Telegram's Dark Trade: Unpacking Brazil's Data Leak Surge

Yuri do A. N. Maia[1], **Manuel Sánchez Rubio**[1]

[1]Department of Computer Science – University of Alcalá (UAH)
Alcalá de Henares – Madrid – Spain

yuri.amaral@edu.uah.es, manuel.sanchez@uah.es

Abstract. *The exploitation of leaked personally identifiable information (PII) has become a key enabler of social engineering attacks. While dark markets are commonly associated to traditional dark Web forums, Telegram emerges as an easy-to-use meeting place for users and vendors. This study investigates a specific ecosystem within Telegram known as "Pull Groups" (PG), where large amounts of sensitive personal data are requested and shared. During a six-month monitoring period, we systematically collected and analyzed data, extracting more than 12 million PII records. To assess the potential impact of this exposure, we propose the Leak Exposure Index (LEI), which combines the leak volume with he group size to highlight high-risk environments. Our findings show that a small set of groups are responsible for the majority of leaked data, underlining the need for countermeasures and informed risk assessment strategies. This is also a call for public awareness of this threat.*

1. Introduction

In Brazil, 50% Internet users are concerned or very concerned about records of their activity when accessing websites and social networks [CETIC.br 2024]. This concern is not without reason, as a trend has been observed since 2020 in which physical robbery is being replaced by forms of cybercrime [Fórum Brasileiro de Segurança Pública 2024]. Between 2022 and 2023 alone, electronic fraud increased by 13.6%. In this scenario, Telegram surges as a dark market vector for selling (or freely distributing) personally identifiable information (PII).

Despite traditional dark Web environments like Tor being more commonly associated with illicit trading, Telegram is increasingly emerging as a popular platform for dark markets [Garkava et al. 2024]. It allows criminals to easily set up marketplaces to sell different products and services to a large number of users through channels and groups. Unlike the often complex Tor setup, Telegram enables users to connect, communicate, and trade with little knowledge. Along with this, Telegram has been known for its alleged protection of user anonymity, providing the grounds to be a meeting point for vendors and clients of illicit goods.

A major commodity in dark markets is breached data [Liu et al. 2020]. These breaches can encompass a variety of sensitive data types, including credentials, credit/debit cards, browser cookies, and PII. Each can be further exploited for many different crimes, for instance, network intrusions, fraudulent transactions, and identity theft. The versatility of this kind of product for crimes is underscored by their significant presence in dark markets, where they are clearly the foremost product offered [Georgoulias et al. 2023].

Leaked PII, including name, e-mail, phone number, and parents' names, is a powerful tool for social engineering and electronic fraud. Fraudsters take advantage of ongoing and fake government programs to create phishing and scam pages, while also leveraging leaked PII to enhance credibility and effectively deceive their victims [Santini et al. 2025]. Once the fraudsters have achieved their goal, the impact on the victim goes beyond financial loss [Kayser et al. 2024]. The emotional distress impacts the victim's mental health, which can lead to a lifetime anxiety. Victims may also experience anticipated stress due to concerns about the malicious usage of their PII in the future.

Notwithstanding its importance, the extent of the impact goes far beyond the individual financial loss. In 2024, Brazil's total losses in financial scams surpassed US$ 1.7 billion [Folha de S.Paulo 2025], an increase of 17% from 2023. This economic impact can undermine public trust in institutions, which could lead to systemic effects such as a decrease in economic activity.

In this study, we address a significant source of free leaked PII known in the Brazilian underground as ”pulls” (Portuguese: *puxadas*) or ”queries” (Portuguese: *consultas*). To the best of our knowledge, this is the first study to shed light on the activity of these groups in Telegram. We investigate the following research questions: how to systematically identify and detect groups dedicated to illicit PII trading and how to assess the potential real-world impacts of the leaked data?

Our research quantifies the availability of PII for exploitation and proposes the Leak Exposure Index (LEI), thereby setting the stage for fraud risk assessment and public awareness campaigns to empower individuals against victimization.

2. Related Work

Dark markets have been the subject of extensive research. [Georgoulias et al. 2023] analyzed cybercrime-related products traded on popular dark Web forums. The fraud category represents 71% of listed products, which includes carding and account information. [Liu et al. 2020] investigated the risk of PII exposure in both the dark Web and the clear Web. They collected nearly 1.2 billion records, most of them being stolen account credentials. [Garkava et al. 2024] researched the operation of dark markets on Telegram. It has been noted that Telegram's characteristics foster its usage among dark market users.

