

UNIVERSITY OF TWENTE

**So long and thanks for all the (big) fish**  
**Exploring cybercrime in Dutch Telegram groups**

M.Sc. thesis by

Kitty Boersma

June 2023

Under supervision of:

Dr.Ir. F.W. Hahn (University of Twente)

Dr. R.S. van Wegberg (TU Delft)

Ir. H.L.J. Bijmans (TNO)

in the

Faculty of Electrical Engineering, Mathematics and Computer Science

## *Abstract*

According to prominent law enforcement agencies and cyber security organisations around the world, cybercrime on social media is a growing problem. This work addresses the issue by contributing to a deeper understanding of Telegram as a cybercrime platform. We propose a method for data collection and preprocessing, and create a dataset containing the messages and users of public Dutch Telegram groups used for cybercrime. The dataset is used to create holistic profiles of the Telegram groups and discover relations between them using new and existing approaches. The resulting highly connected mention network suggests that the collective network should be seen as a professionally used cybercrime platform, transcending the perception of isolated groups. Additionally, this work introduces a novel approach for identifying relations between active users through overlap in sent advertisements, revealing that multiple user accounts are likely managed by the same (groups of) individuals. Building upon these findings and leveraging related work, we successfully identify influential Telegram groups and users, the so-called big fish. Overall, this study evaluates the Dutch cybercrime landscape on Telegram, while enhancing our understanding of Telegram as a cybercrime platform through the implementation of novel profiling techniques and the identification of interconnected relationships.

# *Acknowledgements*

Writing this thesis and conducting the research has been quite a journey. And while the saying “*It’s about the journey, not the destination*” usually makes me a little sick, I’m convinced this journey might have been one of the most interesting ones in my life so far. It has definitely been the longest project I worked on - and quite possibly also the one that required the most motivation - which is a feat in itself if you ask me. It is because of this that I’ve started to wonder if the topic of my thesis might have grown from being just the topic of my thesis into a passion which I can pursue in the future. That said, the destination, being the work you have in front of you right now, is something that I’m equally proud of. I hope you will enjoy reading it, feel inspired, and maybe even learn something.

This work would not have been the same if it wasn’t for my supervisors Rolf, Hugo, and Florian. It was Rolf who came up with the idea of investigating cybercrime on Telegram after I asked him if he could help me brainstorm topics for my thesis. It were Rolf and Hugo who found a place for me at TNO and helped give my research direction - a daunting task in a mostly unexplored topic. And it were Hugo and the members of the cybercrime team at TNO who always notified me of new publications and asked how my research was going. Finally, it was Florian who took the role of chair with unmatched enthusiasm, even though the topic was as new to him as it was to me. So, thank you, Rolf, Hugo, Florian, and cybercrime team at TNO. You really helped me through it.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research questions & objectives . . . . .	2
1.2 Contributions & motivation . . . . .	3
1.3 Organisation . . . . .	4
<b>2 Background &amp; related work</b>	<b>5</b>
2.1 Cybercrime . . . . .	5
2.1.1 Types of cybercrime in literature . . . . .	6
2.1.2 Where does cybercrime take place? . . . . .	8
2.1.2.1 Dark web markets & underground forums . . . . .	9
2.1.2.2 Shift to social media . . . . .	11
2.1.2.3 Local embeddedness of cybercrime . . . . .	12
2.2 Social media . . . . .	13
2.2.1 Social relations . . . . .	14
2.2.2 Cybercrime & social media . . . . .	16
2.3 Telegram . . . . .	16
2.3.1 Users and bots . . . . .	19
2.3.2 Cybercrime on Telegram . . . . .	19
2.4 Conclusion . . . . .	20
<b>3 Approach</b>	<b>22</b>
3.1 Data collection . . . . .	22
3.1.1 Ethics . . . . .	24
3.1.2 Scraper . . . . .	24
3.1.3 Snowball . . . . .	25
3.1.4 Starting point . . . . .	25
3.2 Preprocessing . . . . .	26
3.2.1 Anonymisation . . . . .	27
3.3 Classification . . . . .	27
3.3.1 Ground truth . . . . .	28
3.3.2 Classifier training . . . . .	29
3.3.3 Classification . . . . .	31

---

3.3.4	Performance . . . . .	32
3.4	Results . . . . .	32
3.4.1	Data collection . . . . .	32
3.4.2	Preprocessing . . . . .	34
3.4.3	Classification . . . . .	35
3.5	Conclusion . . . . .	37
<b>4</b>	<b>Phase 1: Groups</b>	<b>39</b>
4.1	Approach . . . . .	39
4.1.1	Basic info . . . . .	39
4.1.1.1	Clustering . . . . .	42
4.1.2	Mention network . . . . .	42
4.1.2.1	Community detection . . . . .	45
4.1.3	Advertisement and member overlap . . . . .	45
4.2	Results . . . . .	46
4.2.1	Group descriptives . . . . .	46
4.2.1.1	Aliases and crime markets . . . . .	48
4.2.1.2	Channels . . . . .	49
4.2.1.3	Clustering . . . . .	50
4.2.2	Mention network . . . . .	51
4.2.3	Advertisement and member overlap . . . . .	58
4.2.4	Community detection using Louvain method . . . . .	60
4.3	Conclusion . . . . .	61
<b>5</b>	<b>Phase 2: Users</b>	<b>62</b>
5.1	Approach . . . . .	62
5.1.1	Basic info . . . . .	62
5.1.1.1	Roles . . . . .	63
5.1.1.2	Previously active users . . . . .	63
5.1.2	Relations . . . . .	63
5.1.2.1	Most-sent message . . . . .	64
5.2	Results . . . . .	65
5.2.1	Basic info . . . . .	65
5.2.1.1	Bots . . . . .	66
5.2.1.2	Membership . . . . .	67
5.2.1.3	Activity . . . . .	68
5.2.1.4	Previously active members . . . . .	68
5.2.2	Relations . . . . .	69
5.2.2.1	Message overlap . . . . .	70
5.3	Conclusion . . . . .	78
<b>6</b>	<b>Phase 3: Influence</b>	<b>80</b>
6.1	Groups . . . . .	81
6.1.1	Influence based on group properties . . . . .	81
6.1.2	Influence of groups in the mention network . . . . .	83
6.1.3	Approach . . . . .	84
6.1.4	Results - groups . . . . .	84

---

6.1.4.1	Advertisements . . . . .	84
6.1.4.2	Correlation . . . . .	86
6.1.4.3	Influential groups . . . . .	86
6.2	Users . . . . .	88
6.2.1	Important users in the network . . . . .	88
6.2.2	Important users within a single Telegram group . . . . .	90
6.2.3	Approach . . . . .	91
6.2.4	Results . . . . .	92
6.3	Conclusion . . . . .	95
<b>7</b>	<b>Discussion</b>	<b>97</b>
7.1	Telegram as a platform for cybercrime . . . . .	97
7.2	Limitations . . . . .	99
7.2.1	Approach . . . . .	100
7.2.2	Phase 1 . . . . .	102
7.2.3	Phase 2 . . . . .	103
7.2.4	Phase 3 . . . . .	104
7.3	Implications . . . . .	105
7.4	Recommendations . . . . .	106
<b>8</b>	<b>Conclusion</b>	<b>108</b>
<b>A</b>	<b>Appendix</b>	<b>111</b>
	<b>References</b>	<b>115</b>

# Chapter 1

## Introduction

Worldwide, an estimated 4.9 billion people use the Internet (International Telecommunication Union, 2021). It is, therefore, unsurprising that cybercrime rates are skyrocketing, even though crime rates have declined in most Western countries (Dupont, 2017; Harkin, Whelan, & Chang, 2018; Kemp, Miró-Llinares, & Moneva, 2020). It is difficult to estimate its total costs, as the majority of cyber fraud and cyber crime is never reported to the police (Dupont, 2017). However, its impact is large. That said, law enforcement agencies still regularly arrest and stop malicious groups. For example, joint law enforcement efforts have blocked (*New major interventions to block encrypted communications of criminal networks*, 2021) and taken down (*800 criminals arrested in biggest ever law enforcement operation against encrypted communication*, 2021) encrypted communication systems used by (cyber)criminals across the world. Another great example is the takeover of Hansa Market by the Dutch Police, where the police ran the illegal marketplace for three weeks to gather information about its users before shutting it down.

According to Europol (2021), EU law enforcement identified the fragmentation of the Dark Web as a threat to the current cyber landscape. With this, they mean that there is an increase in single vendor shops and small TOR markets. Moreover, they reported an increase in the usage of encrypted communication platforms outside of Dark Web Markets for the sale of illegal goods and services. Specifically Telegram and Wickr were mentioned as popular platforms.

Research on traditional dark web markets and their response to law enforcement operations coexist with research on the impact of social media applications on communication. Snapchat user patterns (Lukasz Piwek, 2016), group-based communication in WhatsApp (Seufert, Hoßfeld, Schwind, Burger, & Tran-Gia, 2016), and the characteristics of viral messages on Telegram (Dargahi Nobari, Sarraf, Neshati, & Erfanian Daneshvar, 2021) have been studied recently. Although research has been conducted on the use of social media for cybercrime, such as the sale of drugs (Blankers, van der Gouwe, Stegemann, & Smit-Rigter, 2021; Moyle, Childs, Coomber, & Barratt, 2019), no research has been conducted on the cybercriminal landscape on a single social media platform in

a specific region. Moreover, no published work has mapped the relationships between the players in these networks and determined the significance of the players.

The study of cybercrime is often at the intersection of criminology and computer science, with research topics such as dark web markets (Celestini, Me, & Mignone, 2016; Van Wegberg & Verburgh, 2018) and online forums (Motoyama, McCoy, Levchenko, Savage, & Voelker, 2011; Zamani, Rabbani, Horicsányi, Zafeiris, & Vicsek, 2019). ENISA and Europol have both reported the growing use of social media as a platform for crime. However, research on social media platforms often focuses on different social media platforms, their use, and the relations between their users. Little research has been done on social media use for crime. Despite the research, most of the research platforms are focused on dark web markets and underground forums.

## 1.1 Research questions & objectives

In light of the lack of research on the topic, we decided to narrow the gap by investigating the use of the public portion of Telegram for cybercrime. Our objectives are twofold: to gain a better understanding of the use of Telegram for cybercrime and to gain insights into the Dutch cybercrime landscape. This is done by introducing a method that aims to point to the more interesting users and Telegram groups. Telegram was chosen as our focus due to its use for the promotion and sale of illicit goods and services, the availability of public data in its public groups, and the relative ease with which data can be collected using Telegram's API.

Aligned with our objectives, the main question of this study is *How can we identify influential users in Dutch Telegram groups used for cybercrime?*. The research consists of three phases, with each phase serving as a smaller study in its own right. The first phase focuses on examining the Telegram groups, while the second phase investigates the users within these groups. Finally, the third phase aims to determine the influence of both the groups and the users. The approach and findings are presented for each phase, with subsequent phases potentially drawing on the outcomes of preceding phases.

The first step of this research is to familiarize ourselves with the context in which cybercrime takes place on Telegram. We intend to do this by characterizing the Telegram groups in which cybercrime takes place and mapping the relations between them. For this purpose, we propose the following research questions:

1. To what extent can we create a profile of the Dutch Telegram groups used for cybercrime?
2. To what extent can we define the relations between the Telegram groups?

The second step of this work is to categorize the members of the Dutch Telegram groups. This provides us with insight into the players in the Dutch cybercrime landscape, as well as that it helps our efforts in identifying characteristics of influential users. The following research questions help us with this objective:



1. To what extent can we categorize the members of the Dutch Telegram groups that are used for cybercrime?
2. To what extent can we define the relations between the group members?

The final step of the proposed research is then to use the information from the previous phases to develop a method for determining the influence of the Telegram group members, and ultimately applying the method to a dataset. To achieve this, we define the following research questions:

1. What methods can we use to determine the influence of subcommunities (groups)?
2. What methods can we use to determine the influence of group members?
3. To what extent can we combine these methods and apply them to the Dutch Telegram groups?

## 1.2 Contributions & motivation

This study aims to address the knowledge gap surrounding the use of Telegram for cybercrime in the Netherlands. By reducing or closing this gap, multiple parties can benefit. Firstly, it could aid law enforcement agencies by enabling them to gain insights into the Dutch cybercrime landscape on Telegram. This would equip them with theoretical knowledge about cybercrime and Telegram, as well as a practical application for profiling both Telegram groups and users. Not only could this assist law enforcement in identifying key players on Telegram, but a similar methodology could also be adapted for use on other social media platforms. Secondly, reducing the knowledge gap could also benefit academics. The ability to point to influential users on social media platforms has broad-ranging implications for research across multiple fields. Additionally, this research could offer novel perspectives on cybercrime, given that its transition to social media platforms has occurred only recently. As such, it could pave the way for a new wave of cybercrime research.

This work:

- Gives an overview of the context of cybercrime on social media, including potential actors, types of cybercrime, the local embeddedness of cybercrime, and the appeal of using Telegram for this purpose.
- Proposes a method to collect and preprocess data that can be used to gain insight into Telegram groups, a network of Telegram groups, and individual users on Telegram in the context of cybercrime.
- Proposes a method to find interesting or influential Telegram groups and users, i.e. the big(ger) fish.

- Describes the use of Telegram as a platform for cybercrime, which, to our knowledge, has not been done before in academic literature.
- Offers insight into the current cybercrime landscape on Telegram in the Netherlands. This is done by showing the size and scale of the platform, providing evidence that the platform is deliberately created and carefully used, and giving examples of the professional operations of vendors on the platform.
- Discusses the implications of the findings of this study and provides recommendations to law enforcement and policymakers.

### 1.3 Organisation

The organisation of this thesis is as follows. In Chapter 2, we provide the necessary background information and use related works to create a frame of reference that is used in the continuation of this work. Then, we describe the approach taken in this study in Chapter 3, followed by a description of the classifier and the resulting dataset used in this work. The first phase, Chapter 4, dives into the Telegram groups in our dataset, exploring both the individual group properties and the relations between the groups in a resulting mention network. The next step is to look into the users in our dataset, which is done in Chapter 5. The third phase, Chapter 6, tries to determine the prominent Telegram groups and users in the dataset while taking a closer look at the nature of both in the context of Telegram as a cybercrime platform. Finally, we discuss the results in Chapter 7 and conclude our work in Chapter 8.

## Chapter 2

# Background & related work

With the growth of the Internet, challenges that come with this growth have emerged – including challenges related to crime in online environments (Abiodun Raufu and Lucy Tsado and Emmanuel Ben-Edet, 2021; Cascavilla, Tamburri, & Van Den Heuvel, 2021). The Internet has an enabling effect on previously existing crimes, as well as on the invention of entirely new criminal acts (Jewkes & Yar, 2013; E. R. Leukfeldt, Lavorgna, & Kleemans, 2017; Wall, 2017). Cybercrime fundamentally alters the nature of traditional (offline) crimes. Correspondingly, cybercriminals use virtual meeting places much in the same way traditional criminals meet in physical locations to relax, meet new people, or exchange information (Soudijn & Zegers, 2012). As Chandra and Snowe (2020) illustrates: *“a security issue is a cybersecurity issue, if it involves the use of a computer, its related technology and the networked system in which it functions to inflict harm—tangible or intangible—on the victim.”* (p.1).

This chapter aims to provide the reader with an understanding of the context in which the study takes place. We start with a description of cybercrime by relating to studies in the field of criminology and computer science in Section 2.1. In Section 2.2 we focus our attention towards social media and highlight related studies about the roles and relations of social media users as well as studies about cybercrime on social media. Finally, in Section 2.3 we include a description of Telegram and detail specific elements that are relevant to this work.

### 2.1 Cybercrime

The cybercrime research community studies online crime, Internet-facilitated crime, and the behaviour of cybercriminals, amongst others (Katsikeas, Johnson, Ekstedt, & Lagerström, 2021). This includes, for example, crime happening on online black markets and vendors selling illegal goods on social media.

It is difficult to find a definition of cybercrime that is set in stone. Abiodun Raufu and Lucy Tsado and Emmanuel Ben-Edet (2021) states that cybercrime is generally seen as an unlawful action that creates, distributes, changes, steals, misuses and destroys

data through the manipulation of computer software which is done against the will of the victim. C.-B. Europol (2017) uses the working definition “*any crime that can only be committed using computers, computer networks or other forms of information communication technology (ICT)*” (p.18). According to E. R. Leukfeldt, Veenstra, and Stol (2013), the Dutch police did not have a single definition of cybercrime in the time leading up to its publication in 2013. Definitions differed from ‘only crimes that are committed on computer systems’, such as hacking or spreading viruses, to ‘crimes where the perpetrator has merely used a digital component’. E. R. Leukfeldt et al. (2013) continues that the term ‘cybercrime’ can be defined as “*the overarching concept for all kinds of crime whereby ICT plays a significant role in the committing of the offence.*” (p. 3). We adopt this definition, as it seems it can be applied to the topic of this work.

Often, a distinction is made between cyber-dependent crime (i.e. crime targeted at computers or online environments) and cyber-enabled crime (i.e. crime facilitated by computers and online environments) (Broadhead, 2018; Brown, 2015; Harfield & Schofield, 2020; E. R. Leukfeldt, Kleemans, Kruisbergen, & Roks, 2019). Examples of cyber-dependent crime are DDoS attacks, malware and ransomware attacks, botnets, and network attacks. Cyber-enabled crime encompasses a broader range of crimes, such as scams, payment fraud, and selling illegal goods such as drugs or weapons. Since this thesis looks at cyber-enabled crime, this is what we mean when using the term cybercrime.

Recent reports have observed the emergence of cybercrime-as-a-service (Malin, Gudaitis, Holt, & Kilger, 2017; van Wegberg et al., 2018). In this model, specialized services, products or resources are catered to criminal entrepreneurs. Hence, they can focus on what to buy instead of needing to learn it themselves. This commoditization lowers the barriers for criminals, as a criminal act requires less technical expertise for the criminal to be successful. A great example of this is the wide range of phishing kits that are on offer online. Research has shown that there are off-the-shelf, easy-to-employ phishing kits available for whoever is willing to pay. They often come with a manual on how to use the kits, a step to step explanation on how to use the kit, and phishing websites that are ready to be deployed (Bijmans, Booij, Schwedersky, Nedgabat, & van Wegberg, 2021).

### 2.1.1 Types of cybercrime in literature

Cybercrime includes a broad range of different crimes. Crimes can be motivated by passion, ideology, or revenge, but also by economic gains for the attacker (E. R. Leukfeldt et al., 2017). We might go even further and follow Barn and Barn (2016) conceptual model of cybercrime, where motivation is divided into intrinsic and extrinsic. Intrinsic motivation, they argue, encompasses mischief, ego, challenge, or morality. Extrinsic motivation is espionage, ideology, financial gain, or revenge.

When looking to create a taxonomy of the different types of cybercrime, Chandra and Snowe (2020) took a victim-centric approach. They determined that the victims of cybercrime could be of the technological kind; being divided into computer systems,

related technology, and networked systems. If cybercrime victims are not of the technological kind, they can be divided into (a group of) natural person(s), crimes against property<sup>1</sup>, and crimes against the government. It is important to realise that the victim of a crime is not need not always the target of the crime. A victim is an entity that is affected by a cybercriminal act, while the target is the entity that the cybercriminal act is specifically directed at (Donalds & Osei-Bryson, 2019). To illustrate, a massive DDoS attack targeted DNS provider Dyn in 2016 (*Famous DDoS attacks*, n.d.), which created disruptions for many major sites, such as Netflix, PayPal, Visa, Amazon, Reddit, and GitHub, and could last up to a whole day. Anyone unable to access these sites can be seen as a victim in this particular case, while the target of the attack was Dyn.

It is a well-known fact in the criminological field that most criminals cooperate quite intensively with other criminals. (E. R. Leukfeldt, Kruisbergen, Kleemans, & Roks, 2020) However, in literature, partly opposed points of view on the organizational structure of cybercriminals can be found. Some imply that cybercriminals prefer to work alone, while others point out that long-lasting, job-sharing cybercriminal networks exist (Odinot, Verhoeven, Pool, & de Poot, 2017). Recent empirical research has pointed out that a number of organized cybercriminal networks have a structure similar to that of traditional criminal networks (E. R. Leukfeldt et al., 2020). An example used by E. R. Leukfeldt et al. (2020) is that most cybercriminal networks studied consisted of a group of core members that committed offences together for a prolonged period of time. Earlier research by E. R. Leukfeldt, Kleemans, and Stol (2016); Odinot et al. (2017) has shown that real-life relationships are important for the origin and growth of cybercriminal networks, despite the options that digitization offers. In a study into organised financial cybercrime, E. R. Leukfeldt et al. (2017) found that none of the studied cybercriminal networks had a strict hierarchical structure. They found that all networks displayed dependency relationships and different functional roles, such as core members, enablers and money mules. It must be noted, however, that definitions of organized crime can be very broad and that many of the relations in a (cyber)criminal network will not be advertised publicly. In other words, we might be unaware of other roles in typical cybercriminal networks.

A literature review by Cascavilla et al. (2021) has looked into cybercrime threat intelligence. They noted that research on this topic has focused on seven different criminal activity types (Cascavilla et al., 2021), namely:

1. Hacktivism groups or virtual criminal networks: hacktivism attacks are often ideologically motivated, while virtual criminal networks are often dedicated to regular crime using online means.
2. Child abuse: exploiting minors for malicious intents and purposes, including human trafficking
3. Cyber-terrorism: crime activities with the goal of causing harm and often ideologically motivated.

---

<sup>1</sup>crimes against property other than the computer ecosystem.

4. Cyber-espionage: criminal activities with the goal of obtaining (secret) information
5. Cyber-warfare: operations in the cyber domain to achieve an operational advantage of military significance.
6. DDoS/Spam/Phishing: crimes against critical infrastructure of individuals, organizations or countries.
7. Cyber drug-trafficking: producing, transporting and selling/buying illegal substances.

That is not to say that these are the only types of cybercrime. For example, cyber drug trafficking as defined above is a subset of the producing, transporting and selling/buying of all kinds of illegal substances.

Due to the nature of the digital ecosystem, cybercrimes are not bound to one platform or location. However, [R. Leukfeldt, Kleemans, and Stol \(2017\)](#) points out that many online meeting places can be viewed as criminal markets where all kinds of stolen goods and services are offered and requested. They distinguish three categories of marketed products:

- Stolen data, such as data from credit cards, bank accounts, PayPal accounts, and identity documents.
- Tools for cybercrime, such as phishing kits, malware, hacking tools, and botnets.
- Cybercriminal services, including escrow services, exchangers, money mule services, bulletproof web hosting, and other cash-out services that cash out money made through criminal activities.

Other studies looked into products and services offered on dark web markets. Typically, dark web markets offer products and services in the following categories: drugs, pharmaceuticals, fraudulent (identity) documents, digital identities, malware and exploit kits, counterfeit goods and other contraband, pornographic material, and weapons ([Ball & Broadhurst, 2021](#); [Celestini et al., 2016](#)). In other words, dark web markets offer a combination of physical illegal products, digital products that can be used to perform more criminal activities, and digital products to be used by the end-user.

### 2.1.2 Where does cybercrime take place?

Traditionally, physical meeting places play an important role in the processes and growth of criminal networks. These *offender convergence settings* are a way for newcomers to connect with other members, enter criminal networks or form new alliances, according to [R. Leukfeldt et al. \(2017\)](#). [Soudijn and Zegers \(2012\)](#) mention that cybercriminals, just like traditional criminals, make use of offender convergence settings - locations to meet new people, relax with friends, exchange information, sell stolen goods, or plan new criminal acts - albeit the digital version of them. According to law enforcement,

both physical and digital offender convergence settings should be suppressed. (Soudijn & Zegers, 2012) However, this is easier said than done in the case of digital settings, such as forums, as digital forums are often located in countries that do not cooperate with requests to take down servers. Nevertheless, digital forums also have advantages for law enforcement; virtually all digital information, including contacts and discussions, is saved somewhere. For example, the forum that was subject of Soudijn and Zegers (2012) served two purposes: exchanging knowledge and offering or requesting products and services.

Soudijn and Zegers (2012) mention several characteristics of the virtual aspects that online forums have over traditional meeting places. First of all, there is no physical contact. This means that people can get to know each other using nicknames and that users are disciplined in a different way in case of undesirable behaviour. In case of conflicts or dissatisfaction, users can give each other bad reviews. Another characteristic is that online forums are not influenced by bad weather and commutes, nor are they affected by closing times. Furthermore, they can be accessed from virtually anywhere through mobile phones or laptops. Finally, an online forum is not bound by a physical room that only fits a number of people, allowing the space to fit many more users than would be feasible in a physical room.

In their study, R. Leukfeldt et al. (2017) distinguishes three functions of online forums used by cybercriminals. Most online meeting places studied in their research were used as a criminal markets where illegal goods and services are offered and requested, the function being a *market function*. Next to that, online meeting places can have a *social function*. They can be used as a place where criminals can meet new people and form alliances, eventually facilitating the growth and development of criminal networks. Finally, they distinguish a *learning function*, as online meeting places are used for communication between cybercriminals and studies have shown that information sharing is part of the cybercrime subculture.

### 2.1.2.1 Dark web markets & underground forums

Marketplaces on the dark web have been a popular place for cybercriminals to meet, sell, and buy illegal goods. (R. Leukfeldt et al., 2017; Soska & Christin, 2015; van Wegberg et al., 2020) While these marketplaces started out on underground chatrooms and discussion rooms, a second generation of illicit online marketplaces, known as cryptomarkets or dark web markets, started to gain traction around 2011 (Chen, 2011). These marketplaces generally look like regular e-commerce websites such as Amazon or eBay (Décarry-Hétu, Paquet-Clouston, & Aldridge, 2016; Soska & Christin, 2015), and offer a plethora of goods and services. Dark web markets are known for their security features. For example, they are accessed by Tor browsers, which guarantee anonymity if used correctly, and many markets allow for payments in cryptocurrencies, using the properties of these coins to provide anonymous payment options.<sup>2</sup> Most markets also offer escrow

<sup>2</sup>It must be noted that not all cryptocurrencies provide anonymity and that the level of anonymity depends on the cryptocurrency used.

services that prevent financial risk and remove the need for trust between vendors and buyers. All in all, dark web markets are designed to protect the identity of their users.

The security features are characteristics that make dark web markets appealing for (cyber)criminals. Not only can one be anonymous on dark web markets due to its use of Tor browsers, but the markets are also not indexed by search engines, such as Google. Additionally, many markets have features that make it difficult (but not impossible) for researchers and law enforcement to scrape and index the platform, and it is not unheard of that users can only access markets when they are invited. In other words, the markets are often designed to keep their users safe and to keep law enforcement at bay. Another characteristic of dark web markets that is advantageous for cybercriminals is the vendor rating system that most markets have (Décary-Hétu et al., 2016; Soska & Christin, 2015). Vendors can use the rating system to build a reputation that can ultimately provide them with more business. As mentioned by Van Wegberg and Verburgh (2018), many vendors migrated from Alphabay to Hansa Market after the 2017 Alphabay takedown. They often kept the same username and/or PGP key (i.e. by which buyers could identify them) in order to keep the reputation they had built. Less than a month later, Hansa Market was taken offline and the Dutch police announced that they had been running the market for several weeks. After this announcement, Van Wegberg and Verburgh (2018) found, several vendors traded their reputation for anonymity now they had to migrate markets once again, suggesting a panic reaction.

The need for technical knowledge is often described as one of the downsides of dark web markets. (Moyle et al., 2019) It can be difficult or daunting for potential users to access the markets. Next to that, users may be scared due to law enforcement's interest in the market places. Additionally, buyers more often carry the risk at the point of delivery, as goods still need to be delivered or picked up. Lucky for the not-so-technically inclined, cybercrime is spreading to other platforms.

Before cryptomarkets or dark web markets came into existence, underground forums were the place to be for (cyber)criminals. These forums allow users to maintain profiles, create buddy lists, post to message boards, and engage in private messaging. Zamani et al. (2019) distinguishes between three types of forums. First, they state, there are public forums that are accessible through regular web search engines, for example, Reddit. These, we argue, do not qualify as underground forums. The second type of forum they define are dark web forums, which refers to the part of the Internet that is not indexed by regular search engines, for example, dark web markets mentioned in the previous paragraphs (e.g. Alphabay, Silk Road, or Dream Market) or forums related to paedophilia or porn. Thirdly, they define semi-dark web forums to describe forums that were initiated in the public domain but later moved to the dark web (and possibly moved back and forth). An example of this type of forum is 8chan, which does not restrict the content of the posts. Underground forums are often used for buying and selling illegal goods and services, or function as a place to exchange knowledge between participants and allow users to find new trading partners. Over time, several of these forums changed to "closed" forums, meaning existing users needed to vouch for new users (Motoyama et al., 2011).



Zamani et al. (2019) found that there tend to be a few very active users on dark web forums, while most users show little activity. They compare this to public forums, which show a much more homogeneous distribution of user activity. Motoyama et al. (2011) looks deeper into the interaction of users of underground forums and defines three relationships between users: a buddy relation between two users, users sending private messages, and users replying to each other on forum threads.

### 2.1.2.2 Shift to social media

Cybercriminals are quick to adopt new technological developments, among which social media. Research by Moyle et al. (2019) finds that the sale of drugs is moving from dark web markets to messaging applications. According to Europol (2021), Dark Web users are increasingly using Wickr and Telegram as communication channels or as sales platforms to bypass market fees.

Studies usually distinguish between social media, social networks and (instant) messaging applications, while many of them use a slightly different definition for each term. According to Merriam-Webster, social media can be defined as “forms of electronic communication (such as websites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (such as videos)” (*Definition of SOCIAL MEDIA*, n.d.). Examples of social media used in academic literature are Facebook, Instagram, Snapchat, WeChat, Whatsapp, Telegram, Wickr, and Reddit, amongst others (Buntain & Golbeck, 2014; Church & de Oliveira, 2013; Dargahi Nobari, Reshadatmand, & Neshati, 2017; Moyle et al., 2019; Obar & Wildman, 2015; Park, Kee, & Valenzuela, 2009). Within the overarching term social media, a distinction can be made between social networks and messaging applications. Generally, a social network is a service that allows users to create a (semi-)public profile that can be used to connect and communicate with other users on the network, usually allowing for a specific type of communication (Adedoyin-Olowe, Gaber, & Stahl, 2013; Dargahi Nobari et al., 2021; Moyle et al., 2019). Instant messaging applications allow for chat-based communication between individuals or groups, usually through text and the sharing of media, such as pictures, videos or documents. Whatsapp, Telegram, Signal, and Wickr are well-known examples of instant messaging applications.

For example, Moyle et al. (2019) shows that social media and messaging applications appear to have become a convenient method of connecting buyers and sellers of drugs. They find that social networking spaces are used to advertise drugs, whereas (encrypted) messaging platforms allow buyers to communicate with known sellers and arrange transactions. The main advantage of buying drugs through social media apps, they found, was that social media apps were easy to use when compared to alternatives (e.g. Dark Web Markets or local street dealers). Next to that, social media apps made it easier to find local dealers (as some make use of location features) and make it easier to check the quality of the product through visual proof, which gives a feeling of trust. Adding to that feeling of trust is the promise of encryption or security by most social media platforms, albeit some end-users are overly trusting. At the same time, trust is one of

the disadvantages, as the research shows that users often have a misleading sense of security. This is founded by the false belief that all social media apps are end-to-end encrypted, that law enforcement agencies are not interested in that user specifically, or that law enforcement cannot legally access the data of the app. Another disadvantage - especially compared to Dark Web Markets - is that social media apps have no vendor rating system, making it difficult for the end-user to determine the legitimacy of the vendor and the product.

E. R. Leukfeldt and Roks (2021) shows street offenders in the Netherlands have taken the new and additional opportunity that social media brings. Although they state offline contacts keep playing an important role in street-crime, traditional street crimes are diversifying through the use of technology. This can, for example, be seen in the recruitment of money mules, who are recruited both offline and online. In other words, not only cybercriminals are moving towards easier, more accessible platforms, also traditional criminals are using social media more and more. Therefore, cybercriminals in this research can have all kinds of backgrounds ranging from traditional cybercriminals, street-level criminals, or members of organised crime groups.

### 2.1.2.3 Local embeddedness of cybercrime

The Internet is a global phenomenon. However, it would be too quick to assume that everything happening on the Internet is happening on a global scale. Local embeddedness is a term used to describe a phenomenon being bound to, limited to, or dependent on a location. Although little research has been done into the local embeddedness of cybercrime, there are some studies that have found evidence that cybercrime is bound to a location more than one would think. For example, E. R. Leukfeldt et al. (2019) found that in the Netherlands, networks dealing with cybercrimes are characterized by core members from the Netherlands that get to know each other offline. Next to that, the core members also recruit others within their offline social circle. Furthermore, they found that the Dutch online sellers of drugs generally limit themselves to selling in the Netherlands and European countries that are within "driving distance" from the Netherlands, such as Belgium and Germany. Supporting these findings, Bakken and Demant (2019) found that social media markets, for example for the sale of drugs and other illegal goods, are often locally bound. They found that these markets often cater to the same region, city or even neighbourhood, which often removes the need for postal shipping or decreases delivery time. In another study by E. R. Leukfeldt et al. (2020), it has been shown that phishing attacks and malware attacks on payment transactions are locally embedded. Money mules act as a middleman in these attacks; providing the account to which money is transferred before it is laundered. When victims of phishing or malware attacks transfer money to a money mule whose account is not located in the same country as the victim's, the bank notes the transaction as unusual and might block the transaction altogether. Local embeddedness of money mules is therefore crucial for the crime to stay undetected. Moreover, in their research into phishing kits on the Dutch market, Bijmans et al. (2021) concluded that the manual included in the phishing kit was often also written in Dutch. This leads us to assume that these phishing

kits were targeted to be used by Dutch cybercriminals. Finally, research by [Han, Kheir, and Balzarotti \(2016\)](#) indicates that phishing kits are often targeted at a specific group of victims and that many phishing kits received the majority of their victims from a specific country. It seems only logical to assume that phishing kits, and maybe even more kinds of financial cybercrime, are locally embedded.

[Van Buskirk, Naicker, Roxburgh, Bruno, and Burns \(2016\)](#) speculates that Dutch consumers of illicit substances would have seemingly less motivation to source drugs from cryptomarkets, as there was a relatively large degree of domestic production of these substances. This implies there are other, possibly easier ways to obtain the substances as a local consumer. They also argue that sellers would have more motivation to export to foreign consumers for the same reason. This claim is backed by [Christin \(2013\)](#), who found that the Netherlands was a popular country of origin for vendors on the cryptomarket Silk Road. Although law enforcement have taken down Silk Road in 2013, several other cryptomarkets have appeared since then. For example, [Van Buskirk et al. \(2016\)](#) have found that the Netherlands was still a popular country of origin on the cryptomarket Agora, both by the number of listings and the number of listings per seller. [Décary-Hétu and Giommoni \(2017\)](#) have shown that the result of a ‘regular’ cryptomarket takedown is that users migrate to other markets and carry on with their business. Therefore, we assume the Netherlands might still be a bigger player in the trade of illicit substances.

## 2.2 Social media

Section 2.1.2.2 introduces the shift of cybercrime to social media and mentions the study by [Moyle et al. \(2019\)](#) to illustrate why this is happening. The reasoning is simple: social media, especially apps, are convenient, accessible, easy to use, and often free. Diving deeper into the world of cybercrime on social media helps us understand what is happening and allows us to develop methods to counter it. This section highlights some recent studies about social media users, relations in online social networks, and cybercrime on social media.

When we want to know more about the users on social media, we should start by profiling the users. Maybe the most simple form of profiling of an individual can be done by looking at demographics, such as gender, age, income, and education), as is done in ([Thach & Olsen, 2015](#)), for example. However, this information may not always be available on social networks. In such a case, one could build a profile of the properties that can be measured. For example in [Buntain and Golbeck \(2014\)](#), the authors identify users with an answer-role within their sub-community on Reddit by looking at their activity. Another measurable property can be the communities a user is a member of; if we can determine a user is a member of several sub-communities related to *gaming*, we can deduce that it is likely this is something the user is interested in.

In order to create a solid image of a user using the subcommunities they are a member of, it is vital to create a profile of the sub-community itself. A logical place would be

to start with the topic of the sub-community, followed by measurable properties, such as the time of activity and the type of activity. Research by Qiu et al. (2016) found that groups on the instant messaging app WeChat could be divided into short-term groups and long-term groups. They found that in a short-term group, the members were added in a short time and they were often added by one user. In a long-term group, members were invited at multiple moments and they could be invited by any user that was already a group member. They also found that short-term groups were often used for a one-time event, and were quiet when whatever the group was created for was arranged, while long-term groups implied a strong(er) relation between the group members, such as friends or family.

Another interesting factor of social media research is the groups users are members of. Research has shown that groups on social messaging platforms bring on a stronger sense of community than traditional text messaging and that social messaging platforms have catalyzed the formation of social groups (Church & de Oliveira, 2013). Moreover, this study found that groups on social messaging platforms are more frequently and habitually created than that group-level social engagement takes place in daily life for most users. Besides, social messaging groups have a relatively shorter lifespan - ranging from hours to months - than groups in social networking sites such as Reddit (Buntain & Golbeck, 2014) or Facebook (Park et al., 2009), which can easily exist for up to years.

Qiu et al. (2016) studies the group lifecycles of groups in the social messaging platform WeChat. They find a difference between short-term groups, with a lifespan ranging from hours to a few days, and long-term groups, which can sustain much longer than short-term groups (often longer than 30 days). Interestingly, they found that short-term groups are often event-driven, while long-term groups are often relationship driven. From a triad count (Holland & Leinhardt, 1971), they observed that long-term groups show stronger underlying friendship dynamics, while short-term groups are less likely to develop friendships over time. Furthermore, they observed that the membership invitations of short-term groups differ from long-term groups. Long-term groups usually have more members and have a bigger cascade depth in the invitation tree (meaning that existing members also invite new members to the group). Moreover, it is observed that membership in short-term groups usually happens in a broadcast fashion, where most of the invitations are managed by the root node. Altogether, they concluded that the lifecycle of messaging groups on WeChat is largely dependent on the social role of the group and the function of the group in the users' daily social experiences and specific purposes.

### 2.2.1 Social relations

Social media platforms often not only contain user profiles, but they are also specially created for user interaction. The term online social networks (OSNs) often points to online platforms used for social interaction. Platforms like these, for example, well-known social media platforms like Facebook, Twitter, LinkedIn, Reddit, etc., are used by billions of users worldwide. It is therefore not surprising that academic literature has taken an interest in the topic. For example, relations between users (Holt, Strumsky,

Smirnova, & others, 2012; Wilson, Boe, Sala, Puttaswamy, & Zhao, 2009), communities on the platforms (Katsikeas et al., 2021; Mislove, Marcon, Gummadi, Druschel, & Bhattacherjee, 2007), the spread of information (Schlette, van Prooijen, Blokland, & Thijs, 2022; Yang et al., 2021) and methods for identifying influential users (Al-Garadi et al., 2018) have been studied.

Relations between people can often be displayed as a social network - which is not to be confused with a social network in the sense of a social media platform such as Reddit. However, both can be displayed in the same manner: in a graph using nodes as people or users and using edges to indicate relations between two nodes. For example in Katsikeas et al. (2021), a citation graph was generated in which all authors were linked to each other according to citations. The authors were represented by nodes and the undirected edges between nodes indicate that an author has cited another author at least once. In this study, the size of each node is related to the number of citations an author has. An abstraction like this allows one to display the relations and define network properties of nodes, such as popularity or influence, by measuring the in-degree, out-degree, and centrality of a node. (Bonacich, 1972; Yang et al., 2021). Al-Garadi et al. (2018), for example, discusses several network properties of nodes in a social graph in relation to a node's influence.

Katsikeas et al. (2021) also uses their author-citation graph to search for communities. They use the Louvain algorithm to find communities of authors publishing about different topics within the cyber security field. They then apply the algorithm a second time on each found community to look for subcommunities. As a result, they were able to detect 12 research communities and for each they were able to discuss their evolution, the most cited articles and the subcommunities. In other words, their usage of communities and subcommunities provided them with valuable insights in a very large social network.

Another interesting example of capturing social relations of online platforms is by Motoyama et al. (2011). In their work, they research relations between users of underground forums, which they argue do not encode pre-existing social relationships, unlike traditional OSNs. Given the forums they study, they define three types of relationships between users: buddy links between two users that have sent and accepted a friend request, private messaging links between two users that have exchanged private messages, and thread relationships between two users that post in the same sub-forum thread.

Finally, we introduce Holt et al. (2012) as an example; in this work, relations between users in different Russian hacker networks are displayed in a graph structure by adding a node for each user and adding an edge between two nodes when the two users they represent are members of the same hacker forum. The study assumes there exists a many-to-many relation between members of the same forum and is, therefore, a great example of a method that could be applied when the researcher has no knowledge or evidence of underlying social relations (in this case because the medium, forums, does not allow it). Additionally, they coloured each node based on the smallest group the user was a member of. A benefit of this method is that it works very well for visualizing the different groups and visualizing users that are part of multiple groups. Another

added benefit is that it does not depend on a platform's built-in social relations, such as Facebook's friend relations or LinkedIn's connection relations.

