Security and Privacy in Online Social Network

Lab 01 - OSN Analytics

Due **6:00pm on 5/19/2023 tentatively**

POLICIES:

1. **Coverage**
   OSN analytics: data collection and analysis
2. **Grade**
   Lab 01 accounts for **40%** of the final grade
3. **Individual or Group**
   Individual based, but group discussion is allowed and encouraged
4. **Academic Honesty**
   Violation of academic honesty may result in a penalty more severe than zero credit for an assignment, a test, and/or an exam.
5. **Submission**
   Submit your ***code if any, collected data, and a report*** including data collection methodology, data analysis, findings, and discussion, to **course.zju.edu.cn**
6. **Late Submission**
   Deduction ceases upon zero;

Lab Goal:

Data Collection and Data Analysis

Telegram is a popular OSN platform, where has become a major venue for people selling illicit services or sharing information about underground crimes.

Lab 01 aims to master techniques for collecting real-world cybercrime-relevant data from Telegram and conducting analytics.

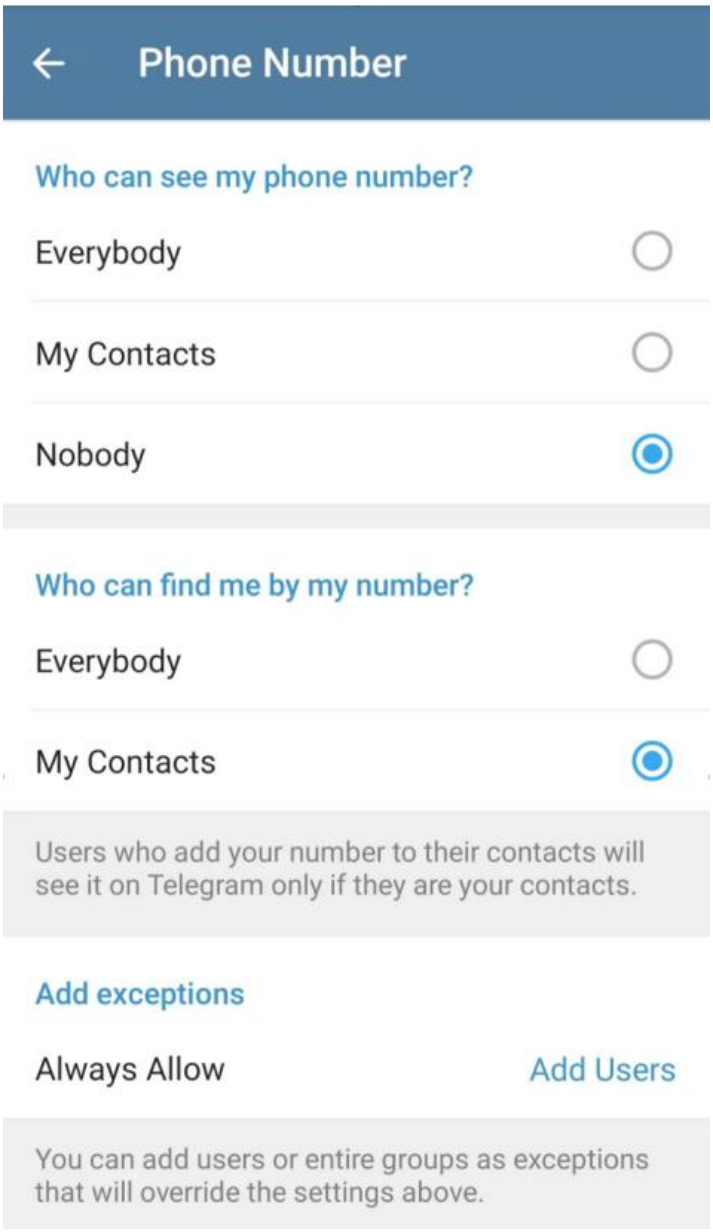
Lab Steps:

Note that **you will need a VPN throughout the lab**.

