Topic-wise Exploration of the Telegram Group-verse

# Topic-wise Exploration of the Telegram Group-verse


Alessandro Perlo
Politecnico di Torino
Torino, Italy
alessandro.perlo@studenti.polito.it

Giordano Paoletti
Politecnico di Torino
Torino, Italy
giordano.paoletti@polito.it

Nikhil Jha
Politecnico di Torino
Torino, Italy
nikhil.jha@polito.it

Luca Vassio
Politecnico di Torino
Torino, Italy
luca.vassio@polito.it

Jussara Almeida
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
jussara@dcc.ufmg.br

Marco Mellia
Politecnico di Torino
Torino, Italy
marco.mellia@polito.it



## Abstract

Although Telegram is currently one of the most popular instant messaging apps in the world, previous studies have mainly focused on analysing discussions on specific angles and topics. In this paper, we present a broad analysis of publicly accessible groups that cover a wide range of discussions, including Education, Erotic, Politics, and Cryptocurrencies. How do people interact with different topic groups? Is there any common or peculiar behaviour? We engineer and offer an open-source tool to automate the collection of messages from Telegram groups, a non-straightforward problem. We use it to collect more than 51 million messages from 669 groups. Here, we present a first-of-its-kind, per-topic analysis, contrasting the users' activity patterns from different angles — the language, the presence of bots, the type and volume of shared media content, links to external platforms, etc. Our results confirm some anecdotal evidence, e.g., indications of spamming behaviour, and unveil some unexpected findings, e.g., the different sharing patterns of video and message length in groups of different topics. Our research provides a horizontal analysis of the public group in Telegram across various general topics, establishing a foundation for future studies that can delve deeper into user interactions and content dynamics within this unique messaging environment.



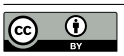
## CCS Concepts

• **Information systems** → **Social networks**; • **Human-centered computing** → **Social media**.

## Keywords

Telegram groups, Multimedia sharing, Topic characterization, User behavior

**ACM Reference Format:**
Alessandro Perlo, Giordano Paoletti, Nikhil Jha, Luca Vassio, Jussara Almeida, and Marco Mellia. 2025. Topic-wise Exploration of the Telegram Group-verse. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701716.3717506



*WWW Companion '25, Sydney, NSW, Australia*



ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717506


## 1 Introduction

Telegram has experienced remarkable growth in the past years, becoming one of the most popular instant messaging apps in the world. In July 2024, it surpassed the mark of 950 million monthly active users worldwide[1]. Telegram offers several features to its users, who can organize themselves into different spaces of communication such as private chats (one-to-one), groups (many-to-many) or channels (one-to-many).

Yet, the literature on Telegram is still limited in breadth. Targeting publicly accessible groups and channels, most prior works focused on *textual* content (e.g. news, hate speech), specific groups (e.g., terrorists [42]) or countries (e.g., Iran [20]), and a single topic of discussion (e.g., far-right politics [37]).

In contrast, this paper is driven by the following research questions: (i) how do users behave on Telegram groups in terms of the features they most often make use of to interact with others (e.g., video sharing, link sharing, use of reactions, polls, etc.)? (ii) How do such platform usage patterns change across groups discussing different topics?

We develop an open-source crawler designed to access public Telegram groups, collect historical messages up to a specified date, and continuously update the message collection over time, to build a longitudinal archive. We use the crawler to gather data from more than a thousand open groups and focus on those with at least 100 active users, distributed across 10 different topics of discussion including *Politics*, *Cryptocurrency*, *Darknet*, *Erotic*, *Video and Films*, etc. In total our data covers around 51.6 M messages and 1.4 M distinct users over a two-month observation period.

We analyse our data aiming to measure and contrast user activity patterns in groups across different topics. We analyse the mix of various languages, the footprint of official Telegram bots, the diverse habits in sharing media (e.g., videos, audios, images, GIFs) and external content via URLs. We overall witness peculiar behavioural patterns; some might be expected (e.g., the presence of automated user behaviour to programmatically share context) while others are more surprising (e.g., users in *Darknet* post much longer messages than users in the other topics).

This paper is structured as follows. Section 2 lists related work. in Section 3, we present the architecture of our crawler. In Section 4, we detail the data collection process and present a first-of-its-kind characterisation of platform usage across different discussion topics. Section 5 further elaborates on the analysis of video content

[1]Telegram CEO Pavel Durov https://t.me/durov/337

and URLs sharing patterns, together with a deeper analysis on the time elapsed between a YouTube video publication and its first appearance in Telegram groups under observation. Last, we draw conclusions in Section 6 and discuss ethics of the work in Section 7.

All in all, this work shows the heterogeneity of usages people exhibit on Telegram, corroborating known facts and exposing surprising findings. We hope this work stimulates other works in exploring some of the highlighted results. For this, we make both the data and the crawler open to the community.[2]

## 2 Related work

The characteristics and dynamics of messaging platforms have attracted a lot of attention. Notably, prior studies analysed content properties [26, 32, 33] and information spread [15, 31] on WhatsApp's groups, hinting at the catalytic role of the platform in various real-world events [14, 18, 25].

More recently, attention has been dedicated to groups and channels on Telegram, as the platform's popularity increases across the globe. Some studies were interested in the inner workings of the mobile application [11, 35], and its use by particular user populations, such as Iranian immigrants [30], terrorist organizations [10, 42], extremist groups [21, 24], or particular countries (e.g., Iran and Russia [9, 19, 20]). Others analysed the formation of communities within Telegram channels [36, 37] and their connection to information spread [16, 22, 38]. Some other efforts studied content properties, limiting to textual content and usage patterns, with attention given to news content [29], hate speech and abusive language [40], as well as the presence of fake channels (i.e., those impersonating important services or persons) [28]. The use of Telegram to perform illicit activities (e.g., pump-and-dump activities in cryptocurrency markets [41], manipulation of social media popularity [39]) has also been previously addressed. Overall, previous works focused on the information people exchange on Telegram, in groups and channels of a specific topic. Only Morgia *et al.* [28] used TGStat to gather channels associated with multiple topics. Yet, they did not distinguish between such topics and aggregated all of them to discover fake channels.

In contrast, we here offer a topic-wise analysis of common user activities in Telegram groups. Rather than focusing on the type of information exchanged, or how it spreads, we show how people leverage different features (e.g., media types, links to external sites) to interact with each other, and how such patterns differ depending on the topics and goals of group discussion.

## 3 Crawler and data collection

Given our interest in *user* behaviour, we focus our data collection effort on Telegram *public groups*, i.e., public chats where all the members can send messages.

To collect the data, we design an open-source, two-stage crawler that we offer to the community.[2] At the first stage, the tool periodically crawls the TGStat website to discover public Telegram groups on various topics. At the second stage, the tool crawls Telegram by joining the discovered groups and collecting all messages.