Many studies look into finding important users in an OSN. In a social network displayed as a graph, where users are denoted by nodes and relations between users by edges, the importance of a user can be calculated using metrics from graph theory (Al-Garadi et al., 2018; Mislove et al., 2007; Yang et al., 2021). However, Motoyama et al. (2011) argues that while the underground forums that are the topic of their study are a form of OSNs, the communication doesn't simply encode pre-existing social relationships like in traditional OSNs.

### 2.2.2 Cybercrime & social media

As mentioned briefly in Section 2.1.2.2, social media is not exempt from being used for cybercrime. On the contrary, the use of social media for cybercrime is only growing. However, to our knowledge there does not exist a lot of academic literature dedicated to this topic at the moment of writing. The following research papers are found that relate to this subject.

In their work, Moyle et al. (2019) conducted over 400 interviews to gather information from drug users about their use of social media apps for buying and selling drugs. As mentioned before, they found that social media was widely used for this purpose, especially since social media is convenient and easy to use. Bijmans et al. (2021), on the other hand, introduces a method for capturing and following the lifecycle of phishing kits found and sold on Telegram. They find 70 different Dutch phishing kits and identify 10 kit families. Blankers et al. (2021) also studies cybercrime on Telegram but focuses instead on the trade of psychoactive substances in the Netherlands during the COVID-19 pandemic. They find that primarily sellers are active and that the sale of psychoactive substances may have been affected by the pandemic. Finally, Al-Rawi (2019) studies the promotion of the drug fentanyl on Instagram and Twitter. They found that drugs dealers often include other types of drugs in their post next to fentanyl. Locations are also often included in promotional posts. Another finding is that drug dealers make use of platforms like Instagram and Twitter to promote their products, the sellers often refer to contact details of other, more private platforms such as Gmail, Whatsapp, Wickr, Kik, and Telegram, or regular phone numbers.

## 2.3 Telegram

Telegram is a cloud-based instant messaging service that is used to send text messages, stickers, files, and media, such as photos and videos. Dargahi Nobari et al. (2021) states that Telegram has formed as a new kind of social media platform that stands between instant messaging applications and social networking applications. In terms of security, it promises end-to-end encryption and even self-destructive messages. At the time of

writing, Telegram has over 500 million monthly active users and is one of the ten most downloaded apps worldwide (*Telegram FAQ*, n.d.).

Telegram allows for one-to-one messaging between two users, one-to-many messaging in channels, and many-to-many messaging in groups (Figure 2.1). In other words, it offers two methods of broadcasting messages: channels and groups (*Channels, supergroups, gigagroups and basic groups*, n.d.). The function of each is described below:

- A **group** is a space where all members of the group can send and receive messages in the group. The members of a group are usually interested in the same topic; the topic of the group. Users can become a group member either by joining the group through an invite link or by being invited by another member of the group.
- A **channel** is a space where public messages are broadcasted to the channel's subscribers. The messages can only be published by the channel's administrator(s), of which there usually are just a few. Channels can have an unlimited number of subscribers and users can subscribe to a channel either through a channel username or through its join link.

Telegram's API distinguishes three types of groups: basic groups, supergroups and gigagroups (*Channels, supergroups, gigagroups and basic groups*, n.d.). Basic groups can have up to 200 members and can be migrated to a supergroup, which can have up to 200,000 members. Gigagroups are something between a group and a channel, as it is a supergroup that can have more than 200,000 members, but where only admins can send messages.

Different from other social media networks, Telegram does not use friend relations. It does, however, make use of a private contact list. Nonetheless, according to Dargahi Nobari et al. (2017); *Reinventing Group Chats: Replies, Mentions, Hashtags and More* (2015) there are two other sorts of relationships to associate users, groups, and channels:

- Messages can be **forwarded** by a user or channel to a different user, channel or group. The content of the message stays the same, but a forwarded message includes the name and link to the original creator of the message. It is also possible that a comment is added to the forwarded message. Telegram's message forwarding can be compared to forwarding an email or retweeting a tweet on Twitter.
- Users, channels and groups can be **mentioned** in a message. This can be used to either get their attention or to provide a link to that user, group or channel. A mention is often formatted the following way: *@groupname* or *(https://)t.me/groupname*. A mention is a clickable link to the mentioned Telegram group or user.

Dargahi Nobari et al. (2017) presents one of the first attempts in academic literature to analyse structural and topical aspects of messages in Telegram. Next to classifying advertisement and spam messages, they extract a mention graph of mentions between Telegram channels. In the mention graph, channels are represented by a node and an

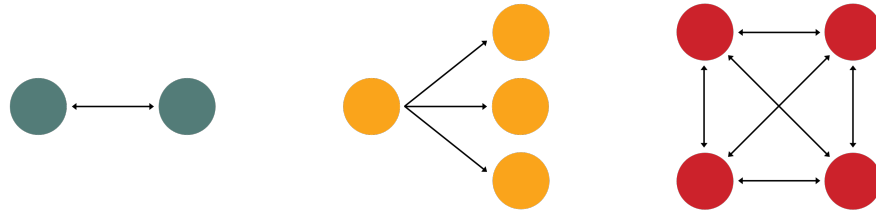


FIGURE 2.1: Messaging relation in chats, channels and groups on Telegram

edge is drawn between two nodes if a message in one channel mentions another channel. The three types of mentions they define in their data are self-mentions, spam mentions, and ham mentions. They define self-mentions as a channel (administrator) including a mention of the channel in a message send in said channel, so it is included in the message when it is forwarded. A spam mention, they say, is a message that mentions one or more channels without adding any other content to the message. A ham mention is defined as a message that includes content of the current channel but mentions another channel.

In the same work, [Dargahi Nobari et al. \(2017\)](#) defines a Telegram channel's popularity by the number of followers it has. They do not find a correlation between the number of followers of a channel and its degree in the mention graph, in spite of the web graph. They even found that many popular Telegram channels are standalone nodes or nodes with a low degree, and many of the nodes with a high degree are not popular channels. Next to that, they find that the traditional PageRank method ([Brin & Page, 1998](#)) cannot be used to detect high-quality channels in Telegram.

Furthermore, relations between messages within groups can be defined by the following elements:

- A user can **reply** to a specific message in the chat, group, or channel they're in. When replying, a preview of the message the user is replying to is shown above the new message.



- **Hashtags** are used to highlight a term in a message. Hashtags in messages are clickable and can be used to quickly search a conversation, group or channel for messages with the same hashtag.

### 2.3.1 Users and bots

To become a Telegram user, one needs to create an account using a phone number. If a user's contacts are also Telegram users, one can use Telegram to message them directly. It's possible to add a public username to an account, such that people can find the user by searching Telegram for the username. A Telegram user can also find another user when they are both members of the same group (*Telegram FAQ*, n.d.). A user's screen name, username, and profile picture are always publicly available. However, a username and profile picture are not mandatory. All other user information can be made public or private by the user.

We distinguish "regular" users and bot users. We define a regular user as a person that manually operates an account. Bot users, on the other hand, are like small programs that run on Telegram. (*Telegram FAQ*, n.d.). Users can interact with bots by sending them messages, requests and commands. A bot can send customized notifications, create custom tools, and accept payments from Telegram users, amongst others. Channels and groups have become popular platforms for mobile advertisement. Generally, repeated messages, such as advertisements, on Telegram are enabled by bots. (Dargahi Nobari et al., 2017)

An example of such a bot is Miss Rose (*Miss Rose*, n.d.). This is a popular bot that can be used for a plethora of tasks in Telegram groups and channels. For example, Miss Rose can help with admin tasks by banning users or she can assist with tasks such as language translation, weather updates, and providing quick access to news articles or other relevant information. Users can also create their own bots using Telegram's publicly accessible API.

### 2.3.2 Cybercrime on Telegram

While Telegram is primarily an instant messaging service, it has several added social networking features that make it an interesting platform for cybercrime. For example, part of the reason for its popularity is that Telegram messages are end-to-end encrypted and can be set to self-destruct (Blankers et al., 2021; Moyle et al., 2019), giving users a sense of privacy. Next to that, users can choose to make their Telegram profile public - usually just their username - while keeping their phone number hidden. This feature allows other Telegram users to search for usernames and message other users directly, without the necessity of having exchanged phone numbers. This not only gives the benefit of relative anonymity but also allows users to message others outside their direct contacts. Furthermore, Telegram can be downloaded through an app store and requires almost no setting up. It is as easy to use as any other chat application, which is a huge benefit for anyone without the technical expertise that is required to use the Dark Web.

Telegram has become increasingly popular as a platform for cybercrime. For example, [European Monitoring Centre for Drugs and Drug Addiction and Europol \(2020\)](#) reports that social media and secure encrypted communication applications appear to have played a more prominent role in the buying of drugs by end-users during the COVID-19 pandemic. Recent work by [Bijmans et al. \(2021\)](#) has shown that phishing kits are largely available in multiple Telegram channels. [Blankers et al. \(2021\)](#) has found that Dutch Telegram groups are mostly used as a seller's market for psychoactive substances.

In (Dutch) Telegram groups used for cybercrime, several crime markets can be identified, amongst which: 1) drugs([Al-Rawi, 2019](#); [Blankers et al., 2021](#); [Moyle et al., 2019](#)), 2) fraud ([E. R. Leukfeldt et al., 2019](#)), 3) cybercrime([Bijmans et al., 2021](#)), 4) money laundering([E. R. Leukfeldt et al., 2019](#)), 5) weapons([Bakken & Demant, 2019](#)), 6) fireworks, 7) fake documents or money, and 8) stolen goods. While there is some overlap between the crime markets, each supplies a unique range of products and services to those who are interested.

Looking at cybercrime on Telegram in the Netherlands, it makes sense to split the category drugs into hard drugs and soft drugs, as these are two separate categories in the Dutch Opioid Law (*Opiumwet*, n.d.). Examples of soft drugs are cannabis-based products and examples of hard drugs are cocaine, heroin, ecstasy, etc.

Messages from Telegram groups and channels can be downloaded through the Telegram desktop client, as in ([Blankers et al., 2021](#)), or scraped through the Telegram API (*Telethon API*, n.d.). [Blankers et al. \(2021\)](#) uses the search function of the Telegram desktop client to input a search term to find Telegram groups that match the criteria of their research and [Bijmans et al. \(2021\)](#) continues the search for new Telegram groups and channels by looking through recently found groups and channels for mentions of new groups and channels (i.e. a snowball approach). ([Dargahi Nobari et al., 2017](#)) developed a crawler to gather the messages of Telegram channels and groups and looks through the messages in these channels and groups to extract mentions, join links and forwards to find new Telegram channels in groups.

## 2.4 Conclusion

This chapter presents the context in which this study takes place. Looking at recent academic publications in the criminology field we notice the various definitions of cybercrime and conclude that the current work focusses on cyber-enabled crime. Previous studies also highlight the importance of real-life relationships in the origin and growth of cybercriminal networks, along with the presence of core members and assigned roles within these networks. Next to that, these studies present that online meeting places are important for cybercriminals, fulfilling market, social, and learning functions. Traditionally, meeting places took the shape of underground forums, characterised by a small number of very active users, and dark web markets, more professional looking places that resemble well-known ecommerce websites. However, recent trends indicate cybercriminals are shifting towards social media platforms as meeting and marketplaces.

While social media platforms offer advantages, such as accessibility, ease of use, and the fact that it is often free, users often overestimate the level of privacy and protection provided by the platforms. Nonetheless, local embeddedness remains prevalent in cybercrime, with, for example, recruitment of money mules and the sale of illicit goods being primarily conducted within local contexts.

Social media platforms contain vast amounts of user information, allowing for the creation of user profiles and the study of user behaviour. However, the structure of social media platforms allow researchers to also study the relationships between users and use these to identify influential users through network structures and graph properties. In similar fashion, related studies have examined relations between users in the context of underground forums. Studies focusing on the use of social media for cybercrime highlight its widespread use for illicit substance transactions, with, for example, public platforms like Instagram and Twitter being utilized for promotional purposes, linking to more private social media platforms to conduct the actual sales and contact.

This work focusses on the social media platform Telegram. Telegram is characterized by its combination of instant messaging features and the ability to create public groups and channels. The option for relative anonymity and the availability of customizable bots make Telegram an intriguing platform for many cybercriminals. Cybercrime on Telegram was recently mentioned as a growing problem. Notably, psychoactive substances as well as phishing kits have been found to be sold over Telegram in the Netherlands.

In conclusion, the comprehensive review of background and related work has laid the foundation for understanding the research context and identifying gaps and trends in the field of cybercrime. The subsequent chapters of this thesis will build upon this knowledge and address the research objectives in light of the identified limitations and opportunities presented by the current state of research and the unique characteristics of Telegram as a social media platform.

# Chapter 3

## Approach

This chapter aims to give an overview of the general approach used in this study, after which the approach for the data collection process is described. As mentioned in Section 1.1, the research is divided into three phases. The first phase studies Telegram groups, the second phase studies its users, and the third phase combines the information from the previous phases to determine what could make a Telegram group or user influential in the context of Telegram being used as a cybercrime platform. These phases can be found in Chapter 4, Chapter 5, and Chapter 6 respectively. The approach for data collection and preprocessing and the resulting dataset are described in the current chapter. An overview of the general approach of this study is visualized in Figure 3.1.

For each of the phases, code is written in Python 3.7 (Van Rossum & Drake, 2009) and executed in a Jupyter Notebook (*Project Jupyter*, n.d.). Next to that, most data manipulation and analysis is done using Pandas (McKinney & Others, 2010) and Numpy (Harris et al., 2020). Most visualizations are made using Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017).

### 3.1 Data collection

Our goal is to study the public side of Telegram to gain more insight into the use of Telegram as a cybercrime platform. Therefore, it makes sense to try and obtain as much information as we can about the public side of Telegram. We are specifically interested in message content and information about users, as these are the main sources of information in a messaging platform such as Telegram. The decision was made to base this study purely on publicly available data to make it easy to recreate or continue this study.

Public messages on Telegram can be found in public groups and channels. While both public groups and channels can be accessed through the Telegram desktop and mobile app, the Telegram API only allows admins to scrape a channel's messages. To scrape the messages and users from a public Telegram group, on the other hand, one does not even need to join the group, let alone be an admin (see Section 2.3). Since it is not

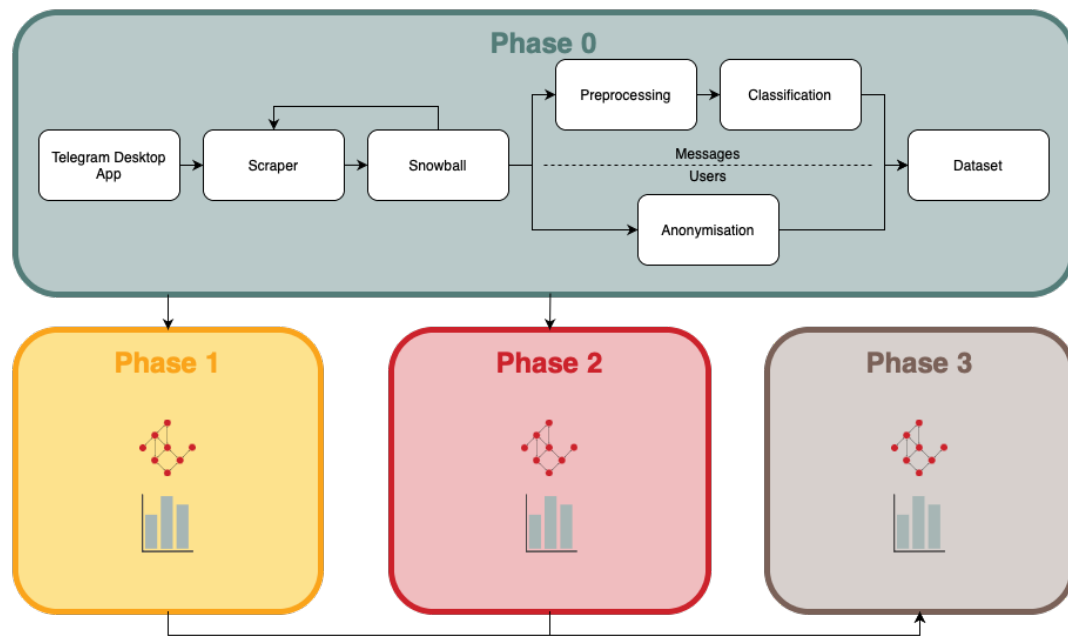


FIGURE 3.1: Visualization of the approach of this study

feasible to become an admin of existing Telegram channels without drawing suspicion to ourselves and since it's not feasible or desirable to start our own Telegram channels, we decided to leave Telegram channels out of our research altogether. Focusing our research on Telegram groups also has the benefit that it allows us to approach the data collection process group-wise; i.e. it allows us to collect data from a whole Telegram group, instead of needing to collect individual messages. It also allows us to use Telegram's natural group structure later in this research.

Collecting the data and transforming it into a usable dataset takes several steps. An overview of the steps can be found in Figure 3.1. First, we obtain all messages and members of Telegram groups using a scraper. A snowball sampling approach is applied to find new Telegram groups to scrape in the messages of the group that was just scraped. This process repeats until the snowball is finished. Then, the collected data goes through several preprocessing steps: we make sure the data contains no messages after a chosen date, we remove groups that don't fit the topic, we remove groups with missing or faulty data, and finally, we anonymize the user data. Finally, a classifier created for a similar project is used to label the messages in the dataset.

We created a Telegram account with a throwaway phone number to use the Telegram API and the Telegram desktop app. All code is written in Python 3.7 (Van Rossum & Drake, 2009) and executed in Jupyter Notebooks (Project Jupyter, n.d.). The Telethon (*Telethon API*, n.d.) library is used for the connection to the Telegram API, and all data analysis steps use a combination of the Numpy (Harris et al., 2020) and Pandas (McKinney & Others, 2010) libraries. Furthermore, Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017) are used to create most visualizations.

### 3.1.1 Ethics

The collection and use of data for this thesis are guided by ethical considerations. We recognise our ethical responsibility in conducting research that respects the privacy of the individuals involved and realise that a work such as the current one should be careful not to cause any harm.

This work uses publicly available data from Telegram, which is well-documented in the platform's privacy policy (*Privacy Policy*, 2019). By signing up for Telegram, a user agrees with the privacy policy. The privacy policy states, for example, that "*everything you post in public will be accessible to everyone.*" (*Privacy Policy*, 2019). Therefore, we argue that users can be expected to be reasonably aware that their posting in public groups is publicly accessible. Next to that, the privacy policy also mentions what user information is stored. The user information that is always publicly available are a screen name, a username (not mandatory), and a profile picture. All other user information can be made public or private by the user themselves. This study only uses the user ID and the username, which we argue are not personal information as they do not contain any identifiable information about the users. Furthermore, we took steps to protect the privacy of the users by using pseudonyms when addressing users and by not reporting any individual-level data that could lead back to a user.

To ensure that our research does not create further harm, we do not name any Telegram groups or vendors in our report to avoid creating publicity for them. Moreover, we also try to only report aggregate data where possible and always use pseudonymous data.

Furthermore, we sought approval from the TNO ethics board, which was granted. The data is stored on TNO servers.

### 3.1.2 Scraper

We use a scraper to collect the messages and users from public Telegram groups. As mentioned above, the scraper is written in Python and executed in a Jupyter Notebook. It uses the Telethon (*Telethon API*, n.d.) library to connect to the Telegram API and uses Telethon's predefined functions to scrape all messages and members of a Telegram group given the group's name. The obtained messages and users of each group are stored in JSON files.

The function that scrapes a group's messages scrapes text messages and media messages, but it does not include the media itself in case of a media message (i.e. it does not include photos or videos). When media is sent with an accompanying text, only the text and the media's metadata are stored. The scraper also stores a message's metadata, such as the date and time the message was sent and by whom it was sent. An overview of all stored metadata can be found in Table A.1 in Appendix A.

Between the messages, a Telegram group contains what Telegram calls message services. Message services keep track of things happening in the group, such as the group's creation or users joining and leaving. The scraper includes these message services. We leave

these message services in our dataset because we believe they may contain interesting information for our research.

Moreover, we use Telethon’s predefined function for scraping users that are a member of a given group. This allows us to get an overview of the passive members in a group in addition to the active members who can be found as senders of the messages in the dataset. As for the messages, the scraped information is stored in a JSON file. For each user, we include the information found in Table A.2 in Appendix A. It must be noted, however, that only the user ID and the ‘is.bot’ fields are included for every user. All other fields can be filled in and made public by users by choice. Therefore, not all users in the resulting dataset have visible usernames, phone numbers, and names.

### 3.1.3 Snowball

A so-called snowballing sampling method is used to find new Telegram groups to scrape. A snowball approach is commonly used when parts of the desired data are hidden. This can be done in social settings using human research subjects, described by Paul Vogt (1999) as follows “A technique for finding research subjects. One subject gives the researcher the name of another subject, who in turn provides the name of a third, and so on.”(p. 300). However, we use the approach to find hard to reach digital data, as is done by Bijmans et al. (2021). They followed shared links in Telegram chats to discover related Telegram channels to extend their dataset. This method allows one to eventually reach saturation in data collection.(Atkinson & Flint, 2001) We apply the snowball approach as follows: after a Telegram group is scraped, our code looks through all messages of the group that was just scraped and finds all mentions of other Telegram groups. If our dataset does not include these groups yet, we add them to the list of new groups that are scraped in the next iteration of the scraper. This process continues until the snowball is finished (i.e. does not find any new Telegram groups) or can be stopped manually.

Our code takes into account the following formats of mentions of Telegram groups: @groupname or (https://)t.me/groupname. Usually, a mention in this format is a clickable link to the mentioned Telegram group. These formats were chosen after a manual inspection in the Telegram Desktop app showed that these were the prevalent forms of mentions of other Telegram groups within a group.

### 3.1.4 Starting point

The snowball approach mentioned above needs the name(s) of one or more Telegram groups to start the snowball. We use the search function in the Telegram desktop app to find groups we can use as a starting point. For example, if we search for the Dutch word *fraude* (English: fraud) in the Telegram desktop app, we find several Telegram groups containing messages advertising fake money or fake documents, or selling methods for swindling others. The found groups are then scraped and the snowball method is applied.

At the start of our research, we received a list of names of Dutch Telegram groups that were used for cybercrime in 2020 from TNO experts. These groups serve two purposes: first, they are used as starting points for our snowball approach, and second, we manually inspect the content of these groups to find search terms we can use in the Telegram desktop app to look for other Telegram groups. Potential search terms were added to the list if they were fraud or crime related (e.g. words like *fraud*, *fgame*, *weapons*, etc.), related to trade, trading, or business (e.g. words like the Dutch words *handel* or *marktplaats*), or related to a location in the Netherlands (i.e. regions or cities, such as *Amsterdam* or *Brabant*). The complete list of search terms we assembled can be found in Table A.3 in Appendix A.

The search terms are used in the search function in the Telegram desktop app. This app returns public groups, channels and users that are found with the searched term. However, it must be noted that the Telegram desktop app returns a limited number of search results. Therefore, we cannot claim this method returns a complete overview of groups. Given that we use this manual search method in combination with the snowball method, however, we believe this limitation to be of minimal effect.

## 3.2 Preprocessing

The scraping and snowball process can take quite some time. Therefore, a resulting dataset may contain Telegram groups of which the moment of scraping is quite far apart. This may cause the messages of some groups to be influenced by an event that had not yet taken place when the messages of the first Telegram groups were scraped. By introducing a cut-off date that is the same for all scraped groups we believe we can avoid potential skewing caused by the groups being scraped at different times. We decide to allow our dataset to contain messages sent up and until the day before we started the scraping and snowball process.

The second preprocessing step is to take a closer look at which scraped Telegram groups fit the topic of the research and remove those that do not fit our research. This research has two criteria for its dataset; a Telegram group needs to contain evidence of cybercrime and it needs to be Dutch or focus on the Netherlands. Not only does this allow us to focus this study, it is also in line with the claim that cybercrime is locally embedded, presented in Chapter 2. To enforce these criteria, we wrote some Python code in a Jupyter Notebook to help us do the following:

1. Randomly select 10 messages from a group to be our sample. For this, the Pandas sample function (*pandas.DataFrame.sample* — *pandas 2.0.2 documentation*, n.d.) is used with a fixed seed to ensure the process can be repeated. If the group contains fewer than 10 messages, it displays all messages.
2. We ignore clear bot messages from our sample, such as "user has been kicked/muted/banned/", as popular admin bots are almost always in English and can skew our results if many of these messages appear in our random sample. To illustrate,



if there are 3 clear bot messages in our sample of 10, this means that the following steps are applied to the 7 remaining messages.

3. We manually check if half of the messages in the sample (at least 5 messages out of 10) are in Dutch, contain Dutch geographic locations (e.g. cities, provinces, regions), or contain elements that reference a Dutch entity (e.g. Dutch phone number). This includes messages that reference a user whose Dutch details are mentioned in another message in our sample. If it does, we assume the group is focused on the Netherlands.
4. We manually check if at least one message in the sample mentions a form of cybercrime (e.g. phishing panels, accounts, hacks, etc.) or cyber-enabled crime (e.g. sale of drugs, stolen goods, documents, cashout, fireworks, etc.). If this is the case, we accept that as evidence that cybercrime occurs in the group.
5. If both criteria are met, the Telegram group is accepted. If either one of the criteria is not met, it is removed from the dataset.

The last step of the preprocessing process is to check our dataset for faulty data. This can be, for example, data that is incomplete or where something went wrong. Faulty data can skew results generated in a later stage of the research, so we believe it wise to avoid this from the start. In this research, manual checking of all collected data for faults is deemed infeasible. However, the continuation of this research serves as an effective means of identifying data that is not in the expected format or shape. For instance, missing entries are likely to lead to errors. It is recommended that anyone attempting to replicate this method should be mindful of erroneous data and implement their own strategies to mitigate such errors.

### 3.2.1 Anonymisation

We do not preprocess the user data, except anonymizing it. The details included in the user data can be found in Table A.2 (Appendix A). While for some studies it may be that all of these details are necessary, this research only benefits from a user's user ID's and username. Usernames can provide an additional perspective in this study, as they might present us with clues about the users. However, no usernames are mentioned in this thesis. We also include the user ID's. As this is a unique identifier for each user, we believe this to be essential to our research, as the user ID is the only thing linking messages and users that sent them. However, no user ID's are mentioned in this work.

## 3.3 Classification

We use a classifier to label the messages in our dataset. Each message gets a label to indicate the message type. The message types we distinguish are described in Table 3.1. If a message is labelled as an advertisement (either an offer or request), it receives a

Message type	Explanation
Chat message	A regular chat message
Advertisement - offer	Presenting illegal goods or services with the goal to sell or promote them
Advertisement - request	Requesting illegal goods or services with the goal to buy them
Bot message	An automatic message sent by a bot, often performing administrative tasks

TABLE 3.1: The message types used throughout this study

Advertisement category	Example
Soft drugs	Soft drugs as classified by the Dutch Opium Act ( <i>Opiumwet</i> , n.d.), e.g. cannabis, cigarettes, alcohol
Hard drugs	Hard drugs as classified by the Dutch Opium Act ( <i>Opiumwet</i> , n.d.), e.g. cocaine, heroin, xtc
Pharmaceuticals	Xanax, anti-depressants, erection pills, sleeping pills
Cashout	Fraud, money laundering, (stolen) credit cards
Cybercrime	Phishing panels, hacking automated systems, digital scams, digital subscriptions (e.g. Netflix, Spotify, etc.)
Stolen goods	Stolen cars, clothes, electronics, scooters, bikes
Licences and personal documents	Fake driving licences, personal identification documents, certificates of good conduct, college degrees
Fireworks	Illegal fireworks
Firearms and explosives	Guns, C4, hand grenades, munition
Weapons	Knives, brass knuckles, tasers
Other	Prostitution, (child) pornography, match fixing, and other advertisements that do not fall under the previous categories

TABLE 3.2: The advertisement categories used throughout this study

label indicating the advertisement category. An overview of all advertisement categories, including examples, can be found in Table 3.2.

The classifier has been developed by a team of four, including myself, as part of a course of the Cyber Security Master’s programme at the University of Twente (D. Lummen, K. Boersma, R. Dijkstra, J. Sustromk, 2021). The approach described in Section 3.3.1 also originates from this work.

The labels produced by the classifier play an important role in this research. They are used in every phase, as they tell us something about the nature and content of the messages while the actual content of the messages becomes redundant. Since our dataset is of significant size, being able to omit the content of the messages speeds up the processes enormously.

The code for the classifier is written in Python 3.7 (Van Rossum & Drake, 2009) and runs in a Jupyter Notebook (*Project Jupyter*, n.d.). Next to that, the SciKitLearn (Pedregosa et al., 2011) library is used for classification, the Pandas (McKinney & Others, 2010) library for data manipulation, and Seaborn (Waskom et al., 2017) and Matplotlib Hunter (2007) for data visualization.

### 3.3.1 Ground truth

The classifier uses labelled data to train its models. The labelled data originates from the dataset used during the project that created the classifier and consists of the messages of

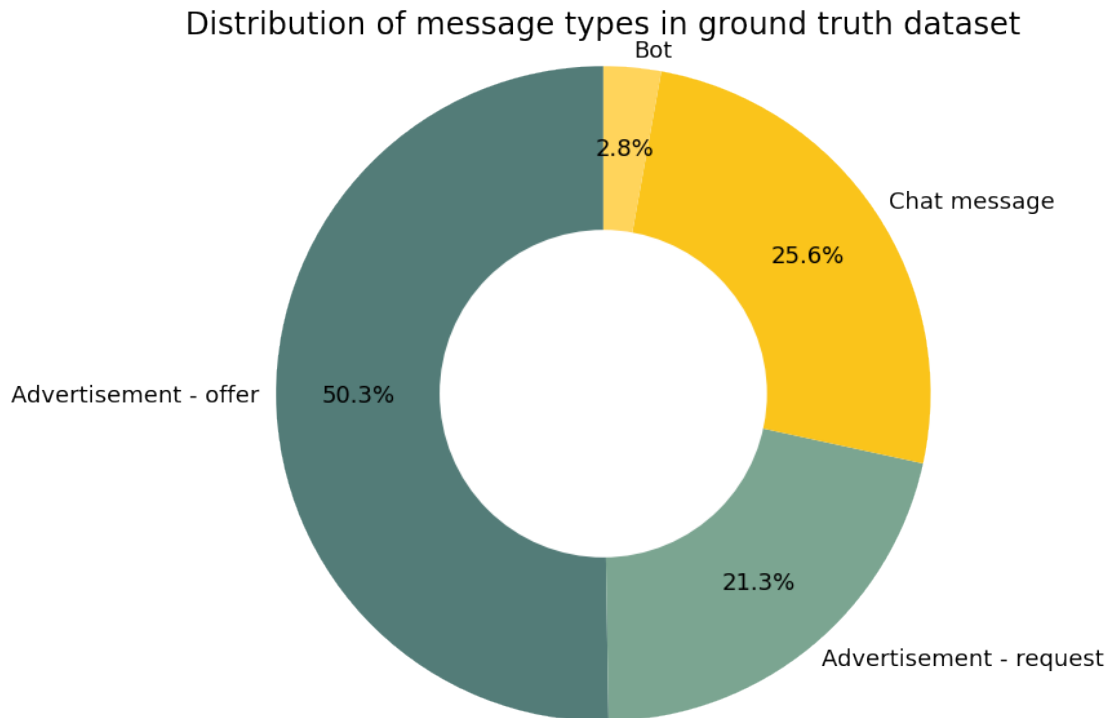


FIGURE 3.2: Distribution of message types in ground truth dataset

48 Dutch Telegram groups showing evidence of cybercrime. Roughly 0.5% of all messages in the dataset were labelled manually, where we aimed to label 0.05% of messages of each Telegram group in the dataset with a minimum of 225 messages (or all messages if the group contained fewer than 225 messages) and a maximum of 900 messages per group to avoid skewed results (some groups were very big). Eventually, over 18,000 messages were labelled. We gave each message a message type, to define if it was either an advertisement request, an advertisement offer, a chat message, or a bot message. If a message was labelled as an advertisement (either request or offer), it received an additional label to indicate the category of the product or service it contained. The category labels were: Cashout, Weapons, Firearms and Explosives, Soft drugs, Hard drugs, Pharmaceuticals, Cybercrime, Licences and Personal Documents, Fireworks, Stolen Goods, or Other.

The labelled data used to train the classifier contains 18,836 labelled messages. As Figure 3.2 shows, about 50% of these messages were manually labelled as advertisements offering a product or service. An additional 21% of messages were labelled as an advertisement requesting a product or service. Figure 3.3 displays the distribution of advertisement categories over the messages labelled as advertisements in the ground truth dataset. We see that the majority of the messages are labelled as cashout. Next to that, there are many messages labelled as either soft drugs or hard drugs. Cybercrime is also a prevailing advertisement category in the ground truth dataset.

### 3.3.2 Classifier training

The classifier takes the following steps:

1. load the training data
2. clean the data
3. extract features to use for classification (tokenizing)
4. train the model
5. evaluate performance
6. label the data

We use Ski-Kit Learn's `CountVectorizer` (`sklearn.feature_extraction.text.CountVectorizer`, n.d.) function to preprocess and tokenize the labelled data. Preprocessing, in this case, consists of replacing capital letters with lowercase letters and removing numbers and symbols. Tokenizing includes tokenizing each unique word and 2-gram. Afterwards, the tf-idf (term frequency times inverse document frequency) is calculated for each token using SkiKitLearn's `TfidfTransformer` function (`sklearn.feature_extraction.text.TfidfTransformer`, n.d.). Term frequency tells us something about the occurrence count of each token independent of document length. Adding inverse document frequency allows us to downscale weights for words that occur in many documents in the corpus and are therefore less informative than those that occur only in a smaller portion of the corpus.

The next step is to train the models that are going to classify the messages in our dataset later on; one model is used to classify the message types and one is to classify the advertisement categories. An Support Vector Machine model using hinge loss and an L2-penalty was used to model the labelled data (`sklearn.linear_model.SGDClassifier`, n.d.). To avoid skewed results, we randomly downsampled the 'Advertisement - Offer' message type in the model to classify the message types, as it appears very frequently.

Distribution of advertisement categories in ground truth dataset

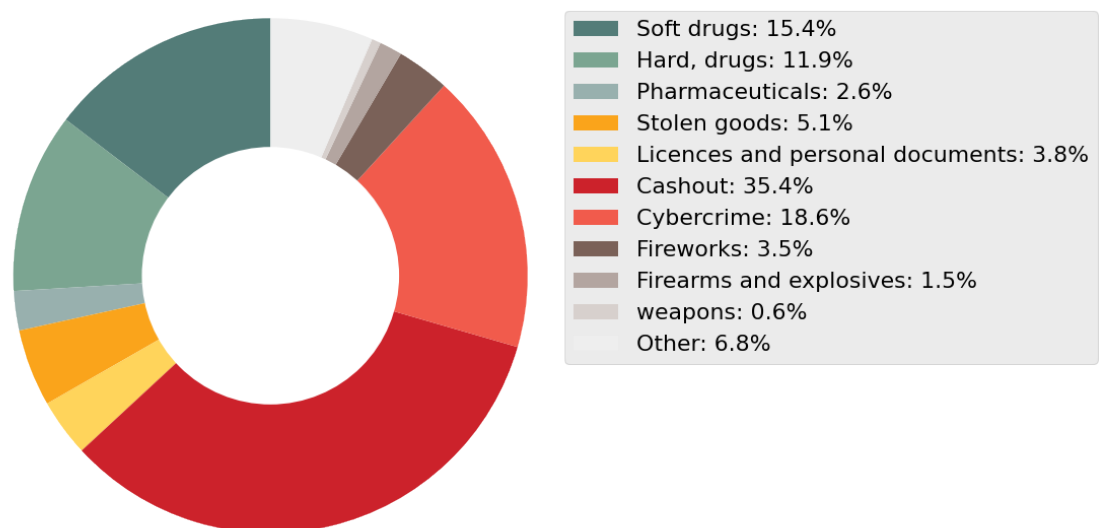


FIGURE 3.3: Distribution of advertisement categories in ground truth dataset

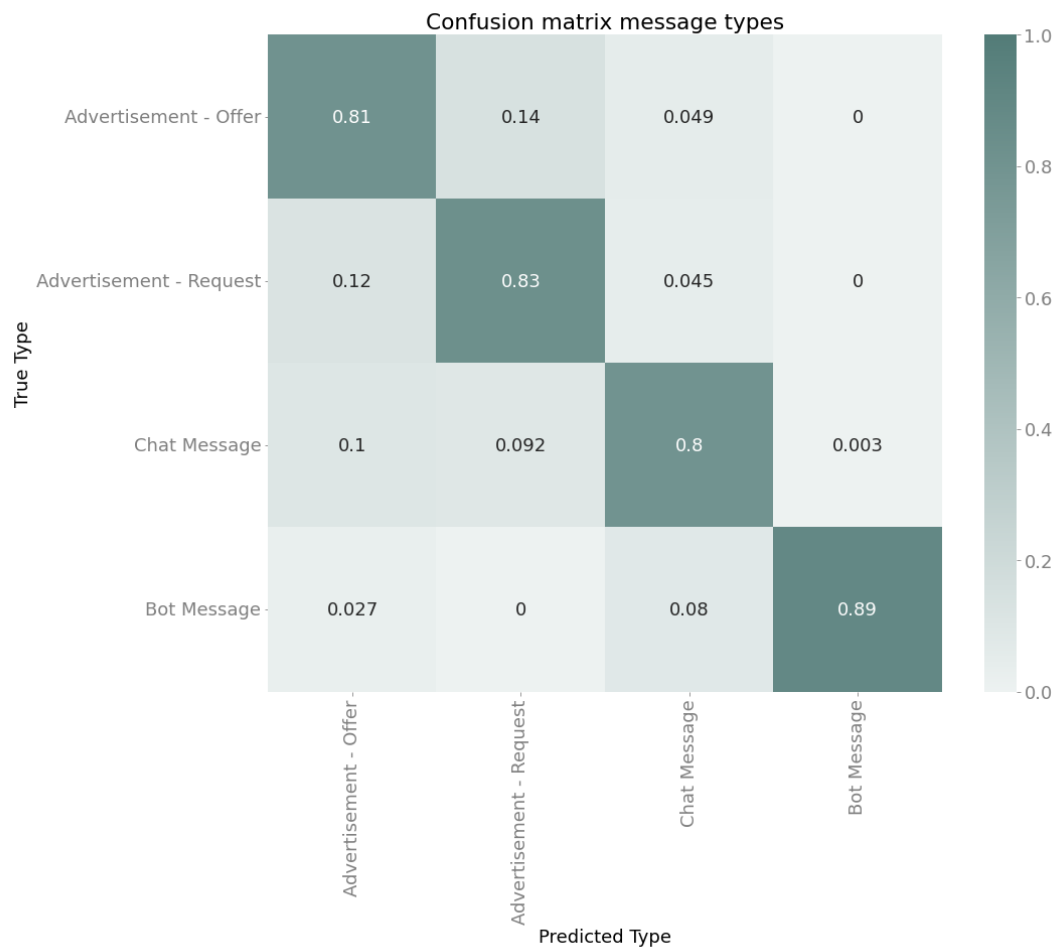


FIGURE 3.4: Confusion matrix of the predicted message type

### 3.3.3 Classification

The models trained in the previous step are used to classify the messages in our dataset. Afterwards, we use the message classification to assign an alias and labels to all Telegram groups in our dataset.

In the rest of this work, we use a group’s alias instead of its name when talking about a specific Telegram group. Before we assign the aliases, we sort the Telegram groups in our dataset alphabetically. While looping over all groups, we determine the prevalent message type in each group. If the prevailing message type of a group is a chat message, we call the group chatgroup-XX, where XX is the number of chat groups starting at 1 (i.e. the first chatgroup gets alias chatgroup-1, the second chatgroup gets alias chatgroup-2, etc.). If the prevailing message type is bot messages, the group gets the alias bots-XX. If the prevailing message type is either advertisement - offer or advertisement - request, we use the prevalent advertisement category as an alias, for example, cashout-XX or hard.drugs-XX, where XX is a count of the groups with that same prefix.

Additionally, we assign a message type and advertisement category label to all Telegram groups in our dataset. To do this, we check what message type and advertisement category is prevalent in a group. If a group has no messages classified as advertisements, we leave the advertisement category label empty.

### 3.3.4 Performance

The evaluation of the models was done using a test sample of 20% of the labelled data. Looking at the model that predicts the message type, we get a precision score and recall score of respectively 85.2% and 83.5%. The confusion matrix of this model can be found in Figure 3.4. The figure shows us that the model sometimes confuses advertisement offers and advertisement requests and it also sometimes predicts a chat message is an advertisement. However, we also see that it rarely happens that an advertisement is classified as a chat message or a bot message.

The model predicting the advertisement category has a precision score of 83.7% and a recall score of 74.1%. The confusion matrix for this model is displayed in Figure 3.5. In this figure, it stands out that the model seems to be performing worse than average when it comes to identifying advertisements for weapons, as we see these advertisements are sometimes classified as firearms and explosives, which makes sense as these categories are closely related, or as cashout. We see the same thing happening between soft drugs, hard drugs, and pharmaceuticals. Confusion between these three categories can also be explained by the categories being closely related and the fact that there are multiple occasions where one advertisement contained a combination of these products. However, it also stands out that cashout has a relatively high number of false predictions. We suspect this may be caused by overfitting, as over a third of the messages in the ground truth dataset fell into the category cashout (see Section 3.3.1).

