TeleScope: A Longitudinal Dataset for Investigating Online Discourse and Information Interaction on Telegram

Susmita Gangopadhyay, Danilo Dessì, Dimitar Dimitrov, Stefan Dietze

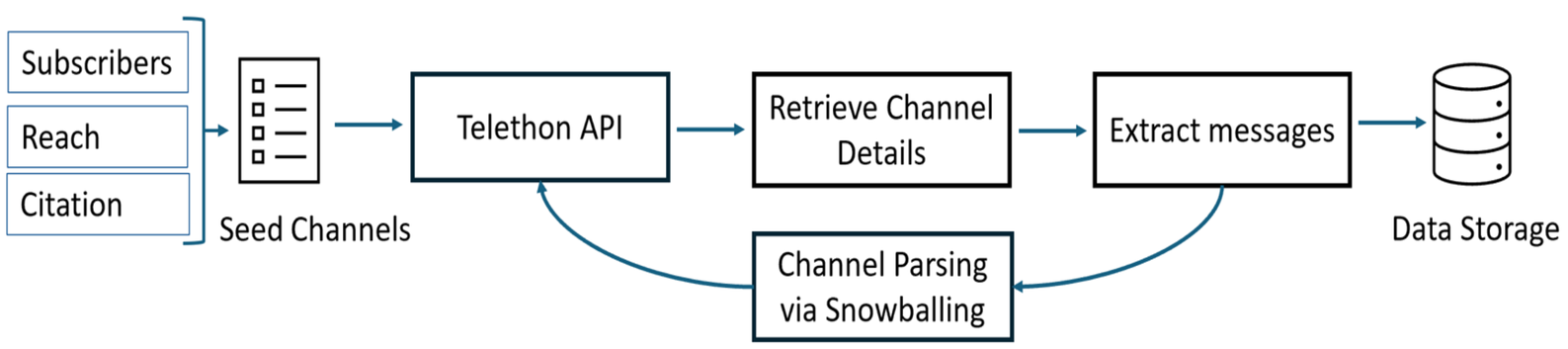
1. Motivation and Problem Statement

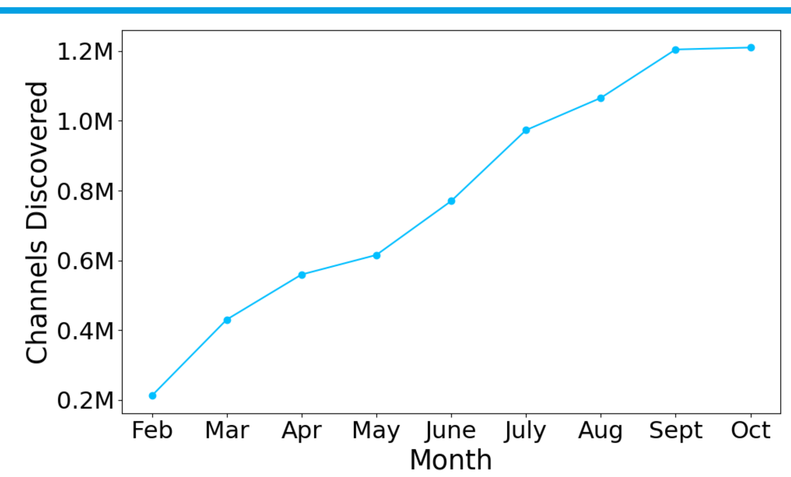
- Telegram is a widely used messaging platform renowned for its privacy and social features.
- Telegram channels can influence public opinion and contribute to the spread of misinformation.
- X/Twitter API restrictions highlight the need for alternative social media data sources.

Problem: How do we enable years of X/Twitter research to be replicated on Telegram?

