# SoK: Digging into the Digital Underworld of Stolen Data Markets

Tina Marjanov
*Department of Computer Science and Technology*
*University of Cambridge*
*tm794@cam.ac.uk*

Alice Hutchings
*Department of Computer Science and Technology*
*University of Cambridge*
*ah793@cam.ac.uk*

***Abstract*—Over the past few decades, the issue of stolen data has expanded from a nuisance caused by few opportunistic individuals to a thriving, highly organised, and profitable economy. As such, it spawned a thread of research trying to document and understand the underground economy. We look back at the past 15 years of research on stolen data markets to uncover the underlying patterns and trends, documented by researchers. We examine the economy and find a changing landscape, both in terms of popular stolen data types as well as the platforms housing the marketplaces. Additionally, we record a consistent decrease in market lifespans and as well as observation periods. We highlight a number of research patterns and potential shortcomings, in particular the low coverage of markets included in research and the low diversity of languages featured in the marketplaces. Finally, we propose a number of directions for future research to better understand the true cost of the economy and the mismatch between data breaches and data appearing on markets. Future research will also need to stay on top of the changing landscape and focus on timely identification of new trends and community movements across platforms.**

***Index Terms*—stolen data, underground markets, dark web, systematization of knowledge**

## 1. Introduction

Data has become an increasingly valuable asset in today's digital economy. Alongside the legally obtained and traded data in conventional markets, there exists a parallel economy where stolen data is traded in underground markets. This illicit trade represents a fast-paced economy that has spurred equally rapid research efforts aimed at documenting, understanding, and disrupting it. The first documented cases of digital data theft date back to the late 1980s and 1990s [1], [2]. While early cases were rare, primarily opportunistic, and perpetrated by individuals, we now live in a world where data theft is driven by a highly organised economy and each individual can expect to have their data or identity stolen multiple times [3]. Unlike most other illicit activities such as, for example, drug or arms trade, the stolen data economy is easily scalable and does not have the same location constraints. Additionally, it can have a profound financial and emotional impact on victims, which in turn may have relatively little influence on whether their data will be stolen and how it might be used. This makes the stolen data problem a costly one for the individuals that have to deal with the aftermath, as well as the data controllers and law enforcement trying to put a stop to it. As such it has attracted academic and industry attention alike.

In this paper, we present a comprehensive review of the recent academic research, documenting the trends, both in the underground economy and the scholarly efforts to study it. Applying an inductive approach to recent literature we explore the types of data that are attractive to offenders, how they are stolen and traded on the underground markets, and how they are later exploited. Furthermore, we investigate the structures, characteristics, and pricing mechanisms of these markets. We also aim to understand how this economy affects the victims. Finally, we critically examine the literature to understand potential shortcomings and suggest directions for future research.

The focus of this paper is academic research published after 2010. We allow a generous interpretation of stolen data and consider any information that has been illicitly acquired without the permission of the rightful owner or information that can be used to steal from the victim or hijack their digital assets, accounts or identities. Similarly, we allow a generous interpretation of the markets and underground economy to capture every dimension covered by research.

Our summarised findings are:

- We find positive trend in terms of **research ethics**, with the majority of more recent research at least partially addressing ethical issues.
- We find changing trends over time with regard to the online platforms housing the marketplaces. Earlier research heavily features **forums**, which are later overtaken in popularity by **paste sites** and short-lived **shops**.
- Earlier economy favoured stolen **credit and debit cards**, while more recently, **credentials** and stolen **personal information** are more prevalent.
- The underground economy landscape is quickly changing. We find a steady **decrease in market lifespans** as well as the length of **observation windows** featured in studies.

- We find that research based on marketplace snapshots covers only a small portion of the full **lifetime** of the average marketplace, threatening the generalisability of research conclusions.
- We find very little diversity in terms of the **languages** of marketplaces featured in research.
- We find a mismatch in stolen data types appearing on markets and the data stolen in data breaches.

The remainder of the paper is structured as follows. In §2 we first provide a short glossary of the specialised terminology. In §3 we outline the scope and present our methodology. We then examine the state of the research field, focusing on the characteristics of research that has been done since 2010 (§4). Next, we define stolen data and zoom in on its properties (§5.1). We then sketch out the general pipeline through which the stolen data passes, from stealing (§5.2), selling (§5.3) and exploiting the data (§5.4), highlighting important observations made by recent research. In §6 we outline the aftermath and consequences of the stolen data market. Finally, in §7 we critically examine the recent research, investigate how it reflects the state of the economy and highlight some under-explored directions in §8.

## 2. Background and terminology

We intentionally refrain from explicitly defining *stolen data* and instead allow the scope to emerge from the reviewed literature and, subsequently, from the economy itself. Broadly, when we talk about stolen data, we refer to any information or digital assets that were acquired illicitly and without the consent of the data subject or data controller. We provide a more detailed overview of the actual stolen data categories appearing on underground markets in §5.1. Similarly, we avoid narrowly defining *underground stolen data markets*, instead using the term broadly to refer to any online venue where the trade of stolen data happens, regardless of the specific format.

A portion of the underground economy operates on the *Tor anonymity network*, also known as the dark web or dark net.[1] This stands in contrast to the "regular" internet, also referred to as *surface web* or clear web. The Tor anonymity network consists of unindexed websites that utilise a more private onion routing system and are accessible only through specialised software. It should not be confused with the deep web, which is also unindexed, but accessible using "normal" browsers and internet protocols. We emphasise that although illicit activities do occur on the Tor anonymity network —among others— they do not define it, nor do they represent the majority of its traffic [4].

Regardless of whether they operate on the surface web or through the Tor anonymity network, three main website types are commonly used for trading:

- **forums** facilitate discussion and trade by structuring users' posts into threads, and threads into topical boards
- **shops** allow sellers to create listings for their goods and advertise in a more structured way
- **paste sites** allow users to share plain text content, often in a simple, unformatted way

## 3. Methodology

Our goal is to present a comprehensive snapshot of the research on underground markets for stolen data and identify the underlying trends and patterns related to the markets and the data sold on them. We utilise an inductive approach, inspired by grounded theory [5], allowing the observations to emerge from the set of annotated research papers.

We begin by collecting papers related to stolen data markets from Google Scholar. We perform a search for keywords related to underground (illicit, underground, dark, black, illegal, dark net, dark web, Tor anonymity network) marketplaces (market, economy) and stolen data (stolen data, personal information, personal data, data breach). We consider research that appears on first 10 pages on Google Scholar[2] and is published after 2010. We exclude any research on underground markets that may contain some stolen data, but primarily focus on other services or goods (e.g. drugs, malware, violence). Finally, we apply the snowball method and collect additional works that are referenced by papers in the original set and that satisfy inclusion criteria. We identify 65 papers that satisfy our conditions.

We further process a subset of papers that use a *data-driven* method to examine marketplace(s) intended for dissemination of stolen data, excluding more theoretical works and works that closely focus on a single aspect of the problem (e.g. malware used for stealing, novel methods for data collection). The final set includes 31 recent works. To identify the emerging patterns and characteristics, we annotate this final set of papers. Specifically, we aim to capture the characteristics of the studies themselves, the marketplaces analysed, and the stolen data offered on the marketplaces. Table 1 contains the code book of main groups and their descriptions. Table 2 contains the list of analysed studies and their coded characteristics, in chronological order.

## 4. The state of the research field

### 4.1. An overview

We first look at the main research directions in the field. While the common thread is, naturally, the stolen data economy, we find several different research focuses, reflecting the interdisciplinary nature of the problem and combining perspectives from computer science, economics, criminology, law ,and psychology.

[1] More recently, efforts are being made to replace such charged terms with more considerate language.

[2] The searches were performed between December 2023 and February 2024.

| Code Group | Code | Description |
|---|---|---|
| Data collection | Automated | Studied data was collected using crawling and scraping |
| | Manual | Studied data was collected manually |
| | Leak | Studied data comes from leaked databases of marketplaces |
| | Honeypot | Studied data was intentionally leaked to marketplaces by researchers |
| Analysis | Quantitative | Analysis using primarily numerical, statistical or machine learning methods |
| | Qualitative | Analysis using descriptive and content-based methods |
| Data collected | Posts | Posts and threads from online forums or paste sites |
| | Listings | Listings and advertisements from shops |
| | Stolen data | The actual stolen data being traded on marketplaces |
| | Messages | Messages from instant messaging platforms |
| | Leaked data | Leaked database of a marketplace; may contain posts, listings, private messages and trade records |
| Ethics | Addressed | Ethical concerns addressed, the research has been approved by ethics review board |
| | Partly addressed | Some ethical concerns addressed, but ethics review is not explicitly mentioned |
| | Not addressed | No mention of any ethical concerns or steps taken to address them |
| Own data collection | | Data collected directly by the researchers involved in the study |
| Stolen data access | | Researchers have direct access to stolen data, as opposed to just listings or posts *about* it |
| Multiple markets | | The dataset collected and studied includes multiple different marketplaces |
| Observation period | | Refers to the time period for which the marketplace data is collected and analysed |

(a) Study characteristics

| | | |
|---|---|---|
| Transparency | Surface web | Studied data comes from surface web platforms and websites |
| | Tor network | Studied data comes from platforms and websites requiring Tor |
| Marketplace format | Forum | Studied marketplace is a forum |
| | Paste site | Studied marketplace is a paste site |
| | Shop | Studied marketplace is a dedicated online shop |
| | Messaging | Studied marketplace is a messaging platform |
| Language | EN | The main language of the marketplace is English |
| | RU | The main language of the marketplace is Russian |
| | DE | The main language of the marketplace is German |
| Stolen data type | Bank cards | Stolen credit and debit card information |
| | Credentials | Stolen credentials for online accounts |
| | PI | Stolen personal or private information |
| | Other | Stolen data that is none of the above |

(b) Marketplace and data characteristics

TABLE 1: Codebook

First, we have research contributing tools and techniques for more effective monitoring of the landscape, data collection and identification of stolen data. Researchers in this domain work on methods and tools for timely identification of relevant marketplaces and stolen data within them [15], [27], as well as crawling and scraping techniques that evade the detection of marketplaces and can adapt to the constantly changing and often hostile landscape [36], [37], [38]. While many make primarily methodological contributions, some instead focus on sharing the raw datasets, enabling future research to those without the means to collect own data [39].

Building on that is research that presents a high level, quantitative snapshot of one or multiple marketplaces [25], [3], [33]. Such works provide insights into the number of marketplaces, their offerings, prices and membership numbers by reporting various statistics or using social network analysis [8], [30], econometric analysis, and profit estimations [14]. The raw data is often messy and valuable information may be obscured and difficult to extract. The researchers therefore also turn to various machine learning and natural language processing techniques to more easily identify notable marketplaces or stolen data, automatically generate reports about them [40], and cluster or classify stolen data types, discussion topics and interactions between offenders [36], [41], [22].

Another strand of research focuses on understanding the functioning of underground stolen data markets on a deeper level, their organisation, trust mechanisms, socio-cultural characteristics and economic forces that drive demand and pricing [6], [8], [10], [20], [42], [12]. Here, research relies primarily on qualitative techniques and borrows heavily from sociology and criminology. Methods used in research include criminal event perspective [25], panel design and consumer research [34], crime script analysis [9], [13], [35], and interviews [30]. Closely related are also case studies, which tend to provide a more comprehensive picture of an incident, starting with the data breach and reasons for it, the dissemination of the stolen data, and the associated costs to

| **Study** | | | | | | **Data collected** | | | | | **Data collection** | | | | **Analysis** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ethics | Own data collected | Multiple markets | Stolen data access | Posts | Stolen data | Listings | Marketplace leak | Messages | Manual | Automated | Leak | Honeypot | Qualitative | Quantitative |
| Vomel et al. [6] | 2010 | ○ | ✓ | | | | | | | ● | | ● | | ● | ● | ● |
| Holt and Lampke [7] | 2010 | ○ | ✓ | ✓ | | ● | | | | | ● | | | | ● | |
| Motoyama et al. [8] | 2011 | ○ | | ✓ | | | | | ● | | | | ● | | | ● |
| Soudijn and Zegers [9] | 2012 | ○ | | | | | | | ● | | | | ● | | ● | |
| Yip et al. [10] | 2013 | ○ | | ✓ | | | | | ● | | | | ● | | ● | ● |
| Décary-Hétu and Laferrière [11] | 2015 | ◐ | | | | | | | ● | | | | ● | | | ● |
| Allodi et al. [12] | 2015 | ○ | ✓ | ✓ | | | | | ● | | | | ● | | ● | ● |
| Hutchings et al. [13] | 2015 | ○ | ✓ | ✓ | | ● | | | | | ● | | | | ● | |
| Holt et al. [14] | 2016 | ◐ | ✓ | ✓ | | ● | | | | | ● | | | | ● | ● |
| Butler et al. [15] | 2016 | ○ | ✓ | ✓ | ✓ | | ● | | | | | ● | | | | ● |
| Décary-Hétu and Laferrière [16] | 2016 | ○ | ✓ | | | ● | | | | | | ● | | | | ● |
| Lazarov et al. [17] | 2016 | ● | ✓ | | ✓ | | ● | | | | | | | ● | | ● |
| Onaolapo et al. [18] | 2016 | ● | ✓ | ✓ | ✓ | | ● | | | | | | | ● | ● | ● |
| Haslebacher et al. [19] | 2017 | ● | ✓ | ✓ | | ● | | | | | | ● | | | ● | ● |
| Dupont et al. [20] | 2017 | ◐ | | | | | | | ● | | | | ● | | ● | |
| Smirnova and Holt [21] | 2017 | ◐ | ✓ | ✓ | | ● | | ● | | | ● | | | | ● | ● |
| Thomas et al. [22] | 2017 | ◐ | ✓ | ✓ | ✓ | | ● | | | | | ● | | | ● | |
| Bernard-Jones et al. [23] | 2018 | ● | ✓ | ✓ | ✓ | | ● | | | | | | | ● | | ● |
| Onaolapo et al. [24] | 2019 | ● | ✓ | ✓ | ✓ | | ● | | | | | | | ● | | ● |
| Madarie et al. [25] | 2019 | ○ | ✓ | ✓ | | ● | | ● | ● | | | ● | ● | | ● | |
| Steel [3] | 2019 | ○ | ✓ | ✓ | | | | ● | | | ● | | | | | ● |
| Campobasso and Allodi [26] | 2020 | ◐ | ✓ | | | | | ● | | | | ● | | | | ● |
| Liu et al. [27] | 2020 | ○ | ✓ | ✓ | ✓ | ● | | ● | | | | ● | | | | ● |
| Aliapoulios et al. [28] | 2021 | ● | | | ✓ | | | | ● | | | | ● | | | ● |
| Onaolapo et al. [29] | 2021 | ● | ✓ | ✓ | ✓ | | ● | | | | | | | ● | | ● |
| Ouellet et al. [30] | 2022 | ○ | ✓ | ✓ | | | | ● | | | | ● | | | | ● |
| Howell et al. [31] | 2023 | ○ | ✓ | ✓ | | | | ● | | | | ● | | | | ● |
| Campobasso and Allodi [32] | 2023 | ● | ✓ | | | | | ● | | | | ● | | | | ● |
| Georgoulias et al. [33] | 2023 | ◐ | ✓ | ✓ | | | | ● | | | | ● | | | | ● |
| Madarie et al. [34] | 2023 | ● | ✓ | | | | | ● | | | | ● | | | ● | |
| Garkava et al. [35] | 2024 | ● | ✓ | ✓ | | | | | | ● | ● | | | | ● | |

TABLE 2: Studies on stolen data markets and their characteristics

**Ethics**: not addressed (○), partly addressed (◐) or addressed (●).

the original data controller and the affected individuals [43], [44], [45]. Theoretical approaches include rational choice theory [21] and signalling theory [46]. Findings from such research are particularly valuable when devising interventions and disruptions [47], [30], [11].

Finding what happens to the stolen data after it has been sold is very hard due to the very nature of the problem. Attempts to understand this have been made by a cluster of research that utilises honeypots to monitor the attackers' interactions with intentionally leaked data. In such studies, researchers create artificial, but realistic-looking data, which they then intentionally leak on one or more marketplace(s) and then monitor offenders' interactions with it [23], [24], [29], [17], [18]. Additionally, researchers also turn directly to the victims and use interviews and surveys [48] to understand their experiences and the harms caused.

### 4.2. Research ethics and legal issues

Conducting research that involves stolen data or interactions with criminal content and offenders comes with a set of practical, ethical and legal challenges. Any research relying on data collection from underground markets is highly unlikely to involve obtaining informed consent from participants. Therefore it is important to consider and minimise harms and contrast them against the possible benefits [49]. It is also important to minimise the researchers' own contributions to the economy by not paying for the data or driving profits or attention to the offenders in some indirect way. On the other hand, researchers should also minimise any adverse effects their data collection may have on the marketplace or underlying networks. For example, researchers should consider if their data collection puts a significant strain on the Tor network and act accordingly [50]. Finally, when interacting with stolen data directly, such data should be sufficiently anonymised, analysed on a group level, and no attempts should be made to identify any individual. The same is true also when research is based on a leaked marketplace database and researchers might have access to offenders' private information, conversations or transactions.

There is no clear universal consensus on which ethical standards research on underground forums should meet.

Many US universities, for example, do not require any ethical approval for data research, while many European institutions do. Of the 31 studies examined, 10 explicitly mention approval from an institutional review board, ethics committee or similar, and a further seven at least partly address some ethical concerns, but do not explicitly mention any ethical review process. There is a noticeable trend towards explicitly addressing ethical questions, with the more recent research at least partly addressing ethical concerns. In general, researchers mention that they refrain from interacting with the attackers directly or making any purchases. Some researchers also refrain from publicising the name of the forum of marketplace, to avoid driving further traffic towards it or causing any retaliatory action, although this may come at the cost of replicability. Worth mentioning are the efforts in protecting offenders' identities as well. Very few researchers have access to the stolen data, typically citing ethics as the main reason.

In addition to ethics, researchers might also be concerned about the legality of their research, in particular in relation to scraping or possessing potentially illegal content. Scraping is often against the terms of service of most websites or platforms and informed consent cannot be gained from all members. Regional differences further complicate the issue. For example, under the British Society of Criminology's Ethics Statement [51], informed consent may not be required for research into online communities where the data is publicly available, and the research outputs focus on collective rather than individual behaviour. However, scraping remains a grey area with US court ruling that scraping of publicly accessible data does not violate the Computer Fraud and Abuse Act [52], but then also reversing the decision and ruling that scraping is in breach of the user agreement [53].

A mix of ethical and legal challenges contributes to many researchers not openly sharing their datasets or even revealing the markets investigated. Researchers and their institutions need to consider the many implications and potential liability that comes with sharing sensitive and potentially criminal data. For example, research institutions need to consider what will happen if law enforcement demands a copy of the data containing offenders' personal information, when sharing would clash with the ethical principles of research. Additionally, the differences in legal frameworks and ethical constraints across countries and research institutions add to the challenge of openly sharing the datasets with the broader community. Many researchers simply do not have the resources to tackle such problems, instead preferring to delete the data after the study or only sharing minimal information about the community or marketplace in question.

Despite the challenges, some research groups do make (parts of) their datasets or related artefacts available. For example, Cambridge Cybercrime Centre[3] specifically collects cybercrime data and has a legal framework in place to allow data sharing with researchers upon request and after signing a data sharing agreement [39]. There are also platforms set up to streamline the sharing of cybercrime-related datasets, such as ImpactCyberTrust[4], while some researchers make their data available directly, upon request [54].

***Takeaway 1.*** Ethical concerns are becoming increasingly important and are being actively addressed by researchers.

***Takeaway 2.*** There is no clear consensus on how to collect, handle and share research data from a legal standpoint.

## 4.3. Data collected and analysed

The types and sizes of the datasets analysed by researchers vary greatly, from as little as five spreadsheets, to billions of records. The majority of studies include some form of own data collection. Researchers typically do not have direct access to the stolen data traded on the markets. The few notable exceptions include honeypots, where researchers intentionally leak credentials or documents on various marketplaces and observe interactions with the leaked data [17], [18], [23], [24], [29]. In some cases, a part or whole marketplace database is leaked or stolen by a competing marketplace and publicly released. Such datasets typically include forum discussions or shop listings, but often also private conversations, messages and sometimes trading information. Such datasets, however, do not normally contain actual stolen data. While they are technically stolen, we treat them separately from the rest of stolen data. We find eight studies that investigate such leaked marketplace databases. Data on marketplaces analysed by researchers have been collected from both marketplaces on surface web (24) and also Tor anonymity network (13). The larger datasets disproportionately originate from the surface web research.

When the data analysed by researchers does not come from honeypots (6) or a marketplace leak (8), the researchers instead collect the data directly from the marketplace. The collection of data for research is typically automated using crawling and a custom scraper (13). Alternatively, some researchers opt for manual data collection (6). In such cases the dataset is typically smaller (under or around 1,000 data points) and often analysed qualitatively, with the aim of identifying more nuanced actions, motivations and consequences. The same is true for research using honeypot data. On the other hand, studies based on leaked marketplace databases or automatically collected datasets vary greatly in size, reaching millions or even billions of data points. Overall, ten of the datasets used in reviewed studies contain fewer than 1,000 data points, ten contain 1,000–100k data points, four contain 100k–1M data points and six contain more than that.

Figure 1 shows the observation periods reported by the studies examined (we exclude the three studies where collection dates were not explicitly reported). In most cases, the observation period refers to the time span covered by the collected data, which usually, but not always, matches

[3] https://www.cambridgecybercrime.uk/process.html

[4] https://www.impactcybertrust.org/

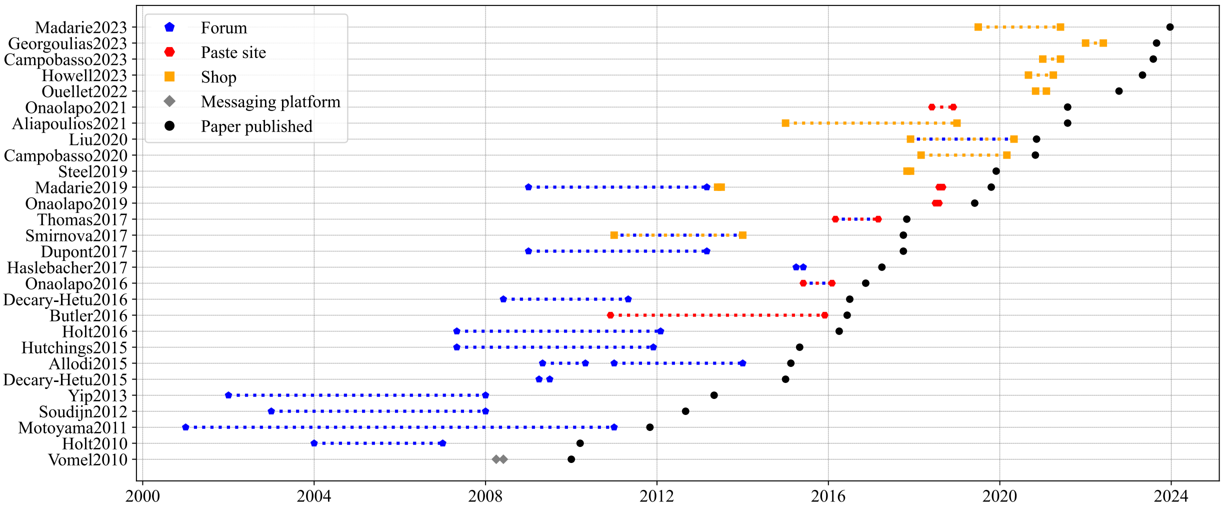


Figure 1: Observation periods reported in the studies and the time paper was published. Forums are marked in blue, paste sites in red, shops in orange and messaging platforms in grey.

the time during which the data was gathered. In some rare cases, the data was collected over a period that is shorter than the time span of the final dataset. For example, forum posts typically remain online as long as the forum is online so researchers are able to collect posts from several months or years ago. On the other hand, shop listings may only be online for a short time until the goods are sold and as such researchers are only able to collect the listings available at the time. The distinction is not always clear, so we chose to group them and treat them as equal. We see that the duration periods vary greatly. The average observation period reported by the studies is 27 months, while the median is 12 months.

We caution the reader that not all observation periods are equal; some studies focus on a *single* marketplace and carefully monitor it over a shorter period, while others investigate multiple marketplaces over several years, often only capturing incremental snapshots of the marketplace. Focusing only on the observation periods reported, regardless of the number and size of marketplace(s), we find that the earlier studies report longer observation periods compared to the more recent ones. Formally, observation start time and observation period are negatively correlated (Pearson correlation, $r = -0.59, p < 0.001, N = 31$).

It is not clear if the shorter observation periods are the result of researcher choices or if they reflect changes in markets over time. We therefore further look into the differences in observation periods across several dimensions. First, we compare observation periods across different data collection methods. We find that research using leaked datasets cover, on average, the longest periods (41 months), followed by manual data collection (37 months). We attribute the large temporal coverage of manually collected data to the fact that such research is often performed on a smaller set (often under or around 1,000 data points) of "cherry-picked" forum posts. Next, we have automated data collection (16 months), and finally, honeypots (4 months).

Next, we turn to different platform characteristics, where we also find notable differences. Research on surface web platforms on average covers 32 months, while research on Tor anonymity network covers on average 15 months. Additionally, we find that observation periods are the longest when forums are researched (36 months), followed by paste sites (18 months) and shops (17 months). Finally, we find that observation periods are the longest when carding is involved (35 months), followed by credentials (24 months), with personal data being the shortest (16 months).

While the patterns are clear, we are unable to make causal claims. Additionally, without the ground truth for different marketplaces' lifetimes, any comparisons related to observation period are not very informative on their own. Therefore, we provide further context by looking at full marketplace lifetimes in §7.

The state of the underground economy is highly volatile. Frequent law enforcement action means that few (prominent) markets survive even a few years. In fact, researchers sometimes report that they were unable to replicate very recent studies due to the takedown of the marketplace in question, with several researchers mentioning that the marketplace investigated was shut down before the study was even released. Our data confirms the claim - of the 56 marketplaces named in the examined literature, only 14 appeared more than once. We highlight the importance of clearly reporting the marketplace(s) observed along with the observation period in the rapidly changing landscape.

***Takeaway 3.*** Surface web is more extensively studied than Tor.

***Takeaway 4.*** There is large heterogeneity in terms of the sizes of the snapshot and observation periods of the studied marketplaces.

***Takeaway 5.*** The observation periods are longer in: (i) earlier studies, (ii) studies that rely on leaked market databases and manually collected data compared to automatically collected data and honeypots, (iii) surface web compared to Tor, (iv) forums compared to paste sites and shops, (v) carding compared to credentials and personal data.

## 5. The stolen data economy

We now turn towards the economy itself and outline the stolen data pipeline, highlighting notable research insights related to each stage. We begin by first providing insights into the types of stolen data that are traded (§5.1). We then follow the typical journey of stolen data from the rightful owner (§5.2) towards the marketplace (§5.3) and beyond (§5.4). The pipeline begins when the data is obtained by the threat actors. Sometimes, the data will be profiled and examined by the original attacker before it is passed on. The data will then appear on the marketplace, where it will be sold, traded or sometimes released as a sample or for free. Naturally, some data will be used by the original attacker directly, without changing hands through a marketplace, but such cases are out of scope for this study. In the final stage, the buyer exploits the data, typically for financial gain. At this stage, the victim is most likely to observe the consequences. The stages broadly mirror the (more recent) specialisations of actors: acquisition, sales, and monetisation of stolen data, as observed and defined in the literature [16], [3].

### 5.1. Stolen data

We find that the majority of studies do not explicitly define stolen data, allowing multiple interpretations. The broad categories of data most commonly mentioned include **credentials** (18), **credit/debit card information** (18) and datasets or lists containing miscellaneous **personal or business data** (9).[5] We find a slight shift in patterns over time (see Figure 2). In the early 2010s the research focus was on *carding* -referring to the cloning and unauthorised use of credit cards-, while more recent studies investigate a much more diverse set of stolen data, focused around credentials and personal information. Partly, the shift could be explained by the rise of online banking and other financial services, which expand the attack surface for financially motivated attackers beyond carding. We further look into each of the broader categories.

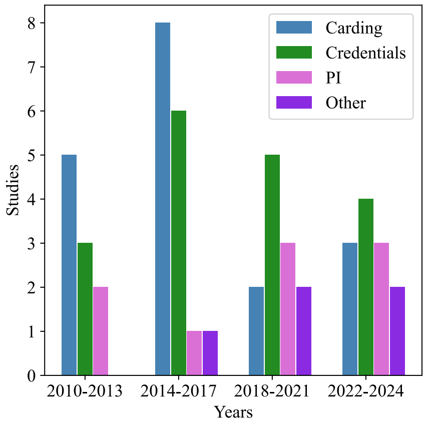


Figure 2: Data groups included in research over time

**5.1.1. Credentials.** We find a lot of variety among stolen credential. Particularly attractive are credentials for financial accounts such as bank accounts, Paypal, and–recently–cryptocurrency wallets or exchanges. In addition, many studies report the popularity of credentials for video and music streaming services (e.g. Netflix, Spotify), gaming services (e.g. Steam), pornographic content, dating apps, VPN services and app stores. Also popular are credentials for email accounts and social media, which can serve as a gateway into other accounts or give an attacker access to the victim's personal information and contacts–new potential victims. Credentials are often offered in *combolists*, collections of (typically) thousands of credentials, usually in a `username:password` format.

As part of the recent Impersonation-as-a-Service offerings, credentials can be accompanied by a complete behavioural and digital fingerprint (e.g., the victim's session cookies, form fill, browser history and other data stored by the browser, environment characteristics, etc.), making up the so-called *infostealer log*. The logs allow the attacker to recreate the victim's environment and fly under the radar of Risk-Based Authentication (RBA) mechanisms [26].

**5.1.2. Carding.** Carders are primarily interested in two products: the so-called *dumps* containing magnetic stripe information of the card and *CVV/CVV2* containing the information printed on the card along with the security code (CVV) printed on the back. CVV2 information is typically used for online purchases through card-not-present (CNP) transactions, which makes it less risky and more desirable [28].

On the border between carding and identity theft lie the so-called *fullz*, the holy grail of stolen data. Fullz often contain the aforementioned financial information sought after by carders, but also additional personal information such as date of birth, social security number, address and identification documents. The amount of information in fullz allows an attacker to impersonate a victim or even steal their identity.

[5] We assign the category based on the *dominant* data type discussed in the study. Multiple data types can be covered by a single study.

#### 5.1.3. Personal and business information.

The final larger category includes hacked databases or lists that typically contain personal information or sometimes adjacent business information. Such databases typically come from breaches of various data controllers (e.g., commercial companies, organisations, schools, hospitals). Such datasets might contain personal information such as customer/employee email addresses, personal document scans, sensitive medical information and drug prescriptions. Threat actors are also interested in non-personal information, such as bank records and other business records.

***Takeaway 6.*** The most frequently studied stolen data types are credentials, carding and personal or business information. We see a shift in popularity from primarily carding to all data types.

### 5.2. Obtaining stolen data

The majority of stolen data originates from malware, phishing, data breaches, or card skimming. While the scope and the economy around stolen data may have evolved over time, the means of obtaining it have not fundamentally changed; attackers have been relying on the same phishing techniques and malware kits since the mid-2000s [22]. In particular, easy access to ready-made data stealing tools also available on the underground markets has lowered the barrier to entry, but also led to a specialisation of roles throughout the ecosystem [55].

Phishing typically involves indiscriminately and opportunistically spreading scam messages with the hope that victims will click on the included link and enter their credentials or other private information on the fake website. A rarer, but more dangerous form of phishing is the so-called spear phishing or manual hijacking. Here, the attackers spend a significant amount of effort to profile the victims and approach them with personalised and emotion provoking stories to raise the odds of a successful attack and maximise the profits [45].

Malware can take many forms, from keyloggers that steal credentials from infected personal machines [22], to mass infections on companies' servers, allowing almost unlimited access to company databases. The latter, often alongside targeted efforts towards a specific data controller, may lead to a larger data breach. More sophisticated malware allows the attacker to not only extract credentials, but also impersonate users by obtaining a detailed fingerprint of their environment and behaviour [26], [56]. In more recent years, we observe sharing of data on underground markets as part of so-called double or triple extortion [57], a recent ransomware adaptation that not only threatens to encrypt the victim's data, but also leak it online, if the ransom is not paid. The risk is therefore no longer just the loss of availability, but also the confidentiality of data.

Finally, attackers interested in obtaining credit or debit card *magnetic stripe* information, may do so using card skimmers physically installed at ATMs, PoS terminals and gas pumps [28].

***Takeaway 7.*** Stolen data is primarily obtained through malware, phishing, data controller breaches and card skimming. The methods have not fundamentally changed in the past 15 years.

### 5.3. Stolen data markets

We now turn our attention to the marketplaces enabling the stolen data economy. We investigate the general characteristics of the platforms on which the markets reside, how they operate and finally, the pricing of stolen data. Table 3 shows the main properties of the marketplaces, as reported by the studies analysed. Note that several studies investigate more than one marketplace and the properties may refer to a group of different marketplaces.

#### 5.3.1. Platforms and their characteristics.

Earlier research sometimes references **IRC** (Internet Relay Chat), but it is only researched by a single study published after 2010. Instead, it is evident that stolen data trade in the early 2010s primarily takes place on **forums**. Yip et. al. [10] and Allodi et. al. [12] argue that forum-like marketplaces help alleviate the issue of asymmetric information that IRC users faced by providing more transparency, better feedback mechanisms, and reputation tracking. Overall, forums are examined in 16 of the studies. These forums initially operate on the surface web (16), sometimes even allowing participation without registration. However by the late 2010s forums on Tor anonymity network (4) gain popularity, alongside frequent law enforcement action and takedowns. Forums residing on the Tor anonymity network tend to be more secretive, requiring registration and sometimes a vouch from an existing member.

In the late 2010s, forums largely give way to **paste sites** (8) and later dedicated **shops** (11), split roughly equally between surface web and Tor anonymity network. Most recent works observe a move away from forums in favour of dedicated shops and encrypted messaging services, in particular **Telegram**, which seem to be more agile and host smaller communities compared to previously observed platforms. This may be partly in response to the common takedowns and arrests made by law enforcement and partly due to the (perceived) privacy such messaging services provide [59], [60]. However, the move away from the more structured, forum-like design might reintroduce the information asymmetry and trust issues that the markets of the 2010s helped alleviate. We are also seeing adaptations by vendors who are active across multiple platforms to secure their market in case one platform or channel goes down [33]. More research is necessary to investigate the magnitude of the movement back to messaging platforms and the consequences on the stolen data markets.

Not all platforms are equal in terms of their data offerings. According to the examined research, cards and personal information are primarily sold on forums (13 and 3, respectively) and shops (6 and 3, respectively), with no documented cases of either appearing–in significant numbers–on paste sites. On the other hand, credentials appear across

TABLE 3: Marketplaces and data for sale

| Study | | Marketplace format | | | | | Language(s) | Stolen data type |
|---|---|---|---|---|---|---|---|---|
| | | Transparency | Forum | Paste site | Shop | Messaging | | |
| Vomel et al. [6] | 2010 | - | | | | ● | EN | carding, credentials, PI |
| Holt and Lampke [7] | 2010 | ○ | ● | | | | EN | carding, credentials, PI |
| Motoyama et al. [8] | 2011 | ○ | ● | | | | EN, DE | carding, credentials |
| Soudijn and Zegers [9] | 2012 | ○ | ● | | | | EN | carding |
| Yip et al. [10] | 2013 | ○ | ● | | | | EN | carding |
| Décary-Hétu and Laferrière [11] | 2015 | ○ | ● | | | | EN | carding |
| Allodi et al. [12] | 2015 | ○ | ● | | | | EN, RU, DE | carding, credentials |
| Hutchings et al. [13] | 2015 | ○ | ● | | | | EN, RU | carding |
| Holt et al. [14] | 2016 | ○ | ● | | | | EN, RU | carding |
| Butler et al. [15] | 2016 | ○● | | ● | | | EN, RU | credentials |
| Décary-Hétu and Laferrière [16] | 2016 | ○ | ● | | | | EN, RU | carding |
| Lazarov et al. [17] | 2016 | ○ | | ● | | | EN | documents |
| Onaolapo et al. [18] | 2016 | ○ | ● | ● | | | EN, RU | credentials |
| Haslebacher et al. [19] | 2017 | ○● | ● | | | | EN, RU, DE | carding, credentials |
| Dupont et al. [20] | 2017 | ○ | ● | | | | EN | carding, PI, credentials |
| Smirnova and Holt [21] | 2017 | ○● | ● | | ● | | EN, RU | carding |
| Thomas et al. [22] | 2017 | ○ | ● | ● | | | EN | credentials |
| Bernard-Jones et al. [23] | 2018 | ○● | | ● | | | EN | credentials |
| Onaolapo et al. [24] | 2019 | ○● | | ● | | | EN, RU | documents |
| Madarie et al. [25] | 2019 | ○● | ● | ● | ● | | EN | credentials |
| Steel [3] | 2019 | ● | | | ● | | EN | PI |
| Campobasso and Allodi [26] | 2020 | ○ | | | ● | | RU | credentials, digital fingerprint |
| Liu et al. [27] | 2020 | ○● | ● | | ● | | EN | carding, credentials, PI |
| Aliapoulios et al. [28] | 2021 | ○ | | | ● | | EN | carding |
| Onaolapo et al. [29] | 2021 | ○● | | ● | | | EN | credentials |
| Ouellet et al. [30] | 2022 | ● | | | ● | | EN | PI, carding |
| Howell et al. [31] | 2023 | ● | | | ● | | EN | PI, carding |
| Campobasso and Allodi [32] | 2023 | ○ | | | ● | | RU | credentials, digital fingerprint |
| Georgoulias et al. [58] | 2023 | ● | | | ● | | EN | carding, credentials, databases |
| Madarie et al. [34] | 2023 | ● | | | ● | | EN | credentials |
| Garkava et al. [35] | 2024 | - | | | | ● | EN, RU | credentials, PI, health |

Notes to columns:
**Transparency**: refers to whether the marketplace(s) appear on surface web (○) or Tor anonymity network (●). If the study does not explicitly specify, we assume surface web.
**Language**: English (EN), Russian (RU), or German (DE)
**Stolen data**: refers to the main data type(s) being traded on the marketplace or investigated closely by the study.

all examined platforms: forums (9), paste sites (6), shops (6) and messaging platforms (2). It is worth noting that even when the same general data type is involved, there may be significant differences in offerings depending on the platform. For example, Madarie et al. [25] compare credential offerings on a forum, shop and a paste site and find financial credentials primarily on paste sites and none on the forum, while credentials for entertainment, webshop, social and adult accounts primarily appear in the shop. While the study has a relatively small sample size, limiting generalisability, it hints at different specialisations of markets across different platforms. Finally, looking at the types of data for sale that various studies collected, we find that carding is disproportionately more common on surface web platforms, while the sale of personal information is disproportionately more common on the Tor anonymity network.

***Takeaway 8.*** The popularity of platforms involved in the studies changed over time - from forums, to paste sites, to shops.

**5.3.2. Market organisation and trust mechanisms.** Modern stolen data markets are highly organised, mirroring many structures, incentives and mechanisms of legal markets. Parallels have also been made with ordinary "office jobs", where the schedule follows a traditional work week and the employer provides the tools and training [45]. Knowledge about the economy is dispersed through forum posts, free or paid tutorials and even apprenticeships [13].

While it is clear that the economy is highly organised, there is mixed evidence on whether participants specialise in specific products or services (e.g. stealing, trading, monetising). While Soska and Christin [61] identified several specialisations of sellers on more general underground markets, Haslebacher et al. [19] did not find a significant specialisation on carding forums. Another piece of mixed evidence concerns the number of markets that participants are active on. While Haslebacher et al. [19] found little evidence that users are active on more than one carding forum at the same time, more recent research on Telegram marketplaces by

Garkava et al. [35] suggests the opposite.

Unsurprisingly, the scam and theft economy itself is also filled with scammers and fraudsters. Ablon et al. [1] suggest that on the lower tier markets, 30% of sellers can be classified as so-called *rippers*. 'Ripper' is used to describe a seller or buyer who does not fulfil their part of the deal, either by not providing the agreed upon goods, or by providing bad quality of merchandise. Nixon et al. [62] also report the existence of recycled or fake data where malicious actors announce a breach and provide fake data as evidence.

With little recourse and no official channel to turn to in case of a scam, participants in stolen data markets use a number of mechanisms to protect themselves. Many such mechanisms are provided (or required) by the marketplace itself. For example, marketplace might require a sponsor/vouch to introduce the new person, some initial payments or membership to participate, proof of technical skills and experience [9] and the use of marketplace-recognised contracts [20]. Additionally, successful markets often establish systems to track feedback, reputation and experience metrics [63], for sellers to have their data verified [13], resolve disputes [58] and offer escrow [7]–all enforced through a team of administrators and moderators [64], [65].

When a newcomer is entering a marketplace with many of such mechanisms in place, they might face the so-called *cold start problem*, the inability to start participating in the economy due to lack of trust by the community. In such cases, they may turn to individual mechanisms in order to prove themselves. Such mechanisms include the use of argot (a jargon associated with experience) [66], smaller [67] or more public [68] transactions or providing free samples and ample information for the data offered [46].

***Takeaway 9.*** The economy is highly organised and can mirror legal professional environments.

***Takeaway 10.*** Trust plays a crucial role in the economy. To gain trust, participants can turn to market-level mechanisms (reputation or experience metrics, feedback or rating, use of contracts, escrow, data verification, moderation, etc.) or individual-level mechanisms (use or argot, higher transparency, smaller transactions, free samples, etc.).

**5.3.3. Pricing.** Research points at several layers of the economy, from the more "entry level markets", where data that is easy to obtain, but difficult to monetise is readily available for cheap [42], to closed and specialised exclusive markets [26], [35]. Pricing is set accordingly. Certain marketplaces also turn to bidding and auctions instead of a fixed price, with bids starting as low as $100-$400 for hundreds of thousands of credentials [25].

Factors that affect the pricing include the data type, time since breach, the quality/richness of the dataset, and user characteristics [1], [33]. Specifically, richer countries or regions (e.g. USA and Europe) routinely fetch higher prices, as well as certain banks that are considered less protected and easier to exploit [7]. In the case of Impersonation-as-a-service, accounts with more features (e.g. geographic location, amount of stolen cookies, number of platforms that can be accessed) are also more attractive and therefore more expensive [26]. Higher credit score was also found to be positively correlated with the price, while larger markets tend to have lower prices [3]. Additionally, the ease of monetisation also affects the price of goods. Dumps and fullz, for example, were found to be more than three times as expensive as credit card numbers (CVVs), which require more effort to monetise [19].

Research on forums observed between 2004 and 2007 found the average price of a dump was $56, with prices ranging between $1 and $500 [7]. The same study also found the average price of a PayPal account to be around $13, with eBay accounts available for $1-$3 and a database containing millions of e-mail addresses for $400-$1,500. Similarly, in 2008, the prices for credit cards ranged between $2 and $8 per credit card, or $200 for a bundle of 30 cards [6]. Steel [3] finds that in 2017, the prices of *fullz* were as low as $1 for 250 identities, with a mean of $24 per identity. For the period 2015-2019, Aliapoulios et al. [28] recorded magnetic stripe prices between $0.21 and $260, with $14 median, and CVV2 prices between $0.93 and $49, with $13 median. In 2019, the mean asking price of $13.57 and the median of $6 were recorded for credentials [34]. Finally, listings for carding, hacked databases, and account credentials in 2022 ranged from $1.90 to $267, with an average of $15. While the data cannot be compared directly for the aforementioned reasons, there seems to be a slight downward trend in prices over time, likely driven by the large supply of stolen data.

***Takeaway 11.*** The prices vary significantly. Factors that affect the price include the type of data, volume, time since breach, ease of monetisation, quality of the dataset and user characteristics.

## 5.4. Exploitation and monetisation of stolen data

Exploitation usually happens in two main steps. First, the attacker may perform some form of profiling and exploration of the data. For example, when dealing with stolen accounts, they may look for financial data, linked account credentials and personal information that could be used for blackmail [45]. The exploration may be followed directly by exploitation, but in some cases the stolen data will first change hands through the market.

Next comes the exploitation. There is clear consensus among previous research that financial gain is the main motivator behind the trade and sale of stolen data. Sometimes reputation gain is also mentioned as a motivator, although one could argue that reputation is a necessary intermediate step towards eventual financial gain. While the main motivator –financial gain– is relatively straightforward, there is more variety in how criminals actually monetise the stolen data.

When dealing with stolen debit or credit cards and online banking accounts, the obvious goal is to transfer the victim's funds to the attacker, either directly or through some proxy. Before the widespread availability of online payment options, threat actors had to come up with creative tactics, such as betting and “losing into attacker’s pocket” on an online gambling site [69]. Alternatively, various *cashing out* and money laundering techniques have been and remain available for attackers to transform funds stolen from credit and debit cards or banking accounts into cryptocurrencies, PayPal, gift cards or some other format [6], [70]. Finally, some offenders use the credit cards to buy goods to resell, potentially with the help of a complicit merchant who can process fraudulent transactions [13].

Czeschik [71] suggests that sensitive (medical) information could be used for extortion, which is in line with the search terms performed on the stolen account, as identified by honeypot research. Additionally, stolen medical and healthcare information can be used by uninsured people to access health services, obtain prescription drugs and file fraudulent insurance claims. Finally, lists of emails or phone numbers are used for spam or phishing, with the aim of reaching new victims.

A valuable insight is offered by the research where artificial data was created and intentionally “leaked” on a number of marketplaces. Specifically, researchers leaked Google spreadsheets [17], [24], email credentials [18], [23], and Facebook accounts [29]. The findings are rather inconclusive and show that many criminals either access leaked accounts and documents without taking any actions or perform minor actions such as searching for information or defacing the documents. In particular, attackers search for attachments, banking information, payments, investments, cryptocurrencies, institutions and sensitive private information. They are also easily able to overcome language barriers when facing non-English information [23]. Apart from the searches, there is little clear evidence of systematic monetisation of the stolen data. This is potentially due to the limitations of the research, imposed due to ethical constraints: no real accounts were provided, the data was relatively isolated, and there was no real data available to directly monetise. Additionally, the supposedly stolen data was offered for free on relatively open marketplaces, which may have played a role in the self-selection of people interacting with it; as previously mentioned free data might attract “curious amateurs”, rather than professional cybercriminals.

Honeypot research also gives us a valuable insight into the timing of stolen credential exploitation. The timings of access range from minutes after the breach or release on the marketplace, to weeks. Research using social media honeypots found that half of recorded account accesses happened within first 25 days and 80% within a month [29]. Access patterns may differ by marketplace platform. For example, accounts released on a paste site, forum, and stolen by malware recorded 80%, 60% and 40% of accesses within first 25 days, respectively [18]. Research using spreadsheets recorded few accesses within the first 22 hours, followed by a sudden surge to 80% by the 25th hour [24]. On the other hand, research using Google’s proprietary data recorded 20% of access attempts only 30 minutes after the attack, 50% within seven hours, and 70% within one day [45]. While specific access patterns vary across platforms and data types, previous literature agrees that the majority of accessed happen under a month, showing strong preference for fresh data.

***Takeaway 12.*** The main documented motivation is financial gain.

***Takeaway 13.*** The freshness of the stolen data is important. The majority of accesses to stolen data happen within a month of it appearing on the market.

## 6. The aftermath and consequences

When talking about the consequences, two main groups need to be considered - the individual victims whose personal data was stolen and the data controller who lost the data. While the two might overlap, this is typically not the case.

When account credentials are concerned, the success of the attack may rely on the victim not realising their account has been compromised or them not being able to intervene. Bursztein et al. [45] find that, to retain the stolen account, the attackers will often lock the victim out by changing the password, delay account recovery by changing the secondary email, recovery phone number or secret questions, and redirect future communications from the victim to the attacker. However, doing so may also alert the victim. Therefore, the attackers try to cover up their tracks, for example, by setting up filtering or forwarding of specific traffic where possible.

Shay et al. [48] document victims’ experiences with account hijacking. They find that alongside more practical concerns such as having to change the password or dealing with the consequences of their email being used to propagate spam, the victims often suffer emotional harms as well. This is because the compromised accounts are often valuable to victims as they are used daily, and for personal communication.

The original data controller involved in a data breach faces additional consequences. They typically include fines by data protection authorities or lawsuits, loss of reputation and fall in stock prices. Additionally, the data controller may have to provide compensation or pay for credit screening services or equivalent [43].

***Takeaway 14.*** Costs and harms are both direct and indirect, and impacts are felt by the data controller and those whose data was stolen.

## 7. A critical perspective

The analysis of relevant literature highlights a number of distinct temporal patterns, such as the changing popularity of platforms and stolen data types over time. However, this raises the question: are the trends a faithful reflection of the true state of the underground economy or are they biased measurement artefacts, driven by research trends? We attempt to answer this question in the coming section. As we naturally do not have the ground truth, we instead rely on a number of proxies.

### 7.1. Marketplace languages

We first turn to the languages of marketplaces featured in research, where we find very little diversity. English is the most represented in research (29 of 31 examined papers), followed by Russian (12) and German (3). However, studies with access to some form of geographic attacker data often find that attacks originate from a very diverse set of locations. While real locations of the attackers can be partly obfuscated by the use of proxies or Tor, several countries stand out. Bursztein et al., for example find that the majority of attackers involved in manual hijacking of Google accounts come from only five countries: China, Ivory Coast, Malaysia, Nigeria, and South Africa [45]. Additionally, Lazarov et al. [17] find clicks on their honeypot data originate from 35 countries, and Onaolapo et al. [29] find clicks on honeypot Facebook accounts originate from 53 countries.

We highlight a distinct gap between the countries where attackers reportedly originate from and the forums analysed in majority of literature. It is not clear if the lack of diversity is caused by selection bias or if the majority of discussions and underground dealings indeed happen on English or Russian speaking marketplaces. However, just as certain services or platforms are mostly popular in one region (e.g. Baidu in China, VK in Russia), we would expect there exist more regional underground marketplaces. Additionally, previous research has shown that cybercrime can be locally embedded [72], [73]. Paired with the observations about the origin of attackers, we argue more effort should be invested into diversification of research beyond just English, Russian and German marketplaces.

***Takeaway 15.*** Research primarily features English and Russian-speaking marketplaces.

### 7.2. Marketplace coverage

We now direct our attention back to the temporal coverage of the economy. Earlier, we presented the observation periods reported by studies; we now contrast this against the full lifetime of the market(s) they investigate. To do this, we collected the start and end dates for any marketplace named in one or more of the examined studies. If marketplace was still online as of 2024-06-01, we consider that as the end date. When dates are explicitly reported by the studies or marketplaces themselves we use those. Otherwise, we consider Internet Archive's Wayback Machine [6], public domain registration information, research reports, or news articles. We exclude marketplaces where either start or end date could not be identified. While this is our best attempt, we acknowledge it may result in a somewhat biased picture. We are able to confidently identify the start and end dates for 46 marketplaces appearing across 15 studies. Of these, 29 marketplaces are no longer active, while 17 are–mostly more general forums and paste sites, which are also used for dissemination of stolen data, but that is not their main use case.

Figure 3 plots the full lifetimes for each of the available marketplaces in blue, along with the period during which the marketplace was observed by some study, in red. We previously reported the average and median observation periods *per study* - 27 months and 12 months respectively. When we instead look at observation periods *across the marketplaces investigated* (i.e. taking into account the fact that some studies investigate multiple marketplaces), we get an average of 12 months and a median of just above six. We contrast this against the average lifetime of a marketplace, which is significantly longer at 78 months, or around six and a half years.[7] We find that an average study covers only a very small portion of a market's lifetime - around 17%. This opens a question of how representative the observed snapshots are. In particular, we see that the majority of snapshots cover the middle of the marketplace's life, with some covering the beginning as well. Only few studies cover the full market's lifetime, including the final stages. Additionally, the sample is skewed towards more short-lived markets.

We also find that, just like the observation periods, the lifetimes of the marketplaces have been getting shorter. Formally, marketplace start dates and their lifetime are negatively correlated (Pearson correlation, $r = -0.38, p < 0.05, N = 29$).[8] We do not have a large enough set of different markets to formally check the relationship between marketplace lifetime and other marketplace characteristics such as platform type or stolen data offered. Future research should further investigate any such relationships.

***Takeaway 16.*** The studies only cover a small portion of the marketplaces' lifetimes, with fewer studies observing the final stages.

## 8. Open questions and future research

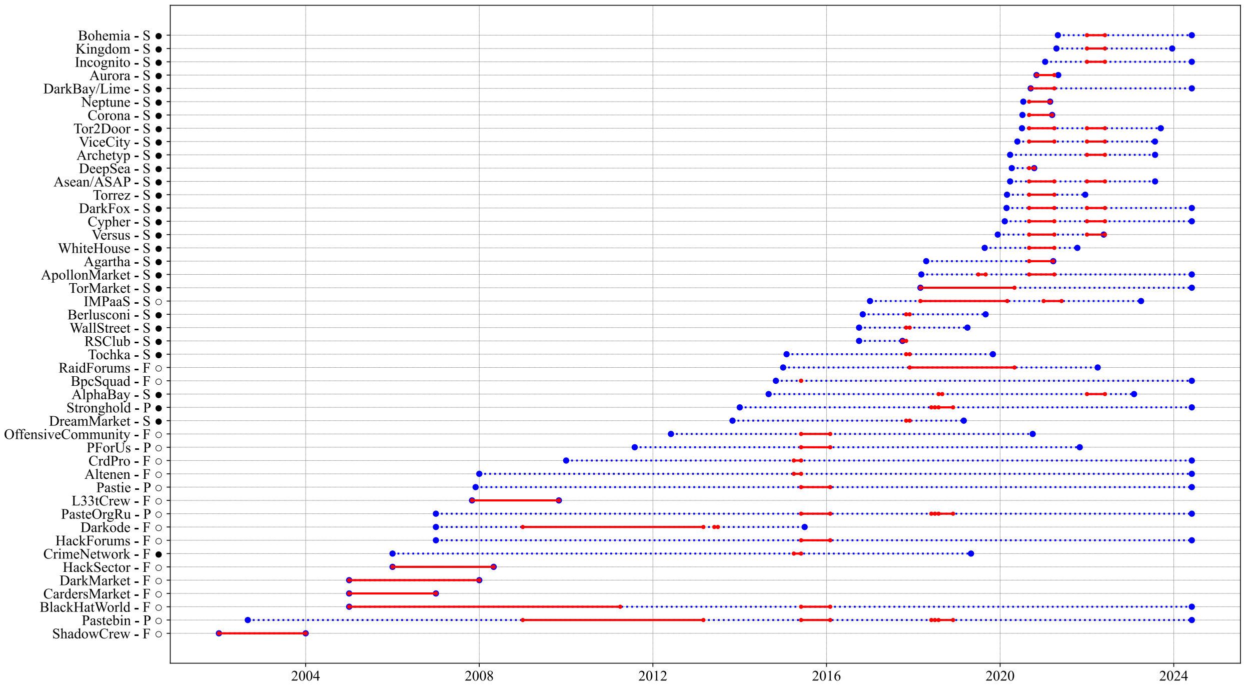
Given the rapidly evolving nature of the economy, future research should focus on timely identification and documentation of new trends, while paying attention to the potential biases introduced by the selection of observed marketplaces. Focusing on the actionable takeaways, we present a number of open questions and recommendations for future research.

Having discussed in length the types of stolen data that have been found on underground markets, we now turn towards the data that *has not*. We found no research documenting the presence of consumer behaviour, tracking data or data scraped from public sources. Such data falls into the grey area and might be valuable to legitimate businesses, competitors, advertisers or malicious actors alike. Additionally, we found no evidence for stolen data appearing on

[6] https://web.archive.org/

[7] The average duration includes marketplaces that are currently online. If we exclude those, the average market lifetime drops to 48 months, or around four years.

[8] We remove marketplaces that are still alive in our test; including them gives us an even stronger correlation (Pearson correlation, $r = -0.60, p < 0.001, N = 46$).



Marketplace format: shop (S), forum (F), paste site (P)
Transparency: accessible on surface web (○), accessible through Tor (●)

Figure 3: Full lifetime of a marketplace (blue line) observation period(s) (red line) reported by studies

marketplaces as a result of deliberate intellectual property theft, industrial espionage, and nation state actions. This suggests the existence of a different, more closed economy, where stolen data bypasses markets, which future research should look into.

Next, in terms of both the costs as well as potential harms to the victims, data breaches in healthcare are among the most problematic and costly [74], [75], [76]. However, we find surprisingly little research that explicitly documents any *widespread* presence of medical data on underground markets or its systematic monetisation [71].[9] That is despite a volume of academic literature documenting data breaches in the healthcare sector [77], [78]. Instead, the majority of research on the presence and value of medical data on underground markets comes from (often anecdotal) news reporting and private companies with potentially divergent incentives and motivations. It is not clear if this is an under-researched area or there actually is relatively low presence of medical data on underground forums. Future research should look into this and confirm or reject the presence of such mismatch.

Additionally, healthcare providers are often targeted with ransomware, where the primary goal is to receive a payment from the data controller directly and the threat of data appearing on the underground markets is used as a bargaining chip, rather than an end goal of itself. Another factor is that medical data might not be easy to monetise on mass scale, compared to credit card information, credentials and other personal data. More research is needed to understand stolen medical data, as well as the relationship between data breaches and data featured on underground markets in general. More generally, a question follows: are we assigning resources to the right problems? How else is data acquired through a breach used, if it does not appear on the underground marketplaces? Knowing that the offender motivations are of financial nature, we might want to consider which data can be monetised easily and prioritise our actions with that in mind.

This leads us to a broader potential issue which is related to studying the value and pricing of stolen data, in particular the size of the economy as the whole. Our literature search uncovered several references to various reports, blogs, and news articles by non-academic institutions and organisations (e.g. Ponemon Institute, Symantec, Trustwave, Identity Theft Resource Center), but relatively little *academic* research on the pricing. Such research, naturally, can be of a very high standard and should not be discounted, however the question of incentives remains relevant. A similar question motivated the seminal study on the cost of cybercrime by Anderson et al. [79] which looked at the ecosystem as a whole to surmise the claims made by the security industry

[9] We explicitly focus on data of medical nature and not other personal information that can also be obtained from breaching a healthcare data controller.

were overblown. More work is needed to understand the problem on a more nuanced level, especially when it comes to stolen data appearing on underground markets.

The main documented motivations for trade of stolen data are financial in nature. However, incidents fuelled by revenge, activism and other personal, political or ideological motivations tend to receive much public attention [80], [81], [82], [83]. More research is needed to understand how prevalent such incidents are and how they compare to the typical, financially motivated stolen data commerce.

We also document the quickly changing landscape, preventing detailed investigations of the same marketplace and reducing the replicability chances. However, we also find that some studies investigate the same marketplace. It is not clear how that effects the validity of the findings. In particular, we do not know whether the marketplaces are aware they are being observed and whether they change their behaviour in any way once they have been named in research.

We also uncover significant gaps in the coverage of marketplaces, in particular towards the end of a market's lifetime. Future research is necessary on the late stages of a marketplace to understand how markets collapse, exit scam or get shut down by law enforcement. Additionally, what are the signs, is any, that a market is approaching the end of its life? Future research should also try to better understand what factors affect how long a marketplace will thrive - are the long lasting markets doing something better or are they just not interesting enough for law enforcement to take action? Or, in case of exit scams, are there any signs that a market is approaching an exit scam? Answers to these questions can help us better allocate attention and resources to where they have the biggest impact.

Existing research is also heavily centred around English- and Russian-speaking marketplaces. While English is widely recognised as a global language, future studies should broaden their scope to include a more diverse range of languages, in particular the most widely spoken ones, like Mandarin, Hindi, Arabic, and Spanish. Doing so will help us better understand whether studying only a single languages gives a representative picture and, if not, give us a more complete view of the economy.

To have the most impact, researchers also need to stay on top of the changing trends and emerging patterns. We have seen several platforms rise and fall in popularity over the past two decades, with some evidence suggesting stolen data shops and encrypted messaging services such as Telegram are on the rise, while forums are falling. While we could not entirely disentangle the relationship between convenience sampling and real trends, we still believe the research captures real shifts of the market back towards messaging services.

Finally, we offer a broader commentary on the overall state of the field. Our review identified several trends, for which we could not establish causality. It is imperative to understand whether they are a reflection of the economy or simply research trends, driven by ease of access or the novelty of the research topics. While our study makes an attempt, we were unable to conclusively rule out the latter. With no standardised way of data collection or reporting, it is hard to quantitatively compare the studies or get a holistic view of the economy long term. With the goal of producing robust, insightful, interpretable findings that benefit the general public, we highlight the following good practices:

- consideration and reporting of ethical issues,
- timely data collection,
- longer observation periods,
- clear reporting of the collection process, including the observation period, time span of the data, granularity, completeness of the dataset, and marketplaces information, if possible,
- data sharing, if possible.

## 9. Conclusion

In this paper, we collected and organised the past 15 years of research on stolen data markets. We find a rapidly changing economy, both in terms of the data that is available as well as the platforms giving home to the marketplaces. We also highlight a number of potentially problematic research patterns such as the low coverage of the markets analysed and low diversity of marketplace languages. Finally, we suggest several directions for future research to better understand the true cost of the economy or why there is a mismatch between data breaches and data appearing on markets. Future research should also remain vigilant of the evolving landscape and focus on timely identification of new trends and community dynamics across various platforms.

## Acknowledgements

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant No 949127, and by an unrestricted gift from Meta. Tina Marjanov is supported by King's College, Cambridge and the Cambridge Trust.

## References

[1] L. Ablon, M. C. Libicki, and A. A. Golay, *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*. RAND Corporation, 2014. [Online]. Available: https://www.rand.org/pubs/research_reports/RR610.html

[2] T. J. Holt and A. M. Bossler, Eds., *The Palgrave Handbook of International Cybercrime and Cyberdeviance*. Springer International Publishing, 2020. [Online]. Available: https://doi.org/10.1007/978-3-319-78440-3

[3] C. M. Steel, "Stolen identity valuation and market evolution on the dark web," *International Journal of Cyber Criminology*, vol. 13, no. 1, pp. 70–83, 2019. [Online]. Available: https://www.cybercrimejournal.com/pdf/Steelvol13issue1IJCC2019.pdf

[4] B. Collier, *Tor: From the Dark Web to the Future of Privacy*. MIT Press, 2024. [Online]. Available: https://doi.org/10.7551/mitpress/14907.001.0001

[5] B. Glaser and A. Strauss, *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, 2017. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.4324/9780203793206/discovery-grounded-theory-barney-glaser-anselm-strauss

[6] S. Vomel, T. Holz, and F. C. Freiling, "“I'd like to pay with your Visa card”: An illustration of illicit online trading activity in the underground economy," University of Mannheim, Tech. Rep. TR-2010-004, 2010. [Online]. Available: https://madoc.bib.uni-mannheim.de/3048/1/underground2.pdf

[7] T. J. Holt and E. Lampke, "Exploring stolen data markets online: products and market forces," *Criminal Justice Studies*, vol. 23, no. 1, pp. 33–50, 2010. [Online]. Available: https://doi.org/10.1080/14786011003634415

[8] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the ACM Internet Measurement Conference*. ACM, 2011, pp. 71–80. [Online]. Available: https://doi.org/10.1145/2068816.2068824

[9] M. R. J. Soudijn and B. C. H. T. Zegers, "Cybercrime and virtual offender convergence settings," *Trends in Organized Crime*, vol. 15, no. 2-3, pp. 111–129, 2012. [Online]. Available: http://link.springer.com/10.1007/s12117-012-9159-z

[10] M. Yip, N. Shadbolt, and C. Webber, "Why forums?: an empirical analysis into the facilitating factors of carding forums," in *Proceedings of the Annual ACM Web Science Conference*. ACM, 2013, pp. 453–462. [Online]. Available: https://dl.acm.org/doi/10.1145/2464464.2464524

[11] D. Décary-Hétu and D. Laferrière, "Discrediting vendors in online criminal markets," in *Disrupting Criminal Networks: Network Analysis in Crime Prevention*. De Gruyter, 2015, pp. 129–152. [Online]. Available: https://doi.org/10.1515/9781626372573-009

[12] L. Allodi, M. Corradin, and F. Massacci, "Then and now: On the maturity of the cybercrime markets the lesson that black-hat marketeers learned," *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 1, pp. 35–46, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7044581/

[13] A. Hutchings and T. J. Holt, "A crime script analysis of the online stolen data market," *British Journal of Criminology*, vol. 55, no. 3, pp. 596–614, 2015. [Online]. Available: https://doi.org/10.1093/bjc/azu106

[14] T. J. Holt, O. Smirnova, and Y. T. Chua, "Exploring and estimating the revenues and profits of participants in stolen data markets," *Deviant Behavior*, vol. 37, no. 4, pp. 353–367, 2016. [Online]. Available: https://doi.org/10.1080/01639625.2015.1026766

[15] B. Butler, B. Wardman, and N. Pratt, "REAPER: an automated, scalable solution for mass credential harvesting and OSINT," in *Proceedings of the APWG Symposium on Electronic Crime Research*. IEEE, 2016, pp. 1–10. [Online]. Available: https://doi.org/10.1109/ECRIME.2016.7487944

[16] D. Décary-Hétu and A. Leppänen, "Criminals and signals: An assessment of criminal performance in the carding underworld," *Security Journal*, vol. 29, no. 3, pp. 442–460, 2016. [Online]. Available: http://link.springer.com/10.1057/sj.2013.39

[17] M. Lazarov, J. Onaolapo, and G. Stringhini, "Honey Sheets: What happens to leaked Google spreadsheets?" in *Workshop on Cyber Security Experimentation and Test*. USENIX Association, 2016, pp. 1–8. [Online]. Available: https://www.usenix.org/system/files/conference/cset16/cset16-paper-lazarov.pdf

[18] J. Onaolapo, E. Mariconti, and G. Stringhini, "What happens after you are pwnd: Understanding the use of leaked webmail credentials in the wild," in *Proceedings of the ACM Internet Measurement Conference*. ACM, 2016, pp. 65–79. [Online]. Available: https://doi.org/10.1145/2987443.2987475

[19] A. Haslebacher, J. Onaolapo, and G. Stringhini, "All your cards are belong to us: Understanding online carding forums," in *Proceedings of the APWG Symposium on Electronic Crime Research*. IEEE, 2017, pp. 41–51. [Online]. Available: https://doi.org/10.1109/ECRIME.2017.7945053

[20] B. Dupont, A.-M. Côté, J.-I. Boutin, and J. Fernandez, "Darkode: Recruitment patterns and transactional features of “the most dangerous cybercrime forum in the world”," *American Behavioral Scientist*, vol. 61, no. 11, pp. 1219–1243, 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0002764217734263

[21] O. Smirnova and T. J. Holt, "Examining the geographic distribution of victim nations in stolen data markets," *American Behavioral Scientist*, vol. 61, no. 11, pp. 1403–1426, 2017. [Online]. Available: https://doi.org/10.1177/00027642177342

[22] K. Thomas, F. Li, A. Zand, J. Barrett, J. Ranieri, L. Invernizzi, Y. Markov, O. Comanescu, V. Eranti, A. Moscicki, D. Margolis, V. Paxson, and E. Bursztein, "Data breaches, phishing, or malware?: Understanding the risks of stolen credentials," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1421–1434. [Online]. Available: https://doi.org/10.1145/3133956.3134067

[23] E. Bernard-Jones, J. Onaolapo, and G. Stringhini, "BABELTOWER: How language affects criminal activity in stolen webmail accounts," in *Companion Proceedings of the The Web Conference*. ACM, 2018, pp. 991–999. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3184558.3191529

[24] J. Onaolapo, M. Lazarov, and G. Stringhini, "Master of Sheets: A tale of compromised cloud documents," in *European Symposium on Security and Privacy Workshops*. IEEE, 2019, pp. 414–422. [Online]. Available: https://ieeexplore.ieee.org/document/8802418/

[25] R. Madarie, S. Ruiter, W. Steenbeek, and E. Kleemans, "Stolen account credentials: an empirical comparison of online dissemination on different platforms," *Journal of Crime and Justice*, vol. 42, no. 5, pp. 551–568, 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0735648X.2019.1692418

[26] M. Campobasso and L. Allodi, "Impersonation-as-a-Service: Characterizing the emerging criminal infrastructure for user impersonation at scale," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2020, pp. 1665–1680. [Online]. Available: https://doi.org/10.1145/3372297.341789

[27] Y. Liu, F. Y. Lin, Z. Ahmad-Post, M. Ebrahimi, N. Zhang, J. L. Hu, J. Xin, W. Li, and H. Chen, "Identifying, collecting, and monitoring personally identifiable information: From the dark web to the surface web," in *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2020, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ISI49825.2020.9280540

[28] M. Aliapoulios, C. Ballard, R. Bhalerao, T. Lauinger, and D. McCoy, "Swiped: Analyzing ground-truth data of a marketplace for stolen debit and credit cards," in *Proceedings of the USENIX Security Symposium*. USENIX Association, 2021, pp. 4151–4168. [Online]. Available: https://www.usenix.org/system/files/sec21-aliapoulios.pdf

[29] J. Onaolapo, N. Leontiadis, D. Magka, and G. Stringhini, "SocialHEISTing: Understanding stolen Facebook accounts," in *Proceedings of the USENIX Security Symposium*. USENIX Association, 2021, pp. 4115–4132. [Online]. Available: https://www.usenix.org/system/files/sec21-onaolapo.pdf

[30] M. Ouellet, D. Maimon, J. C. Howell, and Y. Wu, "The network of online stolen data markets: How vendor flows connect digital marketplaces," *The British Journal of Criminology*, vol. 62, no. 6, pp. 1518–1536, 2022. [Online]. Available: https://doi.org/10.1093/bjc/azab116

[31] C. J. Howell, T. Fisher, C. N. Muniz, D. Maimon, and Y. Rotzinger, "A depiction and classification of the stolen data market ecosystem and comprising darknet markets: A multidisciplinary approach," *Journal of Contemporary Criminal Justice*, vol. 39, no. 2, pp. 298–317, 2023. [Online]. Available: https://doi.org/10.1177/10439862231158000

[32] M. Campobasso and L. Allodi, “Know your cybercriminal: Evaluating attacker preferences by measuring profile sales on an active, leading criminal market for user impersonation at scale,” in *Proceedings of the USENIX Security Symposium*. USENIX Association, 2023, pp. 553–570. [Online]. Available: https://www.usenix.org/system/files/usenixsecurity23-campobasso.pdf

[33] D. Georgoulias, R. Yaben, and E. Vasilomanolakis, “Cheaper than you thought? a dive into the darkweb market of cyber-crime products,” in *Proceedings of the International Conference on Availability, Reliability and Security*. ACM, 2023, pp. 1–10. [Online]. Available: https://doi.org/10.1145/3600160.3605012

[34] R. Madarie, C. De Poot, and M. Weulen Kranenbarg, “Criminal clickbait: a panel data analysis on the attractiveness of online advertisements offering stolen data,” *Frontiers in Big Data*, vol. 6, pp. 1–15, 2023. [Online]. Available: https://doi.org/10.3389%2Ffdata.2023.1320569

[35] T. Garkava, A. Moneva, and E. R. Leukfeldt, “Stolen data markets on Telegram: a crime script analysis and situational crime prevention measures,” *Trends in Organized Crime*, pp. 1–25, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12117-024-09532-6

[36] M. Schafer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, “Blackwidow: Monitoring the dark web for cyber security information,” in *Proceedings of the International Conference on Cyber Conflict*. IEEE, 2019, pp. 1–21. [Online]. Available: https://doi.org/10.23919/CYCON.2019.8756845

[37] K. Turk, S. Pastrana, and B. Collier, “A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments,” in *European Symposium on Security and Privacy Workshops*. IEEE, 2020, pp. 428–437. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/EuroSPW51379.2020.00064

[38] G. Pantelis, P. Petrou, S. Karagiorgou, and D. Alexandrou, “On strengthening SMEs and MEs threat intelligence and awareness by identifying data breaches, stolen credentials and illegal activities on the dark web,” in *Proceedings of the International Conference on Availability, Reliability and Security*. ACM, 2021, pp. 1–7. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3465481.3469201

[39] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, “CrimeBB: Enabling cybercrime research on underground forums at scale,” in *Proceedings of the World Wide Web Conference*. ACM, 2018, pp. 1845–1854. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3178876.3186178

[40] R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, “Tools for automated analysis of cybercriminal markets,” in *Proceedings of the International Conference on World Wide Web*. ACM, 2017, pp. 657–666. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3038912.3052600

[41] U. Akyazi, M. J. G. Van Eeten, and G. C. Hernandez, “Measuring Cybercrime as a Service (CaaS) offerings in a cybercrime forum,” in *Workshop on the Economics of Information Security*, 2021, pp. 1–14. [Online]. Available: https://weis2021.econinfosec.org/wp-content/uploads/sites/9/2021/06/weis21-akyazi.pdf

[42] C. Herley and D. Florêncio, “Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy,” in *Economics of Information Security and Privacy*. Springer US, 2010, pp. 33–53. [Online]. Available: https://doi.org/10.1007/978-1-4419-6967-5_3

[43] F. Pigni, M. Bartosiak, G. Piccoli, and B. Ives, “Targeting Target with a 100 million dollar data breach,” *Journal of Information Technology Teaching Cases*, pp. 9–23, 2017. [Online]. Available: https://doi-org.ezp.lib.cam.ac.uk/10.1057/s41266-017-0028-0

[44] D. Kolevski, K. Michael, R. Abbas, and M. Freeman, “Cloud data breach disclosures: the consumer and their personally identifiable information (PII)?” in *Proceeding of the Conference on Norbert Wiener in the 21st Century*. IEEE, 2021, pp. 1–9. [Online]. Available: https://doi.org/10.1109/21CW48944.2021.9532579

[45] E. Bursztein, B. Benko, D. Margolis, T. Pietraszek, A. Archer, A. Aquino, A. Pitsillidis, and S. Savage, “Handcrafted fraud and extortion: Manual account hijacking in the wild,” in *Proceedings of the ACM Internet Measurement Conference*. ACM, 2014, pp. 347–358. [Online]. Available: https://dl.acm.org/doi/10.1145/2663716.2663749

[46] T. J. Holt, O. Smirnova, and A. Hutchings, “Examining signals of trust in criminal markets online,” *Journal of Cybersecurity*, pp. 137–145, 2016. [Online]. Available: https://doi.org/10.1093/cybsec/tyw007

[47] A. Hutchings and T. J. Holt, “The online stolen data market: disruption and intervention approaches,” *Global Crime*, pp. 11–30, 2017. [Online]. Available: https://doi.org/10.1080/17440572.2016.1197123

[48] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo, “”My religious aunt asked why i was trying to sell her Viagra”: Experiences with account hijacking,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 2657–2666. [Online]. Available: https://doi.org/10.1145/2556288.2557330

[49] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, “Ethical issues in research using datasets of illicit origin,” in *Proceedings of the ACM Internet Measurement Conference*. ACM, 2017, pp. 445–462. [Online]. Available: https://doi.org/10.1145/3131365.3131389

[50] J. Martin and N. Christin, “Ethics in cryptomarket research,” *International Journal of Drug Policy*, pp. 84–91, 2016. [Online]. Available: https://doi.org/10.1016/j.drugpo.2016.05.006

[51] British Society of Criminology, “Statement of ethics,” https://www.britsoccrim.org/ethics/, 2015.

[52] United States Court of Appeals for the Ninth Circuit, “*hiQ Labs, Inc. v. LinkedIn Corp.*” No. 938 F.3d 985, 2019.

[53] ——, “*hiQ Labs, Inc. v. LinkedIn Corp.*” No. 17-3301, 2022.

[54] Y. Jin, E. Jang, Y. Lee, S. Shin, and J.-W. Chung, “Shedding new light on the language of the dark web,” *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 5621–5637, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.412/

[55] B. Collier, R. Clayton, A. Hutchings, and D. R. Thomas, “Cybercrime is (often) boring: maintaining the infrastructure of cybercrime economies,” in *Workshop on the Economics of Information Security*, 2020, pp. 1–25. [Online]. Available: https://weis2016.econinfosec.org/wp-content/uploads/sites/8/2020/06/weis20-final10.pdf

[56] H. W. J. van Rijn, L. Allodi, M. Campobasso, and T. Ozcelebi, “In-depth analysis of AZORult infostealer malware capabilities,” Master’s thesis, Eindhoven University of Technology, 2021. [Online]. Available: https://pure.tue.nl/ws/portalfiles/portal/199812497/Rijn_H.pdf

[57] Q. Kerns, B. Payne, and T. Abegaz, “Double-extortion ransomware: A technical analysis of maze ransomware,” in *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 3*. Springer, 2022, pp. 82–94. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-89912-7_7

[58] D. Georgoulias, J. M. Pedersen, M. Falch, and E. Vasilomanolakis, “A qualitative mapping of darkweb marketplaces,” in *Proceedings of the APWG Symposium on Electronic Crime Research*. IEEE, 2021, pp. 1–15. [Online]. Available: https://doi.org/10.1109/eCrime54498.2021.9738766

[59] A. Akbari and R. Gabdulhakov, “Platform surveillance and resistance in Iran and Russia: The case of Telegram,” *Surveillance and Society*, vol. 17, no. 1/2, pp. 1–9, 2019. [Online]. Available: https://doi.org/10.24908/ss.v17i1/2.12928

[60] K. Boersma, “So long and thanks for all the (big) fish: Exploring cybercrime in Dutch Telegram groups,” Master’s thesis, University of Twente, 2023. [Online]. Available: https://essay.utwente.nl/96173/

[61] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *Proceedings of the USENIX Security Symposium.* USENIX Association, 2015, pp. 33–48. [Online]. Available: https://www.usenix.org/system/files/sec15-paper-soska-updated_v2.pdf

[62] A. Nixon, "Vetting leaks: Finding the truth when the adversary lies," Deloitte Development LLC, Tech. Rep., 2014. [Online]. Available: https://krebsonsecurity.com/wp-content/uploads/2014/10/vetting_leaks_final.pdf

[63] M. Yip, C. Webber, and N. Shadbolt, "Trust among cybercriminals? carding forums, uncertainty and implications for policing," *Policing and Society*, pp. 516–539, 2013. [Online]. Available: https://core.ac.uk/download/pdf/9646081.pdf

[64] J. Lusthaus, "Trust in the world of cybercrime," *Global crime*, pp. 71–94, 2012. [Online]. Available: https://doi.org/10.1080/17440572.2012.674183

[65] ——, "How organised is organised cybercrime?" *Global Crime*, pp. 52–60, 2013. [Online]. Available: https://doi.org/10.1080/17440572.2012.759508

[66] J. Hughes, A. Caines, and A. Hutchings, "Argot as a trust signal: Slang, jargon & reputation on a large cybercrime forum," in *Workshop on the Economics of Information Security*, 2023, pp. 1–11. [Online]. Available: https://weis2023.econinfosec.org/wp-content/uploads/sites/11/2023/06/weis23-hughes.pdf

[67] A. V. Vu, J. Hughes, I. Pete, B. Collier, Y. T. Chua, I. Shumailov, and A. Hutchings, "Turning up the dial: the evolution of a cybercrime market through SET-UP, STABLE, and COVID-19 eras," in *Proceedings of the ACM Internet Measurement Conference.* ACM, 2020, pp. 551–566. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3419394.3423636

[68] T. Marjanov, K. Ioannidis, T. Hyndman, N. Seyedzadeh, and A. Hutchings, "Breaking the ice: Using transparency to overcome the cold start problem in an underground market," in *Workshop on the Economics of Information Security*, 2024, pp. 1–12. [Online]. Available: https://www.repository.cam.ac.uk/items/a7c91f75-e0d0-4628-ac1a-823858e06824

[69] A. Shulman, "The underground credentials market," *Computer Fraud & Security*, pp. 5–8, 2010. [Online]. Available: https://doi.org/10.1016/S1361-3723(10)70022-1

[70] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy, "Mapping the underground: Supervised discovery of cybercrime supply chains," in *Proceedings of the APWG Symposium on Electronic Crime Research.* IEEE, 2019, pp. 1–16. [Online]. Available: https://doi.org/10.1109/eCrime47957.2019.9037582

[71] C. Czeschik, "Black market value of patient data," in *Digital Marketplaces Unleashed.* Springer, 2018, pp. 883–893. [Online]. Available: https://doi.org/10.1007/978-3-662-49275-8_78

[72] E. R. Leukfeldt, E. R. Kleemans, E. W. Kruisbergen, and R. A. Roks, "Criminal networks in a digitised world: On the nexus of borderless opportunities and local embeddedness," *Trends in Organized Crime*, vol. 22, pp. 324–345, 2019. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-3-319-78440-3_65

[73] E. Leukfeldt, E. Kruisbergen, E. Kleemans, and R. Roks, "Organized financial cybercrime: Criminal cooperation, logistic bottlenecks, and money flows," *The Palgrave handbook of international cybercrime and cyberdeviance*, pp. 961–980, 2020. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s12117-019-09366-7.pdf

[74] L. Cheng, F. Liu, and D. Yao, "Enterprise data breach: causes, challenges, prevention, and future directions," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 5, p. e1211, 2017. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/widm.1211

[75] D. Celeny, L. Maréchal, E. Rousselot, A. Mermoud, and M. Humbert, "Prioritizing investments in cybersecurity: Empirical evidence from an event study on the determinants of cyberattack costs," in *Workshop on the Economics of Information Security*, 2024, pp. 1–38. [Online]. Available: https://www.ssrn.com/abstract=4717020

[76] S. Wairimu and L. Fritsch, "Modelling privacy harms of compromised personal medical data-beyond data breach," in *Proceedings of the International Conference on Availability, Reliability and Security*, 2022, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3538969.3544462

[77] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar, and R. Ahmad Khan, "Healthcare data breaches: insights and implications," in *Healthcare*, vol. 8, no. 2. MDPI, 2020, p. 133. [Online]. Available: https://doi.org/10.3390%2Fhealthcare8020133

[78] V. Liu, M. A. Musen, and T. Chou, "Data Breaches of Protected Health Information in the United States," *JAMA*, vol. 313, no. 14, pp. 1471–1473, 2015. [Online]. Available: https://jamanetwork.com/journals/jama/articlepdf/2247135/jld150008.pdf

[79] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. J. G. Van Eeten, M. Levi, T. Moore, and S. Savage, "Measuring the cost of cybercrime," *The Economics of Information Security and Privacy*, pp. 265–300, 2013. [Online]. Available: https://link.springer.com/10.1007/978-3-642-39498-0_12

[80] A. Aravindan, "U.S. citizen leaks data on 14,200 people in Singapore with HIV," Reuters. [Online]. Available: https://www.reuters.com/article/idUSKCN1PM17T/

[81] K. Zetter, "Hackers finally post stolen Ashley Madison data," Wired. [Online]. Available: https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/

[82] L. Franceschi-Bicchierai, "A teen hacker is targeting Russian sites as revenge for the MH17 crash," Vice. [Online]. Available: https://www.vice.com/en/article/pgkp57/a-teen-hacker-is-targeting-russian-sites-as-revenge-for-the-mh17-crash/

[83] I. Steadman, "Anonymous hacks police site, releases list of South African whistleblowers," Wired. [Online]. Available: https://www.wired.com/story/south-africa-whistleblower-leak/