[2] The code and data are available at https://anonymous.4open.science/r/TopicWiseTelegram-7A81

### 3.1 TGStat crawling

TGStat is a service that catalogues popular Telegram groups and channels worldwide. Currently, TGStat's database covers almost 1.9 M channels and groups [1], which are categorised into 48 predefined topics. For each topic, TGStat shows the lists of the top-100 groups according to various metrics. These lists are continuously and dynamically updated. TGStat characterises each group by some metadata, including the group name, topic, language, and the monthly Active Users (AU), i.e., the number of unique users who have shared messages in the group in the past month.

We extract information from TGStat engineering a Python-based crawler using the BeautifulSoup package [2]. We periodically run it to automatically extract the lists of groups in various topics. This allows us to grow the group lists in those topics of interest to us (see discussion in Section 4.1).

Most prior studies of Telegram searched for links to existing groups in social media, news and even word-of-mouth [13, 16, 21]. This approach can demand extensive crawling, particularly for diverse topics. TGStat streamlines this process by offering categorized groups with real-time activity metrics, facilitating targeted selection and per-topic analysis. Prior studies [28, 36, 37] leveraged TGStat but relied on static snapshots. In contrast, we continuously expand our dataset over multiple days. Furthermore, given our focus on per-topic analysis, we assess the reliability of TGStat's categorization—a crucial step overlooked in past work (see Section 4.1.2).

### 3.2 Telegram crawling

Given a list of previously discovered groups, our crawler automatises the group join and message collection tasks. We rely on the Telethon Python package [7] and design a scalable tool based on threads: a master instructs workers to *join* (and leave, if desired) a group, *check* if a pending request for join has been accepted, *collect* new messages, or just *wait*. For scalability, we use multiple Telegram IDs, each associated with a worker. The master keeps a list of groups to collect messages from and instructs workers to do so from a desired initial date until the present. We store the collected information in a MongoDB database for later processing.

We instrument our crawler to join and stay in groups as we discover them on TGStat. To refresh the collection of messages, workers download only the new messages since the last retrieved snapshot. For every group and message, the crawler stores all the returned information in JSON format in the MongoDB instance. In this paper, we focus on the following message information: sender user's identifier, message body, message time, and media contained in the message (image, video, GIF, poll, etc.).

### 3.3 Crawler design challenges

Telegram implements several countermeasures to avoid API abuse, notably: i) a limit of 500 groups a given Telegram ID can join; ii) an unspecified upper limit on the rate to join new groups which, if not respected, causes a lengthy temporary ban [5]; iii) a without-any-notice permanent ban of novel-activated Telegram IDs that start to interact with the platform with high frequency. Respecting these limitations requires ingenuity when designing the crawler. First, we declared our intentions to the official Telegram support channel. Second, we carefully controlled the group joining rate to limit the

temporal ban. Third, we used multiple already-active Telegram IDs, each associated with a worker thread to scale the data gathering.

Telegram offers the possibility of setting up administration bots (known as *Telegram bots*) that ease group management. Captcha protection bots are popular for filtering fake user bots, i.e., actual Telegram accounts used to programmatically spam messages in open groups. Such captcha protection bots may kick users out if they do not solve the captcha after a specific time. Other bots or administrators might enforce different rules or criteria for group participation. Whenever we were removed from a group, we respected the administrators' willingness and did not try to join the group again. Similarly, to respect the privacy indications of the group administrators, we only consider groups where the participants' messages are persistent and not automatically removed ("auto-delete" functionality removes messages 24 hours or 7 days after sending).

Our crawler can collect up to thousands of messages per second per worker and join tens of groups per hour without overcoming Telegram rate limitations.

# 4 Topic Characterisation

## 4.1 Data collection and filtering

*4.1.1 Topic selection.* On April 1st, 2024 we collected the top-100 groups for all TGStat 48 topics. Out of these, we select the 12 topics in which we were able to join at least 10 English language groups, a condition that allows us to manually validate the accuracy of TGStat's topic labelling. We keep crawling TGStat every week to refresh the lists of top-100 groups for these 12 topics to observe how those lists change over time. We stop on May 1st, 2024, discovering 1,368 groups in total. The growth in the number of groups discovered over time is notable in specific topics such as *Erotics*, *Cryptocurrencies* and *Bookmaking*, where we find around 20% new groups every week. This illustrates that taking a single snapshot from TGStat, as prior work [28, 36, 37], would limit the lists of discovered groups.

Feeding these growing lists to the Telegram crawler, we find that 8.6% of groups have the auto-delete function enabled. We abandon them immediately. We also fail to join 18.8% of groups because the group (i) did not exist anymore, (ii) changed their name before we could join it, or (iii) are moderated and either the administrator did not admit us or a bot kicked us out after joining. We thus successfully join and monitor 993 groups. For each tracked group, we collect all messages starting from March 1st to April 30th.[3] In total, we collect over 50 M messages, with about 1 M new messages gathered each day.

*4.1.2 TGStat topic verification.* Next, we verify if the per-topic categorisation provided by TGstat is reliable. To that end, we check if the *actual* topic of discussion is (i) coherent with the topic assigned by TGStat and (ii) consistent across time. We pick all English groups (206) and for each group, we select three sets of 30 consecutive messages, each set separated by the others by ten days. The 206 groups where randomly assigned to three human evaluators, each required to independently assess whether each set of messages was, collectively, addressing a subject that was consistent with the topic of the group. To measure the agreement among evaluators, we use the Fleiss' Kappa [17]. The labelling showed an agreement of 0.71, indicating substantial agreement between evaluators.

TGStat's topic assignment proves mostly correct, with two exceptions: the *Courses and guides* groups are mostly filled with spam; and the *Economics* groups mostly host discussions about cryptocurrencies, for which a dedicated topic already exists. We thus discard the 166 groups in these two topics, ending up with 827 groups.

*4.1.3 Selecting active groups.* To guarantee groups are active and with enough diversity, we keep groups that have at least 100 active users, i.e. users who sent at least one message in the two-month observation period. From the 827 groups, we discard 158 of them, ending with 669 groups, as detailed in column 3 of Table 1. For comparison, the total number of groups discovered in TGStat in each topic is shown in column 2. The table also details the average number of users per group, the percentage of active ones, the total number of messages and the average number of messages per active user. Figures vary widely, showing already very different interests, engagement and activity levels across the topics.

For the sake of completeness, Figure 10 in the Appendix A details the breakdown of the various cases we faced when trying to collect messages from Telegram. Depending on the topic, we observe various failure cases which may significantly reduce the number of groups to follow.

