigenvector centrality for a vertex v is defined in $C_E(v) = \frac{1}{\lambda} \sum_{u=1}^{n} A_{u,v} C_E(v_u)$, where $\lambda \neq 0$ is some constant. The eigenvector value of vertex v is weighted by the sum of degree centralities of adjacent vertices.

2.3 Case study design

The database dump[1] used in our analysis is from an online forum (accessible from the clearnet) for distributing cracked software and trading stolen credentials. It is a 9.45 GB file, which was leaked 12.05.2016, with details about 599 085 user accounts, including 800 593 private and 3 495 596 public messages. It was used as a substitute for less available darknet forums datasets because forum users in both dark- and clearnet rely on electronic messaging to communicate, plan and organise. Although similarities between dark- and clearnet forums have not been shown in previous research, it is not unlikely to expect they are formed by similar social forces.

Table 2.1 show a full list of database (DB) tables and fields used to extract the needed information for constructing the graphs. The resulting two graphs were then exported in a Graph Exchange XML Format (GEXF), to ease later analyses.

[1] http://leakforums.net/thread-719337

Table 2.1: Database tables and fields of interest

Table	Fields
topics	tid, posts, starter_id, starter_name, forum_id
posts	pid, author_id, author_name, topic_id, new_topic
message_topics	mt_id, mt_starter_id, mt_to_count, mt_to_member_id, mt_replies

Two DB tables were combined to construct the public communication graph. DB table *topics* contain information on the author of forum threads, so field *starter_id* is treated as source vertex. For each forum thread (topic), field *topic_id* was used to retrieve all messages posted on that topic ID from DB table *posts*. Field *author_id* was treated as the target vertex, and for each message found the edge weight between two vertices was incremented.

DB table *message_topics* hold the metadata for private communication, where field *mt_starter_id* is used as source vertex, *mt_to_member_id* as target vertex, and *mt_to_count* + *mt_replies* as edge weight. The edge weight is the sum of messages sent to the recipient and the number of replies. The data extraction and analysis was performed on an Ubuntu 15.10 desktop computer, with Python scripts that we wrote for this purpose. The software used in this case study was MySQL and Python, with packages Networkx and MySQLdb.

2.4 Results

This section contains the result from four centrality measures on two (undirected) graphs, which are divided into public and private, as seen in Table 2.2 and 2.3 respectively. Tables are sorted in descending order by their centrality value because higher values indicate more central positions in the respective measures. They are limited to the first five results due to page limitations; however, it is enough to demonstrate that users are ranked differently according to centrality measures' interpretation of important.

The values have been normalised, so networks of different sizes can be compared with each other. Networkx can normalise the results for us. All values in Table 2.2 and 2.3 have been normalised in the range $[0, 1]$, according to equations found in [8].

We started the analysis on users that occupied similar ranks between each centrality measures and type of communication, to understand why they get their ranks. It was performed by manually inspecting the message contents, and it revealed that many of these individuals had roles such as administrator and moderators in

Table 2.2: Top ten public centrality results

ID	Degree	ID	Closeness
15398	0.31449	15398	0.51280
1337	0.06518	1337	0.44481
5481	0.03564	334	0.42281
16618	0.03036	3507	0.42001
410101	0.02872	2902	0.41946
ID	**Betweenness**	**ID**	**Eigenvector**
15398	0.50134	15398	0.47951
1337	0.07594	1337	0.21054
5	0.02790	334	0.14043
5481	0.02403	5481	0.11948
411677	0.02365	4782	0.10452

Table 2.3: Top ten private centrality results

ID	Degree	ID	Closeness
1	0.09466	1	0.37928
15398	0.03441	334	0.35757
1337	0.03275	1471	0.35631
1471	0.03194	1337	0.35437
51349	0.03074	51349	0.35118
ID	**Betweenness**	**ID**	**Eigenvector**
1	0.17174	193974	0.48531
15398	0.05871	61078	0.47249
1337	0.04811	51349	0.29031
1471	0.04593	315929	0.24046
334	0.03985	336307	0.16937

Nulled.IO. In addition to having responsibilities and being active on the forum, they also contributed with cracked software (mostly cheats for games) and distributing user credentials. Users in eigenvector centrality, in Table 2.3, differed from users in the other centrality measures as the two highest-ranking users (ID 193974 and 61078) was selling services of converting or trading between currencies.

Users with ID 1 and 15398 is ranking highest for degree centrality in both tables, up to 2.75 and 4.82 times larger than the second-highest values respectively. But they get their values because they are connected to more neighbours than other users. This indicates that they are very active in the hacker forum by communicat-

ing with many different users.

In Table 2.3, the user with ID 1 is 2.92 times larger than the second-highest value in betweenness centrality, which indicate that this user is sitting in between a lot more users. However, results from closeness centrality indicate that the network is more connected. As it shows that users have about equally short path to all other users – as it only decreases by 0.053 after 100 users. There was only user with ID 15398 in Table 2.2 that had significant values in all of the centrality measures. Because of our approach to constructing the public graph, this would indicate that the threads created by user ID 15398 are very popular, with 70 906 edges connecting to other users.

2.5 Conclusion

Organised criminal groups use anonymisation techniques to operate and move their illegal activities online. Where they form communities that are used as marketplaces for illegal materials, products and services. This drives the underground economy by providing services that facilitate almost any type of cybercrime. The challenge for law enforcement investigators is to know which individuals to focus their efforts and resources on, in order to disrupt the services provided by cybercriminals.

In this paper, we assessed the performance of four graph-based centrality measures, for their ability to identify important individuals which provide valuable services to other cybercriminals. Centrality measures have previously been used to study small and structured data sets (for example, Enron). However, we tested them on a newly leaked dataset that is larger, more loosely structured and a network with similarities to cybercriminal forums. Our result shows that well established graph-based measures have weaknesses when applied to this new dataset. For example, some individuals are ranked high and appear to be important to the forum. However, they actually had a less important contribution to the community, and their removal would have a low impact on illegal operations from the criminal forum.

Investigators already have centrality measures available in tools they use, such as IBM i2 Analyst's Notebook. However, they need to understand that it is not a silver bullet that automatically identifies important users. To avoid accusing someone of being the leader in a cybercriminal network, further analysis is needed to confirm they are really important for investigator's goals. Focusing on wrong individuals can be illustrated with a real-world example: In 2013, Silk Road was taken down after arresting some of their administrators. After law enforcement interference, a dozen new marketplaces spawned and took Silk Road's place.

Another important aspect of improving the results is to pre-filter the dataset before analysis. This can be done by removing dependent vertices (i.e. $C_D(v) = 1$), which can improve betweenness centrality results. The Nulled.IO forum should also have been analysed as a directed graph. Then additional centrality measures such as *in-* and *out-degree* would be able to identify users with high popularity and expansiveness, respectively.

2.6 Bibliography

[1] Kim-Kwang Raymond Choo. Organised crime groups in cyberspace: a typology. *Trends in Organized Crime*, 11(3):270–295, September 2008. http://link.springer.com/10.1007/s12117-008-9038-9.

[2] Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, 2005.

[3] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2014. Technical report, 2014. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf.

[4] J. S. Hardin, Ghassan Sarkis, and P. C. Urc. Network analysis with the enron email corpus. *Journal of Statistics Education*, 23(2), 2015.

[5] Reece Howard. Using Social Network Analysis: an example using the Enron corpus, July 2015. https://cambridge-intelligence.com/using-social-network-analysis-measures/.

[6] Valdis Krebs. Uncloaking Terrorist Networks. volume 7, page 4. First Monday, 2002. http://journals.uic.edu/ojs/index.php/fm/article/view/941.

[7] National Crime Agency. *Cyber Crime Assessment 2016*. July 2016. https://goo.gl/tKcxDN.

[8] C. Prell. *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2012. https://books.google.com/books?id=p4iTo566nAMC.

[9] Malcolm K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991. http://www.sciencedirect.com/science/article/pii/037887339190008H.

Chapter 3

Article II - Identifying central individuals in organised criminal groups and underground marketplaces

Jan William Johnsen and Katrin Franke. In International Conference on Computational Science. Springer Cham, 2018. pp. 379-386.

Abstract

Traditional, organised criminal groups are becoming more active in the cyber domain. They form online communities and use these as marketplaces for illegal materials, products and services, which drives the Crime as a Service business model. The challenge for law enforcement of investigating and disrupting the underground marketplaces is to know which individuals to focus effort on. Because taking down a few high impact individuals can have more effect on disrupting the criminal services provided. This paper presents our study on social network centrality measures' performance for identifying important individuals in two networks. We focus our analysis on two distinctly different network structures: Enron and Nulled.IO. The first resembles an organised criminal group, while the latter is a more loosely structured hacker forum. Our result shows that centrality measures favour individuals with more communication rather than individuals usually considered more important: organised crime leaders and cybercriminals who sell illegal materials, products and services.

3.1 Introduction

Traditional Organised Criminal Groups (OCGs) have a business-like hierarchy [11]; with a few leaders controlling the organisation and activities done by men under then, which does most of the criminal activities. OCG are starting to move their operations to the *darknet* [1, 5], where they form *digital undergrounds*. These undergrounds serve as meeting places for like-minded people and marketplace for illegal materials, products and services. This also allows the criminal economy to thrive, with little interference from law enforcement [5].

The transition between traditional physical crime to cybercrime changes how criminals organise. They form a loosely connected network in digital undergrounds [7, 12]; where the users can be roughly divided into two distinct groups [5]: a minority and a majority. The division is based on an individual's technical skill and capabilities, and the group names reflect how many individuals are found in each group. The minority group have fewer individuals; however, they have higher technical skills and capabilities. They support the majority – without the same level of skills – through the Crime as a Service (CaaS) business model [5]. The consequence is that highly skilled criminals develop tools that the majority group use as a service. This allows entry-level criminals to have greater impact and success in their cyber operations.

The challenge is identifying key actors [13] in digital undergrounds and stopping their activities. The most effective approach is to target those key actors found in the minority group [5]. We represent the communication pattern between individuals as a network, and then use Social Network Analysis (SNA) methods to investigate those social structures. An important aspect of SNA is that it provides scientific and objective measures for network structure and positions of key actors [4]. Key actors – people with importance or greater power – typically have higher centrality scores than other actors [4, 13].

We substitute the lack of available datasets of OCG and digital undergrounds with the Enron corpus and Nulled.IO, respectively. The Enron corpus has been extensively studied [3, 2, 6], where Hardin et al. [6] studied the relationships by using six different centrality measures. While Nulled.IO is a novel dataset for an online forum for distributing cracked software, and trade of leaked and stolen credentials. SNA methods have also been used by Krebs [10] to analyse the network surrounding the aeroplane hijackers from September 11th, 2001.

The dataset types in these studies are highly varied: ranging from a few individuals to hundreds of them; networks that are hierarchical or are more loosely structured; and complete and incomplete networks. The *no free lunch* theorem [15] states

that there is no algorithm that works best for every scenario. The novelty of our work is that we evaluate the results of centrality measures for two datasets with distinctly different characteristics. Our research tries to answer the following research questions: (i) How does centrality measures identify leading people inside networks of different organisational structures and communication patterns? And (ii) How good are they to identify people of more importance (i.e. inside the smaller population)? The answers to these questions are particularly important for law enforcement, to enable them to focus their efforts on those key actors whose removal has more effect for disrupting the criminal economy.

3.2 Materials and methods

3.2.1 Datasets

Although the Enron MySQL database dump by Shetty and Adibi [14] is unavailable today, we use a MySQL v5 dump of their original release[1]. The corpus contains 252 759 e-mail messages from 75 416 e-mail addresses. Nulled.IO[2] is an online forum which got their entire database leaked on May 2016. The forum contains details about 599 085 user accounts, 800 593 private messages and 3 495 596 public messages. The distinction between private and public is that private messages are between two individuals, while public messages are forum posts accessible by everyone. These datasets have very different characteristics: Enron is an organisation with a strict hierarchical structure, while Nulled.IO is a flat and loosely connected network.

The challenge of analysing our datasets is the large amount of information they contain. Every piece of information would be of potential interest in a forensic investigation; however, we limit the information to that which represents individual people and the communication between then. We use this to create multiple *directed graphs* (digraphs), where individuals are modelled as *vertices* and the communication between them as directed *edges*. A digraph G is more formally defined as a set V of vertices and set E of edges, where E contains ordered pairs of elements in V. For example, (v_1, v_2) is an ordered pair if there exists an edge between vertices v_1 and v_2, called *source* and *target* respectively.

3.2.2 Centrality measures

Centrality measures are graph-based analysis methods found in SNA, used to identify important and influential individuals within a network. We evaluate five popular centrality measures for digraphs: *in-degree* (C_{D^-}), *out-degree* (C_{D^+}),

[1] http://www.ahschulz.de/enron-email-data/

[2] http://leakforums.net/thread-719337 (recently became unavailable)

betweenness (C_B), *closeness* (C_C) and *eigenvector* (C_E). They are implemented in well-known forensic investigation tools, such as IBM i2 Analyst's Notebook [8].

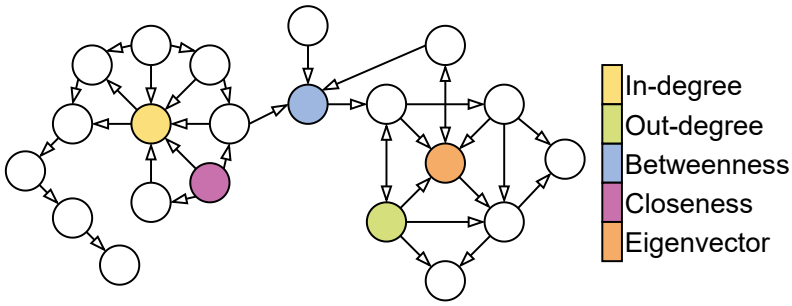


Figure 3.1: Highest ranking vertices in a digraph

The centrality measures differ in their interpretation of what it means to be 'important' in a network [13]. Thus, some vertices in a network will be ranked as more important than others. Figure 3.1 illustrate how vertices are ranked differently. The number of vertices and edges affects the centrality values. However, normalising the values will counter this effect and allow us to compare vertices from networks of different sizes. Our analysis tool *Networkx* uses a scale to normalise the result to values $[0, 1]$.

3.3 Experiment

We first constructed three weighted digraphs to represent the communication between users in Enron and Nulled.IO. Only a few database (DB) tables and fields had the necessary information to build the digraphs: *message (sender)* and *recipientinfo (rvalue)* for Enron, and *topics (starter_id)*, *posts (author_id)* and *message_topics (mt_starter_id, mt_to_member_id, mt_to_count, and mt_replies)* for Nulled.IO. The digraph construction method can be generalised as: find the sender and receiver of messages. Represent them as unique vertices in a digraph (if not already exists) and connect them with an edge from sender to receiver. These operations were repeated for every public/private and e-mail message. Finally, edges' weights were initialised once it was first created and incremented for each message with identical vertices and edge direction. The data extraction and analysis was performed on a desktop computer, with a MySQL server and Python, with packages Networkx v1.11 and PyMySQL v0.7.9.

Pre-filtering and population boundary

The digraph construction included a bit more information than necessary, which have to be removed before the analysis. However, we want to find a balance

between reducing the information without removing valuable or relevant information. To analyse the hierarchical structure of Enron (our presumed OCG), we have to remove vertices which do not end with '@enron.com'. Additionally, we removed a few general e-mail addresses which could not be linked to unique Enron employees. A total of 691 vertices was removed by these steps.

We have previously identified user with ID 1 as an administrator account on the Nulled.IO forum [9], used to send out private 'welcome'-messages to newly registered users, which can skew the results. Thus, we only remove edges from ID 1 to other vertices – when its weight equals one – to achieve the goal of information preservation. The private network that originally had 295 147 vertices and 376 087 edges, was reduced to 33 647 (88.6%) and 98 253 (73.87%), respectively. The public thread communication network did not undergo any preprocessing. From its original 299 702 vertices and 2 738 710 edges, it was reduced to 299 105 (0.2%) and 2 705 578 (1.22%), respectively.

The final preprocessing step was to remove isolated vertices and self-loops. Isolated vertices had to be deleted because they have an infinite distance to every other vertex in the network. Self-loops was also removed because it is not interesting to know a vertex' relation to itself. The reduction in vertices and edges for the Nulled.IO public digraph was a consequence of this final step.

3.4 Results

The results are found in Tables 3.1 and 3.2, sorted in descending order according to vertices' centrality score. Higher scores appear on top and indicate more importance. The results are limited to the top five individuals due to page limitations.

The goal of our research is to identify leaders of OCG or prominent individuals who sell popular services; people whose removal will cause more disruption to the criminal community. To evaluate the success of centrality measures, we first had to identify people's job positions or areas of responsibilities in both Enron and Nulled.IO. We combined information found on LinkedIn profiles, news articles and a previous list[3] to identify Enron employees' position in the hierarchy. For Nulled.IO users, we had to manually inspect both private and public messages to estimate their role or responsibility. The total number of possible messages to inspect made it difficult to determine the exact role for each user.

3.4.1 Enron

sally.beck is within the top three highest-ranking individuals in all centrality measures, except for eigenvector centrality. Her role in Enron was being a Chief Oper-

[3] http://cis.jhu.edu/~parky/Enron/employees

ating Officer (COO); responsible for the daily operation of the company and often reported directly to the Chief Executive Officer (CEO). Her result corresponded to expectations of her role: a lot of sent and received messages to handle the daily operation.

Table 3.1: Top ten centrality results Enron

UID	C_{D^-}	UID	C_{D^+}
louise.kitchen	0.03374	david.forster	0.07393
steven.j.kean	0.02900	sally.beck	0.06559
sally.beck	0.02884	kenneth.lay	0.04982
john.lavorato	0.02803	tracey.kozadinos	0.04955
mark.e.taylor	0.02685	julie.clyatt	0.04907
UID	C_B	**UID**	C_C
sally.beck	0.02152	sally.beck	0.39612
kenneth.lay	0.01831	david.forster	0.38500
jeff.skilling	0.01649	kenneth.lay	0.38362
j.kaminski	0.01555	julie.clyatt	0.38347
louise.kitchen	0.01145	billy.lemmons	0.38293
UID	C_E		
richard.shapiro	0.37927		
james.d.steffes	0.33788		
steven.j.kean	0.27800		
jeff.dasovich	0.27090		
susan.mara	0.25839		

kenneth.lay and *david.forster* are two individuals with high rankings in all centrality measures, except for eigenvector centrality. They are CEO and Vice President, respectively. *kenneth.lay* and his second in command *jeff.skilling* was the heavy hitters in the Enron fraud scandal.

Although there were a few CEOs in the Enron corporation, many of the higher ranking individuals had lower hierarchical positions. Most notably, this occurred in eigenvector centrality; however, this is because of how this measure works. Finally, our result also shows that centrality measures usually ranks the same individuals as being more important than others.

3.4.2 Nulled.IO

Unique Identifier (UID) 0 in the public digraph appears to be a placeholder for deleted accounts because the UID does not appear in the member list and the username in published messages are different. UID 4, 6, 8, 15398, 47671 and 301849,

among others, provides free cracked software to the community, with most of them being cheats or bots for popular games. While UID 1337 and 1471 appears to be administrators.

Table 3.2: Top five public and private centrality results Nulled

Public centrality results					
UID	$\mathbf{C_{D^-}}$	**UID**	$\mathbf{C_{D^+}}$	**UID**	$\mathbf{C_B}$
15398	0.23695	1471	0.00393	0	0.00959
0	0.16282	8	0.00321	15398	0.00855
1337	0.06466	193974	0.00294	1337	0.00461
4	0.05656	47671	0.00273	1471	0.00334
6	0.04276	118229	0.00266	193974	0.00219
UID	$\mathbf{C_C}$	**UID**	$\mathbf{C_E}$		
1471	0.03564	1337	0.28764		
8	0.03553	0	0.27157		
118229	0.03542	15398	0.25494		
169996	0.03540	334	0.23961		
47671	0.03520	71725	0.22798		
Private centrality results					
UID	$\mathbf{C_{D^-}}$	**UID**	$\mathbf{C_{D^+}}$	**UID**	$\mathbf{C_B}$
1	0.08412	1	0.42331	1	0.41719
1471	0.05028	51349	0.00773	1471	0.02369
1337	0.04289	88918	0.00695	334	0.02286
8	0.03970	47671	0.00617	1337	0.02253
15398	0.03967	334	0.00600	15398	0.02129
UID	$\mathbf{C_C}$	**UID**	$\mathbf{C_E}$		
1	0.40665	61078	0.45740		
51349	0.28442	51349	0.30353		
88384	0.28102	1	0.24505		
10019	0.28080	88918	0.21214		
61078	0.28043	193974	0.19651		

UID 1 in the private digraph is found on top of (almost) all centrality measure. Although this account appears as a key actor, it was mostly used to send out thousands of automatic 'Welcome'-messages, 'Thank you'-letters for donations and support and other server administrative activities.

UIDs 8, 334, 1471, 47671, 51349 and 88918 in the private digraph cracks various online accounts, such as Netflix, Spotify and game-related accounts. They usually go after the 'low hanging fruit' that have bad passwords or otherwise easy to get.

Most of the users have low technical skills; however, they are willing to learn to be better and to earn more money from their scriptkid activities. They want to go into software development for economic gains or learn more advanced hacker skills and tools to increase their profit.

3.5 Discussion and conclusion

Law enforcement agencies can disrupt the CaaS business model or OCG when they know which key actors to effectively focus their efforts on. However, implementations of centrality measures in forensic investigation tools are given without any explanation or advice for how to interpret the results; which inadvertently can lead to accusation of lesser criminals of being among the leaders of criminal organisations. Although the centrality measures do not perfectly identify individuals highest in the organisation hierarchy, our result shows that potential secondary targets can be found via them. Secondary targets are individuals that any leader relies on to effectively run their organisation.

Contemporary centrality measures studied here most often identified individuals with a natural higher frequency of communication, such as administrators and moderators. However, going after forum administrators is only a minor setback, as history has shown a dozen new underground marketplaces took Silk Road's place after it was shut down. Thus, the problem with current centrality measures is that they are affected by the network connectivity rather than actual criminal activities.

Our result demonstrates their weakness, as centrality measures cannot be used with any other definition for their interpretation of 'importance'. There is a lack of good interpretations of current centrality measures that fit for forensic investigations. Interpretations which are able to effectively address the growing problem of cybercrime and the changes it brings. We will continue working on identifying areas where already existing methods are sufficient, in addition to developing our own proposed solutions to address this problem.

3.6 Bibliography

[1] Kim-Kwang Raymond Choo. Organised crime groups in cyberspace: a typology. *Trends in Organized Crime*, 11(3):270–295, September 2008. http://link.springer.com/10.1007/s12117-008-9038-9.

[2] Jana Diesner. Communication Networks from the Enron Email Corpus: "It's Always About the People. Enron is no Different.". 2005. https://pdfs.semanticscholar.org/875b/59b06c76e3b52a8570103ba6d8d70b0cf33e.pdf.

[3] Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on*

Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA, 2005.

[4] David Décary-Hétu and Benoit Dupont. The social network of hackers. *Global Crime*, 13(3):160–175, 2012. http://www.tandfonline.com/doi/abs/10.1080/17440572.2012.702523.

[5] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2014. Technical report, 2014. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf.

[6] J. S. Hardin, Ghassan Sarkis, and P. C. Urc. Network analysis with the enron email corpus. *Journal of Statistics Education*, 23(2), 2015.

[7] Thomas J. Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology*, 6(1):891, 2012.

[8] IBM Knowledge Center. Centrality and centrality measures, 2017. https://www.ibm.com/support/knowledgecenter/en/SS3J58_9.0.8/com.ibm.i2.anb.doc/sna_centrality.html.

[9] Jan William Johnsen and Katrin Franke. Feasibility Study of Social Network Analysis on Loosely Structured Communication Networks. *Procedia Computer Science*, 108:2388–2392, 2017. http://linkinghub.elsevier.com/retrieve/pii/S1877050917307561.

[10] Valdis Krebs. Uncloaking Terrorist Networks. volume 7, page 4. First Monday, 2002. http://journals.uic.edu/ojs/index.php/fm/article/view/941.

[11] Vy Le. Organised Crime Typologies: Structure, Activities and Conditions. *International Journal of Criminology and Sociology*, 1:121–131, 2012.

[12] Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. Identifying Digital Threats in a Hacker Web Forum. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pages 926–933, Paris, France, 2015. ACM Press. http://dl.acm.org/citation.cfm?doid=2808797.2808878.

[13] C. Prell. *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2012. https://books.google.com/books?id=p4iTo566nAMC.

[14] Jitesh Shetty and Jafar Adibi. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4, 2004. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.9477&rep=rep1&type=pdf.

[15] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997. http://ieeexplore.ieee.org/abstract/document/585893/.

Chapter 4

Article III - The impact of preprocessing in natural language for open source intelligence and criminal investigation

Jan William Johnsen and Katrin Franke. In 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019. pp. 4248-4254.

Abstract

Underground forums serve as gathering place for like-minded cybercriminals and are a continued threat to law and order. Law enforcement agencies can use Open-Source Intelligence (OSINT) to gather valuable information to proactively counter existing and new threats. For example, by shifting criminal investigation's focus onto certain cybercriminals with a large impact in underground forums and related criminal business models. This paper presents our study on text preprocessing requirements and document construction for the topic model algorithm Latent Dirichlet Allocation (LDA). We identify a set of preprocessing requirements based on literature review and demonstrate them on a real-world forum, similar to those used by cybercriminals. Our result shows that topic modelling processes need to follow a very strict procedure to provide a significant result that can be useful in OSINT. Additionally, more reliable results are produced by tuning the hyper-

parameters and the number of topics for LDA. We demonstrate improved results by iterative preprocessing to continuously improve the model, which provide more coherent and focused topics.

4.1 Introduction

OSINT exploits publicly available data such as pictures, video and text to piece together factual data – i.e. information – for an end goal. Two overlapping developments have particularly influenced the growth of OSINT: expansion of social media and big data [13]. Social media is a good example of big data in practice, as tons of user-produced videos and texts are uploaded onto the Internet every day. Information gathered from open sources can give insights into world events; however, piecing together relevant data from the vast sea of materials can be difficult. Furthermore, big data majorly consists of unstructured data, which current traditional analytical tools are not built to handle.

Researchers frequently repeat the '80 per cent rule', which refer to the quantification of open-source contribution to intelligence [10, 18]. It is difficult to put an estimate on how much OSINT contribute to an intelligence operation, and the 80 per cent number is generally considered a mischievous red herring [10]; however, it provides an opportunity where OSINT can offer *significant value to proactive Cyber Threat Intelligence (CTI) to organisations about threats they were not previously aware of* [5, 17]. Consequently, data acquisition from OSINT are largely automated and can cause an increase in false positives [17]. In other words, the result of automated processes can have a negative effect on information reliability.

Law enforcement agencies have primarily used reactive approaches in criminal investigations for decades. New proactive approaches and utilising the vast amount of unstructured data can assist law enforcement agencies in prevent crime and upholding the law. Information is key to any criminal investigation [2], where information is constructed from data. However, correctly structuring, analysing and extracting useful knowledge or facts from unstructured data is a challenge. The goal is to gather sufficient information to accurately and adequately explain the circumstances of a situation or incident. Additionally, *the reliability and validity of data can change with attributes to the data source and the methods used to process the data* [2].

One goal of OSINT is to make sense of a lot of unstructured data, e.g. by automatically analyse various discussion forums to understand new trends or progression of malware development. Natural Language Processing (NLP) is used to process and analyse large amounts of natural language data, where LDA is one of the more popular algorithms. LDA is a generative statistical model, commonly used to cat-

egorise a set of observations (i.e. text) into unobserved groups that explain why some parts of the data are similar. LDA is described further in Section 4.3.

Every algorithm, including LDA, is susceptible to the expression 'garbage in, garbage out'. In other words, results will be incorrect if the input is erroneous, regardless of the algorithm's accuracy. The way these LDA models are trained, and in particular, how their inputs are preprocessed (if at all) is something we find missing in previous research. Therefore, our research concentrates on improving our current understanding of how to best construct documents as input for the LDA algorithm. We first briefly explain how the Machine Learning (ML) and the forensic process model can be linked, and then we define which requirements must apply for using LDA in a digital forensic context. With these requirements in mind, we will cross-validate three different document construction methods for LDA and study it in detail on the Nulled dataset. We primarily focus on OSINT in the context of digital forensics, but it will hold the same for intelligence operations.

Recently, LDA has been widely studied from a digital forensic perspective. Anwar et al. [3] analyse authorship attribution for Urdu text; Porter [14] splits his dataset into time intervals to find the evolution of hacker tools and trends; Caines et al. [6] uses ML and rule-based classifiers to automatically label post type and intent from posts in an underground forum; Samtani et al. [15] designed a novel CTI framework to analyse and understand threats present in hacker communities; L'huillier et al. [12] combine text mining and social network analysis to extract key members from dark web forums.

Text preprocessing varies widely in these studies, e.g. grammatical mistakes and word preferences are relevant in authorship attribution [3] or hacker forums contain atypical language [14]. They have a few issues, such as using Google Translate to convert text into English [15] or not checking model fit [12]. Additionally, they frequently do not describe how they structure the LDA input.

This article is structured in the following way: Section 4.2 describes previous and relevant work for our research, linking the ML and forensic process model and defining LDA preprocessing requirements; Section 4.3 and 4.4 report any preprocessing on the data, define the LDA document construction and provide results of our real-world scenario demonstration. We discuss the significance of our results and give a recapitulation of this article in Section 4.5.

4.2 Previous work

Data preprocessing is an integral step from the perspective of the ML process model – as described by Kononenko and Kukar [11] – where data quality directly affects the ability of ML models to learn. Furthermore, a survey by Crowd-

Flower [8] found that 60 per cent of the professionals spend much of their time cleaning and organising data. The same emphasis on data quality also holds for digital forensics. Andersen [2] gives details of the digital forensic process, in relation to criminal cases. He points out that information is crucial, and it should be reliable to have any value in a court of law. It is beyond this article to have a complete comparison of both process models, but there is a mutual understanding in both domains that the preprocessing phase is the most crucial step. Data preprocessing is a time-consuming and crucial step that consolidate and structure data to improve the accuracy of results.

Both the user of a system and the system itself have some requirements for it to be accurate and precise, i.e. reliable. We focus our requirements from the user's perspective: what they need to do to adeptly use the system, such as LDA in a digital forensic context. Text analysis typically begins with preprocessing the input data, but related literature varies widely with regards to which preprocessing method they utilise. Requirements should improve the algorithms' ability to identify interesting or important patterns in the data, instead of noise. The following list is composed of some common recommendations for cleaning the data [9, 14].

- **Word normalisation:** Inflected languages modifies words to express different grammatical categories. *Stemming* and *lemmatisation* are two methods to normalise text, as they help find the root form of words. Stemming removes suffixes or prefixes used with a word, without considering the resulting word belongs to a language. Lemmatisation reduces the inflected words properly while ensuring that the root word belongs to the language.
- **Stop word removal:** Words that are generally the most common words in a language, which tend to be over-represented in the result unless removed. They do not contain any important significance. However, removing stop words indiscriminately means you can accidentally filter out important data.
- **Uninformative word removal:** Similar to stop word removal, however, it is a domain-specific list of uninformative words. It can be quite long and depend on the domain producing the text in question.
- **Word length removal:** Remove words that have fewer than x (e.g. three) characters.
- **Document de-duplication:** Eliminating duplicate copies of repeating data, i.e. removing identical documents that appear frequently.
- **Expanding/replacing acronyms:** Acronyms are used quite often and may need some subject matter expertise to understand.

- **Other:** Convert everything to lowercase and remove punctuation marks/special symbols. Finally, remove extra white-spaces.

Requirements which reduce the vocabulary size has clear advantages for the quality. For example, removing stop words leave remaining terms that convey clearly topic-specific semantic content. Schofield et al. [16] looked at some of the common practices we have listed and found that many have either no effect or a negative effect. For example, (i) effects from document duplication were minimal until they had a substantial proportion of the corpus; (ii) stop word removal (determiners, conjunctions and prepositions) can improve model fit and quality; and (iii) stemming methods perform worse.

4.3 Methodology

There are several topic modelling algorithms [1]; however, we selected LDA because it is typically more effective and generalises better than other algorithms. This is beneficial as our proposed method may generalise to more specific domains, such as those of underground forums. Furthermore, LDA can extract human-interpretative topics from a document corpus, where each topic is characterised by the words they are most associated with. LDA [4] is a way of 'soft clustering' using a set of documents and a predefined k number of topics. Each document has some probability of belonging to several topics, which allow for a nuanced way of categorising documents.

The three hyper-parameters k, α and β adjust the LDA learning. Where k is a predefined amount of topics and α and β regulate two Dirichlet distributions. These Dirichlet distributions adjust the LDA model document-topic density and topic-word density, respectively. More specifically, LDA models assume documents consists of fewer topics at low α values, while higher α values documents can consist of more than one topic. Higher values will likely produce a more uniform distribution so that a document will have an even mixture of all the topics. Hyper-parameter β works similarly, but adjust the word distribution per topic. Thus, topics consist of less words at low β values and more words at higher values. LDA is most commonly used to (i) shrink a large corpus of text to some sequence of keywords, (ii) reduce the task of clustering or searching a huge number of documents, (iii) summarise a large collection of text or (iv) automatically tag new incoming text by the learned topics.

We use the previously mentioned requirements and preprocessing recommendations from Schofield et al. [16], such as removing about 700 of the most common English stop words. Following their recommendations, we decided to not remove duplicated documents nor use stemming, as this was reported to have little effect.

We removed additional text such as HTML tags (incl. their attributes), HTML entities (e.g.), symbols and extra spaces. Finally, we removed all rows with an empty text field and converted everything to lower case characters.

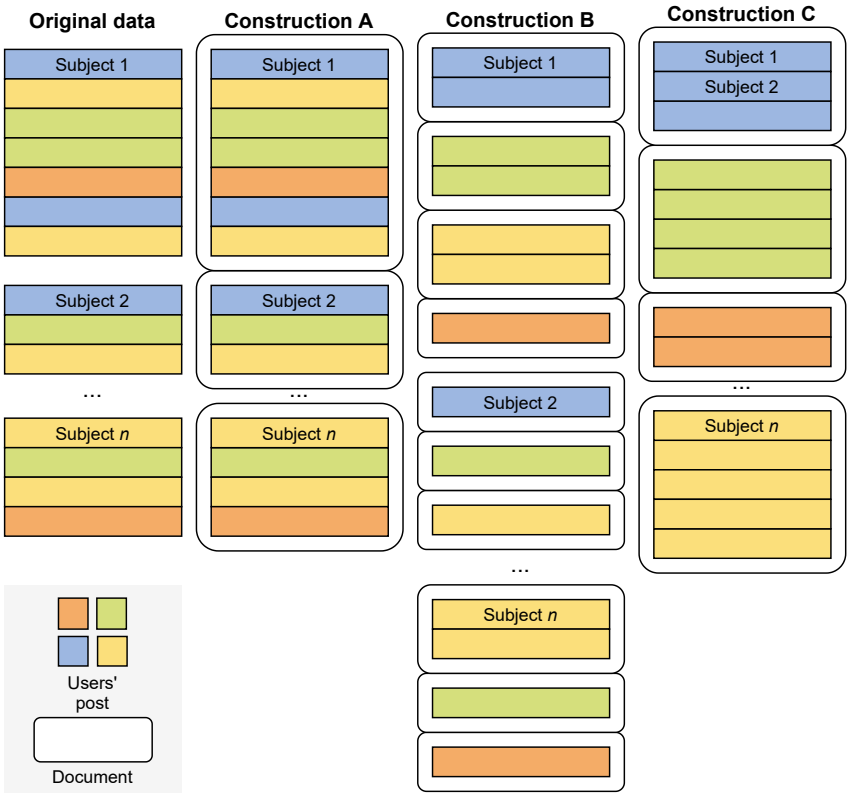


Figure 4.1: Document construction approaches analysed in this article. Construction A is subject-centred; construction B is subject-user-centred; while construction C is user-centred. Unique users are marked with different colours.

Users can write public *posts* to communicate with other forum users. These posts can have two distinctions: a *subject* is started by an initial post by a user, while other users are able to *reply* with their own posts to subjects. There is always zero or more replies associated with each subject. Figure 4.1 illustrate this type of interactions between users, where each user is depicted with different colours.

We focused our document construction method on the criteria to include all available posts found on the forum and ended up with identifying three distinct ways that we named: A, B and C. Figure 4.1 also portrays these document construction methods, where A is subject-centred, B is subject-user-centred, and C is user-

centred. Other construction approaches, than those shown in Figure 4.1, can be created and would yield different results. However, we decided not to consider them further as they would have too much information loss due to ignoring many posts.

Construction A keeps the original subject-structure found on the forum. In other words, one document is the combination of the subject starter and all its replies. Construction B builds upon this idea of being subject-centred. However, this approach combines the posts from users in a subject into separate documents. Finally, construction C combines all posts for distinct users into a separate document; i.e. one document consists of all posts that have been written by a specific user.

The motivation for construction A is to capture the overall activity on the forum, to get a high-level overview of topics that users are talking about. However, combining all posts from various users per subject might obscure the result. Therefore, we designed construction B to be subject and user-centred, as this could produce a more accurate result. While construction C is user-centric and should capture more of the interests for forum users.

The number of latent topics, k, is a parameter we have to set in LDA models; we explore $k = 10$, 20, 30, 40, 50 and 60 in our experiment. The other parameters α and β are either inferred from the data ($\frac{1}{k}$ when they are set to *None*) or set to the values 0.05, 0.1, 0.5, 1, 5 and 10.

Finally, we have to evaluate the model quality after the unsupervised learning process. We use k-fold cross-validation to assess how well the LDA models will generalise to an independent data set. For each analysis, we split the data into five folds: each fold is used for training the LDA model four times and testing the model one time. We use *perplexity* to objectively measure how well our model predicts the testing fold, where a low perplexity score indicates a better model. Furthermore, we use mean perplexity (i.e. the arithmetic mean for each fold) to compare all 882 ($k \times \alpha \times \eta$ combinations) models between each other. We select models with the lowest perplexity for further manual inspection.

4.4 Experiment and results

We explicitly concentrate our attention on data preprocessing in this research article, where LDA document construction is centre. It is, therefore, out of our scope to focus on the data gathering process, such as running web scraping tools to extract OSINT from real-world underground forums. Instead, we will use a dataset of 'Nulled' that was leaked in May 2016. It is a hacker forum on the deep web, that facilitate the brokering of compromised passwords, stolen bitcoins and other sensitive data. Nulled's Structured Query Language (SQL) database was leaked in

its original form, without any filtration or preprocessing. Their database contained details about 599 085 user accounts, 800 593 private messages and 3 495 596 public messages. We imported it to a MySQL server and exported the necessary information from tables and fields with a Python script, using the Pandas package. More specifically, we stored information found in database tables 'topics' and 'posts' (columns: 'author_id', 'post', 'topic_id') in a file for further analysis.

We used Pandas to group the three construction methods following the design described in Section 4.3 and depicted in Figure 4.1. The text column 'post' was further processed (described in Section 4.3) to make it suitable for LDA and document generation. We fit the LDA algorithm from the Scikit-learn package, for all the possible parameter combinations. All three document construction approaches were analysed using 294 distinct combinations of LDA hyper-parameters. We ran a total of 882 (294×3) LDA analyses to find the optimal combination of parameters. Table 4.1 shows the best ten models with the lowest perplexity.

Interestingly, our best result had very low hyper-parameters and 10 topics. While Samtani et al. [15] found an optimal topic number ranging from between 80 and 100. More importantly, Chang et al. [7] found that perplexity is not strongly correlated to human interpretation, as they found that the most frequent words in topics usually do not describe a coherent idea for those topics. A human forensic analyst would at least manage to interpret and understand fewer topics than something like 80 and 100 topics. However, fewer topics with a low perplexity score are not guaranteed to be easier interpreted by a human analyst. An important note is that low hyper-parameters also result in a slower convergence rate. While this solution might not be suitable for any time-critical criminal investigation, it could be applied to proactive OSINT gathering.

Table 4.2 show the five most frequent words from each topic, from the three best models which were manually inspected. These topics are not sorted in any particular order. Some words appear in multiple topics, such as *hide*, *color*, *http/https* and numbers, which does not provide any meaningful interpretation of topics. For example, 'hide' is a tag in the BBcode lightweight markup language, commonly used to format posts in many message boards. It is frequently used to withhold information until a visitor creates a user account on the forum and gains privileges to view the hidden content.

The various document construction methods (as seen in Table 4.2) does not show much variance in the identified keywords. The main difference was the number of documents that the LDA could learn from. Document construction A has 120 875 documents, B contains 2 794 304, and C has 272 023. Although document construction B had 2 212 per cent greater number of documents to learn from than

Table 4.1: Ten best models with hyper-parameter combinations

Construction A					Construction B				
#	α	η	k	Perplexity	#	α	η	k	Perplexity
1	0.05	0.05	10	5855.00	1	None	0.05	10	7088.24
2	0.10	0.05	10	5886.47	2	0.10	0.05	10	7133.69
3	None	0.05	10	5960.00	3	0.50	0.05	10	7133.89
4	0.50	0.05	10	6035.86	4	0.05	0.05	10	7268.40
5	0.05	0.10	10	6279.13	5	1.00	0.05	10	7484.53
6	1.00	0.05	10	6299.63	6	0.05	None	10	7763.43
7	None	None	10	6325.16	7	None	None	10	7768.09
8	0.10	None	10	6354.32	8	0.05	0.10	10	7870.99
9	0.10	0.10	10	6354.98	9	None	0.10	10	7877.45
10	0.50	0.10	10	6476.63	10	0.50	None	10	7937.83

Construction C				
#	α	η	k	Perplexity
1	None	0.05	10	8111.34
2	0.05	0.05	10	8276.60
3	0.10	0.05	10	8344.80
4	0.50	0.05	10	8492.75
5	0.10	0.10	10	8687.27
6	None	None	10	8785.00
7	0.50	None	10	8865.91
8	0.05	None	10	8889.34
9	None	0.10	10	8930.34
10	1.00	0.05	10	8947.48

method A, it didn't produce any significant differently result. Thus, it can be recommended to go with the two other document constructions (A and C) as they produce a similar and faster result using fewer documents.

