MASTER

Characterization of illegal dark web arms markets

Ubbink, J.R.

Award date:
2019

Link to publication

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
Characterization of Illegal Dark Web Arms Markets

*Masters’ Thesis*

Jeroen Ubbink

*Supervisor:*
Dr. Luca Allodi

*Assessment Committee:*
Dr. Luca Allodi
Dr. Alexander Serebrenik
Dr. Decebal Mocanu

Eindhoven, September 2019
Abstract

The nature of online underground gun markets on the dark web has been relatively under-researched in comparison to those regarding drugs or malware. This work attempts to improve the general understanding of the nature of these markets, with a longitudinal assessment of the market as a whole. From this assessment, the various properties that characterize the market such as overall sales and the breadth of items on offer can be catalogued and compared against offline markets, or other online markets.

In addition to this longitudinal study, the online communities surrounding the sale of firearms were identified, with topic models fit to the datasets spanning approximately five years, with the intent of characterizing and comparing them to each other in a more structured manner. Once the topic models were generated, documents were drawn from before and after mass shooting attacks. These documents were then labeled by the separate topic models, and then contrasted and compared against each other in order to assess the reactions of these communities to traumatic events, thus observing if there were clear patterns of behavior universal across these communities.

Online underground arms markets were found to be generally thin, albeit larger in scale than a few years before, and appear to be predominantly focused on the sale of rifles, pistols, and custom orders. Gun communities online were observed to differ depending on the strictness of moderation of their parent communities, though still have a number of shared topics, such as gun legislation or usage. Furthermore, the assessed communities varied heavily in their reactions to attacks, further highlighting their differences.
Preface

The following work is the thesis “Characterization of illicit dark web arms markets”, a paper that attempts to characterize the sale of illegal arms markets on the dark web, in addition to the online communities that surround them. This work has been written as a thesis fulfilling the graduation criteria for the Computer Science and Mathematics Masters program at the Technical University of Eindhoven, and took approximately ten months from inception to completion, from November 2018 until September 2019.

My experience with the dark web before this work was theoretical at best, and as far as an introduction to the scale and nature of this particular subset of the online world as gone, it could not have been more engaging or challenging. Although deciding on the research questions was tough due to the sheer scale of the potential topics and fields of research that could be combined, I could not have been happier with the final outcome.

First and foremost, I would like to thank Dr. Allodi for his guidance and advice throughout the entire process, and for the opportunity to explore a truly unique area of research. I would never have explored this topic, nor the entire field of natural language processing if not for his advice and recommendations. I always said that I wanted to do something interesting, and this definitely far surpassed my expectations!

I am immensely grateful to Katinka and Jesper for their unwavering support over the past few months, never even thinking twice about sacrificing their time and energy to help me through the most stressful time in my life. Whatever I needed, they did not hesitate to offer, and I could not have done this without them. I would also like to thank Tim, for always offering his advice and expertise with Python programming where my admittedly lackluster skills in that department fell short.

Finally, my thanks go to my girlfriend and family, for being patient and grounded throughout the long road this thesis has taken, and for always being there for me, no matter where in the world you were.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Research questions &amp; Overview of approach</td>
<td>1</td>
</tr>
<tr>
<td>2 Background</td>
<td>3</td>
</tr>
<tr>
<td>2.1 Weapon classifications and taxonomy</td>
<td>3</td>
</tr>
<tr>
<td>2.1.1 Small arms</td>
<td>3</td>
</tr>
<tr>
<td>2.1.2 Light Weapons &amp; Explosives</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Offline gun markets</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2.2.2 How do firearms enter underground markets?</td>
<td>9</td>
</tr>
<tr>
<td>2.2.3 Market “thin”-ness</td>
<td>10</td>
</tr>
<tr>
<td>2.2.4 Weapon characteristics</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Online weapon markets</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Overview</td>
<td>13</td>
</tr>
<tr>
<td>2.3.2 Where are weapons sold?</td>
<td>14</td>
</tr>
<tr>
<td>2.3.3 Logistic aspects</td>
<td>14</td>
</tr>
<tr>
<td>2.3.4 Who buys firearms on the dark web?</td>
<td>15</td>
</tr>
<tr>
<td>2.3.5 Size of online weapon markets</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Comparison of offline and online weapon markets</td>
<td>16</td>
</tr>
<tr>
<td>3 Methodology</td>
<td>17</td>
</tr>
<tr>
<td>3.1 Dark web market identification</td>
<td>17</td>
</tr>
<tr>
<td>3.1.1 Berlusconi market</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2 Single vendor markets</td>
<td>19</td>
</tr>
<tr>
<td>3.2 Social community identification</td>
<td>19</td>
</tr>
<tr>
<td>3.2.1 Community selection</td>
<td>20</td>
</tr>
<tr>
<td>3.2.2 Private communities</td>
<td>22</td>
</tr>
<tr>
<td>3.3 Analysis methods</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1 Preprocessing &amp; Latent Dirichlet Allocation</td>
<td>23</td>
</tr>
<tr>
<td>3.4 Topic analysis</td>
<td>24</td>
</tr>
<tr>
<td>3.4.1 Generating list of shocks</td>
<td>24</td>
</tr>
<tr>
<td>3.5 Social communities</td>
<td>26</td>
</tr>
<tr>
<td>3.5.1 List of shocks</td>
<td>27</td>
</tr>
<tr>
<td>4 Results &amp; Analysis</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Dark web markets</td>
<td>29</td>
</tr>
<tr>
<td>4.1.1 Berlusconi market</td>
<td>29</td>
</tr>
<tr>
<td>4.1.2 Single vendor market</td>
<td>33</td>
</tr>
<tr>
<td>4.1.3 Discussion</td>
<td>35</td>
</tr>
<tr>
<td>4.2 Analysis of community discussion topics</td>
<td>39</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Over the past decade, the development and growth of anonymity networks have birthed the “dark web”, an area of the internet un-indexed by traditional search engines such as Google. Instead, in order to connect to resources and websites hosted on the dark web, access to the “Tor” network is required, which conceals communications between various parties using the network. Due to the anonymity that comes with using Tor, the dark web has developed a reputation as a hotbed for the transaction of illegal goods and services, purchased through ostensibly anonymous cryptocurrencies such as Bitcoin. These marketplaces (known as cryptomarkets) are similar in purpose and format to eBay or other clear web auction sites as the primary nexuses of trade for the dark web.

It is however unclear how these markets differentiate offers of illegal goods; specifically, while most studies have focused on the nature of the drug trade, very little has been done on the online arms trade. Part of the reason appears to be the relatively small size of the arms market in comparison [37], leading to the belief that markets are not significant in the overall landscape of the global illegal arms trade. In fact, an argument can be made that the relationships and nature of these markets and their communities are especially relevant, considering the alleged use of weapons purchased online in recent attacks [65]. Therefore, the case is made in this thesis that such markets for illegal weapons being generally thin is intrinsic to their nature, regardless of location, and thus the true relevance and relationship of these markets to criminal activity remains to be explored. Furthermore, when looking at other underground markets on the dark web, many of them have communities surrounding them where buyers and sellers interact and transact. For example, the rise of malware as a service [54] has been documented, with a number of underground markets for malware developers having been documented and observed [17]. Underground gun markets may have similar communities, and no studies thus far have attempted to shed light on their existence, let alone nature. Moreover, with the increased incidence of mass shootings across the globe, the reaction and response of these communities to an attack may be telling, and could even potentially allow for prediction of some attacks, in the best possible case scenario.

1.1 Research questions & Overview of approach

As the above statements are extremely broad, it is prudent to explicitly state the intent of this thesis. In short, the objective of this thesis is in order to improve the characterization of illegal dark web arms markets and the individuals that purchase weapons from them. As this is still rather loose in terms of concrete objectives, the following research questions have been formulated:

RQ1 How do the economic characteristics of illegal dark web arms markets change over time?

RQ2.1 What topics are communities that are more likely to have buyers of illegal arms discussing?

RQ2.2 How do these topics change when firearm-related terrorist attacks occur?
CHAPTER 1. INTRODUCTION

The goal of research question 1 is to improve on the fairly spare research regarding dark web arms markets, in the hope of identifying and characterizing their aspects at present time, and providing a snapshot that can be used for future study.

These questions will be answered and substantiated through a bipartite methodology. the first method, in order to answer research question one requires identifying and infiltrating the markets involved with the illicit trafficking of arms on the dark web. A comprehensive search of the dark web (insofar as it is possible) will be performed in order to identify various dark web markets that are involved with arms sales. Using a scraper in conjunction with a parser, the various properties of the available listings and vendors will be collected, recorded, and then analyzed.

In parallel, the communities surrounding online arms discussions will be infiltrated and observed. Any public posts or comment threads will be recorded and then parsed using topic analysis tools in order to identify the primary topics of discussion.

Initially, however, it is important to contextualize the background academic research for both offline and online gun markets, in order to establish a baseline understanding of the nature of both dark web markets and offline arms markets, and comprehend why dark web arms markets are so under-researched in comparison.
Chapter 2

Background

This chapter establishes the context and background of the sale of illegal firearms in both offline and online contexts. A surface overview of categories of often trafficked firearms is established together with potentially unfamiliar terms, followed by a summary of the characteristics of offline gun markets.

2.1 Weapon classifications and taxonomy

Commonly trafficked weapons online are generally limited to firearms that fall within the criteria of “Small Arms and Light Weapons” [52], defined as man-portable lethal weapons that expel or launch, are designed to expel or launch, or may be readily converted to expel or launch a shot, bullet of projectile by way of explosive action. The vast majority of these weapons can be loosely classified into the following groups, and will be further explained in detail in each subsection.

- Handguns
- Shotguns
- Rifles
- Carbines
- Sub-machine Guns
- Personal Defense Weapons
- Light Machine Guns

2.1.1 Small arms

Small arms are defined as weapons designed for individual use, and are split into the following commonly accepted categories with individual definitions to follow.

In general, firearms require the following components in order to function: A chamber to hold the cartridge or round. A cartridge or round is the packaging of a projectile (either a shell, bullet, or shot) together with propellant and an ignition device. The chamber is attached directly to one or more barrels, which may or may not be rifled. There is commonly a trigger assembly which contains the actual trigger, and initiates the firing action, usually done by spring loaded mechanism that creates an impact between the firing pin and the back of the round that is currently loaded within the chamber. When the trigger is pulled, The ignition device is struck, igniting the propellant and firing the projectile out of the end of the barrel, with the remains of the now spent cartridge resting in the chamber. Loading the next cartridge into the chamber is then done, cycling out the remains of the previous, which can either be done through manual action or automatically, depending on the type of weapon. If automatic, the remains are cycled...
CHAPTER 2. BACKGROUND

by using the exhaust gases from the previous round to clear the chamber, and to load the next round from an externally easily replaceable magazine.

Automatic weapons are defined as weapons that will continue to fire and replace chambered rounds for as long as the trigger is depressed or until there are no more rounds in the magazine. Semi-automatic weapons share many characteristics with automatic weapons, such as being magazine fed and automatic cycling the next round, but will usually only fire one round per trigger pull. This functionality is similar to manually loaded weapons, but they can often be toggled to automatic fire mode through a selector switch that allow automatic, semi-automatic, and burst fire modes. The latter is where the weapon fires a small number of rounds (usually three) per trigger pull, in order for the shooter to better control the recoil of the weapon, which can be a useful trade off between accuracy and firepower. The recoil of a weapon tends to be a major limiting factor of weapons when it comes to accuracy, since the greater the recoil of the weapon, the harder it is for the shooter to re-center their weapon on target. Recoil is determined by three factors, the mass of the gun, the mass of the projectile, and the speed of the projectile, thus, the heavier the weapon, the less the recoil, while the mass and speed of the projectile increase the recoil.

**Handguns**

Handguns are commonly defined as short-barreled firearms designed to be fired with only one hand, and include revolvers, semi-automatic pistols, and machine pistols. Revolvers are handguns with a rotating chamber that hold a limited number of individual rounds, with one or more barrels, allowing the shooter to fire multiple times without reloading. Handguns are small, portable firearms, and are thus easily concealable and used for either personal protection reasons by individuals (such as police officers), or by criminals. They fulfill a niche as a secondary weapon for heavily armed military personnel in the field, in case their primary weapon malfunctions.

A semi-automatic pistol is defined as a singly chambered and barreled weapon that uses the energy of a fired shot in order to eject the spent shell and load the next cartridge from a magazine. This class of weapon normally operates under a single-trigger pull for a single round policy, thus depressing the trigger once (regardless of how long the trigger is held) will only fire one round. The rounds are loaded through a magazine, usually inserted in the grip, which must be manually reloaded when emptied.

Machine pistols are a type of fully automatic handgun. For machine pistols the restriction of a single trigger pull per shot is removed, resulting in shots being fired for as long as the trigger is depressed or until the magazine is empty. Machine pistols are rarely used due to their small size and high recoil making it difficult for inexperienced shooters to control accurately. As a result, these weapons are generally considered a special purpose weapon primarily for use by either military or police special forces. However, there is minor demand from individuals or groups that are interested in weapons as a status symbol more than for precision shooting. This is likely due to their high rate of fire resulting in a very striking effect when fired as part of display of violence.
CHAPTER 2. BACKGROUND

Figure 2.2: A semi-automatic glock pistol

intended for psychological effect moreso than the actual number of rounds accurately fired at a target, e.g. a drive-by shooting by a gang.

Figure 2.3: A fully automatic Heckler & Koch MP7 Machine Pistol

Shotguns

Shotguns are firearms that are shoulder fired, discharging either a shell that contains a number of small spherical pellets (commonly referred to as shot) or a single solid projectile (commonly known as a slug). Beyond this core definition, there are a large number of differences between classifications of shotgun. Shotguns tend to have either one or two barrels, and can be manually or automatically loaded. The manually loaded designs utilize one of the following methods for loading: breech-loading, pump-action, lever-action, and revolver actions. Breech-loading is where the shooter must manually load each cartridge directly into the chamber after every shot. Pump-action loading mechanisms are where a forend is moved (pumped) forward then backward in order to eject the remains of the spent shell and load a fresh shell into the chamber after every shot. Lever-action loading mechanisms are similar to pump-action, where after every shot the shooter manually pulls a lever to eject the spent round and load a fresh round into the breech. Revolver action shotguns are similar to revolver handguns as in section 2.1.1, where multiple rounds are in a rotating cylinder that cycles with every trigger pull, which loads a fresh round into the chamber. Finally, there are also semi-automatic and fully automatic varieties, which do not differ in function from the mechanisms outlined in section 2.1.1, merely scale. Generally, manual loading mechanisms are more durable in inclement conditions, such as sand, water, or other climactic extremes [64], but are slower in terms of rate of fire than semi-automatic or fully
automatic mechanisms.

Shotguns are bought by civilians for home defense, sport shooting or hunting purposes. They are suited to these roles due to the spread of shot pellets upon release from the barrel increasing the chance of hitting a target even at range. When the target is a disc such as in skeet shooting or similarly small and fragile targets such as birds or wildlife, a single pellet impacting is likely sufficient for the shooter’s purposes. The spread pattern also makes shotguns extremely useful as close quarter weapons due to the more forgiving accuracy, since having a majority of shot pellets (instead of a singular projectile) make it more likely to stop a target at close range.

Rifles

Rifles are long-barreled firearms that are characterized by the ‘rifling’ of the barrel. Rifling is a spiral has been machined into the inside of the barrel in order to impart a gyroscopic spin to the projectile as it is fired, improving accuracy by preventing it from tumbling in flight. This innovation allowed for more aerodynamic shapes than spheres to be used as projectiles (leading to the rise of the classic ‘bullet’ shape and distinguishing rifles from the previous unrifled musket design), resulting in better range and accuracy.

Modern rifles generally fall into two categories: Manual and automatic, labeled as such by the loading mechanism. This categorization is similar to shotguns, with the addition of bolt-action rifles to the manual category, where after the shooter fires, they then retract (via a handle) a bolt that secures the round during firing in order to extract the spent round and load the next round (either directly or from a magazine, depending on design). Manual loading designs are less prone to jamming, where the casing of the previous round has been improperly ejected. A jam prevents the new round from being loaded, and must be cleared before the weapon can fire again. Most hunting rifles or sniper rifles that are generally prized for accuracy and range are manually loaded.

