# Stayin' Alive: How Global Stolen Data Markets Thrive on Telegram

Tina Marjanov
*University of Cambridge*

Taro Tsuchiya
*Carnegie Mellon University*

Konstantinos Ioannidis
*University of Cambridge*

Jack Hughes
*University of Cambridge*

Nicolas Christin
*Carnegie Mellon University*

Alice Hutchings
*University of Cambridge*

## Abstract

Stolen data acts as a catalyst for many cybercriminal activities, such as spam campaigns, spear phishing, and identity theft. Studying online communities that serve stolen data helps combat those criminal activities. While anonymous marketplaces and forums have traditionally been the primary venue for stolen data, the chat-based messaging application Telegram has emerged as a popular alternative. Given Telegram's increased accessibility to the general public, it remains unclear how stolen data communities adapt their operations to this platform, circumvent moderation efforts, and create resilient communities. In this work, we characterize: i) where stolen data communities appear within Telegram's ecosystem, ii) what types of stolen data they offer, iii) where they operate from, and iv) how they evade detection. This paper offers four main contributions. First, we provide one of the largest longitudinal datasets of Telegram stolen data channels. Over one year, we manually curate 1,521 channels and collect 14 million messages and 3.6 million shared files. We show that the stolen data communities are largely disjoint from other communities on Telegram. Second, we categorize the types of stolen data with the aim of understanding the potential cybercrime they enable. Third, while existing literature focuses on English-speaking communities, we find that many channels operate in non-English languages and source stolen data from non-English markets. Fourth, those communities deploy various techniques to evade regulation. Notably, "gateway channels" that provide links to other stolen data channels play a crucial role in increasing longevity and growth rate. We conclude by providing implications not only for academic researchers but also for Telegram and law enforcement agencies across different jurisdictions seeking to monitor and moderate those activities.

## 1 Introduction

Over the last decade, anonymous online marketplaces and forums have been the primary venue for trading illegal goods, both tangible and intangible. Among those illicit goods, stolen data serves as a foundation for various cyberattacks, such as spear phishing, spam campaigns, account hijacking, and impersonation. Understanding how cybercriminals sell stolen data is critical for developing preemptive measures against those attacks before they happen.

There has been a gradual evolution of stolen data marketplaces over time [23]. Some are deployed over an anonymous network to conceal users/vendors' location and use PGP encryption to facilitate secure and private transactions and communication [13]. The use of cryptocurrencies and escrow payment services has also been observed [9, 16, 30]. While such technologies enhance the security and privacy of those marketplaces, they present a significant barrier to entry for many less technical users.

In recent years, the Telegram messaging application has emerged as an alternative platform for stolen data markets. Factors for Telegram's popularity may include (i) how easily it can be installed on a mobile phone or a laptop requiring only a phone number, (ii) resilience to law enforcement take-downs and quick resurgence compared to forums and marketplaces, and (iii) enhanced privacy compared to other messaging platforms. However, it remains unclear how cybercriminals operate on Telegram, given its accessibility and visibility to the general public (with 1 billion active users[1]).

This study investigates the stolen data communities on Telegram and provides the largest longitudinal dataset of Telegram stolen data markets to date, for future research. We further characterize what type of stolen data cybercriminals advertise/sell, the languages they use, whom they target, and how they bypass content moderation and maintain a resilient community. We address the following four research questions.
**RQ1**: *How can we identify channels involved in the distribution of stolen data?* We identify stolen data channels through keyword search, references from anonymous forums, as well as Telegram communities themselves. We perform manual filtering and channel selection to ensure desired scope and

[1] https://telegram.org/faq//#q-what-is-Telegram-what-do-i-do-here

eliminate false positives. We curate over 1,500 channels and collect 14 million posts and 3.6 million files shared on those channels, which we expect will spur future research on stolen data. We find that more than 99% of identified channels do not appear in existing general-purpose Telegram datasets.

**RQ2**: *What types of stolen data do we observe? What types of attacks are possible?* We code channels based on the data type, and scale up annotation with an LLM (Large Language Model). We find six categories of stolen data: payment cards, personal documents, account credentials, infostealer logs[2], breached databases, and personal information. A large portion of the channels are highly specialized, only focusing on one or two data types. Some combinations of data types appear more often, potentially enabling downstream cybercrimes (e.g., credentials & infostealer logs for account takeover, carding & personal information for stealing assets).

**RQ3**: *What languages are used in the stolen data channels, and from which regions do cybercriminals source data?* We find that many channels host non-English communities (Chinese, Spanish, Portuguese, Russian, Farsi), in contrast to existing literature that predominantly focuses on English communities [23]. Some channels procure data from a specific country. For instance, we find a disproportionately high concentration of carding data from Argentina, while the USA leads other data types.

**RQ4**: *Which factors contribute to longevity and growth of stolen data channels? Do Telegram policies matter?* We observe a multi-stage community structure: 155 channels work as a gateway to provide links to other private or new channels. These gateway channels appear to survive longer. Some channels also prepare a backup channel before the main dies, use misspellings or different encoding of certain words to avoid detection, and claim to comply with the law. Through statistical modeling, we find that channels linked to by gateway channels survive longer and grow faster; gateway channels can be a potential choke point for moderation efforts. We also find that regulation (i.e., Telegram announcing cooperation with law enforcement) inhibits longevity and growth. Based on our findings, we address the implications for researchers, threat intelligence, law enforcement, and Telegram.

The remainder of the paper is structured as follows. We introduce Telegram and previous research on stolen data markets in §2, explain our data collection methodology in §3, and the resulting dataset in §4. We next present the six main types of stolen data in §5, illustrate how they appear and are sold in individual channels in §6, and characterize each data type in §7. Finally, we investigate the factors associated with the channel lifetime in §8. We discuss the implications and limitations of our work in §9.

## 2 Background

### 2.1 Telegram

Telegram launched as a messaging platform in August 2013. With the introduction of social media features, bots, mini-apps, and the ability to share large media files, it quickly gained popularity as an alternative to more established platforms. Its perceived privacy and refusal to cooperate with law enforcement made it particularly popular among communities struggling to exist elsewhere. In September 2024, the platform announced a change in its policy, which includes providing certain user data to law enforcement. In January 2025, the first reports [10] of increased channel bans came out.

In the last few years, researchers have increasingly examined problematic content in Telegram, such as extremist content [33, 35, 38], propaganda [15, 19], conspiracies [17], price manipulations [25, 26, 37], and phishing [4, 8, 11]. Closely related, some have studied the communities involved in credential compromise, piracy, hacking, and related cybercriminal activities [5, 14, 29]. Importantly, the same features that enable harmful activity simultaneously attract millions of legitimate users circumventing censorship or mobilizing protests [32, 34, 36].

There are two main types of communities on Telegram: groups and channels. They mainly differ in size and the way their communication works. Channels are primarily meant for one-way broadcasts, whereas groups are more community-focused. For this research, we refer to them jointly as channels. Channels can be public or private. Public channels have a public username, are discoverable through Telegram's search field, and have a public invitation link (i.e., `t.me/username` or `@username`). Private channels have special, typically short-lived invitation links (i.e., `t.me/+a1b2c3d4e5f6g7h8`) and cannot be discovered through normal search; a new user can only join a channel using the invite link. A practical feature of Telegram is that a new user joining a channel can see all previous non-deleted messages and media.

### 2.2 Stolen data markets

From the opportunistic beginnings in the 1980s to the industrialized scale we see today, breached datasets, banking information, and personal data have always been valuable resources for malicious actors [2]. The landscape of stolen data is constantly evolving, with different types of stolen data-related cybercrime dominating at different times; the 2010s saw the dominance of carding and credential compromises, while more recently we see an increased availability of personal information and generally a more diverse landscape [23]. Methods for stealing data have evolved over time, for example, from simple username-password-based attacks [28] to much more powerful infostealers [7]. Various formats of stolen data markets have been documented on both the clear web and the

[2] Infostealer logs are a set of private information typically extracted by malware, such as user credentials, browser cookies, and machine metadata.

Tor network, including online forums, dedicated shops, and paste sites. However, stolen data offered in languages other than English and Russian are relatively understudied [23]. Reflecting this changing landscape, recent work has identified the growing presence of stolen data on Telegram [12, 29], motivating our research into this platform.

Closest to our research is work by Roy et al. [29], investigating the cybercriminal activity on Telegram, including but not limited to stolen credentials. In our work, we focus solely on stolen data, resulting in a manually curated sample, allowing us to address the problem of stolen data with more granularity. Additionally, our sample includes a diverse set of languages, addressing the previously identified lack of research on non-English markets.

# 3 Data collection

## 3.1 Channel selection criteria

We collect data at scale from Telegram marketplaces (i.e., channels and groups) *involved in the trade of stolen data*. Concretely, we only consider a channel if (i) it contains messages, files, or media with stolen data; (ii) it advertises the supply of stolen data; or (iii) its primary function is to link to a channel that does the aforementioned (i.e., gateways, see Section 6.2). We manually check whether each channel satisfies any of those conditions, and distinguish (iii) gateway channels for the later analyses. We collect data from channels from any region communicating in any language. There is no minimum size limit for the channels we collect.

To make the data collection more efficient and allow the timely collection of the most relevant content, a channel is excluded if: (i) it discusses stolen data, but does not contain, or provide a direct path towards obtaining it; ii) it primarily offers related tools, guides, or services (e.g., malware, infostealer software, tools for parsing and monetizing the data); iii), its function is coordination or attack planning; or iv) it is used for general discussions and trust building (e.g., reviews, vouches, scam reporting) as part of a group of interconnected channels that are included.

## 3.2 Sampling of Telegram channels

To ensure broad coverage and high data quality, timely discovery of new channels is essential. Publicly available lists of Telegram channels are generally insufficient, as they often miss smaller or emerging channels. We therefore identify new stolen data channels through a two-step process, repeated approximately every ten days. First, we search for candidate channels across diverse sources. Second, we apply snowball sampling by tracking channels referenced within previously collected messages.

### 3.2.1 Data sources

**Keyword search.** We compile a list of keywords related to various types of stolen data in 12 languages.[3] We create the initial set of keywords based on recent academic literature, industry sources (articles, reports, news), and terminology used on forums and known marketplaces. We further expanded the list of words when we encounter new terms in already collected channels. Native or proficient speakers familiar with the relevant context created the translations. We use the list of keywords to discover new channels in Telegram's search box and search bots, as well as to help determine channels' relevance during the snowballing process.

**Cybercriminal forums and markets.** We find candidate channels from cybercriminal forums or marketplaces that specialize in the trade of stolen data. There, the seller often directs potential buyers to their Telegram channel for more information, and the forum members can see lists of prominent channels. For this purpose, we use the continually updated CrimeBB dataset [27], available through data sharing agreements with the Cambridge Cybercrime Centre.[4] CrimeBB includes over 129 million posts collected from 40 English, Russian, Spanish, German, Arabic, and Vietnamese-language forums from 2002 to the present day.

**Other sources.** We consider candidate channels appearing in news, reports, previous research, public lists, repositories (GitHub, Reddit), and Telegram's recommendations. We do not explicitly search for these candidates; we consider them when we become aware of them. The occurrence of such sources is low and, as such, is not further categorized.

### 3.2.2 Snowball sampling

We perform recurrent snowballing to expand the set of channels. In each snowballing round, we collect candidate channels appearing in previously collected messages, channel descriptions, and the data sources described above. We also search for invite links, including from private channels, as well as channel and user mentions. When a message is forwarded, the original channel also becomes a candidate. This process gives us a set of candidate channels. We manually evaluate each candidate by examining its recent activity. If we determine that the channel fits the inclusion criteria, we observe the channel until it becomes unavailable.

In an average snowballing round, there are more candidates than can be evaluated manually. We therefore use three heuristics to ensure the best possible coverage while ensuring a manageable manual effort. First, we order the candidates based on the number of references from each source and always examine the top 20% most commonly mentioned candidates. Second, we examine channels that contain at least

[3] The list contains keywords in English, Chinese, Russian, Spanish, Arabic, French, German, Japanese, Italian, Greek, Estonian, and Dutch. The keyword list is available in Appendix A.1.

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

one of the keywords in their public username. Third, we randomly examine the remaining candidates until we find ten consecutive channels that we have already collected or channels that are irrelevant. After this process, we consider the set of channels saturated and conclude the snowballing. We skip channels that do not exist anymore, links that have expired, or lead to a user profile (rather than a channel).

Across all the snowballing periods, we identify 21k candidates and examine roughly half of them.[5] In total, we join over 1,500 channels that fit the inclusion criteria.

### 3.3 Collected data

We collect data about the channels and their users, as well as messages and files shared within the channels. As one of the primary goals of our work is to share the dataset with the broader scientific community, we present all collected information regardless of its use in our analysis.

**Channels.** For each channel, we collect its name (typically longer and more descriptive, e.g., Country Leaked Databases) and, if available, its username (a short, unique identifier, e.g., `country_leak`).[6] Additionally, we collect its profile picture, description, participant count, and latest pinned message; we also monitor and record any changes to these.

**Messages.** We collect all available messages exchanged on the channel. For each message, we record the author identifier, timestamp, content, number of views, and list of reactions. Additionally, we record whether the message is forwarded, and if so, how many times, and the original creator. Finally, we store the names of any files contained in the message.

**Files.** We retrieve metadata of any file present in a message, namely its name, format, size, and description. We also download a subset of files and images. Specifically, we collect JPG images and files in TXT, CSV, and SQL formats. To reduce the strain on the ecosystem and ensure timely data collection, we record metadata but do not download files larger than 100MB. We compute and store a SHA-512 hash of each downloaded file and store a single copy of a file corresponding to a given hash to avoid duplicate storage of identical files.

**Users.** We only collect minimal information about users to allow inference about user numbers and membership networks. We collect user identifier[7] and, where available, their username, first name, and last name.[8] For each observed user, we also record their channel membership.[9]

### 3.4 Data collection implementation

We collect data by accessing Telegram's official API using a custom-built Telethon data collection tool.[10] We store the data in a PostgreSQL database running on a secure server with an encrypted filesystem. We collect all available messages from a given channel, including messages posted before joining.

The data collection tool operates in two modes with different stop conditions. *Minimal mode* collects all *previously unseen* data in reverse chronological order. Upon encountering a previously collected message in a given channel, the data collection tool moves to the next channel until all channels are exhausted. The cycle repeats after a short period of rest. *Full mode* collects *all* available data and all available channels, including previously collected data. The cycle does not repeat on its own.

The regular data collection mode is set to a rate-limited *minimal* to minimize the impact our data collection has on the channels and Telegram infrastructure. Additionally, minimal mode allows more timely data collection, reducing the chance that the channel will be banned or messages removed before they are collected. We do a monthly *full* data collection to update statistics such as the number of views, forwards, and reactions.

## 4 Dataset overview

We collect data using a combination of regular *minimal* and periodic *full* data collection modes between August 2024 and August 2025. The collected messages were posted between January 2016 and August 2025. Due to the low frequency of older content, we only analyze content between the beginning of July 2021 and the end of July 2025. To allow future research, we make the dataset available to the research community through data sharing agreements (see Open Science statement in Section 10).

### 4.1 Channels and messages

The final dataset contains approximately 14 million messages across 1,521 channels, of which 1,073 (70%) are public and 448 (30%) are private. Those channels consist of 155 gateway channels and 1,366 non-gateway channels. We record approximately 252k distinct users. We also collect more than 316k channel state snapshots, capturing live changes in participant counts, channel descriptions, profile photos, and pinned messages. We start tracking channel states with participant counts in October 2024.

To recognize the language(s) of the collected messages, we use *langdetect*.[11] We observe seven languages present in

[5] A candidate corresponds to one invite link or linked channel. Channel owners can create multiple distinct invite links for the same channel, so the actual number of channels is likely lower.

[6] If a channel is private, we can only observe the name.

[7] User identifier is a unique number assigned to a user by Telegram and does not correspond to their phone number.

[8] First and last name are rarely available. Users predominantly use the field to provide more descriptive nicknames with emojis and special characters not allowed in usernames, rather than their real names.

[9] Most channels display the number of participants, but do not make the full list of members publicly available. In such cases, we can only infer a certain user's presence in a channel when they post in that channel.

[10] https://docs.telethon.dev/en/stable/

[11] Langdetect is a Python library that can run locally, downloaded from `https://pypi.org/project/langdetect/`. There are cloud-based language recognition tools, but we do not use them due to the sensitive nature

at least 140k messages (1% of overall messages): English, Chinese, Spanish, Portuguese, Russian, Farsi, and Indonesian. Other notable observed languages appearing less commonly include Arabic, Urdu, Italian, German, Dutch, Hindi, and Bengali.[12] We calculate the percentage of the seven most popular languages each week from July 2021 until July 2025. Langdetect could not recognize the language of 32% of messages because they are either too short or contained primarily special characters, emojis, numbers, or intentionally misspelled words.[13] Figure 1 illustrates the evolution of the languages used in our channels. The percentages are over the total number of messages per week. English usage is around 70% at the beginning of our observation period and plateaus around 50% later. This reflects the globalized nature of stolen data markets, with Chinese and Spanish being the next most frequently used languages in our dataset.

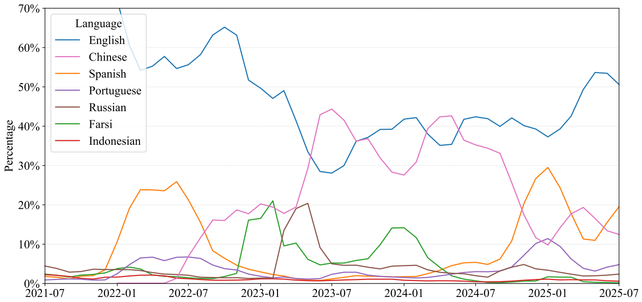

Figure 1: Languages in Telegram messages over time

## 4.2 File formats

Our dataset contains 3.6M media files: 1.97M images (55%), 680k text files (20%), and 950k other formats (25%). Public channels dominate overall volume, but file distribution differs: 22% of public-channel messages include a file compared to 61% in private channels. This result aligns with our expectation that channel owners are more likely to share stolen data privately.

We collected metadata (e.g., filename, size) for all attachments and successfully downloaded 1.34M images (68%) and 374k text files (55%). Text files account for only 8% of the total size as the distribution is skewed by a few very large files. Table 1 summarizes encountered and downloaded files by format. The following subsections detail each type.

**Images.** Of the 1.97M downloaded images, only 278k are unique. Most (189k, 68%) appear once, while 35k (13%) occur twice and 13k (5%) three times. The most frequent images are usually channel or group logos with little informational value. The single most repeated image appears 151k times.

| | Number of files | | Total file size | | Median file size | |
|---|---|---|---|---|---|---|
| **File format** | Total | Downloaded | Total | Downloaded | Total | Downloaded |
| Image | 1.97M | 1.34M | 230GB | 116GB | 71KB | 69KB |
| Text (TXT) | 603K | 314K | 58.5TB | 4.3TB | 1.1MB | 179KB |
| Text (CSV) | 77K | 61K | 2.8TB | 0.54TB | 527KB | 391KB |
| Text (SQL) | 17K | 11K | 3.5TB | 0.3TB | 30MB | 12MB |
| Other | 930K | – | 725TB | – | 29MB | – |
| **All** | 3.6M | 1.8M | 790TB | 5.3TB | 106KB | 77KB |

Table 1: Descriptive statistics of files by file format

**Text files.** Of the 680k downloaded text files, only 185k are unique. Most (142k, 77%) appear once, 20k (11%) appear twice, and 7k (4%) appear three times. Compared to images, there is less overall repetition (re-sharing) of text files and more unique files. A few text files recur frequently, with the most repeated one appearing 384 times. Without inspecting contents for ethical reasons, filenames and related messages suggest that these widely shared files typically contain method instructions or HTTPS/SOCKS proxies.

**Other media.** The most commonly encountered media types are images and plaintext files. The third and fifth most common media types are compressed files—RAR (218k or 6% of all media) and ZIP (187k, 5%), respectively. Filenames of compressed files most commonly refer to *logs* or *databases*; they typically contain the name or website of the breached entity. The fourth most common type of media is video files, which we do not download. According to filenames and accompanying messages, they contain instructions, proofs of breach or validity of data, displays of luxury, political statements, news, entertainment, memes, reactions, and displays of emotion.

## 4.3 Comparison with other Telegram datasets

We finish this section by comparing our dataset against three general purpose Telegram datasets to identify where stolen data channels exist in the broader Telegram ecosystem. Pushshift's [3] data collection began with right-wing and cryptocurrency-related channels and performed snowball sampling to attain a more general set of 29,661 channels by October 2019. TGDataset [21] contains 120,979 channels (until July 2022) and covers a wide range of topics, such as religion, news, business, and entertainment. As those two datasets may be outdated, we also compare against our newly updated third general-purpose dataset. This dataset was constructed through continuous snowball sampling of channel URLs over a three-month period (May 19th to August 18th, 2024). We have queried new messages from the channels already present in Pushshift and TGDataset, as well as newly discovered 67,868 channels. In total, the three datasets together comprise 208,652 distinct Telegram channels, which can be regarded as a publicly accessible "top-of-the-ocean"

of our data.

[12] Languages using non-Roman script are more difficult to detect with out-of-the-box automated language detection tools due to the common use of Roman script online, rather than the traditional scripts (e.g., Hinglish). This means we likely under-report those languages.

[13] We can identify at least one language for the vast majority of *channels* with only 40 (less than 1%) having no language detected at all.

sample of the platform.

We compare our stolen data channels (i.e., 715 stolen data channels that had existed before August 18th, 2024) against these three datasets. We find, based on channel IDs, that the overlap is only **six** stolen data channels, i.e., we can only identify 0.8% of the total stolen data channels through the (extended) publicly available dataset. Thus, stolen data channels may not be easily reachable through public links and are rather disjoint from other channels. This result justifies our approach of collecting stolen data channels through keyword searches and traditional forums/marketplaces.

## 5 Partitioning stolen data types

To partition the stolen data ecosystem and allow further analysis, we classify the stolen data type(s) each channel offers. The data labels were established by three researchers and were informed by literature, categories commonly appearing on known markets (i.e., discussion boards in cybercriminal forums), and manual inspection of the collected data. The codebook is presented in Table 2. We define six labels describing the types of stolen data present or discussed in the channels: *credentials*, *infostealer logs*, *carding*, *databases*, *personal information*, and *personal documents*.

### 5.1 LLM annotation

After manually establishing the codebook, we scale up the annotation of stolen data types to all channels over time. Due to the size of our dataset (14M messages), we classify a subset of messages from each channel. We (i) split each channel into one-week periods, (ii) randomly select one message per period, and (iii) merge the following 20 messages or up to 1,000 characters into a weekly-channel-string.[15] We experiment with different context windows[16], and manually inspect a sample of these strings and confirm that the selected thresholds adequately capture the message contexts.

We perform the classification using a Large Language Model (LLM) classifier. For ethical reasons, we use local models instead of cloud-based LLMs to avoid sending stolen and potentially sensitive data to third-party servers. Gemma 3 [31] offers the best balance between accuracy and speed among various locally run LLMs we have tested. We use Gemma 3 27B with GGUF Q4_0 quantization and a context size of 10k tokens. Due to resource constraints, we perform inference on CPUs rather than GPUs. Quantized Gemma 3 could feasibly run in several weeks, rather than larger models, which would have taken several months to run on our hardware.

We instruct Gemma 3 to identify whether any of the codes (i.e., data types) defined in the codebook are present in the sample. We allow up to three codes per sample, ordered by their prevalence in descending order. We do not translate the original messages as Gemma 3 supports multilingual inputs. The full prompt is available in Appendix A.2.

To evaluate the performance of the classifier, two human coders independently annotated a sample of 90 weekly-channel-strings. Human coders and the LLM used the same prompt and viewed the samples in their original language. However, human coders could translate the samples using Google Translate, after removing any sensitive data.

We measure inter-coder agreement using the Jaccard similarity index [22], which quantifies the proportion of shared labels between coders relative to the total distinct labels they assigned. This measure is appropriate for our multi-label setting, where each case could receive multiple labels, as it captures partial label overlap. The agreement rate between the two human coders was 0.89, and between all three coders (i.e., two humans and LLM, treated equally) was 0.77, which is considered high given that the chance agreement with seven labels (i.e., six data types and other) is low. We manually confirm that most disagreements originate in adjacent label categories (e.g., credentials and infostealer logs). We do not observe any difference in performance between messages originally in English and messages in other languages.

### 5.2 Classification results

Table 2 presents the stolen data types, their descriptions, and aggregated weekly occurrences within channels. Of the 56,417 labels assigned to 1,521 channels across 213 weeks, 51,545 belonged to one of the six data types and 4,872 were classified as *other*. Among the six data types, the most commonly occurring is *credentials* (22%), followed by its close relative *infostealer logs* (21%). Next are *carding* (19%) and *databases* (18%), followed by *personal information* (14%). The least common are *personal documents*, appearing in 4% of classifications.

Beyond aggregate statistics, we further examine how each data type evolves over time. Figure 2 displays the number of channels that contain each data type each week between July 2021 and the end of July 2025, after removing *other* label. The figure also shows the total number of channels observed each week in gray. The two vertical dashed lines indicate Telegram's announced policy changes (September 2024) and first news reports about the increased moderation (January 2025). First, we notice a sharp decline in the number of channels after Telegram changed its policy to collaborate

[14] While leads may come from legitimate sources and be used for legitimate purposes (e.g., an individual expresses interest in a specific service through a sign-up form), it is often implied by the sellers that the data was obtained without the consent of data subject by scraping public information, or from breached data. Additionally, it is often advertised alongside mass spamming tools (e.g., bulk SMS). We therefore consider leads found through Telegram in scope.

[15] Channels have on average 311 messages per week.

[16] We test a time windows between one day and one month, number of messages between 10 and 100, and string lengths between 1,000 and 5,000 characters.

Table 2: Stolen data types category definitions and frequencies

| Data type | Occurrences | Description | Associated keywords (EN) |
|---|---|---|---|
| Credentials (Cred) | 11,478 (22%) | Usernames and passwords or password hashes, sometimes together with the URL of the website, service or application to which the credentials belong. | username:password, email:password, username:hash, combo, login, ULP (URL:login:password), configs |
| Infostealer logs (Logs) | 10,912 (21%) | User credentials together with cookies, browser history, digital (behavioral) fingerprints, form fill information, system configuration, and any other information obtained through malware. | log, infostealer, cookies |
| Databases (DB) | 9,513 (18%) | Datasets or proprietary information stolen from companies, government organizations, educational institutions, etc. in a data breach. This includes non-personal documents such as trade secrets, designs, schemas etc.. | breach, leak, spy, secret, database, dump, names of breached companies |
| Personal information (PI) | 7,348 (14%) | People's personally identifiable information such as name, email address, residence, country, phones, interests, employment status, medical history, and more. This data can be compiled from breached databases or publicly available information. | leads[14], fullz, consumer info, user info (name, DOB, phone, address, email) |
| Carding (Card) | 9,988 (19%) | Any information required to make direct transactions and purchases, such as credit cards, debit cards. Includes access to online banking that allows direct cashout. | CVV, credit, debit, CC, BIN, NFC, Visa, MasterCard, dump, expiration, lists of numbers representing banking information, including the balance |
| Personal documents (Doc) | 2,306 (4%) | Photos or scans of official documents used for identification and travel of individuals. | passport, driving license, ID, KYC, selfie, verify |
| Other (Oth) | 4,872 | None of the above or not enough information. | - |

with law enforcement. We statistically examine the impact of this policy in Section 8.1. Second, we find the distribution of data types is relatively stable over time with a few notable patterns. We observe infostealer logs steadily rising from appearing in around 15% of offerings in 2021 to a stable 30% in 2022, with a further increasing trajectory in mid-2025. The rise in infostealer log offerings is juxtaposed by a decrease of a similar rate in credential offerings. We also record a rise in carding activity starting in late 2024, alongside a fall in database offerings, with a sharp drop around the same time.

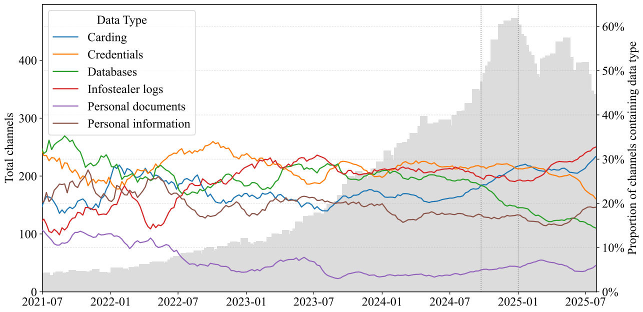

Figure 2: Data types in Telegram channels over time

## 6 Categorizing individual channels

In the previous section, we have looked at the overall trend of data types (i.e., *across* channels). Next, we are interested in a more granular view of data types *within* each channel. Specifically, we examine whether channels specialize in one or multiple data offerings. Additionally, we identify a set of channels that do not actually contain any stolen data, but rather link towards channels that do ("gateway channels").

### 6.1 Channel specializations

To investigate data types within channels, we compute the distribution of labels within each channel aggregated over time. A typical channel (median) sells three data types across its lifetime; the most frequent label is typically assigned to 67% of its activity, the second 20%, and the third 6%. We classify channels based on the frequency of up to two dominant data types as follows:

- **single-specialized**: one data type accounts for more than 67% of its activity
- **double-specialized**: the top two data types together account for more than 67% of activity, with each contributing at least 33%
- **non-specialized**: do not meet any of the criteria above

Our dataset contains 587 single-specialized channels, 228 double-specialized, and 467 non-specialized channels. Table 3 shows the co-occurrence of data types across specialized channels. We find certain data types typically appear in single-specialized channels, while others more commonly appear alongside others. Credentials and logs most commonly appear alone, appearing in 151 and 177 specialized channels, respectively. However, there is also a cluster of channels that double-specialize in credentials and infostealer logs (43), likely due to their shared goal of account takeover. On the other hand, databases and personal information often occur together (55), but more rarely on their own (40 and 19, respectively). With the line between databases and personal information somewhat blurry, this result is not particularly noteworthy. However, personal information also commonly appears alongside carding, which may imply its usage to aid in impersonating the victim when cashing out. The somewhat high co-occurrence of databases with credentials may also help explain the origin of some of the credentials, namely, breached datasets.

Table 3: Specialization across data types co-occurrences

| | **Cred** | **Logs** | **DB** | **PI** | **Card** | **Doc** | **Oth** |
|---|---|---|---|---|---|---|---|
| **Cred** | 151 | 43 | 21 | 8 | 10 | 1 | 8 |
| **Logs** | 43 | 177 | 8 | 1 | 6 | 0 | 4 |
| **DB** | 21 | 8 | 40 | 55 | 1 | 2 | 10 |
| **PI** | 8 | 1 | 55 | 19 | 29 | 3 | 2 |
| **Card** | 10 | 6 | 1 | 29 | 144 | 8 | 9 |
| **Doc** | 1 | 0 | 2 | 3 | 8 | 23 | 0 |
| **Oth** | 8 | 4 | 10 | 2 | 9 | 0 | 33 |
| Total | 242 | 239 | 137 | 117 | 207 | 37 | 66 |

## 6.2 Gateway channels

Both users and channels face a relatively low barrier to entry, allowing participants or communities to quickly rejoin Telegram even after being banned. However, this introduces the problem of continuity and trust. To maintain a stable presence, channel owners often set up *gateway* channels, also referred to as links, portals, adapters, redirects, or relay points. Gateways are separate channels created to provide links to private or new channels, or to share announcements. By remaining separate from channels containing stolen data, they are less vulnerable to bans, while serving as a stable point of reference and enabling the community to quickly migrate to a new channel if a ban does occur. Gateway channels typically contain a very low number of messages—sometimes a single message, which is occasionally edited to contain the latest link to the active stolen data channel. Sometimes, the channel description contains the link instead. We only consider a channel as a gateway when it links to a separate channel (i.e., not itself).

Upon data collection, we manually identify 155 gateways, of which 26% are private and 74% are public. This matches the wider dataset, where roughly 30% of channels are private and 70% are public. Notably, the links included in the gateways overwhelmingly lead to private channels (over 91% of all observed links). We have joined 145 channels linked in gateways; we refer to these channels as *linked channels*.[17]

# 7 How is stolen data obtained and used

After identifying channel specializations, we examine each data type in more detail (e.g., the language or data files that appear). To enable a cleaner comparison, we focus only on channels with single or double specializations. Table 4 shows the countries most frequently mentioned in file names across data types.

## 7.1 Credentials and infostealer logs

We look at credentials and infostealer logs jointly. While they likely differ in the means through which the data is stolen and what they offer to an attacker, they both serve a key goal: breaking into and taking over someone's account. We identify 151 specialized credential channels, 177 that specialize in infostealer logs, and 43 that specialize in both.

Participants in those channels communicate primarily in English. The channels commonly refer to themselves as "Clouds," e.g., "Aardvark Cloud"—out of 253 channels with cloud in their name, 165 specialize or co-specialize in credentials or infostealer logs. Such channels tend to be more professional and structured in terms of their operations, posting regular (daily or even hourly) updates on items in stock, which users can typically access through a subscription. Subscription prices range around \$40–\$70 for weekly access to the data sharing channel, \$120–\$250 for monthly access, \$250–\$500 for 90 days, and upwards of \$800 for a lifetime subscription. Some clouds offer much more expensive "premium" or "live traffic" options. Customers looking to purchase data from clouds can also do so in a highly automated way without direct interaction with the channel owners; much of the data is available through bots, which provide information, manage inventory or memberships, and accept payments.

Credential and log channels contained 1.7M messages and shared 807k media files, out of which 250k are images, 351k text files, and 206k other files, primarily compressed archives. Many of the filenames are uninformative, containing dates (presumably of when the data was obtained), channel names, or claims of freshness and validity. We extract the 50 most frequently appearing account domains, mentioned roughly 22.8k times, and classify them into broad groups according to their primary function. Among the top appearing domains, 28 (16k occurrences) belong to email providers, eight (2.3k) belong to social accounts, seven (2.6k) to gaming, three (1k) to financial services, and two (500 and 400, respectively) to streaming and shopping. While the USA appears most commonly, other top appearing countries are primarily European.

## 7.2 Carding

We identify 144 channels specializing and 63 co-specializing in carding, with 8.9M messages shared between members. Only 12% of the messages contained any media file, 88% of

[17] There are likely more linked channels in our dataset, but we either did not encounter their gateway, or could not confirm it was at some point linked through a gateway, e.g., due to expired links.

Table 4: Top countries of origin for different stolen data types.

| Accounts | Count | Carding - BINs | Count | Carding - text | Count | Databases | Count | PI & Docs | Count |
|---|---|---|---|---|---|---|---|---|---|
| USA | 16.8k | Argentina | 4.2M | Germany | 727k | USA | 3k | USA | 10.5k |
| France | 13.4k | USA | 1.1M | USA | 477k | China | 1.5k | China | 5.2k |
| Germany | 12.2k | Australia | 329k | Canada | 361k | India | 900 | France | 2.8k |
| Italy | 7.8k | Nigeria | 83k | Australia | 360k | Russia | 860 | India | 2.6k |
| EU | 5.7k | Canada | 56k | Taiwan | 254k | France | 820 | Germany | 1.9k |
| Poland | 5.5k | Spain | 56k | Hong Kong | 239k | Germany | 820 | Italy | 1.7k |
| UK | 5.2k | Mexico | 55k | Singapore | 225k | Japan | 710 | UK | 1.6k |
| Brazil | 4.6k | Great Britain | 43k | Japan | 224k | Bangladesh | 690 | Russia | 1.6k |
| Japan | 4.1k | Oman | 41k | Albania | 161k | Israel | 630 | Israel | 1.4k |
| Canada | 3.6k | Italy | 38k | United Kingdom | 142k | Indonesia | 520 | Japan | 1.4k |
| Total | 78.9k | Total | 6M | Total | 3.4M | Total | 10.5k | Total | 30.7k |

which were images. In carding channels, samples or "free" data are rare; instead, members commonly share screenshots of successful transactions or images of purchased goods. The languages most frequently identified in carding-related channels are English, Spanish, and Chinese, which appear with roughly equal frequency, followed by Portuguese. Sellers typically share lists of partial payment card information, such as available funds, card issuer, and the country. Card offerings often appear in similarly structured messages and might be accompanied by offers of personal information as well, presumably to increase the chance of a successful transaction. We do not report on the item price due to less structured messages, preventing us from reliably extracting the price.

Partial card listings often include card BINs (bank identification numbers). A BIN is a set of typically six digits, used to identify the financial institution issuing the card, and appears at the beginning of a payment card number. BINs also carry public information about the card brand (e.g., Visa, MasterCard), category (e.g., classic, premier), type (e.g., credit, debit), and the issuing bank's country.

We extract 6.5M BINs advertised in carding channels and use public records to extract public card information.[18] The majority of identified BINs belong to either Visa (67%), MasterCard (28%), American Express (1.8%), Maestro (1.3%), and Discover (1%), and include premier and platinum cards. 52% of cards are credit, 17% debit, with the remaining cards unspecified.

Finally, we look at the issuing bank's countries. We find the most common countries appearing in advertised BINs are Argentina (65.3%), the USA (17.1%), Australia (5.1%), Nicaragua (1.3%), and Canada (0.9%). The relatively high appearance of Middle and South American countries such as Argentina and Nicaragua (as well as Spain, Mexico, Ecuador, Colombia, Brazil, and Paraguay, all among the 20 most frequently appearing countries) points to Telegram being a popular platform for carding markets in certain regions, consistent with the higher relative count of Spanish and Portuguese-speaking carding channels.

Beyond BINs, we extract countries appearing in the text of messages. We find Germany, the USA, and Canada leading in appearances, followed by a cluster of countries from the Asia-Pacific region. Similar to previously identified clustering of Middle and South American data in Spanish-speaking channels, this shows localized activity in the proximity of Chinese-speaking regions.

Notably, Spanish-speaking channels tend to share BINs directly, while Chinese-speaking channels more often describe their offerings in text, suggesting a difference in preferred formats of sharing.

## 7.3 Datasets

We identify 40 channels specializing and 97 co-specializing in databases. The channels contain 1.1M messages, 87% of which include media files. Among these, 32% are photos, 25% are text files, and 43% are other types, primarily compressed archives. The majority of communication in database channels happens in English, followed by Spanish, Portuguese, and Farsi.

To better understand the offerings, we extract the most frequently appearing keywords and country names from file names in database channels. We find filenames often contain names of breached entities or their websites. We refrain from naming any datasets or entities to preserve their privacy and instead focus on broader characteristics. We manually examined top entries and confirmed they appear in previously documented breaches.

We identify frequently appearing descriptives referring to breached entities. Most commonly, they refer to businesses and corporations (e.g., corporate, B2B, business, industry), and particularly financial aspects (bank, crypto). We also emphasize the relatively high presence of *gov* and *edu* keywords, while we find a relatively low number of references to the healthcare and medical sector.

Additionally, filenames often reveal the country of origin. The most frequently mentioned countries include the USA

[18] Multiple card numbers can share the same BIN, so identical BINs do not necessarily indicate the same card. As a result, our reported count may include repeated card numbers. In total, we record 64.6k distinct BINs.

(3k mentions), China (1.5k), India (900), Russia (860), and France (820). Filenames often contain references to the size of the data (e.g., lines, MB, GB), composition (sample, full, mix), and contents (users, consumers, admin).

### 7.4 Personal information and documents

We evaluate personal information and documents jointly as they both target individuals. There are 3.1M messages exchanged within 42 single-specialized and 112 double-specialized channels. Within them, we measure 615k images, 132k text files, and 282k other files. Notably, media files appearing in specialized personal document channels are overwhelmingly images—selfies, photographs, or scans of official documents such as ID cards, passports, and driving licenses. While it is unclear how such images are obtained, discussions in the channels specializing in personal documents suggest images are intended for bypassing mandatory user verification processes, such as Know Your Customer (KYC) checks by financial institutions. Channels providing personal documents are overwhelmingly in English, whereas channels providing personal information are in Spanish first, followed by English, Portuguese, and Farsi.

We further examine frequent keywords appearing in filenames. We are particularly interested in which victim characteristics are advertised. We extract keywords appearing alongside frequently used terms referring to people such as “people”, “leads”, “users”, “consumers”, “information”, and “details”. Resulting matches often refer to countries, nationality, or citizenship (Indian, Chinese). Additionally, we find keywords referring to people’s gender (female), age (old), working status (worker, employee, student, education), financial status (e.g., rich, owners), and activities (casino, forex, investors, crypto, gambling, gaming). This implies personal information is used for targeted phishing or spamming of specific individuals according to their characteristics. The distribution of top-mentioned countries is similar to database channels; the USA leads in mentions (10.5k), followed by China (5.2k), France (2.8k), India (2.6k), and Germany (1.9k).

## 8 Factors contributing to longevity and growth

In this section, we aim to quantify the harms caused by Telegram’s stolen data channels. We are not analyzing the content of the files shared on the channels, so we cannot directly measure the harm caused. We assume that channels that persist over time and have many members have a larger impact potential *on average* compared to channels that are very short-lived or have very few members. Thus, we approximate their impact based on their lifespan (how long a channel remains active) and their size (the number of members in a channel). On average, channels are active for 35 weeks and have 2.6k members. However, the bottom 10% of channels were active for less than a month whereas the top 10% are active for roughly 1.5 years. Similarly, the bottom 10% of channels have at most 130 participants whereas the top 10% have more than 5k members. Given the substantial heterogeneity observed, we present both quantitative and qualitative analyses of the factors that affect longevity and membership of channels.

### 8.1 Quantitative analysis of channel lifespan and membership growth

We have 1,282 non-gateway channels that were classified as containing stolen data, out of which 276 (21%) remain active at the end of our observation period, and 1,006 (79%) are inactive/banned. We also have membership counts for 1,059 of the channels. Among the 276 channels that *remain active*, 210 (76%) have their highest membership recorded during the last observation week, suggesting that they continue to grow over time. 39 channels (14%) appear quite stable as their numbers decreased by less than 170 members, corresponding to, on average, less than 1% of their peak membership, and could be interpreted as ordinary temporal variation. 27 channels (10%) lost on average more than 2,000 members, corresponding to roughly one third of their peak membership. This pattern suggests that even without getting banned, some channels naturally die out over time, with three of those channels losing more than 90% of their members.

Among the remaining 783 channels that *became inactive* during our observation period, 592 (75%) had their highest membership during the last observation week. We believe those channels were growing, but were then suddenly banned. 112 (15%) were relatively stable and only lost on average 15 members from their peak, corresponding to a decrease of less than 1%. 78 (10%) channels were already on a downward trajectory when they went inactive, recording losses of more than 1,300 members, corresponding to more than 25% of their membership. The largest drop recorded was 19,828 members in a single week, with the corresponding channel ceasing activity four weeks later.

#### 8.1.1 Channel lifespan

We are interested in the determinants of channel lifespan duration. We use methods that can account for the joint effects of multiple factors. Our data includes both characteristics that vary over time, such as message activity and number of members, as well as fixed characteristics, such as whether a channel is linked to a gateway channel.

Since Telegram allows us to collect messages going back in time, characteristics such as files shared and language of messages can be estimated ex-post, but characteristics such as membership cannot. Additionally, the full dataset may be biased as only channels that remained active until the beginning of the data collection can be observed. Such biases are less likely to be present on the more recent subset of the channels we observe from their inception.

We estimate Cox proportional hazards models [18] to understand how each channel characteristic influences the probability that a channel will remain active for an additional week. Since we do not directly observe the exact time a channel changes status, we model the likelihood of a channel becoming inactive as interval-censored. In other words, we assume it can occur at any time between two consecutive weeks where the status changed from active to inactive. To ensure robustness of our analysis, we vary whether we use the full data or only the more recent data, resulting in two models in total. We relegate detailed estimation results to Appendix A.3 (columns 1–2), and present here notable qualitative conclusions.

Across both models, the most influential predictor of channel durability is Telegram's policy change. Telegram announced a policy change to cooperate more closely with law enforcement in September 2024. The policy was implemented gradually, with the first measurable effects being reported in January 2025. Both of those dates play an important role. After the policy change, channels face a significantly higher risk of being banned. Whether a channel is linked to a gateway also has a significant positive effect. The language used and the data types shared have a minor effect on the longevity of a channel, whereas whether a channel is public, the number and format of files shared, and the number of messages have no effect. It is also worth emphasizing that we find no relation between the speed at which a channel grows/shrinks and the likelihood of a channel becoming inactive or banned.

### 8.1.2 Channel growth rate

We are also interested in the determinants of channel growth rate. We proxy the growth rate by the change in membership count between consecutive weeks, which can be positive if a channel is growing or negative if a channel is shrinking. We estimate three models that differ in the assumptions about the correlation of weekly changes in membership. The first model assumes that weekly changes are independent, the second assumes that the correlations over time are exchangeable (i.e., the correlation between two weekly changes is the same no matter how far apart they are), and the third model assumes they follow an AR(1) autocorrelation process (i.e., the correlations decay over time, weeks closer together have more correlated changes compared to weeks further apart).

Detailed estimation results are relegated to Appendix A.3 (columns 3–5 for each model, respectively.) The only factor that has a consistent effect on growth rate is whether a channel is linked; linked channels grow significantly faster. It is also important to highlight the effect of Telegram policy changes. We only have membership data after Telegram announced its policy change. We observe a significant negative shock to growth rates once this policy started being implemented and received broad news coverage in January 2025. However, this direct effect is short-lived, as once we incorporate autocorrelations in the growth rates, it loses significance. This implies that the news coverage of the policy implementation acted as a shock to the ecosystem and triggered a negative spiral. Once participants started leaving channels, more participants followed in the following weeks.

### 8.1.3 A closer look at gateways and linked channels

Having identified the important role that gateway channels play in both longevity and growth rate of channels, we take a closer look at how linked channels differ from the rest of the ecosystem. We find that linked channels have a significantly higher participant count (mean 3,324, median 1,328), compared to non-linked channels (mean 2,265, median 719).

Additionally, we also find a much higher growth rate among linked channels; they gain, on average, 25 new participants per day, compared to nine within non-linked channels. We might overestimate the growth rate for linked channels because we may detect them earlier through the gateway than we do for other channels. To account for this bias, we compute average growth rates across the full lifetime of channels by dividing their highest participant count by days since the first message. Using this auxiliary metric, we still find a much steeper growth of linked channels compared to non-linked—ten and four new participants per day, respectively. Overall, both metrics show more than two times higher growth rate of linked channels.

Finally, we also find that linked channels are almost exclusively involved in trading with credentials and infostealer logs, and rarely carding, and hardly ever deal with datasets, personal information, and personal documents. Overall, the evidence shows that, that gateways cannot protect a channel from being banned, but they help the community quickly recover when new channels are created.

## 8.2 Qualitative evidence on additional resilience mechanisms

This subsection presents a qualitative description of additional strategies that channels use to ensure the survival of their communities and to avoid detection.

### 8.2.1 Backup channels

Newly linked channels occasionally start as *backup* channels, recognizable from keywords in their name such as backup, reserve, new, v2, or 2.0 (or v3 or higher in case of multiple channels bans). Backup channels typically mirror their sibling channel in terms of their properties, except for their activity; in that respect, they typically lie dormant until the sibling channel is banned, and then pick up where it left off.

Backup channels do not use consistent naming and might overlap with the main channel, making them difficult to reliably quantitatively analyze[19]. Instead, we present one rep-

[19] We reliably identified 41 backup channels. There are likely more that started as a backup, but we did not capture in that stage.

resentative case to illustrate the operation. The two channels are involved in the trade of infostealer logs; following the common naming convention, but anonymizing the channels, we refer to them as “Aardvark Cloud” and “Aardvark Cloud Backup”. Figure 3 shows the recorded membership history of both channels.

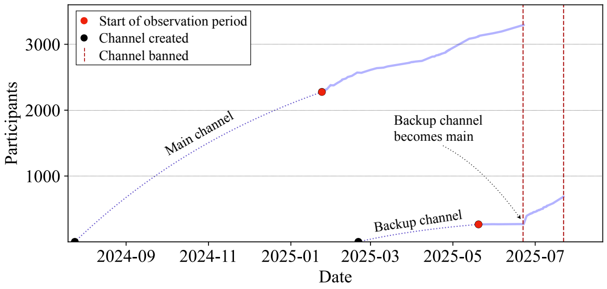

Figure 3: Aardvark main and backup channel membership

Aardvark Cloud was established in July 2024. By the time we first collected participant statistics in January 2025, it had reached over 2,000 members.[20] We see a relatively steady growth between then and June 2025, when the channel was banned. Earlier, in February 2025, Aardvark Cloud Backup was created. The backup channel attracted a small fraction of the main channel’s membership, initially just under 10%. However, when the main channel was banned in June 2025, membership in the days that followed showed a discrete jump, followed by steady growth over time, mirroring the pattern observed for the now-banned Aardvark Cloud. At the same time, the name of the active channel was changed from Aardvark Cloud Backup to Aardvark Cloud, clearly signaling the shift in function. After the change, the new main channel remained active for another month, when it was also banned, having reached just under 700 members (20% of the previous channel’s peak membership). We could not reliably confirm the presence of another backup channel, and we therefore consider Aardvark Cloud extinct.

#### 8.2.2 Ban evasion tactics

We also identify numerous tactics used to disguise the true purpose of channels. First, some channels use special characters, fonts, misspelled words, or spacing in problematic words, most likely to avoid detection. Second, we find channels that include disclaimers in the channel description, explaining that the channel is for educational purposes, ethical reasons, or for awareness and research. Some channels claim to be role-playing channels or discussing illicit activities in the context of a video game. Third, some channels contain a direct address to Telegram and its owner, explaining that their channel does not violate the law as it does not contain pornographic or “rude” content. The specific jurisdiction to which it refers remains unclear, but it is typically written in Russian.

Focusing on the most recent available 1050 channels (without empty descriptions), we identify 53 channels that address Telegram or its owner, 50 using special characters or spelling to obscure the text, and 44 explaining they are for educational, role-playing, or research purposes. As we only look at the last available description, these are a lower bound, with the mentioned tactics likely occurring more frequently across the dataset.

Finally, during the data collection period, we record 180 channels that have changed their names and branding. 73% of these channels only rebrand once. The remaining channels rebrand up to six times, with a single channel that changed names 18 times. A small portion of these changes is relatively minor: changes in capitalization, emojis, wording (e.g., cc becomes carding), or functionality (e.g., backup channel becomes main). However, some channels completely rebrand in a way that makes them disconnected from their previous identity.

## 9 Discussion, implications, and limitations

### 9.1 Clustering and sub-ecosystems

The stolen data ecosystem exhibits a high diversity of data types, with a relatively stable presence over time. Notably, infostealer logs are gaining momentum. This is particularly alarming, as they are primarily obtained through malware on the victim’s machine, and allow much more effective attacks compared to previously possible credential compromises.

Our analysis suggests the presence of three distinct sub-ecosystems within the stolen data landscape on Telegram. The first comprises professional and profit-driven credential and infostealer log “Cloud” channels characterized by subscription-based access and automated purchasing via bots. These channels exhibit resilience through gateways and operate primarily as one-way communication channels with limited discussion, which also prevents us from quantifying how stolen data is used.

The second sub-ecosystem centers on carding, sometimes supported through personal information. It is also highly specialized and motivated by profit. The carding sub-ecosystem is dominated by closely connected channels operating with a local focus—evidenced by country analysis—with Spanish, Chinese, and Portuguese languages prominently represented. Despite a smaller number of channels, carding constitutes the most active segment of the ecosystem accounting for roughly two thirds of message volume.

A third, less specialized and less industrialized sub-ecosystem exists, characterized by seemingly more opportunistic motives, a lack of clear linguistic patterns, more lively discussion, and a prevalence of databases, personal informa-

[20] The membership count before is estimated from the average growth rate of primary and backup channels, respectively.

tion, and documents. Common co-occurrence of databases and personal information suggests a more supportive role, helping attackers identify victims and understand the attack vectors.

We emphasize that the two specialized sub-ecosystems operate on granular data – individual payment card numbers or account credentials. This data is characterized by high volume and a steady supply via automated threats like malware, card skimming, and phishing campaigns. This is unlike the less specialized third sub-ecosystem, which relies more on larger bulk datasets and personal data, likely obtained in breaches that often require more manual effort on the attacker's side.

The fragmented nature of data sales within this messaging platform also aligns with the economic incentives. Selling high volumes of granular data (e.g., individual card details or credentials) presents a lower barrier to entry and potentially faster turnover compared to monetizing large, static breached datasets. Furthermore, the subscription model—offering access for as little as $40/month—reduces financial risk for buyers in a fragmented and unstable ecosystem. It also minimizes the need for extensive data cleaning and validation, issues often associated with larger, less curated datasets purchased in bulk. The platform's limited persistence likely contributes to this preference for continuous, smaller transactions.

## 9.2 Implications

Guided by our findings, we identify a set of implications that should guide future research, policies, and practitioners.

**Channel discovery.** The stolen data ecosystem is largely disjoint from the rest of the Telegram ecosystem, suggesting that general-purpose data collection efforts are unlikely to successfully reach these specific channels. Researchers, threat intelligence providers, and law enforcement officials seeking to understand stolen data communities on Telegram must anticipate this and invest effort into timely discovery. The incremental snowball approach taken in this study—built on candidate discovery from several diverse sources and a set of keywords in different languages—proved effective at identifying channels otherwise invisible in less targeted data collection efforts. Future research should investigate the automation of the snowballing process, in particular, candidate evaluation.

**Languages.** While English is the most frequently used language overall, at times it represents as little as 30% of measured weekly activity. In particular, we find clusters of specialized activity centered around Spanish, Chinese, Portuguese, Russian, and other languages. Additionally, certain languages are over-represented in specific parts of the market. Any research focusing on English-only channels is likely to miss important segments. When investigating these communities, we should carefully consider the linguistic dimension, acknowledging that language selection may significantly shape research findings.

**Resilience.** We recognize that gateway channels are crucial for stability and growth rate, especially for credential and log channels. They allow communities to quickly recover following a ban. The presence of gateways suggests that future ban attempts will likely see a short-term success and a swift recovery of the banned community. Moderating gateway channels by Telegram remains essential.

**Platform policy.** Telegram's policy changes, such as their decision to cooperate with law enforcement, can drastically affect the ecosystem, as both the lifespan and the growth rate of the channel are affected. This highlights the important effect that top-down platform policies can have on the communities they host and the activities they facilitate.

## 9.3 Limitations

Our study has several limitations that should be considered when interpreting the results. First, as with all measurement-based research, dataset representativeness is a concern. Especially, we might miss short-lived channels due to the retroactive nature of data collection (i.e., survivorship bias). While we collect data from multiple sources and aim for broad linguistic coverage, we did not include some widely spoken languages in our set of keywords, such as Indian languages, Portuguese, and Vietnamese. Consequently, our findings should be understood as reflective of the observed sample rather than the entire ecosystem.

Second, our analysis is limited only to publicly accessible content. Activities occurring in closed, or paid channels remain unobserved (except for those that provide public invitation links). This necessarily leaves parts of the ecosystem outside the scope of our study.

Third, a limitation of this study is the manual evaluation of candidate channels. While this approach allows a more nuanced and flexible filtering of candidates matching our stolen data definition, it prevents full-scale automatic discovery and evaluation of potential candidates (e.g., through LLMs). Instead, we only use LLMs during the analysis phase, when working with a static one-year snapshot of the ecosystem.

Fourth, regarding the data type classification, we classify channels on a weekly basis rather than by individual message. While we believe our results are representative at a high level, in some cases, the sampling process may have captured outliers (e.g., a week where a primarily carding-focused channel discussed personal information). However, we did not observe many such cases. Additionally, we use an LLM to scale up the classification of stolen data types. While the model demonstrated satisfactory performance on the subset evaluated by human coders, applying it to the full corpus introduces a level of uncertainty.

Finally, we cannot make claims about the authenticity or quality of the stolen data. Our observations are limited to advertisements and publicly available data, which may be outdated, inaccurate, or deliberately falsified. Furthermore, for ethical reasons, we did not analyze the stolen data itself.

## 10 Conclusion

In this work, we investigate the newly emerging stolen data markets operating on the Telegram platform. This study seeks to characterize these markets, analyze the types of stolen data commonly traded within them, and identify which factors influence their lifespan and sustainability.

This work presents the largest dataset to date of stolen data markets on Telegram, which we are releasing to the research community to facilitate further research. We discover how disconnected the stolen data ecosystem is from the broader Telegram platform. We find a diverse set of communities, with English as the dominant language, but significant activity also in Chinese, Spanish, Portuguese, Russian, and other languages. We identify specialized, automated, and more professional communities centered around carding and account stealing, facilitated by credentials and infostealer logs. Additionally, we identify a less specialized cluster of communities involved in sharing stolen datasets, personal information, and personal documents.

We also observe several ban evasion and resilience mechanisms, the most notable of which is the use of gateway channels. We find that gateways are an effective way of ensuring long-term presence, even when ban action is taken. On the contrary, Telegram's policy change in September 2024 to cooperate with law enforcement hindered channel longevity, membership growth, and overall membership numbers. Acknowledging the inherent challenges and limitations of empirical data, we propose actions for future researchers and policymakers to improve detection, observation, and responses to the illicit data economy on Telegram.

## Acknowledgments

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. Taro Tsuchiya is supported by the Nakajima Foundation and the Cylab Presidential Fellowship. Konstantinos Ioannidis is supported by the Leverhulme Trust. We are grateful to our colleagues Hugo Bijmans, Annalina Buckmann, Mariella Mischinger, Marilyne Ordekian, Yanna Papadodimitraki, Yangheran Piao, Harshini Sri Ramulu, Roy Ricaldi, Anna Talas, Anh V. Vu, and Luna Wang for their help with keyword creations and helpful feedback on earlier versions of this paper. We also thank the anonymous reviewers and the committee members for helpful feedback and insightful comments.

## Ethical Considerations

We observe a collection of Telegram channels dedicated to the trade of stolen data. Our research involves observing and analyzing spaces used for illegal activities. As such, it requires careful consideration of the stakeholders, possible harms to them, and the precautions taken to minimize the harms.

**Ethics committee approval.** This research project was granted ethics approval from the department's Ethics Committee (Ethics Review #2407, Department of Computer Science and Technology, University of Cambridge). The research involves the collection and analysis of stolen data originating from data breaches, malware, and similar sources. Recognizing the sensitivity of this data, we carefully considered our obligations under the General Data Protection Regulation (GDPR). While we recognize that legal compliance is distinct from ethical sufficiency, we view regulations such as the GDPR as a practical guideline for the handling of sensitive personal data. We justify the collection and processing under the legitimate interest basis, specifically the substantial public benefit derived from understanding the illicit online ecosystem and informing strategies to mitigate future harms. We acknowledge the deontological perspective on the ethics of our research, which would suggest that the individuals in our dataset have an inalienable right to privacy, which is violated when their data is involved in a study [20]. We ultimately take a consequentialist view, which is that the public benefit overweighs the potential risk to individuals [24].

**Stakeholders and harm considerations.** We identify a number of stakeholders and possible harms to them.

*Participants in the markets*. Individuals actively involved in the observed markets risk legal repercussions if identified. Even perceived association with illicit activity could lead to consequences, and the expectation of anonymity within these closed groups is a significant concern.

*Telegram platform*. Public exposure of the prevalence of these markets could damage Telegram's reputation and attract increased scrutiny from regulators and law enforcement. This may lead to demands for greater content moderation, potentially impacting user privacy more broadly.

*Researchers*. Exposure to disturbing content carries the risk of psychological distress and secondary trauma for the research team. Further, there is a potential, albeit low, risk of retaliation or harassment from individuals involved.

*Victims, whose data was stolen.* Research highlighting the continued trading of their stolen data could constitute revictimization and exacerbate emotional distress. Publicizing details of specific breaches may also increase the risk of further fraud or identity theft. In case of data breaches, our findings might highlight systemic failures in data protection practices, potentially leading to investigations and fines levied against organizations responsible for the original data breaches, leading to financial and reputational consequences. Finally, we recognize the risk that sensitive data collected during our observations could be misused by malicious parties, potentially leading to further harm for data breach victims.

**Precautions taken.** We take the following steps to minimize harm to the identified stakeholders. The dataset was collected through the official Telegram API and only contains publicly accessible information. We join public and private channels whose invite links are advertised on forums or another channel. We do not lie or pretend to be an interested buyer to be admitted into groups. We also do not attempt to join any groups that require payments or vouching by an existing member. We do not buy anything and only act as neutral observers, minimizing any interaction with the community.

We could not obtain informed consent from all members of the observed communities. According to the British Society of Criminology's ethical guidelines [6], informed consent may be waived for research examining publicly accessible online communities, particularly when the focus is on collective patterns of behavior rather than individual actions. In the United States, publicly available data without personal identifiers generally does not qualify as human subject research and are therefore exempt from IRB review, unless multiple data sources are combined to infer personal information. We do not attempt to de-anonymize any channel participants or victims at any point. We also do not name specific channels or breached organizations. We also clarify on several occasions that Telegram is not inherently a cybercriminal platform. Researchers involved in the research had access to mental health resources or debriefing sessions to address potential psychological impacts from exposure to disturbing content.

We do not foresee any negative impacts resulting from the publication of this research, as our analysis focuses on aggregated trends and systemic patterns rather than individual identification or exposure. All data handling procedures were conducted with great care and adherence to relevant guidelines. The collected data is stored in a secure, encrypted server and is only accessible to researchers working on this project. The dataset is available to academic researchers through a license agreement to allow more careful control and limit access for research purposes only. We do not access or analyze the contents of the files, except to compute their hashes required to recognize distinct files. Any automated analysis, including LLM classification, is performed locally, rather than using cloud-based services, ensuring all data processing remains fully under our control.

## Open Science

The dataset collected for the purpose of this study is available from the Cambridge Cybercrime Centre [1]. Due to the sensitive nature of the collected data, we are unable to fully release the database publicly as an artifact. Instead, the data can be shared with academic researchers for appropriate projects through a license agreement between the researchers' institution and the University of Cambridge.

## References

[1] Cambridge Cybercrime Centre. Legal framework. https://www.cambridgecybercrime.uk/data.html. 2016.

[2] Lillian Ablon, Martin C Libicki, and Andrea A Golay. *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar*. RAND Corporation, 2014.

[3] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. The Pushshift Telegram dataset. In *Proceedings of the International Conference on Web and Social Media*, pages 840–847. AAAI, 2020.

[4] Hugo Bijmans, Tim Booij, Anneke Schwedersky, Aria Nedgabat, and Rolf van Wegberg. Catching phishers by their bait: Investigating the Dutch phishing landscape through phishing kit detection. In *Proceedings of USENIX Security Symposium*, pages 3757–3774. USENIX Association, 2021.

[5] Kitty Boersma. So long and thanks for all the (big) fish: Exploring cybercrime in Dutch Telegram groups. Master's thesis, University of Twente, 2023.

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

[7] Michele Campobasso and Luca 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*, pages 1665–1680. ACM, 2020.

[8] Federico Cernera, Massimo La Morgia, Alessandro Mei, and Francesco Sassi. Token spammers, rug pulls, and sniper bots: An analysis of the ecosystem of tokens in Ethereum and in the Binance smart chain (BNB). In *Proceedings of USENIX Security Symposium*, pages 3349–3366. USENIX Association, 2023.

[9] Nicolas Christin. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the International Conference on World Wide Web*, pages 213–224. ACM, 2013.

[10] Forbes. The wiretap: Telegram handing data on thousands of users to law enforcement across the world. https://www.forbes.com/sites/thomasbrewster/2025/01/07/telegram-hands-data-on-thousands-of-users-to-law-enforcement/, 2025. Accessed: 2025-08-25.

[11] Bingyu Gao, Haoyu Wang, Pengcheng Xia, Siwei Wu, Yajin Zhou, Xiapu Luo, and Gareth Tyson. Tracking counterfeit cryptocurrency end-to-end. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, pages 1–28. ACM, 2020.

[12] Taisiia Garkava, Asier Moneva, and E Rutger Leukfeldt. Stolen data markets on Telegram: a crime script analysis and situational crime prevention measures. *Trends in Organized Crime*, pages 1–25, 2024.

[13] Dimitrios Georgoulias, Ricardo Yaben, and Emmanouil 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*, pages 1–10. ACM, 2023.

[14] Yanhui Guo, Dong Wang, Liu Wang, Yongsheng Fang, Chao Wang, Minghui Yang, Tianming Liu, and Haoyu Wang. Beyond app markets: Demystifying underground mobile app distribution via Telegram. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, pages 1–25. ACM, 2024.

[15] Hans WA Hanley and Zakir Durumeric. Partial mobilization: Tracking multilingual information flows amongst Russian media outlets and Telegram. In *Proceedings of the International Conference on Web and Social Media*, pages 528–541. AAAI, 2024.

[16] Alice Hutchings and Thomas J Holt. A crime script analysis of the online stolen data market. *British Journal of Criminology*, 55(3):596–614, 2015.

[17] Vincenzo Imperati, Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Francesco Sassi. The conspiracy money machine: Uncovering Telegram's conspiracy channels and their profit model. In *Proceedings of USENIX Security Symposium*, pages 5229–5246. USENIX Association, 2025.

[18] John D Kalbfleisch and Douglas E Schaubel. Fifty years of the Cox model. *Annual Review of Statistics and Its Application*, 10(1):1–23, 2023.

[19] Klim Kireev, Yevhen Mykhno, Carmela Troncoso, and Rebekah Overdorf. Characterizing and detecting propaganda-spreading accounts on Telegram. In *Proceedings of USENIX Security Symposium*, pages 161–180. USENIX Association, 2025.

[20] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. Ethical frameworks and computer security trolley problems: Foundations for conversations. In *Proceedings of USENIX Security Symposium*, pages 5145–5162. USENIX Association, 2023.

[21] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. TGDataset: Collecting and exploring the largest Telegram channels dataset. In *Proceedings of the ACM SIGKDD Conference*

*on Knowledge Discovery and Data Mining*, page 2325–2334. ACM, 2025.

[22] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971.

[23] Tina Marjanov and Alice Hutchings. SoK: Digging into the digital underworld of stolen data markets. In *Proceedings of IEEE Symposium on Security and Privacy*, pages 1–18. IEEE, 2025.

[24] James Martin and Nicolas Christin. Ethics in cryptomarket research. *International Journal of Drug Policy*, pages 84–91, 2016.

[25] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, 8(3):607–617, 2021.

[26] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, Maurizio Tesconi, and Emilio Ferrara. Charting the landscape of online cryptocurrency manipulation. *IEEE Access*, 8:113230–113245, 2020.

[27] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the World Wide Web Conference*, pages 1845–1854. ACM, 2018.

[28] Matej Rabzelj and Urban Sedlar. Beyond the leak: Analyzing the real-world exploitation of stolen credentials using honeypots. *Sensors*, 25(12):3676, 2025.

[29] Sayak Saha Roy, Elham Pourabbas Vafa, Kobra Khanmohamaddi, and Shirin Nilizadeh. DarkGram: A large-scale analysis of cybercriminal activity channels on Telegram. In *Proceedings of the USENIX Security Symposium*, pages 4839–4858. USENIX Association, 2025.

[30] Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *Proceedings of the USENIX Security Symposium*, pages 33–48. USENIX Association, 2015.

[31] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[32] Aleksandra Urman, Justin Chun-ting Ho, and Stefan Katz. Analyzing protest mobilization on Telegram: The case of 2019 anti-extradition bill movement in Hong Kong. *Plos One*, 16(10):1–21, 2021.

[33] Aleksandra Urman and Stefan Katz. What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society*, 25(7):904–923, 2022.

[34] Otavio R Venâncio, Carlos HG Ferreira, Jussara M Almeida, and Ana Paula C da Silva. Unraveling user coordination on Telegram: A comprehensive analysis of political mobilization during the 2022 Brazilian presidential election. In *Proceedings of the International Conference on Web and Social Media*, volume 18, pages 1545–1556. AAAI, 2024.

[35] Samantha Walther and Andrew McCoy. US extremism on Telegram. *Perspectives on Terrorism*, 15(2):100–124, 2021.

[36] Mariëlle Wijermars and Tetyana Lokot. Is Telegram a "harbinger of freedom"? The performance, practices, and perception of platforms as political actors in authoritarian states. *Post-Soviet Affairs*, 38(1-2):125–145, 2022.

[37] Jiahua Xu and Benjamin Livshits. The anatomy of a cryptocurrency pump-and-dump scheme. In *Proceedings of USENIX Security Symposium*, pages 1609–1625. USENIX Association, 2019.

[38] Ahmet S Yayla and Anne Speckhard. Telegram: The mighty application that ISIS loves. *International Center for the Study of Violent Extremism*, 9, 2017.

# A Appendix

## A.1 Keywords used for finding candidate communities

| Language | Keywords |
|---|---|
| English | data, leak, breach, database, base, personal data, account, lead, card, cc, db, cvv, acc, combo, fullz, log, document, passport, ID mail, kyc, ULP (url:log:pass), dump, cloud, locker, ransom, NFC, config, spy, secret, UHQ, intel |
| Dutch | leads, combo, duw, duwleads, acccounts, accies, configs, acc, bestelgeschiedenis, saldo, accs |
| Greek | κλεμμένα, κλεμμένοι, διαρροή, υποκλοπή, δεδομένα, κωδικοί, πιστωτικές κάρτες, βάση δεδομένων |
| Spanish | datos robados, informacion personal, cuentas bancarias, tarjetas de credito, credenciales, base de datos, numero de tarjeta, tarjetas, contraseña, contraseñas, informacion secreta, informacion privada, codigo, codigos, clave, claves, clave digital |
| French | carte de credit, mot de passe, carte, banque, carte bancaire, info personnel, infos, banque en ligne, info privé |
| Italian | carta, carta di credito, chiavetta, chiave, banca, banca online, credenziali, truffa, bonifico istantaneo, ricarica, carta prepagata, informazioni, informazione privata, nascosta, nascosto, chiave segreta |
| German | daten, datenbank, persönliche Daten, persoenliche Daten/Informationen, Konto, Kontodaten, Datenbruch, passwoerter, passwörter, nutzerdaten, kreditkarte, dokumente, email adressen, nutzerdaten, benutzerdaten, zugangsdaten, adressen, datenpanne, identitaet, identität, personalausweis, fuehrerschein, führerschein |
| Estonian | andmed, leke, andmeleke, lekitus, andmebaas, isiklikud andmed, personaalsed andmed, konto, kaart, pangakonto, dokument, logi, meil |
| Chinese | 数据，信息，泄露, 泄漏, 脱裤拖库，数据库, 库，个人信息, 个人数据，私人信息，账号, 账户，线索, 电销数据, 目标客户，卡，信用卡，银行卡，借记卡，储蓄卡，安全码，记录, 日志，文件, 文档，邮件, 邮箱, 卡号, 信用卡号, 信用卡卡号, 银行卡号, 银行卡卡号, 轨道料, 账号密码打包, 护照, 身份证, 实名认证, 登录凭证, 数据库泄露, 撞库, 拖库, 云, 云服务, 加密, 勒索, 卡贴, 盗刷 |
| Arabic | ديتابيس فدية, اختراق, تسريب, بيانات, مجموعة |
| Russian | Базы Данных, облак, лог, данные, слив, сливки, slivki, утечка, взлом, комбо, пачка, полные данные, полный профиль, документ, персональные данные, личные данные, аккаунт, счёт, карта, кредитная карта, Фуллз, Переходник |
| Japanese | データ, 漏洩, 侵害, データベース, 個人情報, アカウント, リード, カード, クレカ, クレジットカード, データベース, アカウント, コンボ, ログ, 文書, パスポート, メール, 本人確認, ログインパスワード, ダンプ, クラウド, ロッカー, ランサム, 身代金, コンフィグ, スパイ, 秘密 |

Table A1: Keywords in all languages

### A.2 Data type classification prompt

Below is the complete prompt used for data type classification by Gemma3 as well as human annotators.

You are aiding a cybercrime researcher. You observe a Telegram channel involved in the trade of stolen data. There are 6 types of stolen data you consider:

- credentials: Credentials are usernames and passwords or password hashes, sometimes together with the URL of the website, service or application that credentials belong to. Examples and keywords often associated with it include: username:password, email:password, username:hash, combo, login, ULP (URL:login:password), configs, mail access. If the discussion includes sale of infostealer logs or cookies that might also contain credentials, classify it as infostealer_logs only.
- infostealer_logs: Infostealer logs contain user credentials, cookies, browser history, digital (behavioral) fingerprints, form fill information, system configuration, and any other information obtained through malware. Examples and keywords often associated with it include: log, infostealer, cookies.
- databases: Databases are datasets or proprietary information stolen from companies, government organizations, educational institutions, etc. in a data breach. This includes non-personal documents such as trade secrets, designs, schemas etc. Examples and keywords often associated with it include: breach, leak, spy, secret, database, dump, names of breached companies.
- personal_information : Personal information consists of people's personally identifiable information such as name, email address, residence, country, phones, interests, employment status, medical history, and more. This data can be compiled from breached databases or publicly available information. Examples and keywords often associated with it include: leads, fullz, consumer info, user info (name, DOB, phone, address, email).
- carding: Carding information is any information required to make direct transactions and purchases, such as credit cards, debit cards. Includes access to online banking that allows direct cashout. Examples and keywords often associated with it include: CVV, credit, debit, CC, BIN, NFC, Visa, MasterCard, dump, expiration, lists of numbers that could represent banking information, including the balance. Make sure you do not confuse information about the price of items for sale with carding information.
- personal_documents: Personal documents photos or scans of official documents used for identification and travel of individuals. Examples and keywords often associated with it include: passport, driving license, ID, KYC, selfie, verify.
- other: None of the above data types detected, unknown, or not enough information to make a decision.

Multi-labels are allowed, but should only be used sparingly, and only when multiple data types are featured prominently enough, rather than just being mentioned shortly. Use at most 3 labels. When using multi-labels, make sure the first label is the one that best represents the largest portion of the data and is the most distinct. Only pick the exact label names provided in the dictionary.

Write your answer in the form label1 or for multiple labels label1,label2,label3.

Which of the defined stolen data types is advertised or shared in the channel?

Please ONLY output your response in the following format: CLASSIFICATION:values. Provide no additional information.

- This is the channel name: {channel_name}
- This is the channel username: {channel_username}
- This is the channel description: {channel_description}
- This is the content of a set of recent messages: {message_content}
- These are the names of files that appeared in the recent messages: {filename}

## A.3 Statistical analysis of channel longevity and growth rate

| | Longevity | | Growth rate | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Weeks active | 0.001$^{*}$ | −0.001 | 0.093 | 0.041 | 0.086 |
| Public | 0.266$^{**}$ | 0.195 | −16.732 | −20.138$^{*}$ | −17.556 |
| Linked | −0.556$^{***}$ | −0.444$^{**}$ | 50.959$^{*}$ | 52.978$^{***}$ | 56.475$^{***}$ |
| Single-specialized | 0.190$^{*}$ | 0.085 | 19.285 | 14.032 | 21.139 |
| Double-specialized | 0.271$^{**}$ | 0.074 | −9.823 | −12.118 | −1.651 |
| Number of files shared | −0.001 | −0.001 | −0.002 | −0.001 | −0.002 |
| Percentage of image files | 0.002$^{*}$ | 0.003$^{*}$ | −0.179 | −0.296 | −0.038 |
| Percentage of text files | 0.007$^{***}$ | 0.003$^{*}$ | 0.083 | 0.079 | 0.410 |
| Percentage of other files | 0.004$^{**}$ | 0.001 | −0.017 | −0.067 | 0.467$^{*}$ |
| English | 0.001 | 0.001 | −0.092 | −0.156 | 0.079 |
| Chinese | −0.004 | −0.011$^{*}$ | −1.352 | −1.477 | −0.140 |
| Spanish | 0.008$^{***}$ | 0.006$^{**}$ | 0.109 | 0.039 | 0.576$^{*}$ |
| Portuguese | 0.007 | 0.006 | 0.507 | 0.453 | 0.705 |
| Russian | −0.001 | −0.002 | 0.204 | 0.072 | 0.746$^{*}$ |
| Farsi | 0.001 | −0.001 | −0.293 | −0.265 | −0.439 |
| Indonesian | −0.008 | −0.004 | −1.863 | −1.766 | −0.553 |
| Credentials | −0.006$^{***}$ | −0.001 | −0.098 | −0.091 | −0.101 |
| Infostealer logs | −0.010$^{***}$ | −0.004 | 0.072 | 0.016 | −0.470 |
| Carding | −0.006$^{**}$ | −0.001 | −0.308 | −0.280 | −0.513 |
| Databases | −0.003 | 0.002 | 0.063 | 0.073 | 0.165 |
| Personal information | −0.009$^{***}$ | −0.006 | 0.174 | 0.158 | 0.073 |
| Personal documents | −0.006 | −0.003 | 0.624 | 0.547 | 1.382 |
| Number of messages | −0.001 | 0.001 | 0.003 | 0.002 | 0.003 |
| Percentage of forwarded messages | −0.001$^{***}$ | −0.001$^{***}$ | −0.003 | −0.003 | −0.004 |
| Percentage of admin messages | 0.001 | 0.001 | −0.003 | −0.003 | −0.003$^{*}$ |
| Weekly membership change | | 0.001 | | | |
| Telegram policy change announced | −1.245$^{***}$ | | | | |
| Telegram policy change implemented | −1.955$^{***}$ | −0.741$^{***}$ | −37.479$^{*}$ | −36.577 | 1.101 |
| Observations | 45,596 | 9,702 | 9,702 | 9,702 | 5,920 |

Notes for table:
Dependent variable: Models (1)–(2) binary (channel active/inactive). Models (3)–(5) count (weekly member change)
Model (1): All data, Cox proportional hazards models assuming interval-censored survival time
Model (2): Recent data, Cox proportional hazards models assuming interval-censored survival time
Model (3): Recent data, linear mixed effects model, no assumption on correlations between member changes
Model (4): Recent data, general estimating equation (GEE) model assuming correlations between member changes are exchangeable
Model (5): Recent data, general estimating equation (GEE) model assuming correlations between member changes are autocorrelated AR(1)

Table A2: Effect of channels characteristics on longevity and growth rate of a channel