We need to further improve our result found in Table 4.2 to make the topics more clear for human analysts. We repeat the previous preprocessing steps and adding some new steps to enhance the result. We begin by iteratively identify and remove BBcode tags and additional uninformative words[1] from topics. We also removed numbers during the preprocessing, as numbers had very little meaning other than

[1] http, https, www, gmail, hotmail, yahoo, inurl, ty, font, color, youtube, asp, well, post, myfonts, otf, abc, qwerty, ru, qwe, rar, add, true, beta, day, ip, net, aol, uk, function, live, fr, msn, var, de, br, nulled, menu, wa, time, people, ha, window, thing, start, year, de, site, php, zip, uk, pl, web, edition, lol, work, aspx, xmlrpc, html, view, content, xd

Table 4.2: Five most frequent words for topics

#	Construction A	#	Construction B
1	account, good, help, time, accounts	1	80, 8080, 120, 195, 3128
2	80, 8080, 120, 195, 3128	2	account, http, hide, accounts, kappa
3	ty, thx, nice, man, hide	3	download, hide, bot, https, bol
4	gmail, hotmail, yahoo, net, aol	4	http, site, de, php, net
5	ty, fixed, version, download, bot	5	sharing, testing, script, best, scripts
6	bol, scripts, script, https, legends	6	ty, http, members, 123456a, tx
7	php, inurl, site, v1, 123456789a	7	gmail, hotmail, yahoo, check, thx
8	color, ru, size, http, hide	8	man, php, bro, mate, yahoo
9	http, https, youtube, watch, members	9	test, works, lol, hope, game
10	game, origin, sims, email, games	10	thx, nice, good, work, share

#	Construction C
1	account, hide, http, https, accounts
2	80, 8080, 195, 120, 3128
3	download, bot, version, file, script
4	thx, nice, man, good, bro
5	ty, test, thx, nice, bro
6	gmail, hotmail, yahoo, php, http
7	site, php, color, hide, http
8	game, gmail, hotmail, games, captured
9	tks, unknown, 5900, password, null
10	55336, 123456789a, 123, ruddy, asdf3425j3d

being related to network ports or passwords. Finally, we used lemmatisation due to the frequent similar words such as ‘account’, ‘accounts’, ‘member’, ‘members’ and so forth.

After conducting the iterative preprocessing, we use the previously gained know-

ledge to adjust the hyper-parameters in our experiment. We re-run the experiment for all document construction approaches using low hyper-parameters: where α and are set to values None, 0.05 and 0.1 and k set to values 10 and 20 – resulting in running 54 (18×3) additional analyses. Table 4.3 show that the perplexity increase for the iterative preprocessing steps.

Table 4.3: Iterative ten best models with hyper-parameter combinations

Construction A					**Construction B**				
#	α	η	**k**	**Perplexity**	**#**	α	η	**k**	**Perplexity**
1	0.05	0.05	10	18249.81	1	None	0.05	10	17582.44
2	None	0.05	10	18332.84	2	0.10	0.05	10	17825.29
3	0.10	0.05	10	18343.05	3	0.05	0.05	10	17827.16
4	0.10	0.10	10	19380.88	4	None	None	10	18732.37
5	None	0.10	10	19434.33	5	0.05	None	10	18788.12
6	0.05	None	10	19525.33	6	None	0.10	10	18822.09
7	0.10	None	10	19546.14	7	0.05	0.10	10	18998.47
8	0.05	0.10	10	19576.82	8	0.10	None	10	19077.16
9	None	None	10	19793.34	9	0.10	0.10	10	19131.95
10	None	None	20	22685.33	10	None	0.05	20	22562.58

Construction C				
#	α	η	**k**	**Perplexity**
1	0.10	0.05	10	29255.27
2	0.05	0.05	10	29595.43
3	None	0.05	10	29608.72
4	0.10	None	10	29721.30
5	0.05	None	10	29834.72
6	0.10	0.10	10	30328.28
7	None	None	10	30579.64
8	0.05	0.10	10	30974.97
9	None	0.10	10	31947.12
10	0.05	None	20	41342.10

The more frequent words per topic, as seen in Table 4.4, also show greater coherent ideas per topics after additional iterative preprocessing. For example, there exist topics that: (i) express gratitude or appreciation (work, thx, nice, share, good), (ii) about popular games (lol, battlefield, fifa, sims, origin), (iii) leaking of credentials (username, password), (iv) various malicious tools (stealer, crypter, phisher, rat) and (v) administrative purposes (member, ban, pm).

Document construction A can be suitable to get an overview of what the under-

Table 4.4: Iterative five most frequent words for topics

#	Construction A	#	Construction B
1	account, file, bot, download, link	1	help, crack, link, guy, bol
2	comcast, music, song, sbcglobal, rr	2	share, check, skin, gg, account
3	game, origin, sims, email, github	3	download, dude, bot, update, version
4	capture, type, key, unit, local	4	bro, great, watch, rep, hello
5	mail, password, username, unknown, user	5	thx, nice, test, hope, wow
6	member, wp, pro, stealer, clean	6	file, tnx, download, gonna, password
7	game, play, watch, best, good	7	wub, member, god, omg, gj
8	script, update, enemy, auto, download	8	tks, cool, awesome, wp, tyty
9	account, bol, legend, help, crack	9	good, script, mate, love, best
10	thx, nice, share, test, man	10	account, man, kappa, ban, lot

#	Construction C
1	nice, bro, tnx, tyy, gg
2	tks, ea, member, mail, info
3	account, game, link, crack, free
4	script, bol, update, download, game
5	thx, man, share, nice, good
6	file, download, bot, version, update
7	clean, stealer, rat, crypter, password
8	capture, account, member, gmx, key
9	wp, thnx, pro, unit, local
10	unknown, user, creed, assassin, unite

ground forum is about, as it shows a relation to accounts, leaks of credentials and games. Document construction B show less diverse topics as many of them can be categorised as expressing some gratitude. Thus, making this construction approach less suitable for a digital forensic investigation. Construction C can be suitable for understanding the different users within a forum, including their interest or possibly role on the forum. For example, people with a high proportion of expression of gratitude (thanks, thx, nice, etc.) in their messages might belong to the majority group of less technical skilled cybercriminals. Additional steps for removing unnecessary and less informative words may result in highlighting more skilled cybercriminals.

4.5 Conclusion

Cybercrime continues to be a treat to our economy and the general sense of justice. Law enforcement agencies can exploit OSINT to gather proactive CTI, which might make them more effective to combat cybercriminals. The challenge of OSINT comes from a lot of unstructured data which may result in unreliable information from automated processes. Our research shows that automated algorithms such as LDA must follow a set of requirements to reduce the vocabulary size and improve quality. We recommend repetitive preprocessing steps, e.g. continuously remove common words, until the result contains coherent and clear topics. Data cleaning is invariably an iterative process as there are always problems that are overlooked the first time around.

Contemporary related research mostly focuses on using topic modelling to get a quick overview of a lot of documents. This article tries to reduce the gap between the reliability of automated processes to make them applicable in digital forensic contexts. We identified three distinct ways users' posts could be constructed into documents, and each approach focused on different aspects: subject-centred, subject-user-centred and user-centred. While they did not produce any significant different result in keywords between topics; our result shows that more documents do not necessarily improve the quality of topics.

Data is key to piece together any criminal investigation, and more research is needed to further improve the reliability of automated processes/algorithms. Small changes in the input can produce an unreliable output, which in turn, forensic analysts can misinterpret. Thus, we need to move further than contemporary research's focus on using LDA to produce a general overview of a large corpus of text. For example, by applying techniques described in this article on real-world dark web underground forums. Furthermore, we need to design reliable and automated processes suitable in a digital forensic context. For example, to distinguish between individuals that produce advance tools for cybercrime and from those who simply are consumers of such tools. Finally, similar research as Chang et al. [7] should be conducted to analyse human-understandable topics and evaluation metrics (e.g. perplexity) in a digital forensic context.

4.6 Bibliography

[1] Rubayyi Alghamdi and Khalid Alfalqi. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015. http://thesai.org/Publications/ViewPaper?Volume=6&Issue=1&Code=ijacsa&SerialNo=21.

[2] Stig Andersen. Technical Report: A preliminary Process Model for Investig-

ation. preprint, SocArXiv, May 2019. https://osf.io/z4wma.

[3] Waheed Anwar, Imran Sarwar Bajwa, M. Abbas Choudhary, and Shabana Ramzan. An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution. *IEEE Access*, 7:3224–3234, 2019. https://ieeexplore.ieee.org/document/8558478/.

[4] David M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. page 30, 2003.

[5] Matt Bromiley. Threat Intelligence: What It Is, and How to Use It Effectively, 2016.

[6] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J. Buttery. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):19, December 2018. https://crimesciencejournal.springeropen.com/articles/10.1186/s40163-018-0094-4.

[7] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. page 10, 2009.

[8] CrowdFlower. Data Science Report, 2016. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.

[9] Isuf Deliu, Carl Leichter, and Katrin Franke. Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3648–3656, Boston, MA, December 2017. IEEE. http://ieeexplore.ieee.org/document/8258359/.

[10] R. Dover, M.S. Goodman, and C. Hillebrand. *Routledge Companion to Intelligence Studies*. Routledge Companions. Taylor & Francis, 2013. https://books.google.com/books?id=sTCwAAAAQBAJ.

[11] I. Kononenko and M. Kukar. *Machine Learning and Data Mining*. Elsevier Science, 2007. https://books.google.com/books?id=NUikAgAAQBAJ.

[12] Gaston L'Huillier, Hector Alvarez, Sebastián A. Ríos, and Felipe Aguilera. Topic-based social network analysis for virtual communities of interests in the Dark Web. page 9, 2011.

[13] Matthew Moran. Big data brings new power to open-source intelligence, 2014. http://theconversation.com/big-data-brings-new-power-to-open-source-intelligence-26554.

[14] Kyle Porter. Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. *Digital Investigation*, 26:S87–S97, July 2018. https://linkinghub.elsevier.com/retrieve/pii/S1742287618302020.

[15] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker. Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, 34(4):1023–1053, 2017. https://www.tandfonline.com/doi/full/10.1080/07421222.2017.1394049.

[16] Alexandra Schofield, Måns Magnusson, Laure Thompson, and David Mimno. Understanding Text Pre-Processing for Latent Dirichlet Allocation. page 4, 2017.

[17] Ryan Williams, Sagar Samtani, Mark Patton, and Hsinchun Chen. Incremental Hacker Forum Exploit Collection and Classification for Proactive Cyber Threat Intelligence: An Exploratory Study. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 94–99, Miami, FL, November 2018. IEEE. https://ieeexplore.ieee.org/document/8587336/.

[18] Hamid Akın Ünver. Digital Open Source Intelligence and International Security: A Primer. 2018. http://rgdoi.net/10.13140/RG.2.2.16242.56000.

Chapter 5

Article IV - Identifying proficient cybercriminals through text and network analysis

Jan William Johnsen and Katrin Franke. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019. pp. 1-7.

Abstract

A few highly skilled cybercriminals run the Crime as a Service business model. These expert hackers provide entry-level criminals with tools that allow them to enhance their cybercrime operations significantly. Thus, effectively and efficiently disrupting highly proficient cybercriminals is of a high priority to law enforcement. Such individuals can be found in vast underground forums, though it is particularly challenging to identify and profile individual users. We tackle this problem by combining two analysis methods: text analysis with Latent Dirichlet Allocation (LDA) and Social Network Analysis with centrality measures. In this paper, we use LDA to eliminate around 79% of hacker forum users with very low to no technical skills, while also inferring the forum roles held by the remaining users. Furthermore, we use centrality measures to identify users with hugely popular public posts, including users with very few public posts who receive much attention from their peers. We study various preprocessing methods, wherein we achieve our results by following a series of rigorous preprocessing steps. Our proposed method works towards overcoming current challenges in identifying and interrupting highly proficient cybercriminals.

5.1 Introduction

A few very skilled cybercriminals sell their criminal spoils and technical knowledge to the larger underground market population, through the Crime as a Service (CaaS) business model [1, 8, 10, 18, 24]. Such criminal activities contribute to the estimated cybercrime cost of (at least) 45 billion US dollars in 2018 [22]. Focusing on identifying and taking down a few proficient criminal actors has desirable benefits [8, 9], such as causing more substantial disruption on the CaaS business model. Therefore, both researchers and practitioners alike are developing methods to identify those proficient actors.

In this paper, we study the combination of methodologies from Natural Language Processing (NLP) and Social Network Analysis (SNA) to identify the more prominent and popular users in such underground forum marketplaces, more specifically: Nulled.io hacker forum. By utilising complementary methods, we explore *what* users are talking about and with *whom* they communicate. Our novel approach can exclude around 79% of underground market users who demonstrate low to no technical skills and remove them from further analysis. This reduction in users increases investigation efficiency and effectiveness by reducing algorithms' execution time and focuses the analysis on prominent cybercriminals.

To achieve a reduction in users, we understand that similar users tend to use combinations of equivalent words when they write posts on the forum. Thus, we can exclude users who exclusively express gratitude towards others when we analyse all forum posts made by individual users and assign specific topics to each user. Removing lower-skilled users allows further analysis steps to focus on skilled cybercriminals. Furthermore, we can infer users' role on the forum, because user groups (such as administrators and reverse engineers) use different assortments of words.

We can also identify users with popular public posts using centrality measures; not only in terms of their connectivity but also those users with few public posts that generate a high intercommunication in the forum. A challenge with underground forum marketplaces is that they are imbalanced datasets where a few administrators and skilled users serve many thousands of users. It is, therefore, necessary to follow a series of rigorous steps to achieve our results. In this paper, we present this series of rigorous preprocessing steps to reduce the number of users to investigate in an underground forum.

5.2 Previous work

We refer to our previous article [12] for an overview of related work concerning SNA and network centrality measures, as we focus this section on topic modelling. The topic modelling algorithm LDA ability to find unobserved groups (i.e. identify latent topics) makes it applicable to many areas. LDA is typically used to get an overview over a large corpus of text, but can also be used to identifying key actors and hacker assets in underground forums. For example, Porter [25] utilised LDA to find keywords and trends over a period in darknet markets subreddits.

Although it is challenging to investigate underground forums due to the combination of public, restricted and private sections, Motoyam et al. [20] found that a user's reputations come from being publicly active. Their findings motivated our approach to looking at publicly available posts to identify highly proficient actors. In contrast, Marin et al. [18] used a reputation system in the underground forum to validated their results. They showed that hybridisation of features could identify key hackers more precisely, which we also suggest by combining multiple methods to explain their independent results better.

Researchers [1, 2, 23] have examined various features to identify expert hackers, such as hacker assets (number of attachments), speciality lexicons (vocabulary of a person) and forum involvements (metrics such as number of threads, posts and attachments). They found that older forum members and very active members typically have a higher reputation than other users. Pastrana et al. [23] applied a combination of SNA, topic modelling and clustering to identify features to understand better who is at risk of becoming involved in criminal activities.

Features such as keywords have been used by Benjamin et al. [3, 4, 5] to explore and understand hacker language and to identify keywords for potential threats. While researchers like Li et al. [13, 14, 15] have tried to use sentiment analysis (interpret positive, negative and neutral emotions in the text) to identify and profile top malware and carding sellers. They also mentioned that active hackers comprise of those who are more actively involved in hacker community discussions.

Samtani et al. [26, 27, 28] have utilised SNA techniques to identify key hackers and explore Cyber Threat Intelligence (CTI) and hacker assets. More specifically, Samtani et al. [26] looked at particular classes of networks (bipartite and monopartite) and limited key hacker identification using only betweenness centrality measure. Furthermore, they also focused their research [27, 28] on thread starters, utilising LDA to understand the topic characteristics of hacker assets and identify hacker tools, such as crypters, keyloggers, web and database exploits. In contrast, Nunes et al. [21] tried to find zero-day exploits and vulnerabilities.

Marin et al. [17] used clustering to identify hacker product categories and found that many (nearly) identical items are posted across multiple marketplaces, sometimes under the same vendor username. In comparison, Huang and Chen [11] used clustering to find key members and their roles in the cyber fraud value chain. They used SNA to identify communities and assume key members generally post more content and receive more replies compared to other members.

A challenge with previous research is that many of them gather data through web crawling underground forums. This approach closely mimics real law forensic investigations, but it encounters the same problems from anti-crawling techniques, and only parts of the underground forum may be accessible. Pastrana et al. [24] rectified this problem of relying on incomplete or outdated datasets by capturing data from underground forums resulting in a dataset called CrimeBB.[1] Our access to a leaked hacker forum database has two unique opportunities: (i) we can evaluate the performance of our proposed method on the best-case scenario and (ii) we have some type of ground truth to base our evaluation with access to all data.

Other issues are that they only look at very tiny fractions of a network (e.g. thread starters or bipartite networks) or rely on knowledge from cybersecurity experts, knowledge which may not generalise to other underground forums. Additionally, many of them assume that higher reputations can indicate proficiency, which can be artificially increased by being a very active member or accumulate reputation over time. For example, forum administrators must be quite active to manage their forums, and they do not necessarily have to possess technical knowledge except for running a website. Thus, we disagree that proficient users are those who are extroverts and communicates a lot in the forum. Proficient hackers can also be introverts which only make a few posts with a high impact on the forum.

5.3 Methods and material

Nulled is a hacker forum found in the deep web, that facilitate the brokering of compromised passwords, stolen bitcoins and other sensitive data. We chose Nulled as we believe they closely resemble criminal underground forums, and we had a unique opportunity accessing this leaked database. We have access to all data contained in the database, which allow us to have some type of ground truth to verify our results. The Nulled dataset contains details about 599 085 user accounts, 800 593 private messages and 3 495 596 public messages.

Figure 5.1 depicts the overall process of preprocessing and analysing our dataset. We incorporate many standard text preprocessing practices found in NLP and related literature. We also introduce some measures of our own because some

[1] Available after legal agreement, which has not been pursued in this moment of time.

type of data was producing noise in the analysis. These measures are specific to this type of dataset, which includes removing e-mail and password combinations and removing Uniform Resource Locators (URLs), HyperText Markup Language (HTML) and BBcode tags. Where related research only employ stemming/lemmatisation to normalise words in their data, we actively attempt to normalise it further. We normalised the text by extracting words with repeating patterns and replacing them with their intended word – this section details these preprocessing and analysis steps.

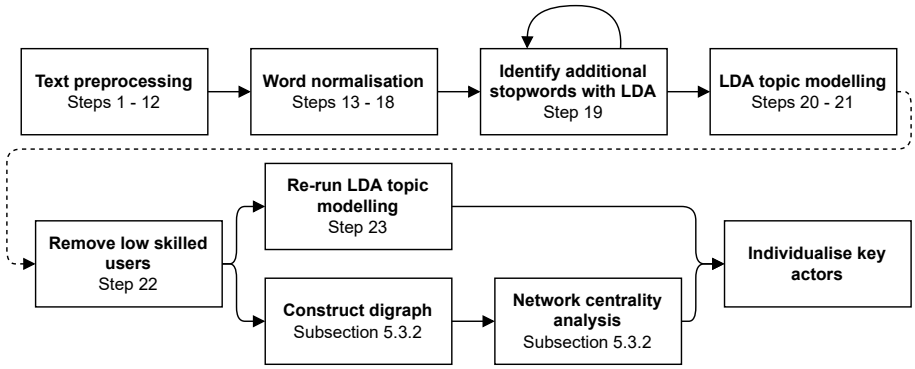


Figure 5.1: Process model

5.3.1 Latent Dirichlet Allocation

We start this section with a general introduction to the LDA algorithm. Then, the following two subsections will describe: (i) our LDA preprocessing steps on forum posts in more detail and (ii) generate an LDA model to find new words to filter out.

LDA is one of the more popular algorithms in NLP because it is typically more effective and generalises better than other algorithms. LDA [6, 16] is a statistical model, commonly used to categorise a set of observations (i.e. text) into unobserved groups that explain why some parts of the data are similar. The result is a set of human-interpretative topics from a document corpus. The generalisability property is particularly beneficial, so our proposed method may generalise to more specific domains such as those of underground forums.

LDA is a way of 'soft clustering' using a set of documents and a pre-defined k number of topics. The two other hyper-parameters α and adjust two Dirichlet distributions. These Dirichlet distributions adjust the LDA model document-topic density and topic-word density, respectively. Thus, LDA allows for a nuanced way of categorising documents, as each document has some probability of belonging to several topics. The biggest problem with LDA is the lack of extracting semantic-

ally meaningful information [7]. However, a human analyst can deduce what the relation of the topics by studying the word-distributions.

The challenges of analysing a dataset such as Nulled includes how frequent users write with a spoken language. This means text tends to include more repetitions, incomplete sentences, slang expressions (such as 'gonna') and using repeating words/characters to add emphasis. They also tend to write short messages, such as a simple 'thanks' to express gratitude or appreciation. Finally, most users are non-native English speakers and sometimes use non-English words or frequently misspell words; either because they do not know how to spell certain words or they simply ignore misspellings.

Latent Dirichlet Allocation dataset preprocessing

We followed standard topic modelling preparation steps, while also adding a few of our own to accommodate for this type of dataset. We ensured to replace anything that was removed with whitespace to guarantee that words are not unintentionally combined. The following list is the order of our initial preparation steps:

1. Remove extra spaces and convert to lowercase.
2. Remove all URLs.
3. Remove all 'e-mail:password'-combinations.
4. Remove all e-mails.
    - Leaking of credentials is one primary focus area of this hacker forum. Consequently, it contains large dumps of e-mail and passwords which dominated (esp. mail hosting domains) the topics.
5. Remove all HTML tags (including its contents).
6. Remove all HTML entities (e.g. ' ').
7. Remove newline/tabulator characters ('\n', '\r', and '\t').
8. Remove all BBcode tags.
    - HTML tags and entities and BBcode tags are particular for this type of dataset. We removed them as they were of little use; however, it could be useful if one wants to preserve, e.g. attack indicators.
9. Remove symbols, including numbers.
10. Run lemmatisation.

- We chose lemmatisation instead of stemming because lemmatisation considers the morphological analysis of words. In other words, it considers the structure and part of words to find their root form.

11. Remove stop words.
    - Removing over 700 of the most common English stop words, such as ‘able’, ‘come’, ‘do’ and ‘during’.
12. Remove any extra white space.

Users on this hacker forum frequently used exaggeration and abbreviations when writing. They show this by repeating characters (e.g. ‘niceeee’ or ‘goooood’) and words (e.g. ‘tytyty’, short for ‘thanks’). Normalising this data could allow us to distinguish between low and high skilled cybercriminals more accurately. Repetition of characters and words, including word misspellings, is an obstacle for LDA [21] – as word variations exist as separate words during the analysis. lemmatisation mitigates many issues of word variations because it converts inflectional forms of words such as ‘studying’ and ‘studies’ to the base form of ‘study’. However, lemmatisation fails to fix the issue word variations from misspelling and repetition.

To solve the challenge with word variations from repeating characters and words, we extracted and replaced those repeating patterns in two different processes. The first part extracts whole words by looking at repeating patterns while minimising the chance of replacing words erroneously. An example is to avoid illogical changes such as ‘r*emem*ber’ to ‘rember’.

After extracting words with repeating characters or words, we found their shortest expressable form by allowing repetitions to occur a maximum of two times. For example, ‘goooood’ would have the short form ‘good’ and ‘tytyty’ would have the short form ‘tyty’. This transformation allowed us to gather repeating patterns of varying length into a collective short form. Finally, the extraction part grouped common short forms and sorted them in descending order of frequency.

13. Extract all words with repeating characters.
14. Extract all words with repeating words.
15. Identify the short form of all extracted words.
16. Group and count the short form words and sort it by frequency in descending order.

17. Inspect and identify replacement words for the first 1000 words.
18. Replace those 1000 new short words with the original words in the dataset.

Our approach to finding the shortest common words made it easier and more effective to replace words with similar repeating characteristics. We manually inspected and changed the first 1000 shortest common words to ensure data quality and control in our experiment. Manual examination allowed us to avoid changing words erroneously. We would either (i) replace the short word with the intended word when it was apparent or (ii) replace it with itself if the intended word was unknown or it was already an existing word.

Finding replacement words for the 1000 most frequent short words allowed us to replace 73% of the words found in that list (around 17% of words in the original dataset). The second part of the process is to use those 1000 replacement words to change the original words identified in the first part. Additionally, we replaced many common synonyms and abbreviations without any repeating patterns, for example, 'ty', 'thx', 'thnx', 'merci', and 'gracias' was replaced with 'thanks'.

Running the Latent Dirichlet Allocation analysis

The previous subsection explained more general preprocessing done to forum posts. However, running the LDA algorithm on this dataset can still produce words that do not provide any significant meaning to our analysis. For example, words like 'haha', 'asp', 'tk', 'content' and 'href' are words without any meaning. To further ensure data quality in our experiment, we had to run the LDA algorithm a few times to identify these words to remove them from the final analysis.

The LDA model is very affected by the document input construction. We identified two distinct document construction approaches: (i) one document is a concatenation of all messages from a single user or (ii) each message is a single document. We call these approaches concatenated and singular, respectively. We keep the hyper-parameters k, α and identical between each construction approach when running the LDA analyses.

We analysed both approaches and concluded that the singular approach had topics with more mixed words and therefore provided less coherent topics for a human analyst. Therefore, we focus the result section on showing the outcome from the concatenated approach. However, the singular approach was suitable when individualising users, as it gave LDA more documents to learn the underlying topics.

19. Run LDA analysis a few times (four times in our experiment) in order to identify additional stop words and remove them from future analyses.

After these steps, we had a suitable LDA model that can identify which topics users were primarily sending messages about. The identification process resulted in a floating-point array for every user. Each element in the array shows the similarity between the user's messages and every topic. Array elements are the sum of similarities and averaged by the number of sent messages. We also convert this array to a binary format, where topic(s) with any positive float value receives a one, and the other topics are set to zero. In other words, the threshold is anything above zero.

We continue by labelling each topic in the LDA model to find the category that best explains those word combinations. We assume that cybercriminals with low technical skills are very dependent on others with higher kills. They will, therefore, more likely, be consumers of information rather than producers of it. This should come from the way they communicate, such as expressing relatively more gratitude in their public posts. Thus, we can distinguish between users who are pure consumers and express gratitude with everyone else.

20. Categorising LDA model topics, distinguishing between the expression of gratitude versus reverse engineering.
21. Identify which topics each user mainly post messages.
22. Distinguish between users who purely express gratitude with everyone else.
23. Re-run LDA and network centrality analysis using the remaining users that are of interest.

Notably, the result of LDA should be improved when lesser skilled forums users are removed from the dataset, as this gets rid of much junk. Network centrality analysis could also benefit from this, possibly by highlighting different key actors. The main benefit for network centrality measures is fewer forum users to go through, instead of wasting time considering lesser skilled individuals. Thus, the result should be attained faster and feasible for investigators finding secondary targets to take down.

5.3.2 Centrality measures

Network centrality measures are graph-based analysis methods found in SNA, used to identify important and influential individuals within a network. However, public forum posts do not have any natural way of constructing a directed graph (digraph). As the digraph construction will affect the centrality analysis results, we need to decide on how to best model the interaction between users. For example,

should edges' direction go out from the thread starter or in towards them? Constructing accurate graphs from these forums are non-trivial, yet essential, to avoid meaningless centrality measures and attribute incorrect significance to users [1].

We denote a set of users V and a set of posts E, as the vertices and edges in a digraph $G = (V, E)$. We chose to construct digraphs with edges from a replying user to the author of forum threads. More specifically, there is a direct edge (v, v'), when user v reply to a thread started by v'. This edge represents an interest to respond on a public thread. We acknowledge this construction method does not truly reflect how forum users interact with other users, as forum threads can be used with multiple purposes, such as asking other users for advice or having unrelated discussions.

We evaluate five popular centrality measures for digraphs: in-degree (C_{D^-}), out-degree (C_{D^+}), betweenness (C_B), closeness (C_C) and eigenvector (C_E). They differ in their interpretation of what it means to be 'important' in a network. Thus, some vertices in a network will be ranked as more important than others, as vertices and edges affect the centrality value. We chose these centrality measures as they are in popular forensic investigation tools such as IBM i2 Analyst's Notebook. We refer the reader to SNA books, e.g. McCulloh et al. [19], for formulas and detailed explanations of centrality measures.

5.4 Experiment and results

We have two goals with the experiment: first, to distinguish the majority from the minority (i.e. consumers of content with those who produce it) and secondly, to find out better which individuals to focus investigations resources.

5.4.1 Latent Dirichlet Allocation topic results

We identified two distinct LDA document construction approaches. The first approach concatenated all messages from individual users into one document, while the second approach treated each message as a document. Due to the page limits, we are only showing the results for the concatenated approach.

We observe that the concatenated approach in Table 5.1 gives more coherent groups of words, compared to the singular approach. For example, topic 1 talks about various popular games (possibly sharing of e-mail/username and passwords to get access to these games); topic 2 confirms that something is working (possibly cracks); topic 3 captures various hacker tools, such as Remote Access Trojan (RAT), crypter (encrypt, obfuscate and manipulate malware, to make it harder to detect by security programs) and stealer (theft of some type of information); and topic 7 and 10 express some appreciation of someone's work or thank them for

sharing. The singular approach was producing less intelligible results because it has a broader mix of words per topics. For example, appreciation words were distributed among several topics.

Table 5.1: Concatenated topics for all users

Topic #	Keywords
Topic 1	game, origin, email, sims, capture, key, battlefield, edition, password, country, unit, username, type, fifa, command
Topic 2	work, download, account, crack, post, file, game, time, link, help, find, update, bot, free, check, script, guy, people
Topic 3	stealer, rat, crypter, tool, phisher, scan, binder, beta, spam, user, lsie, module, power, password, ddos, public
Topic 4	script, bol, update, download, legend, enemy, work, champion, bot, version, auto, game, target, login, vip, combo
Topic 5	account, pm, sell, buy, bump, email, paypal, skype, skin, vouch, price, member, password, ban, level, information
Topic 6	add, attack, bot, troop, clashbot, play, password, base, download, set, version, bln, update, pro, feature, option
Topic 7	nice, good, work, man, share, brother, test, love, thank, hope, job, check, mate, wow, dude, lol, great, awesome
Topic 8	password, lol, xxx, minecraft, dragon, account, thank, class, brazzers, alex, fish, mofos, major, profile, cre, david
Topic 9	site, project, user, lol, password, smtp, unranked, round, username, location, try, game, modifier, key, kid, type
Topic 10	thank, test, nice, brother, work, lol, man, good, account, please, much, very, check, rt, wow, rep, dude, game, help

For each forum user, we run their messages through the LDA model (Table 5.1) and output similarity for which of the ten topics they are most similar, as detailed in Section 5.3. Since this similarity is a binary value, we can exclusively distinguish users into two distinct groups: assumed high skill users and low skilled users with only appreciation posts.

There are a total of 299 719 unique users on this forum that had made at least one public post. Table 5.2 shows that the concatenating approach categorised 24% of them as assumed high-skilled users. This approach probably achieved fewer skilled users because it could group words used in similar situations more appropriately than the singular approach. Thus, the concatenated approach is the best technique to employ for forensic analysts as it reduces the amount of users most.

However, it is hardly the case that every one of the 24% is equally interesting for

Table 5.2: Number of users in high- and low-skill groups

Digraph	Appreciation topics	High skill	Low skill
Concatenated	Topics 7 and 10	62 859	201 924
Singular	Topics 1, 2, 5 and 6	94 755	170 028
	Topics 2, 5 and 6	101 439	163 344
	Topics 5 and 6	115 008	149 775

law enforcement. For example, few individuals can have skills that are very sought after by other cybercriminals or for some other reason are more attractive targets for investigations. Therefore, we need to employ network centrality measures to prioritise proficient users further.

5.4.2 Network centrality analysis results

Each forum user is assigned a unique and incremental Unique Identifier (UID); this value is a positive integer based on the order they registered as users. Furthermore, they receive a rank or group from their peers (typically assigned by moderators/administrators), which indicate their position in a forum. A variety of factors enable this group position to change during a user's lifetime. We obtained the forum groups for the Nulled forum, from their database tables, as seen in Table 5.3. This group overview gives us the ground truth to compare our findings against, which was previously lacking in many related works. The reader can refer back to Table 5.3 to find short names for groups used in this section.

Constructing a network of all the public posts would yield a network size of 299 702 vertices and 2 738 710 edges. However, constructing the same network using only high skilled users (Table 5.2) from the concatenated approach, reduced the number of vertices by 79.6% and edges by 71.7%, for this particular dataset. More specifically, this digraph had 61 127 vertices and 773 983 edges. Consequently, network centrality algorithms took a much shorter time to complete; particularly for betweenness centrality, which has a time complexity of $O(VE)$. This is significant as such digraph can now be used for time-critical investigations.

At first glance, Table 5.4 is almost identical in the ordering of who are most central individuals as our previous research [12]. The reason for this similarity is that we use the LDA results to extract a sub-graph, which retain a selected set of vertices and all their edges. More notably, we could identify a new user (with UID 574289) higher up in our result in this paper.

Many senior ranking members respond to lower-skilled cybercriminals for various reasons, such as helping them with guidance or answering questions, which artifi-

Table 5.3: Forum group overview

Group	Short name	# of members
Donator	Do	1
Moderators	Mo	1
Administrators	Ad	2
Legendary Reverser	LR	2
Senior Moderator	SM	3
VIP_Plus	VIP+	3
Reverser	Re	6
Legendary	Le	7
Contributor	Co	57
Royal	Ro	63
VIP	VIP	2245
Validating	Va	98837
Banned	Ba	111967
Members	Me	385891

Table 5.4: Concatenated top ten centrality results Nulled (forum group overview is found in Table 5.3)

UID	Group	C_{D^-}	UID	Group	C_{D^+}
15398	LR	0.294588	1471	Ad	0.016474
574289	LR	0.136505	8	SM	0.012891
1337	Le	0.100874	193974	Mo	0.01173
4	Re	0.08561	47671	Ba	0.011141
0	N/A	0.076759	334	Ba	0.010503

UID	Group	C_B	UID	Group	C_C
15398	LR	0.029145	15398	LR	0.535465
1337	Le	0.016342	1337	Le	0.472462
1471	Ad	0.012232	0	N/A	0.470536
334	Ba	0.008738	574289	LR	0.456597
574289	LR	0.0087	8841	Le	0.454961

UID	Group	C_E
15398	LR	0.207474
1337	Le	0.177821
0	N/A	0.148149
334	Ba	0.133743
22239	Le	0.13364

cially increase their network centrality scores. We can reduce this spurious effect of responding to many users and have a more accurate representation of important actors. Therefore, we see that results for betweenness, closeness and eigenvector centrality change the most from our previous research [12].

Centrality measures only provide a number to represent how central a user is, as seen in Table 5.4, which compares the relative centrality between users. However, it does not provide any other type of information, such as why they receive this score or their forum group. LDA provides us with this ability to inspect the overall posts made by individual actors, finding whether they are to interest in a continued investigation.

We re-run the LDA analysis for those users that the network centrality measures found as most central. We train this new LDA model using the hyper-parameter values ($\alpha = 0.05$ and $\eta = 0.05$ and $k = 3$) and using the singular approach mentioned in Section 5.3.

Table 5.5 shows a selection of the top actors from each centrality measure. We can distinctly see that UID 1337 and 1471 talks about some administration-related topics. While UID 15398 and 574289 talks about some reverse engineering-related topics. Thus, the individualised LDA results have a correlation with the groups these users has been assigned. Similar to Samtani et al. [26], our result also indicates that many key/central members are those most senior and longest participants in their community, due to their low UID numbers.

Table 5.5: Sample of central individual's topics

UID	Group	Topic	Words
1337	Le	1	data, update, work, thread, nulled, press
1337	Le	2	member, account, post, scam, download
1337	Le	3	ban, post, account, thread, thing, time
1471	Ad	1	game, forum, member, nulled, time
1471	Ad	2	deny, account, member, ban, pm, solve
1471	Ad	3	bump, ban, post, long, text, deny
15398	LR	1	loader, update, open, pipe, work, crack
15398	LR	2	member, rep, allow, hack, update, paypal
15398	LR	3	work, game, inject, bol, crack, nulled
574289	LR	1	bot, application, crack, download
574289	LR	2	bot, wrobot, feature, clashbot, download
574289	LR	3	version, download, troop, improve, add

5.5 Conclusion

We combine LDA and network centrality measures to identify proficient criminals, i.e. key hackers, in a real-world hacker forum. We can remove up to 79% of uninteresting users (who only wrote appreciating messages). This allowed us to focus our investigation on the remaining and presumed high skilled cybercriminals. This reduction allowed network centrality measures to run faster on a smaller sub-graph, as a lot of vertices and edges was removed. Furthermore, utilising a leaked hacker database allowed us to examine these methods in a best-case scenario, where we have the ground truth of forum user's groups.

Recall that our digraph was the product of who responded to which forum thread (i.e. a collection of related posts), instead of the actual relationship between users. Therefore, it is essential to note that centrality measures mostly identified users with viral threads. On the other hand, these threads become popular for a reason; for example, users can acquire the threads without gaining the technical skills to acquire them by themselves. When users post something that other users desire and become famous, they could be seen as having higher skills than the rest.

The contribution of our research is manifold. First, we proposed to study the underground economy through the lens of its participants, uniquely identifying the minority group of highly skilled cybercriminals. The minority group play a pivotal role in the CaaS and observing their behaviours enriches our understanding of it, allowing us to investigate central criminals further. Secondly, we developed advanced text-mining technique capable of identifying and profiling key underground hackers, and we experimented with evaluating its effectiveness.

For future work, we find it interesting to pursue a legal agreement to get access to CrimeBB dataset and repeat experiments, as presented in this paper. Although CrimeBB does not provide us with access to ground truth, it will provide us with other real-life underground forums to analyse.

5.6 Bibliography

[1] Ahmed Abbasi, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. Descriptive Analytics: Examining Expert Hackers in Web Forums. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 56–63, The Hague, Netherlands, 2014. IEEE. http://ieeexplore.ieee.org/document/6975554/.

[2] Victor Benjamin and Hsinchun Chen. Securing cyberspace: Identifying key actors in hacker communities. In *2012 IEEE International Conference on Intelligence and Security Informatics*, pages 24–29, Arlington, VA, 2012. IEEE. http://ieeexplore.ieee.org/document/6283296/.

[3] Victor Benjamin and Hsinchun Chen. Developing understanding of hacker language through the use of lexical semantics. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 79–84, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165943/.

[4] Victor Benjamin and Hsinchun Chen. Identifying language groups within multilingual cybercriminal forums. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 205–207, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745471/.

[5] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 85–90, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165944/.

[6] David M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. page 30, 2003.

[7] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. page 10, 2009.

[8] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2014. Technical report, 2014. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf.

[9] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2019. Technical report, 2019. https://www.europol.europa.eu/sites/default/files/documents/iocta_2019.pdf.

[10] Zhen Fang, Xinyi Zhao, Qiang Wei, Guoqing Chen, Yong Zhang, Chunxiao Xing, Weifeng Li, and Hsinchun Chen. Exploring key hackers and cybersecurity threats in Chinese hacker communities. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 13–18, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745436/.

[11] Shin-Ying Huang and Hsinchun Chen. Exploring the online underground marketplaces through topic-based social network and clustering. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 145–150, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745458/.

[12] Jan William Johnsen and Katrin Franke. Identifying Central Individuals in Organised Criminal Groups and Underground Marketplaces. In *Computational Science – ICCS 2018*, volume 10862, pages 379–386. Springer International Publishing, Cham, 2018. http://link.springer.com/10.1007/978-3-319-93713-7_31.

[13] Weifeng Li and Hsinchun Chen. Identifying Top Sellers In Underground Economy Using Deep Learning-Based Sentiment Analysis. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 64–67, The Hague, Netherlands, 2014. IEEE. http://ieeexplore.ieee.org/document/6975555/.

[14] Weifeng Li, Hsinchun Chen, and Jay F. Nunamaker. Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System. *Journal of Management Information Systems*, 33(4):1059–1086, 2016. https://www.tandfonline.com/doi/full/10.1080/07421222.2016.1267528.

[15] Weifeng Li, Junming Yin, and Hsinchun Chen. Identifying High Quality Carding Services in Underground Economy using Nonparametric Supervised Topic Model. page 10, 2016.

[16] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018. https://www.tandfonline.com/doi/full/10.1080/19312458.2018.1430754.

[17] Ericsson Marin, Ahmad Diab, and Paulo Shakarian. Product offerings in malicious hacker markets. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 187–189, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745465/.

[18] Ericsson Marin, Jana Shakarian, and Paulo Shakarian. Mining Key-Hackers on Darkweb Forums. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 73–80, South Padre Island, TX, 2018. IEEE. https://ieeexplore.ieee.org/document/8367642/.

[19] I. McCulloh, H. Armstrong, and A. Johnson. *Social Network Analysis with Applications*. Wiley, 2013. https://books.google.com/books?id=lDhuJLGRMPEC.

[20] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11*, page 71, Berlin, Germany, 2011. ACM Press. http://dl.acm.org/citation.cfm?doid=2068816.2068824.

[21] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 7–12, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745435/.

[22] Online Trust Alliance (OTA). 2018 Cyber Incident & Breach Trends Report, 2019. https://www.internetsociety.org/wp-content/uploads/2019/07/OTA-Incident-Breach-Trends-Report_2019.pdf.

[23] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. In Michael Bailey, Thorsten Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis, editors, *Research in Attacks, Intrusions, and Defenses*, pages 207–227, Cham, 2018. Springer International Publishing.

[24] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 1845–1854, Lyon, France, 2018. ACM Press. http://dl.acm.org/citation.cfm?doid=3178876.3186178.

[25] Kyle Porter. Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling. *Digital Investigation*, 26:S87–S97, July 2018. https://linkinghub.elsevier.com/retrieve/pii/S1742287618302020.

[26] Sagar Samtani and Hsinchun Chen. Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 319–321, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745500/.

[27] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. Exploring hacker assets in underground forums. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 31–36, Baltimore, MD, USA, 2015. IEEE. http://ieeexplore.ieee.org/document/7165935/.

[28] Sagar Samtani, Ryan Chinn, Hsinchun Chen, and Jay F. Nunamaker. Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*, 34(4):1023–1053, 2017. https://www.tandfonline.com/doi/full/10.1080/07421222.2017.1394049.

Chapter 6

Article V - On the feasibility of social network analysis methods for investigating large-scale criminal networks

Jan William Johnsen and Katrin Franke. [Manuscript submitted for publication]

Abstract

Cybercrime exists in a highly organised form, with a minority of criminals selling their technical expertise to the majority without such abilities. Underground forums are significant hubs in this regard, and they are an essential channel for Crime as a Service (CaaS) suppliers. Law enforcement must target proficient actors within the minority population to disrupt the CaaS business model. The challenge, however, is finding key criminal actors – with high technical expertise and prominence – among several hundred thousand underground forum users. Law enforcement investigators also need computer support to identify key actors in large criminal networks. Due to a lack of better means, they use off-the-shelf Social Network Analysis (SNA) methods, such as centrality measures. This study analyses whether these existing methods are appropriate for the application of criminal network analysis and if they indeed find key criminal actors.

Our research on large-scale criminal network analysis shows that existing centrality measures do not identify high-profile criminals, but instead, they identify 'talkative' individuals. Despite the well-proven application of centrality meas-

ures in other application domains, their use for identifying key criminals leads to procedural inaccuracy because only individuals with high communication frequencies are recognised by these SNA methods. Our result shows that actors with high communication frequencies are mainly linked to users such as administrators and moderators, so we must not equate high network centrality with high criminality. We prove the insufficiency of existing methods for the new application of finding skilled cybercriminals in large-scale and distributed criminal networks and propose an improved method to deal with the new application. We propose a novel method using text analysis to group an underground forum population into skilled technical contributors (the minority) and the remaining low-skilled forum users (the majority). Our method identifies 93% and 89% of the underground forum population as low-skilled actors, which allows law enforcement to focus their resources in further investigations on the remaining minority group.

6.1 Introduction

Cybercriminal underground forums gather like-minded individuals who pursue illicit activities such as malware/exploit development, vulnerability disclosure, hacker tools exchange and distribution, and trading of materials, products and services [32, 10, 1]. These underground forums are platforms for the Crime as a Service (CaaS) business model, where a minority of criminal individuals or groups sell, let or give away their technical skills to the majority of less experienced cybercriminals [9]. The CaaS significantly reduce buyers' need for knowledge and expertise to conduct successful cyber attacks. Consequently, researchers, law enforcement and intelligence agencies target key actors within the minority population to stop the CaaS business model.