Semi-automatic and automatic rifles are generally considered to be assault weapons, which are optimised for size and portability while still maintaining a high rate of fire and stopping power over relatively short ranges. They also tend to have select-fire modes, changing them from single-shot to burst-fire or fully automatic fire for versatility depending on engagement conditions. Note that in gun control discussions, the term assault weapon and assault rifle are not synonymous, with assault weapons being a set of firearms that include assault rifles, but also certain automatic shotguns and pistols that accept magazines. The exact properties that define which weapons are considered 'assault weapons' vary on a per-country basis.
Rifles as a class of small arm are generally used due to their range and accuracy, and the wide range of situations where they are considered to be useful. Because of their increased stability when shoulder fired over a handgun or some smaller weapon, they generally use larger sized (or caliber) rounds. This effect, together with the greater mass of the weapon decreases the overall recoil of the weapon, and gives rifles greater stopping power and range. Stopping power is defined as the ability of a firearm to incapacitate targets, thus stopping them. This effect is a prime concern in military and law enforcement environments, since disabling a target tends to be a more important concern than lethality in order to end the engagement as quickly as possible. Stopping power is determined not by the force of the bullet but by the damaging effects of an impact, namely that of blood loss and the accompanying loss of blood pressure. This effect has led to the development of specialty expanding rounds or hollow point rounds, both of which have been banned by the Hague Convention [6]. Large caliber assault weapons that have higher rates of fire thus have greater stopping power.

**Carbines**

Carbines are a relatively long-barreled firearm that is shorter than a rifle. Many carbines are shortened makes of full rifles, and are generally lighter and easier to handle. Their effective accurate range suffers as a result, but with the change in scope of modern military engagements, carbines excel in urban environments with both short and extended range engagements. Until recently, carbines were mostly used by military special forces units, due to the mix of heavier stopping power than handgun caliber weapons, while not suffering from the weight and unwieldiness of full rifles, especially in close-quarters. However, with the change over recent years in types of engagements
by military forces (generally shifting towards more urban and close-quarter engagements), carbines have become the primary weapon issued by most US military branches [48].

![Figure 2.7: An M4 Carbine](image)

**Sub-machine Guns**

The weapons that fall under this classification are generally unpopular, having been superseded in their role by assault weapons, due to being similar in size to carbines, but suffering from a lack of stopping power in comparison. This change is attested to the difference in caliber between submachine guns and assault rifles, with the former using smaller pistol-caliber cartridges. It is important to note that the line differentiating between machine pistol and sub-machine gun is generally difficult to draw, with many weapons technically falling into both categories.

**Personal Defense Weapon**

Personal defense weapons are a fairly new classification of weapon that are compact, magazine-fed sub-machine gun like firearms. They cross the line between assault rifle and sub-machine gun, using larger, rifle caliber rounds instead of the pistol-caliber rounds for sub-machine guns.

![Figure 2.8: An FN-P90 Personal Defense Weapon](image)

**Light Machine Guns**

Light Machine Guns are defined as automatic weapons designed to be operated and employed by an individual soldier as an infantry support weapon. They excel for the purpose of suppressive fire in order to advance on enemy held territory, and as a result are generally solely used by military forces. They are large in profile and extremely heavy and cumbersome to deploy, and thus are fairly unlikely to be trafficked in any great number to civilians.
2.1.2 Light Weapons & Explosives

This covers a wide range of weapons, most of which are military in scope and are thus generally difficult for a civilian or a commercially licensed gun shop to procure. They are generally designed for use by two or three persons serving as a crew, which already limits the use case and interest for the purposes of this background. Therefore, and in addition to the large number of different classes of weapon within this category, only the definitions of light weapons that are found to be for sale are included.

Insert found classifications here

Explosives

The number of different types of explosives is vast, but for the scope of this project, the actual chemical or eruptive qualities of the explosives on offer are not particularly relevant. With that said, however, the existence of another avenue of obtaining explosives for creating car bombs, pipe bombs, or Improvised Explosive Devices (IED) is worrying due to the terror and fear that such weapons generate when used in terrorist attacks.

Insert found classifications here

2.2 Offline gun markets

2.2.1 Introduction

Legal firearm markets are split into two categories: Primary and Secondary. Primary markets usually are composed of transactions from legally licensed gun sellers, which results in a market of vendors that obey the legal requirements of their geographical jurisdiction. Within the USA, this license is a Federal Firearms License (FFA), and allows an individual or company to engage in the business of manufacture or importation of firearms. Thus, a sale from a FFL licensee involves performing background checks of potential buyers, in addition to reporting to authorities if certain conditions occur, such as multiple arms sales in a single day [50]. This ensures that these markets (in theory) will only sell weapons to individuals or organizations that are legally afforded the right to own firearms.

However, secondary markets generally consist of transactions between unlicensed individuals, meaning person to person deals. The exact distinction between primary and secondary market sellers varies greatly from one country to another, or even between states. In the USA, generally, secondary markets consist of individuals that are not primarily employed in the business of firearm sale and manufacture occasionally selling weapons in their collection to others, and thus are not mandated to perform checks that an FFL holder would be required to. The most publicized avenues for secondary markets sales in the US is that of gun shows, which are events where large arenas are rented and buyers and sellers pay an entry fee in order to transact in weapons sales without the requirements associated with buying or selling as a licensed vendor, even if they are an FFL holder [24].

Note that not every country has legal secondary markets, however. For example, Belgium, a country with an estimated 1.45 million firearms in civilian possession [40], does not legally allow for sale of firearms to individuals either online or via mail, or through any public or private market [20]. Consequently, only 426,939 [40] of those are registered, with the rest being unregistered and thus illegal. Exploring the nature of illegal underground gun markets, therefore, is the focus of the following section, as their characteristics differ significantly from legal gun markets.

2.2.2 How do firearms enter underground markets?

Guns enter illegal underground gun markets through a number of different vectors, ranging from trafficking rings involving nominally authorized dealers, to simple secondary market sales where
the seller is not required to ensure that the buyer is legally allowed to own a weapon. These can be loosely categorized into “point sources” and “diffuse sources”. Point sources are named as such due to the high density of weapons (i.e. each source provides a relatively large number items) that enter the market through these methods, such as trafficking rings or unlawful dealers that have access to numerous weapons. Diffuse sources are more low-level and widespread, such as theft or secondary market sales [57]. It is interesting to note that even in environments with a large legal secondary market and general public acceptance of firearms, illegal sources, i.e. theft, been estimated to be the source of up to 50% of all firearms involved with criminal activity [61]. Unfortunately, the current scientific consensus is still not in accord as to the impact or significance of secondary vs primary markets, or even the overall prevalence of weapons in a society as contributing factors to gun violence.

2.2.3 Market “thin”-ness

Regardless of the various avenues through which weapons enter underground markets, it is apparent that weapons are still remarkably difficult to procure illegally, even in countries with relatively large secondary markets. In order to further restrict This has resulted in those markets being remarkably “thin”, defined as a market with a low number of buyers and sellers. Gun markets are a typical example of a thin market [27], with high price volatility (i.e. rapid price changes depending on supply of goods) and low liquidity, with the demand for guns being limited. This demand is theorized to be due to a number of different factors, the most important of which are police pressure, gang presence, and market setting.

**Police pressure**

With economic and policy incentives aligning heavily with the war on drugs, police operations in areas with high criminality tend to focus on suppressing gangs that traffic in drugs [18]. Considering that the drug trade is far and away is the largest source of income for gangs [43], police pressure that is aimed at reducing the number of guns on the street heavily jeopardizes the income from drug sales if the gang is careless about its storage of and trafficking in weapons [27]. Furthermore, if a gang member is found in possession of a gun by the authorities, the gang that they hold membership of is often cracked down on viciously, which in combination with the low demand, disincentivizes the gangs from buying and selling illegal weapons both to gang members and other criminal elements. This disparity between how the police appear to react to the sale of weapons versus the more “standard” trade in drugs may be because of the criteria police departments use to value seizures of illegal goods, with weapons being rarer and thus more valuable if taken off the street, for example. In short, the highly lucrative trade in drugs has had a serious collateral effect and is therefore a major factor in the restricted supply of weapons in underground markets.

**Gang presence**

Guns in the hands of gangs allow them to maintain a monopoly of deadly force in neighborhoods that they control, which reduces the overall amount of armed violence. This effect allows for nonviolent criminal behavior (i.e. drug dealing) and entrenches the gangs further within their turf without actually needing to use firearms. Seemingly paradoxically, however, gangs tend not to act in a manner that actively reduces the number of gun sales on their turf. Instead, they are more interested in information about who, if anyone, is selling weapons, and taxing those transactions [60]. This directly increases the prices of weapons, ensuring that the vast majority of firearms in a neighbourhood are controlled by gangs, while being able to profit from gun sales without being at risk from police activity. This has the additional side effect of increasing the attractiveness of gang membership to youths that reside in the area, who become members due to wanting the prestige of association and shared ownership of firearms without having to pay the high cost associated with acquiring a firearm. Gang membership promises much coveted access to weapons for a negligible monetary cost, resulting in gangs being able to use the desire to display
social status through gun ownership to control their members. Furthermore, due to the value of
the firearms, the leadership can keep the guns within the gang safe by only lending weapons to
trusted gang members that have been part of the gang for a considerable amount of time, further
incentivizing members that joined due to the glamour of ownership to stay and become further
lodged in the criminal ecosystem. Furthermore, there appears to be a degree of collaboration
between the police and gang leaders. Gang leaders will occasionally report on their own gang
members that have obtained guns for themselves in an “unsanctioned” manner, in order to return
the wayward member to the fold with the gun being confiscated by the authorities and taken off
the street, but without an arrest being made [27], thus preserving the authority and monopoly of
the gang.

Market setting

The end result is that weapons on underground gun markets are hard to obtain, with interviewees
stating that it would take upwards of a week to obtain a weapon on demand, and expensive, with
markups of between $100 to $200 for cheap second-hand handguns. In other US states such as
Massachusetts, handguns were estimated to be marked up 191% over their fair market value when
sold illegally, and 64% over their suggested retail price [28]. In areas with even more stringent
gun control laws, such as the UK, the cost of a new Glock or Beretta 9 mm pistol was around
£1,000 to £1,400, which is at least double if not triple the cost of a handgun from an authorized
vendor [34].

These “unauthorized” sellers vary in profession, from fences to gun brokers, gangs, and drug
dealers. These vendors, when combined with individuals that only sell firearms to buyers that
they have a social relationship with, results in about 70% of total illegal transactions. Gifting guns
is also not uncommon, but for the majority of confiscated illegal weapons, the supply chain runs
through a series of secondary market transactions, which transitions at a certain point to someone
prohibited from having that weapon, and then stays in the criminal ecosystem. Presumably, the
profile of weapon sales in countries where secondary markets are illegal would differ, perhaps
starting with a theft of a firearm instead, a far less likely occurrence and giving some reason for
the significantly higher cost of obtaining a gun in those circumstances. It does appear that the
entry of weapons into underground gun markets tends to be generally through individuals that
resell legally bought weapons, purchased from out of state where gun control regulations are less
stringent, stolen either from gun stores or from individuals, or finally, allegedly from corrupt police
officers [28].

First and foremost, purchasing illegal firearms seems to be primarily founded on trust rela-
tionships [28], with individuals selling or trading primarily with other individuals that they have
met beforehand, and thus have some kind of personal or professional relationship with. It follows,
therefore, that the more relationships an individual has (such as gang association) which can be
leveraged transitorily into more trust relationships are a strong indicator for ease of obtaining a
weapon. Indeed, social networking tools indicate that gang membership decreases the distance
(in terms of number of contacts) to the nearest gun broker by 27%, reinforcing the point made
that gangs are a conduit for firearms into neighborhoods [60]. Furthermore, when looking at the
demographics of individuals purchasing weapons from underground gun brokers, they tend to be
predominantly young men from low income backgrounds. As such, they tend to be geographically
restricted, thus limiting the number of gun brokers that they could purchase weapons from inde-
pendently. This paints a picture that implies that if an individual is not already strongly involved
with illegal activity but is barred for whatever reason from legally purchasing firearms, it will be
(rightfully) difficult for them to obtain a weapon.

2.2.4 Weapon characteristics

Weapons sold in underground gun markets appear to traffic in semiautomatic pistols and revolvers
by an extremely large majority, with 94.2% of guns obtained from gangs over a 6 year period falling
into those categories as seen in table 2.1. Weapons confiscated from non-gang affiliated individuals
show a similar trend, with 84.8% of total weapons confiscated falling into those categories. Interestingly, weapons that are illegal in Massachusetts, namely assault rifles are slightly more likely to be owned by non-gang affiliated individuals than by gang members, with 6.8% versus 2.4%. Furthermore, weapons from disreputable or low quality manufacturers are surprisingly heavily marked up, implying either that buyers are not especially gun savvy, or do not see the need for high quality firearms, and are just obtaining whatever they can get. Another interesting datum is that new weapons did not garner an increased cost from individuals on the street, implying that despite the guarantee that those weapons were not involved in previous criminal activity, the market does not place a premium on that attribute.

To conclude, purchasing weapons from underground gun markets are surprisingly difficult. Purchasing firearms without gang membership is relatively rare, with both gangs and authorities working to limit or curb weapons within a given neighborhood for their own (very different) intents. As a direct consequence, purchasing weapons illegally requires some degree of pre-existing trust, either with gangs, gun brokers, fences, or anyone else on the “street”. This problem is likely exacerbated in locales that already have extremely restricted secondary markets.
2.3 Online weapon markets

The question then arises, what are the options available to an individual that is not legally allowed to purchase a weapon, and unable to create the trust relationships required to purchase one? Theft has always been a solution, but stealing weapons from armed (and thus potentially dangerous) individuals is extremely risky. However, since the birth of the internet, a new potential avenue of obtaining weapons has opened. The internet has been a technology that has become an avenue for discussion about topics that are normally considered taboo, such as weapons, under the cloak of pseudo-anonymity. However, actual sales and purchases of restricted illicit goods over the internet have been relatively easy to trace for authorities, until the emergence of anonymity networks such as Tor creating the so-called “dark web”. In conjunction with other technologies such as cryptocurrencies allowing for a degree of untraceability for financial transactions, the concept of a “cryptomarket” has arisen.

2.3.1 Overview

In February 2011, The Silk Road, the first modern cryptomarket was launched. Cryptomarkets are a portmanteau of cryptocurrency and market, due to their use of cryptocurrencies as units of trade, and their similarity to preexisting clear web markets such as eBay, but with a focus on illicit goods and services instead. Over a period of 8 months, between February and July 2012, Silk Road’s monthly total revenue was approximated at about 1.2 million USD [25], with approximately 70% of the traffic being drug related, an explosive amount of revenue over an extremely short time. Shortly before the end of Silk Road’s existence, it was estimated to be selling $89.7 million worth of drugs, a demand that has likely not decreased since across other markets [15]. In the 8 years since Silk Road, the growth of dark web markets has exploded, in parallel with their popularity, with seven general purpose cryptomarkets currently running [7]. AlphaBay, one of the larger markets post Silk Road’s shutdown, was approximated to have upwards of 600,000 USD per day in total revenue [26]. As a result, these markets have become the subject of research by academics from various fields.

Interestingly, the primary goal of cryptomarkets is to provide a number of trust based amenities, a valuable service in a low trust, anonymous environment. Firstly, they allow feedback comments for buyers and sellers. These comments are presumed to be written only after a successful purchase, but even with some degree of artificial inflation of responses, they provide a certain degree of trust that firstly, a buyer is not being scammed by a seller out of their cryptocurrency, and vice versa, that a buyer can actually trust that they will receive payment for the good or service that is on offer. Every transaction that is listed on the cryptomarket is taxed as part of the service provided, and is the core of the income model that fuels the operation of the cryptomarket. Most markets also provide cryptocurrency escrow services, where they hold the cryptocurrency for a transaction in a third account, until the buyer has verified that they have received their ordered good, at which point the cryptomarket releases the fund to the seller. The assumption is that cryptomarkets value the trust given by buyers and sellers using the platform more than the actual cryptocurrency amount in escrow, although the existence of “Exit Scams” have showed that this is not always true. Exit scams are where cryptomarket operators abscond with the entirety of the currency in the market’s escrow accounts, shutting down the market and profiting from their ill-gotten gains.

Although there are mitigation mechanics that prevent exit scams but still allow cryptomarkets to provide escrow services, such as cryptographically signed multi-signature escrow accounts, which require two of the three cryptographic signatures of the buyer, seller, and market in order to release the escrow funds. However, the number of supposedly cryptographically secure escrow funds are difficult to check, and practically relies on trusting the operator of the cryptomarket.