**Table 1: Dataset statistics for verified topics. Column 2 refers to groups discovered in TGStat, all others refer to joined groups with at least 100 Active Users (AU) with consistent topic.**

| Topic | # Groups | | Average # Users | | # Messages | |
|---|---|---|---|---|---|---|
| | TGStat | Joined | Memb. (k) | AU (%) | Tot. (M) | Per AU |
| *Education* | 115 | 98 | 28.7 | 8.5 | 4.8 | 21.1 |
| *Bookmaking* | 120 | 91 | 16.6 | 13.1 | 9.0 | 47.2 |
| *Crypto* | 123 | 80 | 69.6 | 9.0 | 10.3 | 22.6 |
| *Technologies* | 108 | 67 | 27.0 | 7.8 | 6.0 | 43.0 |
| *Darknet* | 112 | 62 | 11.0 | 19.2 | 5.2 | 41.9 |
| *Software* | 114 | 61 | 14.7 | 7.7 | 3.8 | 55.8 |
| *Video&films* | 114 | 59 | 13.1 | 9.2 | 2.4 | 34.8 |
| *Politics* | 115 | 58 | 4.9 | 26.0 | 4.2 | 57.1 |
| *Erotic* | 124 | 52 | 22.5 | 7.7 | 4.6 | 56.1 |
| *Linguistics* | 109 | 40 | 6.7 | 8.0 | 1.4 | 67.3 |
| Total | 1,154[4] | 669 | 23.7 | 9.7 | 51.7 | 36.4 |

## 4.2 Per-Topic Characterisation

We characterize our dataset by extracting various features from each group and then aggregating them on a per-topic basis (i.e., per-group macro average). This allows us to avoid the bias induced by large groups. Our goal is to explore how differently users interact on each topic.

*4.2.1 Telegram Bot usage and user activity.* Telegram group administrators can incorporate bots[5] into their groups, offering a

[3] For this work, we limit the collection period to avoid overloading the Telegram servers.

[4] The total number of groups adds up to 1,368 if we consider 214 groups from *Courses and guides* and *Economics* that we initially considered and later discarded.

[5] Note that we use the term *bot* to refer exclusively to official bots and not to users exhibiting bot-like automated behaviour.

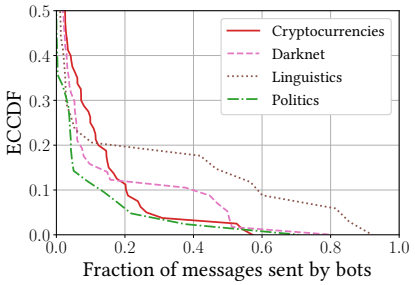



**Figure 1: Fraction of messages sent by bots in a group across topics.**

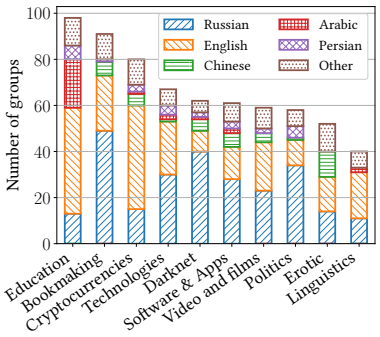



**Figure 2: Most popular language in each group.**

wide set of functionalities, from welcoming newcomers with group rules to responding to user commands, from collecting statistics to moderating join requests and messages. Telegram bots are quite popular: only 10.2% of groups in our dataset do not include bots; in the median, there are four bots per group. Interestingly, 1.5% of groups have 20 bots (the maximum allowed by Telegram). Some bots enjoy significant popularity: *Combot* [4] and *MissRose_bot* [6] are present in 145 and 129 groups, respectively. Both provide moderation services, analytics, and anti-spam features.

Bots footprint is not negligible: they generate on average 8.6% of messages, with notable variations across topics. Figure 1 shows, for 4 topics, the Empirical Complementary Cumulative Distribution Function (ECCDF) of the fraction of messages sent by bots for different groups. The largest fractions of messages sent by bots are in *Linguistics* groups. Some of these bots are integral to learning platforms and merit examinations. For instance *Quizbot* [8] is widely deployed in *Linguistics* and *Education* groups (27.5% and 28.6%, respectively). It generates 39.0% and 13.0% of messages. In one *Linguistics* group it generates 91% of messages. Conversely, *Politics* groups see the smallest fraction of messages generated by bots (green dotted line), possibly testifying to a higher user engagement[6] in political groups than in other topics. Curiously, there is a quite large fraction of groups with bots that simply collect statistics or moderate the group without sending any message (leftmost part of Figure 1). For the remainder of our analysis, we remove messages sent by Telegram bots.

Focusing on the number of messages actual users generate, we observe that there are a few users sending thousands of messages, while the majority are not active (see column 4 in Table 1) or send few messages. Indeed, the Empirical Probability Density Function (EPDF) of the number of messages generated by users follows a heavy-tailed shape that can be fit by a Pareto distribution with $\alpha = 1.9$. Remarkably, the fittings of the per-topic EPDFs are very similar, suggesting the universality of these behaviours, as widely acknowledged in the literature[12, 34].

***Main takeaways:*** *official bots' impact on the activity of Telegram is significant (8.6% of messages), with some moderation bots being present on a wide variety of groups. Some topics are more prone to bots offering also specific features (e.g.,* QuizBot *generating 39% of messages in* Linguistics *' groups.*

[6]For messages sent by a user account, we cannot distinguish between messages sent by a human or by an automated system.

*4.2.2 Language.* The next question we answer is what are the languages people use in each topic and group. For each *textual* message in a group, we associate a language by employing the FastText language identification library [23] to obtain the distribution of languages for each group. In Figure 2 we report the breakdown of the *most popular* language for groups on the same topic:

• English (a global language) and Russian (being Telegram very popular in Russia) are the two most popular languages. Their share changes based on the topic. For instance, most groups in *Bookmaking* and *Darknet* have a lot of Russian groups. Conversely, the majority of *Education* and *Cryptocurrency* groups are in English, possibly due to the worldwide interest in such topics.

• Despite Telegram's restricted use in Iran, some popular groups are in Persian, mostly in *Politics*. This is in line with the claim that Telegram is used to evade state censorship in Iran [9].

• Although Telegram is blocked in China [3], we do find groups with Chinese as the dominant language (up to 21% of *Erotic* groups).

We observe that in almost half of the groups, more than 75% of the messages are written in the same dominant language. Still, we observe messages written in other languages, hinting at a global user population. *Linguistics* is the topic where groups contain the largest mix of languages, which supports the anecdotal observation of users practising foreign languages and mixing messages in their native language in such groups. In contrast, both *Darknet* and *Technology* stand out with more than 58% of the groups having more than 80% of the messages in a single language.

***Main takeaways:*** *English and Russian are the most widely used languages, dominant in 34.1% and 38.4% of groups, respectively. Notably, Persian (3.9%) and Chinese (6.0%) also appear frequently, despite Telegram restrictions in Iran and China. Many groups mix multiple languages within their discussions.*

*4.2.3 Message length.* Figure 3 shows the ECCDF of the length of *textual* messages for each topic. Since message length may be influenced by language, we consider only messages written in English-dominated groups. We observe some great distinctions across topics. On one hand, *Darknet* groups are dominated by very long messages (80% longer than 100 characters). A manual check unveils that most messages contain samples of illicit content people trade, such as bank account credentials, credit cards, passports and SSNs. Conversely, in *Bookmaking* and *Technologies* groups, long messages describe bookmaking websites, betting experiences and results, or

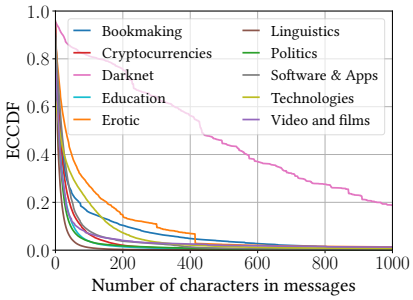



**Figure 3: ECCDF of the number of characters in text messages grouped by topic (groups in English).**

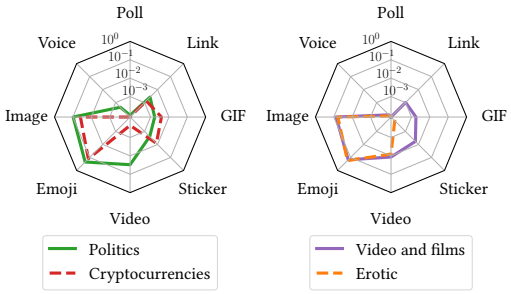



**Figure 4: Median fraction of messages with non-textual elements in selected topics.**

**Table 2: Per-topic detailed metrics of usage of non-textual elements in messages.**

| Topic | Poll (%) | Voice (%) | Image (%) | Emoji (%) | Video (%) | Sticker (%) | GIF (%) | Link (%) |
|---|---|---|---|---|---|---|---|---|
| *Education* | 0.01 | 0.02 | 5.13 | 12.43 | 0.04 | 0.01 | <0.01 | 0.06 |
| *Bookmaking* | <0.01 | 0.03 | 8.67 | 19.81 | 0.40 | 0.72 | 0.38 | 0.04 |
| *Crypto* | <0.01 | <0.01 | 3.89 | 12.54 | 0.04 | 0.81 | 0.39 | 0.14 |
| *Technologies* | <0.01 | <0.01 | 6.12 | 8.22 | 0.10 | 0.09 | 0.05 | 0.29 |
| *Darknet* | <0.01 | <0.01 | 5.44 | 56.51 | 0.34 | <0.01 | 0.03 | 0.02 |
| *Software* | <0.01 | <0.01 | 6.50 | 9.61 | 0.42 | 0.42 | 0.10 | 0.36 |
| *Video&Films* | <0.01 | 0.02 | 7.61 | 15.10 | 1.15 | 0.58 | 0.17 | 0.10 |
| *Politics* | <0.01 | 0.07 | 9.69 | 22.17 | 3.03 | 0.25 | 0.17 | 0.26 |
| *Erotic* | <0.01 | <0.01 | 6.05 | 17.29 | 0.80 | 0.03 | 0.02 | <0.01 |
| *Linguistics* | 0.10 | 0.08 | 2.34 | 15.34 | 0.33 | 0.16 | <0.01 | 0.04 |
| *All Topics* | <0.01 | <0.01 | 6.02 | 14.65 | 0.23 | 0.23 | 0.09 | 0.07 |

offer details about devices for sale. In *Erotic* groups, long messages are used to advertise services and sell content, some of them repeated multiple times (observe the steps in the ECCDF). On the other hand, *Linguistics* and *Politics* groups are dominated by very short messages in which people debate or provide quick suggestions (80% shorter than 30 characters). At last, steps in the distribution might suggest the presence of repeated automated messages (mostly spam), further analysed in Section 5.3.

***Main takeaways:*** *There appears to be a correlation between message length and promotional content, with longer messages linked to advertisements, while discussion-driven groups to shorter messages (e.g., compare Bookmaking and Linguistics in Figure 3).*

*4.2.4 Usage of non-textual elements.* We now broaden our analysis to consider non-textual elements. Specifically, we extract, for each group, the fraction of messages containing images, external links, voice messages, polls, GIFs, stickers, videos, and emojis. How are these elements used? To gauge this, we compute the fraction of messages with such element in a group and compute the average over all groups of a given topic (macro average). Figure 4 visually compares two pairs of selected topics using radar charts. Table 2 and Figure 11 in Appendix A provides the complete set of results. Some interesting findings emerge:

• *Politics* groups represent the typical usage of non-textual elements: 20–30% of messages contain emojis; 10% of messages share an image; stickers are more popular than GIFs; few messages contain voice content; polls are mostly an unused feature (polls are in fact only used in *Linguistics* groups).

• *Cryptocurrencies* groups represent some mixed usage: no videos and voice messages, fewer photos and emojis but more stickers and GIFs than average.

• Groups in *Video and Films* and *Politics* have very similar usage patterns (i.e. radar shape), though, surprisingly, the former has fewer videos (which are present in only 1% of the messages) — see Section 5.

• *Erotic* groups present the minimum usage of non-textual elements: no stickers, no GIFs, while photos are found in 6% of messages. Surprisingly, we see very few links to external platforms. In fact, readers are invited to contact advertisers via private chat.

Notice that sending stickers requires manual actions hardly automatizable. Also, stickers are commonly used as reactions to other messages. Their prominent use in a group or topic may testify to a larger fraction of messages being sent by real users, or to a more confidential exchange: *Erotic*, *Darknet* and *Technologies* have the least fraction of stickers and are dominated by ad-style messages. Conversely, in *Bookmaking* and *Cryptocurrencies* people use more stickers as reactions for suggestions. This argument was highlighted on other platforms such as WhatsApp where, for instance, in political groups, stickers are rarely forwarded, suggesting that users often save and maintain their collection of preferred stickers for future use rather than relying on the platform's sharing tools [27], further highlighting a more human-based interaction pattern.

***Main takeaways:*** *the media-sharing patterns across topics show big differences based on the topic under observation. No common habits emerge.*

## 5 Multimedia and external links

We now delve deeper into the usage of specific non-textual elements, notably shared videos and URLs to external sites.

### 5.1 Video size and duration

We focus our analysis on the three topics with the largest share of videos: *Politics*, *Video and Films*, and *Erotic*. Although Figure 4 suggests some similarities in the amount of videos people directly share on Telegram, a closer examination of the video duration and file size reveals noteworthy differences in goals and types of shared videos, again showcasing notable differences in users' behavioural patterns among different topics.

The left plot of Figure 5 compares the video duration, by showing the fraction of videos by length. Observe how videos shared in

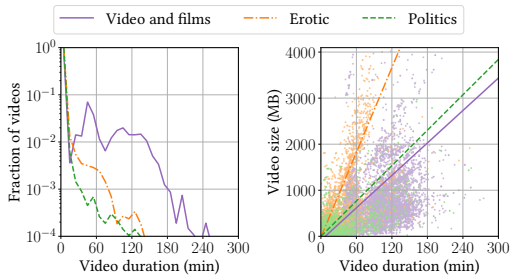



**Figure 5: Distribution of the video duration (on the left, bin every 10 minutes) and comparison between video size and video duration (on the right, with regression line reported).**

*Video and Films* are notably longer than in other topics, with peaks roughly around the 60- and 120-minute marks (notice the log-y scale). Manual inspection confirms that people share entire TV series episodes and movies. Their total volume amounts to 7.45 TB of data. *Erotic* and *Politics* tend to have much shorter videos.

Compare now this video duration with its file size — see the right graph of Figure 5, a scatter plot with regression lines, where each point represents a shared video. Again, we observe a noticeable difference between topics: videos shared in *Erotic* groups have significantly higher video rate (thus quality), despite shorter durations, as evidenced by the steeper slope of the regression line. In contrast, videos shared in *Politics* groups tend to be shorter, with limited attention given to production quality. Their primary function is to serve immediate needs, such as delivering brief announcements, political speeches, or viral clips intended to shape opinions. A sample of these videos reveals content related to the Russian-Ukrainian conflict, speeches by politicians, and clips with political undertones designed to provoke indignation.

***Main takeaways:*** *video quality (Figure 5, left) and duration (Figure 5, right) vary based on the goal of sharing such a video, from entire movies to a few seconds clips to shape the debate.*

## 5.2 Links to external domains

We now extend the analysis to the sharing of links to external websites (i.e., different from telegram.me and t.me) to assess how Telegram users wish to redirect (or drive) attention to other websites, and the interplay with external Web resources .

In Figure 6, we present the average frequency at which a domain appears in a given topic. We focus on the union of the 5 most popular domains across each topic. These cover from ≈20% to ≈50% of links — with a heavy-tailed presence of other platforms (bottom row). The three most frequent platforms are social networks: X, YouTube and Instagram. Usage varies a lot: X sees significant use in *Cryptocurrency* groups; YouTube and, to a lesser extent, Instagram are the two most transversal platforms. Conversely, in line with their discussion topics, *Technology* and *Software and Application* groups show a significant usage of GitHub pages. The telegra.ph, an open and anonymous publishing platform, is very often used in *Software & Applications* groups, e.g., to share anonymously installation guides and tutorials. All in all, we observe a very diverse

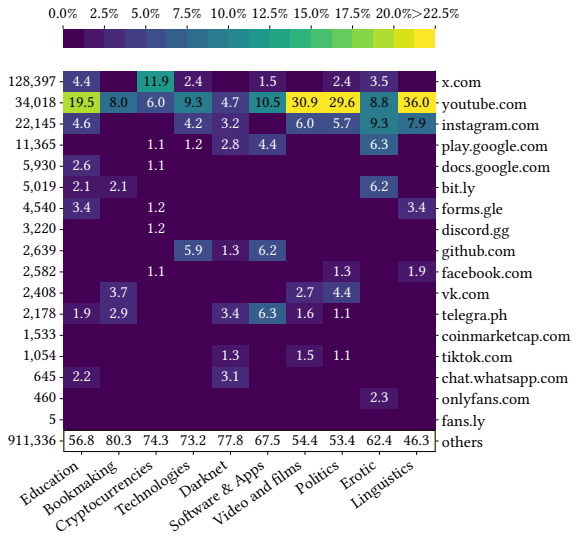



**Figure 6: Per-topic average of the percentage of linked domains within groups. On the left, the number of times a domain appears.**

sharing of information from extensive platforms, with each topic having different preferential means to refer to external outlets.

Finally, we expand `bit.ly` URLs, finding that only 4.2% pointed to popular domains (i.e., those in the union of the five most frequent domains per topic), while 67.4% led to less common websites with few occurrences. Additionally, 28.4% of these URLs directed to expired domains. As `bit.ly` URLs represent a small fraction of the total, their expansion minimally impacts domain rankings but rather highlights the use of URL shorteners as a noteworthy behavior.

***Main takeaways:*** *YouTube consistently serves as the most prevalent domain across all topics to share external content (largest share of links across each but one topic). A notable exception is the cryptocurrency discussions, where X takes precedence (11.9% vs. 6%). Some platforms are popular in some topics, with again few common habits (except YouTube, see Figure 6).*

## 5.3 Repeated sharing behavior

Delving into the analysis of the links to external domains, we now look at possible patterns from the repeated sharing of the same URL. In total, there are around 239,000 unique URLs in our dataset, around 43,000 (18%) of which are shared multiple times during the period of analysis. We here focus on these duplicated URLs.

In Figure 7 we report a scatter plot where every dot represents the sharing of a given URL (y-axis) at a specific timestamp (x-axis). We sort the unique URLs by their first appearances in our data. We distinguish if the URL is shared multiple times by the same sender (*old sender*) or a new one (*new sender*), and if the URL is shared multiple times in the same group (*old group*) or in a new one (*new group*). A sender is *new* when they first send a given URL to a given group. If they send the same URL to the same group again, they become *old*. If they send the same URL to another group for the first time, they are *new* again.

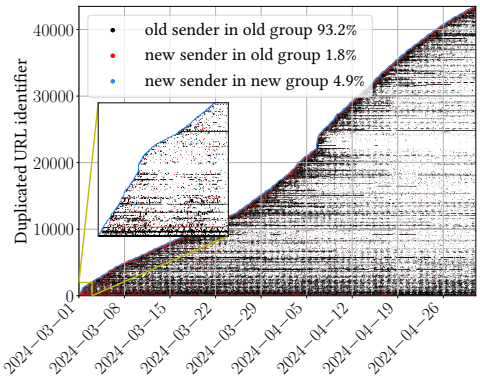



**Figure 7: Cumulative time series of unique URLs appearing at least twice in the dataset, along with all their subsequent repetitions. The zoom highlights the first two days.**

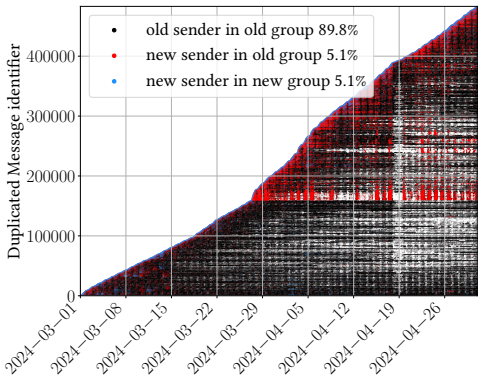



**Figure 8: Cumulative time series of unique long messages (over 50 characters) repeated at least twice in the dataset, including all subsequent identical reposts.**

We find that the large majority (93.2%) of all duplicated URLs correspond to users posting the same URLs in groups they have already posted the URLs in. This hints at a spamming activity by regular accounts that programmatically share the same URL with the same audience, to maximise the reach of the content. Most of these repeated URLs appear in groups about *Bookmaking* (54%), *Cryptocurrencies* (21%) and *Darknet* (14%) topics. These groups are in fact full of ads and spam messages.

By manually analysing the most replicated URLs, we observe that many of them advertise Social Media Managing services, which promise to boost one's interactions (followers, likes,etc.) through the use of likely fake accounts. Curiously, the most repeated URL appears more than 27,000 times in a single group over two months.

Our analysis reveals also suspicious patterns. The red line in the inset of Figure 7 shows an uninterrupted sequence of messages with the same URL sent within the same *Bookmaking* group by 1,498 distinct users. This suggests a coordinated network of likely fake users promoting a betting website. In the first week of April, we also observe a sudden burst of shared URLs that are shared for a few days. These URLs refer to Reddit comments or posts arguably sent by their author who asks for "upvotes" to other members of a Cryptocurrencies group, suggesting an attempt to manipulate content curation algorithms as shown in [39].

After detecting spamming behaviour in URL sharing, we investigate whether similar patterns occur with textual messages. To filter out trivial repetitions (e.g., "ok" or "good morning"), we focuse on messages longer than 50 characters and analyse perfect duplicates with identical content. As shown in Figure 8, duplication is common even for long messages, with nearly 500,000 messages repeated at least once across 97.8% of the groups, accounting for 65% of all long messages. Unlike URLs, duplicated messages are more evenly spread across topics, with the *Darknet* topic showing the highest repetition (19%) and *Linguistics*, *Video and Films*, and *Education* having the fewest (under 1%).

Although Telegram allows message forwarding, only 13.7% of these repeated messages are forwarded; most are posted directly by the same users. Similar to URL patterns, most duplicated messages come from users repeatedly posting the same content in the same group, further indicating spamming behavior.

Suddenly, starting from the last week of March 2024, a great portion of duplicated messages is reposted in the same group by users who had not previously shared them (*new senders old group*, red pattern). Interestingly, this phenomenon happens in a single group. This group ("`kadyrov_95chat`") contributes to 4.3% of all the repeated messages and 84% of all the repeated messages sent by different users in the same group. Predominantly in Chechen, messages often include congratulations, generic wishes, and praise for the Chechen Republic and its government. We conjecture that this behaviour is caused by some sort of spam bot networks, which tend to copy-paste the same message multiple times. However, we acknowledge that we could not retrieve messages sent before March 27$^{th}$, leaving it unclear whether the group was previously inactive or if earlier messages were deleted before our data collection began. This highlights the importance of tracking a group activity over time, to retrieve its historical messages.

***Main takeaways:*** *we observe indications of spamming behaviour in the sharing of both URLs (Figure 7) and textual messages (Figure 8). Most repeated content comes from the same single users (93.2% and 89.8% respectively). Yet, evidence of coordinated group of spammers clearly emerges (e.g., red line in the inset of Figure 7). This should encourage further research to characterise this behaviour and its impact on user participation and group dynamics.*

## 5.4 Time Lag in YouTube Video Sharing

As shown in Figure 6, youtube.com is the most consistent domain shared on Telegram groups across topics. It accounts for the 8.1% of the unique links shared in our dataset. For each posted video link, we collect its publication timestamp on YouTube through the *YouTube Data API v3*.[7] Then, we measure the elapsed time to its first appearance in all of our Telegram groups.[8] In Figure 9, we present the ECCDF for the topics with the most diverse distributions. Considering videos shared in *Education* groups, more than

[7] https://developers.google.com/youtube/v3?hl=en

[8] If appearing in multiple groups, we evaluate the difference multiple times.

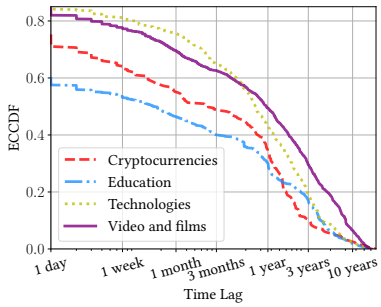



**Figure 9: Elapsed time between a video publication on YouTube and its first appearance in Telegram groups, considering 4 topics (log x-scale).**

40% of them are shared on Telegram less than 24 hours after their publication on YouTube. This suggests a community interested in fresh content. Most videos consist of classes and tutorials offering guidance on how to approach specific exams. The *Video and films* topic, in turn, exhibits a great interest in old videos: more than 30% of the videos have been on YouTube for almost three years (mostly old movies). This indicates that this topic is not recency-bound.

In the *Technologies* groups, we observe a limited interest in new videos, with only 20% of shared content appearing within 10 days of publication. This is surprising, as one would expect a high level of engagement from the community regarding the latest developments in the field. One explanation lies in the predominance of video tutorials on technology topics, which retain their relevance over time. In fact, most URLs refer to educational content or guides in the technology field, while some present consumer technology reviews.

Finally, it is noteworthy to focus on the knee of the ECCDF of videos shared in the Cryptocurrencies groups: here videos older than one year are less and less shared. This can be due to the variety of Cryptocurrencies and market-related suggestion videos, whose content naturally loses interest over time.

***Main takeaways:*** *the distribution of the elapsed time between publication on YouTube and sharing on Telegram (Figure 9) again showcases context-based patterns that depend on the communities interacting on the platform.*

## 6 Conclusion

In this paper, we conducted a comprehensive, topic-wise analysis of Telegram public groups retrieved from TGStat, providing a transversal view of platform usage across diverse discussion topics. Our work contributes to the understanding of how Telegram is used across various domains, shedding light on usage patterns that had not been explored in prior studies. Furthermore, we develop an open-source crawler and dataset that contributes valuable tools for future researchers to explore Telegram usage.

Our analysis reveals a significant heterogeneity in how different communities leverage the platform's capabilities. Bots, for instance, play a crucial role in many groups, with some topics — such as *Linguistics* and *Education*—showing up to 80–90% of messages generated by bots. This highlights the importance of automated interaction in certain topics, where bots are integral to group functionality. Conversely, in *Politics* groups, bot activity is notably lower, suggesting higher levels of direct user engagement in discussions.

Media sharing varies significantly across topics. In *Politics* groups, videos are short and focused on quick messages or viral clips, while in *Video and Films* groups, entire movies and TV shows are shared, contributing to a larger data volume. *Erotic* groups prioritize high-resolution videos despite shorter durations. Non-textual elements like stickers, emojis, and links also differ, reflecting diverse community dynamics and interaction styles.

In terms of user activity, the length of user-generated messages differs greatly. In *Darknet* groups, messages are particularly long, often containing detailed information about illicit goods and services, while in discussion-driven groups like *Politics* and *Linguistics* more concise messages prevail. Moreover, we identified substantial repeated sharing behaviour, particularly in *Bookmaking*, *Cryptocurrency*, and *Darknet* groups, where spamming and coordinated actions (e.g., repeated posting of URLs or messages by different users) are common. We highlighted the presence of bot-like networks or coordinated manipulation efforts within specific communities.

Overall, our study offers a broad, multi-faceted exploration of Telegram group usage that contrasts sharply with previous studies, which were more limited in scope. The diversity of patterns we uncovered demonstrates how versatile and complex Telegram usage can be, depending on the topic at hand.

Future research directions include identifying and quantifying the key factors that drive user engagement and examining the processes of influence and information dissemination. Additionally, further analysis of content through NLP techniques could shed light on shifts in discussion subjects.

## 7 Ethical aspects

In our work, we take ethics under utmost consideration.

- Our data extraction rate on TGStat is below 5 pages per minute, to minimise the load, and we repeat the crawling only once a week.
- We contacted Telegram's support to declare our intention, asking them to share with us any limitation. We received no answer. The privacy policy of the platform does not forbid crawling.
- To respect privacy restrictions imposed by group administrators, we restrict our analysis to public groups.We do not monitor groups where the administrator sets the auto-delete functionality and those where admins refuse or ignore our join request.
- Telegram might be used to share copyright-protected material and illicit content. We avoid storing any non-textual element, only collecting metadata (e.g., video duration and size).

## Acknowledgements

This work has been partially supported by the Spoke 1 "FutureHPC & BigData" of ICSC — Centro Nazionale di Ricerca in High-Performance-Computing, Big Data and Quantum Computing, funded by European Union — NextGenerationEU. It has also been partially supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), both in Brazil.

## References


[1] 2024. API for Telegram channels statistics :: TGStat Statistics API :: TGStat. https://tgstat.ru/api/stat, accessed on February 25, 2025.

[2] 2024. beautifulsoup4 · PyPI. https://pypi.org/project/beautifulsoup4/, accessed on February 25, 2025.

[3] 2024. Censorship of Telegram - Wikipedia. https://en.wikipedia.org/wiki/Censorship_of_Telegram#China, accessed on February 25, 2025.

[4] 2024. Combot. https://combot.org/, accessed on February 25, 2025.

[5] 2024. Error handling. https://core.telegram.org/api/errors, accessed on February 25, 2025.

[6] 2024. Home - MissRose. https://missrose.org/, accessed on February 25, 2025.

[7] 2024. LonamiWebs/Telethon: Pure Python 3 MTProto API Telegram client library, for bots too! https://github.com/LonamiWebs/Telethon/, accessed on February 25, 2025.

[8] 2024. Quizzes. https://telegram.org/tour/quizbot/, accessed on February 25, 2025.

[9] Azadeh Akbari and Rashid Gabdulhakov. 2019. Platform Surveillance and Resistance in Iran and Russia: The Case of Telegram. *Surveillance & Society* 17, 1/2 (March 2019), 223–231. doi:10.24908/ss.v17i1/2.12928

[10] Abdullah Alrhmoun, Charlie Winter, and János Kertész. 2024. Automating Terror: The Role and Impact of Telegram Bots in the Islamic State's Online Ecosystem. *Terrorism and Political Violence* 36, 4 (May 2024), 409–424. doi:10.1080/09546553.2023.2169141

[11] Cosimo Anglano, Massimo Canonico, and Marco Guazzone. 2017. Forensic analysis of Telegram Messenger on Android smartphones. *Digital Investigation* 23 (Dec. 2017), 31–49. doi:10.1016/j.diin.2017.09.002

[12] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207–211.

[13] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Telegram Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 840–847. doi:10.1609/icwsm.v14i1.7348

[14] Jeremy Bowles, Horacio Larreguy, and Shelley Liu. 2020. Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe. *PLOS ONE*, 1–11.

[15] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T Marques-Neto, and Virgílio Almeida. 2019. Characterizing attention cascades in whatsapp groups. In *10th ACM Conference on Web Science*. 27–36.

[16] Arash Dargahi Nobari, Malikeh Haj Khan Mirzaye Sarraf, Mahmood Neshati, and Farnaz Erfanian Daneshvar. 2021. Characteristics of viral messages on Telegram; The world's largest hybrid public and private messenger. *Expert Systems with Applications* 168 (April 2021), 114303. doi:10.1016/j.eswa.2020.114303

[17] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions* 2, 212-236 (1981), 22–23.

[18] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *The Web Conference*.

[19] Hans W. A. Hanley and Zakir Durumeric. 2024. Partial Mobilization: Tracking Multilingual Information Flows amongst Russian Media Outlets and Telegram. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 528–541. doi:10.1609/icwsm.v18i1.31332

[20] Ali Hashemi and Mohammad Ali Zare Chahooki. 2019. Telegram group quality measurement by user behavior analysis. *Social Network Analysis and Mining* 9, 1 (July 2019), 33. doi:10.1007/s13278-019-0575-9

[21] Mohamad Hoseini, Philipe Melo, Fabricio Benevenuto, Anja Feldmann, and Savvas Zannettou. 2023. On the Globalization of the QAnon Conspiracy Theory Through Telegram. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 75–85. doi:10.1145/3578503.3583603

[22] Mohamad Hoseini, Philipe de Freitas Melo, Fabrício Benevenuto, Anja Feldmann, and Savvas Zannettou. 2024. Characterizing Information Propagation in Fringe Communities on Telegram. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 583–595. doi:10.1609/icwsm.v18i1.31336

[23] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 427–431.

[24] Ian Kloo, Iain J. Cruickshank, and Kathleen M. Carley. 2024. A Cross-Platform Topic Analysis of the Nazi Narrative on Twitter and Telegram during the 2022 Russian Invasion of Ukraine. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 839–850. doi:10.1609/icwsm.v18i1.31356

[25] Caio Machado, Beatriz Kira, Vidya Narayanan, Bence Kollanyi, and Philip Howard. 2019. A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections.. In *The Web Conference*. 1013–1019.

[26] Alexandre Maros, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 3005–3011. doi:10.1145/3366423.3380070

[27] Philipe Melo, João MM Couto, Daniel Kansaon, Vitor Mafra, Júlio Reis, and Fabrício Benevenuto. 2024. A Sticker is Worth a Thousand Words: Characterizing the Use of Stickers in WhatsApp Political Groups in Brazil. *arXiv preprint arXiv:2406.08429* (2024).

[28] Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Jie Wu. 2023. It's a Trap! Detection and Analysis of Fake Channels on Telegram. In *2023 IEEE International Conference on Web Services (ICWS)*. 97–104. doi:10.1109/ICWS60048.2023.00026 ISSN: 2836-3868.

[29] Mohammad Naseri and Hamed Zamani. 2019. Analyzing and Predicting News Popularity in an Instant Messaging Service. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1053–1056. doi:10.1145/3331184.3331301

[30] Sarah Nikkhah, Andrew D. Miller, and Alyson L. Young. 2018. Telegram as An Immigration Management Tool. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 345–348. doi:10.1145/3272973.3274093

[31] Gabriel Peres Nobre, Carlos HG Ferreira, and Jussara M Almeida. 2022. A hierarchical network-oriented analysis of user participation in misinformation spread on WhatsApp. *Information Processing & Management* 59, 1 (2022), 102757.

[32] Gustavo Resende, Philipe Melo, Julio CS Reis, Marisa Vasconcelos, Jussara M Almeida, and Fabrício Benevenuto. 2019. Analyzing textual (mis) information shared in WhatsApp groups. In *10th ACM Conference on Web Science*.

[33] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *The Web Conference*. 818–828.

[34] Diego Rybski, Sergey V Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A Makse. 2009. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences* 106, 31 (2009), 12640–12645.

[35] Gandeva Bayu Satrya, Philip Tobianto Daely, and Muhammad Arif Nugroho. 2016. Digital forensic analysis of Telegram Messenger on Android devices. In *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. 1–7. doi:10.1109/ICTS.2016.7910263 ISSN: 2338-185X.

[36] Kseniia Tikhomirova and Ilya Makarov. 2021. Community Detection Based on the Nodes Role in a Network: The Telegram Platform Case. In *Analysis of Images, Social Networks and Texts*, Wil M. P. van der Aalst, Vladimir Batagelj, Dmitry I. Ignatov, Michael Khachay, Olessia Koltsova, Andrey Kutuzov, Sergei O. Kuznetsov, Irina A. Lomazova, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Marcello Pelillo, Andrey V. Savchenko, and Elena Tutubalina (Eds.). Springer International Publishing, Cham, 294–302.

[37] Aleksandra Urman and Stefan Katz. 2022. What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society* 25, 7 (May 2022), 904–923. doi:10.1080/1369118X.2020.1803946

[38] Otavio Venâncio, Carlos H. G. Ferreira, Jussara M. Almeida, and Ana Paula C. da Silva. 2024. Unraveling User Coordination on Telegram: A Comprehensive Analysis of Political Mobilization during the 2022 Brazilian Presidential Election. In *Proceedings of the 18th International AAAI Conference on Web and Social Media*.

[39] Janith Weerasinghe, Bailey Flanigan, Aviel Stein, Damon McCoy, and Rachel Greenstadt. 2020. The Pod People: Understanding Manipulation of Social Media Popularity via Reciprocity Abuse. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1874–1884. doi:10.1145/3366423.3380256

[40] Maximilian Wich, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh. 2022. Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community. *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 1133–1144. doi:10.1609/icwsm.v16i1.19364

[41] Jiahua Xu and Benjamin Livshits. 2019. The anatomy of a cryptocurrency pump-and-dump scheme. In *Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19)*. USENIX Association, USA, 1609–1625.

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


## A Appendix

Appendix includes additional results to contextualize our analysis.

In Figure 10, we show the breakdown of the groups we found on TGStat, and those we actually monitor for the paper (blue pattern). The figure details the fraction of groups we ignored for various motivations.

In Table 3, we provide a qualitative explanation of the topics under observation. In

**Table 3: Empirical topic description. TGStat does not provide any description of the topics.**

| Topics | Description |
|---|---|
| *Education* | Discussion about college and university courses and exams. |
| *Bookmaking* | Discussion about online betting and similar topics. |
| *Cryptocurrencies* | Discussion about cryptocurrencies, market stock and similar topics. Some groups offer official support for crypto-exchanges. |
| *Technologies* | Discussions about consumer electronics, mostly smartphones. Some groups are second-hand marketplaces for consumer electronics. |
| *Darknet* | Trading of mostly illegal content, i.e., credit card numbers, accounts of media platforms, etc. |
| *Software and apps* | Discussion about usage of software, mostly regarding Android modding and app piracy. Some groups discuss software development for specific languages or technologies. |
| *Video and Films* | Discussion about movies and video sharing of movies and TV series and (possibly piracy). |
| *Politics* | Discussion about political news at large in different countries. |
| *Erotic* | Sharing and suggestion of adult content or services. |
| *Linguistics* | Community of users practising a particular language for educational purposes. |
| *Courses and guides* | Some groups share content related to MOOCs or paid online courses, but most groups are filled with spam advertising new courses. |
| *Economics* | Discussion about market stocks and investments, but most groups are actually about cryptocurrencies. |

Table 4, we present additional per-topic detailed statistics.

In Figure 11, we compare the median use of non-textual media in group messages by topic.

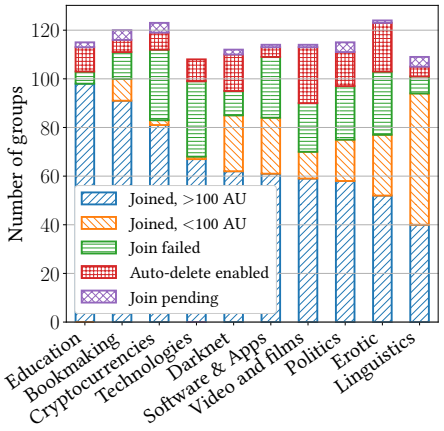



**Figure 10: Per-topic groups of TGStat and results of our crawling.**

**Table 4: Per-topic additional metrics.**

| Topic | English groups | Russian groups | Bot msgs (%) | Avg msg length (ch.) | Avg video size (MB) | Avg video duration (min) |
|---|---|---|---|---|---|---|
| *Education* | 46 | 13 | 8.9 | 43.3 | 49.9 | 4.3 |
| *Bookmaking* | 24 | 49 | 6.4 | 68.3 | 14.1 | 0.8 |
| *Crypto* | 45 | 15 | 8.3 | 44.4 | 12.5 | 0.9 |
| *Technologies* | 23 | 30 | 9.9 | 58.5 | 13.1 | 0.8 |
| *Darknet* | 9 | 40 | 8.0 | 305.4 | 7.8 | 0.5 |
| *Software* | 14 | 28 | 5.4 | 40.1 | 17.4 | 1.4 |
| *Video&Films* | 21 | 23 | 9.6 | 61.1 | 48.6 | 5.1 |
| *Politics* | 11 | 34 | 3.6 | 160.8 | 21.2 | 1.9 |
| *Erotic* | 15 | 14 | 6.6 | 60.1 | 60.1 | 2.8 |
| *Linguistics* | 20 | 11 | 12.6 | 35.0 | 19.5 | 3.4 |
| *All Topics* | 228 | 257 | 7.8 | 65.5 | 26.1 | 2.1 |

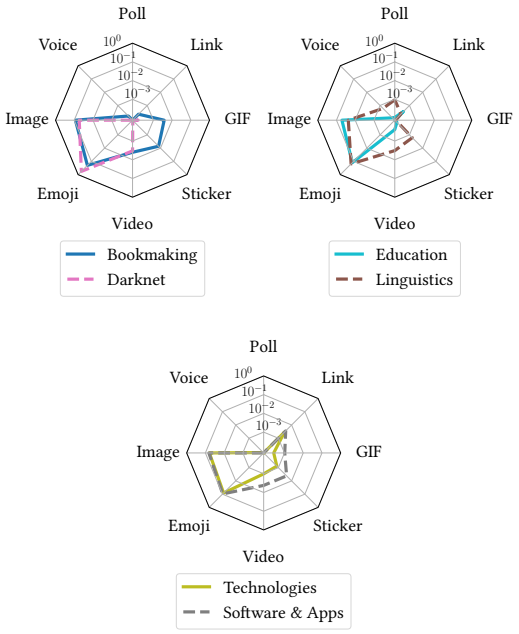



**Figure 11: Median fraction of messages with non-textual elements in topics left out from Figure 4.**