Identifying key actors in small criminal networks can often be done by eye. On the other hand, criminal underground forums are far more complex and involve hundreds of thousands of members and millions of connections. In such cases, forensic investigators and intelligence agents use off-the-shelf tools – which employ Social Network Analysis (SNA) – to obtain a grasp of the members and relationships within the criminal network [12]. SNA-based methods focus on network centrality [36] to find a criminal's prominence and determine the relative importance or influence that members possess within a network [12]. Network centrality is also used by related work as a way of identifying key actors (e.g. leaders) of criminal organisations [41].

As seen in Figure 6.1, forensic science (forensics) is the intersection between application, technology and methodology [11]. Law enforcement applies SNA centrality measures to find the most important or central actors in a network; those actors with more opportunities and fewer constraints. Using centrality measures

might have been sufficient when the application was finding key actors in small, physical and hierarchical criminal networks. However, today's underground forums have several hundred thousand loosely connected actors. The new network characteristics have changed the application domain for SNA, which requires a re-evaluation of the overall investigative methodology [39]. A significant drawback formerly consisted of the lack of scientific work that validated SNA centrality measures methods for the new application of large-scale criminal network analysis.

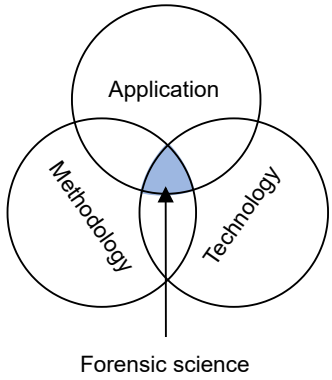


Figure 6.1: Franke and Srihari [11] define forensic science as the cross section of technology, methodology and application. This figure is adopted from their paper. The same technology and method will work fine for one application domain, but it might not work for another application domain. Therefore, we need to cross-validate the technology and methodology to solve challenges in other application domains [39].

Our study addresses this issue, which is urgently required to ensure that law enforcement uses investigative methods correctly. Moreover, the use of unvalidated methods in legal rulings can lead to reduced procedural accuracy and violate the right to a fair trial and human rights [38]. We increase the knowledge of using centrality in forensics by accurately describing how they identify 'key' criminal actors. We can validate the results in this study because of our unique access to two real-world criminal underground forum datasets. The complete datasets give us access to some ground truth information, which has rarely been done by other studies [24].

This research article presents two experiments. The first experiment, detailed in Subsection 6.4.1, conducts a hypothesis testing of the correlation between forum activity and centrality measure ranks. In this part of the article, we demonstrate how earlier network centrality research results can misrepresent results in newer research. The second experiment, detailed in Subsection 6.4.2, is where we present a novel method using Natural Language Processing (NLP) to enable law enforcement and researchers to largely separate the majority and minority population of underground forum users. Our method significantly reduces the number of actors to evaluate. The benefits are that: (i) further research can be aimed at better understanding technically skilled members; and (ii) enables law enforcement to focus

their limited resources on high-impact cybercriminals.

The rest of the paper is organised as follows: Section 6.2 explains this work in relation to related and previous work. Section 6.3 details the experimental process model, methods used in our work, and the preprocessing done to the material. Section 6.4 presents our experiment, exhibits and discuss the corresponding results. Finally, Section 6.5 recapitulates our study and includes suggestions for further work.

6.2 Previous work

Existing methods for identifying key actors fall into two main categories: content-based and social network-based analysis. Content-based analysis refers to mining data generated by users of underground forums, such as activity level and content quality. SNA, on the other hand, can model and analyse user interactions in underground forums. Related SNA research targeting critical network actors focuses on the concept of centrality to identify actors who are somehow central, vital, important, key, or pivotal in a criminal network [37, 36]. Centrality is interpreted differently by the centrality measures to highlight distinct actors as important. The centrality measures rank actors from high to low, where important actors tend to have high centrality scores [24, 1].

Centrality measures typically used to identify important actors in criminal networks are: degree, betweenness, closeness and eigenvector. Degree centrality is an indicator of importance, influence, control, and can signify visibility [27, 5]. Actors with a high degree centrality are considered to influence a large number of people and are capable of communicating quickly with actors in their neighbourhood [28]. Moreover, actors with a high degree centrality are more likely to be arrested [27], found guilty, receive longer sentences, and larger fines [3]. Therefore, actors with a high degree centrality are often considered leaders, experts, or hubs in a criminal network [42].

Betweenness centrality measures information flow through individuals. It is an indicator to show whether an actor plays the role of a broker or gatekeeper in a network. Broker exchanges between two other actors, and a gatekeeper controls (e.g. withhold or distort) information passing between actors [23]. Actors with high betweenness centrality were less likely to be arrested because they were less likely to be part of the criminal network [27]. Closeness centrality measures how easy it is for one actor to be able to communicate with others in the network [23]. Thus, actors with high closeness centrality can reach most or all other actors in the network [5].

Centrality measures have been applied in a diversity of research domains, for in-

stance, analysing the structure of terrorist [21] and criminal networks. Centrality measures are robust in the presence of noise [28] and empirically measure the structural importance of a single actor in a network. Moreover, researchers believe network centrality leads to an 'evidence-based understanding of the overall structure of a criminal network and the positioning of a variety of key actors [27]'. Another use of network centrality is the removal of key actors to disrupt criminal networks and underground forums [30] and decrease the ability of the criminal network to function normally [26, 37].

Pete et al. [30] (similar to our earlier work [18]) identify key actors using network centrality measures and got a similar result: (i) centrality measures were overlapping when ranking some actors higher than others, and (ii) key actors hold administrative positions in the forum and are active forum participants [13]. Similar to what we do in this research article, Pete et al. [30] showed a significant positive correlation between users' centrality and their posting activity by using Pearson's correlation test. We provide a more in-depth and novel understanding of centrality measures and correlation with the number of received replies in Subsection 6.4.1.

Table 6.1 contains a list of previous work that analyse criminal networks. This table shows that our study is significantly more complex and more extensive when comparing the network sizes; previous work had a median network size of 282 actors. The work of Pastrana et al. [29] is the only research that comes close in regards to the number of actors studied. The listed research articles use various resources – typically data from the police – to construct the criminal network. On the other hand, we have access to the complete network and the ground truth from two leaked underground forums.

Previous work often represents a criminal network by constructing undirected or directed graphs, where vertices are the forum members and edges their interactions. The type of interaction takes on different forms (e.g. forum posts [29, 30, 24, 18], co-conspirators [43, 5] or other communication forms [41]) depending on the relationship researchers are trying to model. It is commonly referred to as an interaction network when forum threads and posts are modelled using graphs. Thread starters initiate interactions, and any forum member replying to the thread will take part in the interaction [30]. More formally, researchers define a (un)directed edge (v, u) if there is a post/reply from member v to thread starter u. Researchers have minor variations in their graph construction to account for different posting times [24] or type of data sources [23, 5]. This research article will use an interaction network to model forum communication and analyse the resulting graphs using network centrality measures.

Table 6.1: Comparing network sizes from previous work

Research article	Nodes	Edges
Krebs [21] (2002)	19	
Lu et al. [23] (2010)	23	
Memon [26] (2012)	62	153
Baker and Faulkner [3] (1993)	78	
Grisham et al. [13] (2017)	100	562
Hardin et al. [15] (2015)	156	
Morselli [27] (2010)	174	
Xu and Chen [42] (2003)	164 – 744	
Diesner and Carley [7] (2005)	227	
Holt et al. [16] (2012)	336	
Décary-Hétu and Dupont [8] (2012)	771	
Xu et al. [43] (2004)	924	
Abbasi et al. [1] (2014)	4 576	
Samtani and Chen [33] (2016)	6 796	
Pete et al. [30] (2020)	22 – 16 401	57 – 624 926
Pastrana et al. [29] (2018)	572 000	
Johnsen and Franke [17] (2017)	599 086	371 002
	599 085	2 672 147
Johnsen and Franke [18] (2018)	75 416	319 935
	33 647	98 253
	299 105	2 705 578
Johnsen and Franke [20] (2020)	94 832	490 268
	62 933	794 868
This study	21 432	64 938
	299 701	2 741 464
	185 806	1 794 947

6.3 Material and methods

Assessing the feasibility of centrality measures as forensic techniques requires high-quality data. Good data reduces the factors that can affect the assessment and increase conclusion accuracy. We use two leaked underground forums in our experiment and the well-known Enron corpus. The choice of the Enron dataset allows other researchers to reproduce our findings, while our access to leaked datasets gives us a unique opportunity with three advantages: (i) we can study real-world and criminal communities with actors of various technical skill levels; (ii) the data is stored in a structured manner, which makes extracting and prepro-

cessing easier compared to other data collection methods; and, more importantly, (iii) we sit on the ground truth with all the user-generated information (such as private and public messages and forum roles), which allows us to assess network centrality measures' results better.

The leaked underground forums are Nulled.io and Cracked.to [10]. They are both hacker communities that facilitate the brokering of compromised passwords, provide tools and leaks, and generally act as marketplaces for services, products and materials. Table 6.2 provides a short overview of these two datasets. Nulled is the older forum with almost four years of data, while Cracked is the younger forum with over one year of data. The benefit of analysing two similar – yet distinct – datasets is that we can evaluate the generalisability of our methodology proposed in Subsection 6.4.2.

Table 6.2: Statistics over dataset users and public posts

Dataset	Users	Posts	First post	Last post
Enron	75 416	252 759	30 Oct 1998	3 Feb 2004
Nulled	599 085	3 495 596	26 Nov 2012	6 May 2016
Cracked	321 444	2 459 543	19 Mar 2018	21 Jul 2019

The types of communication in e-mail and forums are obviously different. For example, e-mail communication is more similar to one-to-one direct messages. At the same time, public forum posts are one-to-many or many-to-one, depending on how one would model the communication as a graph. The difference in the result for these two communication types is negligible. A similar result strengthens our argument when we criticise using network centrality to find 'key' criminal actors in a criminal network. The result and criticism are discussed further in Subsection 6.4.1.

6.3.1 Experimental setup

The two experiments detailed in this research article require separate process models for data preprocessing, structuring, analyses and result interpretation. Figures 6.2 and 6.3 show the experimental process model for (i) hypothesis testing, and (ii) using communication contents to eliminate low-skilled forum users, respectively.

Figure 6.2 illustrates the process model for testing our hypothesis: that there is a correlation between the number of replies thread starters receive and how important network centrality measures assess them. In this experiment, we construct a directed graph (digraph) by extracting information from all three datasets about users and how they interact with each other. We carefully handle irregularities in

the extracted data, such as some users being identified by the same user identification (ID) or e-mail alias, as the case for the Enron dataset. This step is crucial for correct results because it ensures a vertex uniquely identifies every user in the digraph. Furthermore, SNA requires its preprocessing steps by removing isolated vertices and self-loop edges. Subsection 6.3.2 details this process, such as data preprocessing and choice of methods.

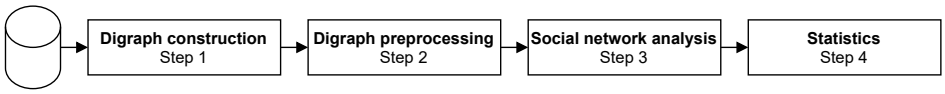


Figure 6.2: Process model for evaluating centrality measures

Figure 6.3 illustrate the iterative process model for our proposed approach to removing low-skilled users from the dataset. In this experiment, we focus on analysing the two underground forums, Nulled and Cracked [10]. We begin the process with standard text preprocessing steps, as well as replacing repeating words and characters to normalise the text further [20]. We generate topic models with Latent Dirichlet Allocation (LDA) or Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM) depending on the iteration. The process continues by identifying users' topic distribution, attained by calculating the similarity between topics and users' posts. The final step is to identify appreciation topic(s) by evaluating topics keywords and users' topic distribution and remove users who have primarily posted in those selected topics before continuing to the next iteration. Section 6.3.3 explains every step of this process in more detail.

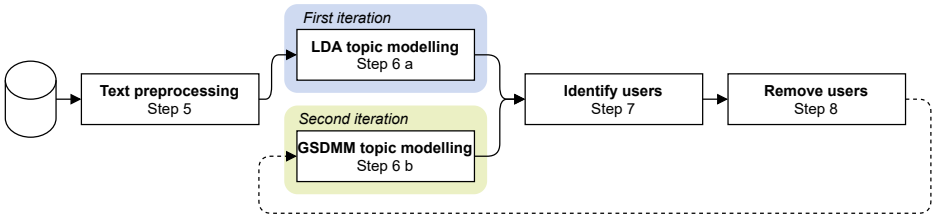


Figure 6.3: Process model for our novel approach

6.3.2 Digraph construction, SNA preprocessing and analysis

Step 1. Digraph construction

Constructing accurate digraphs of communication interactions is essential to avoid erroneous social networks with meaningless centrality measures [1]. This means that we have to account for irregularities in the data, such as e-mail aliases, which

are one or more alternative e-mail addresses that forward messages to another address. Consequently, SNA analysis produces inaccurate centrality results if employees are not uniquely represented by one vertex. 97 out of 149 known employees had one or more e-mail aliases.

It is out of this study's scope to account for all the e-mail aliases inside the Enron corpus. This is better addressed in other research articles or when it is mission-critical for the application, such as in police investigations. However, we normalised the e-mail aliases for 149 known Enron employees and made them the 'focus of our investigation'. Concentrating on a smaller subset of vertices is analogous to police investigations, where the focus is on a few primary suspects, and the complete network structure is unknown. We construct a digraph by extracting a subset of the corpus where these 149 employees (uniquely identified by e-mail aliases) were present as either (i) senders or (ii) receivers of e-mail messages.

Nulled and Cracked datasets contain instances where multiple vertices can represent individuals. For example, individuals can register multiple user accounts on the forum, or the database has conflicting entries. Identifying actors who have registered or are using multiple forum accounts is outside this article's scope and should be addressed by other researchers. We focused our attention on the latter case, as some users had non-unique combinations of user ID numbers and usernames. These conflicting database entries happened for fifteen users in Nulled and three in Cracked. Although database conflicts were low, they had to be addressed to uniquely identify individuals in the social network. We achieved uniqueness of these values by: (i) grouping user ID numbers and usernames, (ii) removing non–duplicated username entries, and (iii) replacing the user ID number with the last occurring ID number for duplicates.

Step 2. Digraph preprocessing

Self-loop edges and isolated vertices are not of interest when analysing a network because users' relationship with themselves are uninteresting and isolated vertices has an infinite distance to other vertices [31]. Self-loop edges and isolated vertices are, therefore, removed before the analysis without affecting the result [22]. Table 6.3 shows the reduction in the number of vertices and edges when removing self-loops and isolates. This table shows that 597 vertices become isolated after removing 33 134 self-loop edges from the Nulled dataset, while 456 vertices were isolated after removing 14 091 self-loop edges from the Cracked dataset. These vertices were initially disconnected from the larger network component because they started forum threads without replies. However, they only become isolated after removing their self-loop edge according to the definition of isolated. Discon-

nected vertices indicate those users had started forum threads without any other interaction on the forum due to the lack of other edges.

A word of warning: removing self-loop edges and isolated vertices are okay in this situation because we check that it would not remove skilled cybercriminals. Other researchers and investigators must be cautious when following this exact approach because proficient cybercriminals can use it against them. For example, cybercriminals can actively create and reply to their forum threads but avoid posting to other forum threads. This is how cybercriminals can sell their services to other underground forum users yet remain hidden during the analysis because their vertices became isolated after removing self-loops and subsequently removed from the graph. However, the vertices will not become isolated if other users reply to their forum threads.

	Before preprocessing		**After preprocessing**	
Dataset	**Vertices**	**Edges**	**Vertices**	**Edges**
Enron	21 432	64 938	21 432	64 845 (-0.1%)
Nulled	299 701	2 741 464	299 104 (-0.2%)	2 708 330 (-1.2%)
Cracked	185 806	1 794 947	185 350 (-0.3%)	1 780 856 (-0.8%)

Table 6.3: Comparing the reduction in vertices and edges

Network centrality analysis is affected by the digraph construction. Moreover, creating accurate digraphs is essential to avoid meaningless centrality scores and attribute incorrect significance to users [1]. Deciding what vertices and edges represent is, therefore, crucial when constructing a digraph from public forum posts that accurately model the interactions. Our construction approach is identical to how previous work have constructed their graphs. We denote a set of users V and a set of posts E, as the vertices and edges in a digraph $G = (V, E)$. The set E contains ordered (v, u) pairs of edge elements, where user v makes a post (i.e. reply) to a forum thread started by user u, and $\{v, u\} \in V$.

Step 3. Social network analysis

Network centrality measures are used in SNA to identify important and influential individuals within a network. Multiple centrality measures interpret different aspects of what it means to be important or influential because this depends on the context of the situation. Thus, vertices receive a centrality score that reflects their importance for a particular centrality measure. Sorting these scores will rank nodes against each other, with some having higher scores than others. If an actor

has a high centrality score, they are relatively more important or influential than others with lower scores.

The most considered centrality measures for digraphs are in-degree (C_{D^-}), out-degree (C_{D^+}), betweenness (C_B), closeness (C_C), and eigenvector (C_E) [31]. • In-degree is often seen as a measure of prestige or popularity. • Out-degree captures the outreach of a user to the community. • Betweenness measures the amount of influence a node has over the flow of information. • Closeness measures how fast information will spread from one node in a network to all other vertices. • Eigenvector is another measure of the influence level. These five measures are implemented in off-the-shelf forensic investigation tools such as IMB i2 Analyst's Notebook. This research evaluates the methods' ability to identify important individuals in a criminal network.

Step 4. Statistics

In our previous studies [18, 20] of network centrality, we saw that an actor's centrality rank could be associated with the number of replies received on their forum threads. Any positive association between these two variables will significantly impact how forensic experts can use network centrality during their investigation. A bi-variate analysis can determine if it exists a statistical association between two variables, the degree of association, and whether one variable may predict the other variable [34].

We use bi-variate analysis to test our hypothesis that there is a relationship between the centrality ranks and the number of replies users receive. Pete et al. [30] used Pearson's correlation test to check this relationship. However, Pearson makes wrong assumptions about the variables when analysing this specific type of problem. In particular, Pearson makes the wrong assumptions that the variables are continuous and their relationship is linear. Therefore, we use Spearman's rank-order correlation to make some assumptions about the data, making it more sensitive to non-linear monotonic relationships and ordinal variables. The Spearman's assumptions better fit our experimental data as seen in the scatter plots of Figures 6.4, 6.5 and 6.6.

The experimental data is gathered by (i) doing the previous process steps 1-3 to find users' scores for all of the five centrality measures, and (ii) counting the number of replies users have received to their forum threads. The first variable x is the centrality score sorted in descending order, and this arrangement illustrates the rank users receive from centrality measures. For example, rank 1 contains the user with the highest score for a particular centrality measure; rank 2 contains

the second-highest score; and so forth. Thus, the x-axis is the users' rank order $(1, 2, 3, ..., n)$, sorted from higher centrality to lower centrality. The second variable y is the number of replies users receive. The x and y variables explained here is found in the scatter plots of Figures 6.4, 6.5 and 6.6.

It is important to note that bi-variate analysis cannot statistically account or control for other variables other than the two studied variables. Consequently, bi-variate correlations alone do not necessarily imply causation [34] because both variables can be associated with a different casual variable. We discuss this problem in more detail in Subsection 6.4.1.

6.3.3 Data preprocessing

Step 5. Text preprocessing

The text must be preprocessed before it can be analysed by topic models such as LDA or GSDMM. Thus, we conduct a series of standard and original text preprocessing steps tailored for the online communication-type of data found in Nulled and Cracked datasets. Importantly, we ensure to replace anything removed from the text with white space. This white space replacement is necessary to guarantee that words are not unintentionally combined and cause the text (or topics) to become unintelligible. The following is a list of our preprocessing steps and their order:

(a) Convert text to lowercase.

(b) Remove special newline, tabular and return characters.

(c) Remove BBcode tags.

- BBcode tags format messages in online forums, usually indicated by a keyword surrounding square brackets.

(d) Remove URLs.

(e) Remove HTML tags (including most content) and HTML entities.

(f) Remove forum-specific text.

- Leaking credentials is one of the primary focus for Nulled and Cracked, which means they contain posts with large dumps of e-mail and passwords. We remove any 'e-mail:password'-combination, e-mail addresses and emojis.

(g) Remove symbols and numbers (anything non-alphabetic).

(h) Use lemmatisation to transform inflected words to be analysed as a single item.

(i) Remove stopwords, around 700 of the most common English stopwords.

(j) Search and replace repeating words and characters.

- Users frequently express exaggeration and other emotions by repeating characters (e.g. 'thaaaaanks') and words (e.g. 'thxthx'). This way of writing creates many unnecessary words that have the same meaning (e.g. 'thanks'). Thus, this step, together with lemmatisation, greatly reduces the number of word variations and, consequently, improves the topic models. We refer readers to our previous research article [20] for more details about this step.

Step 6. Topic modelling

The next problem – after preprocessing the text – is that we now have a document corpus where we do not know the contents. Each document in this corpus is one message or a combination of messages from users. To learn the contents (i.e. topics) of the documents, we train them using two topic modelling algorithms LDA and GSDMM, for the first and second iteration, respectively. These algorithms have distinct advantages when modelling long and short text documents, where LDA have better performance on long documents, and GSDMM perform better on shorter text. In this step, we explain the different document construction approaches required by the algorithms while referring readers to our previous research article [19] for other document construction approaches.

a. LDA topic modelling

LDA is a generative model which learn the joint probability distribution $P(x, y)$. More specifically, LDA try to solve a general problem with an intermediate step by modelling how a particular topic y would generate input data x. The model can subsequently pick the most likely topic by calculating the conditional probability $P(y|x)$. I.e. what is the probability of topic y given the input values of x. In the case for LDA, the inputs x is latent variables aimed to capture abstract notions such as topics [4]. The result is a set of human-interpretative topics, which explain why some parts of the data are similar.

The creation of a LDA model is controlled by three hyper-parameters k, α, and . The hyper-parameters control the number of topics and two Dirichlet distributions for the document-topic and topic-word density. Thus, LDA allow for a nuanced way of 'soft clustering' documents into k topics, as documents can belong to multiple topics. We found that lower hyper-parameter values ($k = 10$, $\alpha = 0.05$, and $= 0.05$) works better for our proposed method. A low k is sufficient because the aim is to create a coarse grouping of topics instead of identifying all possible topics. While low α and values assume that documents contain fewer topics and topics are composed of few words.

Document construction also plays an essential role in the LDA model creation because the document structures affect the learning process. Additionally, result interpretation is improved by understanding and being aware of the document construction. For the first iteration, we construct the documents by combining all messages from a user into a single document. Thus, the number of documents equals the number of users because they have written something publicly, and our preprocessing has not removed their posts.

Researchers commonly build LDA topic models to automatically organise, understand, search, and summarise large corpora. The model is subsequently used to find the topic for unseen data. In this experiment, however, we use the model to identify groups of similar users and use this knowledge to separate low-skilled users from presumably medium-/high-skilled users.

b. GSDMM topic modelling

The LDA algorithm groups users who write similar messages together. This is beneficial for the first iteration because the grouping allows us to exclude many underground forum users who exclusively post appreciation messages. Continuing using LDA for the second iteration proved difficult for two reasons: (i) the concatenated document construction do not distinguish users with mixed posts/topics, and (ii) LDA works poorly on short text (predominantly appearing in our corpus datasets).

The effect of the first issue is that the LDA algorithm is unable to further distinguish users based on the message content due to the concatenated document construction. Thus, we lack the nuanced ways for distinguishing users further. One possible approach can be changing the document construction and treating each message as a separate document. However, we would encounter the common LDA issues with learning from short texts. Another approach can be increasing the α hyper-parameter to identify documents with more than one topic. Ultimately

we chose a different topic model algorithm that could model shorter texts because this is a better solution to further differentiate between users.

LDA makes the sensible assumption that longer texts contain multiple topics. However, the increasing popularity of micro-blogging websites such as Twitter and Facebook challenges this assumption. The topic modelling algorithm GSDMM [44] assumes that a document can only belong to a single topic, which makes it better suited for topic modelling of shorter text. The other difference is that LDA requires the k number of topics to be set in advance, whereas GSDMM only requires an upper bound of k number of topics to infer the number of topics from the data. These changes has shown that GSDMM generally outperform LDA on short text [25].

The GSDMM algorithm is better suited for short text, which demands a different document construction approach. We consider each user's message as a separate document for the second iteration. We keep the other hyper-parameters similar to the first iteration, where $k = 10$, $\alpha = 0.05$, $= 0.05$, and $i = 30$ number of iterations.

Step 7. Identify users

This intermediary step makes the user removal easier by (i) identifying the topics, and (ii) users' association to topics. The topic identification lists out the $m = 20$ topic keywords and interprets the keywords to infer the topics. While associating users with topics is achieved using the previous LDA or GSDMM topic models to calculate how similar users' messages are to each topic in the model. Notably, users' messages must be constructed similarly to the document construction which generated the model.

We can calculate the conditional probability $P(y|x)$ to find the similarity between a document y given a topic x, which results in a k-length vector for every user of floating-point values between 0.0 and 1.0. This approach only works when each user has a single vector, such as in the case of the first iteration. The second iteration, however, resulted in multiple documents for each user. We summed the document vectors for each user and normalised them by the number of messages they sent, which resulted in every user being represented by only one document vector. Finally, we perform binarisation on the document vectors to avoid defining threshold values by which we remove users in the next step. The binarised operation gives a one for the highest topic(s) score(s), while the other topics receive a zero. This step is also described in our previous work [20].

Step 8. Remove users

We can remove users based on the content they produce when we are equipped with the knowledge (gained in the previous step) about: (i) the topic keywords, and (ii) how users are associated with each topic. We want to avoid working with floating-point vectors because this requires us to define threshold values for removing users, which must be defined after further studies. We also see threshold values as a potential weakness because the values may change depending on the dataset being analysed. Thus, we choose to work with the binarised vectors when removing users.

The removal process is: (i) identify topics with semantically coherent appreciation words, and (ii) removes users who predominantly write messages on these topics. For example, let there be $k = 10$ topics numbered from 0 – 9 since tagging them is unnecessary. Assume a human analyst evaluates topics 4 and 7 to contain appreciation keywords (such as 'nice', 'man', 'work', 'thank', 'good', and 'work'), then we remove all users who have a binary value of one in topic 4 or 7 (or both) from the dataset. Results of this process are found in Subsection 6.4.2.

6.4 Experimental results and discussion

This section details the results of our experiments and discusses their significance. It is divided into two subsections, which detail and discuss the two primary goals of our research separately. More specifically, the subsections concentrate on: (i) hypothesis testing the correlation between centrality ranks and number of replies thread starters receive, and (ii) a novel method for discarding uninteresting users from underground forums. Subsection 6.4.1 details the correlation we found between the two previously discussed variables. Additionally, this subsection elaborates on the adverse effect the correlation can have on police investigations into underground forums and other groups of criminals (e.g. gangs, organised crime, and so forth). Although our findings demonstrate why investigators cannot rely on network centrality measures for finding 'key criminal actors', they still need to identify a small group of more prominent criminals to focus their investigations on them. Subsection 6.4.2 details our novel method of discarding low-skilled forum users, which leaves a smaller group of cybercriminals that can be the targets of law enforcement investigations.

6.4.1 Correlation testing

Our research hypothesis is that there is a correlation (i.e. a relationship) between the two variables: (1) the order of centrality ranks, and (2) the number of replies thread starters receive. In other words, users who start popular threads (i.e. threads

with more replies and attention) will be more central in the social network and, consequently, obtain higher centrality ranks. We begin scrutinising our hypothesis by investigating the relationship between the two variables using scatter plots, as seen in Figures 6.4, 6.5, and 6.6 for the datasets Enron, Nulled and Cracked, respectively.

The first sub-figure in Figures 6.4, 6.5, and 6.6 shows the raw data curve where the users are ranked (in descending order) according to the number of received e-mail and forum thread replies. Users with more popular forum threads are found on the left side of the scatter plot, while the popularity quickly reduces towards the right. The Enron dataset has a similar interpretation, where the popularity in this dataset is the number of received e-mail messages.

More importantly, Figures 6.4, 6.5, and 6.6 also show the scatter plots for the relationship between various centrality measures and the number of replies to e-mail messages/thread starters. These scatter plots show that the relationship between the two variables follows a similar non-linear relationship as the raw data sub-figure. Notably, users with more replies rank higher in in-degree, closeness, betweenness, and eigenvector centrality. We perform a Spearman rank-order correlation test to determine the strength and direction of the relationship between the two variables. The Spearman test is appropriate due to the ordinal variable of centrality ranks and the non-linear monotonic relationship revealed by the scatter plots.

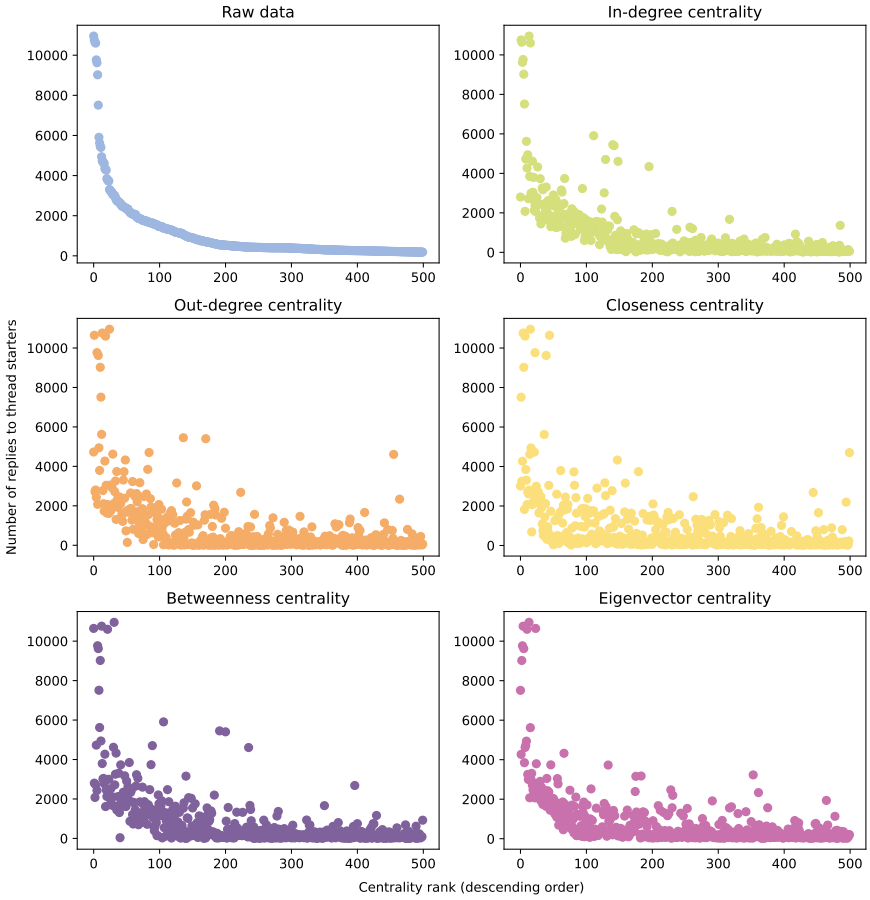


Figure 6.4: Enron statistics

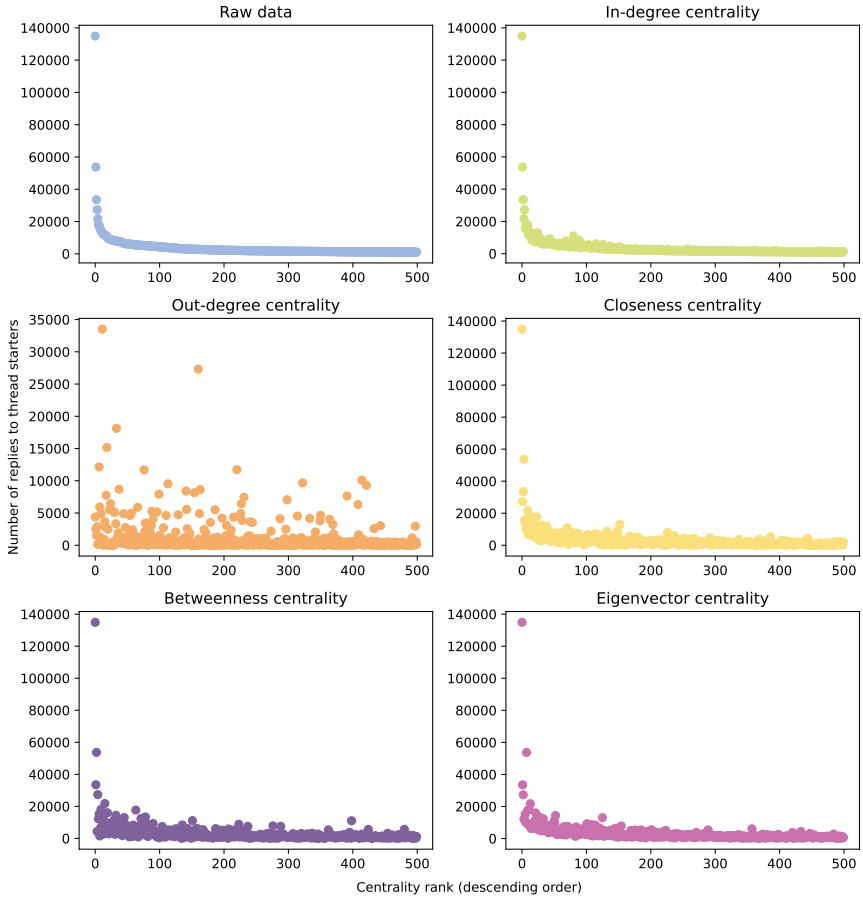


Figure 6.5: Nulled statistics

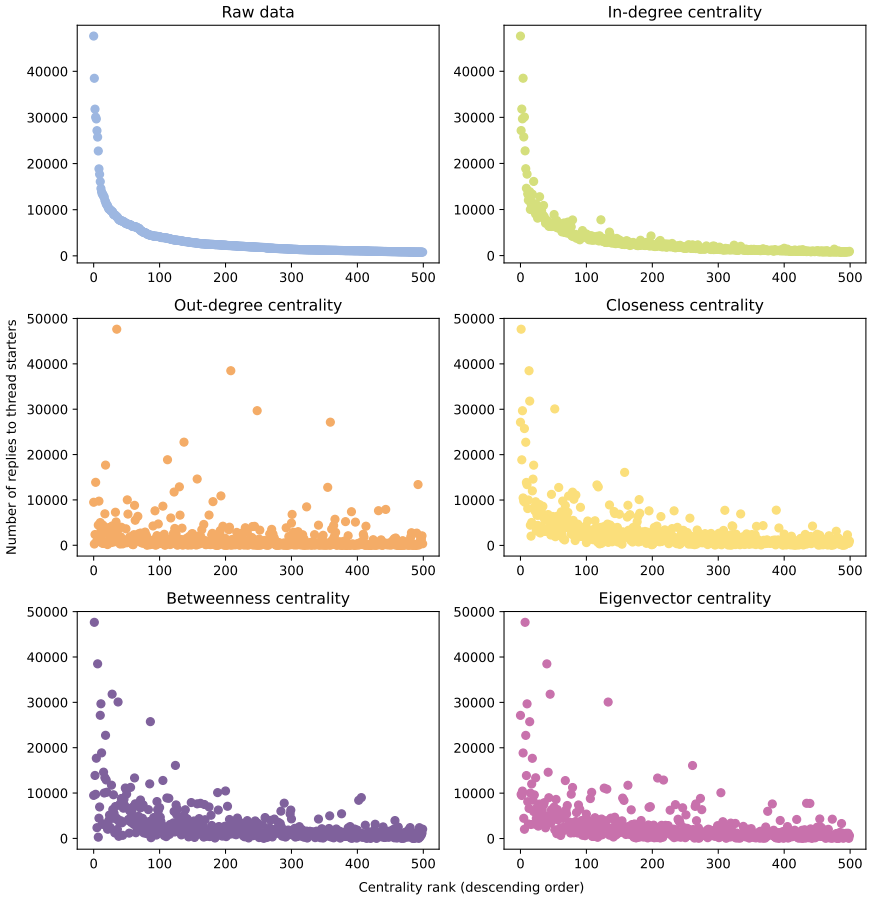


Figure 6.6: Cracked statistics

Table 6.4 shows the results of the Spearman rank-order correlation test. Spearman's r_s value indicates the strength of the relationship between two variables to change in tandem. The relationship strength is typically categorised in five levels: very strong $(1.0-0.9)$, strong $(0.89-0.7)$, moderate $(0.69-0.4)$, weak $(0.39-0.1)$ and no correlation $(0.09-0.0)$. Table 6.4 demonstrates a moderate/strong correlation between the two variables for closeness, betweenness, and eigenvector centrality, while in-degree centrality has a very strong correlation. Moreover, the r_s value indicates a positive relationship, as users with more replies tend to coincide with high network centrality ranks. The strong positive relationships are clearly visible in Figures 6.4, 6.5, and 6.6, as the data points fall to the x-axis.

The p-value is used for hypothesis testing. The default null hypothesis (H_0) says

Dataset	Centrality	r_s	p-value
Enron	C_{D^-}	0.6711	0.00
	C_{D^+}	0.5042	0.00
	C_C	0.4601	0.00
	C_B	0.5004	0.00
	C_E	0.4620	0.00
Nulled	C_{D^-}	0.9956	0.00
	C_{D^+}	0.3841	0.00
	C_C	0.6218	0.00
	C_B	0.7142	0.00
	C_E	0.6518	0.00
Cracked	C_{D^-}	0.9969	0.00
	C_{D^+}	0.2934	0.00
	C_C	0.4633	0.00
	C_B	0.5665	0.00
	C_E	0.4549	0.00

Table 6.4: Spearman rank-order correlation

there is no correlation between the two variables, while the alternate hypothesis (H_1) says there is a correlation between them. The p-value provides statistical evidence to reject or accept the H_0 if the p-value is below or above the 0.05 threshold of probability. The p-values in Table 6.4 is much lower than the threshold of 0.05. Notably, a weak correlation r_s value can have a significant p-value, which means that the weak correlation is not due to chance factors but is representative of the population. The low p-values mean we can reject H_0 and conclude that there is a relationship between the number of replies and centrality ranks. Thus, these findings support our initial hypothesis that these two variables may be associated. In other words, central users are those users with more replies to e-mails or forum threads.

It is important to note that an observed correlation/association does not assure that the relationship between two variables is casual [35]. A casual relationship can be established with well-designed empirical research. Our research shows that the top central 'criminal' actors can be those actors who receive the most attention because of the similar curves seen in Figures 6.4, 6.5, and 6.6. Indeed publicly active actors receive attention from other peers; however, high-skilled actors do not need to be publicly active.

Let us illustrate the issue of identifying actors who receive more attention with an example from related work. Baker and Faulkner [3] reviewed sworn testimon-

ies citing actors who participated directly in price-fixing conspiracies. They find that 'the more direct contacts [(i.e. higher degree centrality)] a conspirator has, the greater the likelihood of a guilty verdict [3].' The centrality measures found more guilty actors in a conspiracy network, but the eyewitnesses were found in the investigation leading up to the testimonies. Unsurprisingly, it is a legal vulnerability [3] when there are eyewitnesses to crimes because they can testify against suspects. Baker and Faulkner find that actors who receive more 'attention' (i.e. 'fingered' by eyewitnesses [3]) are more often found guilty.

Researchers often rely on data from police investigations directly or indirectly via newspapers, web pages, or criminal proceedings. Investigations naturally collect and analyse data centred around a few key actors relevant for the investigation, such as the testimonies in the example above. When reading previous work, it appears that network centrality measures can identify key actors, e.g. leaders in criminal groups or individuals who are more guilty than others. Investigations have an incomplete view of the criminal network, and we argue that centrality measures mirror the findings of experienced investigators instead of identifying the key criminal actors themselves. Baker and Faulkner [3] only achieved their results because eyewitnesses could provide evidence against conspiracy participants, which – by the way – also requires additional evidence to have them convicted. Eyewitnesses and evidence are sometimes not enough because a key actor in a conspiracy escaped indictment due to a technical error by the prosecutor [3].

Centrality measures are simple algorithms that can and cannot identify key criminal actors in right or wrong conditions. It is difficult to list every variable for when they work or not, so our main point is this. Our work demonstrates how forensic investigators and other researchers must be careful when applying network centrality to identify 'key' actors in criminal networks. The notion of 'key' must be interpreted within the graph's context, and that 'key' only refers to a vertex's position in the network structure (depending on the centrality measure used). For example, our digraph is modelled after those who send and receive posts, and the most central actors will be those users with more replies to their threads. Centrality is not a 'criminal score' nor an indicator of an important actor, such as a highly-skilled CaaS cybercriminal or a leader of a criminal network. Forensic investigators risk identifying non-key actors such as secretaries and military staff with numerous contacts or connections but without any authority or influence when uncritically using network centrality measures.

6.4.2 Our newly proposed method

Law enforcement investigators use various off-the-shelf methods to identify criminals of high interest. Network centrality measures are often applied in these situ-

ations to find 'interesting' actors. However, centrality's interpretation differs from what law enforcement investigators consider important. Although researchers and investigators cannot rely on network centrality measures for identifying interesting criminals in a network, they still need to differentiate between high- and low-skilled criminals. This section shows the results of our proposed method, which can identify and remove low-skilled criminal actors on a large scale. Our method uniquely identifies actors who predominantly post appreciation messages and removes them from the dataset, thereby reducing the number of criminal actors to consider.

Tables 6.5 and 6.6 show the result of our process after the first iteration, as described in Subsection 6.3.3. These tables show the ten topics, the number of posts and the keywords associated with each topic. The result obtained here uses an LDA model learned on documents constructed by concatenating users' messages into individual larger documents. The LDA model creates ten 'clusters' that contains users posting similar content. We can see that topics 1 and 9 from Table 6.5, and topics 5, 6, and 8 from Table 6.6 contains words that express appreciation.

Table 6.5: Nulled first iteration (LDA)

Topic #	# of posts	Keywords
0	35715	file set add function enemy type bot attack
1	113068	nice good man work brother test job share wow
2	34037	php inurl http qwerty asp game product cumsot
3	34537	game origin sims email battlefield unknown
4	72924	work account share post check game hope time
5	34335	qwerty qwe xx lol try wsx abc thanks wso kid
6	34043	http site password php username capture point
7	35678	account pm skin sell level bump email price rp
8	43826	script bol work kappa lol update help test thanks
9	151050	thanks test share brother man work nice mate

The next step is to remove users who have posted on these topics, as explained previously in Step 8 in Subsection 6.3.3. We removed 199 786 and 133 454 users from Nulled and Cracked, respectively. An additional 32 166 (10.73%) and 13 754 (7.40%) users were removed because the rigorous text preprocessing had taken out all of their content and LDA could not assign empty users' documents to topics. The first iteration removed a considerably large portion of the forum users, as 67 767 and 38 602 users remained in the datasets. Thus, this iteration reduced the number of users by 77.39% and 79.23%.