1. Registering a Telegram account

(1) Download Telegram from *https://telegram.org/* and register an account.

(2) For your own privacy protection, go to *Settings-->Privacy and Security-->Phone Number*, and make the following choices (see the below image).



2. Obtain *api_id* and *api_hash*

(1) Login to your Telegram account with the phone number of the developer account to use. *If you register with a Chinese phone number, which should be like +86xxxxxxxxxx. You are expected to receive a SMS message as phone verification.*

(2) Click under API Development tools.

(3) A *Create new application* window will appear. Fill in your application details. There is no need to enter any *URL*, and only the first two fields (*App title* and *Short name*) can currently be changed later.

(4) Click on *Create application* at the end. Remember that your **api_id and api_hash are secret** and Telegram won't let you revoke it. Don't post it anywhere!

3. Install Telethon

Telethon is a Python 3 library to interact with Telegram's API as a user or through a bot account (bot API alternative). This library is meant to make it easy for you to write Python programs that can interact with Telegram.

(1) Install Telethon: pip install telethon
(2) We can finally write some code to log into our account!

```
from telethon import TelegramClient

# Use your own values from my.telegram.org
api_id = 12345
api_hash = '0123456789abcdef0123456789abcdef'

# The first parameter is the .session file name (absolute paths allowed)
with TelegramClient('anon', api_id, api_hash) as client:
    client.loop.run_until_complete(client.send_message('me', 'Hello, myself!'))
```

In the first line, we import the class name so we can create an instance of the client. Then, we define variables to store our API ID and hash conveniently.

At last, we create a new TelegramClient instance and call it client. We can now use the client variable for anything that we want, such as sending a message to ourselves.

(3) Configure the proxy (since we are in mainland China)
- For Python >= 3.6 : *install python-socks*
- For Python <= 3.5 : *install PySocks*

(4) Also change the part *TelegramClient('anon', api_id, api_hash)*
to:
TelegramClient('anon', api_id, api_hash, proxy=("socks5", '127.0.0.1', 7890)),
and replace the the protocol, IP and port with those of your proxy.

Refer to the document *https://docs.telethon.dev/en/stable/* for any questions about data collection with Telethon.

4. Search for interested groups and Harvest each group's content

Next, we need to search for the groups of our interest, i.e., the groups probably related to cybercrimes. There are two typical methods for group searching. One method is via some third-party websites or programs, e.g., tgstat.com, telemetr.io, tgram.io, combot.org. They enable users to search Telegram groups

/channels by keywords. The second is via Telegram Bots, which also allow searching for Telegram Group/Channel by keywords.

(1) In this lab, you are required to **leverage Telegram bot for searching** for interested illegal groups.

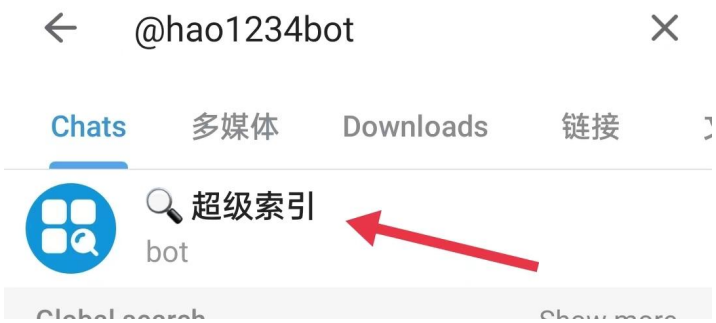
(2) Click on the top-right search button or icon, input one of the following three popular bots, and start using that bot.

@hao1234bot

@SuperIndexBot

@DogeIndexBot

See the below image for illustration.



(3) Then type a keyword, such as “色情”, in the “Message” textbox of the bot window to get a list of groups (see the below image)

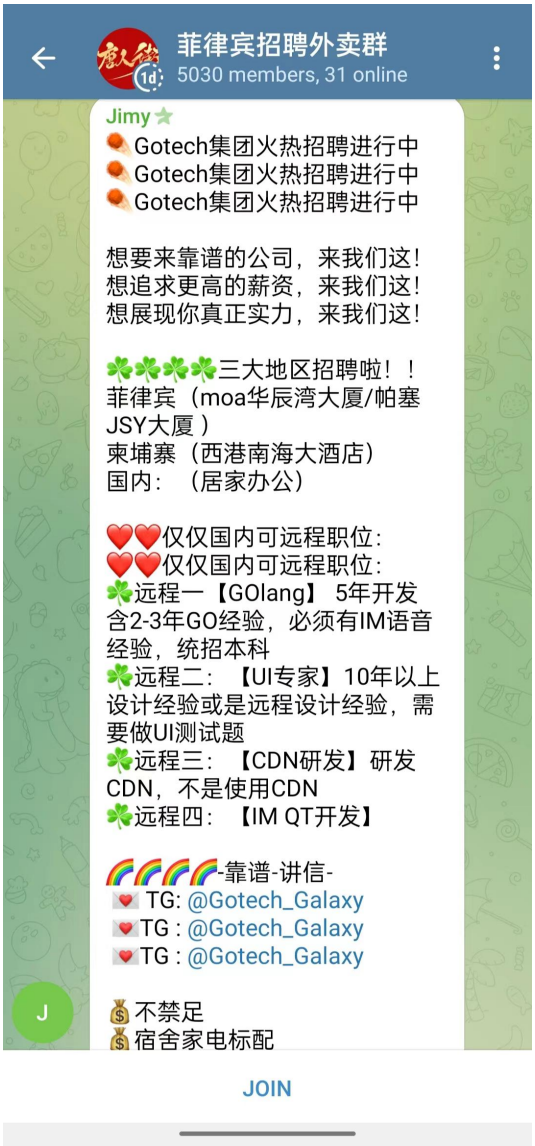
(4) Here is a list of keywords, which are of our interest and could help us find the groups/channels relevant to cybercrime.

0:色情
1:博彩
2:约炮
3:彩票
4:刷单
5:推广
6:话术
7:跑分
8:app
9:破解

(5) **The keyword assigned to you is totally based on the last digit of your student ID**. Suppose the last digit is 0, and then your assigned keyword is 色情.

(6) With the keyword, search for it using one of the three recommended Telegram bots, and use Telethon to **obtain at least 150 unique groups**. For each group, obtain and save its URL link and click on "下一页" to obtain the links of the groups in the next page until you collect 150+ groups' link.

(7) After collecting those groups' link, visit each group, collect its content, the member information, chat frequency, and etc., and save those information locally. *The group content could be text, voice, images and other forms.*



5. Analyze the group content

Analyze the group content, group member information, frequency of chat, and any information of interest, using techniques such as NLP.

Note that the group content could include many duplicated information, so do some data prepossessing and remove the duplicated information.

6. Write a report, based on your data collection and analysis

What should be included in your submitted report?

Your Name (in English) and Student ID #

1. Summary

Please provide a short paragraph including the goal of the lab, the tool you use, the dataset you collected, the tool used for analysis, findings, and etc.

2. Background

Please provide a brief introduction to the Telegram groups you focus on.

3. Data Collection Methodology

Please provide a concise description about how you collect the data.

4. Dataset Description

Please provide statistics about the dataset you have collected.

5. Data Analysis

Please describe how you do the data analysis.

6. Findings

Please present and discuss any findings you make from data analysis, and provide some tables and/or figures if any.

7. Miscellaneous (challenges, solutions, and what you have learned)

Please discuss the challenges you encounter in the lab, the solutions to them, what you have learned, and anything you would like to share with me.

An example report is provided as a separate PDF file.