## 3.4 Results

### 3.4.1 Data collection

In Section 3.1.4, we mentioned using a list of Telegram groups used as a starting point for the scraper and as a source for search terms. This list, assembled by TNO experts in 2020, consists of 48 Telegram groups that were used for cybercrime. A quick search of the groups using the Telegram desktop app told us that 37 of the 48 groups were still publicly available. The other 11 groups were since made private or deleted.

Using the 37 existing Telegram groups, we created a list of 123 search terms. The list contained terms ranging from popular (criminal) slang to names of products or services, and locations. It can be found in Table A.3 in Appendix A. We input each in the search function of the Telegram desktop app and found 220 unique groups to scrape. Applying the snowball method mentioned in Section 3.1.3, another 167 Telegram groups were found.

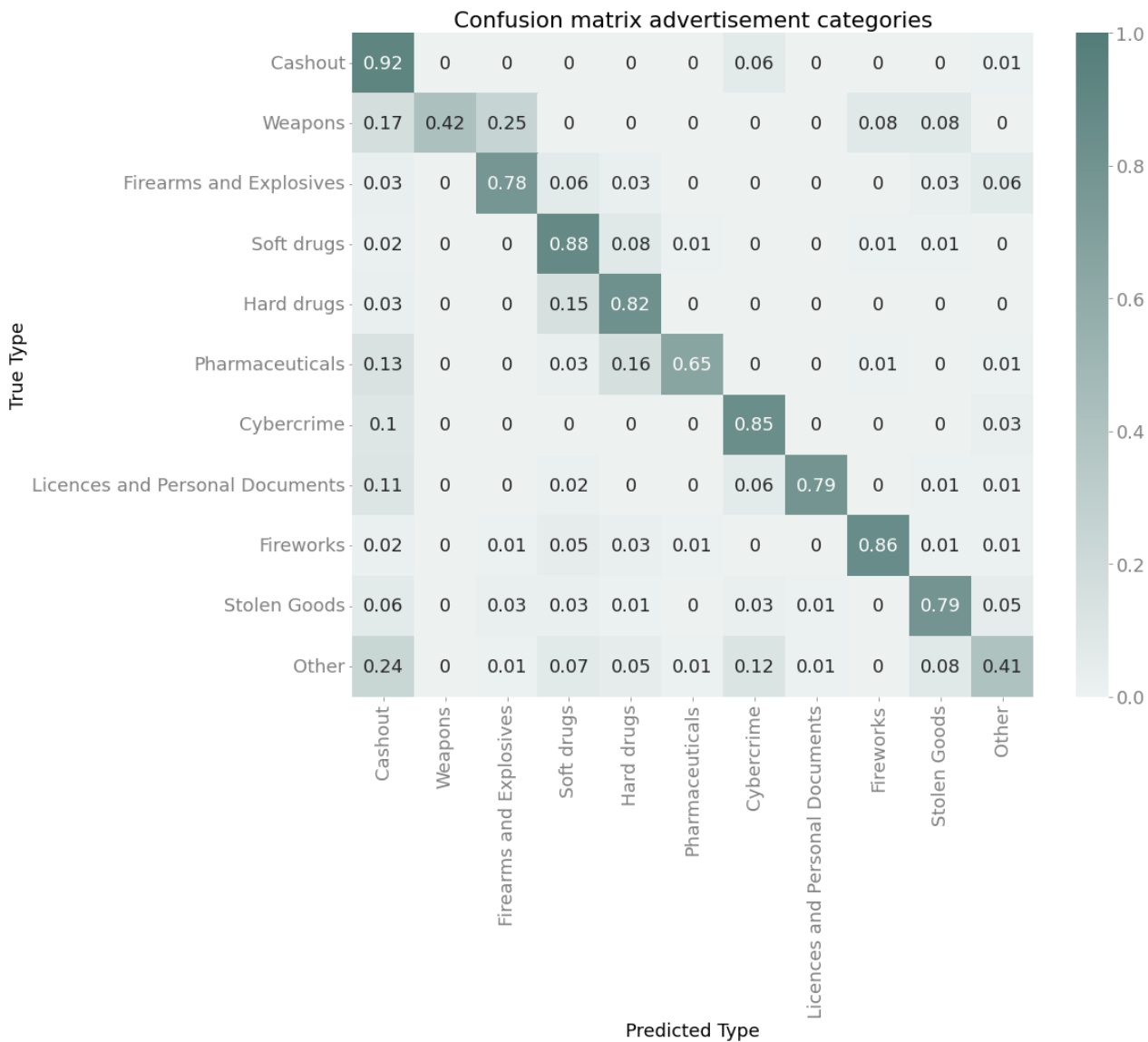


FIGURE 3.5: Confusion matrix of the predicted advertisement category

The messages and members of the Telegram groups were scraped between the 16th of May 2022 and the 20th of June 2022. We scraped all available messages from the groups, starting at the earliest available message up and until the last message sent before the moment of scraping. The earliest available message is often a message service indicating how the group was created or a random message if the group auto-deletes messages after a given time. We also scraped all available members from the groups. However, the Telegram API cuts off the function returning members of a group after 10,000 users have been returned. This can be overcome by setting specific function parameters, but we found out about this happening after the scraper was finished.

The snowball process was active at the same time as the scraping process. It was decided to stop the snowball process manually when it had become clear that the snowball had converged towards groups that did not fit our topic. To illustrate, the groups in the

to-be-scraped list were either focused on conspiracy theories or contained only messages in German, clearly focusing on a German audience. There were also indications that scraping these groups would let our snowball method converge only further towards similar, off-topic groups. While this would be interesting research in itself, we believed this would stray too far from our objectives (i.e. focusing on Dutch Telegram groups used for cybercrime), so it was decided to stop the snowball.

In total, the messages and members of 387 Telegram groups were scraped. This raw dataset is about 65GB.

### 3.4.2 Preprocessing

As we started the scraping process on May 16th, 2022, we chose May 15th, 2022, as our cut-off date. Therefore, we deleted all messages sent after May 15th, 2022, from our dataset. For some Telegram groups, this inherently meant that all messages were removed, for example, if they were created after the cut-off date or did not contain any activity before the cut-off date. There were 24 Telegram groups in our dataset for which this was the case. These groups were removed from the dataset, after which we are left with a dataset containing 373 Telegram groups.

Out of the 373 Telegram groups in the dataset we started with, 78 groups, or 21% of our dataset, did not fit the topic. These groups are removed from the dataset and we are left with 295 Telegram groups. When looking at the topics of the removed groups, we see that several are about (the sharing of) conspiracy theories or the trading or investing of cryptocurrencies. Others are about cybercrime, but target another market, such as the German or British one. Next to that, we also see some Telegram groups that would fit our topic that are removed in this step. Several of these groups consist (mostly) of messages that contain only media without accompanying text (for example a video displaying cocaine or ecstasy). Since we have no access to the media content, these messages are displayed as empty in our topic-fitting process and empty messages do not meet the criteria we set. Therefore, if our random sample contains many empty messages, we may have had to exclude the Telegram group from our dataset. Furthermore, there is also a possibility that the messages in the random sample were not in Dutch or did not include evidence of cybercrime by chance, while the rest of the group would have fit the criteria.

It is interesting to note that the dataset also includes some Belgian groups. The method for topic-fitting presented in Section 3.2 does not allow us to distinguish between Dutch groups and Flemish groups, as Flemish and Dutch are difficult to distinguish in written form. While we could argue that this research is focused on the Netherlands and we should not include Flemish-speaking Belgian groups, we leave these groups in the dataset due to indistinguishability and leave it to Chapter 7 to discuss the effects.

During the later stages of our research, we found that the dataset contained two Telegram groups (out of 295) in which the `from_id` entry of all messages was empty. For both groups, manual inspection seemed to indicate that the causation is the same: only the



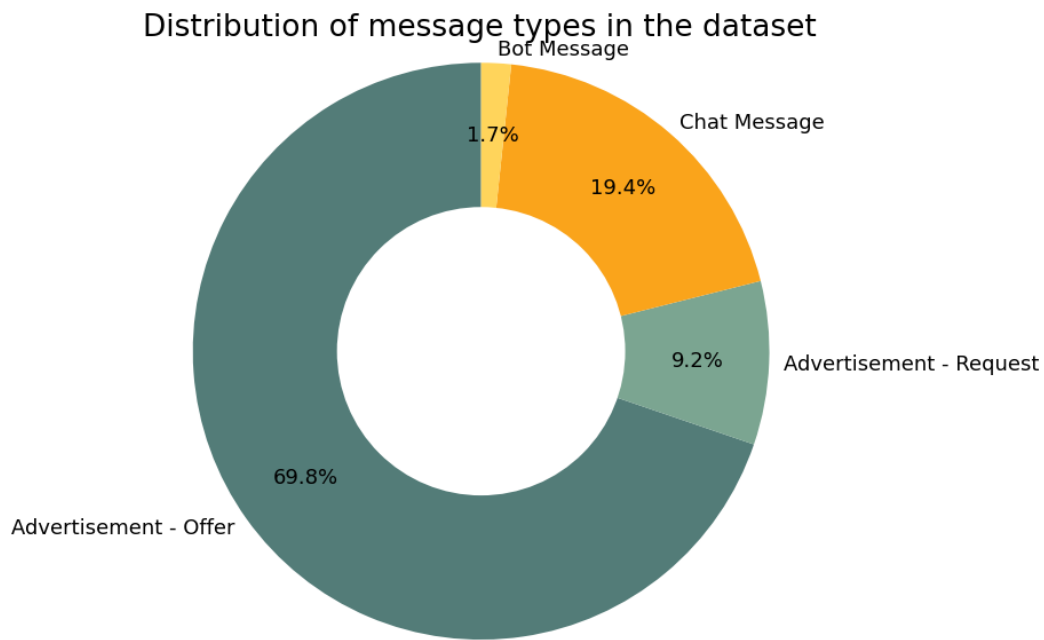


FIGURE 3.6: Distribution of message types over all messages in the dataset

owner or admin of the group has sent messages in the group (as if the group or admin were a separate user itself) and has since deleted their account. At the moment of writing, it is unclear if the reason for the `from_id` entry being empty is rooted in the fact that these users have deleted their accounts or if the `from_id` is always empty if a message is sent by the group or admin. Either way, the missing data may skew the results, so we remove the two Telegram groups where this occurs from the dataset.

### 3.4.3 Classification

When we apply the classifier to our dataset, each message in our dataset is provided with a label indicating its message type (out of the four message types in Table 3.1). Messages classified as an advertisement get an additional label indicating their advertisement category (out of the options displayed in Table 3.2).

The distribution of message types over all messages in our dataset is displayed in Figure 3.6. We see that about 79% of all messages in our dataset are classified as advertisements, the majority of which are advertisements offering a product or service. Next to that, about 19% of messages are chat messages and just under 2% of messages are bot messages.

Figure 3.7 shows the distribution of advertisement categories of all advertising messages in the dataset. We see that cashout is the most advertised category, consisting of almost 33% of all advertisements, closely followed by the combination of soft drugs and hard drugs, consisting of almost 31% of all advertisements. If one were to put advertisements

## Distribution of advertisement categories in the dataset

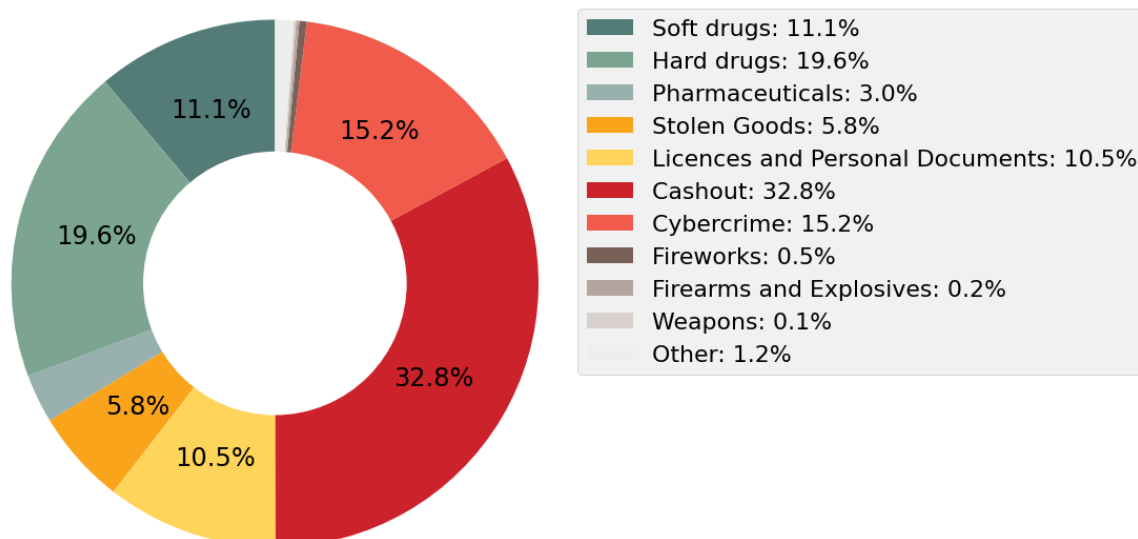


FIGURE 3.7: Distribution of advertisement categories over all advertisements in the dataset

of pharmaceuticals in the umbrella category of drugs, drugs would even be the most advertised category in our dataset.

If we separate our message types and look at the distribution of advertisement categories in our dataset (Figure 3.8), we see that most advertisements requesting something are of the cashout category. Manual inspection of the messages shows good reason for this, as advertisements in the cashout category often ask for bank cards or people that are willing to use their bank accounts to transfer money. While there is a strong possibility that the people behind the accounts are offering money laundering services or the like, they are requesting people to help them or offering people “jobs”. Hence, messages asking for cards or for people willing to use their accounts are often labelled as advertisement - request.

The observant reader might notice that the division of the message types of the ground truth dataset (Figure 3.2) and the dataset used in this study (Figure 3.6) differs; the ground truth dataset contains relatively more chat messages and advertisement requests and fewer advertisement offers than the dataset used in this study. We believe this may be caused by the method used to choose the number of messages that were labelled in each Telegram group in the ground truth dataset (Section 3.3.1, as we labelled a relatively higher percentage of messages in smaller groups and a relatively smaller percentage in bigger groups. Smaller groups, we found, were more likely to contain many chat messages, while bigger groups were likely to contain more advertisement offers. While one could argue that this creates a bias towards chat messages, we believe this approach was a necessary step to assure there be enough chat messages for the classifier to train on. The discrepancy in the percentage of advertisement offers can likely also be explained with reasoning similar to the paragraph above; advertisement requests often

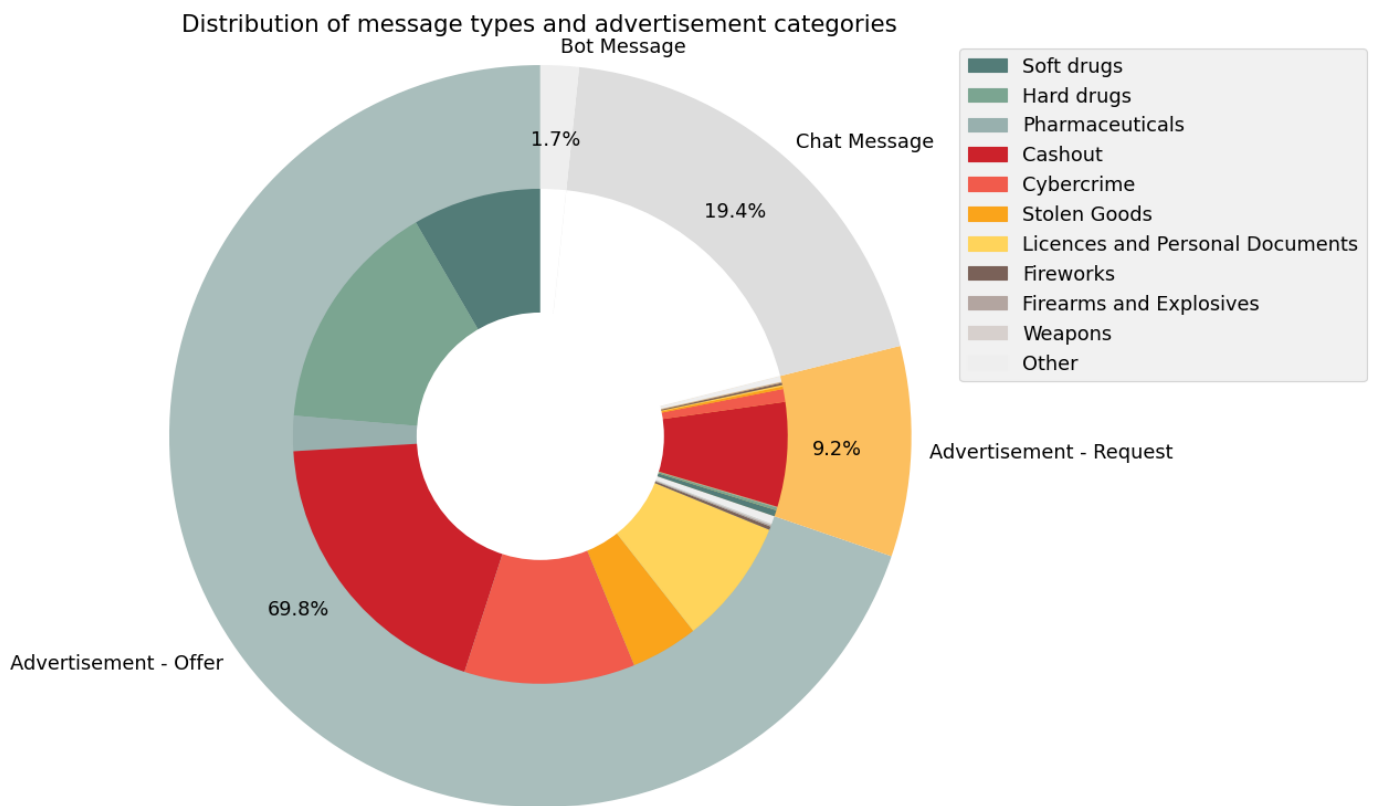


FIGURE 3.8: Distribution of message types and advertisement categories over all messages in the dataset

fall in the cashout category and can often be interpreted as both a request and an offer. It might have happened that these advertisements were often labelled as a request while manually labelling the ground truth dataset, while the classifier labelled the same messages as advertisement offers. Using these two arguments, we believe there is no reason to believe the quality of the labels in the dataset used in this study is any less.

### 3.5 Conclusion

This chapter has presented an approach for this study, mainly focussing on data gathering and preprocessing. The method presented is relatively straightforward and takes into account the scope of the current research and the local embeddedness of cybercrime.

Afterwards, the initial dataset was extended by employing a classifier created in an earlier project, which labelled all messages to be one of four message types. In case a message was labelled as an advertisement (either offering or requesting), the message received a second label indicating the topic or crime market of the message. The classifier information was added to the dataset, resulting in a final dataset that is used in the continuation of this work. The classifier seems to perform well, which allows us to use

---

its results in the continuation of the study. However, it is important to realise that the classifier results need not always be correct.

We end up with a dataset containing 293 Telegram groups, as presented in Section 3.4. The dataset contains slightly less than 27.1 million messages and almost 147,000 unique users. Almost 70% of the messages in the dataset are advertisements offering a product or service, the most prominent categories being cashout, soft drugs and hard drugs, cybercrime, and licences and personal documents. The 9% of the messages labelled as advertisements offering something are almost all offering cashout. Roughly 19% of messages are labelled as chat messages. In other words, with almost 80% of all messages focussing on cybercrime, we confirm that Dutch Telegram groups are used as an advertisement platform for cybercrime and that the public groups in our dataset are likely to be used as a marketplace.

# Chapter 4

## Phase 1: Groups

The first phase of our research dives into Telegram groups. The research questions we wish to answer are as follows:

1. How can we create a profile for a Telegram group?
2. To what extent can we define the relations between Dutch Telegram groups used for cybercrime?

To answer the first question, we gather information about the Telegram groups in the dataset to form metrics, as described in Section 4.1 below. These metrics give us more insight into the Telegram groups, essentially creating a profile for each group. The first question also requires us to take a look at how individual Telegram groups are used for cybercrime. Building from that, the second subquestion looks into the relations between the Telegram groups, allowing us to broaden our perspective from single Telegram groups functioning as cybercrime markets to the role of the connected Telegram groups. In other words, it forces us to create an understanding of the interconnectedness of the Telegram groups in our dataset and allows us to discover whether we should view the connected Telegram groups as one entity. Moreover, the relations can not only help us understand the nature of Telegram as a platform for cybercrime but can also point towards subgroups or communities of Telegram groups in the dataset.

### 4.1 Approach

#### 4.1.1 Basic info

We start the process by obtaining all sorts of information about the Telegram groups in our dataset. An overview of the metrics can be found in Table 4.1. These metrics can help build a profile of the Telegram groups, and may even provide some distinguishing properties. Each of these metrics is trivial to determine given our dataset, but we highlight some below:

Group name
Group ID
The date a group was created
The manner in which a group was created
The name of the group from which the group was migrated (if that is how it was created)
The ID of the group from which the group was migrated (if that is how it was created)
Moment of last activity in the group
The length of the group's existence in days
Total nr. of current members
Total nr. of bots
The nr. of current active members
The nr. of current passive members
The percentage of current members that are active
The total nr. of users that have ever been active in the group
The total nr. of messages sent by a channel
The nr. of channels sending messages in the group
Total nr. of messages
Total nr. of media messages
Total nr. of text messages
Total nr. of unique text messages
Average nr. of messages per day
The nr. of duplicates of the most occurring message
The nr. messages with at least one duplicate
The highest nr. of messages sent by a single member
The average nr. of messages sent by a single member
Total users added
Total users deleted
The percentage of active users that are still a member of the group

TABLE 4.1: Overview of the basic information metrics determined for all groups

- the date the Telegram group was created and the manner in which a group was created point to when and how a group was created. If a group was created by migrating another group, that other group's name and ID are also added.
- the moment of last activity and length of existence days tell us how long a group has existed and what the last moment of activity was.
- the total number of current members and total number of bot members tell us how many members a group has at the moment of scraping and how many of these members are bots.
- total current active members displays the number of current active group members, while total active members all time displays the number of group members that

have ever been active in the group, including those that may not have been a member anymore at the time of scraping. The number of total current passive members displays the number of group members that have not been active (i.e. that have never sent a message in the group). Next to that, we also determine the ratio of current active members against the current total members, allowing us to determine the percentage of active members in a group.

- the total number of messages send by a channel displays the total number of messages in a group that were sent by a channel and the number of channels sending messages displays the number of different channels that have send a message in the group.
- the total number of messages, the total number of text messages, the total number of media messages, and the total number of unique text messages tell us something about the messages in a group. Additionally, we define the average number of messages per day, the number of times the most occurring message occurs, and the number of messages that occur more than once in a group.
- the highest number of messages sent by a single member and the average number of messages sent by single member highest tell us the highest number of messages a single user has sent in the group and the average number of messages a user sends in the group
- the total number of users added and deleted look at the number of users that joined and left the group by looking at the message services. However, we believe these numbers to be quite inaccurate, as they do not seem to correspond to the total number of members a group has.
- the ratio current active members total active members might give an indication of the throughput of active members in a group, as it compares the number of current active members in a group to the number of all users that have ever sent a message in the group.

The information and metrics in table Table 4.1 are combined with the information obtained by the classifier as explained in Section 3.3. For each group, we add the alias, the prevailing message type and advertisement category, and also the division of messages over the message types and advertisement categories, as we feel this has added value for a possible group profile. Building from that, the following metrics are also determined for each group:

- the total number of advertisements: the total number of messages classified as an advertisement in a group.
- the ratio of advertisement in the messages: the ratio of messages that are also advertisements against regular messages, which can provide us with a percentage of advertisements in a group.
- the average number of advertisements per day: the average number of advertisements send in a group per day.

#### 4.1.1.1 Clustering

As an experiment, we use two common unsupervised learning techniques to cluster the Telegram groups in our dataset based on their basic properties. Unsupervised learning is a common approach used to identify clusters of items that exhibit similar characteristics, enabling us to differentiate between these clusters. Using this approach, we aim to label or categorize the Telegram groups according to their basic properties to bring a new perspective to our study. To accomplish this, we must first determine the specific properties that provide meaningful insights about a group, prioritizing clarity and interpretability in order to enhance the comprehensibility of the resulting clusters.

We use KMeans clustering (Lloyd, 1982; MacQueen, 1967) and Gaussian Mixture Model (GMM) clustering as unsupervised learning methods to try to find clusters of Telegram groups in our dataset. KMeans clustering is a commonly used unsupervised learning clustering method, meaning it can attempt to find clusters of data with similar characteristics without needing to use existing labelled data (Sinaga & Yang, 2020). GMM clustering is a similar clustering algorithm, but takes variance into account, allowing for a more precise shape of the clusters, and returns the probability that a data point belongs to a cluster (Patel & Kushwaha, 2020). We use KMeans clustering because we believe it is right to explore this avenue with a common, simple approach. GMM clustering is performed because it is another common clustering method and to be able to compare the results of both approaches.

To prepare the data, we normalize it using a min-max normalization method and replace infinity values with 0. Normalizing the data allows us to compare the values of otherwise incomparable characteristics. Then, using SciKitLearn (Pedregosa et al., 2011), we determine the number of clusters we want to find in our data using Cluster Inertia (also known as 'the elbow method'). Finally, we perform KMeans clustering (*sklearn.cluster.KMeans*, n.d.) using KMeans++ initialization and GMM clustering (*sklearn.mixture.GaussianMixture*, n.d.). The results are visualized in a Seaborn (Waskom et al., 2017) pair plot.

#### 4.1.2 Mention network

In Dargahi Nobari et al. (2017), the authors create a mention network between Telegram channels to determine the relationship between the channels. We take a similar approach to determine the relations between the Telegram groups in our dataset. A mention is defined as when a Telegram group is mentioned in a message in one of the two following formats: *@TelegramGroup* or a variation of *https://t.me/joingroup/TelegramGroup* (which can also be *t.me/TelegramGroup*). Mentions of this kind are formatted by Telegram automatically to become a clickable link that leads to the mentioned group. In the same study, detailed in Chapter 2, the authors define three types of mentions: self mentions, spam mentions, and ham mentions. While we see the added value of differentiating between these types of mentions, we leave it to a future study to distinguish these types, as we believe an argument can be made for the importance of each of these types of mentions.



Still, relations between the groups are challenging to define: it is very clear if a message in a group mentions another group, but it's more difficult to give meaning to a mention. Looking through the messages of some of the Telegram groups in our dataset, we see three different ways another group is mentioned. A group can be mentioned by a single user in a chat conversation, for example, to recommend another group to a (group of) user(s) as part of a conversation. Next to that, there are advertising messages that recommend a single group, sometimes in combination with a vendor promoting their product. Furthermore, we see that some Telegram groups contain advertising messages promoting multiple Telegram groups in one message, implying a stronger relation between these groups (for example a collaboration, an alliance, or groups maintained by the same person or group of people). Although either one of these types of mentions seems to be of a different nature, each of them can be important in this research.

We use the following steps to determine the mentions in a group:

1. Load the messages of a Telegram group
2. Remove empty messages and duplicate messages
3. Check all messages for mentions and keep only unique mentions per message
4. Count the number of unique mentions in this group

We only look at unique messages to determine the mention-relations between the Telegram groups. The main reason for this is that we expect the sheer number of messages using mentions to advertise Telegram groups to be enormous compared to the number of messages in which a mention is used in conversation. We want our mention network to display all mentions instead of just the strongest mention relations, which we achieve by only keeping unique messages. Additionally, we only take into account unique mentions in each message. This is mainly done because some messages contain the same mention over a hundred times. As we want to avoid messages like these skewing our results, we count only unique mentions in a message.

Using the overview of all mentions between all groups, we use the NetworkX library (Hagberg, Swart, & S Chult, 2008) to create a graph of the relations between the Telegram groups. The Telegram groups in our dataset are nodes in the graph and a directed edge is added between two nodes if one group mentions another. For most visualizations, we use NetworkX's Kamada Kawaii layout. The resulting graph structure is used as a basis for the next steps.

We use the basic graph and NetworkX predetermined functions to determine several network properties of the Telegram groups. Research about online social networks (OSN) often uses degree, in-degree, or out-degree to say something about the popularity of a node in a network, referring, for example, to the size of an audience, the number of interactions, or the number of social relationships. For example, a user on a social media platform that uses friend relations has a high degree when the user has many friends. Looking at in- or out-degree in a network of Telegram groups can tell us something

about a group’s promoting behaviour or how a group is promoted in other groups. This might give us an indication of the group’s role in the network. Next to that, we look at shortest path based methods, such as closeness centrality and betweenness centrality. These measures can be used to determine what nodes in an OSN are more central to a network. Furthermore, we look at the eigenvector centrality of the nodes. The eigenvector centrality not only takes into account the number of links between nodes, but it also takes into account the influence of the nodes a node is connected to. Using eigenvector centrality, a node gets a higher score when it’s connected to influential nodes. This could potentially lead us to more influential groups in the mention network. Lastly, we determine the PageRank of each node. Created by Google, PageRank is a well-known algorithm to rank webpages based on their incoming links, allowing webpages that are linked by many other webpages to be returned as the top result. The network properties are then correlated to each other and to several key group metrics determined in Section 4.1.1. In specific, we look at the number of members a group has, the number of messages in a group, the number of advertisements in a group, the ratio of messages that are advertisements, and the average number of advertisements per day send in the group. The results of this are displayed in a heatmap.

We try several graph visualizations to try to find communities of Telegram groups and to find out more about the relations in the network of Telegram groups. Starting with the basic graph of the mention network, we create several other graphs. First, we look at the influence of the number of mentions by colouring the edges in the graph based on the normalized mention count. We normalize the mention count by using *ScikitLearn’s* (Pedregosa et al., 2011) *MinMaxScaler* function. Secondly, we turn the directed graph into an undirected graph keeping only the reciprocal edges, as

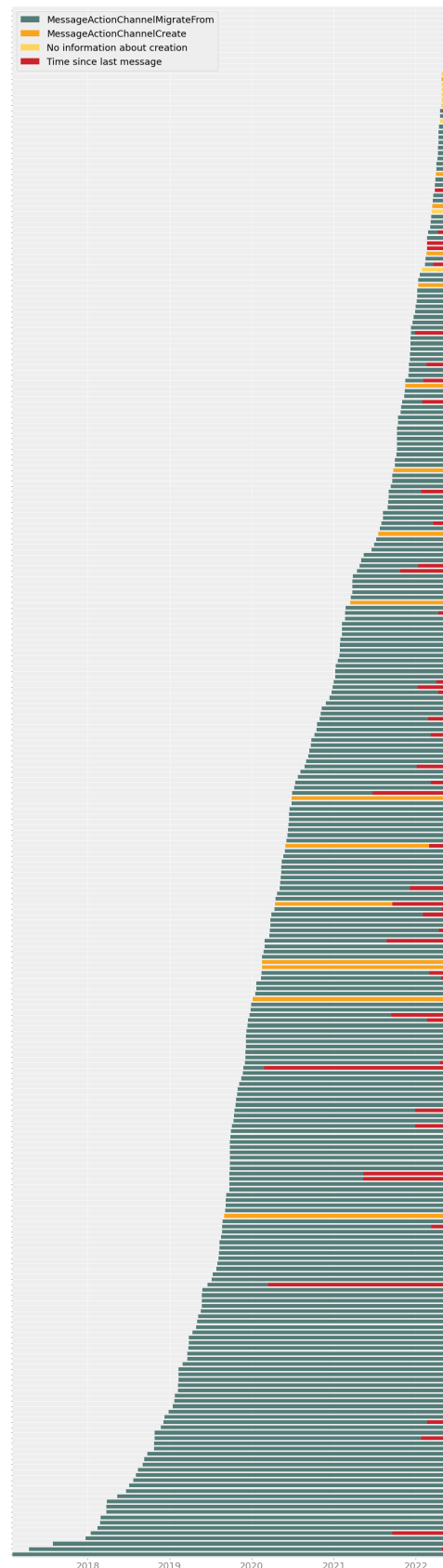


FIGURE 4.1: A timeline of the existence length of the Telegram groups in the dataset

this may give another perspective on existing relations and communities. Furthermore, we create smaller mention graphs of Telegram groups focussing on the same crime market.

#### 4.1.2.1 Community detection

Using the mention network created in the previous step (Section 4.1.2, we try an additional approach to detecting communities in a graph: Louvain Community Detection, as first proposed by [Blondel, Guillaume, Lambiotte, and Lefebvre \(2008\)](#)). This algorithm tries to partition the network into clusters of densely connected nodes, with nodes belonging to different communities only sparsely connected. We use a Python library ([Aynaud, n.d.](#)) created by Thomas Aynaud that was specifically designed for community detection. This library, however, only seems to work for undirected graphs, so we first transform our directed mention network graph into an undirected graph that keeps only reciprocal edges and into an undirected graph that keeps all edges. Both versions of the graph are then used as input for the library function and the output is visualised.

#### 4.1.3 Advertisement and member overlap

Another approach perspective on relations between groups is to see how much they have in common in terms of messages and members. We assume advertisements are only sent by an entity that benefits from it, being it one person, a group, or an organization. We do not necessarily assume that the person or group sending the advertisement is always using the same Telegram account (i.e. we expect that there exist cases where multiple users are sending the same advertisement), so it would not make sense to look at the user sending the advertisements. However, if we assume the same advertisement has the same beneficiary, we believe there might exist a relation between groups containing many of the same advertisements.

To determine the advertisement overlap between two groups, we do the following. First, we take the hashed version of all advertisements in a Telegram group and remove all duplicates. Then, using set theory, we divide the intersection of the advertisements of the two groups by the union of the advertisements of the two groups. In other words, we divide the number of overlapping advertisements by the total number of advertisements. This is done for each combination of Telegram groups in the dataset, resulting in a matrix. A heatmap is created to visualize the result.

We can do something similar to determine the overlap in group members. We assume that there may exist a relationship between groups with high overlap in members. To determine the member overlap, we take all members of a Telegram group and remove all duplicates. Then, we divide the intersection of members by the union of members using set theory. The result is the percentage of member overlap, which is stored in a matrix. A heatmap is created to visualize the results.

## 4.2 Results

### 4.2.1 Group descriptives

To start with, we look at how the groups came into existence. 260 of the Telegram groups in our dataset, which is 89%, were created by migrating another group. According to the Telegram API documentation, migration happens when a basic group is migrated to a channel or supergroup. Since our dataset only contains Telegram groups and no channels, we assume most groups in our dataset are supergroups because of this. Next to that, 16 groups were created as a supergroup as indicated by the `MessageActionChannelCreate` message service as a first message. Furthermore, there are 17 groups of which there is no information about how they were created. This could be caused by our scraper not being able to scrape all messages, but we deem it more likely that these groups have turned on the auto-deletion function for messages in their group. Auto-deletion would cause messages to disappear after a given time. Our suspicion is fueled by the fact that the majority of the groups of which we have no information only contain very recent messages and are quite active given their lack of history. However, there is no way to confirm our suspicion.

If we look at the number of days the groups in our dataset exist, we see that on average a group has existed for roughly a year and nine months at the moment of scraping. Looking at percentiles, the division gets a bit more specific: 25% of the Telegram groups in our dataset have existed 6 months or shorter. The next quartile of the groups has existed between six months and a year and ten months. The third quartile of groups has existed for up to two years and eight months. The last quartile of groups has existed between two years and eight months up to almost five years. This indicates that Telegram groups used for cybercrime have existed at least as long. An overview of the length of existence of the Telegram groups in our dataset can also be found in Figure 4.1. Interestingly, this figure seems to indicate that slightly more Telegram groups came into existence when the first Covid-19 lockdown started in the Netherlands.<sup>1</sup>

Looking at the metrics about members of a group (Table 4.2), we distinguish between the total number of current group members, the total number of active members overall time (i.e. the number of unique members that has ever sent a message in the group), the total number of active members of the current members, the total number of passive members, and the total number of users that were added and deleted. Looking at the total number of current members, we see that the mean of this metric is 1,359 members, with a standard deviation of 2,063. Looking at the division of current members over the quartiles and taking into account the minimum and maximum, we see that the data indicates that few groups have a very large amount of members, while at least half of the groups have fewer than 526 members and at least 75% of the groups have fewer than 1,801 members. In this metric, we also see the phenomenon described in Section 3.4, where the maximum number of group members returned is 10,000. Our dataset contains five groups where this occurred.

---

<sup>1</sup>This first intelligent lockdown was announced on the 12th of March, 2020.

	total nr. current group members	total number of active members overall time	the total number of active members of the current members	the total number of passive members	the total number of users that were added	the total number of users that were deleted
mean	1359	1245	459	900	1795	7
std	2063	2072	695	1587	4345	11
min	1	1	0	0	0	0
25%	75	77	24	39	43	0
50%	526	389	200	209	266	2
75%	1801	1571	563	1048	1442	9
max	10000	12946	5208	9999	33862	77

TABLE 4.2: Description of distribution of Telegram groups over different member-related metrics

Our dataset contains a little over 27 million messages spread over 293 Telegram groups. The group with the fewest messages contains exactly one message, while the group with the most messages contains 918,341 messages. The lowest quartile of groups contains fewer than 929 messages, the second quartile up to 17,850, the third quartile up to 116,066, and the fourth quartile up to 918,341. This implies that our dataset is likely to have a right-skewed distribution when looking at the number of messages sent in a group.

However, we can also look at whether a message contains media, text, or both. Figure 4.2 displays the distribution media and text messages over the whole dataset. It must be

### Distribution of media and text messages in complete dataset

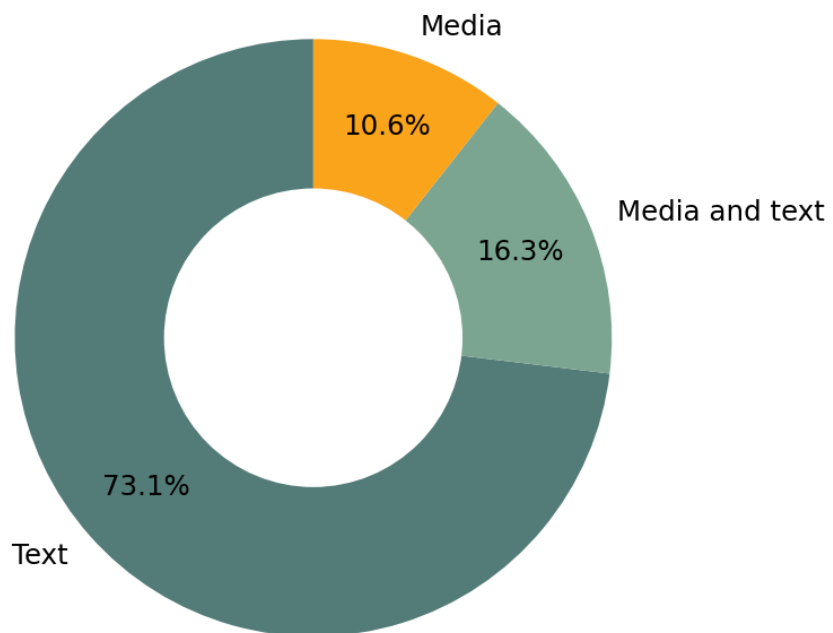


FIGURE 4.2: Distribution of text and media messages

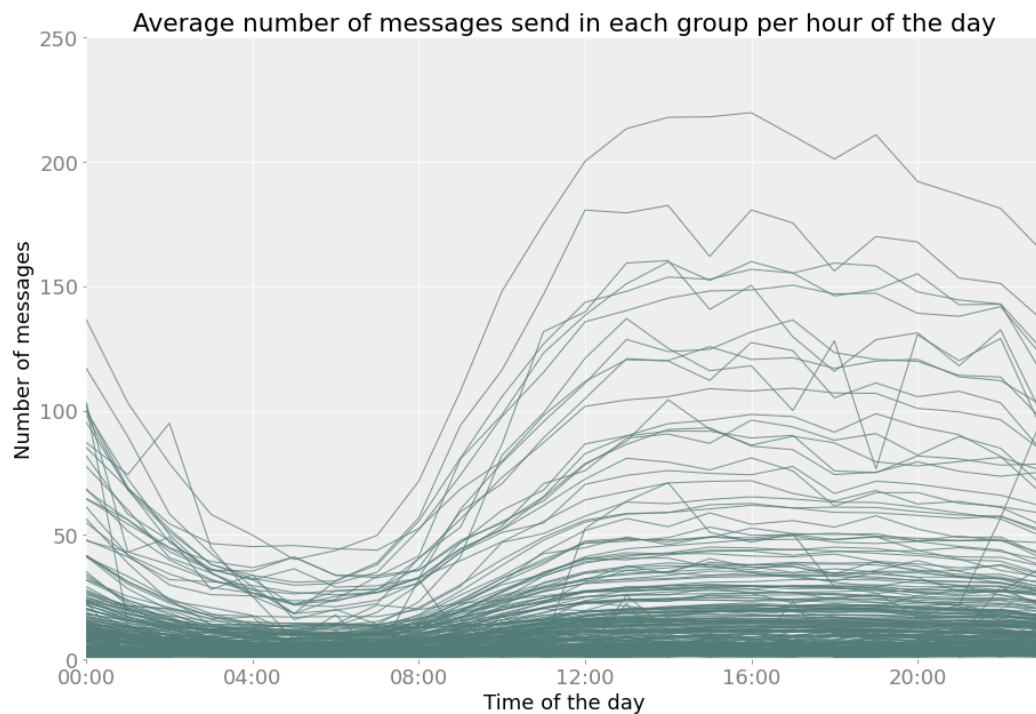


FIGURE 4.3: The average number of messages sent in the Telegram groups per hour of the day

noted that this research is biased towards media messages because we do not have access to the actual media and a Telegram group is more likely to be left out of the research due to the topic fitting process in Section 3.2. Nonetheless, 10.6% of the messages in our dataset contain only media and 16.3% contain media and text.

Still focussing on messages, Figure 4.3 shows the average number of messages sent in each group per hour of the day. It is clear that some kind of day and night rhythm is present, as most groups see a higher average number of messages being sent between 8:00 and 00:00 than between 00:00 and 8:00. We also see that the busiest moments are between 12:00 and 22:00 for most groups. Figure 4.3 also tells us that while there are many groups where the average number of messages sent per hour is below 20, there are some groups where this number is higher than 50 during the day.

#### 4.2.1.1 Aliases and crime markets

Each group was given an alias based on the predominant message type or advertisement category, as explained in Section 3.3.3. Table 4.3 displays the distribution of aliases over the groups in our dataset. We see that 28% of groups in our dataset can be seen as chat groups, 25.6% as a group focusing on drugs (either soft drugs or hard drugs), and 37.2% as cashout. When we compare this to the second column in Table 4.3 we see that the groups with cashout as an alias contain almost two-thirds of the messages in our dataset, groups focussing on drugs (soft drugs and hard drugs combined) contain

Alias	percentage of groups	percentage of groups measured in nr. of messages	percentage of groups measured in nr. of members
chat group	28.0%	3.0%	23.4%
bots	0.3%	1.0%	0.2%
soft drugs	5.5%	2.8%	7.8%
hard drugs	20.1%	21.9%	27.2%
pharmaceuticals	0.0%	0.0%	0.0%
cashout	37.2%	63.6%	36.6%
cybercrime	4.8%	6.4%	2.0%
documents	2.0%	1.2%	1.0%
stolen goods	1.0%	0.1%	0.4%
fireworks	1.0%	0.1%	1.4%

TABLE 4.3: Distribution of aliases over Telegram groups in the dataset, measured taking into account the number of groups, the number of messages in the groups, and the number of members in a group

almost a quarter of the messages, and groups focusing on cybercrime contain slightly more than 6% of messages. It stands out that chatgroups contain relatively few messages and groups focussing on cashout contain relatively many messages. The third column, on the other hand, shows that groups focussing on drugs (soft drugs and hard drugs combined) have relatively many members.

#### 4.2.1.2 Channels

Telegram allows public channels to send messages in public groups. This means that a user can post messages as a channel instead of with their personal account. Our dataset contains 113 groups where a channel has sent a message, where the channel(s) send an average of 248 messages. Looking at percentiles, we see that the distribution is right skewed: the first quartile contains between 1 and 7 messages sent by a channel, the second quartile up to 60 messages, the third quartile up to 257 messages, and the fourth quartile up to 4816 messages. Manual inspection has shown that messages sent by channels can have many purposes, such as (pinned) admin messages explaining the rules of a group, (pinned) advertisements, messages engaging in conversation, or messages enforcing the group rules. When we look at the number of channels sending messages in a group, we see that for 50% of the groups, this number is either 1 or 2. The group with the most channels sending messages has 36 different channels sending messages.

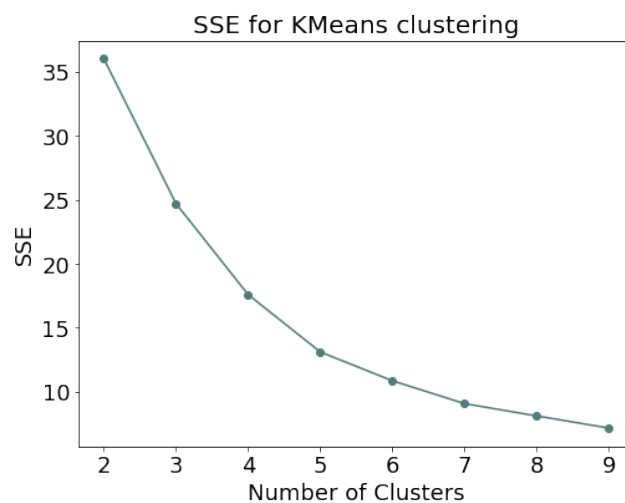


FIGURE 4.4: Cluster Inertia of KMeans clustering

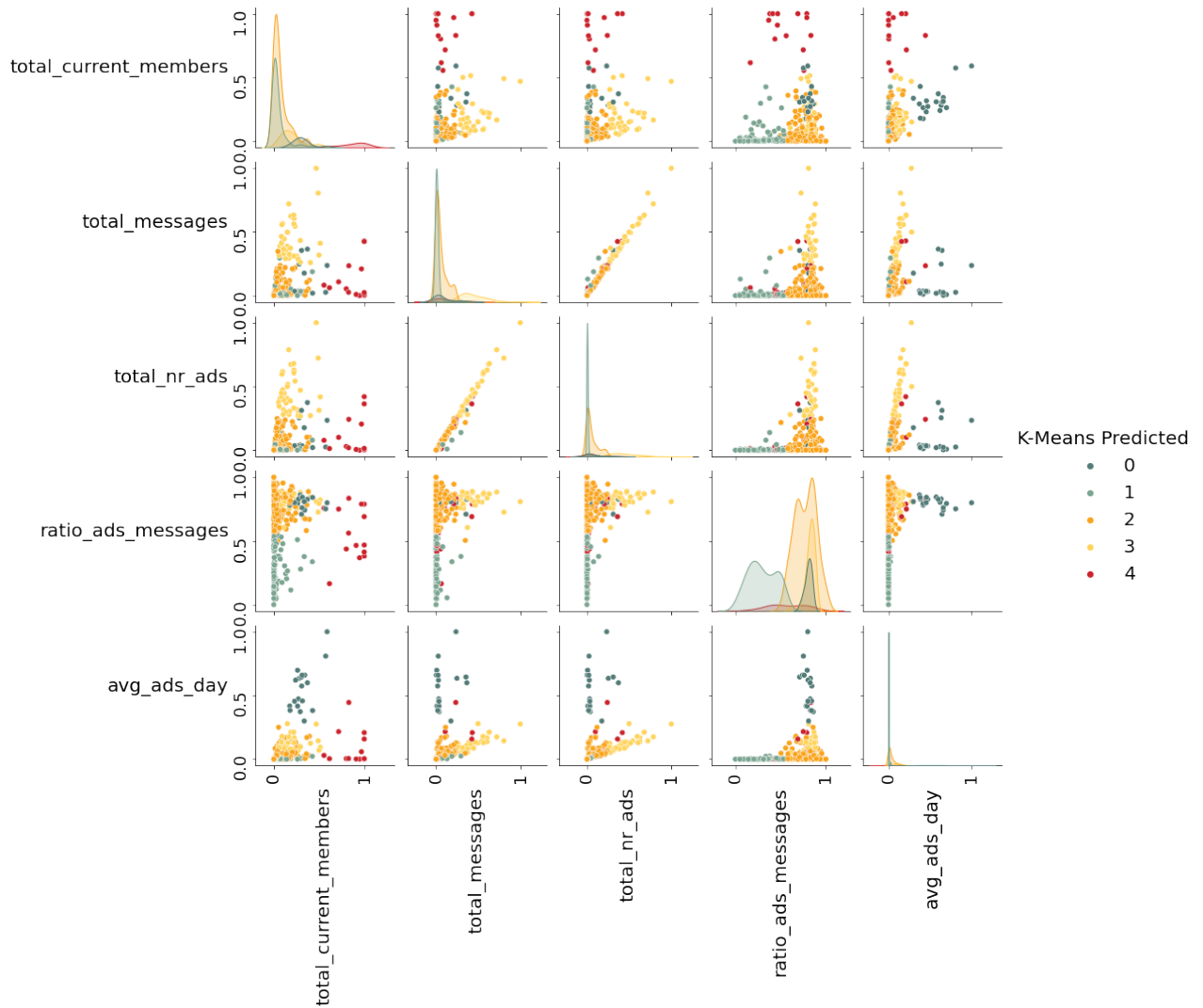


FIGURE 4.5: Pair plot of KMeans clustering into five clusters

### 4.2.1.3 Clustering

We use five properties that seem simply explainable and defining about a Telegram group: the number of members the group has, the number of messages in the group, the number of advertisements in the group, the ratio of messages that are advertisements, and the average number of advertisements per day. It must be noted that the number of messages in a group and the number of advertisements in a group are strongly correlated. We then apply KMeans clustering to the data with a different number of clusters to determine the Cluster Inertia for each number of clusters. Figure 4.4 displays the result and shows us that using 5 clusters is likely to give the best results. Finally, Figure 4.5 displays different characteristics and the found clusters in a pair plot. We see that the found clusters in the data are not very well distinguishable, which can be caused by either the number of clusters, the characteristics that are used, or the fact that this data might just be indistinguishable using this method.



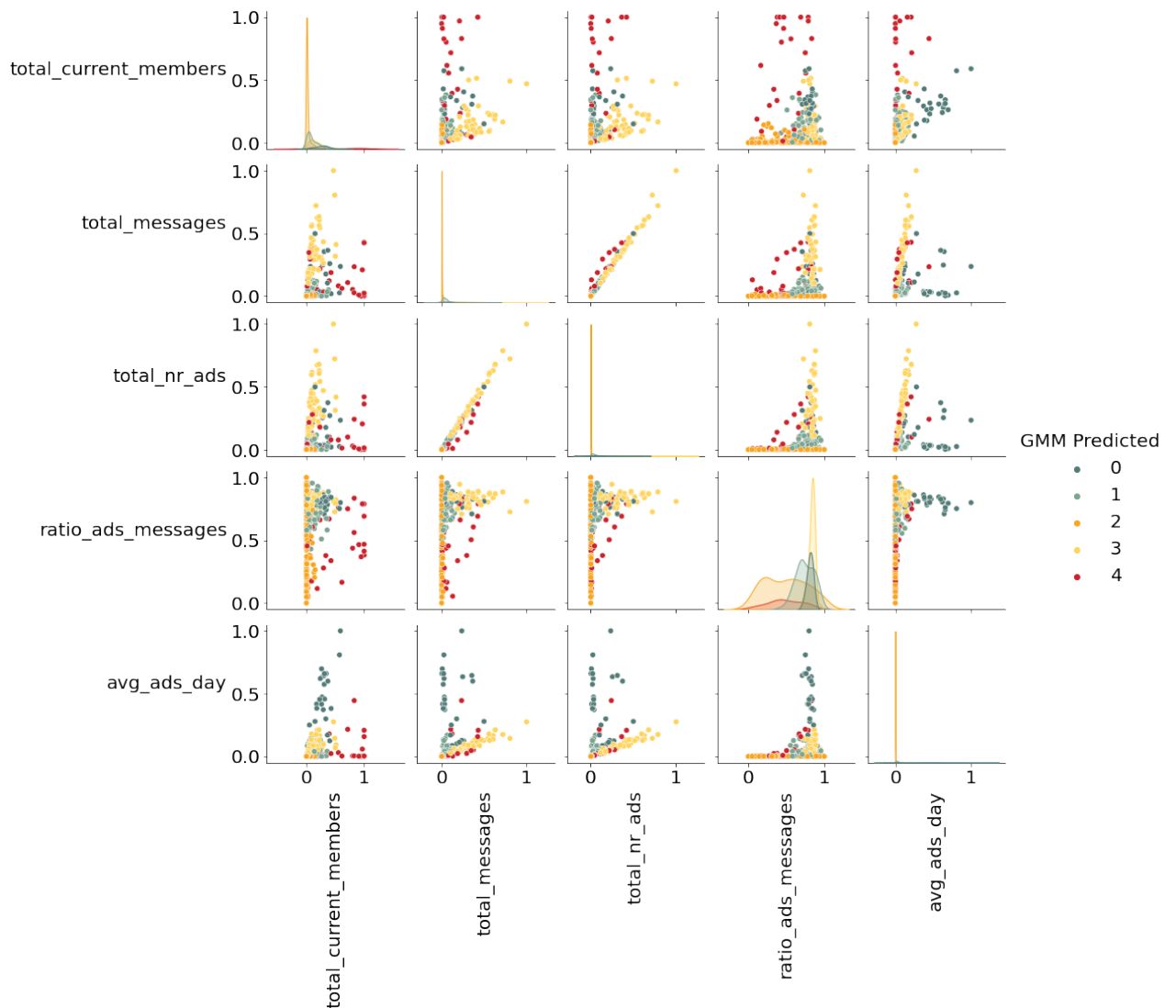


FIGURE 4.6: Pair plot of GMM clustering into five clusters

Applying the same method to GMM clustering yields similar results. Figure 4.6 displays the pair plot of this clustering attempt and we can see that it seems to perform worse than KMeans clustering (i.e. clusters do not seem distinguishable). This leads us to think that either something went wrong in our method or choice of input or this data is simply not suited for this method of clustering.

### 4.2.2 Mention network

Looking at the mention network, we see that 81% of all Telegram groups in our network are mentioned at least once by another group and 80% of all Telegram groups mention another group. Figure 4.7 displays the in and out-degree of the Telegram groups in our dataset, i.e. the number of different groups that mention a given group and the number of different groups that are mentioned in a given group. The group that is mentioned by the most different groups is soft\_drugs-15, which is mentioned by 111 different groups.

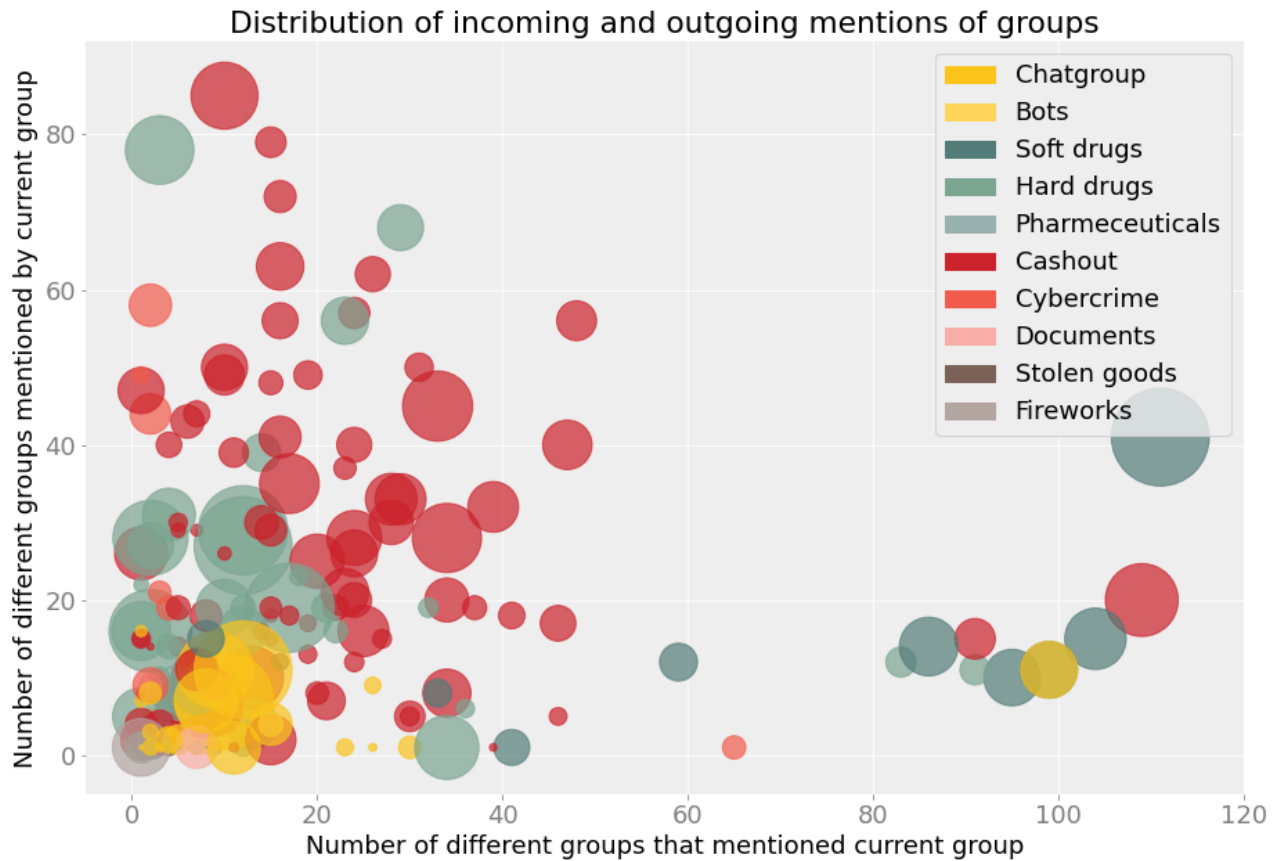


FIGURE 4.7: The in-degree and out-degree of the Telegram groups in the mention network

Mentioning 85 different groups, cashout-46 is the group mentioning most other groups. When we count the individual mentions instead of just the groups, we see that cashout-75 is mentioned 2208 times, which accounts for 10% of all mentions. The group containing the most individual mentions is cashout-23, containing 2928 mentions, or 13% of all mentions.

Figure 4.8 displays a basic overview of the mention network of the Telegram groups in our dataset. Each node is a Telegram group and a node’s size is relative to the number of members in the group. A mention relation between two groups is displayed by a directed edge. The network contains 293 nodes and 3732 edges. It is important to note that the dataset contains 29 Telegram groups with a degree of 0, meaning they have no incoming or outgoing mentions. These groups are not displayed in the network in Figure 4.8. Next to that, 26 Telegram groups are not mentioned by another group while they do mention other groups themselves and 29 Telegram groups are mentioned by other groups while not mentioning another group themselves.

If we draw the same network but colour the edges based on the number of times each mention occurs, we get the network displayed in Figure 4.9. In this network, the darker the edge, the more that specific mention has occurred. The observant reader can see five clusters popping up: two clusters where one Telegram group mentions one other Telegram group, one cluster where many cashout groups mention one other cashout



FIGURE 4.8: Basic graph of the mention network

group, one cluster where many cashout groups mention another cashout group and a chatgroup, and one cluster where four cashout groups mention each other. Interestingly, many of the clusters consist of groups with the same focus. Taking into account that we only include unique mentions in each Telegram group, we suspect there is a reason these clusters of groups appear; we believe there is a chance these Telegram groups actively promote each other and might be working together or be maintained by the same person or group.

If we transform our network to an undirected network where we keep only reciprocal edges, we get the network displayed in Figure 4.10. This network contains 109 nodes (i.e. unconnected nodes were removed) and 467 edges, which is 13% of the edges in the original network. In Figure 4.10 reciprocal, we see some clusters popping up. There



FIGURE 4.9: Basic graph of the mention network, but edges are darker when there is a higher number of mentions between groups

are several “islands” of two or three nodes and two bigger clusters of 9 and 82 nodes. A larger version of both bigger clusters can be found in Figure A.1 and Figure A.2 in Appendix A. The cluster of 9 nodes seems to show one cluster of nodes that are all connected and one node that has a reciprocal mention-relation with two other groups. The cluster of 84 nodes, on the other hand, seems to show many smaller clusters that are connected via a few central nodes. While we suspect that some of the Telegram groups in these clusters have real-life connections, for example through an alliance or by being run by the same people, there is no way to prove it through this data.

Furthermore, we created an overview of the mention-relations between groups of the same crime market (Figure 4.11). We can see that Telegram groups focussing on the

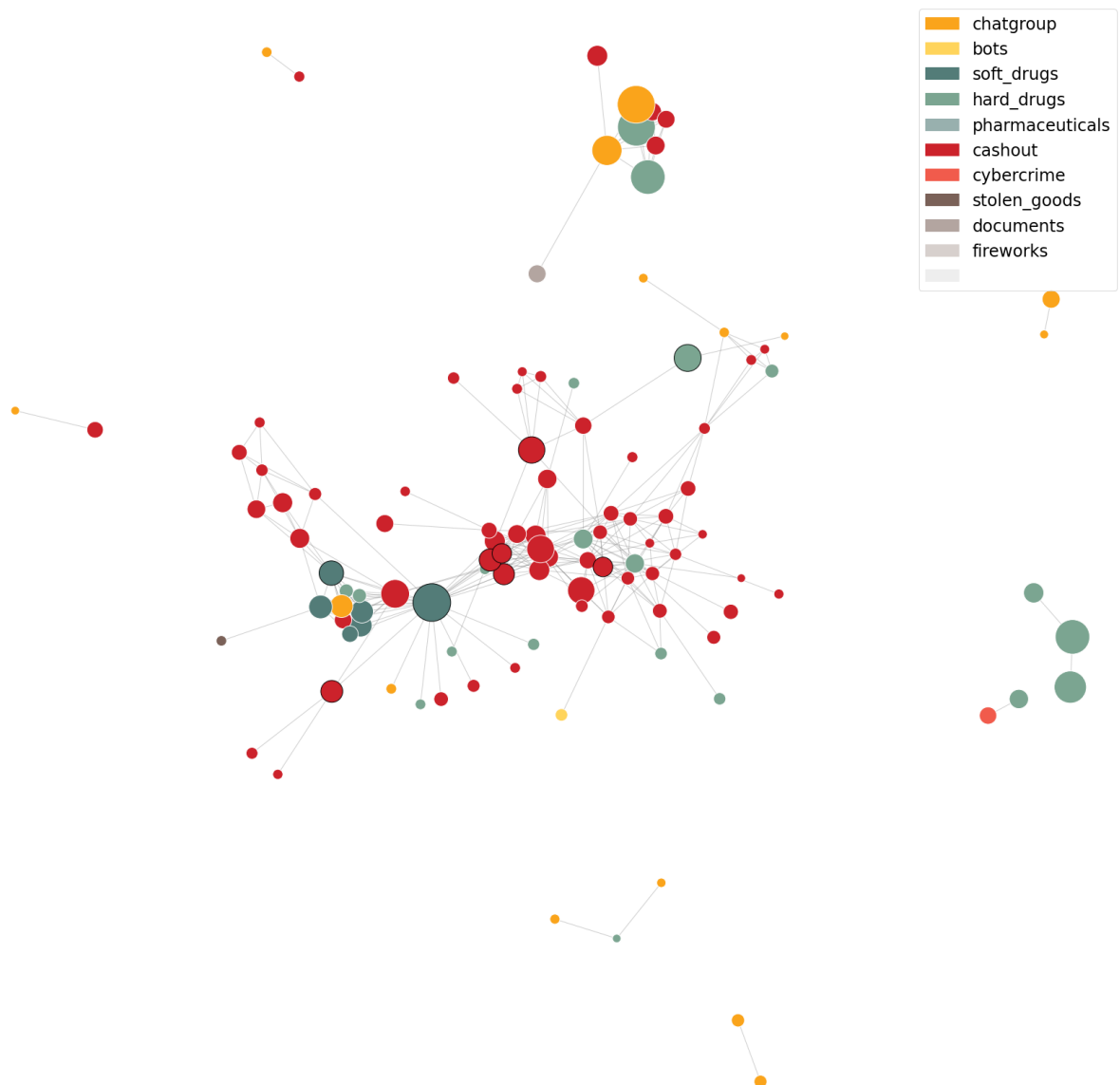


FIGURE 4.10: Mention network with only reciprocal edges

crime market cashout and hard drugs are very connected, the cashout network even displays a cluster of 5 groups. Also, the soft drugs groups are displaying a cluster, but we also see many unconnected groups. We believe this may be caused by soft drugs and hard drugs being so closely related; while the Dutch law distinguishes the two, people buying or selling drugs on the illegal market might not care about this distinction. Figure 4.12 graph supports this idea, as we can see that Telegram groups focussing on soft drugs and Telegram groups focussing on hard drugs form a tight network.

Finally, we take a look at the network properties of the Telegram groups in the mention network. We create a correlation heatmap of the groups' network properties and several key metrics. Figure 4.13 displays a heatmap of the correlation between the properties. In this figure, we see that there seems to be little correlation between the network properties

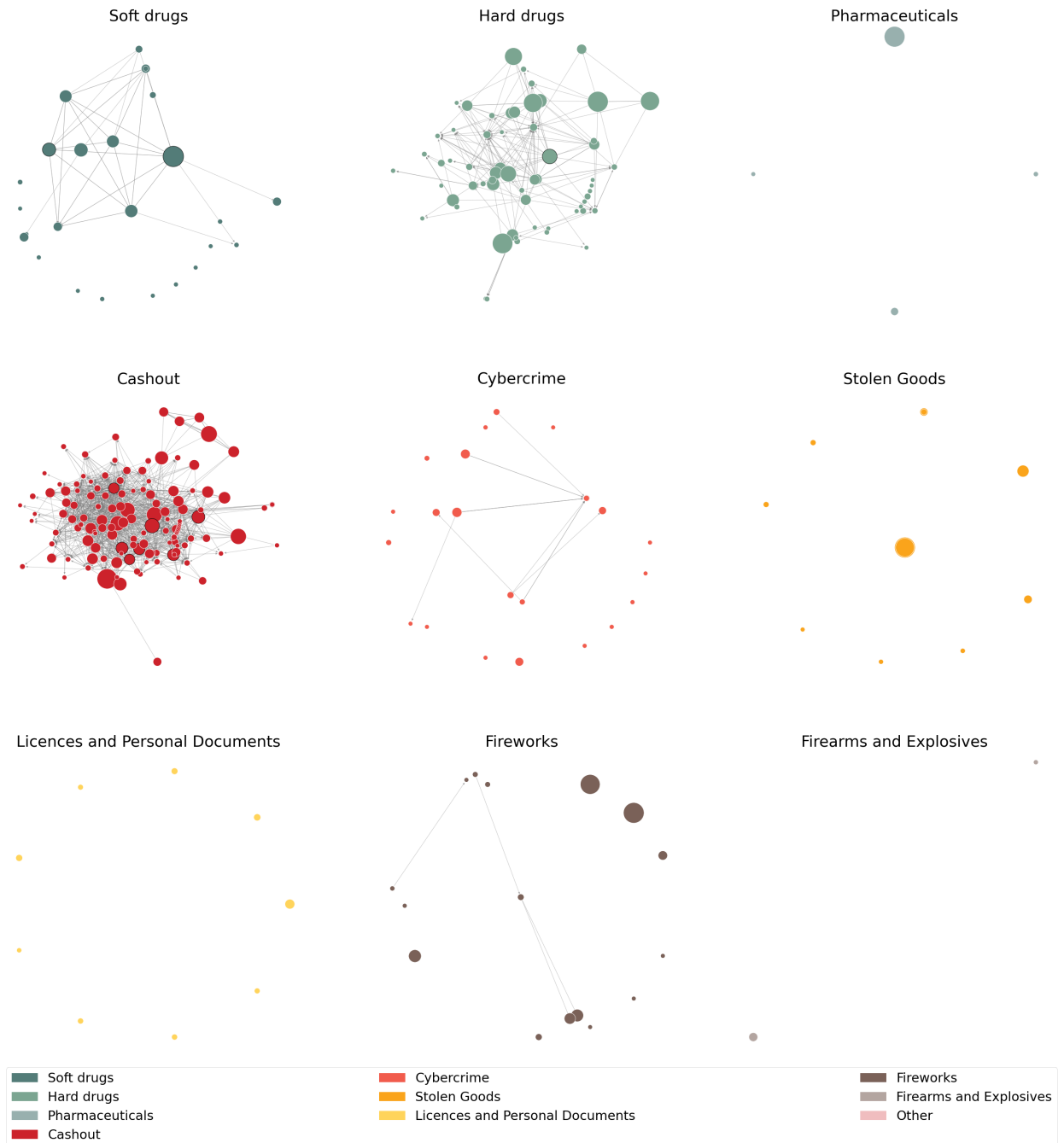


FIGURE 4.11: Mention relations between Telegram groups focussing on the same crime market

and the other properties. The only exception to this is the correlation between out-degree and the number of messages and advertisements and slightly less so for degree and the number of messages and advertisements. We believe this can be explained, though, by the fact that in our dataset it is statistically more likely more Telegram groups are mentioned when a group contains more messages. Since the degree is made up of in- and out-degree and the number of advertisements is closely related to the number of messages in a group, it seems logical that there are more outgoing mentions when a group contains more messages. Next to that, it stands out that degree seems to have a

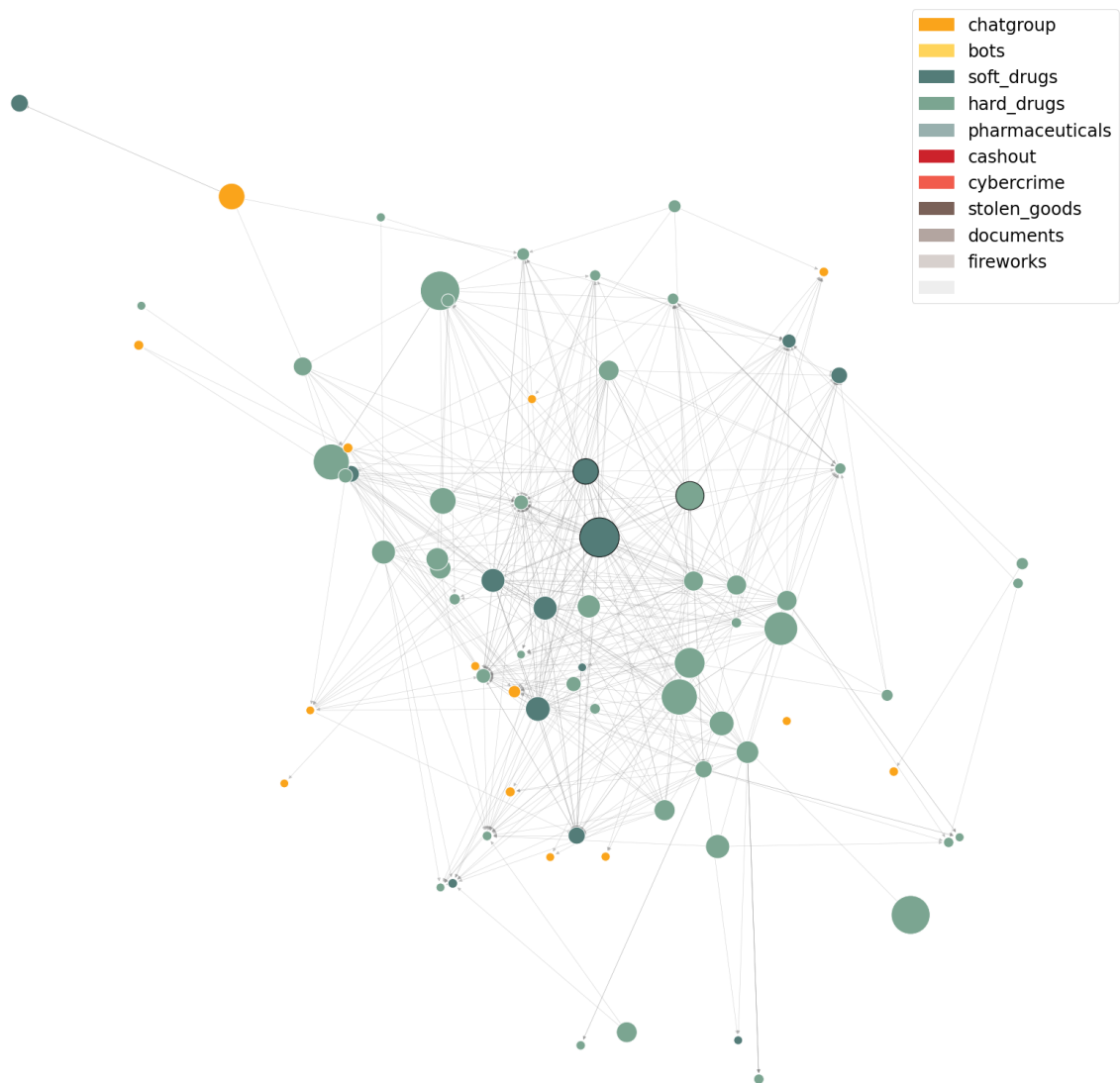


FIGURE 4.12: Mention relations between Telegram groups focussing on drugs

significant correlation with almost every other network property. We suspect that this is due to the fact that all other network properties seem to rely on the number of links of a node in the network. However, this also raises the question of what the different network properties mean. After all, most research into relations on social media is about social relations between two users, while we are looking at mention relations between two chat platforms that are oftentimes used for advertisement. In other words, we cannot just prescribe the meaning the different network properties have in OSN research to the mention relations in our Telegram group network. That is why we argue that degree, backed by its correlation to the other network properties and its simple explainability, in-degree, and out-degree might be the most meaningful network property for now.

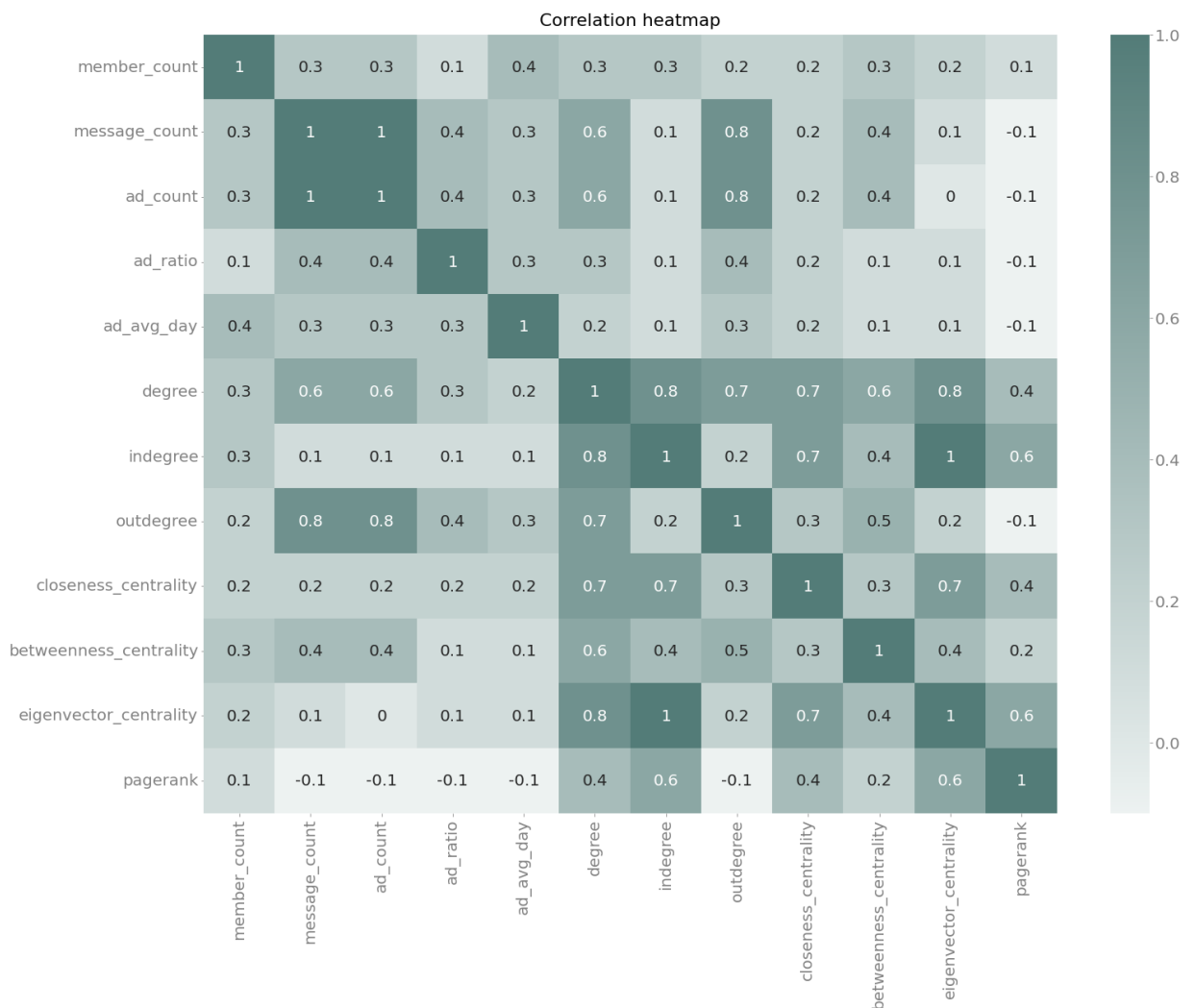


FIGURE 4.13: Correlation between network properties and several key metrics

### 4.2.3 Advertisement and member overlap

Using overlap in messages and members, we get another perspective on relations between Telegram groups in our dataset. After all, if two groups have many of the same members or messages, it is likely that there is some kind of relationship there.

Figure 4.14 displays a heatmap containing Telegram groups with at least 40% overlap in unique advertisements with one other Telegram group. In total, there are 16 groups for which this is the case. We can see some smaller groups forming, for example between cashout-59, cashout-65, and cashout-88, or between hard\_drugs-26, hard\_drugs-29, and cashout-44. Next to that, there are several pairs of groups with a big advertisement overlap. Interestingly, all of these groups are connected in the mention network, except for hard\_drugs-26 and hard\_drugs-29 and hard\_drugs-29 and cashout-44.

A big overlap in unique advertisements is likely to mean one of two things: either the groups are being used by partly the same vendor(s) sending the same messages or



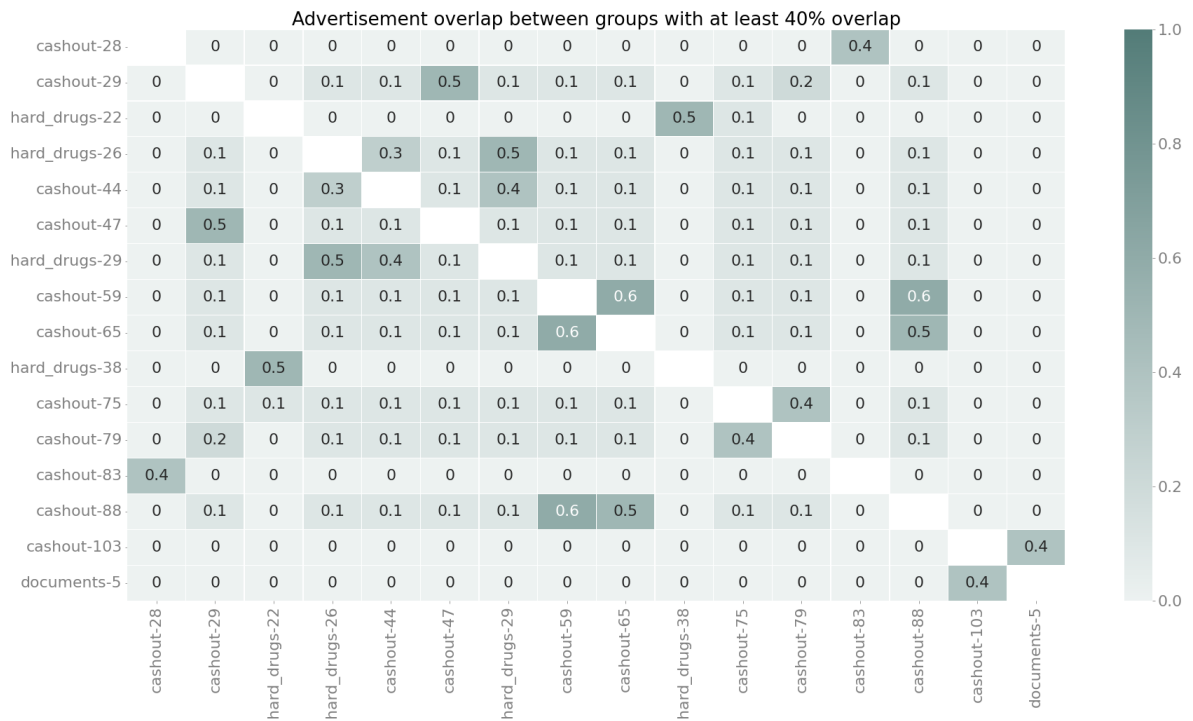


FIGURE 4.14: Advertisement overlap between Telegram groups

the groups contain very few unique advertisements and they happen to coincidentally be similar. We look at the example of `hard_drugs-26`, `hard_drugs-29`, and `cashout-44`: `hard_drugs-29` contains 21872 messages and has 3124 members; `hard_drugs-26` contains 25965 messages and has 2407 members; `cashout-44` contains 24,127 messages and has 5737 members. All three groups we suspect use an auto-delete function for their messages and all three groups have a similar name, containing the Dutch word *handel* (English: trade). Even though we could not find a mention-relation between `hard_drugs-29` and the other two groups, we suspect that there exists a relation between these three groups, given their overlap in advertisements and their very similar properties. However, there is no way to be sure of this without further research. If we look at `cashout-59`, `cashout-65`, and `cashout-88`, we see something similar: next to similar numbers of messages and members, these groups are also suspected to use the auto-delete function for their messages and two of the three groups have similar names.