Table 6.6: Cracked first iteration (LDA)

Topic #	# of posts	Keywords
0	37343	yeah account leech leave post work thread link
1	15722	combo premium subscription status recur
2	16266	premium country family credit spotify false plan
3	17815	pour le file partage je asd de ce da la deep mon
4	15549	game lil platform php cry assassin creed fortnite
5	104954	work man good nice hope great brother mate
6	35985	share brother lot combo man work hope men
7	15867	true rar sb txt xnr php pdf account lik asba
8	33777	thanks dude brother friend much you por best
9	16318	live checked unknown account state united

Bidirectional Encoder Representations from Transformers (BERT) [6] is based on a multi-layer bi-directional transformer architecture and has presented state-of-the-art results in a wide variety of NLP tasks. BERT's bi-directionality distinguishes itself from other language models' sequential processing, and it effectively addresses language ambiguity. It is pre-trained on a large amount of unlabeled data for masked word prediction and next sentence prediction tasks [6]. BERTopic [14] leverages BERT embeddings and other methods to create dense clusters representing each topic. We did initial experimentation with BERTopic using an Associated Press corpus, and it got some promising results. Other researchers' use of BERTopic [2] also persuaded us to use it for the second iteration.

Table 6.7: Cracked second iteration (BERTopic)

Topic #	# of posts	Keywords
-1	234979	premium yeah country spotify live brother
3	74926	version clean bb late storm io newer eform release
974	41247	work hope appreciate good great job love brother
7027	40478	php rar sb xnr txt man hq post pdf hold sync
6938	39959	game platform cry assassin creed ahmad pc arena
6920	32169	mate quote fuck xd hour deap deep lol ban time
60	25097	lil thanks jocker pour partage nice le much je ce
247	22626	share brother nice great man appreciate mate
3651	22176	leech leave leak report dont forget enjoy ban
193	21698	account sims fortnite battlefield commercial
7090	19613	combo thread content reply count course list

Table 6.7 contains the topics found by BERTopic, and we identify 1 and 6 as

the appreciation topics. We follow the removal process as described in step 8, Subsection 6.3.3, and try two training approaches: (i) concatenating messages and (ii) individual messages. We remove 989 users by concatenating their messages and 1749 users when treating the forum posts as individual messages. This result demonstrates that the BERT transformer model has a poor fit for this particular problem, especially when comparing this result with the GSDMM's result. The main problem with transformers is their inability to handle long text sequences. E.g. BERT supports up to 512 tokens. Another challenge is posed by the user-generated content [40], which has many unique characteristics and frequent use of informal language, which typically has short context, noisy, sparse and ambiguous content.

Transformer models have the potential to work well under the right circumstances. For example, BERT performs well on static datasets where labelled data is available [40], however, forensic investigations are dynamic (e.g. new vocabulary, writing styles or contexts). It is out of this research scope to generate labelled datasets or continuously re-train and validate transformer models. Thus, these problems are better addressed in other studies.

Tables 6.8 and 6.9 contains the results for the second iteration. The process is identical to the first iteration: identify appreciation expression topics and remove users who predominantly post such messages. The distinction between these iterations is that we use a GSDMM model for the second iteration, which was learned by treating each public post as a separate document. We can see that topics 0 and 8 in Table 6.8 and topic 6 in Table 6.9 show appreciation.

Table 6.8: Nulled second iteration (GSDMM)

Topic #	# of posts	Keywords
0	25976	thanks share test man check brother nice bump
1	1248	de thanks por account eu se da aporte friend para
2	4712	post ban account rep leech help work forum
3	2182	script good vayne best play work well game
4	2903	game good play well love time lol guy kappa best
5	7003	work download update file crack bol link version
6	2836	account pm sell work email paypal buy free skype
7	707	skin account level te ekk key rler rp champion
8	20904	work thanks good nice hope share man test
9	206	php game http password account email site key

The second and final iteration is where we remove users who have predominantly posted messages in topics identified in the Tables 6.8 and 6.9. We removed an

Table 6.9: Cracked second iteration (GSDMM)

Topic #	# of posts	Keywords
0	7895	quote work post good time vouch forum thread
1	2637	proxy checker download work combo link
2	1818	sdf point df sd dfg awd fd gf dd key fg ng er rt
3	2420	leech leave account post report enjoy thread ban
4	2523	thanks pour le partage je te de asd ekk rler ce
5	1448	commercial obrigado muito ich gim da bom por
6	17739	share brother work bump good thanks man hope
7	813	expires leech bol leave war de ii quote ik fire
8	2698	account discord free work sell buy add link
9	294	premium yeah game lil country account live php

additional 46 678 and 17 552 users from Nulled and Cracked, respectively, and 101 and 187 users we could not assign to any topic. Thus, the second iteration reduced the number of users by 69.03% and 45.95%, resulting in 20 988 and 20 863 remaining users.

The Nulled and Cracked datasets initially had 299 719 and 185 810 users. We rigorously checked our results against the datasets to ensure that our method removed lower-skilled forum users instead of users such as reverse engineers and administrators. Our method consistently removed lower-ranked members (such as active members, banned users, and users with bought ranks) from the dataset. The final result after two iterations is a reduction of 93.00% and 88.77%. The design of the user removal method ensures that higher-skilled individuals are kept for further analysis.

Table 6.10: Reduction in underground forum users

Dataset #	Original size	Reduced size
Nulled	299 719	20 988 (-93.00%)
Cracked	185 810	20 863 (-88.77%)

Our method is limited to the two iterations as described in this paper. Any attempts using GSDMM models for a third iteration resulted in unintelligible topic models, which could not be used to distinguish individuals further and remove additional users. However, we observed that using the LDA algorithm in a particular way could significantly lower the number of potentially interesting actors. The input to the LDA algorithm must use a document construction with individual forum posts. After the first iteration, users are selected and removed by choosing the

$k-1$ topics with the lowest number of posts. In other words, remove users from all topics except for the topic with the most posts. We followed this approach for five iterations and ended up with 1 552 and 1 334 users of potential interest. This is a substantial reduction of 99.48% and 99.28% of low-skilled users (according to their forum ranks). We cannot yet explain this behaviour where the LDA topic modelling clusters higher ranking users in one topic. Thus, it is not a method we can recommend now because it needs further study.

6.5 Conclusion

Law enforcement, intelligence, and researchers efficiently employ off-the-shelf tools to identify key actors in large-scale criminal networks. These tools make use of SNA centrality methods, which past researchers have used to identify leaders and other key actors in small criminal networks. This article addresses the need to evaluate and validate centrality measures as a forensic technique for identifying key actors in large criminal networks and to increase our understanding of using centrality in forensics.

We created three interaction networks – where two of them model the communication found in real-world criminal underground forums – and analysed them using five centrality measures. We evaluated the result using bi-variate analysis to understand better which individuals they identify as more important. Our findings show that network centrality measures strongly correlate with the number of replies thread starter users receive. Thus, centrality measures identify actors with more popular forum threads rather than network leaders or other key CaaS/criminal actors. Although centrality measures give an evidence-based quantification of actors' positions within the *network structure*, they are not an indicator for leaders or highly skilled CaaS actors. Consequently, law enforcement resources may be wasted on non-key actors with popular forum threads.

We propose a novel method of separating less skilled forum users from underground forum datasets using topic modelling. This method removes 93.00% and 88.77% of an underground forum population, enabling law enforcement to focus on the remaining (and arguably more interesting) actors. We suggest that future researchers, law enforcement, and intelligence use other analyses to extract knowledge further and gain insight from the remaining actors to target offenders for removal or develop strategies to disrupt criminal networks.

Acknowledgements

The research leading to these results has received funding from the Research Council of Norway programme IKTPLUSS, under the R&D project 'Ars Forensica -

Computational Forensics for Large-scale Fraud Detection, Crime Investigation & Prevention', grant agreement 248094/O70. We want to thank Dr. Stefan Axelsson and Dr. Patrick Bours for their effort in reading and providing feedback to this research article.

6.6 Bibliography

[1] Ahmed Abbasi, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. Descriptive Analytics: Examining Expert Hackers in Web Forums. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 56–63, The Hague, Netherlands, 2014. IEEE. http://ieeexplore.ieee.org/document/6975554/.

[2] Abeer Abuzayed and Hend Al-Khalifa. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, 189:191–194, 2021.

[3] Wayne E. Baker and Robert R. Faulkner. The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review*, 58(6):837, December 1993. http://www.jstor.org/stable/2095954?origin=crossref.

[4] David M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. page 30, 2003.

[5] David A. Bright, Caitlin E. Hughes, and Jenny Chalmers. Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate. *Crime, Law and Social Change*, 57(2):151–176, March 2012. http://link.springer.com/10.1007/s10611-011-9336-z.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[7] Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, 2005.

[8] David Décary-Hétu and Benoit Dupont. The social network of hackers. *Global Crime*, 13(3):160–175, 2012. http://www.tandfonline.com/doi/abs/10.1080/17440572.2012.702523.

[9] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2014. Technical report, 2014. https://www.europol.europa.eu/sites/default/files/documents/europol_iocta_web.pdf.

[10] Europol. The Internet Organised Crime Threat Assessment (IOCTA) 2020. Technical report, 2020.

[11] Katrin Franke and Sargur N. Srihari. Computational Forensics: An Overview. In *Proceedings of the 2Nd International Workshop on Computational Forensics*, IWCF '08, pages 1–10, Berlin, Heidelberg, 2008. Springer-Verlag. http://dx.doi.org/10.1007/978-3-540-85303-9_1.

[12] G. Gogolin. *Digital Forensics Explained*. CRC Press, 2021.

[13] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 13–18, Beijing, China, July 2017. IEEE. http://ieeexplore.ieee.org/document/8004867/.

[14] Maarten Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics., 2020. Version Number: v0.7.0.

[15] J. S. Hardin, Ghassan Sarkis, and P. C. Urc. Network analysis with the enron email corpus. *Journal of Statistics Education*, 23(2), 2015.

[16] Thomas J. Holt, Deborah Strumsky, Olga Smirnova, and Max Kilger. Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology*, 6(1):891, 2012.

[17] Jan William Johnsen and Katrin Franke. Feasibility Study of Social Network Analysis on Loosely Structured Communication Networks. *Procedia Computer Science*, 108:2388–2392, 2017. http://linkinghub.elsevier.com/retrieve/pii/S1877050917307561.

[18] Jan William Johnsen and Katrin Franke. Identifying Central Individuals in Organised Criminal Groups and Underground Marketplaces. In *Computational Science – ICCS 2018*, volume 10862, pages 379–386. Springer International Publishing, Cham, 2018. http://link.springer.com/10.1007/978-3-319-93713-7_31.

[19] Jan William Johnsen and Katrin Franke. The impact of preprocessing in natural language for open source intelligence and criminal investigation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4248–4254, Los Angeles, CA, USA, December 2019. IEEE. https://ieeexplore.ieee.org/document/9006006/.

[20] Jan William Johnsen and Katrin Franke. Identifying Proficient Cybercriminals Through Text and Network Analysis. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–7. IEEE, 2020. https://ieeexplore.ieee.org/abstract/document/9280523.

[21] Valdis Krebs. Uncloaking Terrorist Networks. volume 7, page 4. First Monday, 2002. http://journals.uic.edu/ojs/index.php/fm/article/view/941.

[22] Pawan Kumar and Adwitiya Sinha. Information diffusion modeling and analysis for socially interacting networks. *Social Network Analysis and Mining*, 11(1):11, December 2021.

[23] Yong Lu, Xin Luo, Michael Polgar, and Yuanyuan Cao. Social Network Analysis of a Criminal Hacker Community. *Journal of Computer Information Systems*, page 12, 2010.

[24] Ericsson Marin, Jana Shakarian, and Paulo Shakarian. Mining Key-Hackers on Darkweb Forums. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 73–80, South Padre Island, TX, 2018. IEEE. https://ieeexplore.ieee.org/document/8367642/.

[25] Jocelyn Mazarura and Alta de Waal. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6, Stellenbosch, South Africa, November 2016. IEEE.

[26] Bisharat Rasool Memon. Identifying Important Nodes in Weighted Covert Networks Using Generalized Centrality Measures. In *2012 European Intelligence and Security Informatics Conference*, pages 131–140, Odense, Denmark, August 2012. IEEE. http://ieeexplore.ieee.org/document/6298823/.

[27] Carlo Morselli. Assessing Vulnerable and Strategic Positions in a Criminal Network. *Journal of Contemporary Criminal Justice*, 26(4):382–392, November 2010. http://journals.sagepub.com/doi/10.1177/1043986210377105.

[28] Daniel Ortiz-Arroyo. Discovering Sets of Key Players in Social Networks. In Ajith Abraham, Aboul-Ella Hassanien, and Vaclav Sná¿el, editors, *Computational Social Network Analysis: Trends, Tools and Research Advances*, pages 27–47. Springer London, London, 2010. https://doi.org/10.1007/978-1-84882-229-0_2.

[29] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum. In Michael Bailey, Thorsten Holz, Manolis Stamatogiannakis, and Sotiris Ioannidis, editors, *Research in Attacks, Intrusions, and Defenses*, pages 207–227, Cham, 2018. Springer International Publishing.

[30] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. A Social Network Analysis and Comparison of Six Dark Web Forums. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 484–493, Genoa, Italy, September 2020. IEEE. https://ieeexplore.ieee.org/document/9229679/.

[31] C. Prell. *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2012. https://books.google.com/books?id=p4iTo566nAMC.

[32] Mayra Rosario Fuentes. Shifts in Underground Markets: Past, Present, and Future. Technical report, Trend Micro Research, May 2020.

[33] Sagar Samtani and Hsinchun Chen. Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 319–321, Tucson, AZ, USA, 2016. IEEE. http://ieeexplore.ieee.org/document/7745500/.

[34] Debra (Dallie) Sandilands. Bivariate Analysis. In Alex C. Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 416–418. Springer Netherlands, Dordrecht, 2014.

[35] Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, May 2018.

[36] Daniel M Schwartz and Tony DA Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009. Publisher: Springer.

[37] Malcolm K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991. http://www.sciencedirect.com/science/article/pii/037887339190008H.

[38] Radina Stoykova. Digital evidence: Unaddressed threats to fairness and the presumption of innocence. *Computer Law & Security Review*, 42:105575, September 2021.

[39] Radina Stoykova. *Standards for digital evidence. An inquiry into the opportunities for fair trial safeguards through digital forensics standards in criminal investigations*. PhD thesis, Dual PhD University of Groningen and Norwegian University of Science and technology, 2022.

[40] Muzamil Hussain Syed and Sun-Tae Chung. MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain. *Applied Sciences*, 11(13):6007, June 2021.

[41] Kamal Taha and Paul Yoo. SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization. *IEEE Transactions on Information Forensics and Security*, pages 1–1, 2015.

[42] Jennifer Xu and Hsinchun Chen. Untangling Criminal Networks: A Case Study. In Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, editors, *Intelligence and Security Informatics*, pages 232–248, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[43] Jennifer Xu, Byron Marshall, Siddharth Kaza, and Hsinchun Chen. Analyzing and Visualizing Criminal Network Dynamics: A Case Study. In Hsinchun Chen, Reagan Moore, Daniel D. Zeng, and John Leavitt, editors, *Intelligence and Security Informatics*, pages 359–377, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[44] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, New York New York USA, August 2014. ACM.

Chapter 7

Article VI - Cyber crime investigations in the era of big data

Andrii Shalaginov, Jan William Johnsen and Katrin Franke. In 2017 IEEE Big Data 1st International Workshop on Big Data Analytic for Cyber Crime Investigation and Prevention. IEEE, 2017. pp. 3672-3676.

Abstract

The amount of data seized in crime investigations has increased enormously. Investigators are more than ever confronted with a vast amount of heterogeneous data, highly-diverse data formats, increased complexity in distributed stored information. With constantly increasing network bandwidth, it makes it extremely challenging to process or even store part of the network traffic. Nevertheless, criminal investigations need to solve crimes in a timely manner. New computational methods, infrastructure and algorithmic approaches are required. Although big data is a challenge for criminal investigators, it can also help them make to source an detect patterns to prevent and solve crimes. This paper aims to raise attention to current challenges in cyber crime investigations – related to big data – and possible ways to approach combating cybercrimes.

7.1 Introduction

Information and Communications Technology (ICT) has developed over several decades to facilitate human-automation and advances. Electronic devices – such

as computers, smartphones, etc. – are strongly integrated into almost every aspect of our everyday life. Although the benefit is enormous, and it is why we invest in such technology, they are being exploited by cybercriminals to their own gains. This also includes any white-collar crime, insiders and other malicious activities or misuse of ICT systems. Therefore, cyber crimes include not only malicious activities, yet also misuse of the functions that the ICT systems were originally designed for.

Recent improvements to ICT and the Internet's availability has boosted cyber criminal's ability to attack computer systems. Cybercriminal now can affect a lot more victims than they could before with victims spread all over the globe. Meanwhile, Cyber Crime Investigations (CCI) has seen an increase in seized data size and complexity of used technologies. However, many of the investigative tools are design only to facilitate manual analysis, such as keyword searches and data representation. As a result, current methodological approaches cannot cope with the large-scale data collection in today's cybercrime investigations. This results in investigations taking months or years before bringing justice and stopping crime [19].

Current methods and tools have only pre-programmed human knowledge and expert opinions in them. They include data processing approaches such as keyword search, log event correlation, data visualisation and manual exploration. On the one hand, nearly exponential growth in data size, complexity and speed they travel with make such approaches inefficient. On the other hand, there exists a need for more efficient models capable of describing phenomena in the data, to process and find evidence. Computer-based methods and modelling are thus, slowly becoming an inseparable part of criminal investigations [6].

The paper provides an insight into big data analytic for cybercrime investigation and prevention. There can be seen a strong need to develop advanced data analytic to facilitate crime investigators. It became infeasible to manually process all case data, and the investigator has to perform rather surface search for any relevant pieces or concentrate on in-depth analysis of any small fraction of data. However, such an approach bears no means of intelligent data analytic when a human-like explanation of the data needs to be given.

The remainder of this paper is organised as follows. Section 7.2 gives an overview of the CCI state of the art and modern classification of cybercrimes. Section 7.3 explains the current challenges of big data in relation to digital forensics. Then applications and future directions of computational forensics are described in Section 7.4. Finally, Section 7.5 states our final remarks.

7.2 Cybercrime investigation

Cybercrime is a fast-growing area of crime. Although there is no universal definition of cybercrime, law enforcement generally makes a distinction between two main types: cyber-enabled and cyber-dependent crimes [13]. The first type is traditional crimes which use ICT to increase their scale or reach, while the second is crimes committed only through the use of ICT. Cyber-dependent crime has – after more than two decades [9] – became one of the major threats to our global economy. Further, the literature overview shows the major categorisation of computer-related crimes (in the book 'Cyber Crime Investigations' [17]), which is also applicable to ICT:

- ICT as a target
- ICT as a tool
- ICT is affiliated with a crime
- Crimes against ICT industry

Furthermore, there was an established *Convention on Cybercrime* or *Budapest Convention on Cybercrime* in 2001 [3], which has been ratified by 52 states as of December 2016. The Convention includes provisions on how to combat such crimes.

Estimates show cyber-dependent crimes results in about 450 billion USD loses in 2016 alone [8]. Some estimates this number to increase up to four times by 2019, to 2 trillion USD [14]. A recent Hiscox Cyber Readiness Report [21] estimates that 72% of US businesses and 43% German firms have been successfully attacked by cyber-dependent crimes, while 42% had to deal with two or more attacks, which in turn increase the expenses used on cybersecurity, reaching on average 11.7 million USD according to Accenture [15]. As a result of the strong integration of ICT in modern society, the cyber crimes have begun to affect ordinary people's everyday life.

Most of the computers in 1990th had storage equal to hundreds of MBytes. This means that most of the files can be reviewed by a single person in a timely manner. In 2017, smartphones have 128 GBytes storage, while computers and laptops hit 2-4 TBytes disk storage level already. Therefore, there is a need for research for new ways of thinking and processing methods such that reduction techniques, data mining and intelligent analysis [16].

To handle these relatively new crimes and the way they are committed, law enforcement agencies had to adapt their forensic methods. This resulted in the emergence of umbrella-term Digital Forensics (DF), which includes sub-areas like mobile devices, memory, network, database and computer forensics. Law enforcement agencies investigative approach had to change too. Thus, a general Digital Forensics Process (DFP) as shown in Figure 7.1 is not only about crime scene work (identification, preparation, approach strategy and preservation), yet also applying advanced data analytic during the lab work (collection, examination, analysis and presentation).

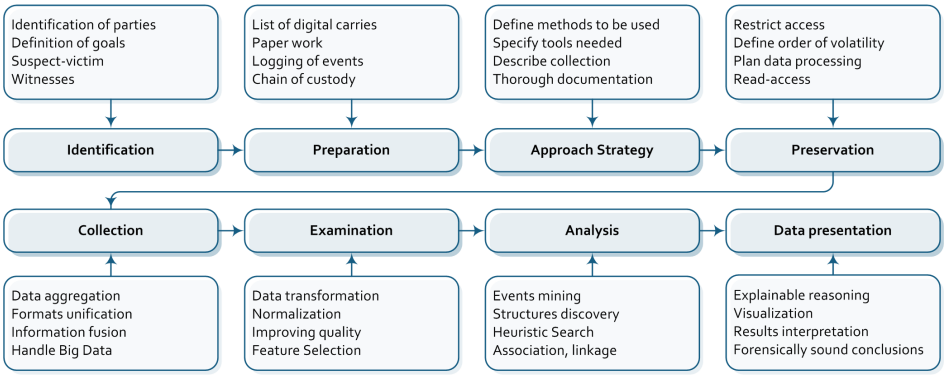


Figure 7.1: The digital forensics process related to data processing and analysis

A successful CCI is also about implementing Digital Forensic Readiness (DFR) [18]. A successful deployment of DFR covers the preparation and execution of policies that solicit fast and cost-effective investigation of incidents. Policies include measures to collect and properly preserve relevant digital evidence; to determine how the incident occurred and provide the necessary documentation to prosecute the perpetrators behind it.

7.3 Big data challenges in digital forensics

DFR allows businesses to plan how to handle incidents. However, it does not necessarily reduce the amount of data examined by human analysts. The fast development in ICT is not going to change and will most likely increase the workload for analysts. The outcome of advancement in ICT is that law enforcement agencies have to prioritise those criminal cases they are able to investigate in a timely manner, which might allow criminal activities to go unpunished.

Big data is a paradigm most often associated with increasing variety, volume, velocity, veracity and value [12]; the so-called five V's. Each of them brings different challenges. Volume and velocity are mostly handled by hardware and software

solutions, capable of strong vast amount of data and transfer it at fast speeds. While there are no easy solutions for variety and veracity in data [23], which encompass incomplete data and varying data formats. The challenge is that current investigation tools and methods are not built to appropriately address today's big data paradigm. This implies a need for new and innovative solutions to analyse huge amounts of data.

An idiom to these challenges is the needle in the haystack; to filter out the noise and discover patterns in large heaps of data to uncover tiny pieces of evidence. Examples on how this can be achieved are through research into data reduction techniques, data mining and intelligent analysis [16]. However, the no free lunch theorem [23] suggests that the same method that finds the needle in one haystack does not necessarily work for another haystack. A particular model can outperform another in a particular situation because it fits better for that problem and not due to the model's superiority. That is, there are no independent reasons to favour one method over another [5]. It is, therefore, important to continuously research new methods and to understand their errors and limitations to extract value for criminal investigators.

While the no free lunch theorem shows that we should not favour one method over another, the ugly duckling theorem [22] states that there is no 'best' feature representation the method can use [5]. Trying to add more features to a model to increase the statistical significance just leads to the curse of dimensionality [11]. Which is a problem caused by the added volume in the feature space; i.e. when more features are added, then the volume of the space increases so fast that the feature space becomes sparse [11].

Over the last few years, authors have been raising the importance of advanced data analytic for DF. Such that DF is already considered to be a big data challenge and requires a complete rethinking of principles [10]. This leads to the application of new tools and the development of new skills to guarantee compliance with guidelines. Another study [24] states new opportunities in the era of big data such that possibilities to efficiently correlate data from different crimes. This may bring new insights and knowledge that has not been previously known. Some authors have been developing new tools to facilitate DFP based on big data-oriented solutions such as NoSQL storage for reports generation [4]. Finally, [1] presented an improvement of the classical DFP, touching specific data processing tasks shown in Figure 7.1.

While there are certain challenges when developing new methods and models, cyber-enabled and cyber-dependent crimes also bring some practical challenges. The interconnected ICT environment means that crimes are no longer gathered in

a finite set of local crime scenes, but rather spread across many distinct systems, with multiple victims and crosses many jurisdictions. It is difficult to have human experts efficiently correlate data from distinct crimes and crime scenes. Therefore, a strong need for advanced data analytic has arisen.

Generally, there are five community-accepted challenges that describe this paradigm [12]: volume, velocity, value, variety, veracity. The first two Vs (data size and network bandwidth) can be handled by hardware and software solutions capable of storing a vast amount of data and transfers at enormous speed. However, there no easy solutions for the last two Vs (incomplete data and a variety of data formats) [23]. This implies a need for intelligent and non-trivial approaches that are capable of extracting a real value of the data for investigators.

For example, most of the computers in 1990th had storage equal to hundreds of MBytes. This means that most of the files can be reviewed by a single person in a timely manner. In 2017, smartphones have 128 GBytes storage, while computers and laptops hit 2-4 TBytes disk storage level already. Therefore, there is a need for research for new ways of thinking and processing methods. For example, research into data reduction techniques, data mining and intelligent analysis [16].

However, the 'no free lunch' theorem [23] suggests that the same method finding the needle in one haystack does not necessarily work for another haystack. It is, therefore, important to research new methods and understand their errors and limitations to extract value for criminal investigators.

However, they are spread across many distinct systems, with multiple victims, and cross more jurisdictions than ever before. It is difficult to have human experts to efficiently correlate data from different crimes and crime scenes. Therefore, a strong need for advanced data analytic has arisen.

Over the last few years, authors have been raising the importance of advanced data analytic for Digital Forensics in their research. Such that Digital Forensics is already considered to be a big data challenge and therefore require a complete rethinking of principles and workflow [10]. This requires the application of new tools and the development of new skills to guarantee compliance with guidelines. Another study [24] states new opportunities in the era of big data such that possibilities to efficiently correlate data from different crimes. This may lead to new insights and knowledge that has not been previously known. It also relates to Cyber Threats Intelligence. Some authors have been developing new tools to facilitate DFP based on big data-oriented solutions such NoSQL storage for reports generation [4]. Finally, [1] presented an improvement of the classical DFP, touching specific data processing tasks shown in the Figure 7.1. So, it can be seen that

there is a strong need for advanced data analytics with some tools already being developed to facilitate CCI.

7.4 Computational forensics

Law enforcement agencies are challenged in their capabilities to analyse large and diverse sources of data. Explicitly programming the knowledge from human experts into forensic tools is not only time consuming, but it also requires it to be correct and reliable. Some sub-fields of data science provides possible solutions to these challenges. More specifically, the sub-fields are machine learning, data mining and pattern recognition.

What these sub-fields have in common is the use of computer power to analyse vast amounts of data. They give machines the ability to learn from data, just as how any human experts would learn, just at a much faster rate and on more data. While humans are limited in how many fields they can be experts in, machines do not have the same limitation. They can learn multiple interdisciplinary fields with the additional benefit of doing it faster than any humans.

Researchers have already started looking at possibilities for applying advanced computational intelligence to assist analysis and criminal investigation. Thus, this research draws upon computational forensics [7] by applying computer-based modelling for forensic science. Computational methods provide tools to support forensic investigators in their daily casework provide a scientific basis and ultimately represent human expert knowledge and reasoning. This is also about looking for abnormalities in sparse and highly-imbalanced data. Moreover, computational forensics complies with the Daubert standards; which provides a rule of evidence regarding the admissibility of expert witnesses' testimony in a court of law. As a result, some of the researches work on the application of human-explainable soft computing and hybrid intelligence as computational forensics methods [20]. So, it can be seen that, despite multiple challenges related to big data in cyber crimes, many promising computational methods have been developed to assist Investigators. Moreover, advanced and cheap hardware with parallel optimisation boosts new approaches to be implemented to tackle real-world investigations such that COPLINK [6] and Hansken [2].

7.5 Conclusion

big data is becoming a challenge to criminal forensic investigators when dealing with cyber-enabled and cyber-dependent crimes. Traditional investigative approaches and digital forensics tools become less efficient, as their capability to provide required results in a timely manner and within resource constraints. One

promising option to cyber crime investigations is to use computational forensics based on advanced data analytic to prevent and combat cybercrime. Therefore, machine intelligence and computer modelling should be an integral part of the investigations. Computational forensics, as one of the solutions, brings fast and efficient ways of analysing data to find tiny evidence in large and unstructured heaps of data.

7.6 Bibliography

[1] Oluwasola Mary Adedayo. Big data and digital forensics. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, pages 1–7. IEEE, 2016.

[2] H. M. A. van Beek, E. J. van Eijk, R.B. van Baar, M. Ugen, J.N.C. Bodde, and A.J. Siemelink. Digital forensics as a service: Game on. *Digital Investigation*, 15:20–38, 2015. http://linkinghub.elsevier.com/retrieve/pii/S1742287615000857.

[3] OF EUROPE COUNCIL. Convention on Cybercrime. *Budapest, November*, 23, 2001.

[4] P. Dhaka and R. Johari. CRIB: Cyber crime investigation, data archival and analysis using big data tool. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pages 117–121, April 2016.

[5] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2012. https://books.google.com/books?id=Br33IRC3PkQC.

[6] Forensic Logic Coplink. *Forensic Logic Announces Acquisition of COPLINK Platform from IBM*. https://forensiclogic.com/forensic-logic-announces-acquisition-of-coplink-platform-from-ibm/.

[7] Katrin Franke and Sargur N. Srihari. Computational Forensics: An Overview. In *Proceedings of the 2Nd International Workshop on Computational Forensics*, IWCF '08, pages 1–10, Berlin, Heidelberg, 2008. Springer-Verlag. http://dx.doi.org/10.1007/978-3-540-85303-9_1.

[8] Luke Graham. *Cybercrime costs the global economy $450 billion: CEO*. February 2017. https://www.cnbc.com/2017/02/07/cybercrime-costs-the-global-economy-450-billion-ceo.html.

[9] David Griffith. *How To Investigate Cybercrime*. November 2003. http://www.policemag.com/channel/technology/articles/2003/11/how-to-investigate-cybercrime.aspx.

[10] Alessandro Guarino. Digital Forensics as a Big Data Challenge. In Helmut Reimer, Norbert Pohlmann, and Wolfgang Schneider, editors, *ISSE 2013 Securing Electronic Business Processes*, pages 197–203. Springer

Fachmedien Wiesbaden, Wiesbaden, 2013. http://link.springer.com/10.1007/978-3-658-03371-2_17.

[11] Eamonn Keogh and Abdullah Mueen. Curse of Dimensionality. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 257–258. Springer US, Boston, MA, 2010. https://doi.org/10.1007/978-0-387-30164-8_192.

[12] B Marr. Why only one of the 5 Vs of big data really matters. *IBM Big Data & Analytics Hub. Available online at www. ibmbigdatahub. com/blog/whyonly-one-5-vs-big-data-really-matters (last accessed February 29, 2016)*, 2015.

[13] M McGuire and S Dowling. *Cyber crime: A review of the evidence–Research Report 75*. London: Home Office, 2013.

[14] Steve Morgan. *Cyber Crime Costs Projected To Reach $2 Trillion by 2019*. January 2017. https://www.forbes.com/sites/stevemorgan/2016/01/17/cyber-crime-costs-projected-to-reach-2-trillion-by-2019/.

[15] Ponemon Institute LLC. Cost of Cyber Crime Study. Technical report, Accenture, 2017.

[16] Darren Quick and Kim-Kwang Raymond Choo. Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digital Investigation*, 11(4):273–294, 2014. http://linkinghub.elsevier.com/retrieve/pii/S1742287614001066.

[17] A. Reyes, R. Brittson, K. O'Shea, and J. Steele. *Cyber Crime Investigations: Bridging the Gaps Between Security Professionals, Law Enforcement, and Prosecutors*. Elsevier Science, 2011.

[18] Robert Rowlingson. A ten step process for forensic readiness. *International Journal of Digital Evidence*, 2(3):1–28, 2004.

[19] Bruce Schneier. *Tracking the Owner of Kickass Torrents*. July 2017. https://www.schneier.com/blog/archives/2016/07/tracking_the_ow.html.

[20] A. Shalaginov. Soft Computing and Hybrid Intelligence for Decision Support in Forensics Science. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 304–306, September 2016.

[21] Hiscox UK. The Hiscox Cyber Readiness Report. Technical report, Hiscox, 2017.

[22] Satosi Watanabe. *Knowing and Guessing a Quantitative Study of Inference and Information*. 1969.

[23] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997. http://ieeexplore.ieee.org/abstract/document/585893/.

[24] Shams Zawoad and Ragib Hasan. Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities. In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pages 1320–1325, New York, NY, August 2015. IEEE. https://ieeexplore.ieee.org/document/7336350/.

Chapter 8

Casework - Digital forensics report for Dagens Næringsliv

Jan William Johnsen and Katrin Franke. In Dagens Næringsliv. 2018. pp. 1-78.

8.1 Executive summary

Dagens Næringsliv (DN) approached the Dagens Næringsliv (DN) to investigate whether some data manipulation had occurred in various log files in its possession. DN advised they are in the process of investigating what is suspected to be the fraudulent manipulation of data in the database of a music streaming service and sought cross-validation on this hypothesis.

The Norwegian University of Science and Technology (NTNU) was asked to investigate whether there was in fact, manipulation of the data, and if so, the scope, methodology, and location of this manipulation. DN suspected there had been manipulation of data due to a spike of user records within specific time periods but did not provide any further details as to why they determined the data to be manipulated and the methods by which it occurred.

Using advanced statistical analysis of the data provided by DN, NTNU determined that there had in fact been a manipulation of the data at particular times due to the large presence of similar duplicate records occurring for a large percentage of the userbase that was active at any given time. In reviewing the data, in isolation from any other records or logs, it was not possible to determine the exact means of manipulation; however, the absence of records with unreadable data suggested it was not an external Structured Query Language injection (SQLi) vector-based attacked, but rather manipulation from within the streaming service itself. Due

to the targeted nature and extent of the manipulation, it is very unlikely that this manipulation was solely the result of a code-based bug or other system anomalies.

The following analysis shows in detail why this conclusion is the most likely conclusion and further, nature and extent it is suspected that the manipulation has affected the accuracy of the data.

8.2 Hypothesis

DN suspect the data it has provided us, is evidence of data manipulation within the records database of a popular music streaming service. DN suspects that this evidence shows an intent by parties within the music streaming service to boost royalty payments and/or dress up the music streaming service as more profitable than it is in reality. DN has asked us to analyse the data and determine:

1. if there has been manipulation of the data;
2. the method of data manipulation;
3. the affected users and numbers thereof;
4. the affected tracks and numbers thereof;
5. the affected artists and numbers thereof; and
6. where possible the difference between actual intentional plays and manipulated plays.

8.3 Assumptions

In order to provide a reproducible and reputable finding, it is important that we state our assumptions on which our findings and analysis is based. In preparing this report, we assume:

- the data was acquired legally;
- the data we have received is complete for the time periods provided;
- DN has not altered the data in any significant manner;
- DN has provided us with the data in its original form;
- any manipulation of the data has either occurred at the server or users' end and not the subject of being altered in transit between both;

- the time provided by the client will be sufficient to determine whether any manipulation of the data has occurred;
- DN has been forthright and honest with us;
- DN is not using this report to discredit us, nor any of our affiliated organisations;
- DN is using this report for journalism and not for illegal or immoral purposes; and
- DN will fairly and accurately report our participation in this project.

We also assume that any fundamental errors or mistakes that have existed in our understanding of the project and the requested report will be addressed with us. While be provided with time for the preparation of an amended report, outlining how these differences alter our opinion and why.

8.4 Data preparation

On February 7, 2018, we received the log files from DN. We first generated MD5 hash sums for each log file immediately after receiving it. These MD5 hashes were compiled into a list and shared via e-mail with DN for back up and cross-validation. The MD5 list was further used by us to preserve the integrity of our work, to always ensure that we worked with the original data. The complete hash list is provided in Appendix A.G.

The files were transferred from the external hard disk to our server via PuTTy Secure Copy Protocol (PSCP). PSCP is a command-line tool for transferring files securely between computers using a Secure Shell (SSH) connection. The data was then rehashed to ensure we transferred the data successfully and in its entirety. Using this data, we performed our analysis, as described in Section 8.5.

After the initial transfer described above, we copied the data to another part of the file system. This copy was reserved as a backup. Both the original and backup files were made read-only to ensure they remained unchanged and avoid unintentional deletion.

8.4.1 Data structuring

MySQL database tables for each log file were created, and we populated each table with the data contained within the corresponding log file. Listing 8.1 shows the MySQL query used to generate the tables used for the analysis. The table name is

replaced with the respective date for each log file, while each field corresponds to a column found in the original Comma Separated Values (CSV) file.

Listing 8.1: Example of query to create database table for date 13.02.2016

```
CREATE TABLE IF NOT EXISTS new_period_2016_02_13 (
  id INT UNSIGNED AUTO_INCREMENT PRIMARY KEY,
  playdatetime DATETIME,
  countrycode CHAR(2),
  systemuserid INT UNSIGNED,
  trackid INT UNSIGNED,
  offlineplay CHAR(1)
);
```

The content of each log file was loaded into their respective MySQL database table via the SQL command/query below. This command ignores the first row which contained headings for each column and then inserts each row into the database without modification.

Listing 8.2: SQL query to load CSV file content into a database table

```
LOAD DATA LOCAL INFILE filepath INTO TABLE table FIELDS TERMINATED
BY ';' LINES TERMINATED BY '\n' IGNORE 1 ROWS (playdatetime,
countrycode, userid, trackid, offlineplay)
```

An example of the first ten rows for the CSV file *all_data_ny_log2016-02-13.csv* (MD5: `EC3D06A81F12990BB0B04EAD9A153E57`) is seen below. All files contained the same column names, separated by semicolons, with the labels `timestamp`, `countrycode`, `systemuserid`, `trackid` and `offlineplay`.

Listing 8.3: First ten rows for date 13.02.2016

```
timestamp;countrycode;systemuserid;trackid;offlineplay
0028-02-12 18:31:19.000;NO;13356188;20659857;Y
0028-02-12 20:25:54.000;NO;13866374;422113;Y
0028-02-12 22:51:47.000;NO;15380556;1647477;Y
0028-02-12 22:58:55.000;NO;15380556;6640790;Y
0028-02-12 23:03:43.000;NO;15380556;6640791;Y
0028-02-12 23:10:23.000;NO;15380556;6640784;Y
0028-02-13 00:19:50.000;NO;13356188;20659854;Y
0028-02-13 00:23:18.000;NO;13356188;20659855;Y
0028-02-13 00:37:23.000;NO;15380556;6640790;Y
```

The `timestamp` column takes the format *YYYY-mm-dd HH:MM:SS.f*, where *mm* and *dd* are zero-padded two decimal number, and *f* is a zero-padded three decimal number for milliseconds. The `countrycode` is a two-letter country code, which is defined in the ISO 3166-1 alpha-2 standard. Both fields `systemuserid` and `trackid` is a decimal number of varying length; we suspect these numbers are unique numeric values that are incremented when a new user or track is created within the system, and each number represents a distinct user or track. Finally,

`offlineplay` holds one of two values: *N*(o) or *Y*(es) for whether the log entry is an online or offline record, respectively. All filed except for `timestamp` can be treated as *categorical* variables.

8.4.2 Data description

We received 74.1 GB in 65 CSV formatted files via an external hard disk. The files contain log entries allegedly representing streamed songs for a total of 65 days over a 110 day period, in two distinct periods of consecutive days. The first period is between 2016-01-21 and 2016-03-03 (43 days), while the second period is between 2016-04-18 and 2016-05-09 (22 days). These periods will frequently be referred to as "period 1" and "period 2", respectively. We did not receive any logs for the intervening days, i.e. between 2016-03-04 and 2016-04-17. See Figure 8.1 for an overview.

There are 1,590,422,377 log entries in total. Figure 8.2 shows the number of entries per day. Figure 8.3 shows the percentage of online versus offline log entries; the total number of tracks played (Figure 8.4) and number of unique system users (Figure 8.5). Keep in mind there is a missing period left out from these figures.

We identified some anomalous timestamps in the data, for example, 0028-02-12. Although it is impossible to have log entries from the year 28, they do not necessarily represent any tampering with the data. The reason for this is that devices playing music without a connection to the log server may use a system clock for a timestamp. Any devices without synchronisation to an external clock may produce an incorrect timestamp for offline playbacks. Another issue is less powerful devices may be incapable of logging in milliseconds, which causes the timestamp to end with ".000". Online log entries have fewer issues as they likely use the log server's time.