While the volume of online fraud is highlighted in criminal reports [Fórum Brasileiro de Segurança Pública 2024], [Kayser et al. 2024] stresses that harm to victims may include financial loss, emotional distress, and reputational damage. The authors underscore the long-term risks associated with the exposure of PII on dark Web. Regarding Brazil, [Santini et al. 2025] analyze the use of advertising platforms by fraudsters. Frauds would use PII to increase credibility of a site. Other studies about Telegram focus on misinformation campaigns. [Maia et al. 2024] address groups related to anti-vaccine activism. [Júnior et al. 2022] track groups related to political topics.

One particular ecosystem is yet to be explored. Pull Groups (PG) are groups that distribute PII to commit fraud. While not exclusive to Telegram, the characteristics of the platform foster the growth of such communities. Most of them have a freemium version, where a user has limited free queries but still have access to the data other users query.

3. Methodology

We propose a systematic approach to assess channels that have exposed PII within Telegram. This process consists of four stages: planning and direction, discovery, monitoring and collection, and processing and analysis.

3.1. Planning and Direction

During this phase, we established the requirements for this study. We define personally identifiable information according to the National Institute of Standards and Technology (NIST) guidelines. PII is considered any information used to identify or trace a person's identity and any other information that can be linked to a specific individual [McCallister et al. 2010]. The data searched includes leaked PII, which encompasses: Brazilian Social Security Number (CPF), name, parents' names, date of birth, phone number, email, address zip code, and employment information, including previous or current employers' Brazilian National Registry of Legal Entities number (CNPJ) and their legal entity name. The monitoring targets are channels and groups on Telegram that are public or open for self-joining. We only consider channels where the exposed data is present in the chat's body or attachments and is related to a leak. To determine whether the data was leaked and related to a PG, we consider channels that have bot commands requesting data. Groups that meet the requirements are joined.

3.2. Discovery

Once the direction is given, the following step involves finding Pull Groups. We identified relevant terms that are indicative of these groups to search in Telegram's global search in the application. The primary search terms were ”puxada” and ”consulta”, along with their inflections. These terms were taken from empirical experience of reading fraud discussion groups.

The first set of groups that resulted from the search was manually examined to check their relation to the topic. They were then searched to identify announcement chats that had invitation links to other groups, applying a snowballing approach. We checked if the link was still valid and if the group was related to pulls. The final set consists of these two sets of groups.

3.3. Monitoring and Collection

Regarding the volatility of the information, Pull Groups displays three different configurations. Groups can either retain the entire history of chats, set the timer for auto-deleting messages (typically with the shortest frame, which is 24 hours), or use a bot to delete messages in shorter periods than the auto-delete setting. This deletion bot is used in conjunction with auto-delete configuration and may be offered as a feature for paid users.

The selected strategy was to retrieve all messages available in the group upon first ingress and, afterward, fetch the messages every 15 minutes. We leverage Telethon[1] API to fetch group information and messages. Telethon allows for seamless interaction

[1] `https://tl.telethon.dev/`

with Telegram API. All data is storage by converting the retrieved object into its text representation.

Once data is retrieved, we store it in a DuckDB[2] database to further process and analysis. The monitoring was conducted over six months, from November 2024 to April 2025.

During the monitoring period, some Pull Groups would migrate to a new group or were banned due to a violation of the Terms of Service. We also noted that some Pull Groups were shadow-banned, but this did not impact our study since we kept the groups' IDs. To maintain the breadth of our observation, we inserted the new migrated groups into the monitoring set when we found them.

3.4. Processing and Analysis

In the processing step, we employ regular expressions to identify and extract the PII as defined in Subsection 3.1. Given the varied formatting of messages from each Pull Group, we use regular expressions to cover most cases. To enhance accuracy, we apply fixed keywords associated with each PII type.

We designed an analytical approach for the examination of collected data to describe the dataset from both a channel-level and a global perspective. We applied descriptive statistical methods to the dataset to provide quantitative insights such as volume, frequency, and recurrence.

4. Leaks Analysis and Discussion

This section presents the analysis of the collected and processed data from the monitored groups. First, general statistics broken down by group are given. Afterward, we analyze the PII present in the data. We then extend this analysis to a global perspective, identifying possible risk indicators and trends, before concluding with a discussion of the results.