Cryptomarkets are not the only markets present on the dark web, however. It is entirely possible for individuals to build and maintain their own private web shops, or single vendor shops in the common parlance, due to the goods and/or services being provided being by a single group or individual. There are a number of advantages to single vendor shops, such as the (usually) lower
CHAPTER 2. BACKGROUND

cost of goods, due to not having the overhead costs of selling on cryptomarkets. Furthermore, from a buyer's point of view, it reduces one's exposure to exit scams. The difficulties however are myriad, the primary issue being a lack of trust. There is no centralized feedback system to verify that a seller is indeed legitimate, and not either a scam or a honeypot run by authorities.

Most cryptomarkets and single vendor markets specialize primarily in the sale of drugs, with the direct sale of arms estimated at around half a percent as per the major publication on the subject, a report by RAND Europe on the trade of weapons on the dark web [55]. This harks back to the previous example of trade on illegal markets, in section 2.2, where gangs tend to strongly prefer trading in drugs over weapons, presumably to the much higher demand, and thus profit margins, combined with the apparent lower risk in trading in drugs.

2.3.2 Where are weapons sold?

Weapons on the dark web are not sold on cryptomarkets to a significant degree. At time of writing, only a single major cryptomarket, Berlusconi market, openly allows for firearm listing and sales, though most others have categories for blueprints for 3D-printed weapons, non-lethal weaponry, and other weapon-related areas. Thus, it is likely that much of the dark web weapons market is very difficult to track by virtue of being done through single market vendors, with little available data to properly track and extrapolate sales to a reasonable degree of accuracy. Regardless, there is a base level of trust that exists, since many single vendor markets also are present on cryptomarkets, with feedback and sale information that prospective buyers may cross reference.

2.3.3 Logistic aspects

Another significant concern when purchasing weapons from the dark web is the delivery of said items. Currently, there are two options that exist for delivery of physical packages, postal services and dead drops [55]. The first works reasonably well for smaller stealthy packages such as drugs or fraudulently obtained identification, due to the scale issues inherent in scanning for essentially nondescript packages. However, weapons are much more identifiable, and thus require extra steps to avoid raising suspicion. These tend to involve the disassembly of weapons into easily camouflaged pieces which are distributed into multiple packages as innocuous deliveries, together with being packaged into innocuous but mechanically "noisy" objects, such as printers [55]. Then upon receiving the entirety of the shipment, the weapon can be reassembled by the recipient. Sending components is by far the most dangerous step for a seller, and considerable effort has gone into risk mitigation efforts [55]. Some of these include sending the various packages containing parts from a variety of geographical locations at different times in order to increase the difficulty of identifying a vendor in case a package is intercepted. Similarly, the shipment step is also very risky for buyers, where picking up a package that has been intercepted, or sent from a vendor that has already been identified and compromised is likely to result in incarceration. For these reasons, the general consensus among dark web weapon vendors is that guaranteeing safe delivery for both buyers and sellers is the largest issue when attempting to successfully transact in illegal firearms from the dark web.

However, an alternative method of delivery of illicit goods, called "dead drops", have arisen in drug markets, where a “dropman” hides shipments in discreet offline locations [5]. These shipments are then listed as available online, and upon purchase, the GPS location is sent to the buyer, who then picks up the shipment. Although this improves safety for both buyer and seller, it somewhat compromises the idea of geographical independence that cryptomarkets afford - if there are no dead drops provided in a buyer's general area, e.g. the buyer lives in a rural area, then actually delivering the firearm that is hidden somewhere becomes a much more difficult proposition. Secondly, there are simply a very limited number of dead drop locations that could fulfill the discretion requirement, and once used, it becomes questionable to use again in the future, which very rapidly puts a ceiling on the scalability of dead drops. Thirdly, competition between criminal organizations is a significant consideration, with the ability to identify and then steal...
CHAPTER 2. BACKGROUND

shipments from dead drop locations. A gang that is in conflict with another could steal packages, or report locations to the authorities and leverage the damage done to the competition in order to further expand their own territory. Similarly, the dropman must be an extraordinarily trusted individual, as they needs to be firstly loyal in order to not rat out locations to the authorities or other gangs, and also to not abscond with the valuable packages for their own personal gain. The risk for firearm sellers is greater as well, due to the higher price per unit of firearms versus drugs, and the larger physical profile of a weapon versus a package of drugs. The risk reward proposition for someone stealing a weapon package would be much higher than a single drug package, due to the general difference in monetary value between a single drug sale versus a weapon sale. Nevertheless, it is important to consider this as an option for firearms sales specifically, due to the risks and difficulties especially inherent to shipping a weapon through the postal system.

2.3.4 Who buys firearms on the dark web?

There are a variety of distinct categories of people that use the dark web to procure firearms, who are generally grouped into terrorists, criminals, and fixated or otherwise vulnerable individuals.

Terrorists

There is evidence that terrorists have attempted to obtain weapons over the dark web in the past, ranging from Islamist groups [62], to right-wing nationalist groups [45]. A Belgian man ordered an assault rifle, a semiautomatic handgun, together with a silencer for the latter and ammunition from the dark web in order to protect his family from asylum seekers from Syria [67]. There is also evidence that the weapons used in the 2015 Paris Attacks [23], the worst in France’s history since World War 2 [36] were bought through the dark web from a dealer in Germany [32].

Criminals

Individuals attempting to purchase weapons from the dark web just to commit crimes are also a separate class of interested party. An Irish police officer was arrested for trying to purchase a 9mm firearm [49] together with a silencer, and drugs [12]. A man from Houston attempted to purchase explosives and remote triggers on the dark web [30], and was also found with an AR-15 semi-automatic rifle. German police have also been particularly effective at finding and arresting both buyers and sellers on the dark web, with a vendor arrested in Berlin for selling weapons and ammunition [46]. There is also the case of a 33 year old Berlin man being arrested for trying to buy weapons on the dark web [51], spending upwards of $2000 USD in bitcoin to purchase a shotgun and ammunition, and another man arrested for trying to buy weapons and ammunition in person [11]. A number of cases that were directly linked to dark web sellers were also present, with an alleged vendor of the weapons used in the November 2015 Paris Attacks [44] having his online identity revealed as “DW Guns”, after having sold assault rifles to the terrorists, with another sixteen weapons found at his house. Finally, three men were arrested in Germany for attempting to purchase weapons from the dark web [33], two of which were registered as legal sports shooters. This last case is of special interest, and provides a basis for speculation that a potential place to find individuals that would buy weapons from the dark web is that of firearm enthusiasts.

Fixated individuals

There is also the category of individuals who purchase weapons for personal reasons that would be considered irrational. A Belgian police officer from Charleroi was arrested for buying weapons with intent to murder the current partners of two of his ex-girlfriends [8]. A British teenager from Newcastle [56] also was arrested and tried after attempting to purchase weapons on the dark web. Similarly, a teenager from Northern Ireland attempted to buy a sub-machine gun [10] and was charged under anti-terrorism legislation [9]. The large number of individuals from different backgrounds and motivations for purchasing illegal weapons from online illegal gun markets makes the concept of community identification and analysis an interesting one to explore.
CHAPTER 2. BACKGROUND

Conclusion
In conclusion, although three major categories of buyer have been identified, it is important to note that there is a great deal of overlap between them. For example, a large number of Islamic terror attacks were performed by lone wolves, with similar attack patterns as fixated individuals [14]. Furthermore, it appears that the individuals that purchased weapons from the dark web lacked the trust relationships to organized crime that make it possible to buy weapons from conventional markets, reinforcing the points made in section 2.2.4. This reinforces the hypothesis that dark web markets are a fresh avenue for the movement of weapons to individuals that would otherwise be unable to do so, such as the teenagers mentioned in section 2.3.4.

2.3.5 Size of online weapon markets
There are a number of criticisms leveled against research into weapon specific dark web markets. The first and foremost is that weapons make up less than half a percent of transacted items on cryptomarkets, heavily outclassed in revenue and number of transactions when compared to drugs, implying that the trade is insignificant enough to be ignored, together with the purported ease of buying weapons legally, as said by Nicolas Christin, a prominent researcher on dark web markets [37]. Conversely, an arms seizure from Spain, where over 3000 weapons allegedly destined to be sold on the dark web [13] were confiscated, implies a significant market belied by the number of transactions. As seen in section 2.3.4, there have been many reported cases just in Germany, with individuals actively selling and buying weapons on the dark web. A main limitation of the authoritative report on dark web arms trafficking, the report by the RAND Europe group [55] is that the analysis done is primarily on cryptomarkets, due to the relative ease of data gathering. However, the disparity between the empirical evidence and number of recorded transactions implies that the largely un-analyzed section of the dark web, namely single vendor markets, should have a reasonably degree of legitimacy, contrary to the opinion that the majority of those weapon related markets are scams.

2.4 Comparison of offline and online weapon markets
In short, after looking at the nature of both offline and online weapons markets, a number of very clear differences and similarities appear. They are both providing an illegal good, thus requiring some degree of trust. Where offline markets requires pre-existing relationships or memberships, online markets replace that through escrow accounts and feedback mechanisms, allowing individuals that would normally be unable to form those trust relationships to buy weapons online. Regardless of the wider potential customer base, the dangers of transacting in weapons online are still very real, and do not solve the core problems of selling weapons. As a result, online weapon markets are also thin especially in comparison to drug markets, mirroring offline illegal markets. Thus, there is significant empirical evidence that weapons have been and continue to be transacted online, and despite their thin-ness, those markets are in deserved need of research.
Chapter 3

Methodology

In order to answer the research questions as laid out in the introduction, two different research methods were required. Firstly, there is a need to identify and collect data on dark web markets in order to characterize the various aspects of the illegal underground weapon markets, and to serve as a comparison baseline to past research. Examining this snapshot will serve to answer research question 1. Secondly, identification and analysis of firearm enthusiast communities was needed in order to better the understanding of communities that may plausibly hold individuals that would purchase weapons from the dark web. Further, in order to appropriately evaluate the characteristics of marketplaces and of the underground arms trade, in addition to shifts in topic discussions following shocks, specific data analysis methods have been identified.

<table>
<thead>
<tr>
<th>Market Analysis</th>
<th>Topic Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlusconi market</td>
<td></td>
</tr>
<tr>
<td>Single vendor markets</td>
<td></td>
</tr>
<tr>
<td>Reddit</td>
<td>✓</td>
</tr>
<tr>
<td>4chan</td>
<td>✓</td>
</tr>
<tr>
<td>Telegram</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.1: Summary table of datasets.

3.1 Dark web market identification

Firstly, in order to improve the characterization of illegal dark web arms markets as per research question 1, a list of the various cryptomarkets that explicitly sold weapons was curated, by using https://www.deepdotweb.com/, a now defunct clear web resource that served as a useful entry point to the dark web, and dark.fail, a website run by a pro-privacy group that lists cryptomarkets and other dark web related news. This resulted in an initial list of 16 cryptomarkets, that was curated by removing non-English language markets, resulting in a curated list of 11 markets, as seen in table A.2. Of these 11 remaining markets, only a single cryptomarket remained with a dedicated category for the sale of weapons, known as “Berlusconi Market”. The other markets were checked via simple queries (e.g. searching for pistol, rifle, and common weapon brands such as Beretta) to check for stray weapon listings, but none were found.

This is a massive reduction in number from the 11 cryptomarkets identified in September 2016 as selling weapons [55], the status of those which can be seen in table A.3 as of September 2019. Of the markets identified in previous literature, only two are accessible as of time of data collection, and do not have any weapon-related listings. Furthermore, the four largest markets were all either seized by law enforcement or voluntarily shut down after being under attack by an extended Denial of Service attack [22]. Although it is far more likely that the sheer number of illegal goods traded...
on the seized cryptomarkets is the reason for law enforcement targeting and consequently seizing those markets, it is not improbable that the trade in arms may have been a contributing factor. Regardless, Berlusconi market is now one of the largest remaining active cryptomarkets, and the only one that is willing to entertain the trade in illegal arms. ¹

In parallel, a list of single vendor markets was generated from thedarkweblinks.com, a clear web site that details and links to dark web sites, in conjunction with a tool called MASSDEAL [21], previously developed by the Security Group at the TU/e. MASSDEAL is a web crawler that automatically explores and indexes dark web links in an effort to catalog various hidden services. From a dump of the MASSDEAL database, all sites that were related to weapon trafficking by the tool were extracted and put into a list, together with the sites from the weapons section in thedarkweblinks.com. This resulted in the list in table A.4, with their statuses. The accessible sites would then need to be scraped and parsed, in order to generate a dataset of single vendor arms markets, the specifics of which can then be contrasted and compared against the dataset gathered from Berlusconi market.

### 3.1.1 Berlusconi market

After the preliminary elimination of dark web markets, only the Berlusconi market cryptomarket remained as a potential market for weapons across all cryptomarkets. Berlusconi market does not have an API in order to efficiently collect listing and vendor information, and thus, a scraper in order to collect and store data was needed. Running the scraper at regular intervals would then be necessary to build up a picture of the market with a longitudinal aspect. After running the scraper over a 21 day time frame, a parser would be required to extract and organise listing and vendor information from the scraped web pages. This would generate a dataset that would serve as a snapshot of arms markets for comparison against previous research databases such as the RAND report [55], or for future research.

For Berlusconi market, the listing data as in table A.5 for each day would be recorded. Likewise, the vendor data as in table A.6 would be recorded. From there, various dimensions could be investigated as to how much prices vary, or price variance per unit per vendor, for example. Furthermore, this data can then be compared to single vendor markets in order to identify differences in features that may entice individuals to purchase on one platform versus another.

¹N.B.: As of July 2019, Berlusconi market no longer allows the sale of weapons.


3.1.2 Single vendor markets

Similarly to Berlusconi market, none of the nine single vendor markets have an API to directly access listing information. They are by and large different in format from a cryptomarket, with much fewer overall offerings, as seen in image 3.2. Thus, a bespoke scraper for single vendor markets also would need to be created that would access and download accessible web pages over Tor, over the same 21 day period as the Berlusconi market scraper. Likewise, a parser to extract data from the raw web pages would then be required. The listing data for every market will be saved as in table A.7. Downloading the Berlusconi and single vendor market data over the same period of time allows direct comparisons of the offerings of both markets to be made, and avoids confounding factors that may hinder the data analysis.

![Image of Guns and Ganja single vendor market](image3.2.png)

Figure 3.2: The index page of the Guns and Ganja single vendor market.

Once both the cryptomarket and single vendor market scraping and parsing is completed, then the pair of datasets containing all found listing and vendor data over the 21 day period can be generated. At that point, the market characteristics, and thus research question 1, can then be investigated, e.g. price markups over conventional legal markets or price variance over time.

3.2 Social community identification

Answering research questions 2.1 and 2.2, however, would appear to be less straightforward. The initial approach involved trying to find communities of individuals that would buy weapons on the dark web. With the assumption that these communities reside on the dark web, attempting to use MASSDEAL and clear web resources proved fruitless. Although communities on the dark web were found, such as “Dread” and various other forums, strictly enforced rules forbidding weapons-related threads were universal, preventing any further progress. Similarly, any initiated conversations in chat rooms or less formal boards were unsuccessful in generating any leads to deeper communities or discussion boards. It was also deemed that directly contacting sellers or buyers of illegal firearms would be a potential ethical concern, and was thus left out of scope.

It therefore became necessary to expand the search beyond solely dark web communities, and into the clear web. Under the assumption that clear web gun enthusiast communities could include individuals are likely to purchase weapons illegally [33], identifying and selecting the most likely communities became the next step.
CHAPTER 3. METHODOLOGY

3.2.1 Community selection

The criteria for selecting these communities was fairly simple. First and foremost, the communities that were selected had to be large in terms of userbase, and overall traffic. This requirement was primarily due to the likelihood of containing individuals that may be active in underground weapon markets. As of January 2019, Reddit was the 13th most visited website [1] in the world, with 4chan at position 439 [2]. 8chan follows suit at position 3,821 [3], and finally voat at 6,653 [4]. All of these communities allow for individuals to create messageboards that other users can then post on, albeit with varying degrees of moderation. Secondly, all of these communities have archives going back years, allowing for a wealth of harvestable data. The individual natures of these boards will be explored in the next few sections.