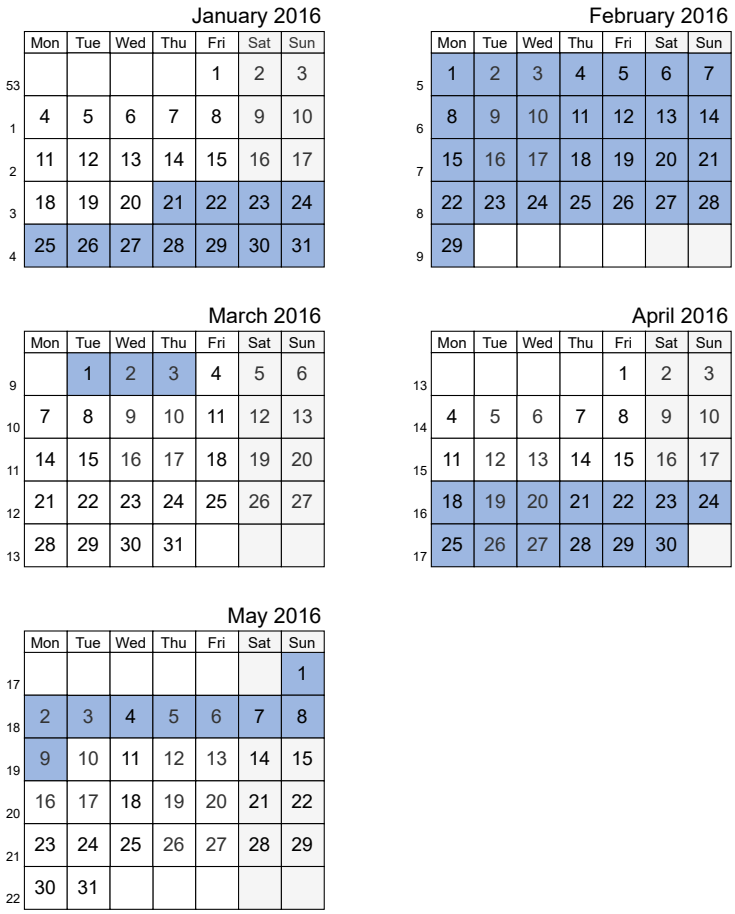


Figure 8.1: Highlighted days represent days with log files

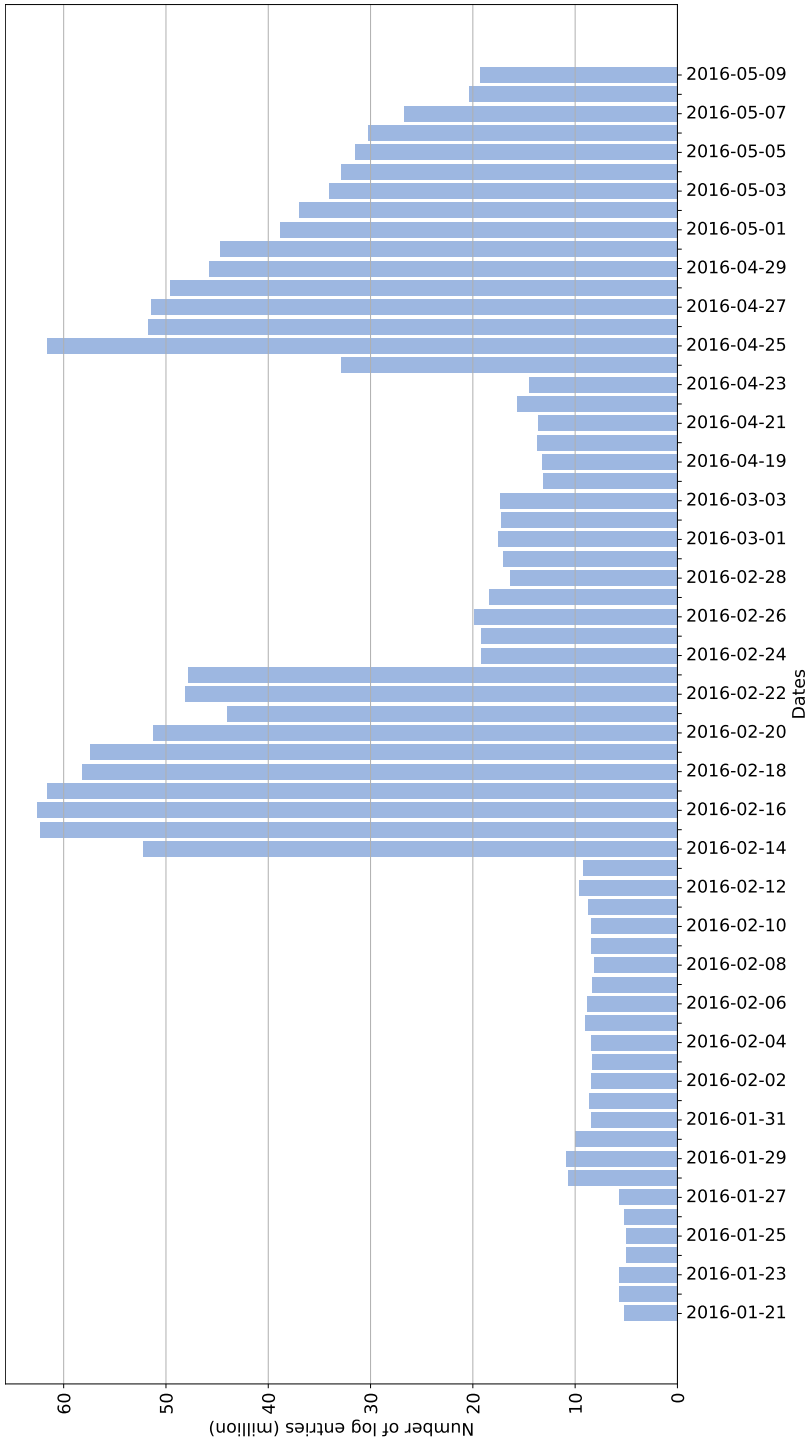


Figure 8.2: Number of log entries

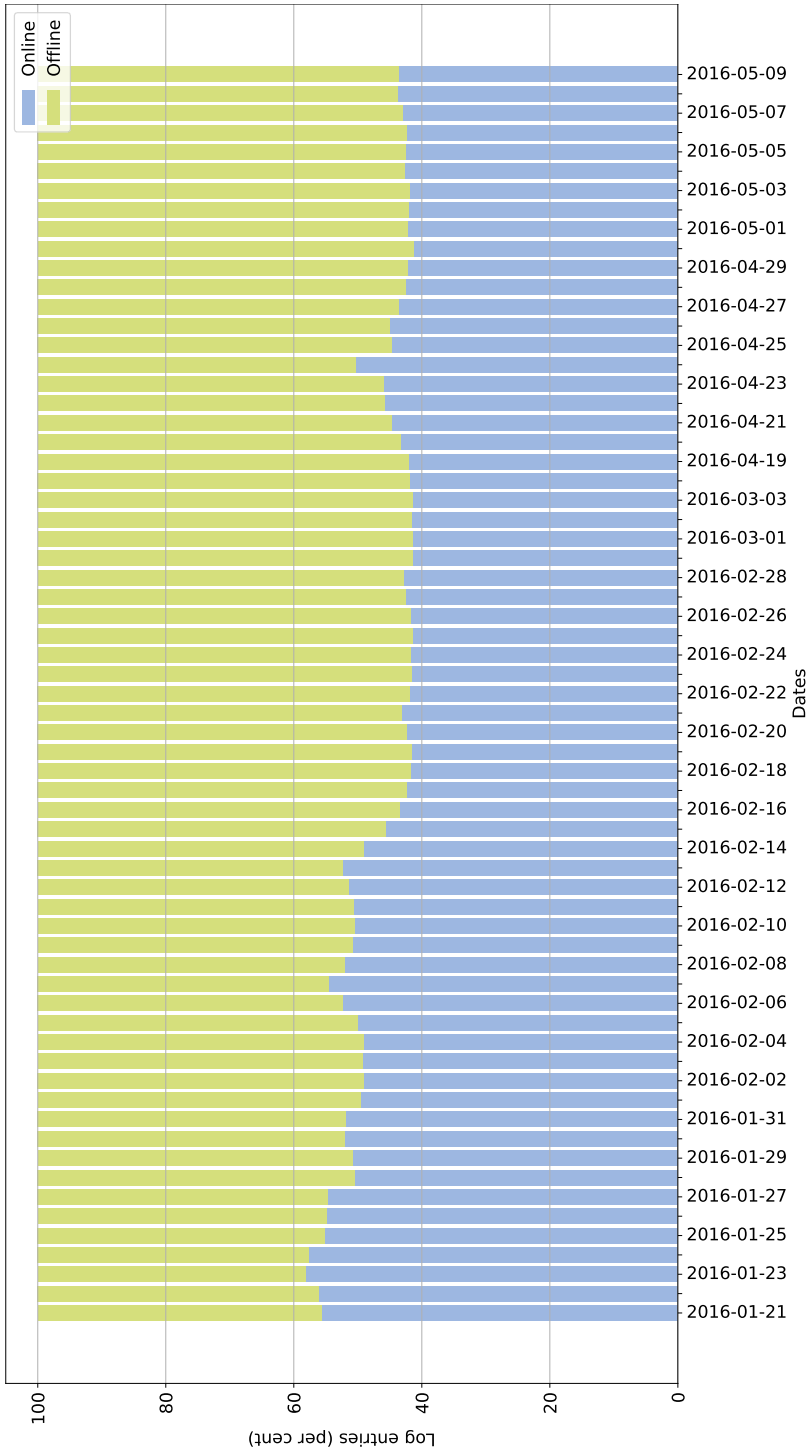


Figure 8.3: Online and offline playbacks

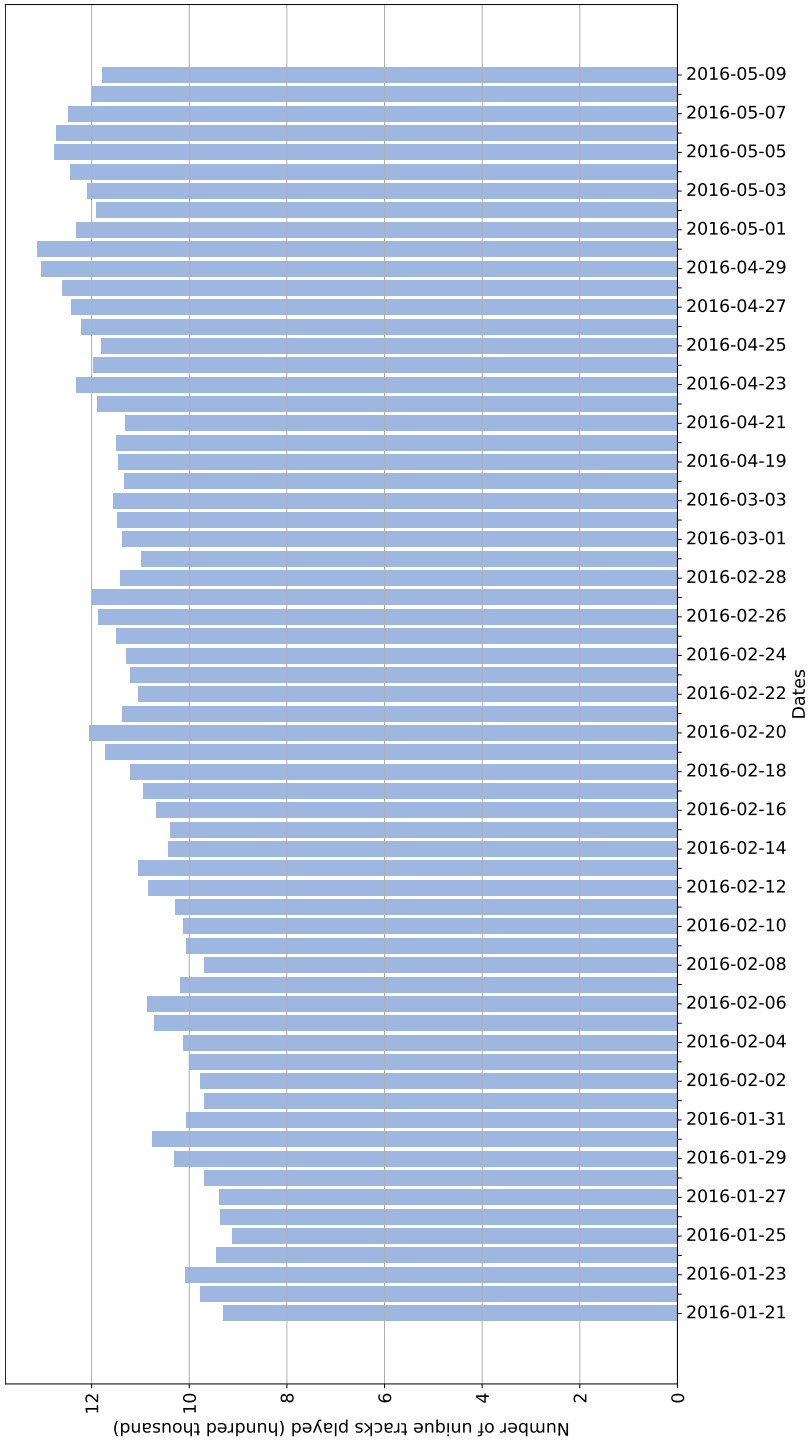


Figure 8.4: Number of unique tracks played

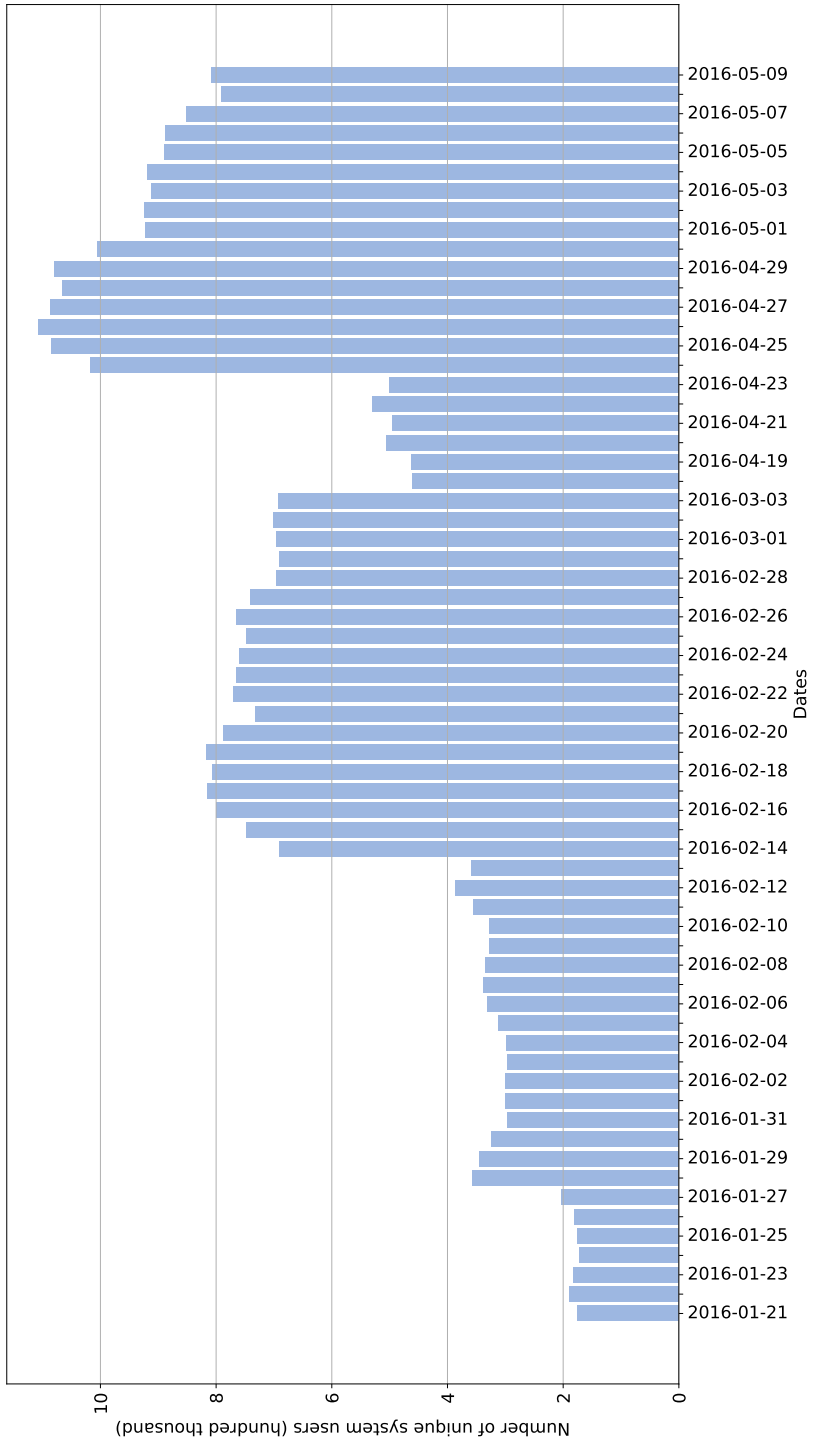


Figure 8.5: Number of unique system users

8.5 Methodology

We documented each step in analysing the data so our results may be reproduced. Each step has its own subsection to describe its purpose/objective, a description, pseudo-code and expected results. Steps with several sub-goals are enumerated.

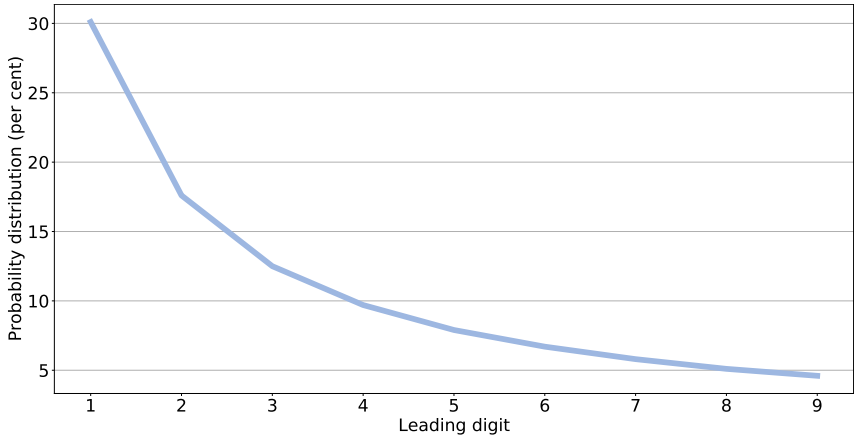


Figure 8.6: Benford's law example

A common expectation is that many of our analysis results should follow Benford's law [1]. Benford's law is the mathematical theory of leading digits that in datasets, the leading digits are distributed in a specific, non-uniform way [2]. For example, the number of people who have listened to five songs would be larger than people who have listened to fifty songs. In other words, lower numbers will appear first in a frequency distribution in many natural cases, as seen in Figure 8.6. Benford's law is used in a variety of fields for the purposes of fraud detection [5, 3, 4].

8.5.1 Descriptive statistical analysis: Analysis method 1

Purpose, objective or hypothesis

The first step in analysing data is a descriptive statistical analysis to better understand the nature of the data we received. We ran several queries to find its dimensions, possible values and so forth.

Verbatim description of the analysis

1. For how many days have we received data? What kind of period does the data give us? Are there any inconsistencies between the filename and the data rows within the file?

2. How large is the data we received (i.e. how many rows)?

3. How many rows are for online and offline plays?

4. How many distinct tracks have been played?

5. How many distinct countries has it been played from?

6. How many distinct users have been playing each day?

7. What are the fields in the CSV file, and what possible values can they take?

Pseudo-code

1. Count the number of files to see how many days we have received, as each file contains data for one day. Then inspect the filename to determine the date and thus the period for which we have data. Finally, check whether the majority of the data rows within the files corresponds to the date found in the filename.

2. Query each database (DB) table to count the number of rows:
   SELECT COUNT(∗) **FROM** *table*

3. Query each DB table to count the number of online and offline plays:
   SELECT (**SELECT COUNT**(∗) **FROM** *table* **WHERE** offlineplay **LIKE** 'Y') **as** offline, (**SELECT COUNT**(∗) **FROM** *table* **WHERE** offlineplay **LIKE** 'N') **as** online

4. Query each DB table to count the number of distinct tracks being played:
   SELECT (**SELECT COUNT**(∗) **FROM** (**SELECT DISTINCT** trackid **FROM** *table*) a) **as** utrack

5. Query each DB table to count the distinct countries:
   SELECT (**SELECT COUNT**(∗) **FROM** (**SELECT DISTINCT** countrycode **FROM** *table*) a) **as** countrycode

6. Query each DB table to count the distinct users:
   SELECT (**SELECT COUNT**(∗) **FROM** (**SELECT DISTINCT** systemuserid **FROM** *table*) a) **as** systemuserid

7. Inspect the CSV header (first row in each file) to determine column names and then inspect a portion of the rows to determine possible values they can have.

Expected results

1. One number for how many files we have received, which should correspond to the number of days. The period is an ordered list of the files. When counting the timestamps found within the files, we expect to find a higher count for the respective date found in the filename. Any large count for other days can be due to user's inability to synchronise their system clocks or that the playbacks have been played on another date but was uploaded/registered for a particular day.

2. One number for how many rows (i.e. distinct playbacks) have been registered by the system on a particular day. In addition, the file size can also be used to get an impression for the received data.

3. Two numbers for the number of online and offline playbacks. We will get to understand the distribution between online and offline playbacks. We expect to find more online playbacks as this is a streaming service. However, a more even distribution would not necessarily be suspicious as it depends on the behaviour of their users. Our expectation is based on the idea of a streaming service is usually online and on-demand.

4. One number for how many distinct songs/tracks are found for each day. Expect that this number would be very even in relation to how many distinct users there are. Releases of new albums should have a low to no impact on this number unless multiple new albums are released.

5. One number for distinct countries represented in the dataset. This number could be affected by services such as proxies or Virtual Private Networks (VPNs), because of the region blocks on music playbacks. Expect to find countries, particularly in Europe and North America.

6. One number for how many distinct system users are found for each day. Comparing this number for multiple days can show whether their customers have grown, shrinks or stays the same over a period.

7. A list with the names of each column found in the files. Inspecting multiple rows will give us an assumption of which possible values each column can have.

8.5.2 Logical impossibilities: Analysis method 2

Purpose, objective or hypothesis

In this method we attempted to find occurrences of logical impossibilities, such as identical or unequal tracks being played at the same time. The client also asked us to do this step, and extract the affected log entries so they could be analysed separately.

Verbatim description of the analysis

1. Find occurrences of two identical track IDs with the same timestamp for a system user ID
2. Find occurrences of more than two identical track IDs with the same timestamp for a system user ID
3. Find occurrences of two unequal track IDs with the same timestamp for a system user ID
4. Find occurrences of more than two unequal track IDs with the same timestamp for a system user ID

Pseudo-code

1. See Appendix A.C
2. See Appendix A.C
3. See Appendix A.D and A.E
4. See Appendix A.D and A.E

Expected results

We do not expect to find a lot of occurrences of logical impossibilities, as they simply should not exist in the data. It should, for example, be impossible to find multiple distinct track IDs being played at the same time. An exception is that there may exist such logical impossibilities from systems with anomalous timestamps. However, this will be considered a limitation for those devices, rather than suspicious activity.

1. A file containing log entries with exactly two identical track IDs which are played at the same time by system user IDs

2. A file containing log entries with three or more identical track IDs which are played the same time by system user IDs

3. A file containing log entries with only two unequal track IDs which are played at the same time by system user IDs

4. A file containing log entries with three or more identical track IDs which are played at the same time by system user IDs

8.5.3 Unique tracks per user: Analysis method 3

Purpose, objective or hypothesis

Count the number of unique tracks played per user (`numuplays`), then count `numuplays` to find the frequency distribution of unique tracks being played. This finds how many *X* tracks has been played *N* times, for example: "436249 tracks have been played 2 times". We try to identify unexpected *N* counts of played tracks. This step can also identify unusual high counts for played tracks that day, for example: "1 track have been played 300 times". Then, we can look further into that particular track with a high play count.

Verbatim description of the analysis

Group based on `systemuserid` and `trackid`, and then `size()` to get the number of times systemuserid u_i have played track t_j. Group by and `sum()` this number (`numuplays`) to count how many times tracks have been played $1, 2, 3, 4, ...n$ times. Will result in an ordered set result, which looks like this: $[1, 3225011], [2, 436249], [3, 113165], [4, 46112], [5, 21189], [6, 12215]$. This is interpreted as "*3 225 011 tracks have been played 1 time*", and so forth.

Pseudo-code

```
df.groupby(['systemuserid', 'trackid']).size().to_frame('numuplays').
    reset_index().groupby('numuplays').size().to_frame('count').
    reset_index()
```

Expected results

We would expect to see a lot of tracks being played once or just a few times, i.e. the frequency distribution follows Benford's law. We consider it suspicious when encountering large "bumps" in the curve.

8.5.4 Tracks per user: Analysis method 4

Purpose, objective or hypothesis

In this method of analysis, the number of tracks played per user is counted (`count`) and the sum of the count calculated. This is used to assess how many *X* users have played how many *N* tracks that day; for example "9778 users have played 1 track". This step can identify unusual or unexpected high counts of *N*, for example: "1 user has played 300 tracks". We can then look further into any users with such high music play counts. However, this is dependent on the final number, as it could just be a large music consumer.

Verbatim description of the analysis

Group based on `systemuserid` (can include grouping by `trackid` to count unique tracks), and then `size()` to get the number of times systemuserid u_i have played track t_j. Group by and `sum()` to get the total played tracks per systemuserid. Finally group by `count` and `size()` to group the number of counts and find how many users have played tracks $1, 2, 3, 4, ...n$ times. These steps will result in an ordered set result looking like this:
$[1, 9778], [2, 7000], [3, 6115], [4, 5879], [5, 5672], [6, 5472]$. This is interpreted as "*9778 users have played 1 track*".

Pseudo-code

```
df.groupby('systemuserid').size().to_frame('count').reset_index().groupby(
    'count').size().to_frame('users').reset_index()
```

Expected results

It is expected that system users are more likely to play tracks a few times, thus have the frequency distribution follow Benford's law. Again, it would be suspicious if we encountered a conspicuous deviation from the curve anticipated by Benford's law.

8.5.5 Popular tracks: Analysis method 5

Purpose, objective or hypothesis

This analysis method was used to provide a further understanding of the most frequent tracks that were played each day. Comparing the results from different dates will also allow us to determine trend-based changes such as when tracks move between being unpopular to popular and vice versa.

Verbatim description of the analysis

In order to analyse the data as described, we grouped `trackid` and used `size()` to count how many times track t_j was played. This data was transferred frame called `frequency` and sorted in descending numerical order (largest to smallest) by `frequency`.

Pseudo-code

```
df.groupby('trackid').size().to\_frame(name='frequency').sort_values(by='
    frequency', ascending=False)
```

Expected results

This will result in a list sorted according to how many times each track has been played. Each distinct playback of a song will count towards one in the frequency. An example of expected results: $[53960289, 12916]$, $[52901260, 12562]$, $[53893678, 11736]$, $[54511031, 9294]$, $[53893676, 6689]$. This list is interpreted as "Track with ID 53960289 has been played 12916 times today".

8.5.6 Number of unique tracks: Analysis method 6

Purpose, objective or hypothesis

We then analysed how many users played a track at least once. The difference between this and analysis method 4 (Subsection 8.5.4) is here we look at distinct tracks being played by all users. While analysis method 4 is more general and looks at all tracks being played. We find that *X* users have been playing *N* unique/distinct tracks that day, for example: "11 794 users have played 1 distinct/unique tracks". The aim of such an analysis is to assess for the presence of unexpected *N* counts of distinct/unique tracks played. This result can be used to further identify track IDs and system user IDs where this occurred.

Verbatim description of the analysis

In order to analyse the data in this way, we counted the number of unique/distinct tracks played per user (`count`) and grouped the resulting data by `systemuserid` to get the total count of unique/distinct tracks played. The final step was to create a new frame with this `frequency`, then group by the `frequency` and `size()` to get the number of users who have played distinct/unique tracks N times. This results in an ordered list like: $[1, 11794]$, $[2, 7912]$, $[3, 7072]$, $[4, 6763]$, $[5, 6612]$. This is interpreted as "11 794 users have played 1 distinct/unique tracks".

Pseudo-code

```
df.groupby(['systemuserid', 'trackid']).size().to_frame('count').
    reset_index().groupby('systemuserid').size().to_frame('frequency').
```

```
    reset_index().groupby('frequency').size().reset_index()
```

Expected results

We expect to see a lot of system users playing a few tracks a few times. In other words, we expect that this frequency distribution to also follow Benford's law. We consider it suspicious when encountering noticeable deviations from the expected curve.

8.5.7 System user frequency: Analysis method 7

The results from the previous analysis step (Subsection 8.5.6), we can see a suspiciously high count of eighteen distinct tracks. It is common knowledge that music albums usually contain approximately thirteen and eighteen tracks. Therefore, it is possible this spike was caused by a new album release. To determine if this was the case, we analysed the users with this abnormal count/frequency found in the previous step. We first had to extract the system users based on the frequency with the following lines of code:

Listing 8.4: Extracting all relevant systemuserid

```
a7_temp = df.groupby(['systemuserid', 'trackid']).size().to_frame('count')
    .reset_index().groupby('systemuserid').size().to_frame('frequency').
    reset_index()
list_systemuserid = a7_temp[a7_temp['frequency'] == 18]['systemuserid'].
    values
df = df[df['systemuserid'].isin(list_systemuserid)]
```

The variable `a7_final` will now contain a subset of the original data. A subset which corresponds to all `systemuserid` with the abnormal frequency of playing exactly eighteen distinct/unique tracks.

Purpose, objective or hypothesis

This analysis focuses on a subset of the records/rows for date 2016-02-14, where the abnormal count in played distinct tracks occurred. We focused our attention on the tracks, system users (including their country) and offline or online plays. The goal is to determine if any tracks have been significantly boosted by these system users, whether offline plays could have played a role, and determine if there was a common origin country for the users.

Verbatim description of the analysis

1. Group by the `trackid` and do `size()` to get the count for how many times that track ID occurs in the dataset. Create a new frame and sort those values, having the most played track first.

2. Extract the songs with the significant higher count of plays, and group the result by `offlineplay` to compare the online plays vs the offline plays.

3. Group by `systemuserid` and `countrycode`, then `size()` and finally group by `countrycode` and `size()` to find the number of unique/distinct systemuserid from which country.

Pseudo-code

1. `top_songs = df.groupby('trackid').size().to_frame('count').reset_index().sort_values(by = 'count', ascending = False)[:18]['trackid'].values`

2. `df = df[df['trackid'].isin(top_songs)]; df.groupby('offlineplay').size().to_frame('count').reset_index()`

3. `df.groupby(['systemuserid', 'countrycode']).size().to_frame('count').reset_index().groupby('countrycode').size().to_frame('count').reset_index()`

Expected results

1. A list with how many times track t_i have been played by those system users. Any abnormally high count could be caused by a release by a new album. Then we can analyse the timestamps to confirm it was caused by the release of a popular new album.

2. Two numbers that count the number of online and offline plays. We would expect to see more online plays from a streaming service. However, previous results have shown an almost even (50/50) split between offline and online plays. Therefore, we expect to find this even split also in these results.

3. Numbers for how many distinct system users are from which country. These country counts may be affected by geolocation obfuscation (through proxies or VPNs) depending on how they determine and store country codes for profiles.

8.5.8 Binning: Analysis method 8

Purpose, objective or hypothesis

Having identified a significant number of timestamps that are not for that specific day in each of the different log files. We investigated the extent these playbacks

affected the results by determining how many they are and whether they show similar anomalies.

Verbatim description of the analysis

To check this, we first split the data into bins based on the year in the timestamp. We created four bins: one for the exact date, one for +/- 2 years from 2016 (i.e. between 2014 and 2018), one for everything prior to 2014 and the last bin for everything post-2018. After sorting the data into four bins, we ran analysis methods 3, 4, 5 and 6 on each separate bin. We also normalised the result values; however, this did not provide any further information. Therefore, we chose not to normalise the results.

Pseudo-code

This is a short snippet for how the dataset was split into separate bins. All rows with the exact date were first put into their own bin before these rows were removed from the data. Then we identified rows within two years of 2016 and put them into another bin, and removed them from the data. This was then repeated for the two final bins until all rows were in their corresponding bin. Finally, we could run our previous analysis methods on each bin.

Listing 8.5: Splitting the dataset into four bins

```
bins_split = {'min': '2014', 'exact': argv[0][11:].replace('_', '-'), 'max
    ': '2018'}
bins = {'below': None, 'around': None, 'exact': None, 'above': None}
bins['exact'] = df.loc[df['playdatetime'].str.startswith(bins_split['exact
    '])].copy().reset_index(drop=True)
df = df.drop(bins['exact'].index.values)
bins['around'] = df.loc[(df['playdatetime'].str[:4] >= bins_split['min'])
    & (df['playdatetime'].str[:4] <= bins_split['max'])].copy().
    reset_index(drop=True)
df = df.drop(bins['around'].index.values)
bins['below'] = df.loc[(df['playdatetime'].str[:4] < bins_split['min'])].
    copy().reset_index(drop=True)
df = df.drop(bins['below'].index.values)
bins['above'] = df.copy().reset_index(drop=True)
df.drop(bins['above'].index.values)
```

Expected results

We expect to find that most of the rows are contained in the bin for the exact date. The `around`-bin (with timestamp +/- 2 years around 2016) should have the second-highest count of rows, as this should incorporate systems without rigorous time synchronisation scheme and smaller devices lacking synchronisation capabilities. A minimum of rows should be contained in the two final bins. However, rows found in those bins cannot be automatically called suspicious, as there could be a

number of explanations for their timestamp. Running the previous analysis steps again, particularly for date 2016-02-14 and analysis method 6 (Subsection 8.5.6), should reveal in which bin the anomalies are found.

8.5.9 Modulo six: Analysis method 9

Purpose, objective or hypothesis

The client noticed that multiple playbacks (with the same track ID) ended on identical seconds and milliseconds, while the hour and minutes differed. From a small subset of system users, they had identified that two timestamps with this characteristic could be divided by six minutes. That is, the time difference between $t_1 =$ 2016-04-28 05:54:26.156 and $t_2 =$ 2016-04-28 07:12:26.156 is 1 hour and 18 minutes, which is evenly divisible by six minutes. Another way to illustrate this is to convert the difference into seconds, *1 hour and 18 minutes = 4680 seconds* and *six minutes = 360 seconds*. Taking *4680 mod 360 = 0* show that the time difference between those two playbacks is evenly dividable by six minutes. The modulo operation finds the remainder after division of one number by another. Modulo shows no remainder after dividing $(t_1 - t_2)$ with six minutes.

Verbatim description of the analysis

Assuming two playbacks have the same system user ID and track ID:

1. Find occurrences of two playbacks which varies by minutes dividable by six minutes (e.g. 6 min, 12 min, 18, min, 24 min, 30 min, etc.).

Pseudo-code

1. See Appendix A.F

Expected results

There is a very low probability that any user would seemingly randomly play the same track on two different times; with a time ending on the same second and milliseconds. Therefore, we do not expect to find a high number of these occurrences, although it would still be statistically possible for this happening for a few users.

8.6 Findings

This chapter summarises the findings for each of the analysis methods; in the same order, the analysis methods were executed. Analysis method 2 is discussed toward the end of this section as while it was a task planned for earlier in the analysis of

the data, requests from the client meant that the analysis was carried out at a later stage. This did not cause any issue or difference in the analysis of the data.

8.6.1 Descriptive statistical analysis findings

The findings from analysis method 1 are described in Section 8.4, including figures found therein. Instead of repeating the findings here, we only describe some interesting observations of the data.

Figure 8.2 shows a steady number of log entries in the first period, until a significant spike on 2016-02-14 until 2016-02-23. The steady number was between 9-10 million, while the spike is as much as 52 million log entries on 2016-02-14. This results in a 477.78% increase in log entries in just one day. This is almost inconceivable and as such is suspicious. The increase in log entries continued for a total of nine days before dropping down to a little under 20 million per day for the rest of the period. A similar spike in log entries occurs in the second period; however, in that case, the number of log entries steadily decline after the initial spike on 2016-04-24/25.

Although the numbers of log entries increase at a large rate, these numbers could be explained by the increase of system users shown in Figure 8.5. The number of unique users went from 358 217 on 2016-02-13, to 691 041 on the day after. Nearly a doubling (92.91% increase) in the number of unique users playing songs per day. However, the next analysis steps will demonstrate that these increases in playbacks are not caused by system users. A doubling in the number of playbacks is a logical conclusion when the amount of system users also doubles. This is consistent with what we see in the dates which are not affected by the sudden spikes in playbacks. For example, 2016-02-13 had about 10 million paybacks, while 2016-02-24 had about 20 million playbacks. Thus, 60 million playbacks for the two suspicious spikes cannot be caused by the system users alone.

We associate music streaming to be on-demand and online, so we would expect to find a lot more online log entries. It is, therefore, an interesting observation that the percentage between online and offline log entries are almost evenly split. We suspect that this is a normal user behaviour since this split remains steady during the entire period. Independently of the large spikes we see in both periods.

8.6.2 Unique tracks per user findings

Analysis method 3 counts how many times unique/distinct tracks has been played per user. For example, “436249 tracks have been played 2 times” for a particular day. The curves for the normal days in the first period (except dates between 2016-02-14 and 2016-02-23) followed our expectations. I.e. users play unique tracks

just a few times, and follow the Benford's law. It is noteworthy that the curve also followed our expectation on 2016-01-28. We will come back to the significance of this date in Subsections 8.6.4 and 8.6.5.

Table 8.1: Example count for date 2016-02-14

trackid	numuplays	count
57273409	3	117533
57273414	3	115518
57273416	3	108465
57273413	3	107218
57273410	3	107195
57273417	3	106766
57273412	3	106219
57273411	3	106055
57273415	3	105983
57317919	3	101238
57273419	3	95899

Figure 8.7 shows the curves for all days in the first period. In the suspicious days, there are distinct frequencies for how many unique tracks are being played per user. This appears to follow some multiplicity of three, with spikes appearing on three, six, nine and twelve. This result shows that system users are recorded to play unique tracks exactly three times, then playing unique tracks two or four times. Table 8.1 demonstrates how many times (`count`) which unique tracks have been played exactly three times.

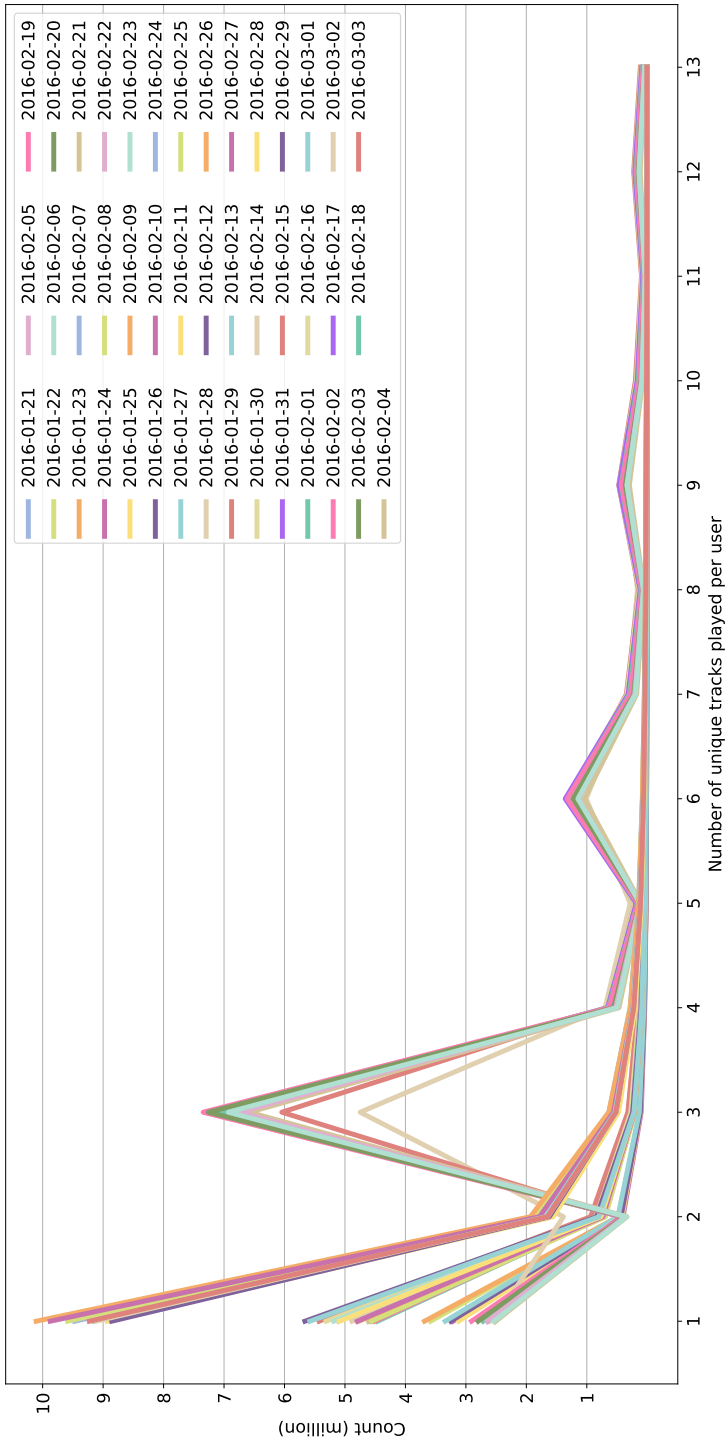


Figure 8.7: Analysis method 3 results period 1

The curves for the second period, as shown in Figure 8.8, generally follow our expectation. A little surprising was occurrences of one individual with a really high number of times playing unique tracks, for example, a user with ID X played track Y a total of 346 times on 2016-05-01. All days had a few users with this high amount of plays for unique tracks; however, we could not find anything suspicious about them.

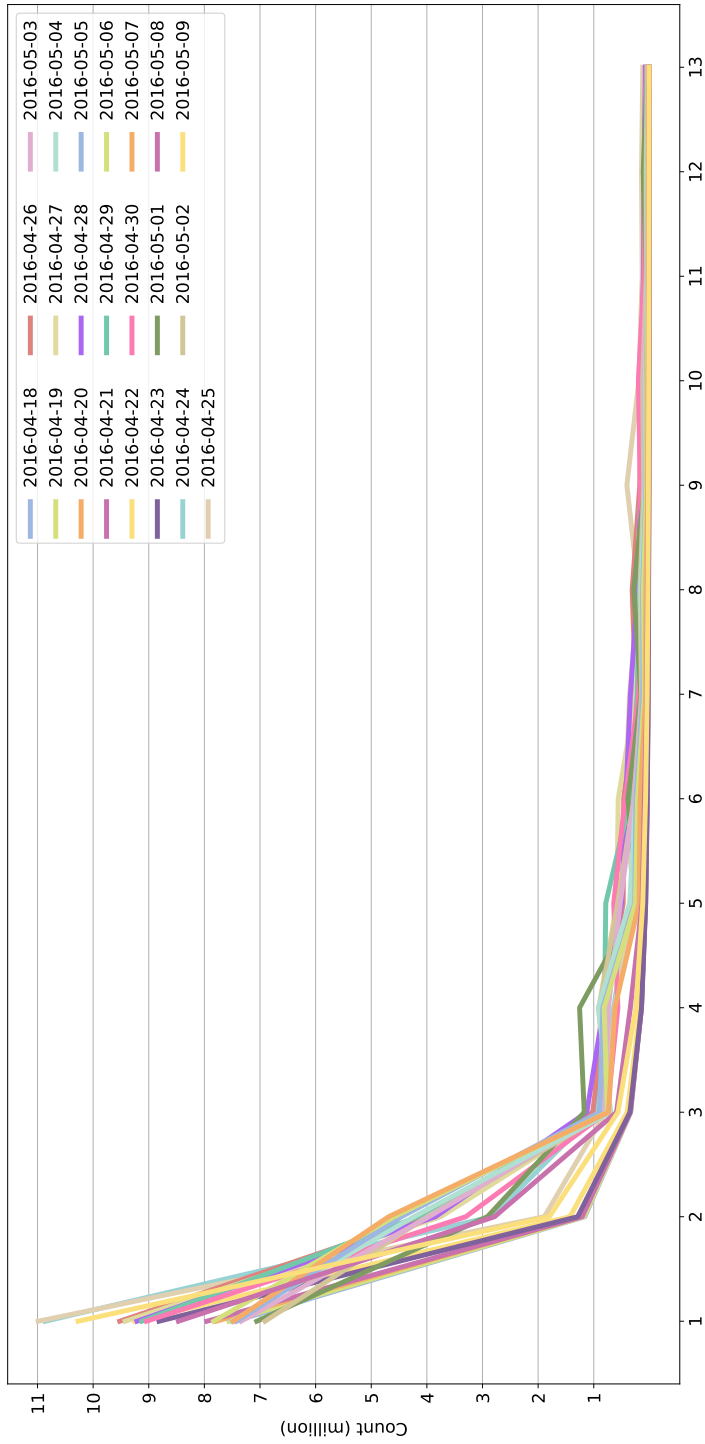


Figure 8.8: Analysis method 3 results period 2

8.6.3 Tracks per user findings

Analysis method 4 counts the number of tracks played per user. For example, "9778 users have played 1 tracks" for a particular day. We again see that the curves for normal days generally follow our expectations. Users normally play a lower number of tracks each day, with a few exceptions by users with up to three thousands plays for a day.