Figure 4.15 displays a heatmap containing Telegram groups with at least 40% overlap in members with one other Telegram group. In total, there are 13 groups for which this is the case. Like before, we see some smaller groups forming. Most prominently, `cashout-14`, `cashout-59`, `cashout-65`, and `cashout-88` seem to all have 40% member overlap. Additionally, `cashout-9` and `cashout-10` have 40% member overlap, and both these groups also seem to have at least 20% user overlap with the group of four cashout groups mentioned before.

Both heatmaps show that it is possible to find some potential relations between Telegram groups in the dataset. In our case, especially the relation between `cashout-59`, `cashout-65`, and `cashout-88` stands out, as these groups have more than 40% overlap in both

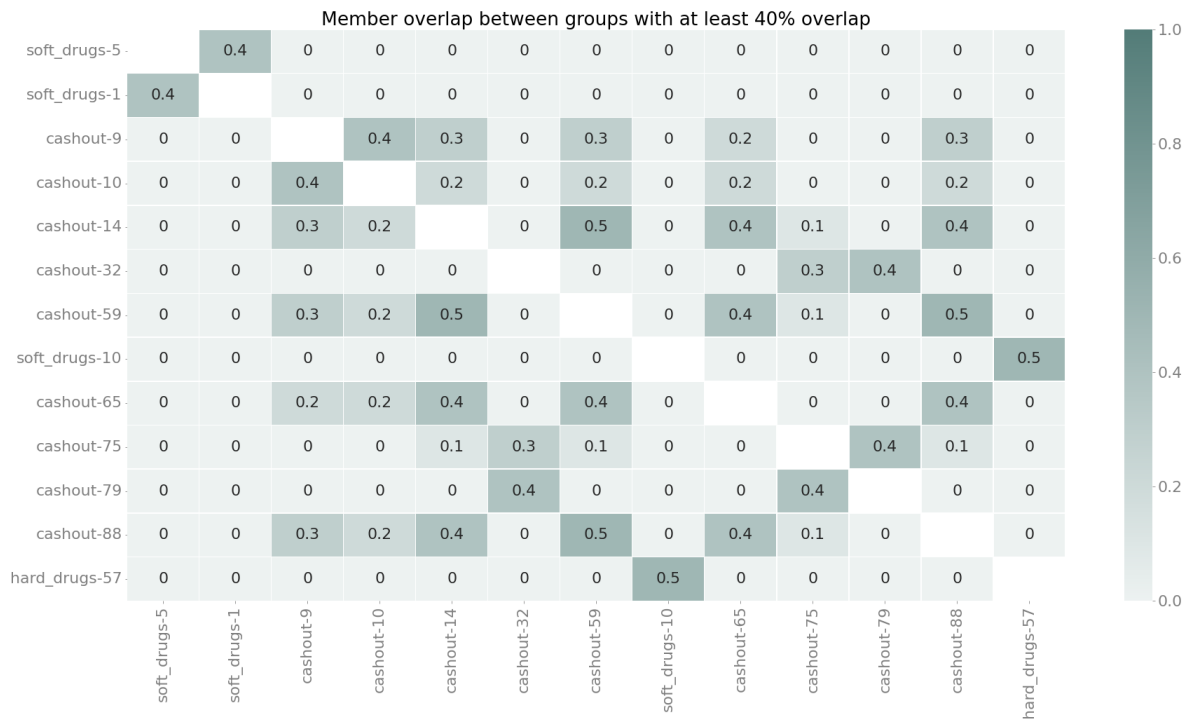


FIGURE 4.15: Member overlap between Telegram groups

advertisements and members. However, it should be noted that this method gives no definitive conclusion whether a relation between groups exists. Next to that, it stands out that there are only a few Telegram groups in our dataset that have significant overlap in advertisements and/or members (respectively 16 and 13 groups), given the density of the mention network the groups form.

#### 4.2.4 Community detection using Louvain method

We first transform the mention network into an undirected graph keeping all edges. This results in the Louvain Community Detection Algorithm detecting 8 communities. However, the graph has a very low modularity, implying that the density of connections within communities is not much higher than the density of connections between different communities. In other words, communities are likely to be hardly present. Next, we use the same method on the mention graph that only kept reciprocal edges when transformed into an undirected graph. While this method resulted in a much higher modularity, implying that there could be more clusters present in the graph, we notice that the majority of nodes in this graph are unconnected and seems to be randomly assigned to a community. Therefore, we deem this approach unsuccessful.

### 4.3 Conclusion

This chapter tries to discover information about the Telegram groups in our dataset by leveraging the information provided by the dataset. This tells us, for example, that the period of time these Telegram groups exist differs and that there are several groups that are likely to be using an auto-delete function for their messages. Next to that, we notice that Telegram groups about cashout are the most popular, followed by Telegram groups focusing on drugs and chat groups. In terms of messages, we see that groups containing the most messages are focused on the topic of cashout, followed by hard drugs and cybercrime. This leads us to believe that both cashout and drugs are topics of interest in the Dutch cybercrime landscape.

Next to that, multiple other properties of the Telegram groups were determined, which could function as a start of a group profile. In this chapter, we attempt to add an additional perspective to the group profiles by clustering the groups in the dataset based on several properties. However, both attempted methods, KMeans clustering and GMM clustering, did not produce sensible, clustered results. Chapter 7 discusses this topic. However, for now, we conclude that this approach was unsuccessful.

Moreover, this chapter introduces an approach to creating a mention network as a way of displaying mention relations between the Telegram groups. There are 29 Telegram groups in our dataset that are unconnected to any other group. However, all other groups are part of the highly connected mention network presented in this chapter. Consequently, we reason that the mention network plays an important role in the use of Telegram as a cybercrime platform in the Netherlands and we should treat it with just as much importance in the continuation of this research as the individual Telegram groups from which it is made up. While some of our attempts to use the mention network to display relations between groups seem to produce some results (for example looking at the number of mentions between groups), often the results are too vague to allow us to treat them as more than a suggestion. While we could view this as a lack of results, we prefer to see it as an extra reason to view the mention network as one entity, instead of only individual groups.

Finally, another approach to finding relationships between the Telegram groups was presented. This method is based on overlapping advertisements and members between groups. While the results suggest that this approach might provide some sensible results, we cannot treat our results as more than a suggestion, as there is no way to validate the found relationships. Nonetheless, we believe this approach could provide an additional perspective to more traditional approaches and has the potential to be an effective approach. The limitations and suggested future work are discussed in Chapter 7.

# Chapter 5

## Phase 2: Users

### 5.1 Approach

This phase dives into the users in our dataset. Just like in the previous phase, we formulate two research questions:

1. How can we create a profile for a Telegram user?
2. To what extent can we define the relations between the members of Dutch Telegram groups that are used for cybercrime?

After introducing the approach in the section that follows, the results show that users come in many shapes and sizes.

#### 5.1.1 Basic info

We create an overview of the group(s) they are members of and the number of groups they're members of. Next to that, we include in the overview if the user is flagged as a bot and the username and id of the user.

We create a DataFrame with, for each user, an overview of the groups they're active in, a count of the groups they're active in, a count of the number of messages the user has sent in each group and a count of the total messages the user has sent.

We determine the number of unique messages a user has sent overall as well as the number of unique messages the user has sent in each group. Next to that, we determine the number of messages each user has sent for each message type and in each advertisement category.

We start the process with a list of all messages that are sent in our dataset, including the user IDs of the user sending it. Messages sent in the past may have been sent by users that are not currently a member of the group. In the current step, we only keep

messages from current members. However, we use the information from messages sent by past members in another step.

#### 5.1.1.1 Roles

To get a better overview of the users, we assign roles to them based on the message types they have sent. For example, a user mainly sending messages of the type advertisement offer can be categorized as a vendor, as they mainly send advertisements offering illegal goods or services. We define the following roles and corresponding message types:

- vendor - advertisement offer
- buyer/recruiter<sup>1</sup> - advertisement request
- chatter - chat message
- bot - bot messages
- lurker - no messages

To assign the roles, we look at the message type of the messages sent by each user. The message type they've sent most determines their role. Additionally, we assign each (active) user a crime market. The crime market, or market of interest, is directly related to the advertisement category of the messages sent by the user. To determine the crime market, we look at the advertisement category in which the user has sent most messages.

It is important to note that our dataset does not allow us to distinguish the buyer and recruiter role from each other, as the distinction is very context dependent. Therefore, we call every user in this category *buyer*.

#### 5.1.1.2 Previously active users

The dataset also contains messages sent by users that were once active but are no longer members of the Telegram groups in our dataset. These users have ceased to be members of the Telegram groups in our dataset or vanished from Telegram completely. As these users are not part of the scraped users as described in Section 3.1.2, we only have their user ID. We create an overview of their composition using the same approach as for the current users, including user activity, roles, and crime markets. We cannot determine these users' membership numbers. When doing this, we must realise that this only concerns (previously) active users and that it doesn't tell us anything about the total number of members of a group at a time.

### 5.1.2 Relations

---

<sup>1</sup>recruiter is a role typically associated with money mules or (drug) couriers

As in Holt et al. (2012), we wish to create a graph that displays the relations between users in the dataset. We reason that the structure of the Telegram groups in our dataset can be compared, in part, to the forums in their research, especially since both platforms promote a many-to-many relation between the users of one group/forum. To achieve this kind of visualization between users, we use NetworkX (Hagberg et al., 2008) to create a graph structure. We add every user as a node and add an edge between two users that are a member of the same group. A graph like this should create clusters of nodes with the same Telegram group memberships.

We use this approach to create a visualization of all users in our dataset as well as multiple subsets of users, such as all active users and users in Telegram groups focussing on specific crime markets. Section 5.2.2 shows (the lack of) results of this approach. Given the large number of users, a graph such as the one we are making requires a lot of capacity. Additionally, visualizations of such a large graph can become very unclear.

### 5.1.2.1 Most-sent message

A message that is sent by multiple users might indicate a link between the users, given the nature of the advertisements. Advertisements often refer to a specific channel, group or user where other users can order the advertised product. We assume that to spread advertisements in such a manner, the users sending them must have something to gain from it, either through affiliation with the advertised channel or through financial gain. Following this train of thought, it would make sense that every user account sending the same advertisement is somehow connected to a person or group managing the advertised goods, services, Telegram group, or user account. It would not necessarily mean that all advertising accounts have a connection with each other: it could be that all advertising accounts are managed by the same person, it could be that the people running the accounts are managed by

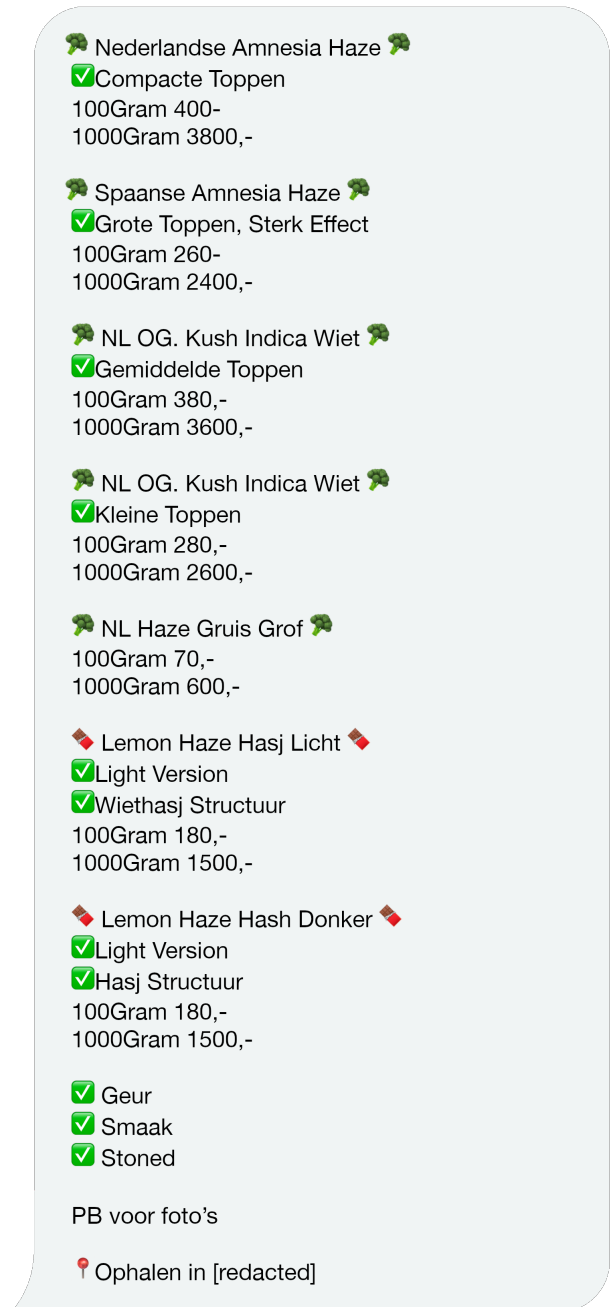


FIGURE 5.1: The most-sent advertisement in the dataset

the same person, or, for example, that the advertising accounts are part of the same organization.

To get a sense of these possible relations between users, we look at the most-sent advertisement in our dataset (Figure 5.1). We could, of course, pick any advertisement, but we reason our chances of finding relations between users based on users sending the same advertisement are highest when looking at the most occurring advertisement. Additionally, we choose to do this process with an advertisement instead of a random text message, as we reason that an advertisement is always sent with a purpose and by someone who has something to gain from it, instead of users coincidentally sending the same, common message, such as “*hey*” or “*pm me*”.

We start our process with a DataFrame containing hashed versions of all messages in our dataset. We then perform some data manipulation to obtain a DataFrame for each hashed message an overview of the users (user ID’s) that have sent the message, the number of users that has sent the message, the number of times the message was sent, the number of times each user has sent the message, the message type, and the advertisement category.

Next, we take the users that have sent the most-sent message and create a DataFrame of all (hashed versions of) advertisements ever sent by at least one of them. Then, we look through all hashed messages again to assemble all users that have ever sent one of these messages. Using NetworkX, we display this information in a graph, where each node is a user and an edge indicates that two users have sent the same advertisement.

We repeat the same process for the second-most-sent advertisement.

## 5.2 Results

### 5.2.1 Basic info

The dataset contains 146,869 unique users. Figure 5.2, displaying the distribution of these users, shows that almost two-thirds of the users in our dataset have never sent a message (in our dataset). Of these passive users, 73.2% are a member of one group, which is 48.5% of all users in our dataset. The other 26.8% of passive users are a member of more than one group, the one who is a member of the most groups having joined 88 Telegram groups in our dataset in total.

The other third of the users in our dataset are considered active users, which is 46,658 users. Of these active users, 30.8% are a member of one group, while 69.2% are a member of more than one group. Looking at activity, we see that 55.7% of the active members are active in one group, while 44.3% of active members are active in more than one group. In the active users, we see a more or less even distribution between vendors, buyers, and chatters.

Furthermore, 2.2% of the current users have been active in a Telegram group in our dataset that they are no longer a member of.

### 5.2.1.1 Bots

There are a total of 116 users flagged as bots in the dataset. 70 of these bots are members of exactly one group and 46 are members of more than one group. Of the 116 users flagged as bots, 55 have not sent a message ever (i.e. they are not active users). This makes it that 61 of the bots are active. Only 26 bots of these are currently active in more than one group. Looking at the bots that are active in more than one group, we see that the Miss Rose bot (*Miss Rose*, n.d.) has joined most groups. Miss Rose is a programmable group management bot that is often tasked with admin tasks, such as enforcing group rules by muting or banning users, sending automatic messages, blocking specified content, and even running multiple groups at once. The second most joined bot in our dataset is GroupHelpBot (*Group Help Bot*, n.d.), a bot that helps admins manage and protect their Telegram groups. The third most-joined bot is Combobot (*Combobot*, n.d.), another moderation, analytics and anti-spam bot. According to *Telegram Bots Rating* (n.d.), these three bots are added to respectively 53728, 26269, and 74675 Telegram chats

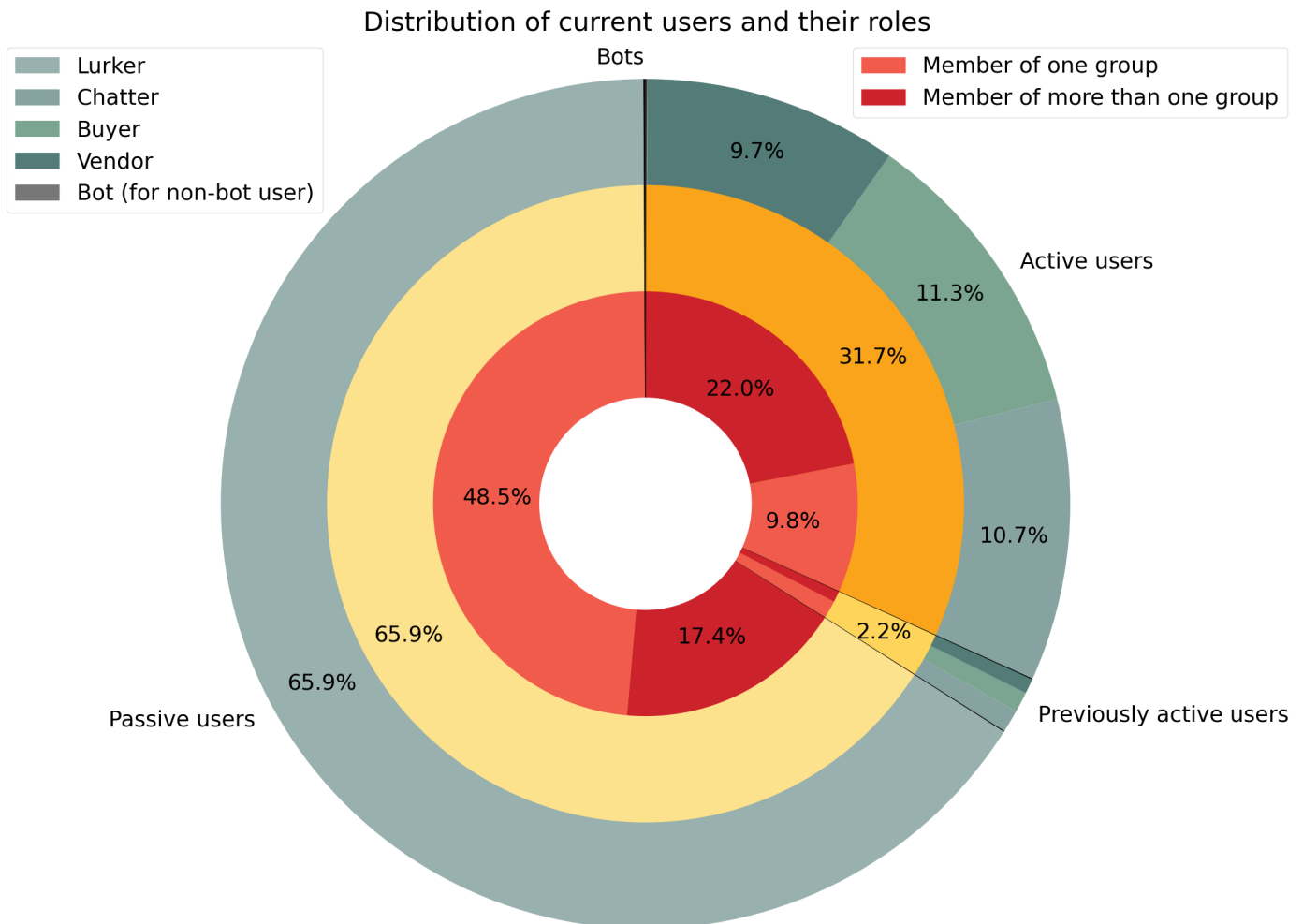


FIGURE 5.2: Distribution of current users and their roles



in total at the moment of writing.<sup>2</sup> Interestingly, these bots all seem to have websites marketing the bots, professional support, and sometimes even payment or subscription plans. For example, Combot is owned by a company and actively promotes that it is used by several well-known names in the cryptocurrency industry. Next to that, they seem to have various clones, either official or unofficial.

Looking at the names of other bots that are active in more than one group, we see that some are general bots performing admin tasks, while others are tasked with a specific assignment, such as banning bots, displaying a user's username history, removing URLs, banning Arabic scrips, hiding join messages, or removing spam.

When we look at bots that are active in one Telegram group, we see a similar combination of bots that seem to be general admin bots and bots that are tasked with a specific assignment. Next to that, we see five groups that seem to have made their own, custom bots, based on the fact that they have the same name as the Telegram group they are active in and that these bots are only active in these groups. Looking at bots that have never sent a message, there are many bots that seem to have the purpose of general admin bots or bots that are tasked with a specific assignment. However, we also see a group that has four bots with very similar names, implying that these are iterations of the same bot. Additionally, we notice several bots whose usernames seem to contain a regular first or last name, some even containing both.

### 5.2.1.2 Membership

The user that has joined most Telegram groups in our dataset is the Miss Rose bot, who has joined 150 groups. The second user that has joined most groups is currently a member of 108 groups in our dataset. Looking at this user's activity, we see that they have sent over 60000 messages spread over 95 groups. Of these messages, only 39 messages are unique and most of these messages are marked as advertisements. This makes us suspect this user is an advertisement bot.

All other users in the top 10 most joined users have joined between 91 and 98 groups of our current dataset. Seven of these follow a pattern we associate with advertisement bots: sending thousands of messages in many groups, while only a few of these messages are unique. The other user has an undefinable pattern: they have sent 50 messages over 9 groups, 36 of which are unique and most of which are classified as chat messages. It could be that this user has turned on auto-delete for all their messages, so that they disappear automatically, but it could also be that they have just sent very few messages.

When we look at the distribution of membership amongst the current users in our dataset, we see that 50% of the users in our dataset is a member of one group. Another 40% have joined between two and six groups. The last 10% has joined between 6 and 150 groups. In other words, a very big part of the users in our dataset is a member of one or a few Telegram groups and a very small part of the users in our dataset is a member of many Telegram groups.

---

<sup>2</sup>While we are unsure about the accuracy of these numbers, their sizes give an indication of the bots' popularity.

### 5.2.1.3 Activity

When we look at the most active users in terms of total messages sent, we see that the Miss Rose bot is also the most active user, having sent over half a million messages. The third most active user is also a bot; it is called Enforcerbot, it is active in ten different Telegram groups in our dataset, and it has sent almost 294000 messages in these chats. The other eight users in the top 10 most active users (i.e. in the number of messages sent) follow the pattern we associate with advertisement bots: a high number of total messages sent in combination with a very low number of unique messages.

Figure 5.2 shows that almost two-thirds of the users in our dataset are passive users.

The dataset also contains users that have been active in a Telegram group that they have left since that moment and have not been active in the groups they are currently a member of. This is the case for 3278 users, or 2.2% of current users in our dataset (Figure 5.2).

### 5.2.1.4 Previously active members

There are 73226 users that we classify as previously active members (i.e. users that have been active but are not members of a Telegram group in our dataset anymore). Manual inspection of several groups shows that a user deleting their account without leaving a Telegram group first is a common occurrence. This can, for example, be seen in the Telegram Desktop app when the sender of a message is named “deleted user” or when “deleted user” has joined or left a group. When this happens, no Message Service is created, as would be the case when a user leaves the group. However, we also notice that there exist several bots that hide the MessageServices about users joining or leaving a group. Therefore, not every group contains evidence of this phenomenon. About the reason why users delete their accounts, we can only speculate. Next to that, there are also users who leave a group or are banned by a group’s admin (or admin bot) for violating group rules. It would be fairly easy for these users to abandon the Telegram account for which this happened and create a new one.

Figure 5.3 shows the distribution of previously active users, their roles and their crime market of interest. We see that 32.5% of users focused on cashout and that cybercrime second most popular crime market amongst previously active users. Next to that, it stands out that drugs are more popular amongst vendors and fireworks amongst buyers. Moreover, 24.4% of the previously active users were chatters who sent only chat messages.

Looking at the activity of the ten most active previously active users (in the total number of messages sent), we see the familiar pattern we associate with advertisement bots; thousands of sent messages over many groups, few of which are unique. For example, the most active user sent 138565 messages over 65 groups, but only 142 of these messages are unique. The other nine users sent between 44134 and 62694 messages of which between 25 and 100 are unique messages ranging from 13 to 69 groups.

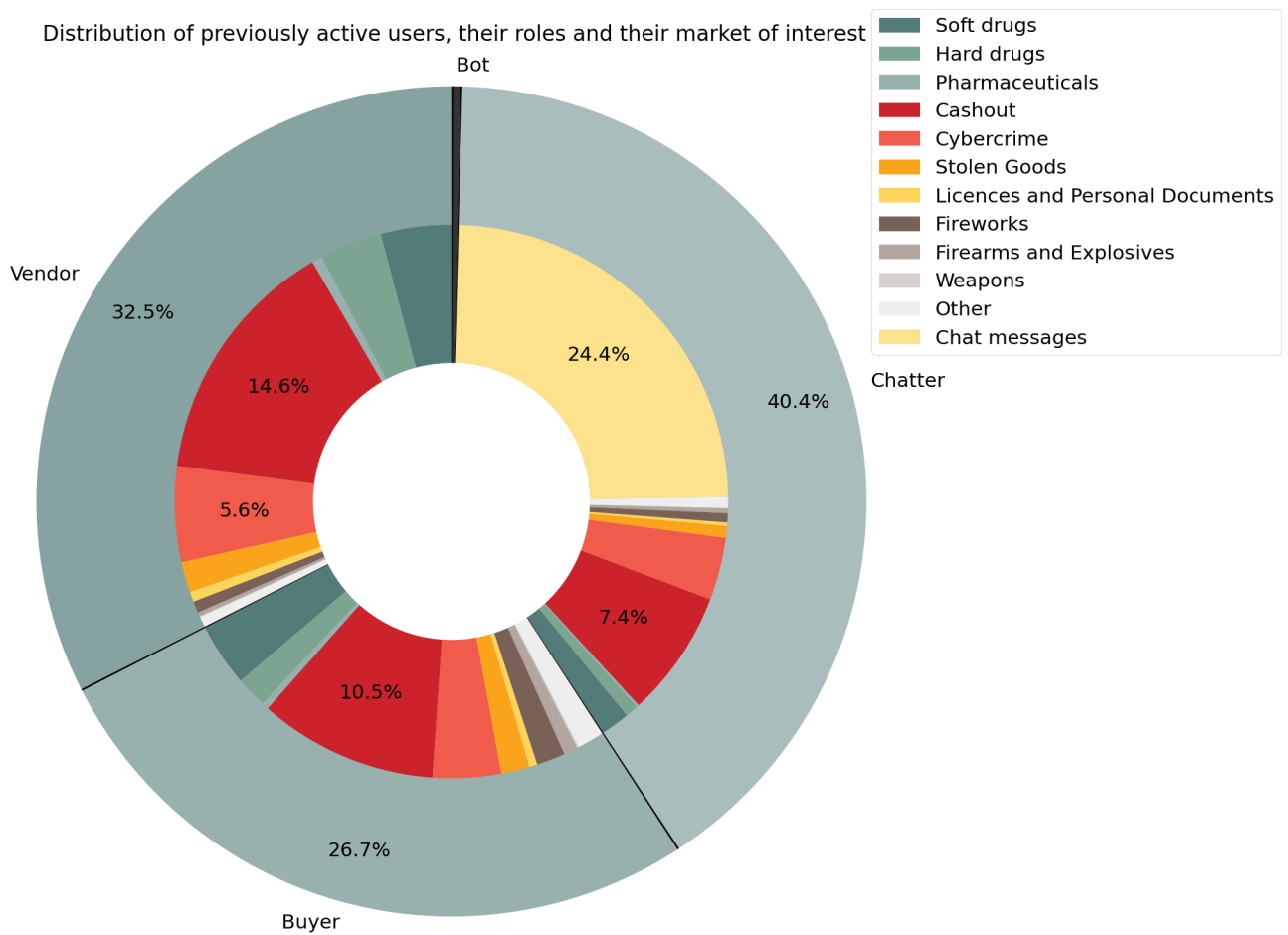


FIGURE 5.3: Distribution of current users and their roles

Looking at all previously active users, we see that 50% are active in 1 group. Another 40% are active in 2 to 6 groups and the last 10% are active in 6 to 77 groups. Next to that, we see that 20% of these users have sent 1 message, 50% have sent between 2 and 13 messages, 20% have sent between 13 and 94 messages, and 10% have sent between 94 and 138565 messages.

### 5.2.2 Relations

We first look into the graphs displaying relations between active users. Figure 5.4 displays the relations between active users in Telegram groups focusing on the pharmaceuticals crime market, consisting of 4 Telegram groups. In this graph, each node is an active user, each edge is a relation between two users that are members of the same group, and the nodes are coloured by a group they are active in. The graph has 159 nodes and 6242 edges in total. We see that many of the active users are active in only one Telegram group in this crime market. However, there is a small cluster of users that are active in two groups.

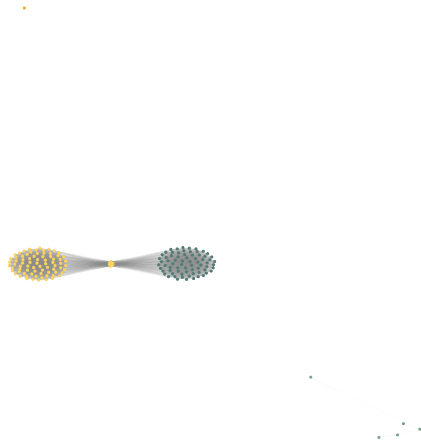


FIGURE 5.4: Network of active members in *pharmaceuticals* crime market

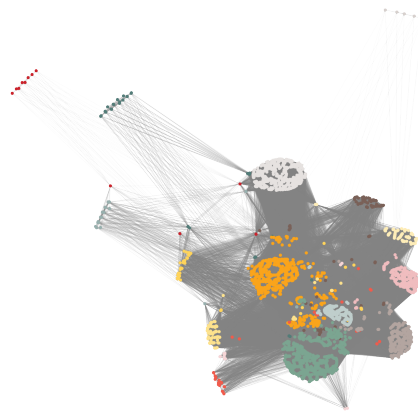


FIGURE 5.5: Network of active members in *fireworks* crime market

Figure 5.5 shows a visualization of the relations between active members in the fireworks crime market. This crime market is serviced by 16 Telegram groups in our dataset and contains 7736 active users with 9.7 million edges between them. While we can see clusters of active users forming, the graph in itself becomes too unclear to draw any conclusions.

The graph visualizing the relations between active members in the pharmaceuticals crime market took a little over 10 seconds to create, while the graph visualizing the relations between active members in the fireworks crime market takes at least 8.5 minutes to form.<sup>3</sup> The same approach was used to create a graph of all active users in the dataset and of the active users in several larger crime markets, such as cashout, soft drugs, or hard drugs. However, the results could not be created due to memory issues.

### 5.2.2.1 Message overlap

The most-sent message in our dataset can be seen in Figure 5.1. It is an advertisement offering large(r) quantities of hash or weed that can be picked up in Amsterdam. It is sent 311392 times in our dataset. The message is sent by 13 users in 87 Telegram groups in our dataset.

Figure 5.6 shows a timeline of the period between which the message was sent in each group. We see that the first appearance of the message happened on the same day in many groups. There also seems to be a second moment it appeared in several other groups at the same time. At the top of this graph, we see a few groups in which the message only seems to be sent recently. However, some of these groups are suspected to use an auto-delete function to delete their messages after a certain number of days (see Section 4.2.1), so we do not know when the message was first sent in these groups.

<sup>3</sup>even though using different hardware would result in different times, the relative increase gives a good indication of feasibility.

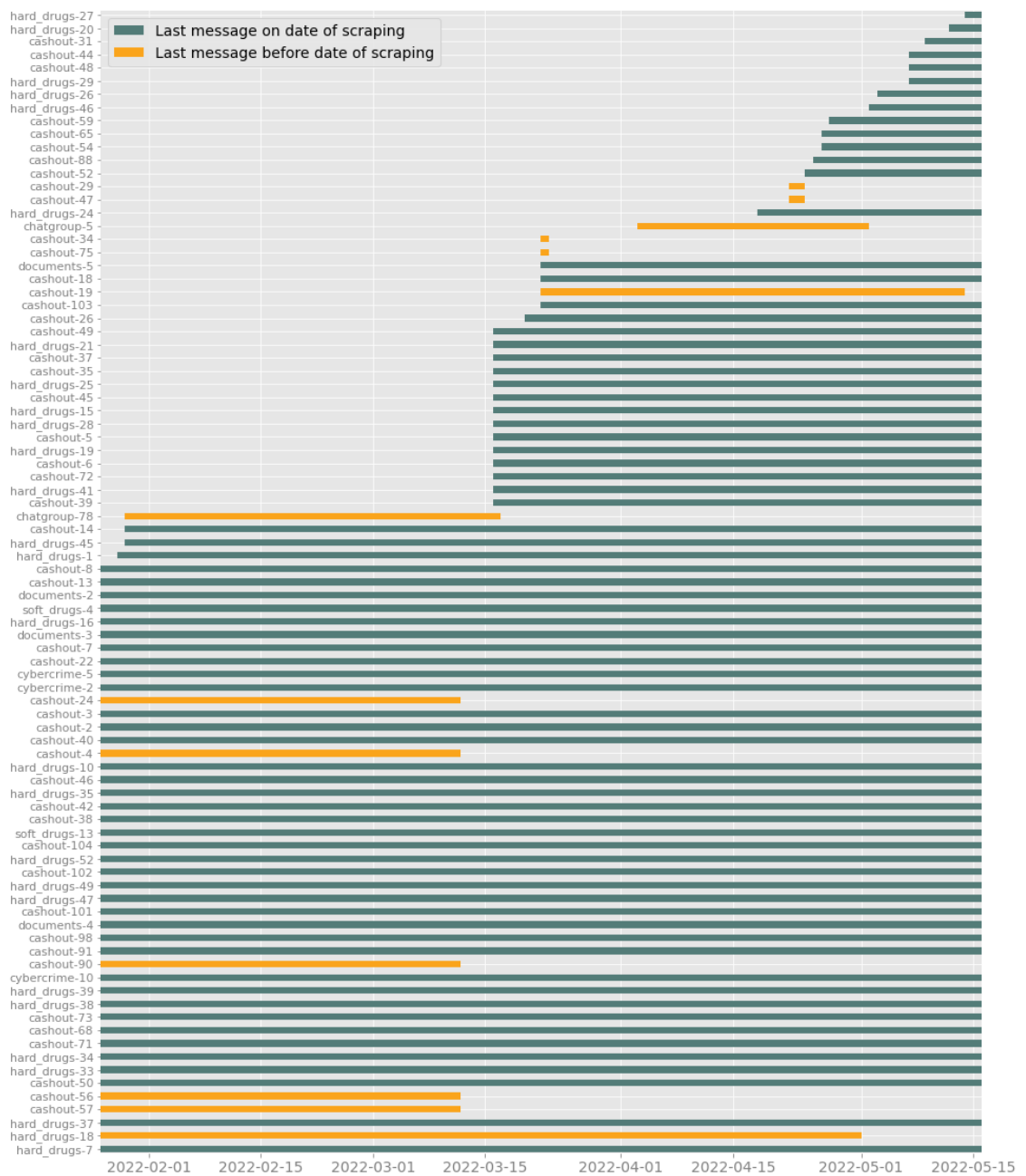


FIGURE 5.6: Timeline of the period in which the most-sent message was sent in different Telegram groups

Furthermore, we see that in most groups the message was still being sent on the day we scraped them. However, the short(er) timelines of some groups imply that a user stopped sending the message in these groups. We suspect the message might have been marked as spam, causing the user(s) sending it to be muted or banned.

Looking at the distribution of the number of times the users have sent the message in each of the groups (Figure 5.7), a clear pattern becomes visible. 6 users have sent the message in many of the groups and in most of these groups, they seem to have sent a

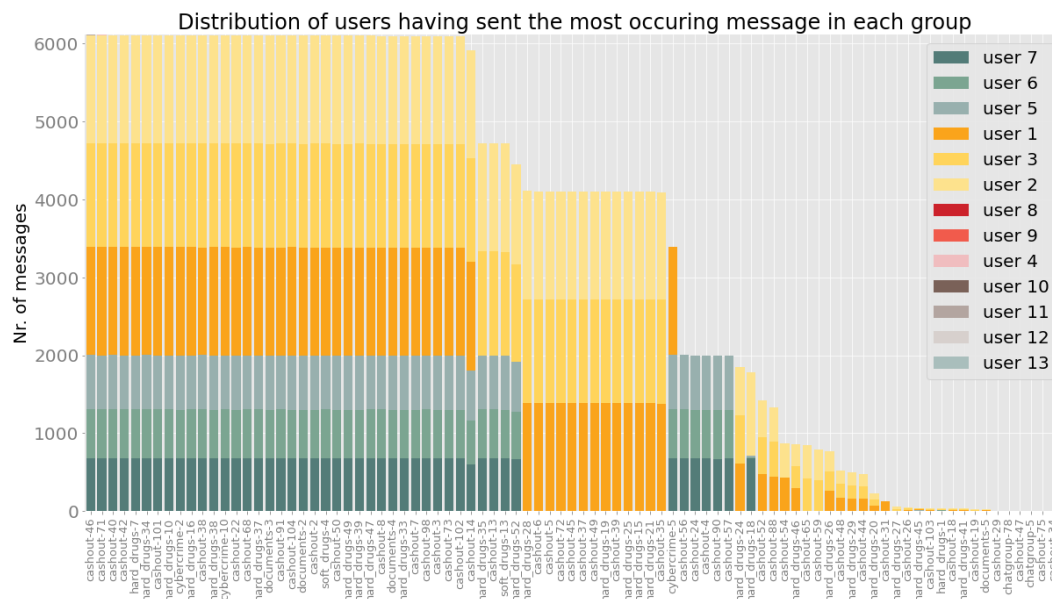


FIGURE 5.7: Distribution of users that have sent the most-sent message in different groups

similar amount. In Figure 5.8, displaying the number of times the users sent the message each day, we see the same pattern of 6 users sending the message over a thousand times each day. However, this figure also shows that user5, user6, and user7 were most active between the end of February and March 13th, while user1, user2, and user3 started their activity three days later and were still sending the message on our cut-off date.

The next step is to take a closer look at the users that have sent the message. In total, 13 different users have sent the message, 11 of which were group members of at least one group at the moment of scraping. There are a few things that stand out:

- User1, user2, and user3 have the same username, only to be distinguished by a number added after the name. They have all sent a little more than 71000 messages in total, of which 1 is unique (i.e. they have only sent the message we are currently looking at). These messages are divided over 74, 73, and 67 groups respectively.
- user5, user6, and user7 all have related usernames, which are also related to the usernames of user1, user2, and user3. Each of these users has sent between 104000 and 106200 messages in total. Of these, respectively 13, 15 and 15 are unique, and they were sent in respectively 49, 48 and 51 groups. Most of the unique messages (all except one) are classified as advertisement offer in the category soft drugs.
- The other users seem to have no correlated pattern when manually inspected. They consist of:
  - User11 has sent slightly more than 15000 messages, of which 114 are unique. This user is classified as offering mainly hard drugs, but also soft drugs and cashout and licenses and personal documents.

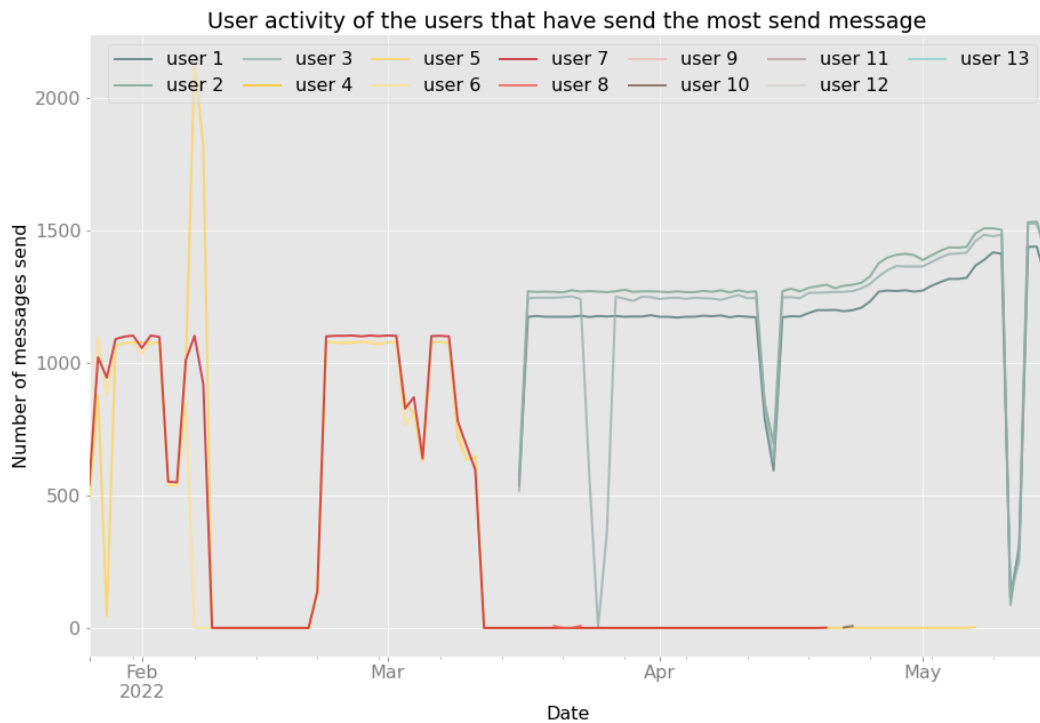


FIGURE 5.8: Activity of users sending the most-sent message over time

- User8 has sent over 43000 messages, 173 of which are unique. They are classified as advertising for cashout.
- User9 has sent almost 16000 messages, 560 of which are unique. They seem to be active mainly in groups whose name indicates that they are meant for *fgame* practises, including fraud and cashout. This user is classified as advertising cashout (about an equal rate of offers and requests, which makes sense for cashout).
- User13 has sent 9 messages (4 unique) in 2 groups and is classified as advertising cashout and soft drugs.
- User4 has sent 91 messages (54 unique) in 9 groups. They are classified as buying soft drugs.
- User10 and user12 are not members of a Telegram group in our dataset anymore. However, when we look at their activity history, we see that user10 has sent 13 messages in total (3 unique) and seems to focus on selling soft drugs. User12 has sent 47 messages (38 unique) and seems to have focused mostly on buying and selling hard drugs, cashout, stolen goods, fireworks, and soft drugs.

The last step in this approach is to look if we can find other users in our dataset that have sent the same advertisement as one of the users above. In other words, we look if we can find clusters of users that have sent the same advertisement, just like the 13 users that have sent the most-sent advertisements, only we take these 13 users as a starting

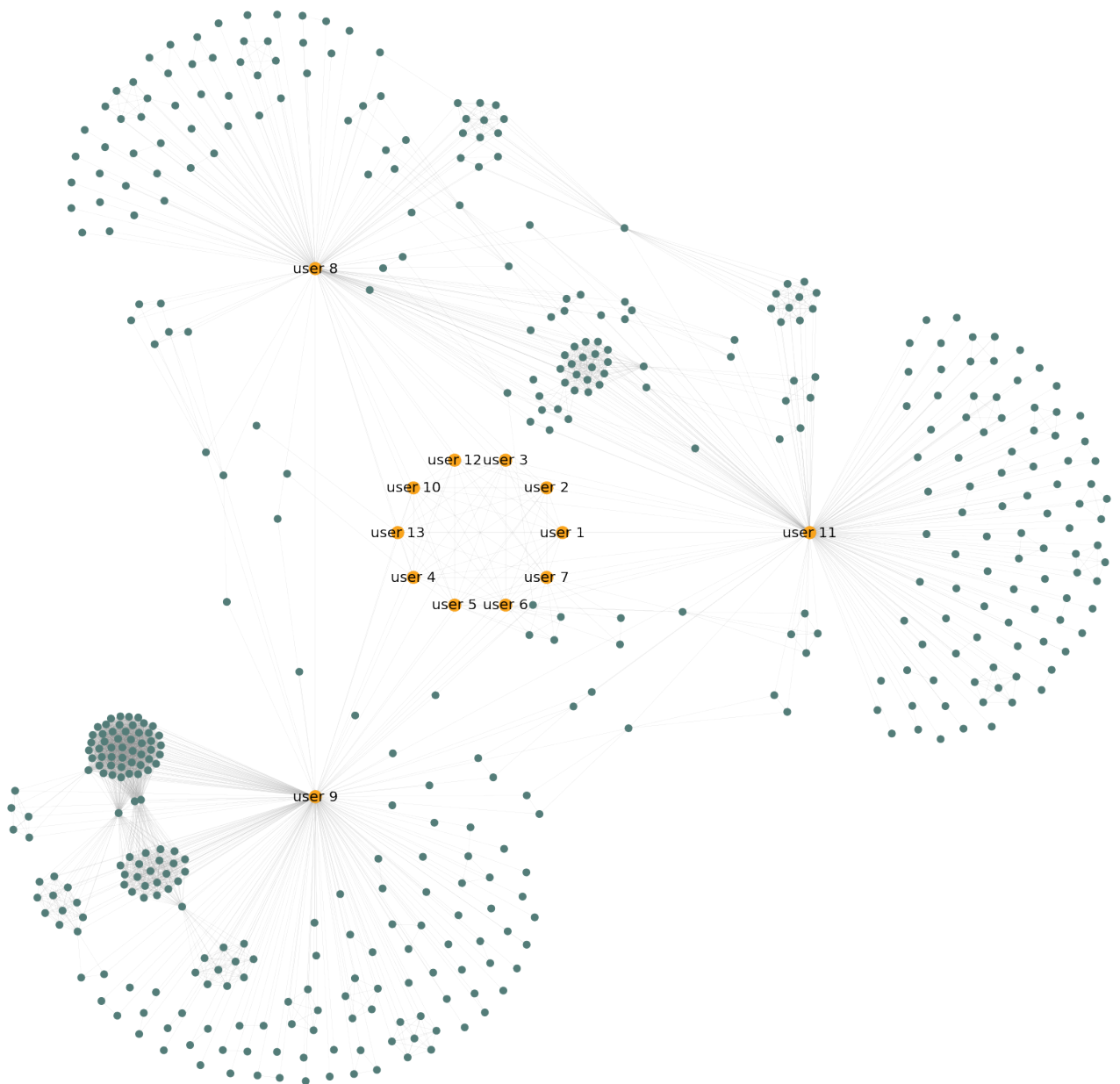


FIGURE 5.9: Relations between users that have sent the same advertisement. The yellow nodes represent users that have sent the most-sent advertisement in the dataset.

point. Figure 5.9 displays the result of this effort. Each node is a user and an edge is drawn between two users when they have sent the same advertisement. The yellow nodes indicate users that have sent the most-sent advertisement. We have removed one message labelled as an advertisement manually (and with that a user cluster), as including it drastically decreased the readability of the graph and since one could question whether the message is really an advertisement, since the message was "Pb me", the Dutch translation of "Pm me".



Message	Nr. of users that sent the message	Nr. of times the message was send
Bunq?	54	122
Wie heeft kaarten	27	46
🌈VEEL WERK VOOR BUNQ!!!! Wie heeft nu een bunq account met betaal verkeer?? 📄, ik zet tussen de €8000,- en €12000,-📄. Op afstand werken is bespreekbaar👉. Heb je nog geen bunq account maar wil je wel één aanmaken is binnen 5 minuten geregeld. Stuur me een bericht als je één hebt of interesse hebt.👉. (vandaag nog je geld💎👉)	13	15619
Wie kan swipe	12	12
Ben jij 18+ en wil je vandaag nog verdienen tussen de 9000/15000 euro 📄bericht me nu om een vivid account te maken is 5 min werk. Vandaag nog geld✅!!!	8	4816
Bunq inlog?	7	52
Money Makers: !! mensen die snel geld willen verdienen dit is je kans!! Je hoeft geen pinpas ofzo te brengen✅ Vandaag nog geld!!!!!! Niet denken gewoon doen👉 Cc me nu✅!! !!mensen die snel geld willen verdienen dit is je kans!!	6	992
MELD ALLE KAARTAS IN ME CHAT! ! ALLES IS AFSTAAN ! 📄LATEN WE VERDIENEN, NIET BERICHTEN ALS JE NIET KAN AFSTAAN!👉⚡	6	1556

TABLE 5.1: Advertisements send by user9 and at least 5 other users

The graph shows that 8 of the users seem to have 0 to 5 edges. However, user8, user9, and user11 stand out by having many edges connected to them. In other words, these users have sent many advertisements that other users have also sent. User9, for example, has also sent the advertisements in Table 5.1, next to the most-sent advertisement displayed in Figure 5.1. We can identify the clusters of users that have sent each message in Table 5.1 in Figure 5.9. For example, the biggest cluster of users connected to user9 has sent the message “*Bunq?*”, asking for Bunq bank cards or accounts, quite possibly to use these accounts for money laundering. Table 5.1 shows us that this user has sent 8 advertisements that are sent by at least 5 other users. Each of these advertisements falls in the cashout category, three of which ask for Bunq cards or accounts specifically. Next to that, one of the advertisement asks for a Vivid account, two of the advertisements ask for cards, and the others do not specify the kind of cards or bank accounts they ask.

Another example is the large cluster of users that is connected to both user8 and user11. These 21 users have all sent the message displayed in Figure 5.10, offering electronics such as mobile phones, laptops, and gaming devices. This advertisement is sent over 61000 times, with two users contributing for the majority of these messages. User8 has sent this advertisement 5 times and user11 has sent it 10 times. Most of the other users have sent it only once, with a few users sending it more than 3 times. Interestingly, the advertisement does not include a way to contact the vendor, implying that the reader might need to contact the sender directly. One can only guess why the advertisement is



FIGURE 5.10: Example of an advertisement



FIGURE 5.11: Example of an advertisement

only send once by so many users; perhaps someone was trying to create an advertisement bot, but it took several tries to let the bot send the message more than once.

As a third example, we add a cluster of 4 users connected to user11; these users have all sent the advertisement in Figure 5.11. User11 has sent this advertisement 9 times in our dataset, while the other four users have sent this advertisement between 1100 and 2600 times. In total, this advertisement was sent almost 8000 times over 34 Telegram groups in our dataset. The advertisement includes references to a Telegram account or channel, a Wickr account, and a Signal account as contact information.

While one could argue that the first advertisement in Table 5.1 (i.e. "Bunq?") is too general to assume there exists a relationship between the users sending it, the table

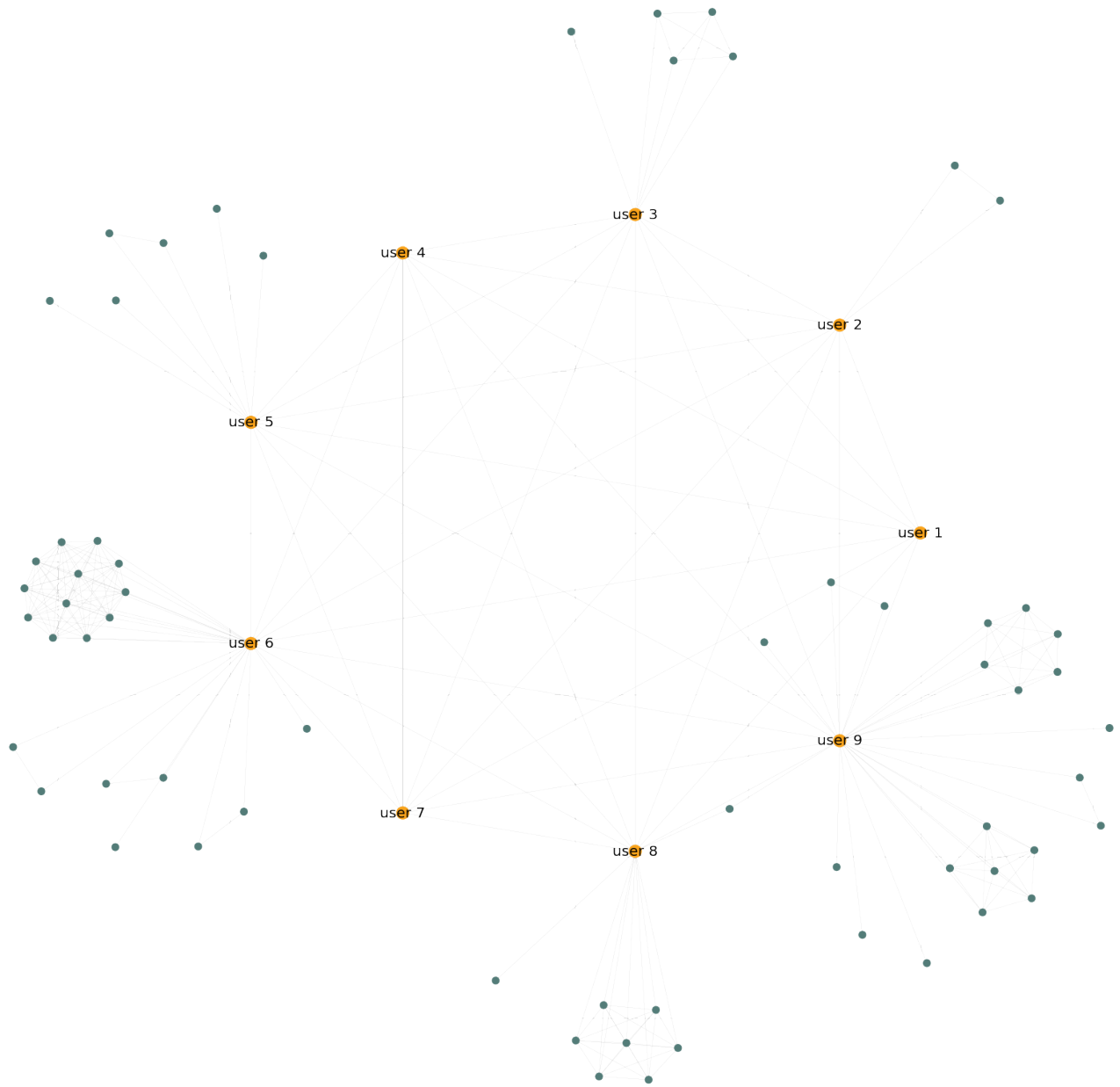


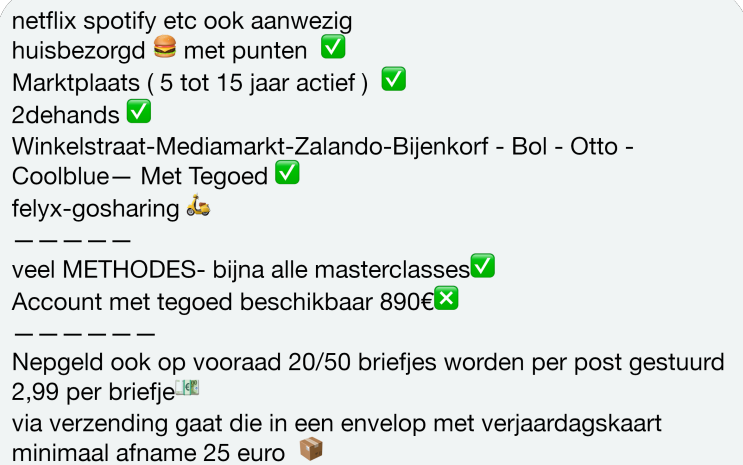
FIGURE 5.12: Relations between users that have sent the same advertisement. The yellow nodes represent users that have sent the second-most-sent advertisement in the dataset.

contains several other examples that should make our reasoning clear. It is hard to imagine users sending these advertisements without benefitting from it, so we assume users benefit in some way by sending the advertisements. A logical way for the users to benefit could be that they are selling their own product or service by advertising it, or that they are working for someone that pays them to send the advertisements. Of course, we cannot determine if groups of users sending the same advertisements are being

controlled by one or multiple persons. Next to that, it is impossible to deduce if user accounts are linked to a single vendor or a whole organization. However, our reasoning continues that, if users are sending the same advertisement, there should be some kind of relation present, whether it is one person running multiple Telegram accounts or multiple people working for the same vendor/organization, and whether it is because they work for the vendor or because they are the vendor. Therefore, a graph such as Figure 5.9 is likely to indicate relationships between users.

To create the graph in Figure 5.12 we use the same approach used to create Figure 5.9, but instead of starting with the users that have send the most-sent advertisement, we start with users that have sent the second-most-sent advertisement in the dataset. Like in Figure 5.9, we manually remove one advertisement, as it decreases the readability of the graph and as we could debate whether it is correctly classified as an advertisement (i.e. the same advertisement as in Figure 5.9). The resulting graph contains 72 nodes and 230 edges. Nine of these nodes display users that have sent the second-most-sent advertisement in the dataset. These users are labeled and given a yellow colour in Figure 5.12.

Comparing the two graphs, we see that the graph in Figure 5.12 is a lot smaller than the graph in Figure 5.9. Nonetheless, several clusters of users that sent the same advertisement are still present. An interesting difference between the two graphs is that Figure 5.9 contains many users that have sent overlapping advertisements with several different users, i.e. there are many users that belong to multiple clusters, while Figure 5.12 contains only one user for which this is the case.



netflix spotify etc ook aanwezig  
 huisbezorgd 🍌 met punten ✓  
 Marktplaats ( 5 tot 15 jaar actief ) ✓  
 2dehands ✓  
 Winkelstraat-Mediamarkt-Zalando-Bijenkorf - Bol - Otto -  
 Coolblue— Met Tegoed ✓  
 felyx-gosharing 🤖  
 -----  
 veel METHODES- bijna alle masterclasses ✓  
 Account met tegoed beschikbaar 890€ ✗  
 -----  
 Np geld ook op voorraad 20/50 briefjes worden per post gestuurd  
 2,99 per briefje 📄  
 via verzending gaat die in een envelop met verjaardagskaart  
 minimaal afname 25 euro 📦

FIGURE 5.13: The second-most-sent advertisement in the dataset

### 5.3 Conclusion

This chapter provides insights into the basic characteristics of the Telegram users in our dataset. We show that almost two-thirds of the users in the dataset are passive users, with around half of all users being passive users that are a member of only one group. The other users have sent at least one message in our dataset, making them active users. The active users are divided roughly evenly over the three roles we defined: chatters, buyers, and vendors. The most active users in the network are bot users used for administrative tasks. We also notice less active bots used for other tasks, such as preventing spam or banning Arabic script. Another category of users we observe within the most active users are

---

accounts we believe are advertisement bots. Their behaviour is characterised by sending a very large amount of messages, only a few of which are unique. The classifier labels often agree that these users are sending advertisements. Additionally, we note a considerable number of previously active users, indicating a potential high account turnover or frequent account switching among individuals.

Although a traditional relationship graph leveraging the many-to-many relations that exist between (active) users in a Telegram group did not yield significant results due to capacity problems, we introduced an alternative approach based on different users sending the same advertisements. Illustrating this approach with the most-sent and second-most-sent advertisements in our dataset, we revealed a network of users who posted the exact same advertisement messages, suggesting a relation between the users for which this is the case. Furthermore, we explored some additional information that could be used to extend the profiles of Telegram groups and users.

## Chapter 6

# Phase 3: Influence

Phase 3 looks into the influence of Telegram groups and users. However, instead of trying to quantify it - which is a daunting task - we explore some simple options to try and identify the more influential groups and users in our dataset taking into account the context of this study. The question is: what makes a Telegram group or user more influential or more interesting than others? Following the objectives of the research (see Section 1.1, we aim to approach this question from the perspective of two important parties that we imagine can benefit from a study such as the current work: law enforcement and academics. While we imagine law enforcement to be more interested in practical applications and academics in the theoretical side, we believe both aspects are intertwined and equally important.

Trying to determine what makes a Telegram group or user important or interesting, we ask ourselves what is impacted by cybercrime on Telegram. From a law enforcement perspective - tasked with keeping society safe - we would say cybercrime impacts society. For example, [Junger, Veldkamp, and Koning \(2022\)](#) found that 41.7% of Dutch people of 16 years and older have experienced at least one attempt to fraud, 70% of which took place online. [E. R. Leukfeldt and Roks \(2021\)](#) adds another perspective by concluding that traditional street-crime uses Internet and social media as a new and additional opportunity. As we see that cybercrime and its societal impact is growing, we add financial impact as an additional area of impact. The reasoning is simple (and perhaps oversimplified): the more illegal goods and services are sold, the more money is made through crime, keeping the illegal economy in place. The money made by criminals is likely to stay in illegal circuits a little longer, as they need to find a way to launder it without raising suspicion ([E. R. Leukfeldt et al., 2020](#); [Weber & Kruisbergen, 2019](#)). The academic perspective identifies the same areas of impact. However, this perspective can also help us understand why people commit crimes and study the effect of crime on individuals and society. Research about cybercrime on Telegram can help develop methods for combating or preventing crime.

While the societal impact is difficult to measure for our dataset, we reason that an individual or group acting as a vendor is more likely to have an impact on society with their criminal activities than an individual who buys a small amount of illegal goods or

products for personal use. For example, an organization selling drugs throughout the Netherlands would have more societal impact than an individual buying a small amount of drugs for personal use. The same reasoning works for financial impact; while we are unable to determine the financial impact of Telegram groups and users in our dataset, it makes sense that active vendors create more financial impact than an individual buying something for personal use. Therefore, we take social and financial impact into account when trying to determine the more important, influential or interesting Telegram groups and users in our dataset.

The rest of this phase is divided as follows. First, we detail the approach used to find the more influential Telegram groups in the dataset, using a group's basic metrics and its place in the mention network amongst others. Next, we discuss the approach taken to identify influential users, both in the whole dataset as well as in single Telegram groups. Finally, we highlight the results.

All code used for this chapter is written in Python 3 (Van Rossum & Drake, 2009) and runs in a Jupyter Notebook (*Project Jupyter*, n.d.). Like in the previous chapters, Pandas (McKinney & Others, 2010) and Numpy (Harris et al., 2020) are used as supporting libraries. Visualizations are created using Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017).

## 6.1 Groups

This subheading focuses on determining the importance or influence of Telegram groups in our dataset. Specifically, we create a method that allows us to pick several Telegram groups from our dataset that are likely to be of influence in the Dutch cybercrime landscape on Telegram. To do this, we first look at the individual group properties determined in Chapter 4. Then, we combine this with the groups' places in the mention networks, as determined in Section 4.2.2. Finally, we obtain some information about the groups, as to get an idea of their properties.

### 6.1.1 Influence based on group properties

Societal impact is a broad concept, but we try to boil it down to be applicable to the Telegram groups in our dataset. Looking at the group properties determined in Chapter 4, we see societal impact come back in the number of members a group has. More specifically, the more members a Telegram group has, the more users are reached by the content of the group, and the more likely it is that more real-life persons know about and have access to the crime in the group. Additionally, we assume that vendors would want to advertise in a Telegram group with more members for the same reason; they reach more potential customers the more members a Telegram group has. One could argue that people in massive Telegram groups or Telegram groups with a lot of spam are more likely to skip over the messages without reading them, but that's something that requires more research.

Next to that, it intuitively makes sense that a Telegram group with a lot of activity (e.g. measured in the number of messages sent in a group) is more likely to impact or influence the Dutch cybercrime landscape than a group with little activity. There are two things to keep in mind, however. First of all, in our dataset almost all evidence of cybercrime is found in advertisements, a subset of all messages, so we need to ask ourselves if we want to look at messages in general or at advertisements specifically. Secondly, our dataset contains groups that have existed for five years, groups that have existed for five days and everything in between. The number of messages metric is therefore highly influenced by the length of a group's existence. Additionally, the number of messages sent in a Telegram group in the past does not necessarily tell us anything about the future development of this group and its impact or influence. In other words, this metric may tell us something that is old news by the moment we notice it and, therefore, needs a sidenote when using it to determine the groups that are more influential.

To combat the possible skewness and inaccuracies that come with using the total number of messages in a Telegram group as a measure of its impact, we take a look at advertisements instead of all messages. As most evidence of crime is found in advertisements for illegal goods and services in the Telegram groups in our dataset, it makes sense to take advertisements into account when trying to determine a group's influence. We create three metrics using advertisements: the total number of advertisements, the ratio of advertisements against messages, and the average number of advertisements per day. Each metric, its meaning, and its advantages and drawbacks are explained in the following paragraphs.

The total number of advertisements says something about the absolute amount of crime that might happen in a Telegram group, making it a useful metric for comparing a group's impact. However, this metric is likely to be influenced by groups that have existed longer and favours groups with a high number of messages over groups with a low number of messages. Therefore, while we believe this metric can be a way of determining a group's impact, the metric by itself makes it difficult to draw conclusions about a group's impact over others.

The ratio of advertisements vs. messages can also be seen as the percentage of messages that are advertisements. Intuitively, it makes sense that a low ratio means that there is less advertisement for criminal goods and services in a group's messages and a high ratio means that there is more advertisement for criminal goods and services in a group's messages. More advertisements for criminal goods and services do not necessarily mean that a group has more financial or societal impact, but chances are higher that it is the case for a group with a high ratio when compared with a group with a low ratio. An additional advantage of this ratio is that it should be less sensitive to a high number of total messages in a group than the total number of advertisements discussed in the previous paragraph. However, the ratio of advertisements vs. messages is skewed by groups with a very low number of messages (e.g. if a group with ten messages contains nine advertisements, the ratio will be 0.9), while reason deems us to say a group with few messages is of smaller influence than a group with many messages.



The average number of advertisements per day seems to be a good indication of the recent number of advertisements sent in a group, making it an interesting metric for a group's recent relevance. The metric is not directly dependent on the length of a group's existence or the total number of messages sent in a group. The drawback is that this metric actually favours groups with a very short existence, preferably a day or shorter.

### 6.1.2 Influence of groups in the mention network

We believe the mention network plays an important role in what makes Telegram an interesting cybercrime platform. One reason for this is that Section 4.2.2 has shown the mention network plays an important part in the functioning of the network as a whole, with different Telegram groups taking on different roles and (suspected) alliances being displayed. The mention-relation network is so highly and intentionally connected (i.e. looking at the evidence of subcommunities forming between groups and relations between users) that we should assume the mention network is, at least partly, an interconnected platform for cybercrime instead of each Telegram group being a separate platform. In other words, it seems that the mention network does not just happen to exist, the users and moderators have purposely created and maintained it, which means we must treat it as such. Therefore, we use the basic graph of the mention relations between the Telegram groups in our dataset created in Section 4.2.2 to look at the groups' influence or to determine the more interesting groups. Specifically, we use the groups' network properties.

Section 4.2.2 reasons that it is difficult to apply the same meaning several studies on OSNs use for the researched network properties to the mention network created in that section, as the nature of the mention network is different from the nature of a traditional OSN. The reasoning in Section 4.2.2 continues that the meaning of degree, in-degree, and out-degree can be explained for the mention network. Simply put, Telegram groups with a high degree are what keep the mention network so highly connected. Taking a closer look at the meaning of in-degree in the mention network, we see that groups with a high in-degree are mentioned by many different groups. They are likely to be actively promoted and are likely to be known by many Telegram users. Telegram groups with a high out-degree, on the other hand, can be seen more as advertisement hubs, for a high out-degree means these groups have mentioned many other Telegram groups in the network. This also means that members of Telegram groups with a high out-degree are likely to have heard of many other Telegram groups in the network. Following this line of reasoning, we would assume that both in-degree and out-degree can point to Telegram groups that are important for keeping the mention network functioning as a platform. Another advantage is that there exists a correlation between degree and the other network properties found in Section 4.2.2, which can make it somewhat of a representative property. Therefore, we use degree (i.e. in-degree and out-degree combined) as a metric for determining what Telegram groups are influential or important in the mention network.

### 6.1.3 Approach

We create an overview of the top 10% of groups in each of the metrics pertaining to individual groups mentioned in Section 6.1.1. To avoid the mentioned pitfalls, we first apply the following rules to filter some Telegram groups from the results:

- The number of days a Telegram group has existed needs to be more than 0 (i.e. they could be just a few hours old). This is to avoid dividing by zero when obtaining the average number of advertisements per day for a group.
- The total number of messages in a Telegram group needs to be at least 100. Reasonably, it makes sense that a group with very few messages is unlikely to influence the Dutch cybercrime landscape. However, the ratio of advertisements against messages (i.e. the percentage of messages that are labelled as advertisements) also favours Telegram groups with a small number of messages.
- The total number of advertisements in a Telegram group needs to be at least 100 ads. Reason deems us to think a group with less than 100 ads wouldn't have much impact.

Implementing the rules described above some Telegram groups from our useful dataset in this step of the research. Therefore, the top 10% might consist of 25 groups instead of 29 from this point on. We then take combinations of the top 10% highest scoring Telegram groups in each of the metrics and create Venn diagrams using Matplotlib. We reason that we do not need all metrics determined in Section 6.1.1, as some are derived from each other. We reason that the total number of members, the average number of advertisements per day, and the degree of a group (see Section 6.1.2) are independent metrics that represent a simple form of importance. The correlation heatmap created in Section 4.2.2 is used to visualize the correlation between the metrics in order to verify that the metrics are indeed independent.<sup>1</sup> We take the three metrics to create a definitive Venn diagram to find important, interesting, or influential groups in the mention network, using the top 20% highest scoring groups for each metric.

Finally, we gather some additional information about the resulting important or influential groups, such as the message-sending pattern over time and the crime market(s) that their content focuses on.

### 6.1.4 Results - groups

#### 6.1.4.1 Advertisements

The Venn diagram in Figure 6.1 displays the overlap between Telegram groups in our dataset when we take the top 10% of each of the metrics described in Section 6.1.3.

---

<sup>1</sup>Even though correlation does not imply causation, we reason that correlation between two metrics would make one of them obsolete when looking for metrics that can imply importance.

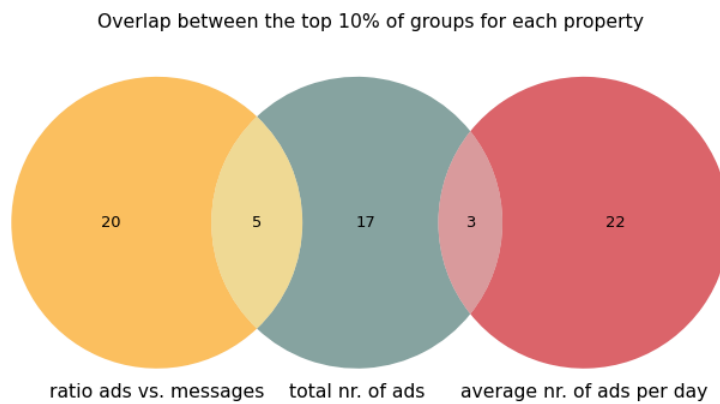


FIGURE 6.1: Venn diagram of Telegram groups that are in the top 10% highest scoring groups in the advertisement properties

We see that there are five groups that are in the top 10% of the highest ratio of advertisements against all messages and also in the top 10% of the highest total number of advertisements. Next to that, there are 3 groups that are in the top 10% of the average number of advertisements per day and in the top 10% of the total number of advertisements. However, there is no group that falls in the highest top 10% of all three metrics.

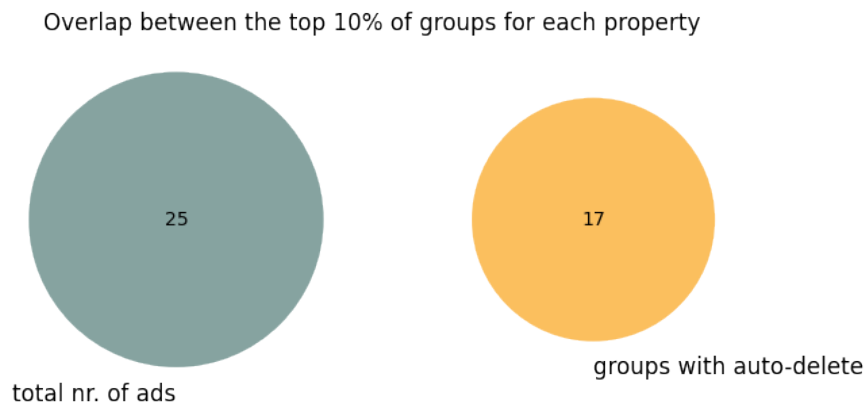


FIGURE 6.2: Venn diagram of Telegram groups that are in the top 10% highest number of total advertisements and groups that we suspect use an auto-delete function for their messages

In Section 4.2, we noted our strong suspicion that there were several Telegram groups in our dataset that use an auto-deletion function for their messages. Under this assumption, we deem it likely that these groups score low in the total number of advertisements metric (i.e. low enough to not make the top 10%). Figure 6.2, displaying a Venn diagram of the overlap between the top 10% of Telegram groups in the number of advertisements and groups using auto-delete, confirms this suspicion. Therefore, if we use the total number of advertisements metric to determine the influence of a Telegram group, we are very likely to exclude the Telegram groups using the suspected auto-delete function, while logic would deem us to say that Telegram groups using auto-delete do not necessarily

have a low societal or financial impact. Figure 6.3 shows a Venn diagram with the groups using auto-delete and the other two advertisement metrics. The figure tells us that all Telegram groups in our dataset that we suspect to use the auto-delete function score in the top 10% of either one of the other two metrics.

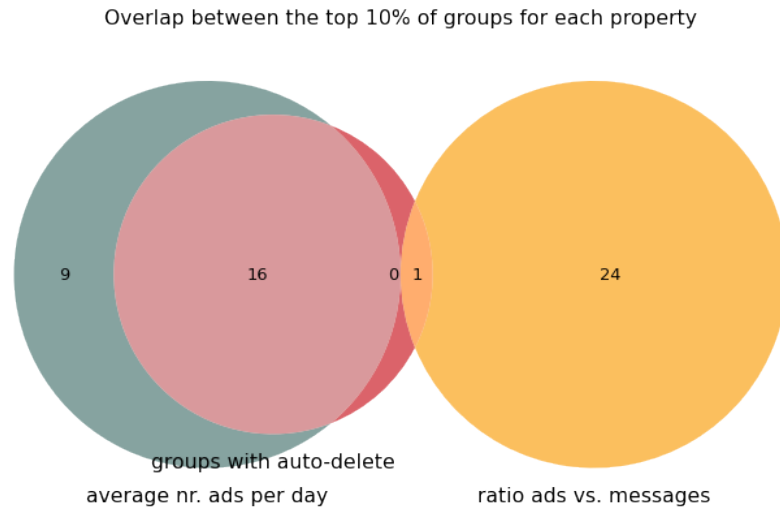


FIGURE 6.3: Venn diagram of Telegram groups that are in the top 10% of average advertisements per day and ratio of advertisements in messages and groups that we suspect use an auto-delete function for their messages

#### 6.1.4.2 Correlation

As mentioned in Section 6.1.3, we use the correlation heatmap in Figure 4.13 created in Section 4.2.2 to examine the correlation between the different properties and to verify whether the chosen properties are independent from each other. The first thing that stands out is that there is almost no correlation between the number of members a Telegram group has and its other properties. Secondly, we see that degree has a slight correlation with the total number of messages and advertisements in a group. This is caused by the out-degree having an even higher correlation with these metrics, which can be explained by the fact that more messages are statistically likely to contain more mentions. Finally, the three key metrics (i.e. the number of group members, the average number of advertisements per day, and the degree of a group in the network) seem to have almost no correlation with each other.

#### 6.1.4.3 Influential groups

Figure 6.4 shows a Venn diagram of the high-scoring Telegram groups when sorting them by the number of members, the average number of advertisements per day, and their degree in the mention network. We look at the top 20% of groups in each metric, as that gives us a balanced number of important or interesting groups.

Next, we take a look at the Telegram groups that are deemed important in the three categories. Figure 6.5 shows the growth of the groups over time measured in the number

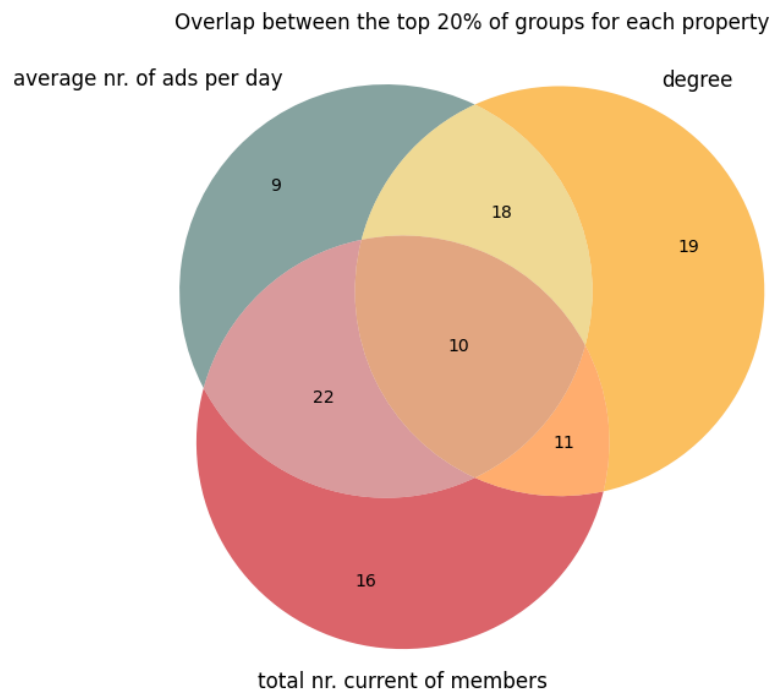


FIGURE 6.4: Venn diagram of Telegram groups that are in the top 20% of the key metrics

of messages in the group. We can see that there is a distinction between groups that have existed for at least a year following a natural growth pattern and groups that have existed for shorter than half a year following a very steep growth pattern. At least two of the groups that have existed for a short time (i.e. cashout-88 and cashout-65) are suspected to use an auto-delete function for their messages, as detailed in Section 4.2.1.

Figure 6.6 shows us the crime markets the advertisements in these groups cater to. At first glance, all groups seem to focus primarily on cashout, drugs, and cybercrime. However, the critical reader may notice that cashout-88 and cashout-65 have a very similar distribution. This may be a coincidence but as Section 4.2.3 points out, these groups have a high advertisement and member overlap, next to being connected in the mention network and having similar names and properties. All of this together leads us to suspect that these groups might be maintained by the same (group of) people. Next to that, the distributions displayed in Figure 6.6 also leads us to think that cashout, cybercrime, drugs, licences and personal documents, and to some extent stolen goods are amongst the most popular publicly accessible criminal markets, while fireworks, firearms and explosives, and weapons seem amongst the least popular markets. Whether this is because weapons, firearms and explosives, and fireworks scare people enough to not want to talk about them on public forums, whether they are indeed less popular, or whether they are just not very prominent overall, we do not know. However, it seems as if these crime markets do not influence the Dutch cybercrime landscape on Telegram as much as the others.

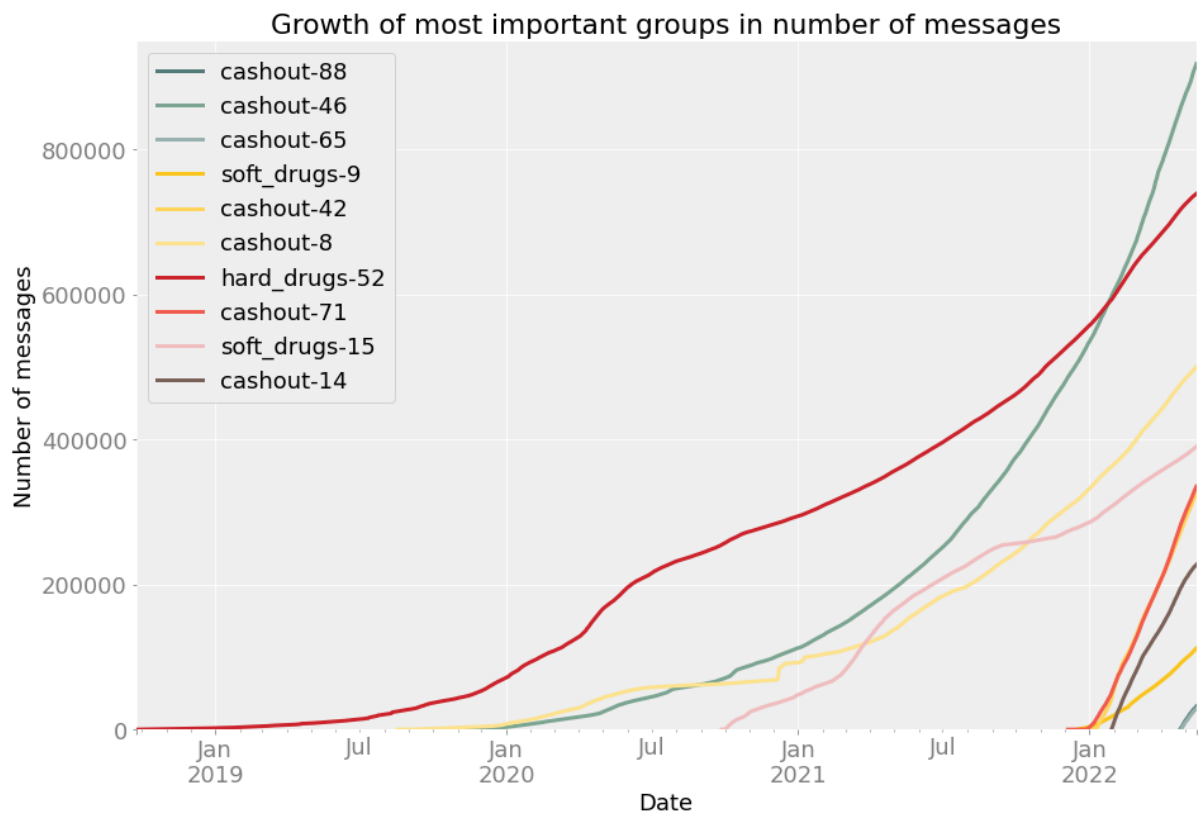


FIGURE 6.5: Growth in number of messages of the important groups over time

## 6.2 Users

First, we determine what makes a user important or interesting and then we determine how we can identify important users in the network. We first look into important users in the network as a whole, after which we dive into identifying users within single Telegram groups.

### 6.2.1 Important users in the network

What comes to mind first is to look at the number of messages users have sent in the groups in the dataset. As discussed in Chapter 2, [Zamani et al. \(2019\)](#) have looked at the level of user activity measured as the number of comments on forum threads. They compare forums on the public web, the semi-dark web, and the dark web. Interestingly, they found that active users are much more active on the semi-dark forums than on the public web, and are even more active on the dark web. [Zamani et al. \(2019\)](#): “*This phenomenon, in the dark web, is due to the high activity of a few users whose aim is to sell or advertise some illegal products, and in 8chan, it might reflect some kind of political fanaticism.*”. As our Telegram groups share some characteristics with forum threads and we know that there are many groups used for the advertisement of illegal

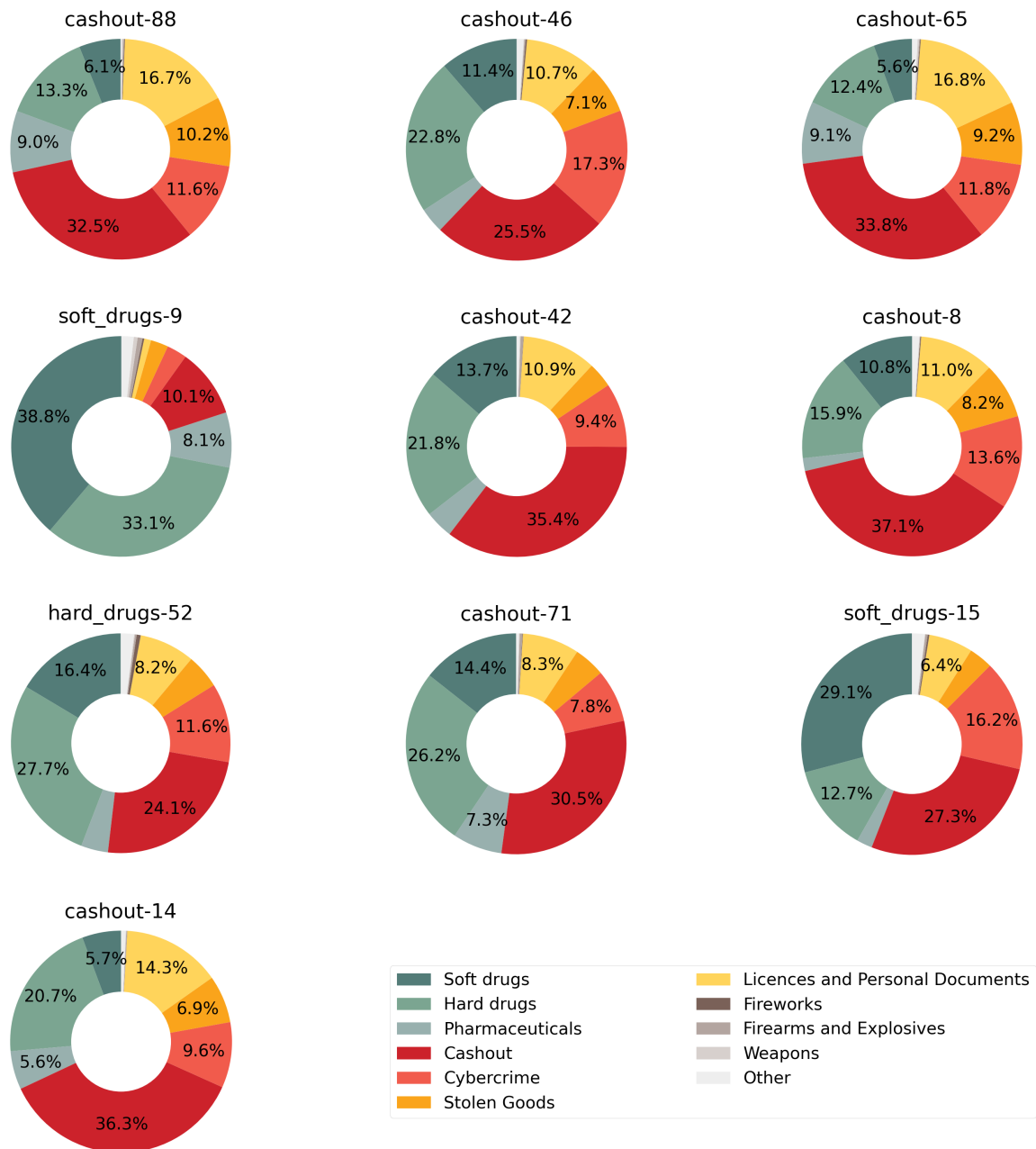


FIGURE 6.6: Important groups: division of messages over crime markets

products, their claim seems to support our idea of looking at active users when trying to find important users (that advertise illegal goods).

We see that the number of messages in the dataset is much larger than the number of messages that are sent by the users in the dataset. The difference can be explained for a large part by the (we assume) large number of users that have deleted their accounts without leaving a group. This phenomenon is seen in different groups where the number of active users throughout the history of the group is higher than the number of total current users. Manual inspection of these groups in the Telegram desktop app showed us messages that said deleted user has entered the chat or displayed a message sent

by deleted user (Section 5.2.1.4). Additionally, a small part of the messages is sent by channels. While it would certainly be an interesting approach to dive deeper into the activity of deleted users, we will not do that in this research. Therefore, we reason we won't compare the activity of the users in our dataset to the total number of messages in our dataset, but only to the number of messages sent by users in our dataset. Another argument for this is that we want to determine the importance of the users in the current dataset, which is per definition relative to the importance of other users in the dataset. It is interesting to note that the dataset being as it is can skew the perception we create of active users in the dataset, because the dataset allows for a lot of historical data. Chapter 7 discusses this in more detail.

At the beginning of this chapter (Section 6.1.1), we determined that the more members a group has, the bigger the audience for a message is. In other words, we reason that a user who sends messages in bigger groups has a bigger audience than a user who sends the same number of messages in smaller groups. If we can find a way to define the audience size a user has for their messages, we can create a potential additional metric that can help us identify interesting or influential users. We must, however, take into account that the number of members a group has is very dynamic (for example demonstrated by the large number of previously active users in Section 5.2.1.4) and changes over time, so a message sent a few months ago might not have the same reach as a message sent right before we scraped the groups. However, at the moment of writing, the dataset does not contain this information.

Section 5.1.2.1 presented a method of defining relations between users that sent the same advertisement. Creating such a network and extending it with the number of times an advertisement is sent or the number of different groups it is sent in could give an indication of what users are interesting to keep tabs on. However, we argue that this method would be more appropriate to detect users working together or being run by the same person(s) in the sense that one could consider the sum of the influence of the users in these alliances to have more impact than the influence of each user separately. Another possibility would be to create a graph as discussed in Section 5.2.2 and determine the degree of each user. We could reason that users with a high degree are more influential in the network, but we also argue that this graph only consists of active members and we could question how influential a vendor is when they (potentially) know of many other vendors because they advertise in the same Telegram groups.

## 6.2.2 Important users within a single Telegram group

Similar to determining what users are influential in the network, we propose using the number of messages sent by a user to determine what users are most influential within a single Telegram group. Simply put, we argue that the more messages a user has sent in a Telegram group, the more they build a presence and the more likely it is many of the members in the group have seen their messages or know of their existence. This is the case in both big and small groups. However, unlike the previous section (Section 6.2.1), we see no sense in trying to determine the reach of a user, as this should be more or less similar for users within the same Telegram group. To only exception to this is a case



where a user was active in a Telegram group when it had fewer members and has not been active since, while it did gain more members. Moreover, given the many-to-many relations between all users in a group, we cannot use relations between users in a single Telegram group to distinguish influential users. There is, however, a case to be made to include users' roles in this, as we can argue that vendors or buyers would be of more interest to most parties than chatters.

Next to that, we argue that there is another group of users that can play an influential role in a single Telegram group: the promoters of the group. The reasoning behind this is that users promoting a Telegram group in another Telegram group play an important role in maintaining or expanding the mention network forming the platform.

### 6.2.3 Approach

First, we create a concentration curve by dividing the active users in the dataset into bins based on their activity measured in the number of messages sent. We do the same for users that are active in groups they were members of at the moment of scraping. For both of these metrics, it follows that the higher the number of messages a user has sent, the more important the user would be.

To determine the reach of a user, we simply sum the number of members of each group in which the user has sent a message. In theory, this is a straightforward manner of determining the number of users that have potentially read the message. However, this does not take into account the fact that we only know the number of members a group had at the moment of scraping, which is not necessarily the same as the number of group members at the moment a message was sent. We add an additional metric we call potential reach to each user. This is the sum of the number of group members of the groups of which the user is a member. We reason that a user could have reached a maximum of that amount of users if they were to send a message in all groups they are a member of.

Next to that, we determine the influential users in single Telegram groups by sorting the users in each group by the number of messages they have sent in the group. While it would be relatively easy to do this for all groups in the dataset, we only apply this method to the ten groups determined in Section 6.1.4.3. To determine which users are the promoters of these Telegram groups, we look at the messages containing mentions, which were assembled in Section 4.1.2. This overview contains the group that was mentioned, the group in which the message was sent, the number of times this exact message was sent in this group and the user that sent it. Using this overview, we filter only the groups for which we want to know the promoters.

Finally, we try to find out more about the influential users in our dataset. We look at one user as an example and for this, we take the most active user that is not an administrator bot. To get more insight into this user's behaviour, we create a graph of their activity over time and a timeline of their activity in the Telegram groups they are active in. Next to that, we take a look at the messages this user has sent.

Distribution of cumulative percentage of active users and the cumulative percentage of messages they have sent

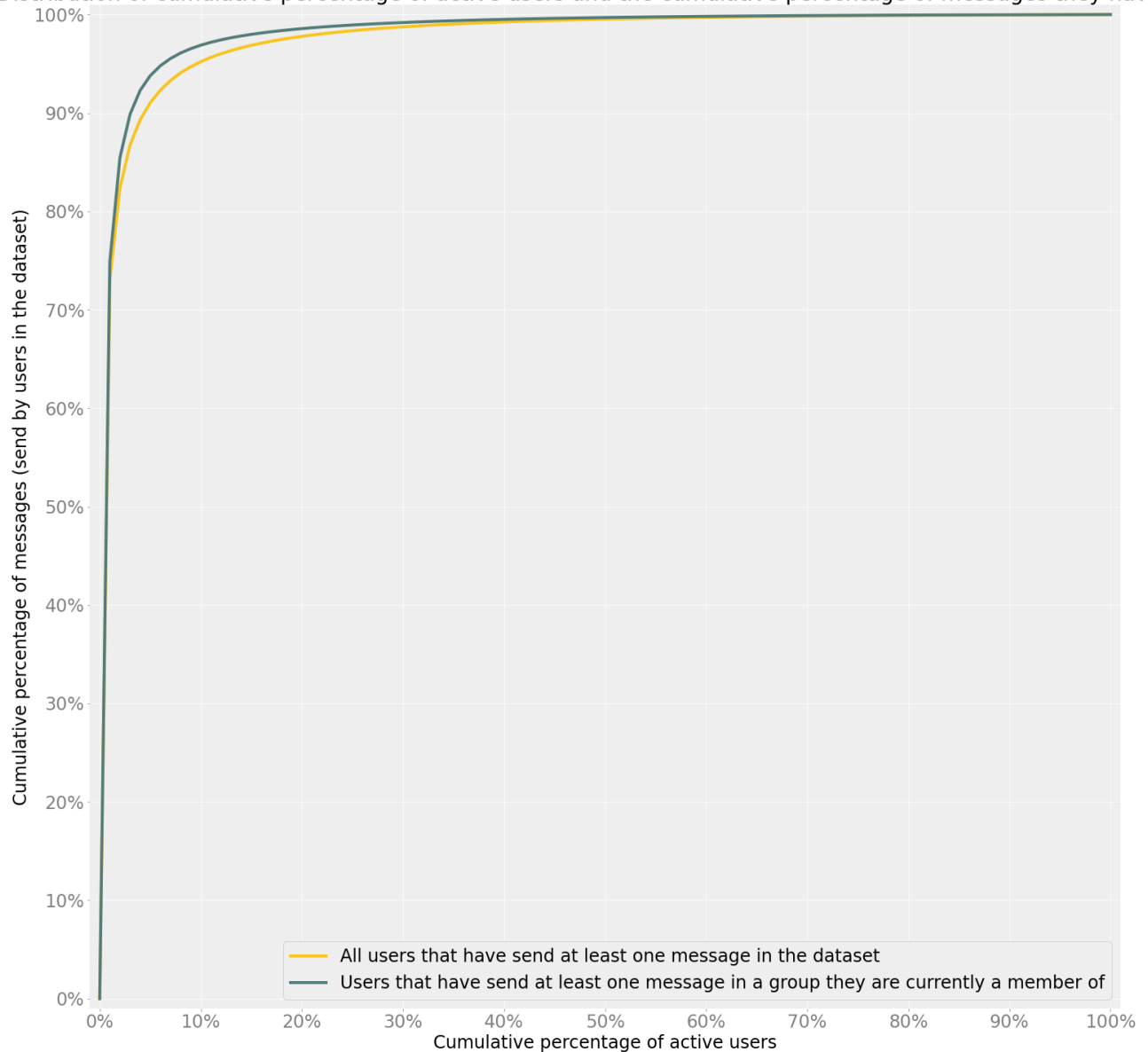


FIGURE 6.7: Cumulative activity over population

#### 6.2.4 Results

The dataset contains just over 27 million messages, of which almost 19 million are sent by almost 50000 active users that are currently still members of at least one group in the dataset. Figure 6.7 displays the distribution of current users and the cumulative percentage of messages they are responsible for. We see that the 1% most active users, i.e. the 500 most active users, are responsible for almost 75% of messages in the dataset and the 10% most active users are responsible for almost 97% of the messages in the dataset. Next to that, the same was done for all users in the dataset, including users that are not currently a member of one of the Telegram groups in the dataset anymore.

There, the top 1% of most active users are responsible for 95% of the messages and the top 1% for almost 73

The ten users with the biggest reach follow the pattern we associate with advertisement bots; they sent thousands of messages in many different Telegram groups, while only a few of them are unique. There is one exception: the Miss Rose bot also makes the top ten, which is to be expected since this bot is used as an administrator in 128 Telegram groups in the dataset.

Looking at potential reach, it gets a little more diverse. In the top ten users with the biggest potential reach, we see five users that have sent very few unique messages and at least 1500 total messages, which leads us to suspect that these accounts are used as advertisement accounts. However, it is unclear whether the advertising is done automatically or manually. For example, we can imagine a user manually posts an advertisement in as many groups as possible by copying a template, or it could also be that the user accounts are recently created and that it is an automated advertisement bot. Next to that, there are two users classified as a lurker in the top ten (i.e. they have not sent a message). Furthermore, the last two users both have usernames that imply that someone is using the accounts for promoting or running a business, but they have sent very few messages.

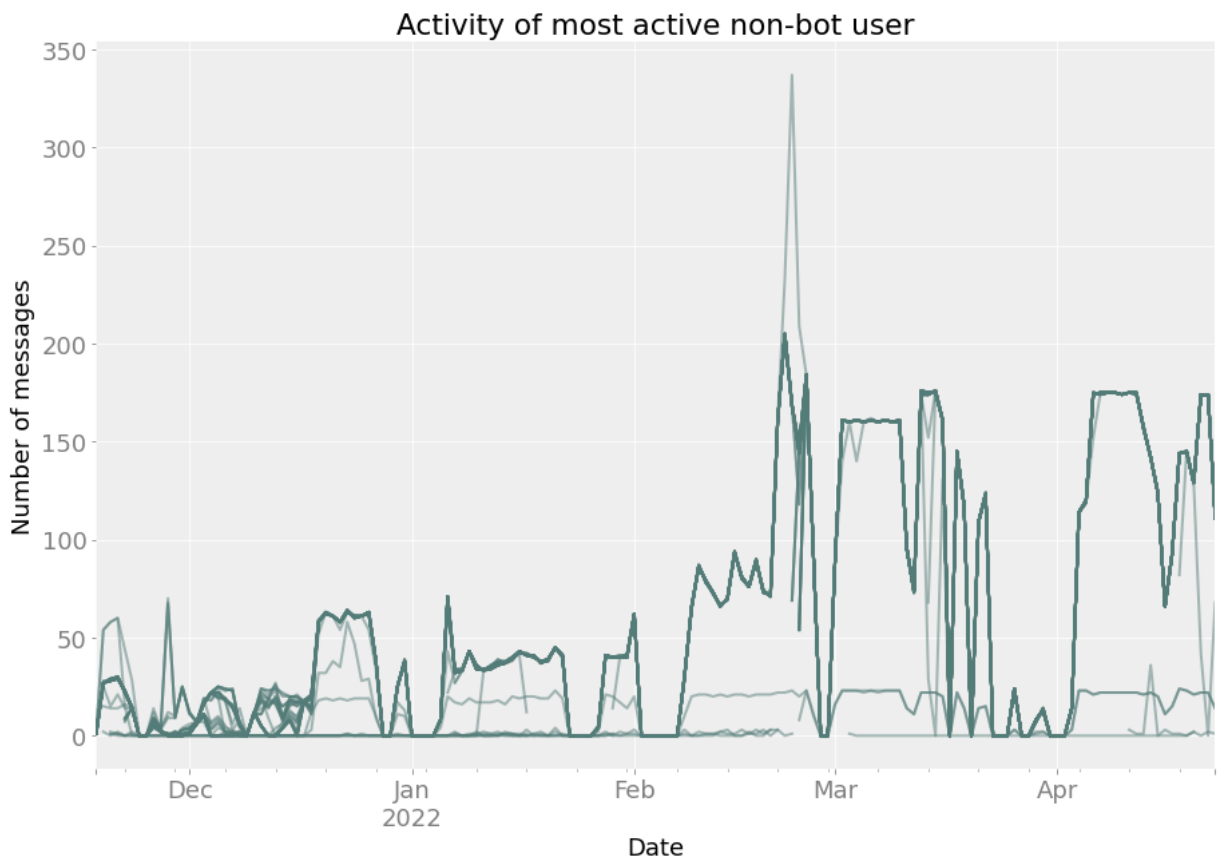


FIGURE 6.8: Activity of most active user over time, where each line is a Telegram group they are active in

Looking at the influential users in the ten Telegram groups that are deemed influential in Section 6.1.4.3, we see that the Miss Rose bot is the most active user in four of the ten groups. Next to that, we notice four groups where the most active users contain two to four different users that have such a similar username they must be run by the same person or organization. We also notice three groups that contain several of the user accounts that sent the most-sent advertisement in our dataset (see Section 5.2.2.1). Furthermore, most of the most active users are classified as vendors. It also stands out that there are many users that are only part of the most active users in one Telegram group, but not of others.

Next, we take a look at the promoters of the important groups. First of all, there are two users that have sent tens of thousands of promotional messages containing a direct link to at least one of our most important groups, one of which is the Miss Rose bot. Next to that, there are an additional five users that have sent between 100 and 220 promotional messages. Furthermore, there are 19 users that have sent between 100 and 10 promotional messages. All other users mentioning one of the important groups have sent fewer than 10 messages doing so.

Another thing that stands out in the results is that cashout-88, cashout-65, and cashout-14 have respectively 22931, 22931, and 22934 promotional messages containing a direct link to the groups. Cashout-14 is promoted by 9 different user accounts, while the other two are promoted by 6 different user accounts. However, the 6 user accounts that promote cashout-88 and cashout-65 are also part of the 9 promoters of cashout-14. It could be the case that the 3 remaining promoters of cashout-14 have mentioned the group in a chat message instead of actively promoting the groups as the other 6 users seem to do. Looking at the promoters of these groups, we see two users that are responsible for most of the promotional messages of these groups, one of which is the Miss Rose bot. Soft\_drugs-15 and soft\_drugs-9, on the other hand,

Telegram: [redacted]  
 Wickr: [redacted] (🇳🇱 Alleen grote bestellingen)

---

🇳🇱 Price list / Prijs lijst 🇳🇱

---

❄️ SNUIF COKE / COCAINE / SOS ❄️

🇳🇱 COLOMBIAANSE COKE 🇳🇱

Voor de ervaren gebruiker 🙌

🇳🇱 0,8 gram	€50
🇳🇱 1 gram	€60
🇳🇱 2 gram	€110
🇳🇱 3 gram	€160
🇳🇱 4 gram	€200
🇳🇱 5 gram	€240
🇳🇱 10gram	€380
🇳🇱 20 gram	€760
🇳🇱 50 gram	€1850

Vaste klanten meer mogelijk

---

💎🔥 240 MG 🔥💎

Bezorgen:

- 🇳🇱 5 stuks €20
- 🇳🇱 10 stuks €40
- 🇳🇱 25 stuks €70
- 🇳🇱 50 stuks €120
- 🇳🇱 100 stuks €150 !! SALE !!
- 🇳🇱 200 stuks € 250

Alleen ophaal!

- 🇳🇱 500 stuks € 500
- 🇳🇱 1000stuks €750 !! SALE !!

---

ALTIJD Puur 🚨

🐱🐱🐱🐱 3MMC 🐱🐱🐱🐱

🐱 1 gram	€ 20
🐱 2 gram	€ 40
🐱 3 gram	€ 50
🐱 10 gram	€ 100
🐱 20 gram	€ 170
🐱 50 gram	€ 300

ALLEEN OPHAAL:

- 🐱 100 gram € 380 !! SALE !!
- 🐱 250 gram € 850 !! SALE !!
- 🐱 500 gram € 1500
- 🐱 1KG — € 2800

Vaste klanten meer mogelijk

FIGURE 6.9: Example of a message sent by the most active user

have many users sending a few messages containing a mention of the groups (i.e. a few being between 2 and 185 messages). A few could users showing this behaviour could be a coincidence, but in this case this happens for at least 50 users. Interestingly enough, there are also three of the most important groups that have a maximum of 11 messages promoting them. We believe there is a good chance these groups have a high out-degree, i.e. they mention many other groups but are not mentioned that much themselves.

When we take a closer look at the activity of the most active user in our dataset who is not an administrator bot, we see that this user has sent a little more than 385000 messages in 35 Telegram groups. Only 25 of these messages are unique and they are a member of 42 Telegram groups. We also see that this user has left some groups they were active in. Almost all messages this user sent are classified as an advertisement offering hard drugs. Figure 6.8 shows us the activity of the user (in the number of messages sent) on each day they since the first day they sent a message in each group they were active in. We see that this user has been active for approximately 5 months and the last moment of activity was April 23rd. Figure 6.10 displays the cumulative number of messages the user has sent in each group. As we can see, the sending pattern is the same for most groups this user was active in, with only a few groups as an exception. Finally, we look at the content of the messages this user sent. Most of the messages are variations of the same advertisement seen in Figure 6.9. This advertisement does not only contain the advertised product but also the contact details of the user for their Telegram and Wickr account. Next to the advertisements, we also identify two chat messages, where the user seems to be responding to another message. This indicates that there is a person actively monitoring this user account.

### 6.3 Conclusion

First of all, this chapter reasons what constitutes influence or importance in a setting such as the one presented in this research. We use two important perspectives of groups that may be interested in a work like this; law enforcement and academics. Looking at Telegram groups, we reason that we must take into account a group's place in the mention network, as Chapter 4 concluded the mention network is an essential part of how Telegram is used as a cybercrime platform. Next to that, we reason groups are more important, the more evidence of criminal activity it contains. We measure this in the number of advertisements per day. Finally, we argue that groups are likely to have more (societal) impact if they have a big(ger) audience. Therefore, we include the number of members a group has in our reasoning. Eventually, this allows us to distinguish several groups in our dataset that could be seen as more influential or important and which are likely to be of interest to interested parties.

Secondly, this chapter tried to determine which users are most important or influential, both in the network and in single Telegram groups. We see that a small group of users is responsible for a very large number of messages in the dataset, leading us to reason

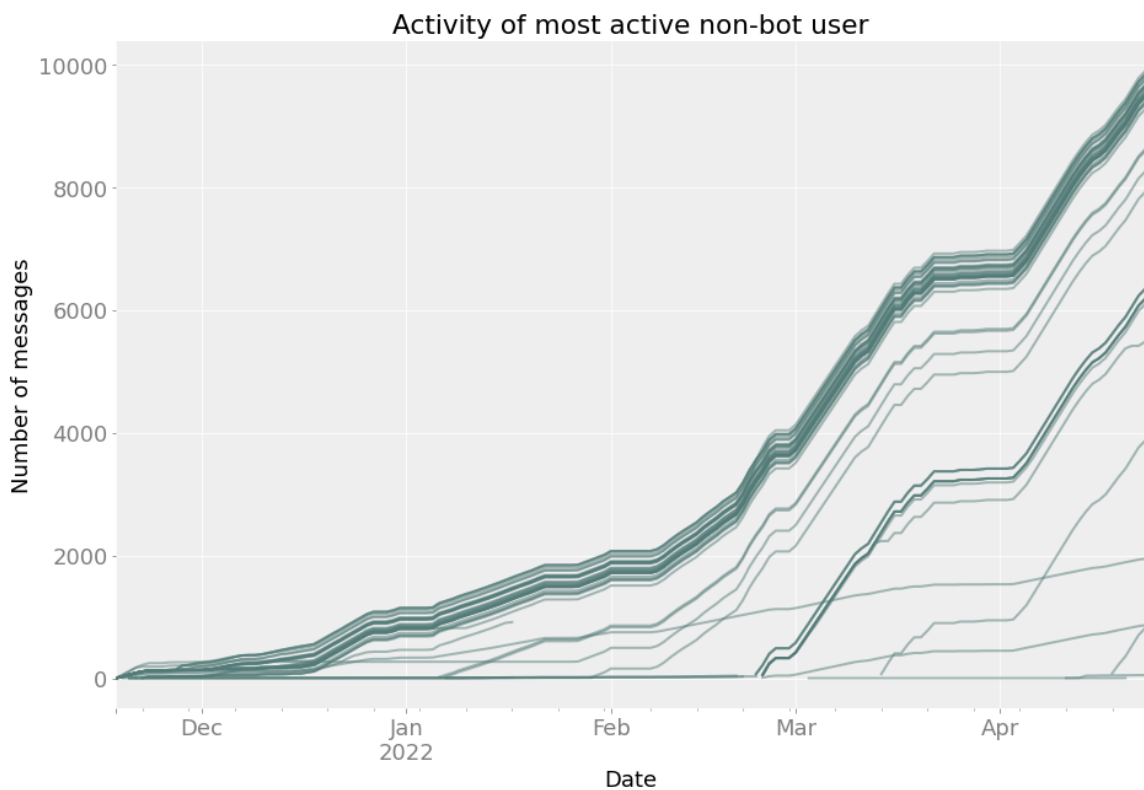


FIGURE 6.10: Cumulative activity (in nr. of messages) of most active user over time, where each line is a Telegram group they are active in

that the important users can be found between the most active users. Most of the most active users follow the pattern we associate with advertisement bots. We also identified promoters of the Telegram groups as important users, as they are (part of) the reason the mention network is the way it is. We can also find out some information about users by looking at the messages they sent, for example, the most active user that is not an administrator bot acts like an advertisement bot most of the time, but has also sent chat messages that reply to other messages, which suggests someone is monitoring the account and able to use it manually as well.

# Chapter 7

## Discussion

This chapter discusses the study presented in this work, including its methods and results, and relates it to the frame of reference created in Chapter 2. We include a description of how we observe Telegram is used as a cybercrime platform in the Netherlands, as we believe this to be an essential contribution of this work. Afterwards, we describe several limitations encountered in the process in Section 7.2. Finally, we round off the discussion with an overview of the implications of the findings of this work as well as some general recommendations to law enforcement and policymakers in Section 7.3 and Section 7.4 respectively.

### 7.1 Telegram as a platform for cybercrime

First of all, we notice many Telegram groups and channels with evidence of cybercrime. This study focused on a commonly occurring mix of traditional crime that takes place online, for example, the sale of illegal goods and services, and low-level cyber-dependent crime, such as the sale of phishing kits. However, there are also Telegram groups and channels focusing on exposing and bullying women, (revenge) porn, leaking music, and trading in cryptocurrency (and quite possibly tricking people into scams). Next to that, there are groups in a grey area, for example focusing on exposing fraudsters, organising protests, spreading anti-vaccination news<sup>1</sup>, or reporting police checks. We could roughly divide the groups in our dataset into three categories. First of all, there were groups where users primarily send chat messages, ask for advice, and ask for or offer products. Most groups focusing on fireworks in our dataset would fit this profile. Secondly, some groups are spammed by advertisements sent by advertisement bots. We see, for example, dozens of new messages appearing each minute and many advertisements are made to look attractive using emojis and formatting. Thirdly, some groups are a combination of the previous two types, where chat messages or users asking for or offering a product alternate more polished advertisements sent by advertisement bots.

---

<sup>1</sup>This was especially prevalent during the COVID-19 pandemic)

The three types of Telegram groups in the dataset make it difficult to give one definition of how single Telegram groups are used for cybercrime. Groups of the first type we would define as chat groups with the main topic being some form of cybercrime. This would allow for these groups to be studied in a similar fashion to other chat applications. Groups of the second type, however, can be viewed more as a messaging board, as these groups see almost no user interaction while many users just put a message out there for anyone to see. This would allow researchers to approach these groups in a fashion similar to studies researching underground forums or dark web markets. However, this forces one to look critically at the similarities and differences between these platforms. Telegram groups of the third kind, the combination of the two types, might be the most difficult, as approaches from both chat groups and underground forums or dark web markets have significant disadvantages. Therefore, we conclude, future research needs to be critical about the data, methods, and objectives of the research, as each influences the results.

As this work has shown, viewing single Telegram groups as single cybercrime platforms is just part of how Telegram is used as a platform for cybercrime. This work has shown that the mention network is created purposefully, and we argue it must be treated that way. This would suggest we must view the mention network as one platform existing of many smaller platforms (i.e. Telegram groups) held together by mentions. This would suggest a structure more similar to underground forums or a platform like Reddit, both of which allow for the existence of many subcommunities within one platform. However, we also believe one must be critical when comparing these structures, as the nature and meaning of mention relations between Telegram groups may differ from that of underground forums. In other words, while we believe this could be a useful comparison, we believe one must be critical about what methods to use.

We see four types of mentions in the dataset that create the mention network. First of all, there is the mention of a group by a user in a conversation, often happening in a group that is focused on chatting. Secondly, we see messages promoting a single group. Thirdly, there exist messages promoting several different Telegram groups at once, which we believe implies an alliance between these groups. Sometimes, this kind of promotional message is accompanied by the promise that other Telegram groups can be added to the list of promoted groups. The fourth type of mention is a mention of a Telegram group or channel included in an advertisement for illegal goods or services. This is often promoted as a way of contacting the seller or having more direct access to the vendor promoting their goods.

Finally, we found roughly three different types of users on the platform: regular users, administrator bots, and advertisement bots. The last two types of users were amongst the most active users in the dataset. Especially, the groups of the second and third types saw a lot of activity from these users (i.e. groups with automated advertisements). This would suggest that vendors and group administrators have thought about their marketing on Telegram and actively create Telegram groups that can be used for this. Regular users can mostly be found in the first and third types of groups (i.e. groups containing regular chat messages). These users also sometimes engage in cybercrime,



but are more likely to ask for or offer a product in a sentence or two than to create a polished advertisement.

Two-thirds of the users in the dataset were passive and about one-third were active, and we saw the common pattern that the most active users were responsible for the majority of messages in the dataset. We were not able to gather a lot of information about the passive users, which makes sense given the dataset. The fact that the most active users were responsible for the majority of the messages in the dataset pulled this work into the direction of the most active users. A disadvantage is, however, that this almost certainly excludes low-level vendors or traders in a platform such as the one presented in this research. We also noticed that the throughput of user accounts seemed high (i.e. there were many deleted accounts). What caused this and the effects of this need to be researched further.

In summary, Telegram groups for cybercrime function as single platforms that differ in nature from chat groups to message boards. The single Telegram groups are bound together by a mention network, which is, at least partly, purposely created by mentions in promotional messages. This work proposes methods to identify sub-communities in the mention network, but the many edges in the mention network suggest it is more useful to consider the whole mention network as a platform.

## 7.2 Limitations

In this section, we discuss both the methods employed in this study and the findings that have been obtained. To begin, we address two overarching limitations that have influenced the research as a whole. These limitations provide insights into the constraints and challenges faced during the study. Subsequently, we delve into a detailed exploration of each research phase, including a discussion of the employed methodologies and their implications. By presenting these limitations and research phases, this chapter contributes to a holistic examination of the study's approach, results, and the subsequent implications for further investigation.

Throughout this study, two significant limitations have been identified that have had a profound impact on the research outcomes. The first limitation pertains to the scarcity of existing research in this particular field. Although Chapter 2 has presented a rich body of criminological literature and highlighted academic interest in underground forums, dark web markets, and social media platforms from both criminological and computer science perspectives, the combination of these three fields remains largely unexplored. Consequently, this study represents a pioneering effort in investigating the convergence of these domains. While this presents an exciting opportunity to uncover new insights, it also poses challenges due to the exploratory nature of the research, limiting the ability to draw definitive conclusions given the vast amount of unknown information.

The second limitation concerns the nature of the dataset used in this study. Snapshot data was collected by scraping messages from Telegram groups starting from the group's inception until a predefined cut-off date. As a result, the availability of historical data

varies among different Telegram groups, as some have existed for longer durations than others. This discrepancy in historical data coverage introduces potential biases in the analysis. Additionally, snapshot data fails to capture certain valuable information, such as user activities (e.g. changing usernames, user accounts or message deletions), Telegram group migrations, and other dynamic elements. To address this limitation, a more comprehensive approach involving continuous data collection could be employed. By periodically monitoring Telegram groups and updating the dataset with new messages, researchers can capture a more accurate representation of user interactions and dynamic changes within the platform.

### 7.2.1 Approach

We acknowledge that there may be limitations to our dataset. The search function in the Telegram Desktop App returns a limited number of search results, so we cannot guarantee that we included every possible Telegram group in the scraper-snowball method for data gathering, as the success of this method depends on the starting point of the snowball method. However, the snowball method is also what guarantees that the data collection process continues until saturation. While we stopped the snowball process manually, because it converged too far from the desired topic, the property of continuing until saturation is what makes us believe we obtained the best possible dataset for this study given the starting point of the snowball. Additionally, we were unable to gather additional information about groups, such as the names of the administrators. Obtaining this information could provide an extra perspective on the groups and their activities.

Another challenge we faced during the data collection process was that the function we used to obtain the members of a group (i.e. `getparticipantsrequest` from the Telegram API) is rumoured to not always return the exact number of members of a channel or group, but about 90% instead. Additionally, the API stops returning users after 10000 users unless explicitly stated. Unfortunately, we only discovered this limitation after collecting data, which led to missing data and potential errors. While we believe this had a negligible impact on the results of our study, we can not claim perfect accuracy. We recommend future researchers be aware of this and implement their own tactics to counter this.

Another challenge we encountered was that the Telegram API marks accounts using the API as spam rather quickly, resulting in timeouts for users that could last anywhere from 30 minutes to 30 hours. This problem occurred regularly during the scraping process and we tried different methods to avoid being marked as spam, but it was not always clear what caused the API to flag our account. As a result, we had to manually restart the scraper after each timeout and let it check the last Telegram group it scraped before continuing. This delay added to the time required for the scraping process.

Moreover, we decided we would not include media, such as pictures or videos, in the study, even though manual inspection of the Telegram groups showed plenty of evidence media was used for cybercrime. This is for two reasons. First of all, including media

in the scraping process would have required substantial storage space, and secondly, we had concerns about potential explicit content being included. While we believe the results presented in this work are not affected by the lack of media, we do realise that not including media excluded Telegram groups from our dataset in the topic fitting process described in Section 3.2 that would otherwise have been included. Future studies could consider focusing on media, or at least include media in their research.

Furthermore, the classification process was a crucial part of this research, and the results presented in Section 3.3.4 indicate that the classifier performed reasonably well. However, we realise that the classifier is not perfect and that every claim we make about the content of messages is dependent on its functioning. Further research can be done in this area to aid future research using large-scale textual content. While we chose a machine learning approach that was trained on manually labelled data, [Blankers et al. \(2021\)](#), for example, chose to filter Telegram messages based on keywords. [Dargahi Nobari et al. \(2017\)](#), on the other hand, compares three machine learning approaches to distinguish Telegram messages between ham and spam and found that all seemed to perform equally. However, they did not use them for textual classification but used several message characteristics to distinguish on (e.g. message length and number of mentions).

We also chose to include Flemish-speaking Belgian groups in our dataset, as the method we used to filter non-Dutch groups did not allow us to distinguish between the two in written form. The suspicion that these groups might be Belgian instead of Dutch is mainly based on the name of the groups and we did not see any difference in content when compared to Dutch groups. There are multiple possible explanations for this, for example, that the administrators of the groups are the same as for the Dutch groups, or the fact that the vendors are not very creative with their messages. However, given the local embeddedness of cybercrime (see Section 2.1.2.3), we believe it most likely that this group has the same or similar vendors as primarily Dutch groups, with the difference that these groups might aim to find their primary clientele just across the border in Belgium. We recognise that including these groups may have some effects on the results of our analysis, and future research could explore these effects.

Finally, we encountered difficulties in distinguishing between requests and offers in advertisements related to cashout. Many cashout advertisements ask for people who are willing to lend their bank accounts or create a new one, which raises the question of whether this should be seen as an offer or a request. A similar problem exists for advertisements related to courier services. Future research could consider whether creating a category called "advertisement" would be more appropriate than distinguishing between requests and offers in these cases, or could focus on improving the accuracy of a classifier.

Overall, these limitations and challenges underscore the importance of being aware of potential biases and limitations when conducting research in this area. Further research could address some of these limitations and improve upon the methods used in this thesis.

### 7.2.2 Phase 1

First, we encountered a problem with clustering algorithms based solely on group characteristics. Our attempts to cluster the groups based on their properties were unsuccessful. It is possible that clustering groups based on their characteristics may not be feasible, but we suspect we may not have chosen the correct group characteristics to cluster on. However, we believe it could be interesting for future research to look into the clustering of Telegram groups based on their characteristics, as this could allow us to distinguish between groups and assign roles. Eventually, we imagine this could, for example, help in predicting the development pattern of Telegram groups based on the type of group or improve attempts that prioritise influential groups.

Secondly, in our analysis of the network, we focused on mentions rather than forwards. This decision was based on the snowball method we used to collect our data, which relied on mentions to identify new groups. While we believe this to be a sound approach, some may argue that our method is like a self-fulfilling prophecy, as it may overemphasise the importance of the mention network. Therefore, exploring the use of forwards in future studies may yield valuable insights, for example by comparing a mention and a forward network. Future work could also look into relations between groups with the use of hashtags or replies, focussing more on related topics and user behaviour.

Thirdly, it could be interesting to create a mention network of Telegram groups and combine it with other sources of information to determine whether the dynamics in the Telegram groups can provide insights into criminal networks. For example, one could combine and compare a mention network with other social media sources, news articles, dark web markets or cryptomarkets, or knowledge of (local) criminal organisations. This could potentially bare relations that would have otherwise gone undetected.

Fourthly, we used the Louvain algorithm to analyze the network structure. However, in the process of transforming our mention network into an undirected network, we lost important information. This loss of information could be a reason for the poor results obtained by our approach. To use this method successfully, one would need to use a different approach or a different implementation of the Louvain algorithm. However, it could also be that the mention network created in this study is just not suitable for this algorithm due to its high number of edges.

Finally, regarding the overlap in advertisements and users between Telegram groups, we expected to find stronger relationships between groups using this method. Especially for overlap in users, we had expected to find more overlap between groups with similar topics. However, we can also imagine it would make more sense for a user to join multiple groups with the same topic if these groups were chat groups than when the groups would be spammed with advertisements. Regarding advertisement overlap, we only looked at unique advertisements, and it makes sense there is only a big overlap in advertisements between two groups if it is primarily used by the same vendors. We suspect that we may get different results if we include duplicates in our analysis. Therefore, future research could explore how accounting for message duplicates could impact our understanding of group relationships.

### 7.2.3 Phase 2

In this section, we discuss several aspects related to the user information used in this research. First, we explain our decision to use only the user ID and username to ensure the anonymity of the users in our dataset. Second, we point out the potential of additional user information that could be gathered from the dataset. Third, we describe our attempts to create a graph of active users and the challenges we faced. Fourth, we discuss the limitations of user information in relation to username changes. Fifth, we suggest a potential approach to filter out advertisement bots from the dataset. Finally, we present the promising results obtained from analysing the most-sent message in the dataset.

We chose to use only the user ID and username of the users in our dataset to protect their anonymity. This means we have no personal information or information that could be used to identify the users in our dataset. However, we noticed that some users post personal information such as phone numbers, locations, names, and social media handles. While we did not use this information in our analysis, interested parties could aggregate this information to gain more insight into the users and the Dutch cybercrime landscape. For example, they could look into crime per region in the Netherlands, identify the locations different vendors cater to, determine popular crime markets per region, or extend user profiles with the information shared.

Without the additional user information mentioned above, we have limited information about the users in our dataset, especially inactive users. While active users are likely to be more interesting in a dataset such as ours, it is difficult to gather more information about a user beyond the Telegram groups they are a member of, their role, and the messages they sent. This could be seen as a limitation of the current work.

Next to that, in the second phase of the research, we attempted to create a graph of active users with many-to-many relations between users in the same group. While this graph could provide insights into users' place in the network, we faced challenges due to memory issues and were unable to successfully run the code needed for this analysis. While the information such a graph provides could have added to the current work, the fact that we could not use this method has forced us to look at the problem from another point of view. We believe the method of looking at users sending the same advertisements potentially contains a lot of information and could be an interesting one to develop in further research. A limitation of this method is that we used hashed versions of advertisements. Using hashes makes it easier to compare advertisements, but has a downside that hashes are different as soon as one character in an advertisement is different. For example, we noticed at least one version of the most-sent advertisement (used to illustrate this method in Section 5.2.2.1) that had a different hash due to a slight change in the message.

It is also important to note that this research eventually focused on advertisement bots, as these types of users seemed to spread the highest absolute number of messages relating to criminal activity. While this makes sense, it does not necessarily mean that these are the most interesting users. In future research, one could consider filtering out

advertisement bots from the dataset, so they could focus on users that do not follow the obvious advertisement bot pattern.

Finally, we found that analyzing the most-sent message in the dataset can provide insights into the relations between users. This approach shows promise when combined with other information gathered from our dataset, such as user activity or the Telegram groups in which the advertisement was sent. Our results in Section 5.2.2.1 support this approach and suggest it could be used to further explore the relations between users in the dataset. Next to that, this approach illustrates how much information can be obtained from a dataset such as the one used in this research. The nature of this research is exploratory and it was our goal to show several possibilities of methods that could be applied to create insights. However, we believe one could go into a lot more detail using a similar dataset.

### 7.2.4 Phase 3

In this section, we discuss several important points related to the third phase of our research. First, we want to emphasize that our choice of influential characteristics does not necessarily imply the most influential groups, but rather groups of interest. For example, [Dargahi Nobari et al. \(2017\)](#) takes into account spam and ham messages, which is something we did not include. However, they also defined a Telegram channel's popularity and quality directly based on the channel's members, a property we believe does not cover the complete complexity of popularity and quality. We recognize that defining influence is a complex task and requires careful consideration of network properties, group properties, and the context of the research. Given the explorative nature of our work, we chose to keep our approach simple. For instance, we did not attempt to make sense of all network properties as we need to be critical about what they mean for this type of network and cannot directly copy approaches used for OSNs. However, future work could look into the meaning of influence in a mention network and further determine characteristics that can help quantify influence.

Moreover, when determining influence, we excluded Telegram groups that are not part of the mention network because of the network's nature and our choice of "influence" characteristics (i.e. we used a group's degree in the network as one of the characteristics). While other characteristics do not justify excluding these groups from being seen as influential just because they are not part of the mention network, we believe the results of the method are still well-founded, as the results are still based on solid reasoning.

Furthermore, in Chapter 2 we discussed the role of reviews or ratings on dark web markets or cryptomarkets. While these markets are specifically designed so that a user does not need to trust anyone, Telegram groups such as the ones in the dataset do not have those features. However, there exist Telegram groups specifically designed to expose fraudsters. We chose not to include these groups in our research because they are not directly related to our research questions. Future studies could investigate the accuracy of accusations made in such groups and their impact on reputation and trust in Telegram groups. Next to that, we suspect that Telegram groups are likely to have

some kind of reputation, as some become popular and others do not. This, we believe, is likely closely related to the promotions of the groups, the moderation activity within a group, and word of mouth between users. Future research could look into this to better understand how a network of Telegram groups develops over time and what are important factors for users to decide what groups they want to become members of.

In Section 6.2.4, we used user activity as a measure of importance. However, we believe it would be more interesting to combine this information with the results from Section 5.2.2.1 to determine the most influential group of users in the network. By identifying the users who seem to work together, we can gain a better understanding of the network's structure and its most significant players.

### 7.3 Implications

Section 2.1.2.3 in Chapter 2 made clear that the Netherlands is a major player when it comes to the sale of drugs on dark web markets. However, it was also implied that it was unlikely for Dutch buyers to buy drugs through dark web markets, as there would likely be more efficient ways to do that given the close geographical proximity of the buyer and the vendor. While the act of buying and selling is not included in this study, we believe this work shows an example of a platform that allows for these local connections between buyers and sellers to form. An implication of this relates to related work mentioned in Chapter 2 as well; buyers no longer have to rely on shady dealers on street corners, but can place an order through social media and can get it delivered wherever they want. In other words, the use of social media platforms for cybercrime allows for easier and more convenient access to illicit substances.

The easy accessibility of social media platforms has more effects. The fact that social media is accessible, easy to use, and often free makes that the platforms are found by an enormous amount of users. This can cause more people to come into contact with cybercrime, such as the advertisements for illegal goods and services presented in this work. Given the ease with which we found Telegram groups containing cybercrime in the data collection process described in Section 3.1.4, we reckon this is the case for the average Telegram user as well. This easy accessibility could cause the audience of cybercriminal vendors to grow, potentially tempting more people to buy illegal goods or services or drawing more people into the (cyber)criminal world, for example by promising them easy money. Therefore, we believe Telegram being used as a cybercrime platform can have a great impact on society.

Another implication of this research is that it shows that cybercrime on Telegram in the Netherlands is not just something that happens sporadically; it is organised in a semi-professional way. This is shown, for example, by the advertisement bots that are used by several vendors to spread the advertisements of their products. Not only that, but the content of the advertisements sometimes changes, including or excluding a product (presumably based on availability) while the rest of the message stays the same. Moreover, this work shows that vendors sometimes use multiple user accounts to spread their advertisements. Additionally, promotional messages of single Telegram groups or alliances

of Telegram groups indicate the groups are deliberately created, aware of each other, and competing for their place in the market. We can only conclude that this work shows evidence of vendors using marketing strategies and advanced tools for aggressive marketing, promotional campaigns, alliances, etc. In other words, cybercrime on Telegram seems to have transcended small-scale street-level crime, or at least professionalised it.

However, law enforcement should be critical and precise when convicting (cyber)criminals. This work proposes methods for finding important users and Telegram groups, the "big fish". These methods are based on reasoning within our current frame of reference. However, it is important that others, such as law enforcement and academics, are critical of the context and its implications on the used methods when using the same methods to find the differentiate the big fish from the mass. For example, metrics used in this study need not be relevant in every other context or a mention-graph of mentions between Telegram groups at another moment in time may result in a different view of the cybercrime landscape on Telegram as it does in this work,

## 7.4 Recommendations

Although it is clear that more research is needed to uncover all the nuances that are present in the way Telegram is used for cybercrime - and about social media and cybercrime in general - this study gives a glimpse of the size and impact of this phenomenon. It is because of this that this section highlights several recommendations for law enforcement and policymakers.

First of all, the context of this study, described in Chapter 2, makes clear that cyber-enabled crime is a growing problem and is not going away soon. Previous studies found that both traditional criminals (i.e. non-digital criminals) and cybercriminals are making use of social media platforms, which means cybercrime is no longer limited to the tech-savvy. Next to that, the popularity of social media and its ease of access allows for a big(ger) audience to come into contact with cybercrime. The scale of cybercrime on Telegram described in this work in combination with the professionalisation of cybercrime on the platform, described in Section 7.3, highlights the need to take the problem seriously. Therefore, we recommend that law enforcement and policymakers be aware of this and use this awareness in future endeavours, for example by using social media research in their standard procedures or including social media as a platform for crime in policies for preventing crime.

Secondly, this research shows that publicly available information on Telegram can provide a wealth of information about cybercriminals. Moreover, we noticed that users tend to share quite a lot of personal information or contact details in their messages or profiles. This is caused, we believe, by the fact that many social media users tend to have a lot of trust in the privacy of the platforms they use or the disinterest of law enforcement agencies in them personally, as is described in Chapter 2. Section 7.2.3 addressed that this study made the explicit choice not to include this personal information. However, including this information and combining it with public information from other platforms could provide law enforcement agencies or researchers with extensive user profiles,



especially when combined with the relationships that this research suggests can be found in the data.

Furthermore, we recommend law enforcement agencies look into the relationships social media platforms allow us to uncover. Social media platforms, being inherently built on social relationships, allow for the discovery of intricate networks of relations between users when approached with relationship graphs or overlapping characteristics, as shown in this work. Being able to find relations between users can allow for the discovery of alliances between criminals or even members of organised crime groups. Recalling Chapter 2 stating that cybercrime is locally embedded and that core criminals often form their relationships offline highlights the importance of this, as it is likely that online relationships between (cyber)criminals have an offline, local impact.

Countering cybercrime on Telegram is not easy, as one of Telegram's fundamental characteristics is the privacy of its users. While one could argue that Telegram's privacy should go, as it allows criminals to conduct their business, there are more sides to this story. For example, Telegram was one of the main platforms that were used by protesters during the protests in Belarus in 2020 to communicate with each other and with the outside world (Aliaksandr Herasimenka, Tetyana Lokot, Olga Onuch and Mariëlle Wijermars, 2020; Onuch, Sasse, & Michiels, 2023; Wijermars & Lokot, 2022; Williams, 2020). As a society, we should ask ourselves if we expect Telegram to change its rules surrounding privacy just so we can combat crime. Next to that, law enforcement agencies are bound by laws about what digital data can and cannot be used when combating crime. Therefore, we believe it is difficult to prevent crime on Telegram altogether and believe researchers, law enforcement agencies, and policymakers should work together on developing new methods of discovering and countering crime on the platform.

## Chapter 8

# Conclusion

Throughout this work, our aim has been to bridge the research gap outlined in Chapter 1 by addressing the main research question: *How can we identify influential users in Dutch Telegram groups used for cybercrime?* This has been achieved by dividing the research into three distinct phases, each with its own set of sub-research questions. In doing so, we have provided a comprehensive understanding of the use of Telegram as a cybercrime platform and have offered valuable insights into the Dutch cybercrime landscape.

This work starts by creating a frame of reference using academic literature to get an idea of the definition of cybercrime, types of actors, and different types of cybercrime. When cybercrime started shifting from its traditional platforms, i.e. underground forums and dark web markets, to social media, it became more accessible for both criminals and potential buyers. Combining this with the knowledge that relationships between core criminals are formed offline and that cybercrime, and specifically cybercrime on social media, is locally embedded, we have a good foundation to approach the problem.

In Chapter 3, we present a method for collecting and preprocessing data. Together with a classifier, this method is used to create the dataset used in this study. This was not a goal in itself, but we believe it is a great contribution to the field, allowing others to recreate the same method.

In this study's first phase, we aimed to create a comprehensive profile for Telegram groups used in cybercrime. Chapter 4 presents the phase's outcomes, which highlight several metrics that are easy to determine for Telegram groups, such as the number of active and passive users in a group, a group's in- and outdegree in the mention network, and the division of the message types and advertisement categories present in the group. Our findings suggest that these metrics provide a holistic view of a Telegram group. However, the profile information did not allow us to effectively categorise the groups. Furthermore, we aimed to determine the relationships between the Telegram groups during this phase. The relationships between the Telegram groups are displayed in a mention network and the connectedness of the mention network suggests that it is purposely created and maintained. Another method presented in the same chapter was identifying relations between groups through advertisement and user overlap. While a

few suspected relationships between groups were identified, the lack of advertisement and user overlap between groups that focused on the same crime market indicates that the platform's diversity might be greater than initially expected. Overall, we are able to create profiles of Telegram groups and detect relationships in the groups. Results such as the mention network also indicate a level of professionalism and competitiveness that transcends street-level crime.

In the second phase of this research, we investigate the users within the dataset. We gather basic information about user membership and activity and assign roles and crime markets of interest to categorize the users. This allowed us to conclude that almost two-thirds of the users in the dataset are passive users. Additionally, we examine previously active users who had sent messages in the groups but were not currently members at the time of scraping. It stood out that there were a lot of these users, implying a high throughput of user accounts or many users leaving the Telegram groups. These users are assigned a role and their activity is recorded, but their profiles are less extensive than those of current users due to the lack of membership or username information. These users were equally distributed over the buyer, vendor and chatter roles. We also find that determining relationships between users is challenging due to the many-to-many relations between users in a Telegram group. For once, we were not able to create a useful graph of relations between all users in the dataset, due to its sheer size. Nonetheless, we introduce a novel method in this section that examines users who have sent the same advertisement, as we believe it is unlikely for this to be a coincidence. We provide two examples using the most-sent and second-most-sent advertisement in our dataset, demonstrating that this approach can create a graph that visualizes users who have sent the same advertisements, allowing for the identification of relationships between users that may not have otherwise been apparent.

In the third phase of this thesis, we utilize the insights gained from the first two phases to determine the influence of both Telegram groups and their users. With regards to groups, we propose that the average number of advertisements per day, the number of members, and the group's degree in the mention network are suitable characteristics for identifying important groups in the network. We also assert that the number of messages a user has sent is the most relevant user characteristic for identifying influential users in the given dataset. This holds true for both individual Telegram groups and the network as a whole. By sorting users according to the number of messages they have sent, our findings in this phase reveal the influential users in both influential Telegram groups and the network, with some overlap between the two. Additionally, we propose that users who promote these influential groups could also be considered influential users, as they contribute to the current state of the Telegram group network. However, our results demonstrate that this only concerns a few users.

In conclusion, this work introduces a method to identify influential users in Dutch Telegram groups used for cybercrime. Although the third phase is crucial for identifying influential Telegram groups and answering the main research question, it would not have been possible without the information gathered in the previous two phases and the data collection phase. Additionally, we aim to have provided readers with a better

understanding of Telegram's use as a cybercrime platform and insights into the Dutch cybercrime landscape.

# Appendix A

## Appendix

Element	Type
id	int
peer_id	Peer
date	date
message	string
out	flag
mentioned	flag
media_unread	flag
silent	flag
post	flag
from_scheduled	flag
legacy	flag
edit_hide	flag
pinned	flag
from_id	Peer
fwd_from	MessageFwdHeader
via_bot_id	int
reply_to	MessageReplyHeader
media	MessageMedia
reply_markup	ReplyMarkup
entities	MessageEntity
views	int
forwards	int
replies	MessageReplies
edit_date	date
post_author	string
grouped_id	long
restriction_reason	RestrictionReason
ttd_period	int

TABLE A.1: Information included in a scraped message (*Message*, n.d.)

Element	Type
id	int
first_name	string
last_name	string
user	string
phone	string
is_bot	flag

TABLE A.2: Information included in a scraped user (*Teleton User*, n.d.)

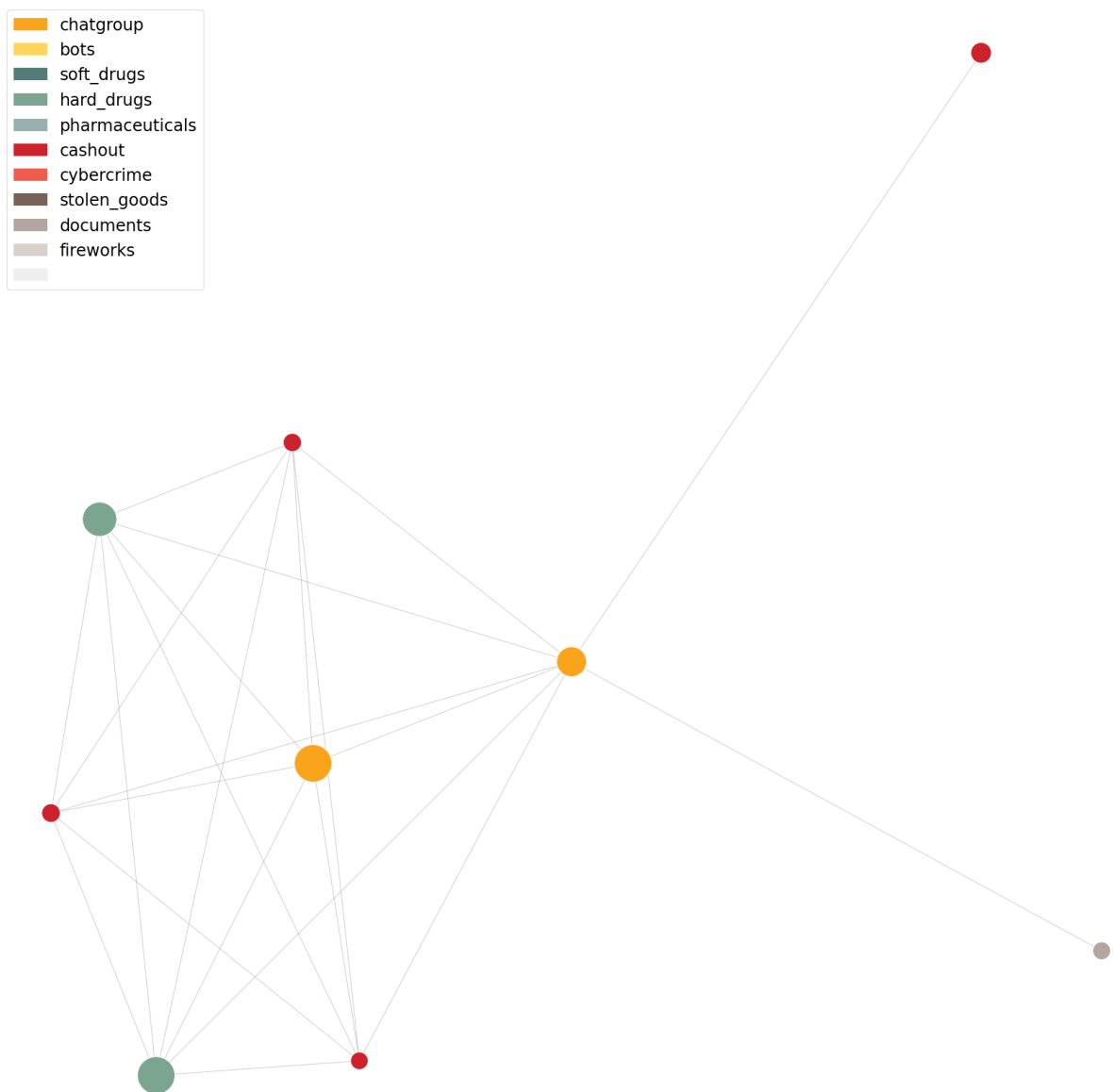


FIGURE A.1: Island of 9 nodes in the mention graph with only reciprocal edges

<b>Jargon</b>		<b>Sold goods</b>	<b>Location</b>
handel	Lounge	wapen	Nederland
fraude	Brigade	Drugs	NL
fgame	Alles	Vuurwerk	Hollandia
group	Grinders	vuurwerkhandel	Hollands
Winkel	Netwerk	vuurwerkvandalen	Noordnederland
Markt	Onderwereld	vw	Zuidnederland
Tele	Werkers	Droga	Westnederland
swipe	Gelekt	Visjes	Oostnederland
Blackmarket	Groothandel	Phishing	Randstad
verborgen	Multiculti	Panels/pannels	MiddenNederland
groep	Oplichters	Druga	Twente
scam	pgroep	Bonk	Amsterdam
Street	Zwartehandel	Bonker	Adam
Straat	Paradijs	Lacasa/lacasa	Rotterdam
market	Geld	Voertuig	Utrecht
scammer	Kech	Wappa	DenHaag
Koop	Ghetto	Gannoe	Friesland
verkoop	Koophandel	Medicijn	Noordholland
anoniem	Zwartemarkt	Accounts	Zuidholland
zakelijk	handelee	Documenten	Brabant
illegaal	handelgrond	Kaarten	Flevoland/Flevo
shop	handelsgroep	Gestolen	020
Crime	telemarkt	Phishingkit	030
misdaad	vanalles	Kit	070
Hossel/hussel	Apotheek	Methode	utrecht030
Hardwork	fgamechat	Cards	
Strijder	fraudeur	Canabis	
Hiyenas	hardewerker	Rijbewijs	
Hyena	markthandel	Diploma	
Verhuur	Multiculti	Vuurwerk2020	
Marktplaats	Swipeouwe	Vuurwerk2021	
jacht		vuurwerk2019	
		Pillen	
		schetsers	

TABLE A.3: The search terms used to in the Telegram API to search for Telegram groups

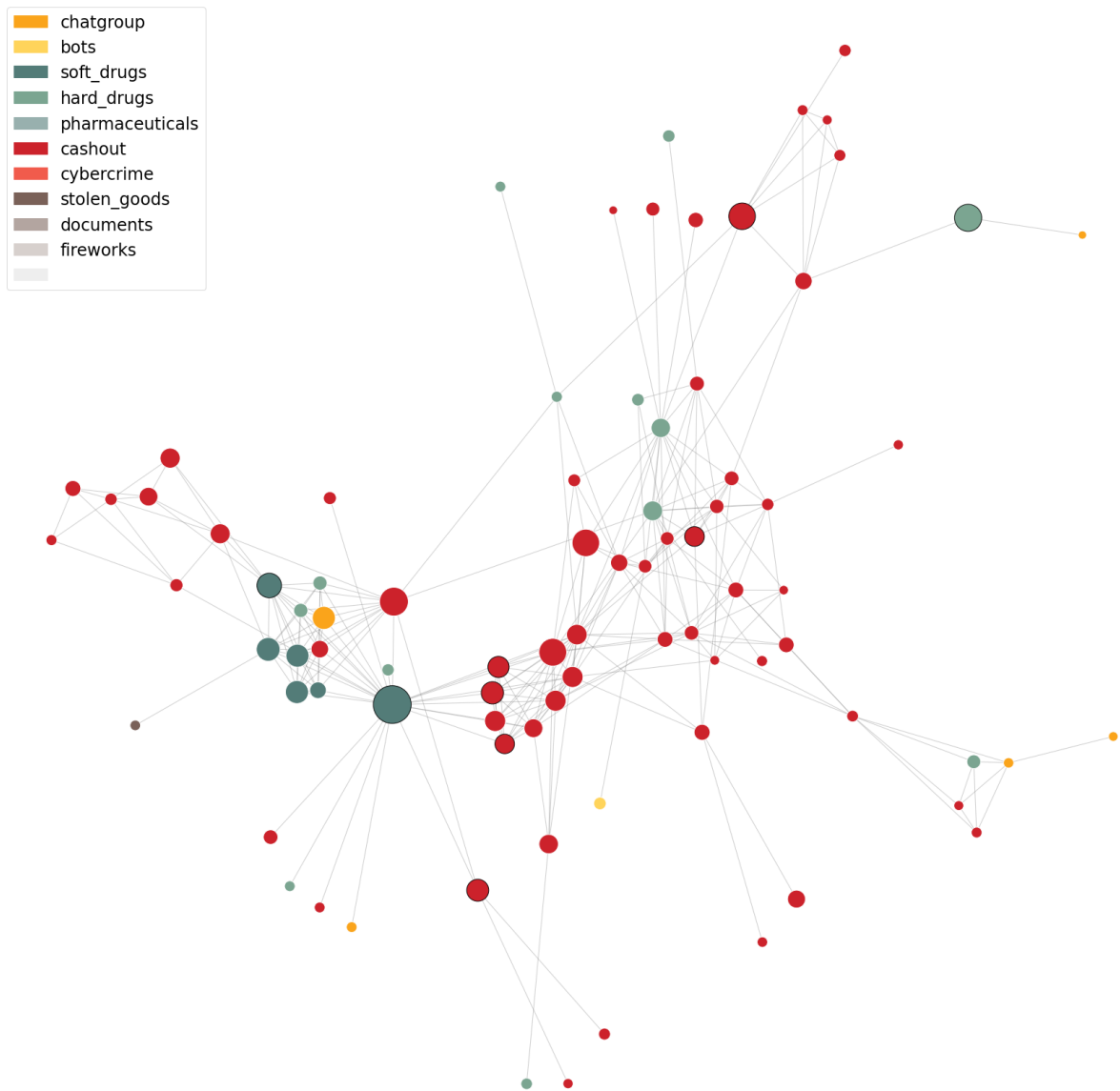


FIGURE A.2: Island of 84 nodes in the mention graph with only reciprocal edges



# References

- 800 criminals arrested in biggest ever law enforcement operation against encrypted communication. (2021, June). <https://www.europol.europa.eu/media-press/newsroom/news/800-criminals-arrested-in-biggest-ever-law-enforcement-operation-against-encrypted-communication>. (Accessed: 2022-5-4)
- Abiodun Raufu and Lucy Tsado and Emmanuel Ben-Edet. (2021, March). *Cybercrimes: Critical issues in a global context* (Vol. 64). Elsevier.
- Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013, December). A survey of data mining techniques for social media analysis.
- Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., Mujtaba, G., Khan, M. U. S., & Khan, S. U. (2018, January). Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Comput. Surv.*, 51(1), 1–37.
- Aliaksandr Herasimenka, Tetyana Lokot, Olga Onuch and Mariëlle Wijermars. (2020, September). There's more to belarus's 'telegram revolution' than a cellphone app. *The Washington Post*.
- Al-Rawi, A. (2019, December). The fentanyl crisis & the dark side of social media. *Telematics and Informatics*, 45, 101280.
- Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, 33(1), 1–4.
- Aynaud, T. (n.d.). *Community detection for NetworkX's documentation — community detection for NetworkX 2 documentation*. <https://python-louvain.readthedocs.io/en/latest/>. (Accessed: 2022-9-29)
- Bakken, S. A., & Demant, J. J. (2019, November). Sellers' risk perceptions in public and private social media drug markets. *International Journal of Drug Policy*, 73, 255–262.
- Ball, M., & Broadhurst, R. (2021, February). *Data capture and analysis of darknet markets*.
- Barn, R., & Barn, B. (2016, June). An ontological representation of a taxonomy for cybercrime. In *ECIS 2016 proceedings*.
- Bijmans, H., Booij, T., Schwedersky, A., Nedgabat, A., & van Wegberg, R. (2021). Catching phishers by their bait: Investigating the dutch phishing landscape through phishing kit detection. In *30th USENIX security symposium (USENIX security 21)* (pp. 3757–3774).

- Blankers, M., van der Gouwe, D., Stegemann, L., & Smit-Rigter, L. (2021, June). Changes in online psychoactive substance trade via telegram during the COVID-19 pandemic. *Eur. Addict. Res.*, *27*(6), 469–474.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008, March). Fast unfolding of communities in large networks.
- Bonacich, P. (1972, January). Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.*, *2*(1), 113–120.
- Brin, S., & Page, L. (1998, April). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117.
- Broadhead, S. (2018, December). The contemporary cybercrime ecosystem: A multidisciplinary overview of the state of affairs and developments. *Computer Law & Security Review*, *34*(6), 1180–1196.
- Brown, C. S. D. (2015, August). Investigating and prosecuting cyber crime: Forensic dependencies and barriers to justice. *International Journal of Cyber Criminology*, *9*(1), 55.
- Buntain, C., & Golbeck, J. (2014, April). Identifying social roles in reddit using network structure. In *Proceedings of the 23rd international conference on world wide web* (pp. 615–620). New York, NY, USA: Association for Computing Machinery.
- Cascavilla, G., Tamburri, D. A., & Van Den Heuvel, W.-J. (2021, June). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, *105*, 102258.
- Celestini, A., Me, G., & Mignone, M. (2016). Tor marketplaces exploratory data analysis: The drugs case. In *Global security, safety and sustainability - the security challenges of the connected world* (pp. 218–229). Springer International Publishing.
- Chandra, A., & Snowe, M. J. (2020, September). A taxonomy of cybercrime: Theory and design. *International Journal of Accounting Information Systems*, *38*, 100467.
- Channels, supergroups, gigagroups and basic groups.* (n.d.). <https://core.telegram.org/api/channel>.
- Chen, A. (2011, June). *The underground website where you can buy any drug imaginable.* <https://www.gawker.com/the-underground-website-where-you-can-buy-any-drug-imag-30818160>. (Accessed: 2023-5-11)
- Christin, N. (2013, May). Traveling the silk road: a measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on world wide web* (pp. 213–224). New York, NY, USA: Association for Computing Machinery.
- Church, K., & de Oliveira, R. (2013, August). What’s up with whatsapp? comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the 15th international conference on human-computer interaction with mobile devices and services* (pp. 352–361). New York, NY, USA: Association for Computing Machinery.
- Combot.* (n.d.). <https://combot.org>. (Accessed: 2023-4-4)
- D. Lummen, K. Boersma, R. Dijkstra, J. Sustronk. (2021, June). *Is telegram the new place to be for criminals?* (Research assignment, available upon request)
- Dargahi Nobari, A., Reshadatmand, N., & Neshati, M. (2017, November). Analysis of

- telegram, an instant messaging service. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 2035–2038). New York, NY, USA: ACM.
- Dargahi Nobari, A., Sarraf, M. H. K. M., Neshati, M., & Erfanian Daneshvar, F. (2021, April). Characteristics of viral messages on telegram; the world’s largest hybrid public and private messenger. *Expert Syst. Appl.*, *168*, 114303.
- Décary-Héту, D., & Giommoni, L. (2017, February). Do police crackdowns disrupt drug cryptomarkets? a longitudinal analysis of the effects of operation onymous. *Crime Law Soc. Change*, *67*(1), 55–75.
- Décary-Héту, D., Paquet-Clouston, M., & Aldridge, J. (2016, September). Going international? risk taking by cryptomarket drug vendors. *Int. J. Drug Policy*, *35*, 69–76.
- Definition of SOCIAL MEDIA*. (n.d.). <https://www.merriam-webster.com/dictionary/social%20media>. (Accessed: 2022-6-29)
- Donalds, C., & Osei-Bryson, K.-M. (2019, March). Toward a cybercrime classification ontology: A knowledge-based approach. *Comput. Human Behav.*, *92*, 403–418.
- Dupont, B. (2017, February). Bots, cops, and corporations: on the limits of enforcement and the promise of polycentric regulation as a way to control large-scale cybercrime. *Crime Law Soc. Change*, *67*(1), 97–116.
- European Monitoring Centre for Drugs and Drug Addiction and Europol. (2020). *EU drug markets: Impact of COVID-19*. Luxembourg: Publications Office of the European Union.
- Europol. (2021). *Internet organised crime threat assessment (IOCTA) 2021*. Publications Office of the European Union.
- Europol, C.-B. (2017). *Internet organised crime threat assessment : IOCTA 2017*. Publications Office of the European Union.
- Famous DDoS attacks*. (n.d.). <https://www.cloudflare.com/learning/ddos/famous-ddos-attacks/>. (Accessed: 2022-7-1)
- Group help bot*. (n.d.). <https://www.grouphelp.top>. (Accessed: 2023-4-4)
- Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (Tech. Rep.). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- Han, X., Kheir, N., & Balzarotti, D. (2016, October). PhishEye: Live monitoring of sandboxed phishing kits. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 1402–1413). New York, NY, USA: Association for Computing Machinery.
- Harfield, C., & Schofield, J. (2020, August). (im)material culture: towards an archaeology of cybercrime. *World Archaeol.*, *52*(4), 607–618.
- Harkin, D., Whelan, C., & Chang, L. (2018, November). The challenges facing specialist police cyber-crime units: an empirical analysis. *Police Pract. Res.*, *19*(6), 519–536.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362.
- Holland, P. W., & Leinhardt, S. (1971, May). Transitivity in structural models of small groups. *Comparative Group Studies*, *2*(2), 107–124.

- Holt, T. J., Strumsky, D., Smirnova, O., & others. (2012). Examining the social networks of malware writers and hackers. *International Journal of*, 6(1).
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, 9(3), 90–95.
- International Telecommunication Union. (2021). *Measuring digital development: Facts and figures 2021*. International Telecommunication Union.
- Jewkes, Y., & Yar, M. (2013). *Handbook of internet crime*. Routledge.
- Junger, M., Veldkamp, B., & Koning, L. (2022, March). *Fraudevictimisatie in nederland* (Tech. Rep.). University of Twente.
- Katsikeas, S., Johnson, P., Ekstedt, M., & Lagerström, R. (2021, November). Research communities in cyber security: A comprehensive literature review. *Computer Science Review*, 42, 100431.
- Kemp, S., Miró-Llinares, F., & Moneva, A. (2020, September). The dark figure and the cyber fraud rise in europe: Evidence from spain. *European Journal on Criminal Policy and Research*, 26(3), 293–312.
- Leukfeldt, E. R., Kleemans, E. R., Kruisbergen, E. W., & Roks, R. A. (2019, September). Criminal networks in a digitised world: on the nexus of borderless opportunities and local embeddedness. *Trends in Organized Crime*, 22(3), 324–345.
- Leukfeldt, E. R., Kleemans, E. R., & Stol, W. P. (2016, February). Cybercriminal networks, social ties and online forums: Social ties versus digital ties within phishing and malware networks. *Br. J. Criminol.*, azw009.
- Leukfeldt, E. R., Kruisbergen, E. W., Kleemans, E. R., & Roks, R. A. (2020). Organized financial cybercrime: Criminal cooperation, logistic bottlenecks, and money flows. In T. J. Holt & A. M. Bossler (Eds.), *The palgrave handbook of international cybercrime and cyberdeviance* (pp. 961–980). Cham: Springer International Publishing.
- Leukfeldt, E. R., Lavorgna, A., & Kleemans, E. R. (2017, September). Organised cybercrime or cybercrime that is organised? an assessment of the conceptualisation of financial cybercrime as organised crime. *European Journal on Criminal Policy and Research*, 23(3), 287–300.
- Leukfeldt, E. R., & Roks, R. A. (2021, November). Cybercrimes on the streets of the netherlands? an exploration of the intersection of cybercrimes and street crimes. *Deviant Behav.*, 42(11), 1458–1469.
- Leukfeldt, E. R., Veenstra, S., & Stol, W. P. (2013, January). High volume cyber crime and the organization of the police: The results of two empirical studies in the netherlands. *International Journal of Cyber Criminology*, 7(1), 1–7.
- Leukfeldt, R., Kleemans, E., & Stol, W. (2017, October). The use of online crime markets by cybercriminal networks: A view from within. *Am. Behav. Sci.*, 61(11), 1387–1402.
- Lloyd, S. (1982, March). Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2), 129–137.
- Lukasz Piwek, A. J. (2016, January). “what do they snapchat about?” patterns of use in time-limited instant messaging service. *Computers in Human Behavior*, 54, 358–367.
- MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66,

- 1, 281-297 (1967).
- Malin, C. H., Gudaitis, T., Holt, T. J., & Kilger, M. (2017, January). 4 - social dynamics of deception: Cyber underground markets and cultures. In C. H. Malin, T. Gudaitis, T. J. Holt, & M. Kilger (Eds.), *Deception in the digital age* (pp. 125–148). Boston: Academic Press.
- McKinney, W., & Others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).  
*Message*. (n.d.). <https://tl.telethon.dev/constructors/message.html>. (Accessed: 2022-7-14)
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007, October). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on internet measurement* (pp. 29–42). New York, NY, USA: Association for Computing Machinery.
- Miss rose*. (n.d.). <https://missrose.org>. (Accessed: 2023-4-4)
- Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011, November). An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on internet measurement conference* (pp. 71–80). New York, NY, USA: Association for Computing Machinery.
- Moyle, L., Childs, A., Coomber, R., & Barratt, M. J. (2019, January). #drugsforsale: An exploration of the use of social media and encrypted messaging apps to supply and access drugs. *Int. J. Drug Policy*, 63, 101–110.
- New major interventions to block encrypted communications of criminal networks*. (2021, March). <https://www.europol.europa.eu/media-press/newsroom/news/new-major-interventions-to-block-encrypted-communications-of-criminal-networks>. (Accessed: 2022-5-4)
- Obar, J. A., & Wildman, S. (2015, October). Social media definition and the governance challenge: An introduction to the special issue. *Telecomm. Policy*, 39(9), 745–750.
- Odinot, G., Verhoeven, M. A., Pool, R. L. D., & de Poot, C. J. (2017). Organised cybercrime in the netherlands.
- Onuch, O., Sasse, G., & Michiels, S. (2023, April). Flowers, tractors, & telegram: Who are the protesters in belarus?: A survey based assessment of Anti-Lukashenka protest participants. *Natl. Pap.*, 1–26.
- Opiumwet*. (n.d.). <http://wetten.overheid.nl/jci1.3:c:BWBR0001941>. (Accessed: 2023-4-12)
- pandas.DataFrame.sample* — *pandas 2.0.2 documentation*. (n.d.). <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html>. (Accessed: 2022-9-10)
- Park, N., Kee, K. F., & Valenzuela, S. (2009, December). Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *Cyberpsychol. Behav.*, 12(6), 729–733.
- Patel, E., & Kushwaha, D. S. (2020, January). Clustering cloud workloads: K-Means vs gaussian mixture model. *Procedia Comput. Sci.*, 171, 158–167.
- Paul Vogt, W. (1999). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. SAGE Publications.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of*

- Machine Learning Research*, 12, 2825–2830.
- Privacy policy*. (2019). <https://telegram.org/privacy>. (Accessed: 2022-NA-NA)
- Project Jupyter*. (n.d.). <https://jupyter.org/>. (Accessed: 2022-7-15)
- Qiu, J., Li, Y., Tang, J., Lu, Z., Ye, H., Chen, B., . . . Hopcroft, J. E. (2016, April). The lifecycle and cascade of WeChat social messaging groups. In *Proceedings of the 25th international conference on world wide web* (pp. 311–320). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.
- Reinventing group chats: Replies, mentions, hashtags and more*. (2015, March). <https://telegram.org/blog/replies-mentions-hashtags>. (Accessed: 2022-6-1)
- Schlette, A., van Prooijen, J.-W., Blokland, A., & Thijs, F. (2022, October). The online structure and development of posting behaviour in dutch anti-vaccination groups on telegram. *New Media & Society*, 14614448221128475.
- Seufert, M., Hoffeld, T., Schwind, A., Burger, V., & Tran-Gia, P. (2016, May). Group-based communication in WhatsApp. *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, 536–541.
- Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means clustering algorithm. *IEEE Access*, 8, 80716–80727.
- sklearn.cluster.KMeans*. (n.d.). <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>. (Accessed: 2022-12-10)
- sklearn.feature\_extraction.text.CountVectorizer*. (n.d.). [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.CountVectorizer.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html). (Accessed: 2022-9-8)
- sklearn.feature\_extraction.text.TfidfTransformer*. (n.d.). [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfTransformer.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html). (Accessed: 2022-9-8)
- sklearn.linear\_model.SGDClassifier*. (n.d.). [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.SGDClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html). (Accessed: 2022-9-8)
- sklearn.mixture.GaussianMixture*. (n.d.). <https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html>.
- Soska, K., & Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th USENIX security symposium (USENIX security 15)* (pp. 33–48).
- Soudijn, M. R. J., & Zegers, B. C. H. (2012, September). Cybercrime and virtual offender convergence settings. *Trends in Organized Crime*, 15(2-3).
- Telegram bots rating*. (n.d.). [tgbots.io](https://tgbots.io). (Accessed: 2023-4-4)
- Telegram FAQ*. (n.d.). <https://telegram.org/faq>. (Accessed: 2022-5-31)
- Telethon API*. (n.d.). <https://tl.telethon.dev/>. (Accessed: 2022-7-15)
- Telethon user*. (n.d.). <https://tl.telethon.dev/constructors/user.html>. (Accessed: 2022-7-14)
- Thach, L., & Olsen, J. (2015, June). Profiling the high frequency wine consumer by price segmentation in the US market. *Wine Economics and Policy*, 4(1), 53–59.
- Van Buskirk, J., Naicker, S., Roxburgh, A., Bruno, R., & Burns, L. (2016, September). Who sells what? country specific differences in substance availability on the agora cryptomarket. *Int. J. Drug Policy*, 35, 16–23.

- Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.
- van Wegberg, R., Miedema, F., Akyazi, U., Noroozian, A., Klievink, B., & van Eeten, M. (2020, April). Go see a specialist? predicting cybercrime sales on online anonymous markets from vendor and product characteristics. In *Proceedings of the web conference 2020* (pp. 816–826). New York, NY, USA: Association for Computing Machinery.
- van Wegberg, R., Tajalizadehkhoob, S., Soska, K., Akyazi, U., Ganan, C. H., Klievink, B., ... van Eeten, M. (2018). Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *27th USENIX security symposium (USENIX security 18)* (pp. 1009–1026).
- Van Wegberg, R., & Verburgh, T. (2018). Lost in the dream? measuring the effects of operation bayonet on vendors migrating to dream market. In *Proceedings of the evolution of the darknet workshop* (pp. 1–5).
- Wall, D. S. (2017). Towards a conceptualisation of cloud (cyber) crime. In *Human aspects of information security, privacy and trust* (pp. 529–538). Springer International Publishing.
- Waskom, M., Botvinnik, O., O’Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., ... Qalieh, A. (2017, September). *mwaskom/seaborn: v0.8.1 (september 2017)*. Zenodo.
- Weber, J., & Kruisbergen, E. W. (2019, September). Criminal markets: the dark web, money laundering and counterstrategies - an overview of the 10th research conference on organized crime. *Trends in Organized Crime*, 22(3), 346–356.
- Wijermars, M., & Lokot, T. (2022, March). Is telegram a “harbinger of freedom”? the performance, practices, and perception of platforms as political actors in authoritarian states. *Post-Soviet Affairs*, 38(1-2), 125–145.
- Williams, S. (2020, August). *Belarus has torn up the protest rulebook. everyone should listen*. <https://www.wired.co.uk/article/belarus-protests-telegram>.
- Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., & Zhao, B. Y. (2009, April). User interactions in social networks and their implications. In *Proceedings of the 4th ACM european conference on computer systems* (pp. 205–218). New York, NY, USA: Association for Computing Machinery.
- Yang, X.-H., Xiong, Z., Ma, F., Chen, X., Ruan, Z., Jiang, P., & Xu, X. (2021, July). Identifying influential spreaders in complex networks based on network embedding and node local centrality. *Physica A: Statistical Mechanics and its Applications*, 573, 125971.
- Zamani, M., Rabbani, F., Horicsányi, A., Zafeiris, A., & Vicsek, T. (2019, July). Differences in structure and dynamics of networks retrieved from dark and public web forums. *Physica A: Statistical Mechanics and its Applications*, 525, 326–336.