Reddit

![Figure 3.3: A post from the r/guns subreddit.](image)

Reddit is an extremely large set of discussion forums with massive appeal and an enormous user base. With the motto “The front page of the Internet”, the continued success of Reddit lies in large part due to the user driven nature of the site. Any user can create a “subreddit”, a submission board for user-submitted and self-moderated (to some degree) content, in any combination of text, video, and image forms, as can be seen in image 3.3. Subreddits can be about anything, with site administrators stepping in only if a subreddit violates any of the site-wide rules or terms of service, such as no child pornography, or revealing private information about another individual, also known as “doxxing”. However, beyond those restrictions, moderators are given a free hand in terms of how subreddits are run. Consequently, the gun-related subreddit serves as a useful baseline representation of gun communities on the internet as a whole.

4chan

As a counterpoint to Reddit, 4chan is a set of discussion boards that was originally created in 2003 by Christopher Poole, and has become one of the core nexuses for the creation and distribution of internet culture. It is best known for being a spawning ground for image macros and other memes, in addition to being the home of the “Anonymous” hacker group, and most recently, being attributed with a degree of credit for the electoral success of President Trump [35]. Consequently, the relatively mild degree of moderation present on 4chan and ability to post without creating an account has resulted in over 22 million monthly visitors, widely spread over its largest boards.

4chan’s weapons board, known as ‘/k/’, consists of text and/or image posts replying to each other in threads, an example of which is in image 3.4 and have a certain number of properties that
differentiate 4chan from discussion boards such as Reddit. As mentioned above, most posters are anonymous, and do not require an account to be made. This anonymity is just with respect to other users and not to either the site operators or authorities unless using anonymity tools such as Tor. Secondly, 4chan is known for its ephemerality, where each board has a finite number of threads, which are deleted after a period of inactivity on that thread. If there are no new posts, the thread falls out of the catalog and is no longer accessible. Accordingly, the former has led to posting patterns on 4chan being far more offensive or shocking for the sake of a response as opposed to other mainstream sites. Furthermore, users will very willingly discuss topics that are normally illegal or taboo, due to the lack of perceived consequence from either legal authorities or board moderators [35].

Although this distinguishes 4chan from Reddit as a home for more outlandish opinions and individuals, it would seem that the transient nature of the discussion board would make it difficult to collect the type of longitudinal data required to answer research question 2.2. Thankfully, due to 4chan’s fairly unique place as the first and largest of the English language imageboards, there are a number of archives that are publicly available that have recorded years worth of threads from various boards, thus serving to mitigate the temporary nature of the board.

8chan

8chan follows in the footsteps of 4chan, and was created in 2013 by Fredrick Brennan due to perceived escalating censorship and restriction of free speech [53] across the internet, and is a haven for individuals that find the already lax moderation of 4chan and other discussion forums too restrictive. It first gained the majority of its userbase after the “GamerGate” controversy, where a group of anti-feminists were ejected from 4chan, who promptly migrated to 8chan. Thus, 8chan has garnered a reputation as a censorship-free haven for controversial opinions, harboring the “alt-right” movement, and contains boards where multiple mass-shooters have posted manifestos shortly prior to the relevant attack. It has a format generally following that of 4chan, as seen in image 3.5, with the added functionality of multiple images per post, as opposed to just one. Indeed, 8chan is no longer listed by Google searches due to harboring child pornography and alleged pedophile networks [54]. In short, 8chan is one of the most extreme communities accessible on the clear web both in opinions voiced, and content. Regardless of the moral stance of the reader, 8chan is indisputably an area of potential interest for research, considering the comparisons that can be made to other imageboards with less unsavory reputations. Unfortunately, however, due to the more chaotic nature, smaller size and shorter history of 8chan, together with the difficulty of hosting an archive that may possibly contain links to child pornography and other illegal material,
3.2.2 Private communities

Unlike the previous datasets, which are publicly available clear web communities, the addition of communities that are nominally private and invite-only offer a unique opportunity for comparative analysis, in this case a Telegram group for European gun enthusiasts on 8chan. Telegram is already being used as a communication platform for drug trading across the dark web [29], due to its reputation as a secure, easy to use mobile messaging application that is not run by an American or European corporation. Telegram is much like “WhatsApp”, but with a major distinction - it does not require a phone number in order to add contacts or to join a group. This group was successfully infiltrated through joining a public link, and then having a short conversation with one of the moderators of the group in order to prove affiliation within the community. The interview was successful, resulting in the opportunity to observe and record communication from its creation until the 20th of August 2019.

3.3 Analysis methods

Once the various social community datasets have been acquired, the question as to how to analyze the datasets to answer research questions 2.1 and 2.2 remains. To do so, it would be necessary to use some form of topic model in order to discover the topics within the extremely large natural language datasets that are a focus of this thesis.

Topic models are statistical models that process large collections of documents (corpora, where each document is a corpus) and attempt to extract semantic data from them [19].

They operate under the basic intuition that if a document is about a particular topic, particular words would show up in that document with a higher likelihood. Documents may have different topics, and thus, a topic model captures the statistical relationships between words in the collection of documents and then attempts to identify what the major topics might be, and what the balance of topics in each document is. The end result allows for a human-readable and comprehensible representation of the found word topics present, which can then be manually labeled in order to have an approximation of the topics discussed by the corpora used to train the model. These topics can then be used together with additional corpora to label them in terms of statistical correlation to the generated topics. This functionality allows for answering research question 2.2, by generating a list of mass shootings (referred to as shocks from this point forward). After this
list has been generated, additional corpora can be drawn from the respective 4chan and reddit datasets within a certain date range of each of the shocks, and then be labeled by the now trained topic model. A summary of the various steps required to turn a collection of comments or posts into the list of topics and probabilities in order to answer research questions 2.1 and 2.2 are in figure 3.6.

Figure 3.6: An example diagram of the various steps necessary in order to build a specific LDA.

3.3.1 Preprocessing & Latent Dirichlet Allocation

First and foremost, whatever textual data is collected needs to be processed and sanitized before being fed into any topic model. This entails a number of steps, known as lemmatizing and stemming, which reduce conjugations and tenses of words to their root forms, at which point they can be tokenized [47].

Stemming reduces words from their conjugated forms to their root forms through stripping suffixes. It is important to note that these root forms are not necessarily actual words, but are just inflection-free versions of the words found. In this case, an algorithmic stemmer is used, known as a Snowball Stemmer, based on the Porter stemmer which does not keep a vocabulary of root words, but instead uses algorithmic rules [58](decided per language) to strip suffixes. An algorithmic stemmer functions better on corpora with slang based vocabularies that are intrinsic to the community, which is especially true of the 4chan dataset. Furthermore, the Porter stemming algorithm is considered to produce the best output, with a low error rate compared to other often used stemmers [39].

Lemmatization ensures that the root form (as created by the stemmer above) actually is a part of the language that it purports to be from. It returns the canonical (thus lemmatized) form of the root word that the stemmer requests.

Finally, a tokenizer is used to reorganise the now lemmatized and stemmed input document into a list of tokens, while also removing any tokens that are stop words. Stop words are generally the most common words in a language, and are often just filler that do not give semantic context to the document. At this point, the processed text

There are a number of different word-topic models that have been developed but for the purposes of this thesis, a Latent Dirichlet Allocation (LDA for short) model will be used. LDA models view documents as a mixture of various topics, where each topic is generated and assigned automatically to documents. Each topic is composed out of a number of different words, but for an LDA versus other topic models, the topics are assumed to be distributed according to a Dirichlet distribution. The details of the distribution are not important, insofar as it attempts to capture that each document only covers a small set of topics, and those topics use only a small set of words frequently, and ensure that the model accounts for that.

The other main section of an LDA pipeline is the vectorization step, which maps the text into quantitative values that the model can then process and use to generate the topics. There are two simple and oft used methods for vectorizing text corpora, the bag of words (BOW) model, and Term Frequency-Inverse Document Frequency (TF-IDF). The BOW model simply just counts the occurrences of various words, and records them. It then builds a matrix of documents and number of tokens, which is then used as the source corpus for the LDA model. TF-IDF, however, attempts to evaluate how important a word is to a document. The assumption is that if a word is more important, it is likely to occur more often, relative to the number of terms in the document.
in order to normalize by document length. The Inverse Document Frequency criteria attempts
to score how rare the word is across documents, in an effort to properly assess how often a topic
appears within the corpus as a whole [59]. This leads to the following criteria:

\[
TF(t) = \frac{\text{number of times term } t \text{ occurs in a document}}{\text{total number of terms in the document}}
\]

\[
IDF(t) = \log \left( \frac{\text{total number of documents}}{\text{number of documents with term } t \text{ within}} \right)
\]

These two values multiplied with each other generates the TF-IDF score, and is generally
accepted to be more accurate in terms of denoting word significance and representation than a
naive BOW model.

In short, an LDA model ingests a collection of documents, which have been turned into vectors,
and proceeds to automatically generate the topics in that collection. Those topics are then assigned
to every document within the corpus, all in a much more efficient and consistent method than a
human can, especially over large collections.

3.4 Topic analysis

Thus, topic modeling and LDA models in particular come in useful when identifying topics present
within a collection of documents. In order to answer question 2.1, this entails building separate
LDA models from each of the Reddit, 4chan, and Telegram corpora. For each community six
topics are generated, and labeled independently, in order to avoid forcing the model to generate
“average” topics over potentially very different communities. Once generated, the outputs can
be compared against each other through calculating the Jaccard index between the models, and
through a visualization tool specifically built for LDA models called LDAvis. This should allow
for some comparison of the topics present within each community.

The generated topics for an LDA model can be compared against those generated by another
model through a host of different methods, which have been reviewed through for accuracy against
a human baseline [16]. In this case however, the Jaccard distance is used

In this case, the generated word-topics from each model are considered to be sets, which allows
for the use of the Jaccard index. The Jaccard index is a measure of similarity between two sets of
data, with a percentage range. The higher the percentage, the more similar the populations. The
formula for calculating such is as follows:

\[
J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}
\]

This generates a similarity matrix, where each element within corresponds to the shared members
between each word-topic. The Jaccard distance is the complement of the Jaccard coefficient.

pyLDAvis is a tool [63] initially built for the R programming language that has been adapted
to Python with the specific intent of visualizing LDA models. It puts forward an additional
criteria called relevance, which is a method that ranks terms within topics, a parameter \( \lambda \) that
allows for user input to modify the weighting of lift, defined as a term’s topic specific probability
versus its marginal probability across the corpus. Furthermore, it allows for an intuitive method
of identifying the overlap or separation between topics within a model, while taking the above
considerations into account.

3.4.1 Generating list of shocks

Once the reddit and 4chan models have been generated, documents can be fed into the models,
returning a list of probabilities that the document belongs to a certain topic. These probabilities
can then be aggregated across a collection of documents, thus generating a topic probability
distribution for that specific corpus. This functionality can be used to answer research question
2.2, through generating a list of attacks, drawing documents from the time periods surrounding
CHAPTER 3. METHODOLOGY

<table>
<thead>
<tr>
<th>North America</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Canada</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>Netherlands</td>
</tr>
<tr>
<td>Norway</td>
<td>Denmark</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>Finland</td>
</tr>
<tr>
<td>Slovenia</td>
<td>Germany</td>
</tr>
<tr>
<td>Belgium</td>
<td>Austria</td>
</tr>
<tr>
<td>Poland</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Romania</td>
<td>Croatia</td>
</tr>
<tr>
<td>Serbia</td>
<td>Portugal</td>
</tr>
<tr>
<td>Russia</td>
<td>Hungary</td>
</tr>
<tr>
<td>Greece</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>New Zealand</td>
</tr>
</tbody>
</table>

Table 3.2: List of Countries where shocks are drawn from.

those attacks, and then plotting the topic distributions in order to note any patterns or common themes.

First and foremost, this required a list of attacks to be generated over the entire period for each community. In order to generate a meaningful list of mass shootings, the first priority was to generate a list of countries that the communities identified above would have a likely response to. The intent is that including every mass shooting that occurs worldwide becomes pointless very quickly when trying to assess the response of a community to a specific event. An individual that is from a western country and is internet-capable is less likely to respond to a shock that occurs within an Arabic country as opposed to one that occurs within western Europe, making shocks with the former criteria effectively noise. Thus, the following constraints for shocks are defined:

Firstly, that the country be English speaking, at least to a reasonable degree. Secondly, that the country has a reasonable degree of web traffic to the communities as described in section 3.2 a list of English speaking countries that contribute to the cultural zeitgeist on these sites. From Alexa ratings, this meant eliminating Africa, South America, and Asia [2], [3], [1], [4]. Then in order to ensure that the first criteria was fulfilled, English speaking countries were selected from https://www.ef.edu/epi/, where if a country was both from a pre-selected region and had an English ranking greater than moderate, it would be added to the list. This resulted in the list as found in table 3.2.

From this list of countries, all attacks that occurred within the dataset would be pulled, their dates recorded, and used as shocks in order to properly answer research question 2.2. Unfortunately, the definition of a mass shooting is not standardized [42], but a variety of news sources have attempted to collate data about the frequency and statistics regarding mass shootings [31]. The criteria for a mass shooting is therefore defined as the following:

- The perpetrator(s) killed at least 3 people.
- The perpetrator(s) used a weapon as outlined in section 2.1.
- The shootings occurred in a public place.
- Perpetrators that died or were wounded during the attack were not included in the list of people killed.

Although countries such as the USA have had lists made of mass shootings [31], both due to the relatively high number of attacks, and thus media focus on the issue, not every country in table 3.2 has had such a report. In most cases, this is due to the rarity of such attacks, although
CHAPTER 3. METHODOLOGY

### Dark Web Data

<table>
<thead>
<tr>
<th>Market Data</th>
<th>Historical dataset comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.g. price shifts over time</td>
<td></td>
</tr>
<tr>
<td>Berlusconi Market Dataset</td>
<td>Y</td>
</tr>
<tr>
<td>Single Vendor Market Dataset</td>
<td>Y</td>
</tr>
</tbody>
</table>

### Community Data

<table>
<thead>
<tr>
<th>Topic Modeling</th>
<th>Topic Modeling around terrorist attacks (LDA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (LDA)</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Reddit /r/guns</td>
<td>Y</td>
</tr>
<tr>
<td>4chan /k/</td>
<td>Y</td>
</tr>
<tr>
<td>8chan Telegram group</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 3.3: A summary of the social community datasets and intended analysis tools.

they tend to be more severe. Consequently, it is likely that attacks are unlikely to be missed by mainstream news sources, mitigating the possibility of missing attacks.

Therefore, the list of events was generated through using the category pages for Mass Shootings in Wikipedia by country and decade, and cross referencing and substantiating them by verifying primary sources, in addition to the criteria established above.

Once done, the archives used to generate the models initially would have to be drawn from the relevant datasets. Four samples per shock would need to be taken (with figure 3.7 as a simplified example of shocks), resulting in samples being extracted from the fortnight before and after an attack, and the week before and after an attack. The topic distributions for each of these shocks, together with a baseline topic distribution i.e. a distribution of topics over the entire training set can then be graphed and compared, with qualitative assessments made between the various samples made per shock. The samples from after a shock allow for qualitative analysis of topics both before and after a shock occurs, to see and visualize how the community reacts to an attack. Together with qualitative comparisons to the severity of a shock in terms of number of victims, or country of occurrence, there is significant opportunity for exploration.

![Figure 3.7: A representation of the three different timelines, Reddit, 4chan, and Telegram, with arrows representing the samples drawn from around shocks.](image)

At this point, the entirety of the methodology was developed sufficiently in order to begin the data gathering phase.

### 3.5 Social communities

From the initial list of 4 social communities identified in section 3.2, these were cut down to 3 viable options. Both Reddit and 4chan have dedicated archival sites, which offer data dumps that can be parsed for relevant posts, comments, and submissions.

For collecting historical data from Reddit, https://pushshift.io/ is an archival project run by Jason Baumgartner, and contains nearly every submission and comment from 2011 to the current day in a series of compressed files.
For 4chan, there are a number of different archives, each of which tends to specialize in a certain subset of boards. Although most of these boards are directly query-able through APIs, the APIs are rate limited and restricted to just 25 posts per query, which makes using the API inadequate for the wholesale data dump required for this thesis. Fortunately, some of these archives have had the foresight to upload snapshots of their databases to internet archival projects such as archive.org. One of these archives is desuarchive.org, which archives /k/, the weapons-related board that is of interest. The archive dump can be found at https://archive.org/download/desu-2017-db, and contains all posts on /k/ from the 1st of July 2012 until the 1st of December 2017. This limits the list of generated shocks to this time period, and also means that the Reddit dataset will be limited to the same time period as well.