Table 8.2: Example count for date 2016-02-14

count	users
3	25012
6	18545
9	13962
12	13760
18	13605
15	12329
21	10703
1	10168
2	10141
24	9994
54	9400

Figure 8.9 show a small bump for date 2016-01-28 (a red curve) at 13, which we will come back to in Subsections 8.6.4 and 8.6.5. This figure demonstrates that a lot of users had listened to tracks with a multiplicity of three. The sawtooth-shaped curves are only found during the suspicious ten days in the first period. Table 8.2 demonstrates how many users had played exactly `count` tracks for a particular day.

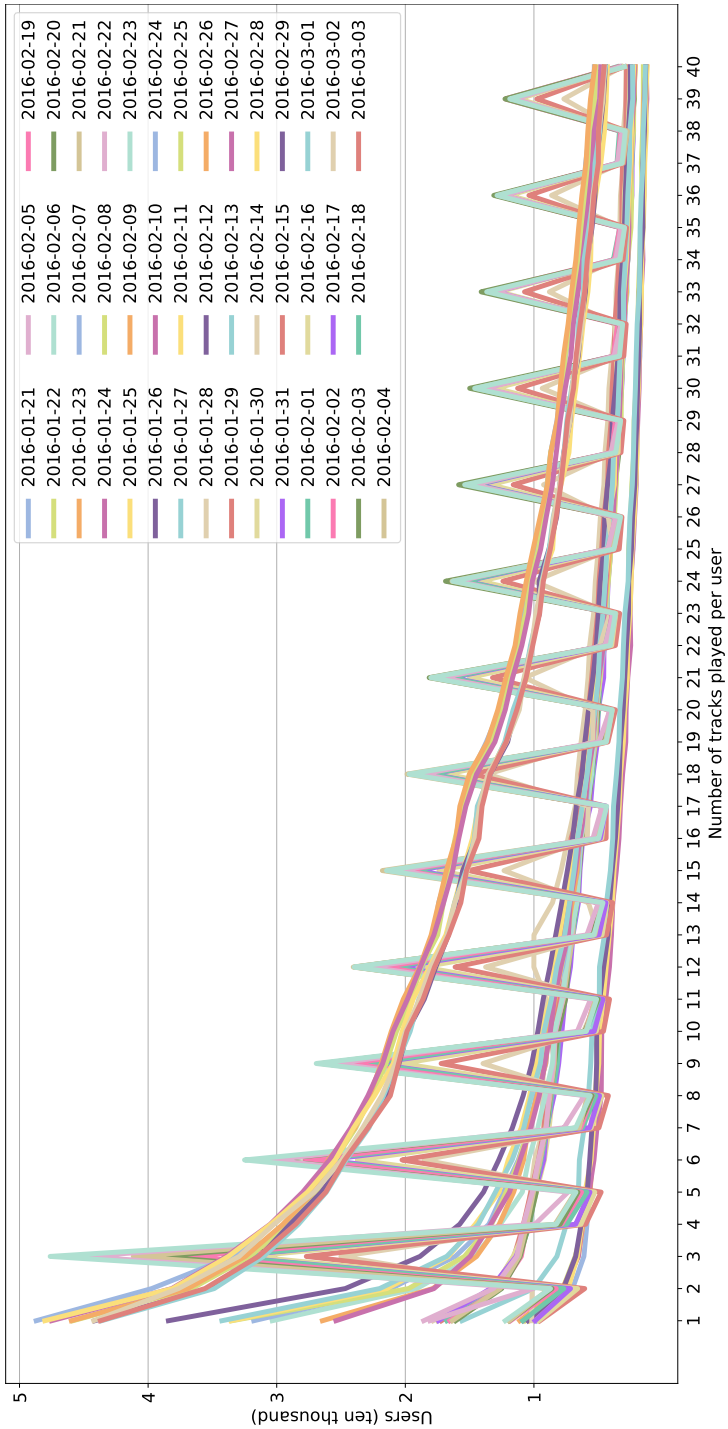


Figure 8.9: Analysis method 4 results period 1

While we did not see anything suspicious abnormalities for the second period in the previous analysis step, this step clearly demonstrates that something is happening on some of these dates. However, this time it appears to be multiplicable of two. The general curve also appears to have been smoothed, so it does not have the same high spikes as in the first period.

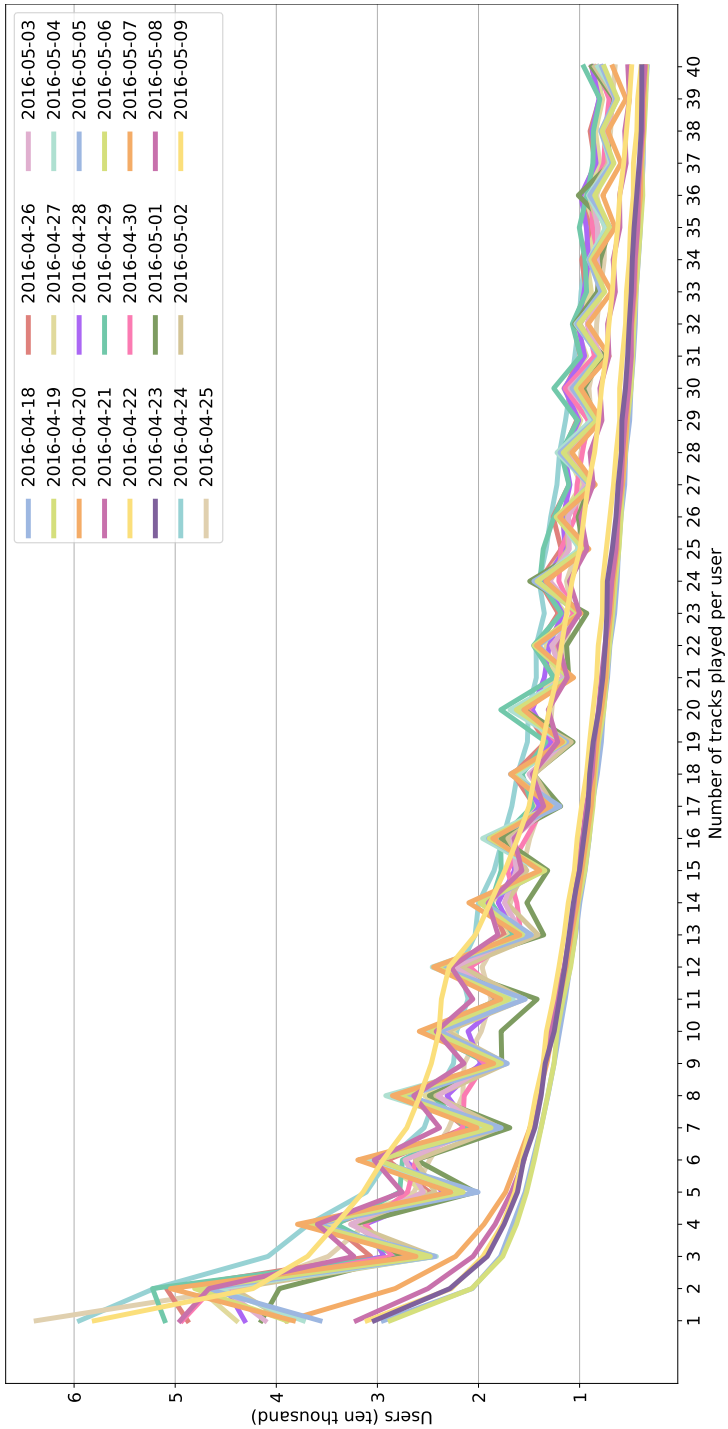


Figure 8.10: Analysis method 4 results period 2

8.6.4 Popular track findings

Figure 8.9 show two album releases during the same period. The first album was 'Anti' by Rihanna on 2016-01-28, which was made available for free digital download on 2016-01-27[1]. The difference between these two dates should come from the timezone of the log server and what is reported on Wikipedia. The second album was 'Life of Pablo' by Kanye West on 2016-02-14[2]. Subsection 8.6.5 also suggests that there were two large albums released in the first period. We have extracted the top played tracks on these two dates using analysis method 5. The results are found in Table 8.3.

The top thirteen tracks played on 2016-01-28 was all from Rihanna's album, with a total of 4 413 802 playbacks on this day alone. The highest track – not related to this album – was Formation (`trackid`: 57034935) by Beyoncé. The top eighteen tracks played on 2016-02-14 was all from Kanye's album, with a total of 33 331 035. Kanye's album had 655.15% more playbacks than Rihanna's album on their release dates. Table 8.4 gives an overview over which track IDs belongs to which album. The track IDs in parentheses are tracks which were added to Kanye's album at later dates. Finally, we also identified the album 'Lemonade' by Beyoncé Knowles from the suspicious days in the second period.

[1] https://en.wikipedia.org/wiki/Anti_(album)
[2] https://en.wikipedia.org/wiki/The_Life_of_Pablo

Table 8.3: Top 30 played tracks

2016-01-28		2016-02-14	
trackid	frequency	trackid	frequency
56677093	475172	57273410	2629970
56677094	404041	57273409	2517090
56677092	401949	57273412	2467229
56677090	383417	57273411	2414477
56677096	378201	57317924	2279144
56677095	353840	57273413	2154190
56677097	335576	57273418	2128204
56677091	313186	57273415	1890142
56677098	310798	57273414	1811385
56677101	295793	57273420	1773354
56677100	277148	57273416	1770112
56677102	242758	57273419	1644453
56677099	241923	57273417	1503686
56638583	67930	57317920	1387581
53960289	14580	57317923	1379380
51004025	14498	57317922	1235945
52901260	14390	57317919	1219114
53893678	14272	57317921	1125579
56290510	13691	57034935	288953
54511031	11280	56681096	155993
56677107	11242	57040670	139501
44094250	10059	57261945	123683
47497148	10028	56681099	114072
48351965	9682	56681097	108889
49671724	8987	56681095	92878
56677108	8925	56681093	87684
56372041	8783	56681100	87681
45323542	8688	56638583	83157
56677106	8401	56681098	80910
51579781	8166	56681101	74448

Table 8.4: Track IDs overview

'Anti' by Rihanna		'Life of Pablo' by Kanye West	
trackid	**Title**	trackid	**Title**
56677090	Consideration (feat. SZA)	57273409	Ultralight Beam
56677091	James Joint	57273410	Father Stretch My Hands Pt. 1
56677092	Kiss It Better	57273411	Pt. 2
56677093	Work	57273412	Famous
56677094	Desperado	57273413	Feedback
56677095	Woo	57273414	Low Lights
56677096	Needed Me	57273415	Highlights
56677097	Yeah, I Said It	57273416	Freestyle 4
56677098	Same Ol' Mistakes	57273417	I Love Kanye
56677099	Never Ending	57273418	FML
56677100	Love On The Brain	57273419	Real Friends
56677101	Higher	57273420	Wolves
56677102	Close To You	57317919	Siiiiiiiilver Surffffeeeeer Intermission
		57317920	30 Hours
		57317921	No More Parties in LA
		57317922	Facts (Charlie Heat Version)
		57317923	Fade
		57317924	Waves
		(58373775)	Frank's Track
		(61872799)	Saint Pablo

'Lemonade' by Beyoncé Knowles	
trackid	**Title**
59727857	Pray you catch me
59727858	Hold up
59727859	Don't hurt yourself
59727860	Sorry
59727861	6 inch
59727862	Daddy lessons
59727863	Love drought
59727864	Sandcastles
59727865	Forward
59727866	Freedom
59727867	All night
59727868	Formation
59727870	Pray you catch me
59727871	Hold up
59727872	Don't hurt yourself
59727873	Sorry
59727874	6 inch
59727875	Daddy lessons
59727876	Love drought
59727877	Sandcastles
59727878	Forward
59727879	Freedom
59727880	All night
59727881	Formation

8.6.5 Number of unique tracks findings

After the results from analysis method 4, in Subsection 8.6.3, we wanted to look closer at the number of unique/distinct tracks played per system user. Figure 8.11 show a small spike when Rihanna's album was released, when she was fully signed with her manager Jay Z's company *Roc Nation*. The next smaller bump at 16 is track IDs from the same Rihanna album but with different track IDs. The largest spike in the first period is exclusively caused by users listening to Kanye's album.

A notable distinction between these albums is that Kanye's 'Life of Pablo' continued to have an unnaturally high number of playbacks over several days. Although both Kanye and Rihanna have somewhat similar popularity, Rihanna's album flattened out much quicker. Even if this particular music streaming platform could have a certain user base, we would expect Kanye's spike to react in a similar fashion as Rihanna's spike. That is, a small spike (not quite as high as indicated by the figure) and then a fast decline and normalisation of the numbers.

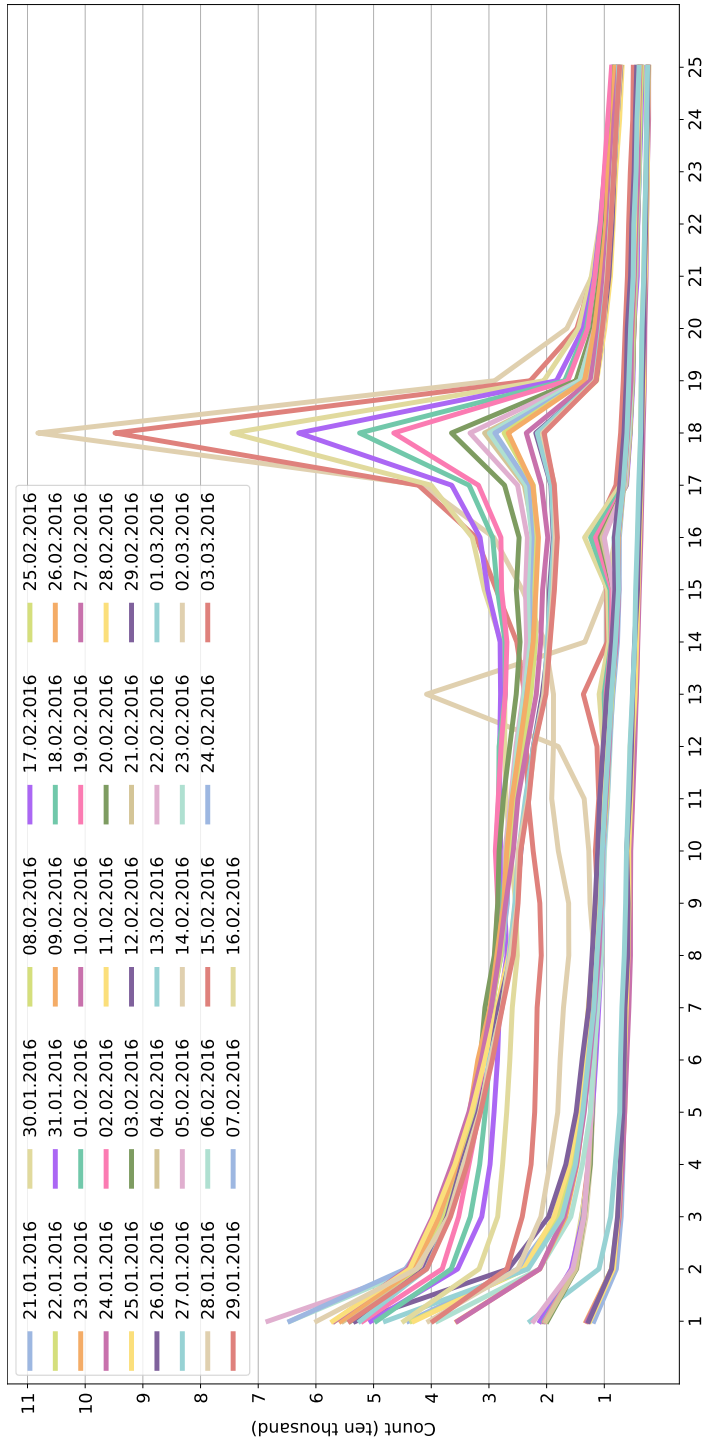


Figure 8.11: Analysis method 6 results period 1

Figure 8.12 show that Beyoncé's album has a similar behaviour as we found for Kanye's album. Beyoncé is married to Jay Z (Shawn Corey Carter). Her album was released on 2016-04-23; however, the time when this occurred in the logs is on 2016-04-24. A difference that can be explained by the different timezones for the log server and the release information found online.

Figure 8.12: Analysis method 6 results period 2

8.6.6 System user frequency findings

It is suspicious the extent to which users would choose to selectively listen to eighteen particular tracks in the first period, and twelve to thirteen tracks in the second period. We would expect them to be more similar behaviour as Rihanna's album: a small spike in interest for the album on release day, and then quickly subside. However, previous analysis steps show that this was not the case for 'Life of Pablo' and 'Lemonade'. The next natural step is to find out more about the system users who listened to these tracks. Figure 8.13 show that tracks from 'Life of Pablo' was played a lot more than any other track ID on 2016-02-14. However, when looking at the offline playback for these top 18 played tracks, we see they are played a lot more offline than online (Figure 8.14). Finally, finding the countries, Figure 8.15, where the users are from gave nothing interesting.

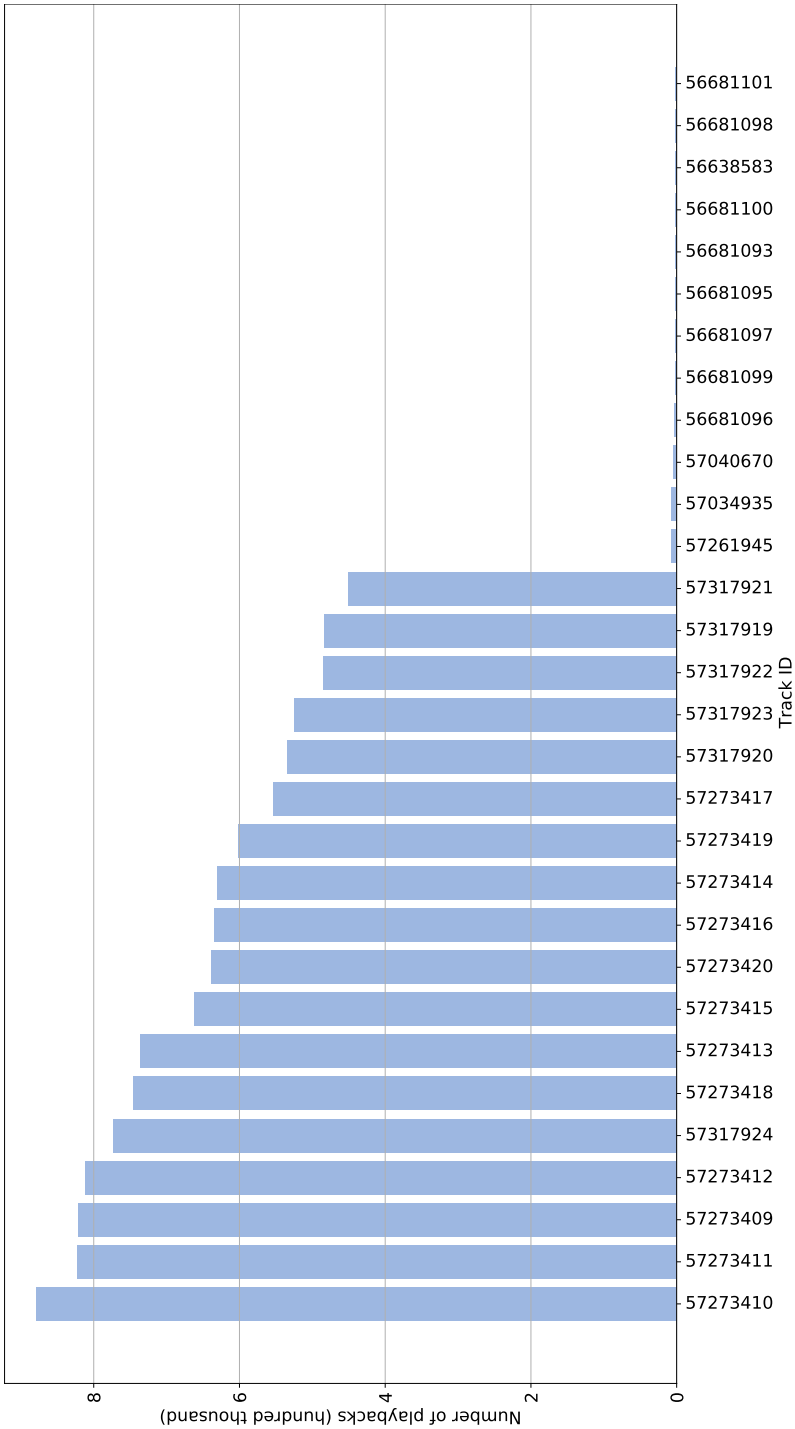


Figure 8.13: Top 30 played tracks on 2016-02-14

Figure 8.14: Top 18 tracks offline play on 2016-02-14

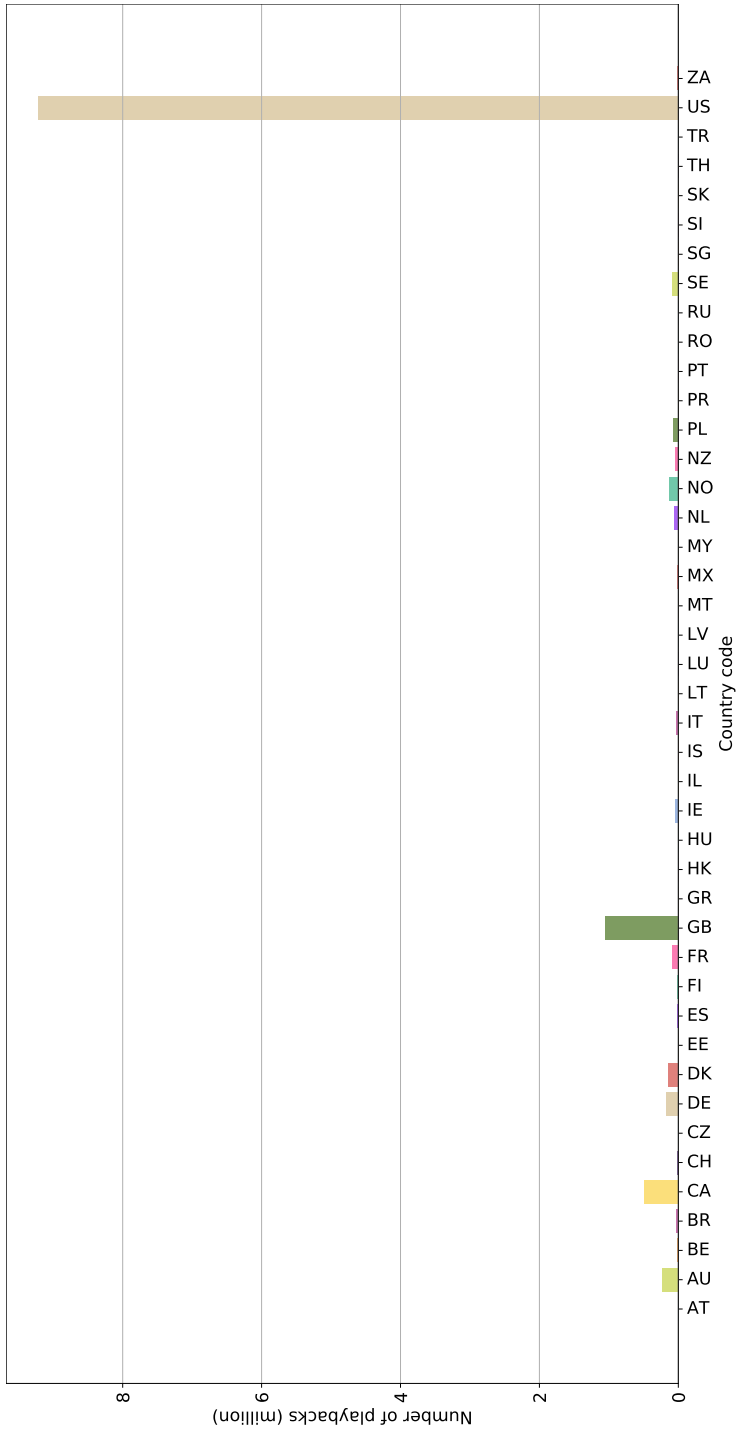


Figure 8.15: Countries playing the top 18 tracks on 2016-02-14

8.6.7 Binning findings

We identified some anomalies in the data, specifically the large number of users listening to them, so it was necessary to understand the system users they affect. For this, we divided the log entries into four distinctive bins based on the year in their timestamp. A short reminder for these bins: *exact* contains log entries for that exact date (respectively for the log file currently under investigation), while *around* contains any rows found +/- 2 years from 2016 (i.e. between 2014 and 2018). The two remaining bins *below* and *above* captures any remaining log entries with years prior to 2014 and post 2018 respectively.

Figure 8.16 show the analysis method 6 for each bin. The bins *below* and *above* is barely visible, while *around* follows our expectations. Note that when looking at the actual number, these three bins do not have the elevated spike on eighteen in frequency. This result suggests that log entries found with weird timestamps have a very low to no effect on the overall results found in previous analysis steps. Furthermore, this shows that those users have not been tampered with. Finally, the figure shows that most of the log entries can be found around or on the exact date, which means that log entries from the two smallest bins (i.e. below and above) can be removed when necessary.

Figure 8.16 show that the characteristics found in analysis method 6 only affects log entries in the *exact*-bin, which means that they were the only ones who showed this abnormal behaviour.

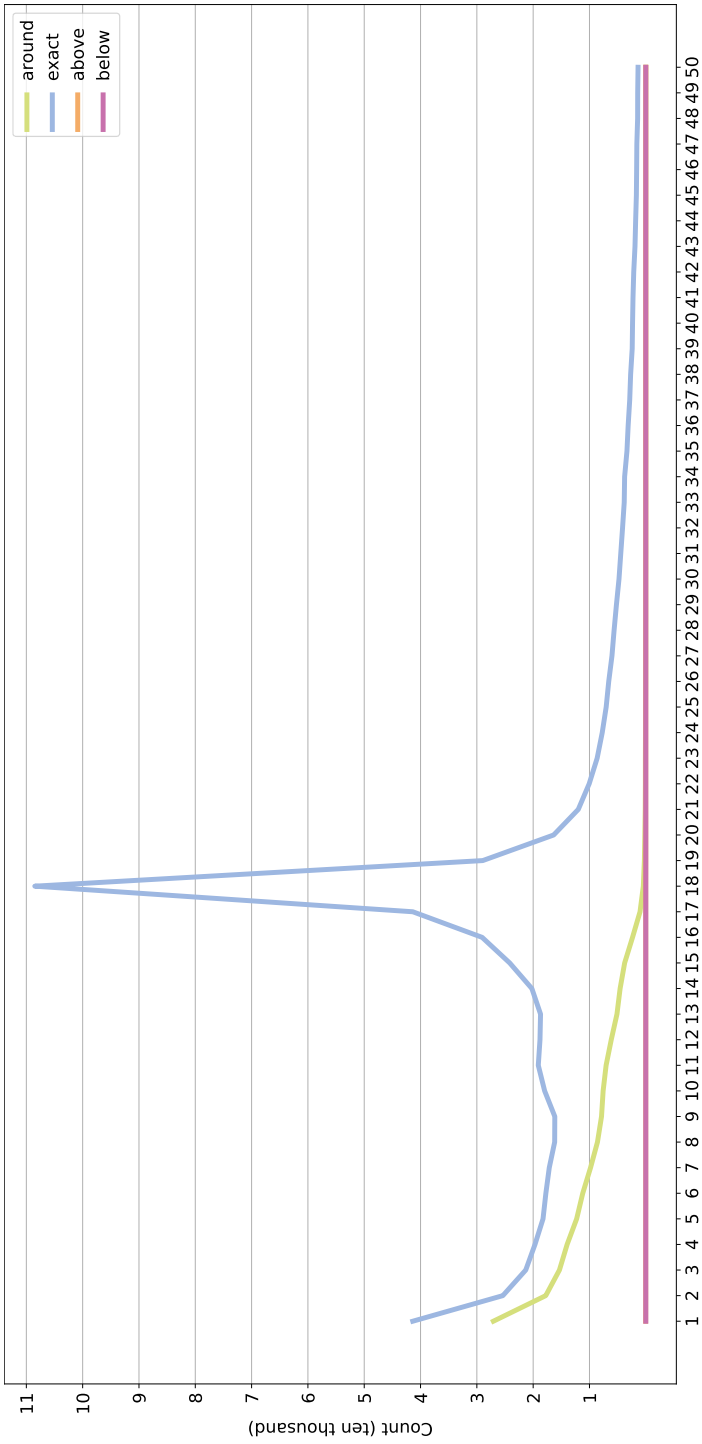


Figure 8.16: Analysis method 8 findings for 2016-02-14

8.6.8 Impossible scenario findings

As previously explained at the beginning of 8.6 Analysis Method 2 was conducted later in the investigation. As such, Analysis Method 2 is presented here, between the findings that preceded and followed Analysis Method 2. This analysis method was conducted later due to specific requests of DN. As each form of statistical analysis is independent of the other, it is of no consequence the order in which the analysis was carried out. For this analysis, we set out to identify occurrences of "impossible" scenarios. This includes four scenarios where the time and system user ID is identical: 1) with two identical tracks; 2) with three or more identical tracks; 3) with two unequal tracks, and 4) with three or more unequal tracks. We refer the reader back to Subsection 8.5.2 for an explanation about the difference between the terms 'identical' and 'unequal'.

The results from this analysis are in three different tables for each impossible scenario. The first table describes the *unique* system users found for each day (i.e. each user is only counted once), and how many of those was affected by the impossibility. An 'online' system user is anyone with at least one online playback for that day, while 'offline' users exclusively have offline playbacks. However, online system users can also have offline playbacks.

The second table focuses on the *unique* playbacks (log entries) found for each day. It will describe how many playbacks was affected by an impossible scenario. In addition, it describes whether the affected playbacks was online or offline. 'Online' playbacks in this table are any log entry marked as being played online, while 'offline' playbacks are marked as offline, 'N' or 'Y' respectively.

The third table enriches information found in the two previous tables. This table shows online users' online and offline playbacks. Thus, we can identify which of their playbacks was affected by these impossible scenarios. It also shows the same information for offline users; however, they do not have any online playbacks because of our definition of offline users.

Finally, the third table also shows how many of those affected playbacks (found in the second table) are from the two albums we previously identified in other analysis steps. More specifically: 'Life of Pablo' by Kanye West and 'Lemonade' by Beyoncé Knowles. A list with the track's IDs is found in Table 8.4. Note that each album is only counted in their respective periods, to understand how many of the affected playbacks are from each individual album.

Table 8.5 contains the results for system users which had exactly two identical duplicates for each day. Users affected with this characteristic occurred about five per cent per each day in both periods. Although we had expected a lower

percentage, we have an understanding that the logging system could sometimes incorrectly log one playback twice. We are unaware of how frequent this system error occurred. However, the next tables can help give an understanding of when it occurred.

While Table 8.5 only looks at the affected users with exactly two identical duplicates, Table 8.6 looks at the affected log entries themselves. Here, we can see that they normally affect about one per cent of the total playbacks per day. It is also noteworthy that they only affected offline playbacks, which is most likely caused by those devices which produce very strange timestamps. However, the exception to our observation is in the suspicious days in the first period, between 2016-02-14 and 2016-02-23. Online playbacks were also affected in this short period of ten days.

Table 8.5 shows that online users (i.e. who have at least one online playback) are affected by having two identical duplicates. However, Table 8.7 contains playbacks for affected users only. This clearly show that this only happened to offline playbacks. Again, with the exception of those ten days.

Table 8.5: Two duplicates system users

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-01-21	174620	7704	4.412	135221	5109	3.778	39399	2595	6.586
2016-01-22	188796	8592	4.551	149491	5841	3.907	39305	2751	6.999
2016-01-23	183237	7460	4.071	148330	5176	3.49	34907	2284	6.543
2016-01-24	171511	6958	4.057	137257	4812	3.506	34254	2146	6.265
2016-01-25	174896	7850	4.488	135543	5127	3.783	39353	2723	6.919
2016-01-26	180917	8020	4.433	140482	5412	3.852	40435	2608	6.45
2016-01-27	202757	28771	14.19	165342	20699	12.519	37415	8072	21.574
2016-01-28	356158	15844	4.449	307284	12788	4.162	48874	3056	6.253
2016-01-29	344910	16998	4.928	277724	12813	4.614	67186	4185	6.229
2016-01-30	324098	14662	4.524	254269	10593	4.166	69829	4069	5.827
2016-01-31	296886	12705	4.279	225798	8696	3.851	71088	4009	5.639
2016-02-01	300715	14548	4.838	222513	9575	4.303	78202	4973	6.359
2016-02-02	300530	14615	4.863	221130	9697	4.385	79400	4918	6.194
2016-02-03	296753	14221	4.792	216748	9409	4.341	80005	4812	6.015
2016-02-04	298583	14647	4.906	217279	9674	4.452	81304	4973	6.117
2016-02-05	312045	15187	4.867	234200	10397	4.439	77845	4790	6.153
2016-02-06	330846	13920	4.207	270729	10284	3.799	60117	3636	6.048
2016-02-07	337341	12624	3.742	287555	9781	3.401	49786	2843	5.71
2016-02-08	334768	14672	4.383	278546	11214	4.026	56222	3458	6.151
2016-02-09	328425	14635	4.456	260760	10516	4.033	67665	4119	6.087
2016-02-10	328089	14660	4.468	257006	10420	4.054	71083	4240	5.965
2016-02-11	355004	14795	4.168	281945	10488	3.72	73059	4307	5.895
2016-02-12	386880	15941	4.12	316405	11733	3.708	70475	4208	5.971
2016-02-13	358217	14293	3.99	291055	10548	3.624	67162	3745	5.576
2016-02-14	691041	363757	52.639	609112	331135	54.364	81929	32622	39.817
2016-02-15	747981	429170	57.377	574979	338434	58.86	173002	90736	52.448
2016-02-16	800171	455094	56.875	585803	343981	58.72	214368	111113	51.833
2016-02-17	815881	458633	56.213	574861	334762	58.234	241020	123871	51.394
2016-02-18	806257	440539	54.64	551943	315342	57.133	254314	125197	49.229
2016-02-19	817275	436639	53.426	556034	311420	56.007	261241	125219	47.932
2016-02-20	787117	399162	50.712	532886	286690	53.799	254231	112472	44.24
2016-02-21	732423	345356	47.153	486737	246436	50.63	245686	98920	40.263

Continued on next page

Continuation of Table 8.5

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-02-22	770746	370496	48.07	533728	271442	50.858	237018	99054	41.792
2016-02-23	764584	373093	48.797	529906	275788	52.045	234678	97305	41.463
2016-02-24	760152	36184	4.76	514411	23142	4.499	245741	13042	5.307
2016-02-25	747083	36595	4.898	496204	23373	4.71	250879	13222	5.27
2016-02-26	764275	37034	4.846	512378	24222	4.727	251897	12812	5.086
2016-02-27	741229	31397	4.236	494773	20657	4.175	246456	10740	4.358
2016-02-28	696194	27857	4.001	459151	17970	3.914	237043	9887	4.171
2016-02-29	691263	33240	4.809	449202	20931	4.66	242061	12309	5.085
2016-03-01	696388	33094	4.752	454878	21059	4.63	241510	12035	4.983
2016-03-02	700917	32573	4.647	461109	20799	4.511	239808	11774	4.91
2016-03-03	691915	32836	4.746	451042	21013	4.659	240873	11823	4.908
2016-04-18	461253	25089	5.439	340662	18112	5.317	120591	6977	5.786
2016-04-19	463153	25454	5.496	343033	18529	5.402	120120	6925	5.765
2016-04-20	504825	27172	5.382	402684	21197	5.264	102141	5975	5.85
2016-04-21	494942	25544	5.161	392501	19465	4.959	102441	6079	5.934
2016-04-22	530312	28734	5.418	429837	22468	5.227	100475	6266	6.236
2016-04-23	500494	24205	4.836	398927	18658	4.677	101567	5547	5.461
2016-04-24	1017351	41223	4.052	937338	36983	3.946	80013	4240	5.299
2016-04-25	1084330	52209	4.815	936931	43237	4.615	147399	8972	6.087
2016-04-26	1106615	83979	7.589	921532	66381	7.203	185083	17598	9.508
2016-04-27	1086552	52833	4.862	878715	40944	4.66	207837	11889	5.72
2016-04-28	1066626	52393	4.912	841933	40006	4.752	224693	12387	5.513
2016-04-29	1079856	51129	4.735	854740	39335	4.602	225116	11794	5.239
2016-04-30	1005141	173312	17.243	778704	132664	17.037	226437	40648	17.951
2016-05-01	922822	36300	3.934	703029	26947	3.833	219793	9353	4.255
2016-05-02	923637	60944	6.598	690843	44603	6.456	232794	16341	7.02
2016-05-03	912692	41667	4.565	668396	29758	4.452	244296	11909	4.875
2016-05-04	919118	41653	4.532	683132	29859	4.371	235986	11794	4.998
2016-05-05	889725	40945	4.602	659028	29640	4.498	230697	11305	4.9
2016-05-06	887489	41820	4.712	658611	30740	4.667	228878	11080	4.841
2016-05-07	851475	37153	4.363	628638	27294	4.342	222837	9859	4.424
2016-05-08	790803	32826	4.151	581512	23845	4.101	209291	8981	4.291
2016-05-09	807714	37653	4.662	590649	26945	4.562	217065	10708	4.933

Table 8.6: Two duplicates playbacks

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-01-21	5183365	58848	1.135	2887796	0	0.0	2295569	58848	2.564
2016-01-22	5699857	63454	1.113	3197499	0	0.0	2502358	63454	2.536
2016-01-23	5706073	56224	0.985	3320763	0	0.0	2385310	56224	2.357
2016-01-24	4990937	55104	1.104	2877218	0	0.0	2113719	55104	2.607
2016-01-25	5034490	58916	1.17	2775176	0	0.0	2259314	58916	2.608
2016-01-26	5211451	57856	1.11	2861114	0	0.0	2350337	57856	2.462
2016-01-27	5699874	175810	3.084	3115105	0	0.0	2584769	175810	6.802
2016-01-28	10716712	94554	0.882	5406227	0	0.0	5310485	94554	1.781
2016-01-29	10860619	102304	0.942	5522217	0	0.0	5338402	102304	1.916
2016-01-30	10008250	89884	0.898	5200705	0	0.0	4807545	89884	1.87
2016-01-31	8445037	78374	0.928	4380686	0	0.0	4064351	78374	1.928
2016-02-01	8595181	91098	1.06	4260810	0	0.0	4334371	91098	2.102
2016-02-02	8469929	91928	1.085	4153867	0	0.0	4316062	91928	2.13
2016-02-03	8386787	90132	1.075	4130617	0	0.0	4256170	90132	2.118
2016-02-04	8410214	96704	1.15	4128894	0	0.0	4281320	96704	2.259
2016-02-05	9001255	102532	1.139	4503223	0	0.0	4498032	102532	2.279
2016-02-06	8841003	106424	1.204	4623472	0	0.0	4217531	106424	2.523
2016-02-07	8375341	89782	1.072	4562874	0	0.0	3812467	89782	2.355
2016-02-08	8145940	89506	1.099	4244968	0	0.0	3900972	89506	2.294
2016-02-09	8439985	94938	1.125	4289622	0	0.0	4150363	94938	2.287
2016-02-10	8437553	94282	1.117	4261212	0	0.0	4176341	94282	2.258
2016-02-11	8748323	102980	1.177	4431643	0	0.0	4316680	102980	2.386
2016-02-12	9591865	103062	1.074	4932532	0	0.0	4659333	103062	2.212
2016-02-13	9207400	99784	1.084	4815322	0	0.0	4392078	99784	2.272
2016-02-14	52264508	6115298	11.701	25683182	2850730	11.1	26581326	3264568	12.281
2016-02-15	62315131	7164866	11.498	28423017	3053831	10.744	33892114	4111035	12.13
2016-02-16	62575347	7341568	11.732	27216556	2938168	10.796	35358791	4403400	12.453
2016-02-17	61635277	7330456	11.893	26096222	2831421	10.85	35539055	4499035	12.659

Continued on next page

Continuation of Table 8.6

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-02-18	58176702	6994444	12.023	24279133	2640953	10.877	33897569	4353491	12.843
2016-02-19	57453384	6920768	12.046	23863079	2570530	10.772	33590305	4350238	12.951
2016-02-20	51238240	6161090	12.024	21695419	2310437	10.649	29542821	3850653	13.034
2016-02-21	44012913	5147548	11.696	19017377	1958538	10.299	24995536	3189010	12.758
2016-02-22	48123726	5730262	11.907	20139829	2136875	10.61	27983897	3593387	12.841
2016-02-23	47818784	5712280	11.946	19908725	2114657	10.622	27910059	3597623	12.89
2016-02-24	19205376	206646	1.076	8020004	0	0.0	11185372	206646	1.847
2016-02-25	19155488	210692	1.1	7933048	0	0.0	11222440	210692	1.877
2016-02-26	19895000	217950	1.096	8316062	0	0.0	11578938	217950	1.882
2016-02-27	18410754	188234	1.022	7817765	0	0.0	10592989	188234	1.777
2016-02-28	16341019	169526	1.037	7006815	0	0.0	9334204	169526	1.816
2016-02-29	17041782	189676	1.113	7049968	0	0.0	9991814	189676	1.898
2016-03-01	17519578	197122	1.125	7264171	0	0.0	10255407	197122	1.922
2016-03-02	17260480	193074	1.119	7182545	0	0.0	10077935	193074	1.916
2016-03-03	17326476	181916	1.05	7175051	0	0.0	10151425	181916	1.792
2016-04-18	13094984	157950	1.206	5480281	0	0.0	7614703	157950	2.074
2016-04-19	13228953	160584	1.214	5561693	0	0.0	7667260	160584	2.094
2016-04-20	13716969	177984	1.298	5943181	0	0.0	7773788	177984	2.29
2016-04-21	13592572	161782	1.19	6080560	6	0.0	7512012	161776	2.154
2016-04-22	15716545	179924	1.145	7204431	0	0.0	8512114	179924	2.114
2016-04-23	14539943	147412	1.014	6686559	0	0.0	7853384	147412	1.877
2016-04-24	32844244	239336	0.729	16533941	2	0.0	16310303	239334	1.467
2016-04-25	61648530	448352	0.727	27532332	0	0.0	34116198	448352	1.314
2016-04-26	51762664	707678	1.367	23271706	1	0.0	28490958	707677	2.484
2016-04-27	51443095	435634	0.847	22441805	0	0.0	29001290	435634	1.502
2016-04-28	49571047	455708	0.919	21060520	0	0.0	28510527	455708	1.598
2016-04-29	45740028	441684	0.966	19304866	0	0.0	26435162	441684	1.671
2016-04-30	44747668	1917000	4.284	18482843	529918	2.867	26264825	1387082	5.281
2016-05-01	38826732	335504	0.864	16407601	0	0.0	22419131	335504	1.497
2016-05-02	36971583	432736	1.17	15561500	0	0.0	21410083	432736	2.021
2016-05-03	34043783	344360	1.012	14283511	0	0.0	19760272	344360	1.743
2016-05-04	32862327	339706	1.034	14013298	0	0.0	18849029	339706	1.802

Continued on next page

Continuation of Table 8.6

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-05-05	31535242	323076	1.024	13427299	0	0.0	18107943	323076	1.784
2016-05-06	30227071	318780	1.055	12798897	0	0.0	17428174	318780	1.829
2016-05-07	26738540	265672	0.994	11477044	0	0.0	15261496	265672	1.741
2016-05-08	20337761	207420	1.02	8893020	0	0.0	11444741	207420	1.812
2016-05-09	19254690	212336	1.103	8408779	0	0.0	10845911	212336	1.958

Table 8.7: Two duplicates affected users playbacks

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-01-21	0	33878	0	24970	0	0.0	0	0
2016-01-22	0	36370	0	27084	0	0.0	0	0
2016-01-23	0	32640	0	23584	0	0.0	0	0
2016-01-24	0	33716	0	21388	0	0.0	0	0
2016-01-25	0	34346	0	24570	0	0.0	0	0
2016-01-26	0	35116	0	22740	0	0.0	0	0
2016-01-27	0	118484	0	57326	0	0.0	0	0
2016-01-28	0	66080	0	28474	0	0.0	0	0
2016-01-29	0	69080	0	33224	0	0.0	0	0
2016-01-30	0	56312	0	33572	0	0.0	0	0
2016-01-31	0	48950	0	29424	0	0.0	0	0
2016-02-01	0	54500	0	36598	0	0.0	0	0
2016-02-02	0	54490	0	37438	0	0.0	0	0
2016-02-03	0	53484	0	36648	0	0.0	0	0
2016-02-04	0	56070	0	40634	0	0.0	0	0
2016-02-05	0	63542	0	38990	0	0.0	0	0
2016-02-06	0	68708	0	37716	0	0.0	0	0
2016-02-07	0	63606	0	26176	0	0.0	0	0
2016-02-08	0	62644	0	26862	0	0.0	0	0
2016-02-09	0	60390	0	34548	0	0.0	0	0
2016-02-10	0	58968	0	35314	0	0.0	0	0
2016-02-11	0	64546	0	38434	0	0.0	0	0
2016-02-12	0	69176	0	33886	0	0.0	0	0
2016-02-13	0	63894	0	35890	0	0.0	0	0
2016-02-14	2850730	2723796	0	540772	4335446	70.895	2020349	2315097
2016-02-15	3053831	2623869	0	1487166	4937780	68.917	2015315	2922465
2016-02-16	2938168	2657648	0	1745752	4780636	65.117	1769607	3011029
2016-02-17	2831421	2605113	0	1893922	4469046	60.965	1537398	2931648
2016-02-18	2640953	2465963	0	1887528	3990740	57.056	1307609	2683131
2016-02-19	2570530	2475424	0	1874814	3694746	53.386	1165928	2528818
2016-02-20	2310437	2220357	0	1630296	3065756	49.76	920394	2145362

Continued on next page

Continuation of Table 8.7

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-02-21	1958538	1797040	0	1391970	2433134	47.268	737798	1695336
2016-02-22	2136875	2147001	0	1446386	2661058	46.439	806081	1854977
2016-02-23	2114657	2181111	0	1416512	2461764	43.096	718301	1743463
2016-02-24	0	121272	0	85374	64172	31.054	0	64172
2016-02-25	0	120010	0	90682	62234	29.538	0	62234
2016-02-26	0	129004	0	88946	60750	27.873	0	60750
2016-02-27	0	111954	0	76280	47036	24.988	0	47036
2016-02-28	0	98810	0	70716	40556	23.923	0	40556
2016-02-29	0	111860	0	77816	52172	27.506	0	52172
2016-03-01	0	114272	0	82850	49746	25.236	0	49746
2016-03-02	0	112806	0	80268	45960	23.804	0	45960
2016-03-03	0	104922	0	76994	42264	23.233	0	42264
2016-04-18	0	97324	0	60626	0	0.0	0	0
2016-04-19	0	98180	0	62404	0	0.0	0	0
2016-04-20	0	121626	0	56358	0	0.0	0	0
2016-04-21	6	105504	0	56272	0	0.0	0	0
2016-04-22	0	120990	0	58934	0	0.0	0	0
2016-04-23	0	95278	0	52134	0	0.0	0	0
2016-04-24	2	194846	0	44488	85528	35.736	2	85526
2016-04-25	0	345322	0	103030	296690	66.173	0	296690
2016-04-26	1	525389	0	182288	526570	74.408	1	526569
2016-04-27	0	301560	0	134074	243360	55.863	0	243360
2016-04-28	0	307510	0	148198	246410	54.072	0	246410
2016-04-29	0	300492	0	141192	223346	50.567	0	223346
2016-04-30	529918	904004	0	483078	1738460	90.686	529918	1208542
2016-05-01	0	213724	0	121780	168930	50.351	0	168930
2016-05-02	0	281466	0	151270	203350	46.992	0	203350
2016-05-03	0	211672	0	132688	133510	38.77	0	133510
2016-05-04	0	208204	0	131502	124288	36.587	0	124288
2016-05-05	0	200344	0	122732	110300	34.141	0	110300
2016-05-06	0	202460	0	116320	90368	28.348	0	90368
2016-05-07	0	170736	0	94936	61690	23.22	0	61690

Continued on next page

Continuation of Table 8.7

Date	Online online backs users play-	Online offline backs users play-	Offline online backs users play-	Offline offline backs users play-	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-05-08	0	128888	0	78532	46584	22.459	0	46584
2016-05-09	0	131302	0	81034	44788	21.093	0	44788

We understand that two identical log entries could be caused by a fault in the system. However, we consider three or more duplicate entries to be very unlikely to occur. Table 8.8 shows that about one per cent of all users was normally affected by this impossibility during normal days in the first period while raising to thirty per cent during the suspicious days.