4.1. Group-level Overview

Initially, a total of 28 Pull Groups were discovered and monitored. By the end of the monitoring, 20 groups were being monitored due to migration or bans. A total of 34 groups compose the collected data. Henceforth, groups will be referred to as PGnn, where nn represents the number attributed to each group. This nomenclature is intended to anonymize the groups and avoid publicizing their activity.

The summary statistics of the groups are presented in Table 1 and discriminated for the top five in Figure 1. Both groups PG05 and PG01 have close to $200,000$ participants, which is the limit unless it is a broadcast group. Nevertheless, they are the third and fifth groups, respectively, when the total number of messages collected is accounted for (Figure 1b). It is important to note that the top five groups comprise 79% of the total messages, as seen in Figure 1a, with the largest group in terms of messages having slightly more than one-fourth of this. Regarding duration, groups with larger lifespan are related to advertising or announcement groups.

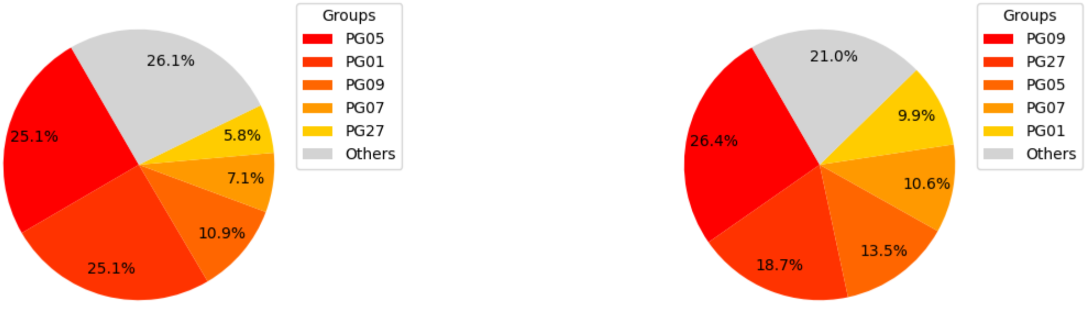
Some PGs changed their group name during monitoring. While the largest groups kept the same name, along with 61% of the groups, one particular cluster of three groups

[2] `https://duckdb.org/`

changed its name nine times. The groups are from the same creator and coexisted during a period.

Table 1. Summary statistics of group participants and messages

Metric ($n = 34$)	**Total**	**Mean**	**Median**	**Min**	**Max**
Participants	797,692	23,462	5,008	153	199,985
Messages	6,535,657	186,733	11,458	42	1,722,895
Duration (days)	-	367	261	55	1340



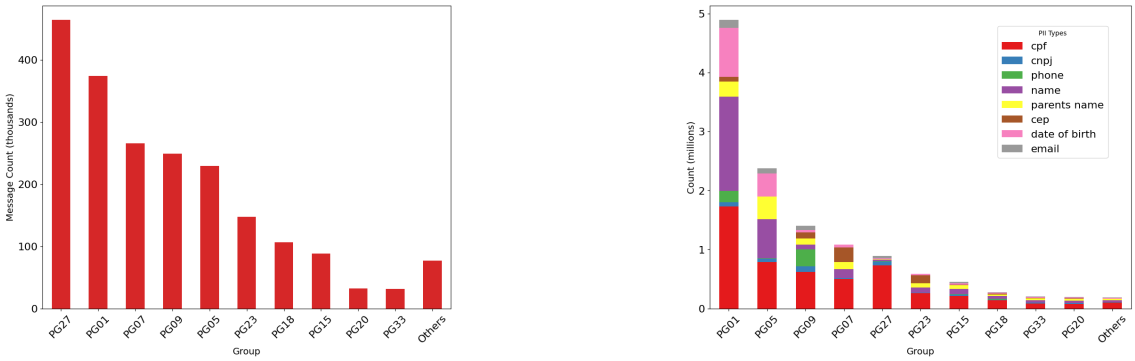
(a) Participants

(b) Messages

Figure 1. Proportion of total from top five groups

Groups may have PII in messages during the monitoring even if their purpose is not to serve as PG. To focus only on PG groups, we set minimum of 1000 PII records in a group to consider it a PG. This cut resulted in 23 PGs remaining. We analyzed the type of PII shared per group, taking into consideration the requirements from Subsection 3.1.

The analysis of PII in the top 10 PG is shown in Figure 2. Figure 2a displays the count of messages (in thousands) that contain any kind of PII. The types of PII are broken down in Figure 2b, on scale of millions. One message can contain more than one type of PII; for instance, a leaked CPF is often found with data such as a name, date of birth, and parents' names. Even though PG27 has the greatest number of messages containing PII, it is only the sixth in PII volume, whereas PG01 has the largest amount of PII.



(a) Messages with PII

(b) PII per type

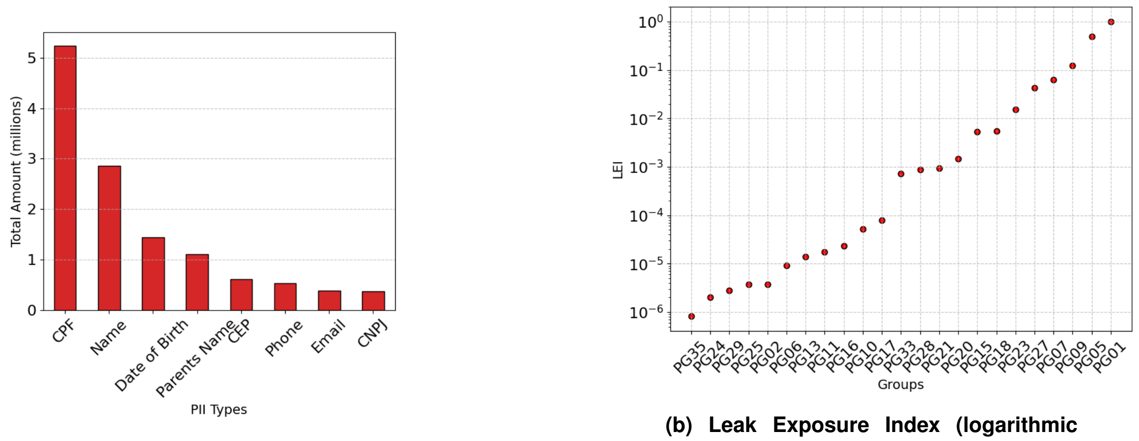
Figure 2. Count per PG

4.2. Global Overview

Analyzing the data from a global perspective, we aggregate insights from the 23 PG. A total of 12,561,248 PII records have been collected. The segmentation of the data is presented in Figure 3a. CPF is the most common data type in the dataset. Three groups – PG01, PG05 and PG09 – represent 69% of the total PII leaked.

To evaluate the risk posed by these groups, we propose a **Leak Exposure Index (LEI)**, defined in Equation 1, where L_i represents the total PII leaked in group i, U_i is the number of participants in group i and $\max(L \times U)$ is the maximum product across all groups. The index serves as a metric for the potential impact of exposure in a group: higher values indicate a greater likelihood that leaked data may be exploited.

$$\text{LEI}_i = \frac{L_i \times U_i}{\max(L \times U)} \tag{1}$$



(a) Count of PII per type

(b) Leak Exposure Index (logarithmic scale)

Figure 3. Global indicators

Figure 3b presents the groups ordered by their LEI value on a logarithmic scale. The indices identify groups PG01, PG05 and PG09 as the most prominent leak sources. Although PG01 and PG05 have similar number of participants (the maximum for this kind of channel), PG01 contains more than double the amount of PII compared to PG05. The difference in LEI between PG05 and PG09 is more significantly influenced by the number of participants.

The highest-ranked group exhibits an LEI value approximately six orders of magnitude greater than that of the three lowest-ranked groups. This significant disparity in LEI values makes it possible to identify distinct tiers of risk for exposure and the potential abuse of the leaked data.

4.3. Discussion

Each group has its own formatting for displaying leaked data. This formatting is not fixed, and groups have different styles even for different commands. During monitoring, it was noted that some groups did not show the results of a query directly in the body of the message. Two observed alternatives are plain text attachments and links to APIs.

Nonetheless, the monitoring retrieved more than 12 million PII with only simple regular expressions.

Furthermore, the observed volume of PII may be an underestimate. This is due to message deletion, as noted in Subsection 3.3, which can occur between data retrieval cycles. To mitigate this, an analysis of replies to non-existing messages can be incorporated, which serve as a proxy for deleted content. This measure could then be integrated into the LEI calculation to provide a more accurate risk assessment.

The LEI effectively identifies a small subset of three PG that are potentially the most harmful. For example, data leaked in PG01 or PG05 is exposed to nearly $200,000$ participants. The index could also be applied to assess the fraud risk of transactions. For a tiered risk categorization, groups with an LEI in the highest order of magnitude could be assigned as ”high risk”, groups in the next order of magnitude as ”medium risk”, and the remaining as ”low risk”.

As LEI can be dominated by extreme values, further analysis is warranted to develop more sensitive metrics, as well as the combination of different dimensions to address specific threats and alternative normalization approaches.

5. Ethical Considerations

The study was conducted in compliance with local data protection law (Lei Geral de Proteção de Dados - LGPD) in Brazil, ensuring the protection of sensitive information, and with Telegram API Terms of Service. Importantly, the data will not be shared and will only be retained for the duration of the research, safeguarding individual privacy. General descriptive statistics may be shared on request.

6. Conclusion and Future Works

The amount of Brazilian leaked PII that is freely available in Telegram PGs underscores the critical need to address these markets. Although limited, the monitoring of Pull Groups has revealed a concerning trend of PII distribution, which poses significant risks to individuals whose data is compromised.

This study is a step toward understanding and tackling this threat. We noted that a small set of groups is responsible for a large amount of leaked data. The findings enable individuals and organizations to better prepare for potential attacks and develop strategies to disrupt them.

Building upon this initial study, future work will focus on two key areas. First, we will leverage this initial dataset to develop an automated methodology for detecting and crawling similar Telegram groups. This involves applying Natural Language Processing techniques and machine learning models to address the first part of our research question. Second, we will explore new metrics to assess different exposure risks depending on the data type and the specific threat (e.g., online scam vs. physical harm). These metrics serve as a more granular assessment of the impact of the leaks.

References

CETIC.br (2024). *Privacidade e Proteção de Dados Pessoais 2023: perspectivas de indivíduos, empresas e organizações públicas no Brasil.* Comitê Gestor da Internet no Brasil, São Paulo.

Folha de S.Paulo (2025). Financial scam losses surpassed $1.7 billion last year, says banking federation. `https://folha.com/me3qsvz9`. Last accessed 08 April 2025.

Fórum Brasileiro de Segurança Pública (2024). *Anuário Brasileiro de Segurança Pública 2024*. Fórum Brasileiro de Segurança Pública, São Paulo, 18 edition.

Garkava, T., Moneva, A., and Leukfeldt, E. R. (2024). Stolen data markets on telegram: a crime script analysis and situational crime prevention measures. *Trends in Organized Crime*.

Georgoulias, D., Yaben, R., and Vasilomanolakis, E. (2023). Cheaper than you thought? a dive into the darkweb market of cyber-crime products. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, ARES '23, New York, NY, USA. Association for Computing Machinery.

Júnior, M., Melo, P., Kansaon, D., Mafra, V., Sá, K., and Benevenuto, F. (2022). Telegram monitor: Monitoring brazilian political groups and channels on telegram. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 228–231. ACM.

Kayser, C. S., Back, S., and Toro-Alvarez, M. M. (2024). Identity theft: The importance of prosecuting on behalf of victims. *Laws*, 13(6).

Liu, Y., Lin, F. Y., Ahmad-Post, Z., Ebrahimi, M., Zhang, N., Hu, J. L., Xin, J., Li, W., and Chen, H. (2020). Identifying, collecting, and monitoring personally identifiable information: From the dark web to the surface web. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6.

Maia, L. R. H., Massarani, L., Santos, M. A. D., and Oliveira, T. (2024). Comunidades de pertencimento, desinformação e antagonismo: processos interacionais em grupos antivacina no telegram no brasil. *Galáxia (São Paulo)*, 49:e64635.

McCallister, E., Grance, T., and Scarfone, K. (2010). *Guide to protecting the confidentiality of personally identifiable information*. Diane Publishing. NIST Special Publication 800-122.

Santini, R. M., Salles, D., Mattos, B., Moreira, A., Mello, D., Haddad, J. G., Dias, B., Gomes, M., Dau, E., Borges, A., and Loureiro, F. (2025). Danos causados pela publicidade enganosa na meta: Anúncios fraudulentos promovem desinformação sobre o pix para lesar cidadãos brasileiros. NetLab – Laboratório de Estudos de Internet e Redes Sociais, Universidade Federal do Rio de Janeiro (UFRJ).