Unfortunately, 8chan does not have a dedicated archival project, and the internal archival mechanism is extremely incomplete, even for solely textual data. Therefore, posts from 8chan has been removed from any further results and analysis, despite the obvious interest in being able to quantify and analyze posts on the so-called “wild west” of the internet. Similarly, despite multiple requests for an API key from the administrators of Voat.co, a lack of response has resulted in the dataset being removed, since there is no external archival service, and writing a scraper for Voat quickly resulted in rate limitations that led to data acquisition being painfully slow.

Therefore, three datasets remain, of which two allow for shock extraction and analysis, namely Reddit and 4chan. The Telegram group stands separately, but provides a text corpus that is interesting to compare and contrast against the two less extreme communities. The technical details for data extraction are reported in appendix A.2.

3.5.1 List of shocks

The list of shocks were generated as per section 3.4.1. The time period was defined as 2012-7-1 until 2017-12-31, resulting in a list of 52 mass shootings as seen in the appendix, table A.1. For each of these shocks, the surrounding documents were extracted as outlined in section 3.4.1.
Chapter 4

Results & analysis

This chapter contains the results and analysis from both dark web markets and social community topic models, with the intent of firstly, characterizing the nature of illegal arms markets, followed by the communities that are likely to contain individuals that may be active on these markets.

4.1 Dark web markets

The section covers the results from Berlusconi market, a cryptomarket which allows users to transact in weapons of various makes and models, in addition to the results from single vendor markets. Single vendor markets are markets that are effectively self-hosted webshops, that sell directly to buyers without the trust related services a cryptomarket provides.

4.1.1 Berlusconi market

The un-processed raw data obtained from scraping and parsing Berlusconi market as in section A.2.1 needed to be processed and sanitized before de-duplication and data analysis could be performed.

This had to be done manually and resulted in 1,638 posts being removed. Furthermore, certain fields had values which were flags and thus indicated values not being present, such as “ERROR”. Some fields also had values that were effectively meaningless, such as the value of “In” for stock, where integers were also used, or “Unspecified” for the ships from field. In these cases, errors are displayed as “NaN”, and the value of In was changed to the integer 1, under the assumption that at least 1 item was available, but with no guarantee of more. Redundant information was also removed, such as the “currency” and “ships to” fields, where both only had a single value across the entire dataset.

The description of the processed dataset is in table 4.1. From the table, a number of points of interest immediately present themselves. Firstly, the differing number of unique listing numbers and titles, implying that there are multiple listings with the same name. This means that de-duplication of the dataset requires removing listings which are semantic duplicates, instead of just assuming that listing number or title is a unique key. Furthermore, the distribution of prices, stock, and number sold are likely to be extremely unbalanced, with a large number of outliers.

However, before de-duplication of listings, checking if there is a significant amount of activity on the market to verify initial claims about the “thin-ness” of a market, such as if new listings were added over the observed period.

As seen in figure 4.1, there were new listings added over the observed time period, with the number of listings active per day varying between 1,600 and 1,826, with a minimum of 0 new posts added in a day and a maximum of 78. Unfortunately there were no discernible trends on which days were popular posting days, with the majority of posts remaining present throughout the entire observed period. Similarly, in figure 4.2, the median price of weapons overall does not
CHAPTER 4. RESULTS & ANALYSIS

<table>
<thead>
<tr>
<th>price_fiat</th>
<th>stock</th>
<th>number_sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>36397</td>
<td>30571</td>
</tr>
<tr>
<td>mean</td>
<td>943</td>
<td>83768</td>
</tr>
<tr>
<td>std</td>
<td>1374</td>
<td>2318554</td>
</tr>
<tr>
<td>min</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>25%</td>
<td>500</td>
<td>50</td>
</tr>
<tr>
<td>50%</td>
<td>600</td>
<td>658</td>
</tr>
<tr>
<td>75%</td>
<td>900</td>
<td>900</td>
</tr>
<tr>
<td>max</td>
<td>26994</td>
<td>100000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>listing_number</th>
<th>title</th>
<th>vendor</th>
<th>ships_from</th>
<th>details</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>36,397</td>
<td>36,397</td>
<td>36,397</td>
<td>22,516</td>
<td>36,397</td>
</tr>
<tr>
<td>unique</td>
<td>2,257</td>
<td>1,212</td>
<td>38</td>
<td>10</td>
<td>776</td>
</tr>
<tr>
<td>freq</td>
<td>21</td>
<td>2,012</td>
<td>11,071</td>
<td>11,215</td>
<td>4,483</td>
</tr>
<tr>
<td>first</td>
<td>2019-05-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>last</td>
<td>2019-06-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Descriptive statistics for the processed Berlusconi market dataset.

Figure 4.1: Variation in number of listings per day on Berlusconi market.

change significantly over time, varying a total of 52.62 Euros between the maximum and minimum values, a variation of about 5.6% from the mean, and well within the standard deviation of 16.03 Euros.

At this point, the dataset is semantically de-duplicated. Instead of simply removing listings with identical numbers, listings with redundant semantic information were removed as well, defined as listings with identical values for title, price fiat, and vendor fields. In conjunction with this de-duplication, the vendor data from table A.9 could also be joined. Due to the large amount of redundant data in the vendor listing dataset, the “last seen”, “vendor seen”, “ships from”, “profile text”, “terms”, and “pgp exists” fields could be removed. The remaining data is the vendor name, which is used as a merging key to join the two datasets, the various feedback fields, and the “orders finalized” field. Furthermore, each listing was then categorized into various classes of weapon, using section 2.1 as a baseline, resulting in the following categories: Pistols, rifles, shotguns, machine gun, machine pistol, grenade launcher, explosives, ammunition, custom, and others. Custom listings are listings that are either bulk or custom orders, and “other” listings are listings that do not fall easily into any pre-existing category.

In parallel, listings were labeled by brand where possible, resulting in the following brands being identified:
CHAPTER 4. RESULTS & ANALYSIS

Figure 4.2: Variation in mean price per day on Berlusconi market.

- generic
- beretta
- kalashnikov
- arsenal
- pistol
- armscor
- taurus
- ATI
- armalite
- cimarron
- lmt
- bushmaster
- glock
- colt
- barrett
- explosives
- ruger
- smith & wesson
- kimber
- cimarron
- smith & wesson
- sig sauer
- walter
- other
- cz
- H & K
- Remington
- century arms
- ammunition
- custom
- makarov
- mossberg
- fn herstal
- rifle

Shipping information

As seen in figure 4.3, the vast majority of listings which have a shipped from value are from the United States, although both Germany, the UK, and Europe as a whole do have significant presences. Considering the primary method of weapons entering illegal markets tends to be through large secondary markets, the near parity between European countries, with relatively strict gun control laws, and the United States and Canada is surprising.

Figure 4.3: Number of listings shipped per country.
CHAPTER 4. RESULTS & ANALYSIS

Stock analysis
The stock field is self reported, and after quick inspection, statistically irrelevant, with a standard deviation more than an order of magnitude greater than the mean, even with values of 1 removed from the dataset. This likely indicates that stock has meaning on Berlusconi market as a metric solely to attract buyers, with no incentive to not lie.

Sales analysis
For sales, however, the range of values is much more realistic, with 21 as a maximum number of sales for a listing. However, 91% of listings have 0 sales listed, skewing the distribution heavily, as seen in figure 4.4. Once the 0 sales have been removed, only 18 of the 38 vendors have a sale listed, with the distribution seen in figure 4.5. Although it is unclear how the sales metric is reported, due to Berlusconi market both allowing users to perform sales through the market’s escrow services, or separately without transferring through cryptocurrency wallets owned by the market. It may be that the former case allows for vendors to confirm sales figures, implying that the vendors with the most sales use the escrow service the most.

Price Analysis
When looking at vendors, starting with the comparison of the median prices of the various listings show that certain vendors are significant outliers, pricing far above the rest of the market, as in figure 4.6. These vendors tend to have very few listings in comparison, with specializations in expensive niche goods, such as rifles from high-end brands, such as Century Arms. Conversely, vendors with the lowest median listing prices tend not to sell weapons at all, but instead sell ammunition.

When it comes to overall number of listings, pistols have by far the most listings, nearly four times the number than the number of rifle listings, as seen in figure 4.7. Furthermore, looking at
the brands on offer, far and away the most common brand of weapon being sold are Glocks as seen in figure 4.8, followed by generic brands and rifles. This reinforces the idea the online trade in weapons are predominantly pistols, especially ones with a reputation for being cheap and robust, much like the most common brand of rifle being sold, Kalashnikovs, the family that the extremely popular AK-47 and its variants belong to.

When looking at figure 4.10, massive outliers appear, with machine pistols and machine guns being far above the other categories in terms of median price. This reinforces the point again that these are generally not easily accessible items due to being extremely rare and expensive. It is also interesting to note that the markups of pistols (which are primarily Glocks) are reasonably low, with legal prices for fresh pistols varying from 450 Euros to 680 Euros, whereas the mean price for a glock on Berlusconi market is 674 Euros, with a median price of 600 Euros. The lack of markup could imply a number of different things: Firstly, that vendors have extremely low costs when obtaining these weapons. This could be because the weapons are second-hand or party to criminal activity, and thus need to be offloaded with minimal traceability. Alternatively, the vendors could also be scofflaw certified gun dealers that obtain the weapons for blue book prices.

When looking at total market value, figure 4.9 shows the breakdown of total value per vendor, with approximately three of the vendors having over 100,000 Euros of goods available on the market.

Figure 4.11 multiplies the mean prices with the number of listings per category, where it becomes apparent that despite their price, there is a low number of offers for any niche weapons. Instead, the expected has occurred, with pistols, rifles, and surprisingly, custom orders being the categories that constitute the majority of the value in the market. This gives a total market value of 1,534,433 Euros, without considering stock or number of items sold.

A naive look at the median price per brand as in figure 4.13 would strongly imply that certain brands are far, far above others in profitability, such as the very rare but expensive Barrett sniper rifles, but once the total number of items of that brand are factored in, the real image becomes far more interesting, as seen in figure 4.12. It confirms the idea that cheap, effective weapons are generally more likely to be purchased on the dark web than effectively a collector’s item.

4.1.2 Single vendor market

For single vendor markets, a similar, though not as involved pre-processing and analysis step was required, converting USD values to Euros, and Bitcoin prices to Euros as well in order to have a standardized currency across markets. The details field was also removed, due to many of the listings not having details fields.

After doing so and starting to analyze the longitudinal data recorded, the price variation over the observed period is 3.29 Euros, which is a 0.4% change. On manually inspecting the original
scraped data, it appears that the only values that change are those for bitcoin, and even that varies very, very little. This may be due to a phenomenon known as “menu pricing”, which comes from the concept that it costs the parent company money to change menus. Although the literal cost of “printing” the listings in this case is basically zero, the possible consequences of buyers becoming displeased or refusing custom after large price variations in either direction for the goods that they are purchasing is potentially massive. Thus, it is possible that vendors on single vendor markets may accept being fairly inflexible to the already meager supply and demand effects on the market, and simply pick a price that they are comfortable selling their goods for, without attempting to optimize further in order to capture a marginally larger section of the market. Furthermore, the number and type of listings does not change. There are zero new listings added, or removed, implying that either the vendors have sufficient stock (or promise delivery with enough of a lead time to obtain their stock) to deliver any purchased goods. Alternatively, there is the possibility that the demand for goods may just be very low, as is the nature of thin markets. Finally, there is a suggestion that the single vendor markets are scams or honeypots, and thus have no real stock to sell or change.

Separately, the mean pricing for the market overall is much lower than that of Berlusconi market, with the median price for goods as a whole being 178.25 Euros less for single vendor markets than Berlusconi market, giving credence to the hypothesis that single vendor markets have a lower operating cost due to not having to pay a middleman.

Alternatively again, it is possible that the single vendor markets are scams or honeypots, and are thus price their non-existent goods aggressively to capture more of the market and thus identify and then convict individuals buying weapons from underground markets.

Furthermore, similarly to Berlusconi market, a similar distribution of popular categories is

Figure 4.7: Categories of weapons sold.

Figure 4.8: Brands of weapons sold.
visible in figure 4.14, where the most popular category is pistols, far and away, with rifles coming in behind. There are proportionally more rifle listings than on Berlusconi market, however.

When looking the mean price per item, however, the distribution of prices on the single vendor markets is much more even, with explosives being surprising cheap, with the other and rifle categories at the top. The other category consists of specialist items such as night vision scopes, and rifles are generally more expensive than other types of firearms, but the relative closeness between the mean price of rifles and pistols is interesting. It may mean that rifles are overvalued on Berlusconi market, or undervalued on single vendor markets. However, it is difficult to assess the value of a rifle, considering the potential difference in asking price for a newly manufactured AK-47 versus one that has been kept in a weapons cache for decades.

When looking at total market value per vendor, figure 4.16 shows the much closer relative difference between vendors than on Berlusconi market, making it difficult to say which of the two is more representative of the market as a whole. The total market value of the single vendor markets is 199,606 Euros.

4.1.3 Discussion

It appears that the overall landscape of underground arms markets on the dark web has changed. Berlusconi market had nearly double the number of listings compared to the last study that attempted to assess the size of underground online markets, with additional signs of activity such as new listings being created, and a very large number of custom bulk orders that appear to be reasonably high in price in comparison to many of the other goods on offer. However, the same
patterns of pistols and rifles being very popular repeats itself through both markets, albeit on a much larger scale for Berlusconi market. Furthermore, Berlusconi market appears to be truly valuable for a much smaller number of vendors, with a much higher total market value, with the total value of listings on Berlusconi market being eight times as high as the single vendor markets. With that said, however, single vendor markets have extremely minimal shifts in pricing, at a significantly lower price with reasonably varied goods on offer in terms of categories. It may be that they are more consistent in terms of availability of items, but with the closure of the weapons dealing section of Berlusconi market, the total value of offers for underground online markets as a whole has dropped heavily.
CHAPTER 4. RESULTS & ANALYSIS

Figure 4.13: Median listing price per brand.

Figure 4.14: Listings per Category.

Figure 4.15: Listings per Category.
Figure 4.16: Value per vendor.
CHAPTER 4. RESULTS & ANALYSIS

4.2 Analysis of community discussion topics

This section contains base analysis of the three models generated from their respective text corpora. This analysis is primarily qualitative in nature, due to the interpretability of topic models. It is done through a number of methods, from manually labeling the discovered topics, the use of a visualization tool called LDAvis which helps to understand and interpret the success criteria that a given model outputs.

4.2.1 LDAvis tool

The reason a tool like LDAvis is used is due to the difficulty of pulling information from topic models in an intuitive manner. When fitted and trained, an LDA model returns a probability mass function, over all possible words within the model for each topic. Represented naively, this results in a massive table of terms and topics, which is difficult to extract meaning and shape from.

![Figure 4.17: The output from the LDAvis visualization tool when prepared with the reddit model, corpus and dictionary.](image)

For a more concrete example, see figure 4.17, which is the default state of the visualization of the reddit LDA model. The visualization is split into two sections, a bubble plot on the left and a list of terms on the right, sorted by overall frequency of that term within the corpus. The bubble plot consists of circles, each representing a topic such that its area is proportional to the proportion of the terms representing that topic across the total number of tokens in the corpus. The distance between the circles is defined by the inter-topic distances as computed by the Jensen-Shannon Divergence.

However, it may be difficult to distinguish different topics if certain terms are prevalent throughout most of the models. The relevance slider in the top right allows for the user to adjust the weighting of $\lambda$, defined as the relevance of a term. Simply put, changing the lambda value changes the value of terms, by putting more weight on the ratio of the frequency of the term given the topic versus the overall frequency of the term, allowing for clearer identification of topics than simply reading a list of weighted topics with their associated terms.

For further information about the visualization used, the initial presentation and paper may provide additional insight [63].
4.2.2 Reddit

The LDA model generated from the sampled Reddit corpus contains the following 6 topics. Each topic is defined by ten keywords and their weights, but must be manually labeled by a human, due to the difficulty in teaching a model semantic linguistic meaning. Do note that the ordering below is arbitrary, since LDA defines no natural ordering of generated topics. These are deterministic per training run, however, and are consistent for the remainder of this document, especially section 4.3, where topics are compared against each other.

1. Going to the range - (0.035, shoot) + (0.025, rifl) + (0.015, gun) + (0.015, rang) + (0.013, sight) + (0.010, pistol) + (0.009, target) + (0.008, point) + (0.007, use) + (0.007, like)

2. Gun ownership laws - (0.024, gun) + (0.013, peopl) + (0.010, state) + (0.009, firearm) + (0.009, law) + (0.008, like) + (0.008, think) + (0.007, right) + (0.007, know) + (0.006, say)

3. Reddit-specific discussion - (0.074, post) + (0.023, comment) + (0.022, link) + (0.019, descript) + (0.017, rule) + (0.016, add) + (0.016, yes) + (0.015, pictur) + (0.015, remov) + (0.015, sub)

4. Gun usage - (0.016, round) + (0.012, gun) + (0.011, like) + (0.011, barrel) + (0.011, mag) + (0.011, shoot) + (0.010, trigger) + (0.007, magazin) + (0.007, use) + (0.006, load)

5. Conversation - (0.032, like) + (0.023, look) + (0.014, fuck) + (0.013, get) + (0.012, shit) + (0.010, think) + (0.009, know) + (0.009, nice) + (0.009, good) + (0.008, thing)

6. Purchasing weapons - (0.023, gun) + (0.022, buy) + (0.015, thank) + (0.012, good) + (0.012, get) + (0.011, price) + (0.011, want) + (0.010, like) + (0.010, time) + (0.010, look)

The topics above can be expected from a community that is focused primarily around the sale and usage of weapons, with sufficiently differentiated terms that topic labeling is straightforward. The only topic that was potentially difficult to label, due to being non-specific is topic 5.

![Image: Reddit model with Topic 1 selected.](Figure 4.18: Reddit model with Topic 1 selected.)

When we look at the visualization in figure 4.18, first and foremost we see a reasonable distribution of topics throughout the space. This implies that there is little relative overlap, and a large number of distinct topics having been found, as intended. Due to LDAvis ordering topics by the greatest relevant term frequency, the ordering of topics has changed from the base unordered set. The ordering is as follows, with the % of relevant terms alongside the topic.
CHAPTER 4. RESULTS & ANALYSIS

1. Gun ownership laws - 21.9%
2. Gun usage - 20.1%
3. Purchasing weapons - 18.7%
4. Conversation - 16.2%
5. Going to the range - 14.1%
6. Reddit-specific discussion - 9.2%

The main points of interest lie in the not unexpected overlap between topics 1, 2, and 5 from the new ordering (which will be used for the rest of this section), namely gun ownership, usage, and going to the range.

Figure 4.19: The output from the LDAvis visualization tool with topic 1 from Reddit selected, and the “kill” term selected for proportional comparison.

Considering the overall theme of this work, selecting aggressive terms such as ‘kill’ should result in reasonably interesting findings, as seen in figure 4.19. The very large overlap between the selected topic and two others is clear as well, showing that it is not an especially topic specific term. Although the term is present through five of the six topics, it becomes clear that it is far and away the most prevalent in Topic 1, as seen in the length of the red versus gray bar on the right.

4.2.3 4chan

The LDA model generated from the sampled 4chan corpus contains the following 6 topics. Similarly to the Reddit topics, they are each defined by ten keywords and their weights, but must be manually labeled.

1. **Gun purchasing** - (0.022, gun) + (0.017, rifle) + (0.015, buy) + (0.014, like) + (0.010, shoot) + (0.009, want) + (0.009, good) + (0.008, round) + (0.008, look) + (0.008, barrel)

2. **Conversation** - (0.034, fuck) + (0.024, like) + (0.021, shit) + (0.014, look) + (0.013, get) + (0.013, guy) + (0.013, know) + (0.013, post) + (0.013, thread) + (0.009, think)

3. **Specific weapon usage** - (0.026, people) + (0.019, gun) + (0.012, fuck) + (0.011, like) + (0.011, think) + (0.010, right) + (0.009, kill) + (0.007, know) + (0.007, say) + (0.007, thing)
CHAPTER 4. RESULTS & ANALYSIS

Figure 4.20: Visualization of Topic 5, with the relevance parameter $\lambda$ tweaked to isolate topic specific terms.

4. **Gun ownership laws** - (0.015, gun) + (0.012, live) + (0.011, get) + (0.010, go) + (0.010, year) + (0.010, state) + (0.009, shoot) + (0.007, time) + (0.006, law) + (0.006, polic)

5. **Weapon usage** - (0.009, use) + (0.009, like) + (0.008, shoot) + (0.006, hit) + (0.006, need) + (0.005, knife) + (0.004, time) + (0.004, bullet) + (0.004, hand) + (0.004, wear)

6. **Military** - (0.013, war) + (0.011, militari) + (0.007, tank) + (0.007, armi) + (0.007, forc) + (0.006, russian) + (0.005, fight) + (0.005, like) + (0.005, oper) + (0.005, countri)

The 2nd topic here is particularly interesting, and serves to differentiate 4chan (and other similar communities) from the majority of discourse on the web. The use of particularly offensive language and in-group jokes and memes serves to “gate-keep” newcomers to the community, such that certain values that define the discourse on that community are preserved [66].

The 3rd and 5th topics were more difficult to classify however, due to additional terms that confounded the classification, such as non-specific words such as “like”, “think”, “right”, “know”, “say”, and “thing”. Again, similarly to the Reddit model, LDAvis was used to better comprehend the greater context of terms within topics 3 and 5 to label them. Similarly to before, the % of relevant terms alongside each topic is as follows:

1. Gun purchasing - 19.2%
2. Conversation - 18.7%
3. Specific weapon usage - 16.2%
4. Violence/attacks - 16%
5. Military - 15.4%
6. Gun legislation - 14.3%

The overall distribution of topics is fairly even throughout the space, with little overlap between the represented topics.
4.2.4 Telegram

Finally, similarly to above, the following 6 topics were created by the LDA fitted to the Telegram data, and solely in this case, re-organized in order of topic relevance. In this case, however, many of the topics share very similar, generic terms. The visualization from LDAvis in figure 4.22 confirms this, with topics 2 through 4 heavily overlapping each other, and topic 1 also being uncomfortably close with terms used. Topic 6 is dominated by two terms, “test” and “text”, and upon checking the original dataset in table A.13, test text is in fact the most sent message text, consisting of 5.7% of total messages sent. The base assumptions that LDA makes for the structure of documents within a corpus appear to be incorrect in this case, likely due to the extremely short length of each document/message, together with the sheer amount of noise present in an informal chat group. A human reading highlights the constant reuse of in-group references, and the fact that most conversations also involve images, videos, and emojis for context and semantic meaning.

When looking at the following labels and topics, the difficulty of the LDA model in classifying the text becomes even clearer.

1. Weapon related - (0.021, like) + (0.015, gun) + (0.010, go) + (0.010, guy) + (0.009, look) + (0.008, say) + (0.007, year) + (0.007, shit) + (0.007, think) + (0.006, peopl)
2. Purchasing - (0.025, like) + (0.020, good) + (0.017, think) + (0.012, look) + (0.012, want) + (0.012, dont) + (0.010, work) + (0.009, know) + (0.009, buy) + (0.008, pretti)
3. Conversation - (0.028, get) + (0.025, fuck) + (0.013, know) + (0.010, shit) + (0.008, want) + (0.008, look) + (0.008, way) + (0.007, lol) + (0.007, like) + (0.007, fun)
4. Gun-related - (0.014, time) + (0.012, right) + (0.010, gun) + (0.009, tri) + (0.009, like) + (0.008, post) + (0.007, need) + (0.007, actual) + (0.006, tfw) + (0.006, sure)
5. In-group slang - (0.041, kek) + (0.028, nice) + (0.025, yeah) + (0.018, yes) + (0.010, need) + (0.009, hot) + (0.008, morn) + (0.007, mean) + (0.007, mag) + (0.007, best)
6. Test messages - (0.219, test) + (0.216, text) + (0.009, oof) + (0.007, like) + (0.006, true) + (0.004, yup) + (0.004, glock) + (0.004, want) + (0.003, nah) + (0.003, need)
CHAPTER 4. RESULTS & ANALYSIS

It may be possible that more topics are required for this specific community or style of community, in order to have more topics with less overlap of relevant terms.

In short, the format of a chat group differs very heavily from that of a forum-style community, and solving those problems through specific chat optimized natural language models [69] may need to be explored further, due to its applicability both in this case and for other media such as Twitter. The collected dataset may still be useful for further research, however.

![Figure 4.22: The output from the LDAvis visualization tool when prepared with the Telegram model, corpus and dictionary.](image)

4.2.5 Reddit vs 4chan

After the models were generated, the Jaccard distance was computed and displayed as per figures 4.23. The first thing to note is that the brightest colors are the topic pairs that are the least similar. Thus, the lower the number, the more similar the topic pair is.

It is important to note the constraints, however. Jaccard distance does not consider the weighting of various terms within each model, nor does it take into consideration semantic similarities between different words, i.e. synonyms or classes of words such as intensifiers or superlatives, which serve to give emphasis to a post or sentence. The former is difficult to solve without using a different difference model, but in this case has been limited in this case by using the 20 most heavily weighted terms, and discarding the rest as not relevant enough to the topic.

Note, that due to the extremely unclear topics generated from the Telegram dataset, the relationships between models were extremely vague, with no definitive statement to be made.

Figure 4.23 displays very clearly the differences between the 4chan and reddit topics, with the vast majority of topic pairings being extremely dissimilar, with fewer than 30% of shared word topics. The following two topic pairs have the greatest similarity, and are representative of the types of topic found.

- Conversation, Conversation: The similarity here is not unexpected, with nearly 45% of terms being shared. This strongly implies that these terms are simply a part of standard discourse on online social communities.

- Gun usage, Gun ownership: With this topic pair, defining the difference between gun usage and ownership topics appears to be community dependent, and affected significantly by weighting, since both topics are essentially about the same overarching topic.
Even considering the loss of granularity in comparing unweighted topic terms against each other, there is definitely some overlap between the Reddit and 4chan weapon communities, although the majority of topic pairs appear to be unrelated, in some cases entirely so. The topic pair of Reddit-specific slang and Military has zero terms in common, a trend that is closely followed by both the column and row. This is not unexpected, considering the wildly different moderating styles and fewer restrictions present for the 4chan community versus the Reddit community. The Reddit slang topic is primarily about the mechanics of posting or comments, whereas the Military topic is entirely about terms that are not directly related to guns, which would likely removed by community moderators on Reddit.

The distribution of relevant terms per topic is also much more even between the topics for 4chan, with a much larger distinction between the largest and smallest frequencies.

![Figure 4.23: A heatmap of the Jaccard distance between the Reddit and 4chan models.](image)

### 4.3 Shock analysis

In this section, visual analysis of the shocks is performed, attempting to observe any patterns in the distributions of topics between around a shock. The intent is to try to find any possible predictors, with emphasis on attacks with high loss of life, for hopefully a more pronounced response. Apart from general observations made after reviewing the separate images, the various extremes of the dataset are also specifically investigated for any interesting features. Specifically, gun-related topics and the shocks where the most documents have been classified into likely being about that topic are of interest. Thus, for each interval, four topic distributions were generated from the posts in a two week interval around each shock, and plotted against the baseline, which is the topic distribution for the entire training corpus. Each graph consists of

#### 4.3.1 Reddit

For the Bataclan shooting, we see a market shift in the Gun laws and Gun usage topics, with a slight drop in the purchasing weapons topic. The topics are loose enough that it is difficult to properly assess which terms specifically caused the various shifts in probability, but the overall trend for this attack is an increase in gun usage topics, with a slight decrease in purchasing weapons and gun laws. It may be due to the predominantly American userbase of Reddit, whereas this
attack happened in France. Thus, due to the lack of immediate relevance to the concerns of most of the subreddit users

However, the response to the Las Vegas Strip massacre was markedly different. The first difference is that the conversation topic dropped heavily from before and after the attack, indicating a much more polarizing effect on the population. The topics shifted towards gun laws, and the possibility of gun legislation. This may be discussing legislation to prevent future attacks, or conversation Gun usage topics stayed about the same from the week before to the fortnight after, although the purchasing weapons topic rose slightly. The latter may be in response to the attack, where one of the major narratives put forward to purchase weapons is that of personal or self defense. Thus, if the populations feel afraid after an attack and feel that their recourse is to purchase weapons to defend themselves, it is reasonable that individuals not previously part of the reddit guns community have questions as to the details of buying weapons. TODO: Compare against other American attacks
4.3.2 4chan

Unfortunately, due to corrupted and/or incomplete data from the dataset, the following shocks were removed for 4chan analysis: The Trestle Trail bridge shooting, Charleston Church Shooting, Chattanooga military recruitment center attack, and Umpqua Community College shooting. TODO: Add notes for Bataclan and LV

In the case of the shocks drawn from the 4chan dataset around the Bataclan attacks, any shift in topics is far less pronounced. Indeed, most remain well within the standard deviation of each shock, apart from the attack topic being a very large outlier two weeks before the event. This is not that strange an outlier, considering the previous attack on the list of shocks was just under two weeks before, namely the Colorado Springs shooting, where there were three casualties.

For the Las Vegas strip attack, there is a shift in topics discussed, primarily that of an increase in the conversation topic, and a drop in the gun ownership topic. The former is especially interesting, due to the conversation topic being a certain level of background noise. An inverse response than
expected could indicate either a lack of response to the attack from the community as a whole, or the need for additional topics for improved granularity. It may be possible that there are subtopics that were not as prevalent or defined enough to be obvious with the final settings for the Topic models.

The number of documents likely about gun ownership has also dropped in frequency after the attack, the opposite result when compared to related topics from the Reddit model for the same attack. This reinforces the point that the weapons related communities on 4chan and Reddit differ heavily in terms of interests and even topic strength or reactions.

If we look further at the Colorado Springs attack in figure 4.28, there is a pattern of a drop in the conversation and military topics, which are likely to be the topics most likely to be independent of attacks. In fact, there is a drop in the gun legislation topic as well, all of which are made up for in the specific weapon usage topic. When researching the event further, it becomes clear that this reaction is not due to the number of casualties, which are not far outside the norm, but likely the location of the attack, at a Planned Parenthood clinic in Colorado Springs. Although it is difficult to tell if the community was incensed by the attack or not, it cannot be denied that the response here is extremely strong.

If we look at the corresponding set of shocks in figure 4.29, it does not appear to be as great a difference, with the indicators for a heavily polarized community, such as large shifts in conversation or range shooting not changing very much. There is a slight increase in discussion about gun laws and corresponding decrease in Gun usage.

4.3.3 Discussion

While comparing the pattern of topics across shocks, it becomes clear that each community reacts very differently to attacks. There appear to be more factors at work than simply the number of casualties, and better controlling and identifying those factors without falling into the fallacy of post hoc reasoning is likely required.

Unfortunately, due to the lack of correlation or definitive pattern of activity around shocks, the optimal objective of finding possible predictors of attacks based on topic modeling is not possible. However, it may still be possible to gain some insight as to how communities react to events, especially with more precisely optimized and tuned models. Furthermore, it may be possible that the sub-communities of online underground arms buyers are not present enough on the clearweb communities that were investigated. It may be interesting to investigate politically motivated
communities on the clearweb instead, and with other natural language tools such as sentiment analysis for aggression or toxicity as well.

Figure 4.29: Reddit Colorado Springs attack.
Chapter 5

Conclusions

To conclude, the objective of this work was to improve the understanding of the nature of underground dark web markets. This involved firstly an extended period of scraping listing information from dark web markets, in order to assess the scale of the market and transaction information. Separately, in order to try identify and measure the communities that may transact on these markets, natural language processing methods were used in order to quantify and process large social communities, with the intent of characterizing these communities. The models generated by doing so enabled understanding how these communities responded to various shocks, or shooting events, and differentiating between them.

5.1 Dark web markets

The illegal markets found by scraping and parsing various listings online resulted in building a dataset of approximately 1,600 active listings over a three week period. The longitudinal aspect showed activity, such as new listings being made, and various metrics such as the number of sales, promoting the idea that these markets were not solely filled with scams, but had genuine traffic and transactions. Of these markets, the one cryptomarket, Berlusconi market proved to have the most data for understanding the categories of items being sold, with a large number of different brands being shipped worldwide from both European and North American dealers with surprisingly low markups in comparison to offline underground markets [27]. Short of actually buying weapons, checking further legitimacy of the vendors on Berlusconi market seems to have hit a standstill, with no other cryptomarkets thus far considering the sale of weapons as a possible income stream. Conversely, the single vendor markets found through scraping the dark web showed remarkably little change over time, with very low prices, especially in comparison to the vendors on Berlusconi market. The reasons for this behavior vary, with either the idea that these vendors consider the cost of constantly changing their offerings being greater than any potential gain, or that these markets are scams, and are competitively priced against Berlusconi market vendors simply as a honeypot for buyers. Finally, the distribution of weapons and brands found online were reasonably similar across both types of markets, selling mostly pistols and rifles, albeit with the consideration that there were a significant number of custom listings on Berlusconi market. The latter point is interesting, and marks a shift from the findings in the RAND report [55]. Where previously, vendors tended to simply list their offerings, the significant number of listings that were custom likely implies a certain degree of communication and thus trust relationships being formed between buyers and sellers. The latter point is especially interesting - in effect, cryptomarkets likely allow for initial trust relationships to be established with minimal risk, after which buyers likely continue to stick with a vendor that has delivered consistently in the past.

In short, although the scale of transactions on the dark web is still relatively small in comparison to the markets for either drugs or malware, the market until recently was alive and well. However, in the time since the collection of primary data, Berlusconi market has closed its doors to weapon...
vendors, with many individual accounts having been banned. Whether this proves to actually stymie the flow of weapons across the dark web remains to be seen, but at least shows that the effect of law enforcement and various police organizations around the world have had some effect on the trade of illegal weapons across the dark web.

5.2 Topic models

From the various communities identified that may contain individuals that may purchase weapons from the dark web, only two valid LDA models, from the Reddit and 4chan communities were effectively generated. The differences between the two communities was very clear, with a far more even spread of topics from 4chan instead of Reddit. The latter had a much more biased distribution towards specific topics such as gun legislation and usage. Despite reasonably heavy overlap between certain topics, the heavy moderation present on the subreddit may be a contributing factor to these biases, by removing posts that are considered to be off topic. On the other hand, both the lack of moderation and more extreme views on 4chan may have contributed firstly to the more even distribution of frequencies of various terms per topic, and secondly to the possible need for models with better hyper-parameter selection in order to distinguish the optimal number of topics, for example, or the learning rate on the models. Regardless, the difference between the two communities was clear through the use of LDA modeling, with 4chan using many extreme and violent terms in comparison to Reddit. Unfortunately, the Telegram community from 8chan that was infiltrated proved to fit very poorly with the LDA model used. It may be possible that LDA models are not well tailored to the format of discourse within a chat group, or that additional topics were necessary, due to the largely non-specific conversations that were recorded. Considering the difference between the Reddit and 4chan results, the difference again between these communities and that of even more extreme boards, such as 8chan, would have been extremely interesting, especially considering the recent removal of 8chan from its various hosting services.

5.3 Shocks

The use of the LDA models from above to generate probability distributions of the posts surrounding mass shootings was interesting as a method of visualizing a community’s response to a shock. The distinction between certain patterns of behavior was evident. Some communities responded strongly to shocks far more strongly than others, with relatively large shifts in topics such as gun legislation, implying a degree of causality. However, for some shocks that one would expect a strong response from, the opposite occurred, with an increase in topics that were expected to either decrease or stay the same, such as the military topic from 4chan or conversation from Reddit. Thus, an overall pattern of responses was not clear or consistent, leading one to believe that there are other factors that have not been taken into consideration.

Despite that, however, these results served to reinforce the differences between the two communities, both in terms of properly assessing what the different overall distribution of topics means, and what shocks these communities care about much more.

5.4 Future work

The scope for future work here is extremely significant. Firstly, for further characterization of arms markets, there is always opportunity for further study. Without a currently active cryptomarket selling arms, it seems not unreasonable that one of the currently growing cryptomarkets may become open to their sale once again. Analyzing these new markets and how various restrictions affect the market may be of interest. Furthermore, the possibility of tracking specific vendors across cryptomarkets may be interesting. The not inconsiderable effort of setting up on a new cryptomarket implies that there is enough trade of arms such that doing so is financially worthwhile. Until the next generation of cryptomarkets becomes open to arms sales, however, the
CHAPTER 5. CONCLUSIONS

single vendor markets can still be monitored for shifts in price or services offered, especially now
since the market has shrunk.

For the LDA models, there are two main areas for future improvement. Firstly, although the
initial generation of results from the text corpora resulted in meaningful results, there is definitely
a great deal more that could have been done to improve the results gathered. LDA models are first
and foremost based upon a statistical estimation of the distribution of terms and topics within a
text corpus. Adding semantic understanding to LDA models would likely improve the quality of
the topics generated, and can be done through the use of a model framework such as word2vec,
or doc2vec. Separately, the models generated for this work could have been tuned significantly
from their current state. There are a number of hyper-parameters that could have been varied,
while using either perplexity, coherence, or accuracy as scoring criteria. The hyper-parameters
include the number of topics generated, the number of passes, and sample size while comparing
the performance of the models on the test sets. Building more accurate models should significantly
improve comparative analysis, and should also aid in identifying community responses to shocks.

In addition to better tuning topic models, use of different natural language processing models
could also be of interest. For example, using sentiment analysis models that are tuned to detect
aggression or toxicity [41] may be relevant, and shed further light on the differences between online
communities. Alternatively, given the intent of identifying sub-communities that are more likely
to either purchase weapons from the dark web or perform future attacks, shifting the focus of
communities either to ones that are more extreme, such as 8chan if it ever returns, or to more
politically motivated communities, such as the infamous /pol/ board on 4chan [35].
Bibliography


[5] dropgangs, or the future of dark markets. 14

[6] Ii.c.9 declaration (iv-3) respecting the prohibition of the use of expanding bullets (29 july 1899). International Law & World Order, page 1–1, Jul 1899. 7


[8] Charleroi policeman arrested for buying explosives from dark net, Oct 2016. 15

[9] Boy, 14, tried to buy sub-machine gun on the dark web, court hears, Apr 2017. 15

[10] Teenage submachine gun buyer hit with a terrorism charge, Oct 2017. 15


[12] Ireland police officer convicted of buying drugs & guns on dark web, Jan 2018. 15


[20] Brussels: Chamber of Representatives of the Kingdom of Belgium. Article 19. law regulating economic and individual activities with weapons 2006-06-08/30, 2006. 9

[22] caliens. Dream market will permanently suspend operations next month. 17

[23] Stefan Candea, Jürgen Dahlkamp, Jörg Schmitt, Andreas Ulrich, Wolf Wiedmann-Schmidt, and et al. The path to death: How eu failures helped paris terrorists obtain weapons - spiegel online - international, Mar 2016. 15


[29] Anthony Cuthbertson. Dark web criminals switch to popular apps to sell drugs, using bots and secret graffiti messages to do business. 22


[33] Südwest Presse Online-Dienste GmbH. Polizei: Waffenkauf übers darknet scheitert, Jan 2017. 15, 19

[34] Gavin Hales, Chris Lewis, and Daniel Silverstone. Gun crime: the market in and use of illegal firearms. 2006. 11


[36] Justin Huggler. Man arrested in germany on suspicion of illegal arm dealing in terror crackdown, Nov 2015. 15

[37] Jessica Hullinger and Jessica Hullinger. Do people really buy guns on the dark web?, Apr 2017. 1, 16


[42] Michelle Ye Hee Lee. Obama’s inconsistent claim on the ‘frequency’ of mass shootings in the u.s. compared to other countries, Dec 2015. 25


[45] DPA/The Local. ‘migrant shock’, a far-right website selling illegal guns, taken down, Feb 2017. 15

[46] Anne Losensky. Berliner darknet-dealer muss vier jahre ins gefängnis, Sep 2016. 15


[48] Kyle Mizokami. The rifle that refuses to die: Why the m4 carbine won’t go away, 2015. 8

[49] RTÉ News. Psni officer accused of trying to procure gun, Sep 2016. 15

[50] Office of the Law Revision Counsel of the House of Representatives. 18 u.s.c. § 923(g)(3. Title 18, Part 1, Chapter 44, Section 923, Subsection (g)(3). 9

[51] Peter Oldenburger. 33-jähriger bestellt waffe im darknet und wird gefasst, Nov 2016. 15

[52] United Nations Office on Drugs and Crime. *International Instrument to Enable States to Identify and Trace, in a Timely and Reliable Manner, Illicit Small Arms and Light Weapons*. United Nations Office on Drugs and Crime, Feb 2013. 3

[53] Howell O’Neill. Schan is home to a hive of pedophiles, Mar 2017. 21

[54] Danny Palmer. Criminals in the cloud: How malware-as-a-service is becoming the tool of choice for crooks, Apr 2016. 1, 21


[56] Thom Patterson. Inside the illegal online weapons trade, Aug 2016. 15


[67] Benjamin Vitáris. Belgian sentenced to community service for ordering weapons of war, Nov 2016. 15


Appendix A

Appendix

A.1 Appendix 1
<table>
<thead>
<tr>
<th>Shock</th>
<th>Date</th>
<th>Country</th>
<th>Deaths</th>
<th>Injured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sikh temple shooting</td>
<td>05/08/2012</td>
<td>USA</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>2012 Moscow</td>
<td>07/11/2012</td>
<td>Russia</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Menznau Shooting</td>
<td>27/02/2013</td>
<td>Switzerland</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Velika Ivanca Shooting</td>
<td>09/04/2013</td>
<td>Serbia</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Belgorod Shooting</td>
<td>22/04/2013</td>
<td>Russia</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Santa Monica rampage</td>
<td>07/06/2013</td>
<td>USA</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Annaberg Shooting</td>
<td>16/09/2013</td>
<td>Austria</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Fort Hood shooting 2</td>
<td>03/04/2014</td>
<td>USA</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Jewish museum of Belgium</td>
<td>24/05/2014</td>
<td>Belgium</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Moncton Shooting</td>
<td>04/06/2014</td>
<td>Canada</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Edmonton Shootings</td>
<td>29/12/2014</td>
<td>Canada</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Charlie Hebdo + Others</td>
<td>09/01/2015</td>
<td>France</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Umpqua Community College shooting</td>
<td>01/10/2015</td>
<td>USA</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>San Bernardino mass shooting</td>
<td>02/12/2015</td>
<td>USA</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>Uhersky Brod shooting</td>
<td>24/02/2015</td>
<td>Czech Republic</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Trestle Trail bridge shooting</td>
<td>11/06/2015</td>
<td>USA</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Bataclan attacks</td>
<td>13/11/2015</td>
<td>France</td>
<td>131</td>
<td>413</td>
</tr>
<tr>
<td>La Loche School Shooting</td>
<td>22/01/2016</td>
<td>Canada</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Zitiste Shooting</td>
<td>02/07/2016</td>
<td>Serbia</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Dallas police shooting</td>
<td>07/07/2016</td>
<td>USA</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Munich Shooting</td>
<td>22/07/2016</td>
<td>Germany</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>Orlando nightclub massacre</td>
<td>12/06/2016</td>
<td>USA</td>
<td>49</td>
<td>53</td>
</tr>
<tr>
<td>Quebec City mosque shooting</td>
<td>29/01/2017</td>
<td>Canada</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Florida awning manufacturer shooting</td>
<td>06/06/2017</td>
<td>USA</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Fort Lauderdale airport shooting</td>
<td>06/01/2017</td>
<td>USA</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>San Francisco UPS shooting</td>
<td>14/06/2017</td>
<td>USA</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Pennsylvania supermarket shooting</td>
<td>07/06/2017</td>
<td>USA</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Las Vegas Strip massacre</td>
<td>01/10/2017</td>
<td>USA</td>
<td>58</td>
<td>546</td>
</tr>
<tr>
<td>Edgewood business park shooting</td>
<td>18/10/2017</td>
<td>USA</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Walmart shooting in suburban Denver</td>
<td>01/11/2017</td>
<td>USA</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Texas First Baptist Church massacre</td>
<td>05/11/2017</td>
<td>USA</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Rancho Tehama shooting spree</td>
<td>14/11/2017</td>
<td>USA</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Rural Ohio nursing home shooting</td>
<td>12/05/2017</td>
<td>USA</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mohawk Valley shootings</td>
<td>13/03/2013</td>
<td>USA</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Sandy Hook Elementary massacre</td>
<td>14/12/2012</td>
<td>USA</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Chattanooga military recruitment center</td>
<td>16/07/2015</td>
<td>USA</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Washington Navy Yard shooting</td>
<td>16/09/2013</td>
<td>USA</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Charleston Church Shooting</td>
<td>17/06/2015</td>
<td>USA</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Fresno downtown shooting</td>
<td>18/04/2017</td>
<td>USA</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Alturas tribal shooting</td>
<td>20/02/2014</td>
<td>USA</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Kalamazoo shooting spree</td>
<td>20/02/2016</td>
<td>USA</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Aurora theater shooting</td>
<td>20/07/2012</td>
<td>USA</td>
<td>12</td>
<td>70</td>
</tr>
<tr>
<td>Pinewood Village Apartment shooting</td>
<td>21/04/2013</td>
<td>USA</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Isla Vista mass murder</td>
<td>23/05/2014</td>
<td>USA</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Cascade Mall shooting</td>
<td>23/09/2016</td>
<td>USA</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Marysville-Pilchuck High School shooting</td>
<td>24/10/2014</td>
<td>USA</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Excel Industries mass shooting</td>
<td>25/02/2016</td>
<td>USA</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Hialeah apartment shooting</td>
<td>26/07/2013</td>
<td>USA</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Baton Rouge police shooting</td>
<td>27/07/2016</td>
<td>USA</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Accent Signage Systems shooting</td>
<td>27/09/2012</td>
<td>USA</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Planned Parenthood clinic</td>
<td>27/11/2015</td>
<td>USA</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Colorado Springs shooting rampage</td>
<td>31/10/2015</td>
<td>USA</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table A.1: List of generated shocks.
APPENDIX A. APPENDIX

<table>
<thead>
<tr>
<th>Market</th>
<th>Source</th>
<th>Reason for removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cannazon</td>
<td>deepdotweb, dark.fail</td>
<td>Only drugs</td>
</tr>
<tr>
<td>Berlusconi market</td>
<td>deepdotweb, dark.fail</td>
<td>Not removed</td>
</tr>
<tr>
<td>CGMC</td>
<td>deepdotweb</td>
<td>Only drugs</td>
</tr>
<tr>
<td>Acropolis market</td>
<td>deepdotweb</td>
<td>Only drugs</td>
</tr>
<tr>
<td>Tochka</td>
<td>deepdotweb, dark.fail</td>
<td>Only drugs</td>
</tr>
<tr>
<td>The Majestic Garden</td>
<td>deepdotweb, dark.fail</td>
<td>Only drugs</td>
</tr>
<tr>
<td>Italian Deep Web</td>
<td>deepdotweb</td>
<td>Italian speaking</td>
</tr>
<tr>
<td>Italian DarkNet community</td>
<td>deepdotweb</td>
<td>Italian speaking</td>
</tr>
<tr>
<td>Hydra</td>
<td>deepdotweb, dark.fail</td>
<td>Russian speaking</td>
</tr>
<tr>
<td>WayAway</td>
<td>deepdotweb, dark.fail</td>
<td>Russian speaking</td>
</tr>
<tr>
<td>RuTor</td>
<td>deepdotweb</td>
<td>Russian speaking</td>
</tr>
<tr>
<td>Empire market</td>
<td>dark.fail</td>
<td>No weapons</td>
</tr>
<tr>
<td>Cryptonia</td>
<td>dark.fail</td>
<td>No weapons</td>
</tr>
<tr>
<td>Genesis market</td>
<td>dark.fail</td>
<td>No weapons</td>
</tr>
<tr>
<td>Apollon market</td>
<td>dark.fail</td>
<td>No weapons</td>
</tr>
<tr>
<td>Monopoly market</td>
<td>dark.fail</td>
<td>No weapons</td>
</tr>
</tbody>
</table>

Table A.2: List of cryptomarkets, their sources, and reasons for removal.

<table>
<thead>
<tr>
<th>Cryptomarket</th>
<th>Reason for closure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabay</td>
<td>Seized by law enforcement</td>
</tr>
<tr>
<td>Dreammarket</td>
<td>Shut down</td>
</tr>
<tr>
<td>Valhalla</td>
<td>Seized by law enforcement</td>
</tr>
<tr>
<td>Hausa</td>
<td>Seized by law enforcement</td>
</tr>
<tr>
<td>Oasis market</td>
<td>Exit Scam</td>
</tr>
<tr>
<td>Python market</td>
<td>Inaccessible</td>
</tr>
<tr>
<td>TheDetox</td>
<td>Inaccessible</td>
</tr>
<tr>
<td>Traderoute</td>
<td>Inaccessible</td>
</tr>
<tr>
<td>Minerva</td>
<td>Inaccessible</td>
</tr>
<tr>
<td>Acropolis</td>
<td>Shut down</td>
</tr>
<tr>
<td>Tochka</td>
<td>No weapon listings found</td>
</tr>
</tbody>
</table>

Table A.3: List of previous markets and their current status.
Single vendor market | Source | Status
--- | --- | ---
Deep web guns store | thedarkweblinks, MASSDEAL | Available
Black market guns and ammo | thedarkweblinks | Available
UK guns and ammo | thedarkweblinks | Available
Euro Guns 1 | thedarkweblinks, MASSDEAL | Available
Euro Guns 2 | thedarkweblinks | Available
Black market guns | thedarkweblinks | Inaccessible
Guns and ganja | thedarkweblinks | Available
Luckp store | thedarkweblinks, MASSDEAL | Available
Darkseid | thedarkweblinks, MASSDEAL | Available
Black market guns (BMG) | thedarkweblinks | Available
Alpha Guns | thedarkweblinks, MASSDEAL | Inaccessible
Guns dark market | thedarkweblinks | Inaccessible
The armory | thedarkweblinks | Inaccessible
BlackGhosts | thedarkweblinks | Inaccessible
Products | thedarkweblinks | Inaccessible
Glocks and Taurus | thedarkweblinks | Inaccessible
BesaMafia | thedarkweblinks, MASSDEAL | Exit Scam
Sweguns | thedarkweblinks, MASSDEAL | Inaccessible

Table A.4: List of single vendor markets, their sources, and their status.

<table>
<thead>
<tr>
<th>Field</th>
<th>Mandatory</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listing ID</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Title</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Price in Bitcoin</td>
<td>N</td>
<td>Float</td>
</tr>
<tr>
<td>Price in Fiat</td>
<td>Y</td>
<td>Currency</td>
</tr>
<tr>
<td>Stock</td>
<td>N</td>
<td>String</td>
</tr>
<tr>
<td>Vendor</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Ships from</td>
<td>N</td>
<td>Country</td>
</tr>
<tr>
<td>Ships to</td>
<td>N</td>
<td>String</td>
</tr>
<tr>
<td>Number sold</td>
<td>Y</td>
<td>Integer</td>
</tr>
<tr>
<td>Details</td>
<td>N</td>
<td>String</td>
</tr>
</tbody>
</table>

Table A.5: Listing fields for Berlusconi market.

<table>
<thead>
<tr>
<th>Field</th>
<th>Mandatory</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Last Seen</td>
<td>Y</td>
<td>Timestamp</td>
</tr>
<tr>
<td>First Seen</td>
<td>Y</td>
<td>Timestamp</td>
</tr>
<tr>
<td>Ships From</td>
<td>Y</td>
<td>Country abbreviation</td>
</tr>
<tr>
<td>Positive Feedback</td>
<td>Y</td>
<td>Integer</td>
</tr>
<tr>
<td>Neutral Feedback</td>
<td>Y</td>
<td>Integer</td>
</tr>
<tr>
<td>Negative Feedback</td>
<td>Y</td>
<td>Integer</td>
</tr>
<tr>
<td>Orders Finalized</td>
<td>N</td>
<td>Integer</td>
</tr>
<tr>
<td>Profile Text</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Terms &amp; Conditions</td>
<td>Y</td>
<td>String</td>
</tr>
</tbody>
</table>

Table A.6: Vendor fields for Berlusconi market.
<table>
<thead>
<tr>
<th>Field</th>
<th>Mandatory</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Name</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Price in Bitcoin</td>
<td>N</td>
<td>Float</td>
</tr>
<tr>
<td>Price in Fiat</td>
<td>Y</td>
<td>Currency</td>
</tr>
<tr>
<td>Currency</td>
<td>Y</td>
<td>String</td>
</tr>
<tr>
<td>Details</td>
<td>N</td>
<td>String</td>
</tr>
</tbody>
</table>

Table A.7: Listing fields for single vendor markets.
A.2 Dark web scraping

As per the methodology, a scraper and parser were written in Python. The scraper was then run daily for a period of three weeks, from the 26th of May 2019 until the 6th of June 2019. From there, a database as per tables A.5 A.6 was generated, in order to compare against single vendor markets.

Similarly, over the same time period, a separate scraper and parser were written in Python to download and parse the web pages for each single vendor market. Each listing was generated as per A.7.

A.2.1 Description of Berlusconi market

This section is a description of the raw, unprocessed data gathered from Berlusconi market dataset over 21 days, from 26/05/2019 to 14/06/2019. Tables A.8 and ?? includes the description of the various fields of the listing dataset. Table A.9 includes the same for vendors.

A.2.2 Single vendor markets

This subsection covers the descriptive statistics for the collected single vendor markets over 21 days, from 26/05/2019 to 14/06/2019.

Although PGP keys were collected, after cross-referencing those from Berlusconi market vendors and single vendor markets, there were no matches, implying that vendors tend to keep to a single type of underground market, and do not run a single vendor market and maintain and advertise their presence on cryptomarkets.
### Table A.8: Descriptive statistics for Berlusconi market listings.

<table>
<thead>
<tr>
<th>listing_number</th>
<th>title</th>
<th>currency</th>
<th>stock</th>
<th>vendor</th>
<th>ships_from</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>36,397</td>
<td>36,397</td>
<td>36,397</td>
<td>35,930</td>
<td>36,397</td>
</tr>
<tr>
<td>unique</td>
<td>2,257</td>
<td>1,212</td>
<td>1</td>
<td>293</td>
<td>38</td>
</tr>
<tr>
<td>top</td>
<td>0a9ca729f77b</td>
<td>BUY Brand...</td>
<td>EUR</td>
<td>In</td>
<td>mrpuff</td>
</tr>
<tr>
<td>freq</td>
<td>21</td>
<td>2,012</td>
<td>36,397</td>
<td>5,359</td>
<td>11,071</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ships_to</th>
<th>details</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>36,167</td>
<td>36,397</td>
</tr>
<tr>
<td>unique</td>
<td>1</td>
<td>776</td>
</tr>
<tr>
<td>top</td>
<td>Worldwide</td>
<td>Brand new Guns and case...</td>
</tr>
<tr>
<td>freq</td>
<td>36,167</td>
<td>4483</td>
</tr>
<tr>
<td>first</td>
<td>2019-05-26</td>
<td></td>
</tr>
<tr>
<td>last</td>
<td>2019-06-15</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>price_fiat</th>
<th>number_sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>36,397.00</td>
</tr>
<tr>
<td>mean</td>
<td>943.70</td>
</tr>
<tr>
<td>std</td>
<td>1,374.76</td>
</tr>
<tr>
<td>min</td>
<td>5.00</td>
</tr>
<tr>
<td>25%</td>
<td>500.00</td>
</tr>
<tr>
<td>50%</td>
<td>600.00</td>
</tr>
<tr>
<td>75%</td>
<td>900.00</td>
</tr>
<tr>
<td>max</td>
<td>26,994.60</td>
</tr>
</tbody>
</table>

### Table A.9: Descriptive statistics for Berlusconi vendor listings.

<table>
<thead>
<tr>
<th>name</th>
<th>last_seen</th>
<th>first_seen</th>
<th>ships_from</th>
<th>profile_text</th>
<th>terms</th>
<th>pgp_exists</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>unique</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>13</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>top</td>
<td>rdc13</td>
<td>2019-06-14</td>
<td>06:26:55</td>
<td>ZZ</td>
<td>Send your Data...</td>
<td>True</td>
</tr>
<tr>
<td>freq</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>22</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feedback_pos</th>
<th>feedback_neutral</th>
<th>feedback_neg</th>
<th>orders_finalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>mean</td>
<td>24.357143</td>
<td>0.660714</td>
<td>0.392857</td>
</tr>
<tr>
<td>std</td>
<td>81.114381</td>
<td>3.491688</td>
<td>2.154729</td>
</tr>
<tr>
<td>min</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>25%</td>
<td>0.750000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>50%</td>
<td>3.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>75%</td>
<td>13.250000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>max</td>
<td>574.000000</td>
<td>25.000000</td>
<td>16.000000</td>
</tr>
</tbody>
</table>

Table A.9: Descriptive statistics for Berlusconi vendor listings.
Table A.10: Descriptive statistics for single vendor market listings.

<table>
<thead>
<tr>
<th>vendor</th>
<th>title</th>
<th>currency</th>
<th>details</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>4,137</td>
<td>4,137</td>
<td>3,339</td>
<td>693</td>
</tr>
<tr>
<td>unique</td>
<td>9</td>
<td>190</td>
<td>2</td>
<td>33</td>
</tr>
</tbody>
</table>

**top**
- **title**: Desert Eagle IMI
- **currency**: EUR
- **details**: Tactical rifle is the ult...
- **date**: 2019-05-29

**freq**
- 1575
- 63
- 1701
- 197

**freq**
- 2019-05-26
- 2019-06-15

<table>
<thead>
<tr>
<th>price_btc</th>
<th>price_fiat</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>1470.000000</td>
</tr>
<tr>
<td>mean</td>
<td>0.085400</td>
</tr>
<tr>
<td>std</td>
<td>0.079697</td>
</tr>
<tr>
<td>min</td>
<td>0.002200</td>
</tr>
<tr>
<td>25%</td>
<td>0.038900</td>
</tr>
<tr>
<td>50%</td>
<td>0.064250</td>
</tr>
<tr>
<td>75%</td>
<td>0.116040</td>
</tr>
<tr>
<td>max</td>
<td>0.395520</td>
</tr>
</tbody>
</table>
A.3 Social communities

From the initial list of 4 social communities identified in section 3.2, these were cut down to 3 viable options. Both Reddit and 4chan have dedicated archival sites, which offer data dumps that can be parsed for relevant posts, comments, and submissions.

For collecting historical data from Reddit, https://pushshift.io/ is an archival project run by Jason Baumgartner, and contains nearly every submission and comment from 2011 to the current day in a series of compressed files.

For 4chan, there are a number of different archives, each of which tends to specialize in a certain subset of boards. Although most of these boards are directly query-able through APIs, the APIs are rate limited and restricted to just 25 posts per query, which makes using the API inadequate for the wholesale data dump required for this thesis. Fortunately, some of these archives have had the foresight to upload snapshots of their databases to internet archival projects such as archive.org. One of these archives is desuarchive.org, which archives /k/, the weapons-related board that is of interest. The archive dump can be found at https://archive.org/download/desu-2017-db, and contains all posts on /k/ from the 1st of July 2012 until the 1st of December 2017. This limits the list of generated shocks to this time period, and also means that the Reddit dataset will be limited to the same time period as well.

Unfortunately, 8chan does not have a dedicated archival project, and the internal archival mechanism is extremely incomplete, even for solely textual data. Therefore, posts from 8chan has been removed from any further results and analysis, despite the obvious interest in being able to quantify and analyze posts on the so-called “wild west” of the internet. Similarly, despite multiple requests for an API key from the administrators of Voat.co, a lack of response has resulted in the dataset being removed, since there is no external archival service, and writing a scraper for Voat quickly resulted in rate limitations that led to data acquisition being painfully slow.

Therefore, three datasets remain, of which two allow for shock extraction and analysis, namely Reddit and 4chan. The Telegram group stands separately, but provides a text corpus that is interesting to compare and contrast against the two less extreme communities.

A.3.1 Data extraction & processing

The available Reddit archive unfortunately is not broken down by subreddit, making the data extraction phase slightly more difficult. Extracting all the comments and submissions to the guns subreddit thus entailed downloading and integrity checking 472 GB of compressed JSON files from files.pushshift.io, and then stream decompressing and extracting the relevant submissions and comments. Once done, this resulted in 4.2 GB of uncompressed JSON files, with 6,412,993 comments and submissions combined, as per table A.11. It is important to note that although the majority of authors in the dataset have since deleted their accounts (or the linkage from their account to a post), the post text may still be present, and removing any post with a deleted author would be premature.

The best available 4chan archive of /k/ was an SQL database dump from desuarchive.org, which was in the format of a CSV file with a total size of 7.3GB. In order to minimize wasted effort, the post information within the CSV was then extracted and converted to JSON files, such that both Reddit and 4chan posts could be fed into the same processing and model generation pipeline. Unfortunately, since CSV formats are not standardized, the database was extracted through an SQL database dumping tool which sanitized newline characters in a manner not immediately parseable by the Python CSV library being used. This required additional pre-processing of the CSV file to remove improperly escaped newline characters and commas. After the CSV sanitization steps and JSON conversion, the total size of the /k/ dataset ended up being 21GB of JSON files, with 22,182,056 posts as per table A.12. This dataset spans over 5 years, from 2012-13-7 to 2017-12-5, and is thus factor defining the date range for the generation of shocks for the final topic analysis section.

The Telegram dataset consists of 62,268 messages dating from 2019-2-7 until the 2019-8-20, dumped in HTML format, as per table A.13. These were parsed, with authors and message text
extracted and converted to JSON in order to package for the LDA model generation pipeline.

At this point all the datasets were pre-processed and fed into the Model pipeline. The pre-processing steps entailed using regular expressions to first remove URLs from any posts, followed by removing deleted posts, punctuation, and numbers. The final step involved deaccenting characters, and converting the text corpus to lowercase. The pre-processed documents were then lemmatized and stemmed to create root forms, as per section 3.3.1.

At this point, a significant concern arose - the Reddit and 4chan datasets were simply too large to build models across the entirety of the corpora. Therefore, a 10% stratified sample was drawn from both the datasets on a per month basis to ensure that post distribution was maintained. This resulted in a training set of 549,945 and 2,213,670 elements for Reddit and 4chan respectively. These samples were used to generate the final LDA models, which can then be compared with each other, and in the case of the Reddit and 4chan models, view topic distributions for the various generated shocks.

<table>
<thead>
<tr>
<th>author</th>
<th>comment</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>6,412,993</td>
<td>6,412,993</td>
</tr>
<tr>
<td>unique</td>
<td>203,645</td>
<td>5,672,212</td>
</tr>
<tr>
<td>top</td>
<td>[deleted]</td>
<td>[deleted]</td>
</tr>
<tr>
<td>frequency</td>
<td>631,951</td>
<td>306,271</td>
</tr>
<tr>
<td>first</td>
<td>2012-01-01 00:00:08</td>
<td></td>
</tr>
<tr>
<td>last</td>
<td>2018-12-31 23:59:56</td>
<td></td>
</tr>
</tbody>
</table>

Table A.11: Descriptive statistics for the full Reddit dataset.

<table>
<thead>
<tr>
<th>comment</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>22,182,056</td>
</tr>
<tr>
<td>unique</td>
<td>21,278,557</td>
</tr>
<tr>
<td>top</td>
<td>\N</td>
</tr>
<tr>
<td>frequency</td>
<td>646,371</td>
</tr>
<tr>
<td>first</td>
<td>2012-07-13 14:32:57</td>
</tr>
<tr>
<td>last</td>
<td>2017-12-05 04:12:24</td>
</tr>
</tbody>
</table>

Table A.12: Descriptive statistics for the full 4chan dataset.
Table A.13: Descriptive statistics for the Telegram dataset.

<table>
<thead>
<tr>
<th></th>
<th>author</th>
<th>comment</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>62,268</td>
<td>62,268</td>
<td>62,268</td>
</tr>
<tr>
<td>unique</td>
<td>43</td>
<td>52,721</td>
<td>61,564</td>
</tr>
<tr>
<td>top</td>
<td>n a p s l</td>
<td>test text</td>
<td>2019-07-27 14:08:16</td>
</tr>
<tr>
<td>frequency</td>
<td>12,979</td>
<td>3,572</td>
<td>3</td>
</tr>
<tr>
<td>first</td>
<td></td>
<td></td>
<td>2019-02-07 19:32:30</td>
</tr>
<tr>
<td>last</td>
<td></td>
<td></td>
<td>2019-08-20 11:22:26</td>
</tr>
</tbody>
</table>