Table 8.9 show that half a per cent of all playbacks during each day was marked as a three or more duplicate. Finally, Table 8.10 show that these playbacks almost exclusively played tracks from the two albums. This impossible scenario even continues after the ten suspicious days in the first period. Although this scenario affected very few playbacks in the second period, most of those affected playbacks were from tracks from 'Lemonade' by Beyoncé.

Table 8.8: Three or more duplicates system users

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-01-21	174620	2092	1.198	135221	1336	0.988	39399	756	1.919
2016-01-22	188796	2241	1.187	149491	1454	0.973	39305	787	2.002
2016-01-23	183237	2021	1.103	148330	1330	0.897	34907	691	1.98
2016-01-24	171511	1855	1.082	137257	1224	0.892	34254	631	1.842
2016-01-25	174896	2217	1.268	135543	1423	1.05	39353	794	2.018
2016-01-26	180917	2274	1.257	140482	1441	1.026	40435	833	2.06
2016-01-27	202757	8447	4.166	165342	6089	3.683	37415	2358	6.302
2016-01-28	356158	3795	1.066	307284	2916	0.949	48874	879	1.799
2016-01-29	344910	4207	1.22	277724	3142	1.131	67186	1065	1.585
2016-01-30	324098	3659	1.129	254269	2636	1.037	69829	1023	1.465
2016-01-31	296886	3255	1.096	225798	2230	0.988	71088	1025	1.442
2016-02-01	300715	3527	1.173	222513	2325	1.045	78202	1202	1.537
2016-02-02	300530	3612	1.202	221130	2381	1.077	79400	1231	1.55
2016-02-03	296753	3675	1.238	216748	2391	1.103	80005	1284	1.605
2016-02-04	298583	3772	1.263	217279	2462	1.133	81304	1310	1.611
2016-02-05	312045	3911	1.253	234200	2635	1.125	77845	1276	1.639
2016-02-06	330846	3504	1.059	270729	2524	0.932	60117	980	1.63
2016-02-07	337341	3182	0.943	287555	2368	0.823	49786	814	1.635
2016-02-08	334768	3638	1.087	278546	2718	0.976	56222	920	1.636
2016-02-09	328425	3726	1.135	260760	2576	0.988	67665	1150	1.7
2016-02-10	328089	3598	1.097	257006	2525	0.982	71083	1073	1.51
2016-02-11	355004	3665	1.032	281945	2576	0.914	73059	1089	1.491
2016-02-12	386880	3995	1.033	316405	2836	0.896	70475	1159	1.645
2016-02-13	358217	3572	0.997	291055	2526	0.868	67162	1046	1.557
2016-02-14	691041	236330	34.199	609112	218266	35.833	81929	18064	22.048
2016-02-15	747981	306128	40.927	574979	244031	42.442	173002	62097	35.894
2016-02-16	800171	317865	39.725	585803	243222	41.519	214368	74643	34.82
2016-02-17	815881	316986	38.852	574861	235908	41.037	241020	81078	33.64
2016-02-18	806257	291439	36.147	551943	213094	38.608	254314	78345	30.806
2016-02-19	817275	280498	34.321	556034	204342	36.75	261241	76156	29.152
2016-02-20	787117	239023	30.367	532886	175506	32.935	254231	63517	24.984
2016-02-21	732423	197588	26.977	486737	145476	29.888	245686	52112	21.211

Continued on next page

Continuation of Table 8.8

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-02-22	770746	223205	28.96	533728	168209	31.516	237018	54996	23.203
2016-02-23	764584	221815	29.011	529906	168688	31.834	234678	53127	22.638
2016-02-24	760152	8199	1.079	514411	5357	1.041	245741	2842	1.157
2016-02-25	747083	8107	1.085	496204	5206	1.049	250879	2901	1.156
2016-02-26	764275	8288	1.084	512378	5632	1.099	251897	2656	1.054
2016-02-27	741229	6807	0.918	494773	4538	0.917	246456	2269	0.921
2016-02-28	696194	6162	0.885	459151	4022	0.876	237043	2140	0.903
2016-02-29	691263	7313	1.058	449202	4735	1.054	242061	2578	1.065
2016-03-01	696388	7537	1.082	454878	4804	1.056	241510	2733	1.132
2016-03-02	700917	7179	1.024	461109	4597	0.997	239808	2582	1.077
2016-03-03	691915	7333	1.06	451042	4706	1.043	240873	2627	1.091
2016-04-18	461253	6295	1.365	340662	4602	1.351	120591	1693	1.404
2016-04-19	463153	6359	1.373	343033	4597	1.34	120120	1762	1.467
2016-04-20	504825	6644	1.316	402684	5109	1.269	102141	1535	1.503
2016-04-21	494942	6069	1.226	392501	4616	1.176	102441	1453	1.418
2016-04-22	530312	7199	1.358	429837	5617	1.307	100475	1582	1.575
2016-04-23	500494	5927	1.184	398927	4490	1.126	101567	1437	1.415
2016-04-24	1017351	9819	0.965	937338	8693	0.927	80013	1126	1.407
2016-04-25	1084330	12092	1.115	936931	10101	1.078	147399	1991	1.351
2016-04-26	1106615	13144	1.188	921532	10570	1.147	185083	2574	1.391
2016-04-27	1086552	12549	1.155	878715	9839	1.12	207837	2710	1.304
2016-04-28	1066626	12460	1.168	841933	9714	1.154	224693	2746	1.222
2016-04-29	1079856	12403	1.149	854740	9653	1.129	225116	2750	1.222
2016-04-30	1005141	11665	1.161	778704	8814	1.132	226437	2851	1.259
2016-05-01	922822	8503	0.921	703029	6376	0.907	219793	2127	0.968
2016-05-02	923637	13502	1.462	690843	10135	1.467	232794	3367	1.446
2016-05-03	912692	10173	1.115	668396	7330	1.097	244296	2843	1.164
2016-05-04	919118	10111	1.1	683132	7316	1.071	235986	2795	1.184
2016-05-05	889725	10070	1.132	659028	7411	1.125	230697	2659	1.153
2016-05-06	887489	10291	1.16	658611	7594	1.153	228878	2697	1.178
2016-05-07	851475	8765	1.029	628638	6442	1.025	222837	2323	1.042
2016-05-08	790803	7993	1.011	581512	5882	1.012	209291	2111	1.009
2016-05-09	807714	9267	1.147	590649	6639	1.124	217065	2628	1.211

Table 8.9: Three or more duplicates playbacks

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-01-21	5183365	30416	0.587	2887796	0	0.0	2295569	30416	1.325
2016-01-22	5699857	32780	0.575	3197499	0	0.0	2502358	32780	1.31
2016-01-23	5706073	33559	0.588	3320763	0	0.0	2385310	33559	1.407
2016-01-24	4990937	28291	0.567	2877218	0	0.0	2113719	28291	1.338
2016-01-25	5034490	33240	0.66	2775176	0	0.0	2259314	33240	1.471
2016-01-26	5211451	34479	0.662	2861114	0	0.0	2350337	34479	1.467
2016-01-27	5699874	74644	1.31	3115105	0	0.0	2584769	74644	2.888
2016-01-28	10716712	49834	0.465	5406227	0	0.0	5310485	49834	0.938
2016-01-29	10860619	52553	0.484	5522217	0	0.0	5338402	52553	0.984
2016-01-30	10008250	48992	0.49	5200705	0	0.0	4807545	48992	1.019
2016-01-31	8445037	43832	0.519	4380686	0	0.0	4064351	43832	1.078
2016-02-01	8595181	46009	0.535	4260810	0	0.0	4334371	46009	1.061
2016-02-02	8469929	49782	0.588	4153867	0	0.0	4316062	49782	1.153
2016-02-03	8386787	44969	0.536	4130617	0	0.0	4256170	44969	1.057
2016-02-04	8410214	47370	0.563	4128894	0	0.0	4281320	47370	1.106
2016-02-05	9001255	48487	0.539	4503223	0	0.0	4498032	48487	1.078
2016-02-06	8841003	48479	0.548	4623472	0	0.0	4217531	48479	1.149
2016-02-07	8375341	43932	0.525	4562874	0	0.0	3812467	43932	1.152
2016-02-08	8145940	45992	0.565	4244968	0	0.0	3900972	45992	1.179
2016-02-09	8439985	49478	0.586	4289622	0	0.0	4150363	49478	1.192
2016-02-10	8437553	48640	0.576	4261212	0	0.0	4176341	48640	1.165
2016-02-11	8748323	46608	0.533	4431643	0	0.0	4316680	46608	1.08
2016-02-12	9591865	55185	0.575	4932532	0	0.0	4659333	55185	1.184
2016-02-13	9207400	46770	0.508	4815322	0	0.0	4392078	46770	1.065
2016-02-14	52264508	10690169	20.454	25683182	4934123	19.211	26581326	5756046	21.654
2016-02-15	62315131	15050131	24.152	28423017	6426009	22.608	33892114	8624122	25.446
2016-02-16	62575347	13856459	22.144	27216556	5482933	20.146	35358791	8373526	23.682
2016-02-17	61635277	12920354	20.963	26096222	4965209	19.027	35539055	7955145	22.384

Continued on next page

Continuation of Table 8.9

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-02-18	58176702	11153454	19.172	24279133	4191485	17.264	33897569	6961969	20.538
2016-02-19	57453384	10274083	17.882	23863079	3806219	15.95	33590305	6467864	19.255
2016-02-20	51238240	7808483	15.24	21695419	2927268	13.493	29542821	4881215	16.523
2016-02-21	44012913	6206960	14.103	19017377	2436149	12.81	24995536	3770811	15.086
2016-02-22	48123726	7557730	15.705	20139829	2861974	14.211	27983897	4695756	16.78
2016-02-23	47818784	7258902	15.18	19908725	2723275	13.679	27910059	4535627	16.251
2016-02-24	19205376	83270	0.434	8020004	0	0.0	11185372	83270	0.744
2016-02-25	19155488	88747	0.463	7933048	0	0.0	11222440	88747	0.791
2016-02-26	19895000	85875	0.432	8316062	0	0.0	11578938	85875	0.742
2016-02-27	18410754	75812	0.412	7817765	0	0.0	10592989	75812	0.716
2016-02-28	16341019	69772	0.427	7006815	0	0.0	9334204	69772	0.747
2016-02-29	17041782	76754	0.45	7049968	0	0.0	9991814	76754	0.768
2016-03-01	17519578	81243	0.464	7264171	0	0.0	10255407	81243	0.792
2016-03-02	17260480	73776	0.427	7182545	0	0.0	10077935	73776	0.732
2016-03-03	17326476	82016	0.473	7175051	0	0.0	10151425	82016	0.808
2016-04-18	13094984	79404	0.606	5480281	0	0.0	7614703	79404	1.043
2016-04-19	13228953	78372	0.592	5561693	0	0.0	7667260	78372	1.022
2016-04-20	13716969	79478	0.579	5943181	0	0.0	7773788	79478	1.022
2016-04-21	13592572	76882	0.566	6080560	0	0.0	7512012	76882	1.023
2016-04-22	15716545	90678	0.577	7204431	0	0.0	8512114	90678	1.065
2016-04-23	14539943	73443	0.505	6686559	0	0.0	7853384	73443	0.935
2016-04-24	32844244	117585	0.358	16533941	0	0.0	16310303	117585	0.721
2016-04-25	61648530	199864	0.324	27532332	0	0.0	34116198	199864	0.586
2016-04-26	51762664	200859	0.388	23271706	0	0.0	28490958	200859	0.705
2016-04-27	51443095	207079	0.403	22441805	0	0.0	29001290	207079	0.714
2016-04-28	49571047	207479	0.419	21060520	0	0.0	28510527	207479	0.728
2016-04-29	45740028	207750	0.454	19304866	0	0.0	26435162	207750	0.786
2016-04-30	44747668	186791	0.417	18482843	0	0.0	26264825	186791	0.711
2016-05-01	38826732	152038	0.392	16407601	0	0.0	22419131	152038	0.678
2016-05-02	36971583	184537	0.499	15561500	0	0.0	21410083	184537	0.862
2016-05-03	34043783	173450	0.509	14283511	0	0.0	19760272	173450	0.878
2016-05-04	32862327	159560	0.486	14013298	0	0.0	18849029	159560	0.847

Continued on next page

Continuation of Table 8.9

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-05-05	31535242	154211	0.489	13427299	0	0.0	18107943	154211	0.852
2016-05-06	30227071	156603	0.518	12798897	0	0.0	17428174	156603	0.899
2016-05-07	26738540	122615	0.459	11477044	0	0.0	15261496	122615	0.803
2016-05-08	20337761	101036	0.497	8893020	0	0.0	11444741	101036	0.883
2016-05-09	19254690	111609	0.58	8408779	0	0.0	10845911	111609	1.029

Table 8.10: Three or more duplicates affected users playbacks

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-01-21	0	17598	0	12818	0	0.0	0	0
2016-01-22	0	18372	0	14408	0	0.0	0	0
2016-01-23	0	17779	0	15780	0	0.0	0	0
2016-01-24	0	14928	0	13363	0	0.0	0	0
2016-01-25	0	19185	0	14055	0	0.0	0	0
2016-01-26	0	18234	0	16245	0	0.0	0	0
2016-01-27	0	49982	0	24662	0	0.0	0	0
2016-01-28	0	33523	0	16311	0	0.0	0	0
2016-01-29	0	36236	0	16317	0	0.0	0	0
2016-01-30	0	30667	0	18325	0	0.0	0	0
2016-01-31	0	26044	0	17788	0	0.0	0	0
2016-02-01	0	26963	0	19046	0	0.0	0	0
2016-02-02	0	27969	0	21813	0	0.0	0	0
2016-02-03	0	24848	0	20121	0	0.0	0	0
2016-02-04	0	24972	0	22398	0	0.0	0	0
2016-02-05	0	29593	0	18894	0	0.0	0	0
2016-02-06	0	29382	0	19097	0	0.0	0	0
2016-02-07	0	26963	0	16969	0	0.0	0	0
2016-02-08	0	28126	0	17866	0	0.0	0	0
2016-02-09	0	29378	0	20100	0	0.0	0	0
2016-02-10	0	30090	0	18550	0	0.0	0	0
2016-02-11	0	28201	0	18407	0	0.0	0	0
2016-02-12	0	34506	0	20679	0	0.0	0	0
2016-02-13	0	27848	0	18922	0	0.0	0	0
2016-02-14	4934123	5089845	0	666201	8947906	83.702	4109839	4838067
2016-02-15	6426009	5751868	0	2872254	12720721	84.522	5336664	7384057
2016-02-16	5482933	5301567	0	3071959	11193926	80.785	4248197	6945729
2016-02-17	4965209	4811611	0	3143534	9847623	76.218	3527806	6319817
2016-02-18	4191485	4139640	0	2822329	8016034	71.87	2777285	5238749
2016-02-19	3806219	3817383	0	2650481	6962242	67.765	2345634	4616608
2016-02-20	2927268	2896475	0	1984740	4815761	61.673	1582917	3232844

Continued on next page

Continuation of Table 8.10

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-02-21	2436149	2219096	0	1551715	3514726	56.626	1209407	2305319
2016-02-22	2861974	2928328	0	1767428	4331472	57.312	1459880	2871592
2016-02-23	2723275	2863272	0	1672355	3816074	52.571	1241066	2575008
2016-02-24	0	48917	0	34353	25499	30.622	0	25499
2016-02-25	0	52995	0	35752	24271	27.349	0	24271
2016-02-26	0	51778	0	34097	23480	27.342	0	23480
2016-02-27	0	45369	0	30443	19152	25.262	0	19152
2016-02-28	0	40244	0	29528	16094	23.067	0	16094
2016-02-29	0	45581	0	31173	19153	24.954	0	19153
2016-03-01	0	44969	0	36274	18257	22.472	0	18257
2016-03-02	0	41986	0	31790	17715	24.012	0	17715
2016-03-03	0	44171	0	37845	17613	21.475	0	17613
2016-04-18	0	51483	0	27921	0	0.0	0	0
2016-04-19	0	46763	0	31609	0	0.0	0	0
2016-04-20	0	53819	0	25659	0	0.0	0	0
2016-04-21	0	51787	0	25095	0	0.0	0	0
2016-04-22	0	61888	0	28790	0	0.0	0	0
2016-04-23	0	47083	0	26360	0	0.0	0	0
2016-04-24	0	94811	0	22774	44928	38.209	0	44928
2016-04-25	0	157064	0	42800	119152	59.617	0	119152
2016-04-26	0	147654	0	53205	106938	53.24	0	106938
2016-04-27	0	145923	0	61156	110788	53.5	0	110788
2016-04-28	0	148338	0	59141	108146	52.124	0	108146
2016-04-29	0	146261	0	61489	95668	46.05	0	95668
2016-04-30	0	127097	0	59694	98623	52.799	0	98623
2016-05-01	0	97661	0	54377	68871	45.299	0	68871
2016-05-02	0	121910	0	62627	76510	41.461	0	76510
2016-05-03	0	105968	0	67482	66110	38.115	0	66110
2016-05-04	0	99748	0	59812	54660	34.257	0	54660
2016-05-05	0	100792	0	53419	49314	31.978	0	49314
2016-05-06	0	98660	0	57943	44189	28.217	0	44189
2016-05-07	0	81499	0	41116	28903	23.572	0	28903

Continued on next page

Continuation of Table 8.10

Date	Online online backs	users play-	Online offline backs	users play-	Offline online backs	users play-	Offline offline backs	users play-	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-05-08	0		64545		0		36491		20781	20.568	0	20781
2016-05-09	0		68468		0		43141		22992	20.6	0	22992

Identical duplicates can occur due to a fault in the system. There are several potential reasons for this, such as cache fault and poor design and implementation. However, *unequal* duplicates should almost never occur in the logs. With a possible exception in log entries received from devices with weird timestamps. Which means that occurrences from this impossible scenario should be low.

Two unequal duplicates mean that two playbacks have the identical timestamp and system user ID, while differing in track IDs. The results in Table 8.11 show that this was normally an infrequent occurrence, with the exception of two suspicious periods. Table 8.12 and 8.13 continues to demonstrate that a lot of users and playbacks was affected by this scenario in those suspicious periods.

Table 8.11: Two unequal duplicates system users

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-01-21	174620	366	0.21	135221	226	0.167	39399	140	0.355
2016-01-22	188796	361	0.191	149491	243	0.163	39305	118	0.3
2016-01-23	183237	361	0.197	148330	244	0.164	34907	117	0.335
2016-01-24	171511	288	0.168	137257	190	0.138	34254	98	0.286
2016-01-25	174896	332	0.19	135543	198	0.146	39353	134	0.341
2016-01-26	180917	338	0.187	140482	212	0.151	40435	126	0.312
2016-01-27	202757	354	0.175	165342	221	0.134	37415	133	0.355
2016-01-28	356158	419	0.118	307284	279	0.091	48874	140	0.286
2016-01-29	344910	438	0.127	277724	301	0.108	67186	137	0.204
2016-01-30	324098	398	0.123	254269	275	0.108	69829	123	0.176
2016-01-31	296886	334	0.113	225798	223	0.099	71088	111	0.156
2016-02-01	300715	348	0.116	222513	207	0.093	78202	141	0.18
2016-02-02	300530	378	0.126	221130	229	0.104	79400	149	0.188
2016-02-03	296753	343	0.116	216748	192	0.089	80005	151	0.189
2016-02-04	298583	383	0.128	217279	247	0.114	81304	136	0.167
2016-02-05	312045	404	0.129	234200	265	0.113	77845	139	0.179
2016-02-06	330846	377	0.114	270729	267	0.099	60117	110	0.183
2016-02-07	337341	327	0.097	287555	214	0.074	49786	113	0.227
2016-02-08	334768	365	0.109	278546	237	0.085	56222	128	0.228
2016-02-09	328425	315	0.096	260760	210	0.081	67665	105	0.155
2016-02-10	328089	372	0.113	257006	228	0.089	71083	144	0.203
2016-02-11	355004	344	0.097	281945	220	0.078	73059	124	0.17
2016-02-12	386880	392	0.101	316405	267	0.084	70475	125	0.177
2016-02-13	358217	342	0.095	291055	240	0.082	67162	102	0.152
2016-02-14	691041	25557	3.698	609112	20396	3.348	81929	5161	6.299
2016-02-15	747981	25678	3.433	574979	17189	2.99	173002	8489	4.907
2016-02-16	800171	31011	3.876	585803	19593	3.345	214368	11418	5.326
2016-02-17	815881	33157	4.064	574861	19563	3.403	241020	13594	5.64
2016-02-18	806257	35677	4.425	551943	20447	3.705	254314	15230	5.989
2016-02-19	817275	37451	4.582	556034	21025	3.781	261241	16426	6.288
2016-02-20	787117	39784	5.054	532886	22463	4.215	254231	17321	6.813
2016-02-21	732423	39207	5.353	486737	21418	4.4	245686	17789	7.241

Continued on next page

Continuation of Table 8.11

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-02-22	770746	38176	4.953	533728	22217	4.163	237018	15959	6.733
2016-02-23	764584	40351	5.278	529906	23721	4.476	234678	16630	7.086
2016-02-24	760152	515	0.068	514411	363	0.071	245741	152	0.062
2016-02-25	747083	515	0.069	496204	374	0.075	250879	141	0.056
2016-02-26	764275	563	0.074	512378	419	0.082	251897	144	0.057
2016-02-27	741229	496	0.067	494773	370	0.075	246456	126	0.051
2016-02-28	696194	434	0.062	459151	293	0.064	237043	141	0.059
2016-02-29	691263	517	0.075	449202	353	0.079	242061	164	0.068
2016-03-01	696388	504	0.072	454878	362	0.08	241510	142	0.059
2016-03-02	700917	486	0.069	461109	336	0.073	239808	150	0.063
2016-03-03	691915	544	0.079	451042	386	0.086	240873	158	0.066
2016-04-18	461253	370	0.08	340662	271	0.08	120591	99	0.082
2016-04-19	463153	383	0.083	343033	275	0.08	120120	108	0.09
2016-04-20	504825	395	0.078	402684	296	0.074	102141	99	0.097
2016-04-21	494942	384	0.078	392501	290	0.074	102441	94	0.092
2016-04-22	530312	455	0.086	429837	345	0.08	100475	110	0.109
2016-04-23	500494	375	0.075	398927	294	0.074	101567	81	0.08
2016-04-24	1017351	1787	0.176	937338	1661	0.177	80013	126	0.157
2016-04-25	1084330	11416	1.053	936931	9331	0.996	147399	2085	1.415
2016-04-26	1106615	7617	0.688	921532	5910	0.641	185083	1707	0.922
2016-04-27	1086552	6707	0.617	878715	4958	0.564	207837	1749	0.842
2016-04-28	1066626	6444	0.604	841933	4668	0.554	224693	1776	0.79
2016-04-29	1079856	4694	0.435	854740	3373	0.395	225116	1321	0.587
2016-04-30	1005141	5881	0.585	778704	4064	0.522	226437	1817	0.802
2016-05-01	922822	5434	0.589	703029	3758	0.535	219793	1676	0.763
2016-05-02	923637	3573	0.387	690843	2545	0.368	232794	1028	0.442
2016-05-03	912692	2592	0.284	668396	1783	0.267	244296	809	0.331
2016-05-04	919118	2396	0.261	683132	1687	0.247	235986	709	0.3
2016-05-05	889725	2388	0.268	659028	1725	0.262	230697	663	0.287
2016-05-06	887489	1972	0.222	658611	1431	0.217	228878	541	0.236
2016-05-07	851475	1671	0.196	628638	1239	0.197	222837	432	0.194
2016-05-08	790803	791	0.1	581512	576	0.099	209291	215	0.103
2016-05-09	807714	381	0.047	590649	293	0.05	217065	88	0.041

Table 8.12: Two unequal duplicates playbacks

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-01-21	5183365	762	0.015	2887796	2	0.0	2295569	760	0.033
2016-01-22	5699857	740	0.013	3197499	3	0.0	2502358	737	0.029
2016-01-23	5706073	734	0.013	3320763	7	0.0	2385310	727	0.03
2016-01-24	4990937	588	0.012	2877218	2	0.0	2113719	586	0.028
2016-01-25	5034490	682	0.014	2775176	2	0.0	2259314	680	0.03
2016-01-26	5211451	694	0.013	2861114	3	0.0	2350337	691	0.029
2016-01-27	5699874	728	0.013	3115105	2	0.0	2584769	726	0.028
2016-01-28	10716712	866	0.008	5406227	4	0.0	5310485	862	0.016
2016-01-29	10860619	908	0.008	5522217	5	0.0	5338402	903	0.017
2016-01-30	10008250	824	0.008	5200705	6	0.0	4807545	818	0.017
2016-01-31	8445037	688	0.008	4380686	6	0.0	4064351	682	0.017
2016-02-01	8595181	716	0.008	4260810	1	0.0	4334371	715	0.016
2016-02-02	8469929	786	0.009	4153867	4	0.0	4316062	782	0.018
2016-02-03	8386787	708	0.008	4130617	3	0.0	4256170	705	0.017
2016-02-04	8410214	794	0.009	4128894	5	0.0	4281320	789	0.018
2016-02-05	9001255	830	0.009	4503223	2	0.0	4498032	828	0.018
2016-02-06	8841003	776	0.009	4623472	5	0.0	4217531	771	0.018
2016-02-07	8375341	666	0.008	4562874	3	0.0	3812467	663	0.017
2016-02-08	8145940	762	0.009	4244968	10	0.0	3900972	752	0.019
2016-02-09	8439985	642	0.008	4289622	6	0.0	4150363	636	0.015
2016-02-10	8437553	758	0.009	4261212	4	0.0	4176341	754	0.018
2016-02-11	8748323	726	0.008	4431643	2	0.0	4316680	724	0.017
2016-02-12	9591865	804	0.008	4932532	8	0.0	4659333	796	0.017
2016-02-13	9207400	712	0.008	4815322	5	0.0	4392078	707	0.016
2016-02-14	52264508	88234	0.169	25683182	53560	0.209	26581326	34674	0.13
2016-02-15	62315131	102684	0.165	28423017	48191	0.17	33892114	54493	0.161
2016-02-16	62575347	123824	0.198	27216556	54760	0.201	35358791	69064	0.195
2016-02-17	61635277	132694	0.215	26096222	52583	0.201	35539055	80111	0.225

Continued on next page

Continuation of Table 8.12

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-02-18	58176702	143182	0.246	24279133	54412	0.224	33897569	88770	0.262
2016-02-19	57453384	150192	0.261	23863079	55431	0.232	33590305	94761	0.282
2016-02-20	51238240	160064	0.312	21695419	61596	0.284	29542821	98468	0.333
2016-02-21	44012913	157886	0.359	19017377	60271	0.317	24995536	97615	0.391
2016-02-22	48123726	153564	0.319	20139829	61372	0.305	27983897	92192	0.329
2016-02-23	47818784	162250	0.339	19908725	66228	0.333	27910059	96022	0.344
2016-02-24	19205376	1078	0.006	8020004	5	0.0	11185372	1073	0.01
2016-02-25	19155488	1064	0.006	7933048	4	0.0	11222440	1060	0.009
2016-02-26	19895000	1172	0.006	8316062	8	0.0	11578938	1164	0.01
2016-02-27	18410754	1060	0.006	7817765	10	0.0	10592989	1050	0.01
2016-02-28	16341019	898	0.005	7006815	6	0.0	9334204	892	0.01
2016-02-29	17041782	1080	0.006	7049968	9	0.0	9991814	1071	0.011
2016-03-01	17519578	1046	0.006	7264171	6	0.0	10255407	1040	0.01
2016-03-02	17260480	1024	0.006	7182545	9	0.0	10077935	1015	0.01
2016-03-03	17326476	1140	0.007	7175051	9	0.0	10151425	1131	0.011
2016-04-18	13094984	762	0.006	5480281	6	0.0	7614703	756	0.01
2016-04-19	13228953	804	0.006	5561693	5	0.0	7667260	799	0.01
2016-04-20	13716969	828	0.006	5943181	8	0.0	7773788	820	0.011
2016-04-21	13592572	816	0.006	6080560	5	0.0	7512012	811	0.011
2016-04-22	15716545	948	0.006	7204431	10	0.0	8512114	938	0.011
2016-04-23	14539943	780	0.005	6686559	8	0.0	7853384	772	0.01
2016-04-24	32844244	3950	0.012	16533941	22	0.0	16310303	3928	0.024
2016-04-25	61648530	39340	0.064	27532332	22	0.0	34116198	39318	0.115
2016-04-26	51762664	26660	0.052	23271706	32	0.0	28490958	26628	0.093
2016-04-27	51443095	16640	0.032	22441805	15	0.0	29001290	16625	0.057
2016-04-28	49571047	16518	0.033	21060520	23	0.0	28510527	16495	0.058
2016-04-29	45740028	11464	0.025	19304866	36	0.0	26435162	11428	0.043
2016-04-30	44747668	22174	0.05	18482843	45	0.0	26264825	22129	0.084
2016-05-01	38826732	13302	0.034	16407601	32	0.0	22419131	13270	0.059
2016-05-02	36971583	8526	0.023	15561500	15	0.0	21410083	8511	0.04
2016-05-03	34043783	6404	0.019	14283511	18	0.0	19760272	6386	0.032
2016-05-04	32862327	5786	0.018	14013298	22	0.0	18849029	5764	0.031

Continued on next page

Continuation of Table 8.12

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-05-05	31535242	5688	0.018	13427299	15	0.0	18107943	5673	0.031
2016-05-06	30227071	4812	0.016	12798897	11	0.0	17428174	4801	0.028
2016-05-07	26738540	3914	0.015	11477044	22	0.0	15261496	3892	0.026
2016-05-08	20337761	1782	0.009	8893020	15	0.0	11444741	1767	0.015
2016-05-09	19254690	792	0.004	8408779	12	0.0	10845911	780	0.007

Table 8.13: Two unequal duplicates affected users playbacks

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-01-21	2	468	0	292	0	0.0	0	0
2016-01-22	3	491	0	246	0	0.0	0	0
2016-01-23	7	487	0	240	0	0.0	0	0
2016-01-24	2	388	0	198	0	0.0	0	0
2016-01-25	2	408	0	272	0	0.0	0	0
2016-01-26	3	429	0	262	0	0.0	0	0
2016-01-27	2	448	0	278	0	0.0	0	0
2016-01-28	4	574	0	288	0	0.0	0	0
2016-01-29	5	617	0	286	0	0.0	0	0
2016-01-30	6	560	0	258	0	0.0	0	0
2016-01-31	6	454	0	228	0	0.0	0	0
2016-02-01	1	425	0	290	0	0.0	0	0
2016-02-02	4	472	0	310	0	0.0	0	0
2016-02-03	3	401	0	304	0	0.0	0	0
2016-02-04	5	505	0	284	0	0.0	0	0
2016-02-05	2	536	0	292	0	0.0	0	0
2016-02-06	5	541	0	230	0	0.0	0	0
2016-02-07	3	429	0	234	0	0.0	0	0
2016-02-08	10	488	0	264	0	0.0	0	0
2016-02-09	6	422	0	214	0	0.0	0	0
2016-02-10	4	456	0	298	0	0.0	0	0
2016-02-11	2	470	0	254	0	0.0	0	0
2016-02-12	8	540	0	256	0	0.0	0	0
2016-02-13	5	497	0	210	0	0.0	0	0
2016-02-14	53560	16644	0	18030	31147	35.3	23580	7567
2016-02-15	48191	20135	0	34358	45415	44.228	18873	26542
2016-02-16	54760	22974	0	46090	59052	47.69	21140	37912
2016-02-17	52583	25161	0	54950	65453	49.326	19624	45829
2016-02-18	54412	27054	0	61716	72195	50.422	20400	51795
2016-02-19	55431	28275	0	66486	74058	49.309	19091	54967
2016-02-20	61596	28232	0	70236	78269	48.899	19757	58512

Continued on next page

Continuation of Table 8.13

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-02-21	60271	25417	0	72198	77779	49.263	19471	58308
2016-02-22	61372	27364	0	64828	70706	46.043	17608	53098
2016-02-23	66228	28494	0	67528	70589	43.506	16530	54059
2016-02-24	5	761	0	312	334	30.983	2	332
2016-02-25	4	776	0	284	292	27.444	0	292
2016-02-26	8	862	0	302	333	28.413	2	331
2016-02-27	10	780	0	270	203	19.151	0	203
2016-02-28	6	606	0	286	201	22.383	1	200
2016-02-29	9	729	0	342	214	19.815	1	213
2016-03-01	6	748	0	292	228	21.797	0	228
2016-03-02	9	709	0	306	241	23.535	0	241
2016-03-03	9	807	0	324	277	24.298	0	277
2016-04-18	6	554	0	202	0	0.0	0	0
2016-04-19	5	573	0	226	0	0.0	0	0
2016-04-20	8	614	0	206	0	0.0	0	0
2016-04-21	5	615	0	196	0	0.0	0	0
2016-04-22	10	714	0	224	0	0.0	0	0
2016-04-23	8	604	0	168	0	0.0	0	0
2016-04-24	22	3660	0	268	2965	75.063	12	2953
2016-04-25	22	32496	0	6822	37288	94.784	13	37275
2016-04-26	32	20740	0	5888	22307	83.672	17	22290
2016-04-27	15	12375	0	4250	12523	75.258	2	12521
2016-04-28	23	12029	0	4466	11550	69.924	7	11543
2016-04-29	36	8224	0	3204	7749	67.594	15	7734
2016-04-30	45	15521	0	6608	19255	86.836	19	19236
2016-05-01	32	9182	0	4088	8456	63.569	3	8453
2016-05-02	15	6119	0	2392	4990	58.527	2	4988
2016-05-03	18	4348	0	2038	2710	42.317	0	2710
2016-05-04	22	4058	0	1706	2295	39.665	4	2291
2016-05-05	15	4047	0	1626	2030	35.689	1	2029
2016-05-06	11	3453	0	1348	1205	25.042	0	1205
2016-05-07	22	2840	0	1052	548	14.001	0	548

Continued on next page

Continuation of Table 8.13

Date	Online online backs	users play-	Online offline backs	users play-	Offline online backs	users play-	Offline offline backs	users play-	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-05-08	15		1273		0		494		286	16.049	0	286
2016-05-09	12		598		0		182		106	13.384	0	106

Three or more unequal duplicates should also never really occur in a log system. However, Table 8.14 illustrate that this was not an infrequent occurrence during the suspicious days. About eighty per cent of the users was affected by this impossibility in the days between 2016-02-14 and 2016-02-23. While only affecting a small percentage of system users during the second period, many of those playbacks were from Beyoncé's album.

Table 8.15 show that about sixty per cent of all playbacks during the first suspicious period was a three or more *unequal* duplicate, most of them apparently playing tracks from Kanye's album.

