Network Analysis of German COVID-19 Related Discussions on Telegram

Valentin Peter[1], Ramona Kühn[1 [0000−0002−9750−0305]], Jelena Mitrović[1 [0000−0003−3220−8749]] Michael Granitzer[1 [0000−0003−3566−5507]], and Hannah Schmid-Petri[1 [0000−0003−2778−4477]]

University of Passau, Innstraße 41, 94032 Passau, Germany
`valentin.peter@hotmail.de`,{`ramona.kuehn, jelena.mitrovic, michael.granitzer, hannah.schmid-petri`}`@uni-passau.de`

Abstract. We present an effective way to create a dataset from relevant channels and groups of the messenger service Telegram, to detect clusters in this network, and to find influential actors. Our focus lies on the network of German COVID-19 sceptics that formed on Telegram along with growing restrictions meant to prevent the spreading of COVID-19. We create the dataset by using a scraper based on exponential discriminative snowball sampling, combining two different approaches. We show the best way to define a starting point for the sampling and to detect relevant neighbouring channels for the given data. Community clusters in the network are detected by using the Louvain method. Furthermore, we show influential channels and actors by defining a PageRank based ranking scheme. A heatmap illustrates the correlation between the number of channel members and the ranking. We also examine the growth of the network in relation to the governmental COVID-19 measures.

Keywords: Corpus Creation · Community Detection · Network Analysis · COVID-19 · Telegram · Querdenker · Social Media

1 Introduction

A global pandemic started in 2020 after the outbreak of the SARS-CoV-2 virus that causes a disease named COVID-19. Governments all over the world implemented restrictions such as curfews, contact restrictions, and the obligation to wear medical masks. A group of people rejecting those measures inevitably formed, denying the existence of the virus and the risks associated with it. In Germany, this movement is called "Querdenken" (*lateral thinking*) and its members "Querdenker" (*lateral thinker*). They show a mistrust towards the established media [7] and are sceptical towards the governmental restrictions related to COVID-19. We will refer to them as COVID-19 sceptics in the following. They are a heterogeneous group, ranging from intellectuals, members of right-wing movements, supporters of alternative medicine, to mothers who are worried about their children [9, 7].

Protests are mainly organized via the messenger service Telegram[1] [5]. Telegram is known for its support of freedom of speech and the lack of censoring. However, Telegram does not provide features for collecting data for research purposes such as Twitter. Telegram offers 1:1 (private chats), 1:n (channels), and n:n (groups) communication. It offers end-to-end encryption and automatic deletion of messages. A group is either public or private and can have up to 200,000 members, while channels allow an unlimited number of members, but only one person to send messages. Channels and groups can reference each other by `mentions` ("@Person"), `forwards` ("Forwarded from ..."), or sharing Telegram `links` ("t.me/*"). Those references help to discover similar groups/channels and to unveil network structures and communities. We will use the terms groups and channels interchangeably, as their structure does not affect our methods.

We present how to efficiently sample relevant data from Telegram. Furthermore, we show how neighbouring channels can be detected. Our focus lies on channels that can be attributed to the COVID-19 sceptics movement in Germany. We create a dataset of 1979 distinct channels and cluster them into communities. Then, we investigate the channels to find influential actors in the German Querdenker movement. The channels are ranked using the PageRank algorithm. With our approach, we provide helpful insights into the organization and evolution of German COVID-19 sceptics' social network actions. Furthermore, we want to facilitate the process of creating datasets from Telegram for other researchers. Following research questions are guiding our work:

RQ1: How can data from Telegram be sampled for further studies?
RQ2: What are the major communities in the Telegram network of the German COVID-19 sceptics movement?
RQ3: What are the most influential Telegram channels or actors of the German COVID-19 sceptics movement?
RQ4: Is the growth of the German COVID-19 sceptics' network related to the governmental measures against COVID-19?

2 Related Work

The Querdenken movement in Germany has been mainly studied from a sociological perspective to understand who participates at the demonstrations. Those studies are based on online and offline surveys [5, 7, 9]. However, the authors remark that it is likely that the protesters at demonstrations do not reveal their true intentions as they fear legal consequences. In Telegram, opinions can be exchanged more freely without fearing any sanctions or negative reputation. However, the challenge here is to detect relevant channels/groups.

Some implementations of scrapers exist, e.g., the TeleGram-Scraper[2], focusing on information about group members. Telescrape[3] is similar to our implementation, with a greater focus on third party comment apps and media files.

[1] https://telegram.org/
[2] https://github.com/th3unkn0n/TeleGram-Scraper
[3] https://github.com/PeterWalchhofer/Telescrape

Jalili and Perc [6] present different measurements for node centrality and cascade models (e.g., PageRank) in networks to rank them. Kwak et al. [8] investigate on Twitter whether users are more influential if their messages are often reposted or mentioned by others. They count the followers and then use the PageRank algorithm on the followers' network allowing them to rank channels by their total number of retweets. Dargahi et al. [2] try to rank channels on Telegram using the `mention` graph to predict the number of participants. Their conclusion is that there is no relationship between the "degree of nodes in the Telegram mention graph and its number of followers". They found out that channels with many followers have the lowest PageRank. The authors point out that an algorithm that properly ranks Telegram channels is needed. In this paper, we will present an effective way for achieving that.

3 Dataset Creation With Exponential Discriminative Snowball Sampling

Data sampling on Telegram is more complicated than on other platforms, as it does not offer straight-forward features to scrape data and due to its structure and a variety of third-party plugins, it is more difficult to find connections between groups/channels. We answer RQ1 by first creating a list of channels (called seed) that serves as a starting point. From those channels, we identify relevant links to other channels with a ranking score. Channels with less than 10 messages are neglected as they are irrelevant.

To retrieve messages from a certain channel, the Telegram API endpoint `channels.getMessages` is used. The endpoint `channels.getFullChannel` provides information about a channel, e.g., number of participants. Global search results can be obtained via the endpoint `contacts.search`. To extract and sample data from Telegram, a scraper[4] is implemented that is based on the Telethon library[5]. The data is parsed to a custom data format containing 23 attributes per message (e.g., channel name, datetime, etc.) and saved to a CSV file. To avoid unnecessary requests, every downloaded message is persisted, so only new messages need to be downloaded. For each scraped channel, the complete history up to the 16th August 2021 is sampled.

Exponential discriminative snowball sampling is a well-known technique to sample hard-to-reach populations and to find relevant channels on Telegram [5, 13]. A channel points to other channels by `links`/`forwardsmentions`. For exponential discriminative snowball sampling one has to first define how to generate the seed/where to start the sampling, second how to evaluate the next potential scraping candidates. The seed generation can bias the sample if one channel is used as starting point, as all of the following channels originate from only one source. Nevertheless, previous studies started with only one channel [13, 5]. Similar to Dargahi et al. [2], we will start with a list of channels as seed.

[4] https://github.com/vali101/telegraph
[5] https://pypi.org/project/Telethon/

This list is created by the combination of three different methods: First, we use the Telegram `contacts.search` function to retrieve channels with the keywords "impfen" (*vaccination*), "maske" (*mask*), "corona" (*covid*), "pandemie" (*pandemic*), or "querdenken" (*German COVID-19 sceptics*). Second, we select channels that are mentioned by e.g., newspapers, famous bloggers, conspiracy news pages, etc. Third, we extract the links to the Telegram groups from the official register of the Querdenken movement[6]. A list of 231 distinct channels is generated that is our starting point for the scraping.

Now, we need to choose the next channels: Holzer [5] starts from one channel and selects the 25 most prominent channels based on `forwards`/`mentions`/`links`. He selects the most prominent channels that were not in the first batch. He stops after two iterations, resulting in 51 distinct channels. This method does not work well when scraping >100 channels as the scraper starts drifting to channels with other topics or other languages, as large channels (>1 million messages) heavily influence the ranking. Urman et al. [13] avoid this problem by counting how many distinct channels reference the channel instead of how often the channel is referenced in total to smooth out the influence of big channels. We will combine both approaches: the ranking score of each channel is determined by the number of distinct channels which refer to it. In each iteration, 200 channels are evaluated and persisted. In the end, our dataset consists of 1979 distinct channels with about 50 million messages in total.

4 Community Detection, Classification, and Ranking of Channels

A community is a combination of different channels that have shared topics, interests, or goals. Grouping the various channels of COVID-19 sceptics into smaller communities can help to better understand the heterogeneous groups. We use the well-established Louvain method for community detection [1].

For the implementation of the Louvain method, we use the Python library Python-Louvain.[7] To construct the network, `forwards`, `mentions`, and Telegram `links` to other channels can be used. In our dataset, every 5th message is a `forward` from another channel. `Mentions` only occur in 7 % of the messages. We will therefore use only `forwards` to construct the network as they are more popular. It would be too time consuming to manually investigate all 1979 channels, however, we assume that the Louvain method does not only create communities based on topological clusters but also on similar discussion topics within the channels.

To define the most influential channels and communities, we do not assume that a high number of participants leads automatically to a high influence, as bots and fake accounts can be members. For each community that the Louvain method defined, we rank the top 5 channels using the PageRank algorithm both

[6] https://app.querdenken-711.de/initiatives-directory
[7] https://python-louvain.readthedocs.io/en/latest/api.html

on `mentions` and `forwards`. This algorithm is a variant of the eigenvector centrality measure and was initially introduced to determine the importance of web pages and to rank them. It mimics a random surfer [10]. With PageRank, we identified the top 20 general channels. High PageRank means that many messages from this channel were forwarded by other influential channels. We code the most prominent channels with the two properties `category_content` and `actor_type`. The `category_content` describes the topics that are discussed in the channel. It is quite challenging to define a category for the discussed topics as they are as heterogeneous as the group of the Querdenker itself. A big German newspapers defined four categories for the topics of the Querdenker [3]. We extend those four categories with two additional ones:

- *COVID-19 sceptic:* Topics are about COVID-19 restrictions, conspiracies about the pandemic, criticism of the politics and established media.
- *QAnon:* Topics about QAnon conspiracy theories.[8]
- *Far-Right:* Channels with members of the far-right movement; official channels of right-wing media/movements. Topics: immigration, nationalism, etc.
- *Conspiracy:* Channels sharing all kinds of conspiracy theories, but no special focus on QAnon conspiracy theories.
- *Spiritual* (New): Channels sharing inspirational quotes, flower images, etc.
- *Alternative* (New): Channels sharing alternative news, often fake or influenced by conspiracy theories.

The `actor_type` expresses the roles of the channel owner/members: *alternative media*, *news aggregator*, *influencer*, *political actor*, or *political initiative*. E.g., all channels from the Querdenken register were coded with `category_content=COVID-19 sceptics`, `actor_type=political initiative`. To answer RQ2, we only consider the seven communities that represent 95 % of all channels. The biggest community with 49 % node share are the COVID-19 protesters, followed by conspiracy theorists/QAnon (17 %) and Trump fanatics (17 %). Community 4 (5 %) are spiritual channels, community 5 (2 %) far-rights in Britain, while community 6 and 7 (both 1 %) have mixed topics.

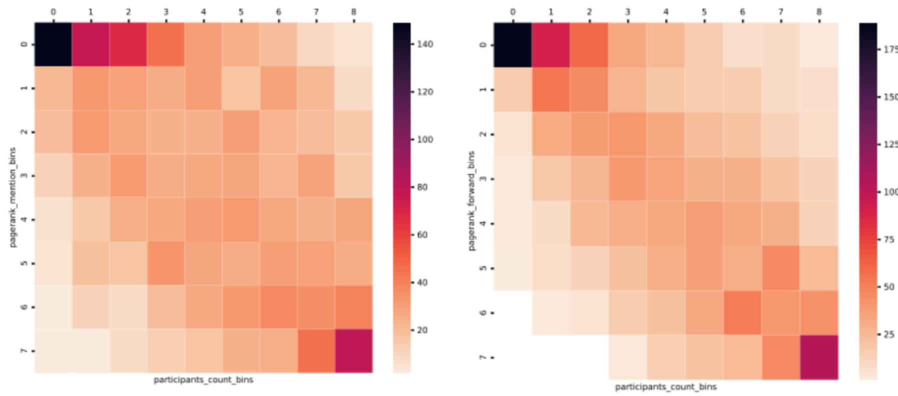
We split the PageRank and participants count into groups using equal frequency binning. Heatmaps show the correlation. In contrast to Dargahi et al., we see a correlation between the number of participants and the PageRank, both for `mentions` (cf. Fig 1a) and for `forwards` (cf. Fig 1b). The identified influential channels in the German COVID-19 sceptics network are shown in Table 1, answering RQ3. Completeness can never be guaranteed as private channels cannot be considered or found.[9]

5 Growth of the Network and Evolution of Measures

We want to answer RQ4 and investigate how the network evolved compared to the measures against the pandemic. The number and severity of restrictions can

[8] E.g., that "the world is run by Satan-worshipping pedophiles" [11].

[9] In the dataset, about 8000 invite links to private channels were found.



(a) `Mentions` (b) `Forwards`

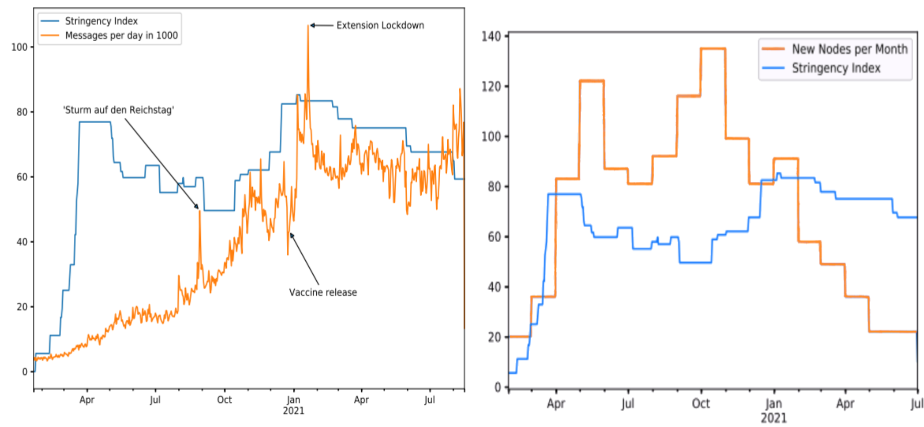
Fig. 1: Correlation between number of participants and PageRank.

Channel name	Rank	Community	`category_content`	`actor_type`	Participants
reitschusterde	0.0283	1	covid_sceptic	media	166,449
epochtimesde	0.0140	2	alternative	media	45,498
ntdeutsch	0.0114	2	alternative	media	11,762
evahermanoffiziell	0.0109	1	conspiracy	influencer	183,440
schubertslm	0.0107	1	other	news aggregator	47,207

Table 1: Most influential channels in the German COVID-19 sceptics network.

be expressed by the Stringency Index: It was developed by Hale et al. [4] and is a COVID-19 Government Response Tracker to compare the different policies against COVID-19 around the globe. They calculate the strictness of the measures using all original containment and closure policy indicators[10], plus an indicator recording public information campaigns. Figure 2a illustrates the relation between message frequency in the German speaking COVID-19 sceptics channels and the stringency index, representing governmental COVID-19 measures. The first peak in the number of messages is reached when the Reichstag was stormed, the second peak when the lockdown and restrictions were extended. After this peak, both the stringency index and the number of messages move sideways. The growth of the Telegram network is expressed by the increase of number of new nodes and the sum of edges in a given time frame [13]. Figure 2b shows the number of new channels/groups joining the network per month.

[10] School/workplace closing, stay at home requirements, travel restrictions, etc.



(a) Relation of COVID-19 countermeasures and number of messages.

(b) New nodes per month.

Fig. 2: Relation of COVID-19 restrictions and network activities.

6 Unknown Recommendation Problem

We want to discuss why the Dargahi et al. [2] approach led to unsatisfactory results, while we achieved good results with the same method. In Telegram, it is unknown who forwarded a message or who mentioned it. One has to find all channels that reference the analysed channel. Unknown channels can stay undetected. We call this the **"Unknown Recommendation Problem"**. Furthermore, we notice a strong correlation (0.55) between the `forward` PageRank and the number of participants. In contrast, Dargahi et al. did not find enough neighbouring channels, and calculated the PageRank on the `mention` graph. Due to their diverse seed and the small number of channels, they discovered channels with many participants but did not collect information about their neighbouring channels.

7 Conclusion

We presented a way to sample data from relevant channels on the messenger platform Telegram. Two different approaches were combined for sampling. The sampling bias was reduced by generating a large seed. We created the largest known dataset of the German COVID-19 sceptics movement on Telegram. We identified communities with the Louvain method and influential channels by using PageRank. We detected a correlation between the number of channel members and the ranking of the channel. In addition, we showed how activities increased in line with governmental COVID-19 measures. Future research can focus on intra-channel or intra-group communication. As some governments blame Telegram for the protests and want to ban it [12], its investigation becomes even more relevant.

Acknowledgements The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The author is responsible for the content of this publication.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
2. Dargahi Nobari, A., Reshadatmand, N., Neshati, M.: Analysis of Telegram, an instant messaging service. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2035–2038 (2017)
3. Ebitsch, S., Kruse, B.: Wer den Hass verbreitet. https://www.sueddeutsche.de/digital/steckbriefe-akteure-telegram-1.5278290 (2021), accessed: 2022-03-03
4. Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., et al.: A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour **5**(4), 529–538 (2021)
5. Holzer, B.: Zwischen Protest und Parodie: Strukturen der "Querdenken"-Kommunikation auf Telegram (und anderswo) (2021)
6. Jalili, M., Perc, M.: Information cascades in complex networks. Journal of Complex Networks **5**(5), 665–693 (2017)
7. Koos, S.: Die "Querdenker". Wer nimmt an Corona-Protesten teil und warum?: Ergebnisse einer Befragung während der "Corona-Proteste" am 4.10. 2020 in Konstanz (2021)
8. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600 (2010)
9. Nachtwey, O., Frei, N., Schäfer, R.: Politische Soziologie der Corona-Proteste (2020)
10. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
11. Rose, K.: What is QAnon, the viral pro-Trump conspiracy theory? www.nytimes.com/article/what-is-qanon.html (2021), accessed: 2022-02-25
12. Stempfle, M.: Was kann der Staat gegen Telegram machen? https://www.tagesschau.de/inland/telegram-verfassungsschutz-corona-leugner-101.html (2021), accessed: 2022-03-10
13. Urman, A., Katz, S.: What they do in the shadows: examining the far-right networks on Telegram. Information, communication & society pp. 1–20 (2020)