2. X/Twitter vs. Telegram

3. Data Collection and Channel Discovery





Monthly discovery of channels

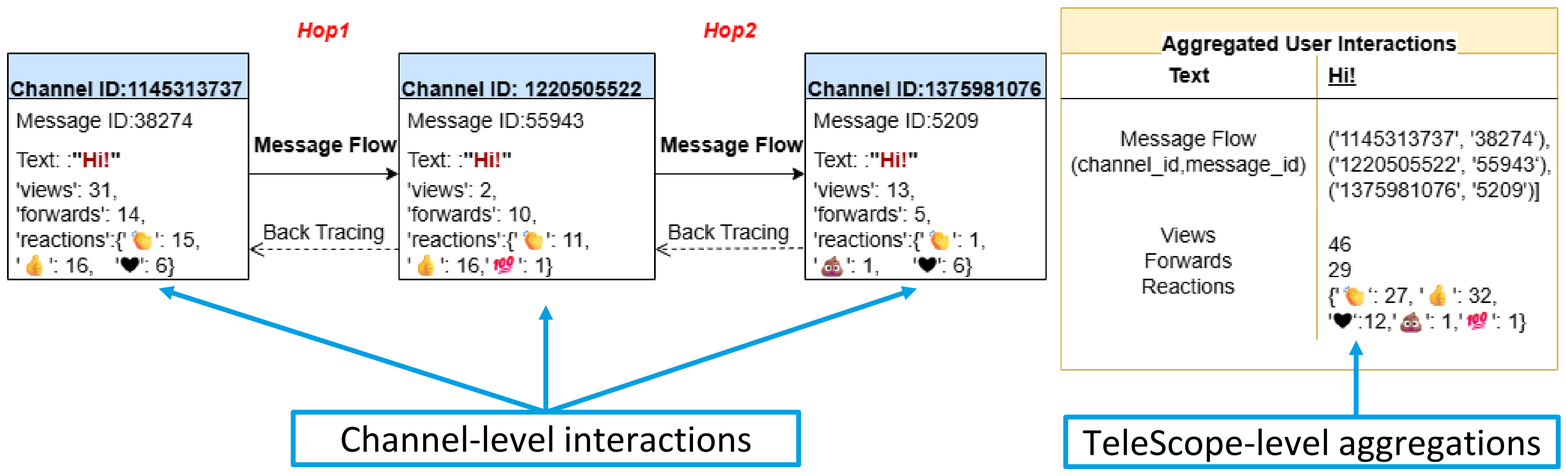
- Seedlist - Top 300 channels collected from tgstat.com, 251 unique channels

4. TeleScope Dataset

4.1 Dataset Statistics

Feature	Value
Time frame	(Feb 1-Oct 29) 2024
Discovered channels	1,210,272
Channels with downloaded metadata	534,137
Fully downloaded public channels	71,048
Number of downloaded messages	120,024,020
Average messages /channel	1689.33
Percentage of forwarded messages	19.6%
Average messages downloaded /hours	20,495
Complete dataset size	76GB(zipped)

4.2 Enrichments



Feature	Value
Total Number of Messages	31,227,109
Number of unique messages	308,147
Smallest message flow	2
Longest message flow	4,810
Average message flow	2,54

Propagation statistics of forwarded messages

Channel-to-channel graph

Language	%
Ru	82.29
Uk	4.6
En	4.2
Fa	2.2
De	1.1
Cannot Determine	0.08
Others	5.53

Language distribution among downloaded public channels

Telegram entities: Top 10 hashtags in messages

5. Use Cases

6. Conclusion and Future Work

- Regular yearly TeleScope releases.
- Focused crawls, i.e., elections, climate change, migration.
- Estimating representativity, i.e., amount and type of channels covered.

Susmita Gangopadhyay (susmita.gangopadhyay@gesis.org)
Danilo Dessì (ddessi@sharjah.ac.ae)
Dimitar Dimitrov (dimitar.dimitrov@gesis.org)
Stefan Dietze (stefan.dietze@gesis.org)

Website Paper Code