Table 8.14: Three or more unequal duplicates system users

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-01-21	174620	2939	1.683	135221	1745	1.29	39399	1194	3.031
2016-01-22	188796	3291	1.743	149491	1997	1.336	39305	1294	3.292
2016-01-23	183237	3166	1.728	148330	1981	1.336	34907	1185	3.395
2016-01-24	171511	2884	1.682	137257	1730	1.26	34254	1154	3.369
2016-01-25	174896	2851	1.63	135543	1584	1.169	39353	1267	3.22
2016-01-26	180917	2951	1.631	140482	1676	1.193	40435	1275	3.153
2016-01-27	202757	2939	1.45	165342	1745	1.055	37415	1194	3.191
2016-01-28	356158	3293	0.925	307284	2072	0.674	48874	1221	2.498
2016-01-29	344910	3462	1.004	277724	2167	0.78	67186	1295	1.927
2016-01-30	324098	3437	1.06	254269	2272	0.894	69829	1165	1.668
2016-01-31	296886	3027	1.02	225798	1895	0.839	71088	1132	1.592
2016-02-01	300715	2960	0.984	222513	1760	0.791	78202	1200	1.534
2016-02-02	300530	2967	0.987	221130	1751	0.792	79400	1216	1.531
2016-02-03	296753	3015	1.016	216748	1768	0.816	80005	1247	1.559
2016-02-04	298583	3063	1.026	217279	1787	0.822	81304	1276	1.569
2016-02-05	312045	3226	1.034	234200	2010	0.858	77845	1216	1.562
2016-02-06	330846	3071	0.928	270729	2024	0.748	60117	1047	1.742
2016-02-07	337341	2874	0.852	287555	1839	0.64	49786	1035	2.079
2016-02-08	334768	2731	0.816	278546	1599	0.574	56222	1132	2.013
2016-02-09	328425	2820	0.859	260760	1710	0.656	67665	1110	1.64
2016-02-10	328089	2787	0.849	257006	1670	0.65	71083	1117	1.571
2016-02-11	355004	2858	0.805	281945	1738	0.616	73059	1120	1.533
2016-02-12	386880	3054	0.789	316405	1916	0.606	70475	1138	1.615
2016-02-13	358217	2991	0.835	291055	1923	0.661	67162	1068	1.59
2016-02-14	691041	514378	74.435	609112	455546	74.789	81929	58832	71.809
2016-02-15	747981	597572	79.891	574979	460063	80.014	173002	137509	79.484
2016-02-16	800171	642455	80.29	585803	473033	80.75	214368	169422	79.033
2016-02-17	815881	653053	80.043	574861	463186	80.574	241020	189867	78.776
2016-02-18	806257	639553	79.324	551943	443794	80.406	254314	195759	76.975
2016-02-19	817275	642100	78.566	556034	443885	79.831	261241	198215	75.874
2016-02-20	787117	607544	77.186	532886	421392	79.077	254231	186152	73.222
2016-02-21	732423	549109	74.972	486737	376486	77.349	245686	172623	70.262

Continued on next page

Continuation of Table 8.14

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-02-22	770746	573237	74.374	533728	402110	75.34	237018	171127	72.2
2016-02-23	764584	577685	75.555	529906	408250	77.042	234678	169435	72.199
2016-02-24	760152	3040	0.4	514411	1772	0.344	245741	1268	0.516
2016-02-25	747083	3067	0.411	496204	1809	0.365	250879	1258	0.501
2016-02-26	764275	3200	0.419	512378	1981	0.387	251897	1219	0.484
2016-02-27	741229	3048	0.411	494773	1948	0.394	246456	1100	0.446
2016-02-28	696194	2810	0.404	459151	1746	0.38	237043	1064	0.449
2016-02-29	691263	2796	0.404	449202	1618	0.36	242061	1178	0.487
2016-03-01	696388	2882	0.414	454878	1744	0.383	241510	1138	0.471
2016-03-02	700917	2793	0.398	461109	1657	0.359	239808	1136	0.474
2016-03-03	691915	2869	0.415	451042	1640	0.364	240873	1229	0.51
2016-04-18	461253	2048	0.444	340662	1184	0.348	120591	864	0.716
2016-04-19	463153	2091	0.451	343033	1195	0.348	120120	896	0.746
2016-04-20	504825	2087	0.413	402684	1259	0.313	102141	828	0.811
2016-04-21	494942	2119	0.428	392501	1319	0.336	102441	800	0.781
2016-04-22	530312	2322	0.438	429837	1549	0.36	100475	773	0.769
2016-04-23	500494	2207	0.441	398927	1461	0.366	101567	746	0.734
2016-04-24	1017351	2300	0.226	937338	1672	0.178	80013	628	0.785
2016-04-25	1084330	2602	0.24	936931	1792	0.191	147399	810	0.55
2016-04-26	1106615	2858	0.258	921532	1977	0.215	185083	881	0.476
2016-04-27	1086552	2547	0.234	878715	1627	0.185	207837	920	0.443
2016-04-28	1066626	2637	0.247	841933	1697	0.202	224693	940	0.418
2016-04-29	1079856	2661	0.246	854740	1749	0.205	225116	912	0.405
2016-04-30	1005141	2906	0.289	778704	1935	0.248	226437	971	0.429
2016-05-01	922822	2269	0.246	703029	1454	0.207	219793	815	0.371
2016-05-02	923637	2261	0.245	690843	1387	0.201	232794	874	0.375
2016-05-03	912692	2242	0.246	668396	1338	0.2	244296	904	0.37
2016-05-04	919118	2325	0.253	683132	1416	0.207	235986	909	0.385
2016-05-05	889725	2199	0.247	659028	1405	0.213	230697	794	0.344
2016-05-06	887489	2276	0.256	658611	1453	0.221	228878	823	0.36
2016-05-07	851475	2194	0.258	628638	1405	0.223	222837	789	0.354
2016-05-08	790803	2014	0.255	581512	1248	0.215	209291	766	0.366
2016-05-09	807714	2044	0.253	590649	1228	0.208	217065	816	0.376

Table 8.15: Three or more unequal duplicates playbacks

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-01-21	5183365	70225	1.355	2887796	0	0.0	2295569	70225	3.059
2016-01-22	5699857	77283	1.356	3197499	0	0.0	2502358	77283	3.088
2016-01-23	5706073	80176	1.405	3320763	0	0.0	2385310	80176	3.361
2016-01-24	4990937	66852	1.339	2877218	0	0.0	2113719	66852	3.163
2016-01-25	5034490	68512	1.361	2775176	0	0.0	2259314	68512	3.032
2016-01-26	5211451	68171	1.308	2861114	0	0.0	2350337	68171	2.9
2016-01-27	5699874	71507	1.255	3115105	0	0.0	2584769	71507	2.766
2016-01-28	10716712	80707	0.753	5406227	0	0.0	5310485	80707	1.52
2016-01-29	10860619	85317	0.786	5522217	0	0.0	5338402	85317	1.598
2016-01-30	10008250	90035	0.9	5200705	0	0.0	4807545	90035	1.873
2016-01-31	8445037	73677	0.872	4380686	0	0.0	4064351	73677	1.813
2016-02-01	8595181	69900	0.813	4260810	0	0.0	4334371	69900	1.613
2016-02-02	8469929	67746	0.8	4153867	0	0.0	4316062	67746	1.57
2016-02-03	8386787	68447	0.816	4130617	0	0.0	4256170	68447	1.608
2016-02-04	8410214	72667	0.864	4128894	0	0.0	4281320	72667	1.697
2016-02-05	9001255	80069	0.89	4503223	0	0.0	4498032	80069	1.78
2016-02-06	8841003	80824	0.914	4623472	0	0.0	4217531	80824	1.916
2016-02-07	8375341	69299	0.827	4562874	0	0.0	3812467	69299	1.818
2016-02-08	8145940	62315	0.765	4244968	0	0.0	3900972	62315	1.597
2016-02-09	8439985	67575	0.801	4289622	0	0.0	4150363	67575	1.628
2016-02-10	8437553	65934	0.781	4261212	0	0.0	4176341	65934	1.579
2016-02-11	8748323	69558	0.795	4431643	0	0.0	4316680	69558	1.611
2016-02-12	9591865	74365	0.775	4932532	0	0.0	4659333	74365	1.596
2016-02-13	9207400	78950	0.857	4815322	0	0.0	4392078	78950	1.798
2016-02-14	52264508	28579867	54.683	25683182	13997322	54.5	26581326	14582545	54.86
2016-02-15	62315131	35837686	57.51	28423017	16318023	57.411	33892114	19519663	57.594
2016-02-16	62575347	36141565	57.757	27216556	15689865	57.648	35358791	20451700	57.84

Continued on next page

Continuation of Table 8.15

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-02-17	61635277	35772188	58.038	26096222	15099751	57.862	35539055	20672437	58.168
2016-02-18	58176702	33716728	57.956	24279133	14032095	57.795	33897569	19684633	58.071
2016-02-19	57453384	33290893	57.944	23863079	13778875	57.741	33590305	19512018	58.088
2016-02-20	51238240	29722290	58.008	21695419	12543292	57.815	29542821	17178998	58.149
2016-02-21	44012913	25462084	57.851	19017377	10983059	57.753	24995536	14479025	57.926
2016-02-22	48123726	27815608	57.8	20139829	11565797	57.427	27983897	16249811	58.068
2016-02-23	47818784	27834973	58.209	19908725	11529780	57.913	27910059	16305193	58.42
2016-02-24	19205376	74641	0.389	8020004	0	0.0	11185372	74641	0.667
2016-02-25	19155488	82384	0.43	7933048	0	0.0	11222440	82384	0.734
2016-02-26	19895000	80020	0.402	8316062	0	0.0	11578938	80020	0.691
2016-02-27	18410754	81353	0.442	7817765	0	0.0	10592989	81353	0.768
2016-02-28	16341019	73829	0.452	7006815	0	0.0	9334204	73829	0.791
2016-02-29	17041782	69102	0.405	7049968	0	0.0	9991814	69102	0.692
2016-03-01	17519578	71552	0.408	7264171	0	0.0	10255407	71552	0.698
2016-03-02	17260480	68163	0.395	7182545	1	0.0	10077935	68162	0.676
2016-03-03	17326476	69478	0.401	7175051	0	0.0	10151425	69478	0.684
2016-04-18	13094984	55019	0.42	5480281	0	0.0	7614703	55019	0.723
2016-04-19	13228953	52600	0.398	5561693	0	0.0	7667260	52600	0.686
2016-04-20	13716969	54451	0.397	5943181	0	0.0	7773788	54451	0.7
2016-04-21	13592572	54086	0.398	6080560	0	0.0	7512012	54086	0.72
2016-04-22	15716545	60662	0.386	7204431	0	0.0	8512114	60662	0.713
2016-04-23	14539943	57355	0.394	6686559	0	0.0	7853384	57355	0.73
2016-04-24	32844244	63288	0.193	16533941	0	0.0	16310303	63288	0.388
2016-04-25	61648530	89955	0.146	27532332	0	0.0	34116198	89955	0.264
2016-04-26	51762664	86455	0.167	23271706	0	0.0	28490958	86455	0.303
2016-04-27	51443095	85111	0.165	22441805	0	0.0	29001290	85111	0.293
2016-04-28	49571047	88136	0.178	21060520	0	0.0	28510527	88136	0.309
2016-04-29	45740028	89346	0.195	19304866	0	0.0	26435162	89346	0.338
2016-04-30	44747668	86771	0.194	18482843	0	0.0	26264825	86771	0.33
2016-05-01	38826732	73514	0.189	16407601	0	0.0	22419131	73514	0.328
2016-05-02	36971583	79715	0.216	15561500	0	0.0	21410083	79715	0.372

Continued on next page

Continuation of Table 8.15

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-05-03	34043783	74078	0.218	14283511	0	0.0	19760272	74078	0.375
2016-05-04	32862327	76266	0.232	14013298	0	0.0	18849029	76266	0.405
2016-05-05	31535242	76458	0.242	13427299	0	0.0	18107943	76458	0.422
2016-05-06	30227071	81452	0.269	12798897	0	0.0	17428174	81452	0.467
2016-05-07	26738540	72107	0.27	11477044	0	0.0	15261496	72107	0.472
2016-05-08	20337761	54173	0.266	8893020	1	0.0	11444741	54172	0.473
2016-05-09	19254690	53743	0.279	8408779	0	0.0	10845911	53743	0.496

Table 8.16: Three or more unequal duplicates affected users playbacks

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-01-21	0	40792	0	29433	0	0.0	0	0
2016-01-22	0	45686	0	31597	0	0.0	0	0
2016-01-23	0	49146	0	31030	0	0.0	0	0
2016-01-24	0	40752	0	26100	0	0.0	0	0
2016-01-25	0	40051	0	28461	0	0.0	0	0
2016-01-26	0	39157	0	29014	0	0.0	0	0
2016-01-27	0	43231	0	28276	0	0.0	0	0
2016-01-28	0	51278	0	29429	0	0.0	0	0
2016-01-29	0	55710	0	29607	0	0.0	0	0
2016-01-30	0	58679	0	31356	0	0.0	0	0
2016-01-31	0	47185	0	26492	0	0.0	0	0
2016-02-01	0	42275	0	27625	0	0.0	0	0
2016-02-02	0	40484	0	27262	0	0.0	0	0
2016-02-03	0	39070	0	29377	0	0.0	0	0
2016-02-04	0	42004	0	30663	0	0.0	0	0
2016-02-05	0	51289	0	28780	0	0.0	0	0
2016-02-06	0	52715	0	28109	0	0.0	0	0
2016-02-07	0	43237	0	26062	0	0.0	0	0
2016-02-08	0	36257	0	26058	0	0.0	0	0
2016-02-09	0	41904	0	25671	0	0.0	0	0
2016-02-10	0	41085	0	24849	0	0.0	0	0
2016-02-11	0	41312	0	28246	0	0.0	0	0
2016-02-12	0	44850	0	29515	0	0.0	0	0
2016-02-13	0	50821	0	28129	0	0.0	0	0
2016-02-14	13997322	12149039	0	2433506	18098334	63.325	8489026	9609308
2016-02-15	16318023	12805452	0	6714211	22864291	63.8	9571626	13292665
2016-02-16	15689865	12800603	0	7651097	21369275	59.127	8121398	13247877
2016-02-17	15099751	12442672	0	8229765	19528939	54.593	6951392	12577547
2016-02-18	14032095	11665625	0	8019008	16902330	50.13	5754559	11147771
2016-02-19	13778875	11600967	0	7911051	15358269	46.134	5045641	10312628
2016-02-20	12543292	10340512	0	6838486	12213349	41.092	3823100	8390249

Continued on next page

Continuation of Table 8.16

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-02-21	10983059	8543361	0	5935664	9720085	38.175	3087757	6632328
2016-02-22	11565797	10020916	0	6228895	10917131	39.248	3469908	7447223
2016-02-23	11529780	10191228	0	6113965	10020697	36.0	3072840	6947857
2016-02-24	0	43793	0	30848	10314	13.818	0	10314
2016-02-25	0	50016	0	32368	10464	12.701	0	10464
2016-02-26	0	50447	0	29573	9925	12.403	0	9925
2016-02-27	0	51983	0	29370	8890	10.928	0	8890
2016-02-28	0	45053	0	28776	7347	9.951	0	7347
2016-02-29	0	40090	0	29012	8432	12.202	0	8432
2016-03-01	0	41522	0	30030	7184	10.04	0	7184
2016-03-02	1	39429	0	28733	6859	10.063	1	6858
2016-03-03	0	38644	0	30834	6586	9.479	0	6586
2016-04-18	0	31672	0	23347	0	0.0	0	0
2016-04-19	0	29671	0	22929	0	0.0	0	0
2016-04-20	0	32864	0	21587	0	0.0	0	0
2016-04-21	0	34060	0	20026	0	0.0	0	0
2016-04-22	0	42045	0	18617	0	0.0	0	0
2016-04-23	0	37293	0	20062	0	0.0	0	0
2016-04-24	0	45059	0	18229	15971	25.235	0	15971
2016-04-25	0	64701	0	25254	40563	45.093	0	40563
2016-04-26	0	60337	0	26118	27330	31.612	0	27330
2016-04-27	0	55496	0	29615	28178	33.107	0	28178
2016-04-28	0	57218	0	30918	27534	31.24	0	27534
2016-04-29	0	58370	0	30976	21940	24.556	0	21940
2016-04-30	0	56992	0	29779	26341	30.357	0	26341
2016-05-01	0	47346	0	26168	20727	28.195	0	20727
2016-05-02	0	47712	0	32003	18467	23.166	0	18467
2016-05-03	0	43154	0	30924	13218	17.843	0	13218
2016-05-04	0	45868	0	30398	12745	16.711	0	12745
2016-05-05	0	48948	0	27510	10815	14.145	0	10815
2016-05-06	0	51520	0	29932	9950	12.216	0	9950
2016-05-07	0	46388	0	25719	7451	10.333	0	7451

Continued on next page

Continuation of Table 8.16

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-05-08	1	33446	0	20726	3843	7.094	0	3843
2016-05-09	0	31643	0	22100	4802	8.935	0	4802

8.6.9 Modulo six findings

The previous tables show that users in the first period are most affected by impossibilities which are described in analysis method 2 (Subsection 8.5.2) than users in the second period. DN have on their own inspected a tiny subset of users in the second period and identified a high count of playbacks which ends on the same second and milliseconds. They suspect that legitimate playbacks have been duplicated by changing the hours and minutes of original playbacks. The amount of change in the timestamp appears to always be evenly divided by six minutes.

It is statistically possible for users to play one track ID on two different timestamps which end on the same seconds and milliseconds, and those two timestamps being divisible by six minutes. Table 8.17 show that around one per cent was affected by this scenario in normal days. A closer inspection of the affected playbacks for normal days confirms our suspicion that they are caused by devices with inadequate ability to record milliseconds as all the records have '.000' as milliseconds. However, during the second suspicious period, this scenario started to also affect records with all types of millisecond numbers. It is an extremely rare occurrence when a legitimate user causes this to happen, and it is unlikely that these numbers seen in the table is caused by system users alone.

An important note is that the increase in occurrences in the first period is caused by duplicates, which happen to be divisible by six minutes. For example, on 2016-02-14 there exists a lot of duplicates playbacks at 01:00:00.000, 04:00:00.000, 01:30:00.000 and 04:30:00.000. Each pair of these timestamps happen to be divisible by six minutes; however, we consider them as being duplicates. In fact, tables in Appendix A.A show that these occurrences were caused by these duplicates.

Table 8.18 show that it is statistically possible for users to have online playbacks for this kind of scenario. However, the frequency for its occurrence sky-rocketed in the second suspicious period. Going from normally less than a hundred occurrences for normal days, up to fifteen million occurrences when 'Lemonade' was released. Finally, Table 8.19 show that these affected playbacks were mostly from this album release.

Keep in mind that we have only accounted for the track IDs and not any videos associated with the album. This means that the affected number of playbacks could be higher. In addition, our result in Table 8.19 accounts for the assumption that first playbacks are legitimate, which means that they are not counted in these tables.

Table 8.17: Six minutes system users

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-01-21	174620	1105	0.633	135221	692	0.512	39399	413	1.048
2016-01-22	188796	1214	0.643	149491	778	0.52	39305	436	1.109
2016-01-23	183237	1016	0.554	148330	660	0.445	34907	356	1.02
2016-01-24	171511	910	0.531	137257	572	0.417	34254	338	0.987
2016-01-25	174896	1125	0.643	135543	688	0.508	39353	437	1.11
2016-01-26	180917	1084	0.599	140482	673	0.479	40435	411	1.016
2016-01-27	202757	1280	0.631	165342	893	0.54	37415	387	1.034
2016-01-28	356158	8511	2.39	307284	7489	2.437	48874	1022	2.091
2016-01-29	344910	6690	1.94	277724	4715	1.698	67186	1975	2.94
2016-01-30	324098	5371	1.657	254269	3125	1.229	69829	2246	3.216
2016-01-31	296886	4152	1.399	225798	2311	1.023	71088	1841	2.59
2016-02-01	300715	4826	1.605	222513	2605	1.171	78202	2221	2.84
2016-02-02	300530	4355	1.449	221130	2367	1.07	79400	1988	2.504
2016-02-03	296753	4029	1.358	216748	2188	1.009	80005	1841	2.301
2016-02-04	298583	3835	1.284	217279	2015	0.927	81304	1820	2.239
2016-02-05	312045	3647	1.169	234200	2065	0.882	77845	1582	2.032
2016-02-06	330846	2989	0.903	270729	2054	0.759	60117	935	1.555
2016-02-07	337341	2955	0.876	287555	2286	0.795	49786	669	1.344
2016-02-08	334768	3448	1.03	278546	2595	0.932	56222	853	1.517
2016-02-09	328425	3562	1.085	260760	2379	0.912	67665	1183	1.748
2016-02-10	328089	3362	1.025	257006	2206	0.858	71083	1156	1.626
2016-02-11	355004	3272	0.922	281945	2050	0.727	73059	1222	1.673
2016-02-12	386880	3301	0.853	316405	2179	0.689	70475	1122	1.592
2016-02-13	358217	2885	0.805	291055	1938	0.666	67162	947	1.41
2016-02-14	691041	546777	79.124	609112	485683	79.736	81929	61094	74.569
2016-02-15	747981	660347	88.284	574979	505980	88.0	173002	154367	89.228
2016-02-16	800171	715331	89.397	585803	523973	89.445	214368	191358	89.266
2016-02-17	815881	732172	89.74	574861	516703	89.883	241020	215469	89.399
2016-02-18	806257	719494	89.239	551943	495053	89.693	254314	224441	88.253
2016-02-19	817275	725205	88.735	556034	496245	89.247	261241	228960	87.643
2016-02-20	787117	695868	88.407	532886	475980	89.321	254231	219888	86.491
2016-02-21	732423	638369	87.159	486737	430320	88.409	245686	208049	84.681

Continued on next page

Continuation of Table 8.17

Date	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
2016-02-22	770746	669531	86.868	533728	467605	87.611	237018	201926	85.194
2016-02-23	764584	677408	88.598	529906	476566	89.934	234678	200842	85.582
2016-02-24	760152	12801	1.684	514411	7295	1.418	245741	5506	2.241
2016-02-25	747083	12531	1.677	496204	7087	1.428	250879	5444	2.17
2016-02-26	764275	12554	1.643	512378	7241	1.413	251897	5313	2.109
2016-02-27	741229	10076	1.359	494773	5670	1.146	246456	4406	1.788
2016-02-28	696194	8387	1.205	459151	4713	1.026	237043	3674	1.55
2016-02-29	691263	10413	1.506	449202	5924	1.319	242061	4489	1.854
2016-03-01	696388	10738	1.542	454878	6087	1.338	241510	4651	1.926
2016-03-02	700917	10015	1.429	461109	5608	1.216	239808	4407	1.838
2016-03-03	691915	9716	1.404	451042	5507	1.221	240873	4209	1.747
2016-04-18	461253	6609	1.433	340662	4514	1.325	120591	2095	1.737
2016-04-19	463153	6564	1.417	343033	4463	1.301	120120	2101	1.749
2016-04-20	504825	6214	1.231	402684	4616	1.146	102141	1598	1.565
2016-04-21	494942	5760	1.164	392501	4159	1.06	102441	1601	1.563
2016-04-22	530312	6503	1.226	429837	4917	1.144	100475	1586	1.579
2016-04-23	500494	5382	1.075	398927	3851	0.965	101567	1531	1.507
2016-04-24	1017351	360335	35.419	937338	353206	37.682	80013	7129	8.91
2016-04-25	1084330	474317	43.743	936931	418428	44.659	147399	55889	37.917
2016-04-26	1106615	678624	61.324	921532	576879	62.6	185083	101745	54.973
2016-04-27	1086552	672915	61.931	878715	553454	62.984	207837	119461	57.478
2016-04-28	1066626	694998	65.159	841933	555214	65.945	224693	139784	62.211
2016-04-29	1079856	701579	64.97	854740	556528	65.111	225116	145051	64.434
2016-04-30	1005141	603375	60.029	778704	470196	60.382	226437	133179	58.815
2016-05-01	922822	636170	68.937	703029	481636	68.509	219793	154534	70.309
2016-05-02	923637	624955	67.662	690843	472798	68.438	232794	152157	65.361
2016-05-03	912692	555316	60.844	668396	416447	62.305	244296	138869	56.845
2016-05-04	919118	587380	63.907	683132	446402	65.346	235986	140978	59.74
2016-05-05	889725	581806	65.392	659028	439989	66.763	230697	141817	61.473
2016-05-06	887489	534149	60.187	658611	407554	61.881	228878	126595	55.311
2016-05-07	851475	473762	55.64	628638	367472	58.455	222837	106290	47.699
2016-05-08	790803	266292	33.674	581512	207387	35.663	209291	58905	28.145
2016-05-09	807714	9267	1.147	590649	6183	1.047	217065	3084	1.421

Table 8.18: Six minutes playbacks

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-01-21	5183365	4734	0.091	2887796	35	0.001	2295569	4699	0.205
2016-01-22	5699857	4739	0.083	3197499	32	0.001	2502358	4707	0.188
2016-01-23	5706073	4517	0.079	3320763	56	0.002	2385310	4461	0.187
2016-01-24	4990937	4364	0.087	2877218	38	0.001	2113719	4326	0.205
2016-01-25	5034490	4598	0.091	2775176	48	0.002	2259314	4550	0.201
2016-01-26	5211451	3825	0.073	2861114	32	0.001	2350337	3793	0.161
2016-01-27	5699874	4822	0.085	3115105	43	0.001	2584769	4779	0.185
2016-01-28	10716712	30827	0.288	5406227	81	0.001	5310485	30746	0.579
2016-01-29	10860619	23421	0.216	5522217	56	0.001	5338402	23365	0.438
2016-01-30	10008250	19929	0.199	5200705	77	0.001	4807545	19852	0.413
2016-01-31	8445037	15194	0.18	4380686	29	0.001	4064351	15165	0.373
2016-02-01	8595181	16261	0.189	4260810	55	0.001	4334371	16206	0.374
2016-02-02	8469929	14746	0.174	4153867	35	0.001	4316062	14711	0.341
2016-02-03	8386787	14464	0.172	4130617	51	0.001	4256170	14413	0.339
2016-02-04	8410214	13956	0.166	4128894	30	0.001	4281320	13926	0.325
2016-02-05	9001255	12552	0.139	4503223	34	0.001	4498032	12518	0.278
2016-02-06	8841003	10903	0.123	4623472	34	0.001	4217531	10869	0.258
2016-02-07	8375341	11647	0.139	4562874	51	0.001	3812467	11596	0.304
2016-02-08	8145940	12581	0.154	4244968	44	0.001	3900972	12537	0.321
2016-02-09	8439985	13191	0.156	4289622	44	0.001	4150363	13147	0.317
2016-02-10	8437553	12209	0.145	4261212	56	0.001	4176341	12153	0.291
2016-02-11	8748323	11819	0.135	4431643	34	0.001	4316680	11785	0.273
2016-02-12	9591865	11552	0.12	4932532	55	0.001	4659333	11497	0.247
2016-02-13	9207400	10506	0.114	4815322	32	0.001	4392078	10474	0.238
2016-02-14	52264508	27636779	52.879	25683182	13463259	52.421	26581326	14173520	53.321
2016-02-15	62315131	36180835	58.061	28423017	16350436	57.525	33892114	19830399	58.51
2016-02-16	62575347	36524250	58.368	27216556	15733229	57.808	35358791	20791021	58.8

Continued on next page

Continuation of Table 8.18

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-02-17	61635277	36208840	58.747	26096222	15176282	58.155	35539055	21032558	59.182
2016-02-18	58176702	34180560	58.753	24279133	14144850	58.259	33897569	20035710	59.107
2016-02-19	57453384	33753034	58.749	23863079	13890396	58.209	33590305	19862638	59.132
2016-02-20	51238240	30170202	58.882	21695419	12661642	58.361	29542821	17508560	59.265
2016-02-21	44012913	25889125	58.822	19017377	11110721	58.424	24995536	14778404	59.124
2016-02-22	48123726	28256470	58.716	20139829	11716634	58.176	27983897	16539836	59.105
2016-02-23	47818784	28293894	59.169	19908725	11683244	58.684	27910059	16610650	59.515
2016-02-24	19205376	42563	0.222	8020004	62	0.001	11185372	42501	0.38
2016-02-25	19155488	40865	0.213	7933048	54	0.001	11222440	40811	0.364
2016-02-26	19895000	41891	0.211	8316062	62	0.001	11578938	41829	0.361
2016-02-27	18410754	33903	0.184	7817765	89	0.001	10592989	33814	0.319
2016-02-28	16341019	27971	0.171	7006815	61	0.001	9334204	27910	0.299
2016-02-29	17041782	34080	0.2	7049968	91	0.001	9991814	33989	0.34
2016-03-01	17519578	34691	0.198	7264171	72	0.001	10255407	34619	0.338
2016-03-02	17260480	33085	0.192	7182545	62	0.001	10077935	33023	0.328
2016-03-03	17326476	32095	0.185	7175051	112	0.002	10151425	31983	0.315
2016-04-18	13094984	24745	0.189	5480281	43	0.001	7614703	24702	0.324
2016-04-19	13228953	26010	0.197	5561693	91	0.002	7667260	25919	0.338
2016-04-20	13716969	23802	0.174	5943181	72	0.001	7773788	23730	0.305
2016-04-21	13592572	21256	0.156	6080560	37	0.001	7512012	21219	0.282
2016-04-22	15716545	23431	0.149	7204431	65	0.001	8512114	23366	0.275
2016-04-23	14539943	20743	0.143	6686559	116	0.002	7853384	20627	0.263
2016-04-24	32844244	8889139	27.065	16533941	4154960	25.13	16310303	4734179	29.026
2016-04-25	61648530	37252611	60.427	27532332	15483150	56.236	34116198	21769461	63.81
2016-04-26	51762664	30028854	58.013	23271706	12924148	55.536	28490958	17104706	60.036
2016-04-27	51443095	30894233	60.055	22441805	12790979	56.996	29001290	18103254	62.422
2016-04-28	49571047	30584075	61.697	21060520	12196311	57.911	28510527	18387764	64.495
2016-04-29	45740028	28088101	61.408	19304866	11067972	57.333	26435162	17020129	64.384
2016-04-30	44747668	28206684	63.035	18482843	10821694	58.55	26264825	17384990	66.191
2016-05-01	38826732	26924745	69.346	16407601	10690627	65.157	22419131	16234118	72.412
2016-05-02	36971583	24016084	64.958	15561500	9739413	62.587	21410083	14276671	66.682

Continued on next page

Continuation of Table 8.18

Date	Total playbacks	Affected playbacks	Affected playbacks %	Total online playbacks	Affected online playbacks	Affected online playbacks %	Total offline playbacks	Affected offline playbacks	Affected offline playbacks %
2016-05-03	34043783	19841920	58.284	14283511	7939668	55.586	19760272	11902252	60.233
2016-05-04	32862327	18701918	56.91	14013298	7602803	54.254	18849029	11099115	58.884
2016-05-05	31535242	17950753	56.923	13427299	7319756	54.514	18107943	10630997	58.709
2016-05-06	30227071	15728090	52.033	12798897	6397223	49.983	17428174	9330867	53.539
2016-05-07	26738540	13342584	49.9	11477044	5760631	50.193	15261496	7581953	49.68
2016-05-08	20337761	5342108	26.267	8893020	2327690	26.174	11444741	3014418	26.339
2016-05-09	19254690	33567	0.174	8408779	96	0.001	10845911	33471	0.309

Table 8.19: Six minutes affected users playbacks

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-01-21	35	2856	0	1843	0	0.0	0.0	0
2016-01-22	32	2871	0	1836	0	0.0	0.0	0
2016-01-23	56	2855	0	1606	0	0.0	0.0	0
2016-01-24	38	2676	0	1650	0	0.0	0.0	0
2016-01-25	48	2736	0	1814	0	0.0	0.0	0
2016-01-26	32	2298	0	1495	0	0.0	0.0	0
2016-01-27	43	3215	0	1564	0	0.0	0.0	0
2016-01-28	81	26818	0	3928	0	0.0	0.0	0
2016-01-29	56	15804	0	7561	0	0.0	0.0	0
2016-01-30	77	11464	0	8388	0	0.0	0.0	0
2016-01-31	29	8286	0	6879	0	0.0	0.0	0
2016-02-01	55	8655	0	7551	0	0.0	0.0	0
2016-02-02	35	7712	0	6999	0	0.0	0.0	0
2016-02-03	51	7445	0	6968	0	0.0	0.0	0
2016-02-04	30	6883	0	7043	0	0.0	0.0	0
2016-02-05	34	6974	0	5544	0	0.0	0.0	0
2016-02-06	34	6898	0	3971	0	0.0	0.0	0
2016-02-07	51	8666	0	2930	0	0.0	0.0	0
2016-02-08	44	9162	0	3375	0	0.0	0.0	0
2016-02-09	44	8524	0	4623	0	0.0	0.0	0
2016-02-10	56	7796	0	4357	0	0.0	0.0	0
2016-02-11	34	6871	0	4914	0	0.0	0.0	0
2016-02-12	55	6649	0	4848	0	0.0	0.0	0
2016-02-13	32	6393	0	4081	0	0.0	0.0	0
2016-02-14	13463259	11832874	0	2340646	18223353	65.939	8536670	9686683
2016-02-15	16350436	12954065	0	6876334	23135700	63.945	9614250	13521450
2016-02-16	15733229	12949491	0	7841530	21667799	59.324	8172642	13495157
2016-02-17	15176282	12590919	0	8441639	19833846	54.776	7003216	12830630
2016-02-18	14144850	11802194	0	8233516	17197908	50.315	5811364	11386544
2016-02-19	13890396	11730727	0	8131911	15643115	46.346	5099225	10543890
2016-02-20	12661642	10457810	0	7050750	12480282	41.366	3879253	8601029

Continued on next page

Continuation of Table 8.19

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-02-21	11110721	8642778	0	6135626	9962709	38.482	3145303	6817406
2016-02-22	11716634	10130726	0	6409110	11141095	39.428	3519729	7621366
2016-02-23	11683244	10310383	0	6300267	10243031	36.202	3118388	7124643
2016-02-24	62	23929	0	18572	25803	60.623	27	25776
2016-02-25	54	22979	0	17832	23822	58.294	14	23808
2016-02-26	62	24186	0	17643	24160	57.673	21	24139
2016-02-27	89	18638	0	15176	17695	52.193	10	17685
2016-02-28	61	15543	0	12367	14004	50.066	11	13993
2016-02-29	91	18647	0	15342	18184	53.357	14	18170
2016-03-01	72	19309	0	15310	17619	50.788	10	17609
2016-03-02	62	17935	0	15088	16246	49.104	8	16238
2016-03-03	112	18021	0	13962	15422	48.051	13	15409
2016-04-18	43	16786	0	7916	0	0.0		
2016-04-19	91	16645	0	9274	0	0.0		
2016-04-20	72	17134	0	6596	0	0.0		
2016-04-21	37	14971	0	6248	0	0.0		
2016-04-22	65	17188	0	6178	0	0.0		
2016-04-23	116	14082	0	6545	0	0.0		
2016-04-24	4154960	4586056	0	148123	4491857	50.532		
2016-04-25	15483150	17683516	0	4085945	31019685	83.268		
2016-04-26	12924148	13004545	0	4100161	17057282	56.803		
2016-04-27	12790979	12993169	0	5110085	18279983	59.17		
2016-04-28	12196311	12815458	0	5572306	17106522	55.933		
2016-04-29	11067972	11818926	0	5201203	14688085	52.293		
2016-04-30	10821694	11659064	0	5725926	17427384	61.785		
2016-05-01	10690627	10691857	0	5542261	13780163	51.18		
2016-05-02	9739413	9335636	0	4941035	10853979	45.195		
2016-05-03	7939668	7597812	0	4304440	7735233	38.984		
2016-05-04	7602803	7181958	0	3917157	6226843	33.295		
2016-05-05	7319756	6956214	0	3674783	5080394	28.302		
2016-05-06	6397223	6249872	0	3080995	3491402	22.199		
2016-05-07	5760631	5261225	0	2320728	1480767	11.098		

Continued on next page

Continuation of Table 8.19

Date	Online users online playbacks	Online users offline playbacks	Offline users online playbacks	Offline users offline playbacks	Affected playbacks from albums	Affected playbacks from albums %	Affected online playbacks from albums	Affected offline playbacks from albums
2016-05-08	2327690	2065418	0	949000	671018	12.561		
2016-05-09	96	21549	0	11922	10869	32.38		

8.6.10 Summary of findings

This subsection summarises our findings for each suspicious period. Tables 8.20 and 8.21 summarises the suspicious days in each period, respectively. While Table 8.22 summarises all suspicious days from the two periods. We can see that 93.79% of users were affected by at least one of the impossible scenarios, while 83.985% of all users were affected in the second suspicious period. It was in total 90.828% of users affected of these impossible scenarios in these two periods.

Tables 8.23 and 8.24 summarises how many playbacks was affected by each impossible scenario. Almost 319 000 000 playbacks were affected during the ten suspicious days in the first period, that is 58.454% of all playbacks occurring during the same period. From these affected playbacks, there are almost 159 000 000 playbacks who have played tracks found in 'Life of Pablo', which means that half (49.851%) of all affected playbacks had played tracks from this album.

Table 8.24 show that 57.683% of all playbacks were affected with at least one impossible scenario. While 67.928% of those affected playbacks had track IDs from the 'Lemonade' album. Note that tables in Subsection 8.6.9 accounts for the first playback as being legitimate; however, tables in this section does not take this into account, which means that numbers in Tables 8.24 and 8.25 contain the first legitimate playback divisible by six minutes.

Table 8.20: Affected users 2016-02-14 - 2016-02-23

Type	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
Two duplicates	1461669	1158967	79.291	1405818	1060506	75.437	715305	467000	65.287
Three or more duplicates	1461669	908178	62.133	1405818	813622	57.875	715305	323737	45.259
Two unequal duplicates	1461669	296867	20.31	1405818	186840	13.29	715305	123268	17.233
Three or more unequal duplicates	1461669	1332871	91.188	1405818	1244367	88.516	715305	619381	86.59
All	1461669	1370899	93.79	1405818	1285582	91.447	715305	647518	90.523

Table 8.21: Affected users 2016-04-24 - 2016-05-08

Type	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
Two duplicates	2086838	498917	23.908	2032841	408721	20.106	956754	153834	16.079
Three or more duplicates	2086838	116652	5.59	2032841	94961	4.671	956754	30492	3.187
Two unequal duplicates	2086838	55346	2.652	2032841	42250	2.078	956754	15449	1.615
Three or more unequal duplicates	2086838	10201	0.489	2032841	8614	0.424	956754	5054	0.528
Six minutes	2086838	1737008	83.236	2032841	1660464	81.682	956754	703241	73.503
All	2086838	1752636	83.985	2032841	1676305	82.461	956754	721858	75.449

Table 8.22: Affected users 2016-02-14 - 2016-02-23 and 2016-04-24 - 2016-05-08

Type	Total users	Affected users	Affected users %	Total online users	Affected online users	Affected online users %	Total offline users	Affected offline users	Affected offline users %
Two duplicates	2915953	1508749	51.741	2858752	1353014	47.329	1450699	592913	40.871
Three or more duplicates	2915953	989679	33.94	2858752	881620	30.839	1450699	349314	24.079
Two unequal duplicates	2915953	346419	11.88	2858752	225863	7.901	1450699	137918	9.507
Three or more unequal duplicates	2915953	1336364	45.829	2858752	1247680	43.644	1450699	622007	42.876
Six minutes	2915953	2638764	90.494	2858752	2533264	88.614	1450699	1233476	85.026
All	2915953	2648493	90.828	2858752	2544783	89.017	1450699	1248529	86.064

Table 8.23: Affected playbacks 2016-02-14 - 2016-02-23

Type	Total playbacks	Affected playbacks	Affected playbacks %	Affected playbacks from albums	Affected playbacks from albums %
Two duplicates	545614012	64618580	11.843	36830106	56.996
Three or more duplicates	545614012	102776725	18.837	74166485	72.163
Two unequal duplicates	545614012	1374574	0.252	644663	46.899
Three or more unequal duplicates	545614012	314173882	57.582	156992700	49.97
All	545614012	318934712	58.454	158991922	49.851

Table 8.24: Affected playbacks 2016-04-24 - 2016-05-08

Type	Total playbacks	Affected playbacks	Affected playbacks %	Affected playbacks from albums	Affected playbacks from albums %
Two duplicates	589300315	7212646	1.224	4299384	59.609
Three or more duplicates	589300315	2531457	0.43	1093581	43.2
Two unequal duplicates	589300315	186960	0.032	136157	72.827
Three or more unequal duplicates	589300315	1176825	0.2	285073	24.224
Six minutes	589300315	335791899	56.981	169390597	50.445
All	589300315	339927886	57.683	230906389	67.928

Table 8.25: Affected playbacks 2016-02-14 - 2016-02-23 and 2016-04-24 - 2016-05-08

Type	Total playbacks	Affected playbacks	Affected playbacks %	Affected playbacks from albums	Affected playbacks from albums %
Two duplicates	1134914327	71831226	6.329	41129490	57.259
Three or more duplicates	1134914327	105308182	9.279	75260066	71.466
Two unequal duplicates	1134914327	1561534	0.138	780820	50.003
Three or more unequal duplicates	1134914327	315350707	27.786	157277773	49.874
Six minutes	1134914327	652885888	57.527	328919435	50.379
All	1134914327	658862598	58.054	389898311	59.177

8.7 Conclusion

We were approached by DN to investigate possible fraudulent manipulation of data in the database of a music streaming service. We have through advanced statistical analysis determined that there has, in fact, been a manipulation of the data at particular times. The manipulation appears targeted towards a very specific set of track IDs, related to two distinct albums.

It is difficult to determine the exact cause and means of the manipulation, but it is likely that several methods were used. The manipulation looks to have become more sophisticated during the period for which we have data. It starts with simple duplication and possible insertion of fabricated playbacks of tracks, to more advanced (and difficult to detect) manipulation at the timestamp by adjusting the timestamps of duplicates with something evenly divisible by six minutes. The advanced manipulation was more difficult to detect because playbacks have not been simply duplicated or inserted into the log files. Our analysis also shows that a considerable amount of system users was affected by the manipulation during these days.

It is very unlikely (but not impossible) the manipulation is the result of an external attack or that an outside source has affected the accuracy of the data. The absence of noise in the data and log files suggests that a Structured Query Language (SQL) based attack was not the cause of the manipulation. Also, unrelated third-party attackers do not have motivation for manipulating the playbacks for very specific tracks. Our analysis also shows a significant number of system users were affected by the manipulation, which may exclude an external or user originated manipula-

tion. As such, the manipulation likely originates from within the streaming service itself. Due to the targeted nature and extent of the manipulation, it is very unlikely that this manipulation was solely the result of a code-based bug or other anomalies.

8.8 Statement of conflicts

The writers of this report confirm that we have no conflict of interest of any kind, other than any which are set out below. The writers will advise DN, if between the date of this report and any further request if there is any change in circumstances which affects this statement.

8.9 Bibliography

[1] Frank Benford. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4):23, March 1938.

[2] Andreas Diekmann and Ben Jann. Benford's Law and Fraud Detection: Facts and Legends. *German Economic Review*, 11(3):397 – 401, 2010. http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1125127&site=ehost-live.

[3] Cindy Durtschi, William Hillison, and Carl Pacini. The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34, 2004.

[4] Dongdong Fu, Yun Q Shi, and Wei Su. A generalized Benford's law for JPEG coefficients and its applications in image forensics. In *Security, Steganography, and Watermarking of Multimedia Contents IX*, volume 6505, page 65051L. International Society for Optics and Photonics, 2007.

[5] Mark J Nigrini and Linda J Mittermaier. The use of Benford's law as an aid in analytical procedures. *Auditing*, 16(2):52, 1997. Publisher: American Accounting Association.

Appendix A

Summary of appendices

Our casework report contains several appendices which are not essential to include in this thesis. Therefore, we refer readers to our casework report at

> Jan William Johnsen and Katrin Franke. 'Digital Forensics Report for Dagens Næringsliv'. Dagens Næringsliv, page 78, April 2018.

Alternatively, access the report by visiting https://www.dn.no/staticprojects/special/2018/05/09/0600/dokumentar/strommekuppet/data/documentation/NTNU-rapport_til_publisering.pdf (last accessed 17. July 2022).

The report contains the following appendices:

A Additional modulo six summaries

- Show the identical duplicates with timestamps at 01:00:00, 01:30:00, 04:00:00 and 04:30:00 on page 56.

B Code for the serial analysis method

- This code consists of a for-loop over all the files and executes the same analysis for each file. This code is found on page 63.

C Code for analysis method 2.1 - 2.3

- The main analysis code for methods 2.1, 2.2 and 2.3 is found on page 66.

D Code for auxiliary analysis method 2.4 - 2.6

- The code for intermediate analysis step for methods 2.4, 2.5 and 2.6 is found on page 67.

E Code for final analysis method 2.4 - 2.6

- The final code for analysis methods 2.4, 2.5 and 2.6 is found on page 68.

F Code for analysis method 9

- The analysis code for method 9 is found on page 70.

G Log files received

- Contains a list of MD5 sums and the size for each file we received on page 73.

Appendix B

Errata list

- Moved the Abbreviations and Glossaries sections to the beginning of the thesis.
- Section 1.2, page 5. Clarifying the thesis' aim to aid in distinguish between proficient cybercriminals and other non-proficient actors. This thesis does not identify the role or exact profile (e.g. reverse engineer, malware/exploit developer, service provider, etc.) of cybercriminals.
- Section 1.4, page 8. Added a new Section 1.4 describing the research methodology.
  - Subsection 1.4.1, page 8. Added a new Subsection 1.4.1 to describe the datasets used in our experiments and the data processing cycle.
  - Subsection 1.4.2, page 12. Added a new Subsection 1.4.2 to describe the general research methodology applied by this thesis.
- Section 1.6, page 14. Added a new subsection in Section 1.6 describing the relationship between publications and research questions. The new subsection includes a graphical illustration of their relationship.
- Subsection 1.7.1, page 19. Moved the description of the Daubert standard into its own subsection so it can be explicitly referenced in the thesis.
- Section 5.3, page 133. Changed the numbering in Figure 5.1 to correctly reference the subsections in this thesis.
- Minor page-layout adjustments to arrange figures with the text description.

ISBN 978-82-326-5817-6 (printed ver.)
ISBN 978-82-326-6833-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology