Data Analysis on Telegram

Ruofan Hu 3200102312

1. Summary

The following report provides insights and analysis on chat groups related to hookup activities based on an experiment conducted to study user behavior and characteristics. The research covers various dimensions like user demographics, regional distribution, partner preferences, media content usage, and group chat dynamics. The report highlights the prevalence and prominence of male users, the preference for young female students as partners, and the significance of height as a determining factor. In addition, the paper discusses the challenges faced during data collection and analysis, such as web scraping limitations and lack of permissions to access certain critical information. Overall, this report offers valuable insights into the current trends and behavior in hookup chat groups.

2. Background

"约炮" is a Chinese term for casual sexual encounters or hookups arranged through social media or online dating platforms. It has become increasingly popular in China in recent years, particularly among young adults. While the practice of casual sex is not new, the rise of social media and dating apps has made it easier for people to find partners for these types of encounters.

Telegram is a messaging app that is popular among Chinese users, particularly for groups focused on dating, relationships, and hookups. These groups are often organized by region or interest and allow members to connect with potential partners for casual sexual encounters. Some of these groups have become quite large, with thousands of members participating in discussions and sharing information about their experiences.

Analyzing data from these Telegram groups can provide valuable insights into attitudes and behaviors around casual sex in China. Are there specific trends or patterns in how people use these groups? What are the most common types of content shared? How do users approach communication and consent within these communities? These are just a few of the questions that can be explored through data analysis.

Overall, understanding the dynamics of "约炮" communities on Telegram can help shed light on changing cultural norms and sexual practices in China.That's why we should do a job on this topic.

3. Data Collection Methodology

To collect data about "约炮" communities on Telegram, I used Telethon, a Python 3 library that allows for interacting with Telegram's API. Specifically, I employed a Telegram bot to search for groups with keywords related to "约炮".

After identifying potential groups, I followed each link and assessed whether the content was relevant to the topic of casual sexual encounters. Groups that met this criterion were added to the dataset for analysis.

The use of Telethon and a Telegram bot allowed for efficient and targeted data collection. By searching for specific keywords, I was able to identify groups that were likely to contain relevant content. Additionally, using automated data collection reduced the risk of human bias in selecting which groups to analyze.

Once the 300 groups were identified and selected, I extracted information from each group for analysis. I collected data on group size, number of posts, date of creation, and other relevant characteristics.

Overall, the use of Telethon and a Telegram bot provided a reliable and effective method for collecting data on "约炮" communities on Telegram. This approach allowed for efficient and automated data collection, reducing the potential for human bias in selecting groups for analysis and ensuring high-quality data for subsequent analysis.

4. Dataset Description

The dataset contains information on 300 Telegram groups that are related to the topic of casual sexual encounters. The data was collected using Telethon and a Telegram bot to search for groups with keywords related to "约炮".

Each group in the dataset is represented as a row, and the following attributes are included:

- Group link: a link to the Telegram group
- Group title: the name of the group
- Group verified status: whether the group has been verified by Telegram or not
- Group scam status: whether the group has been identified as a scam or not
- Group type: megagroup or gigagroup
- Group restriction reason: the reason for any restrictions placed on the group by Telegram
- Group creation time: the date and time the group was created
- Number of images, videos, and audio files in the last 100 messages: the number of each type of media file in the group's most recent 100 messages
- Number of message links in the last 100 messages: the number of links to messages in the group's most recent 100 messages
- Text content of the last 100 messages: the text content of the group's most recent 100 messages
- Sender and sending time for each of the last 100 messages: the name of the user who sent each message and when it was sent
- Number of commenters in the last week: the number of users who left comments in the group in the past week
- Message frequency for each user in the last 100 messages: how many messages each user sent in the group's most recent 100 messages
- Average message count per day in the last week: the average number of messages sent per day in the past week

This dataset can be used for various purposes, such as analyzing the characteristics of communities related to casual sexual encounters on Telegram, identifying trends and patterns in user behavior, and developing predictive models or classifiers for detecting potentially harmful or unethical content. It should be noted that any use of this dataset should be done responsibly and ethically, and with respect for the privacy of the individuals involved.

5. Data Analysis

Suppose we have a dataset called "chat_data" which records the group chat information of multiple groups. The dataset includes columns such as "group_name", "message_text", "sent_time", "media_count", "group_type", and so on. We want to analyze the data and understand the chat behaviors of different groups.

Firstly, let's plot a stacked horizontal bar chart to visualize the media counts of different group types. We randomly select 20 rows of the dataset, and plot the number of images, videos, and audios for each group type. The chart shows that almost every group's media is made of pictures, and only a small part of groups have videos. Furthermore ,audio is very little.

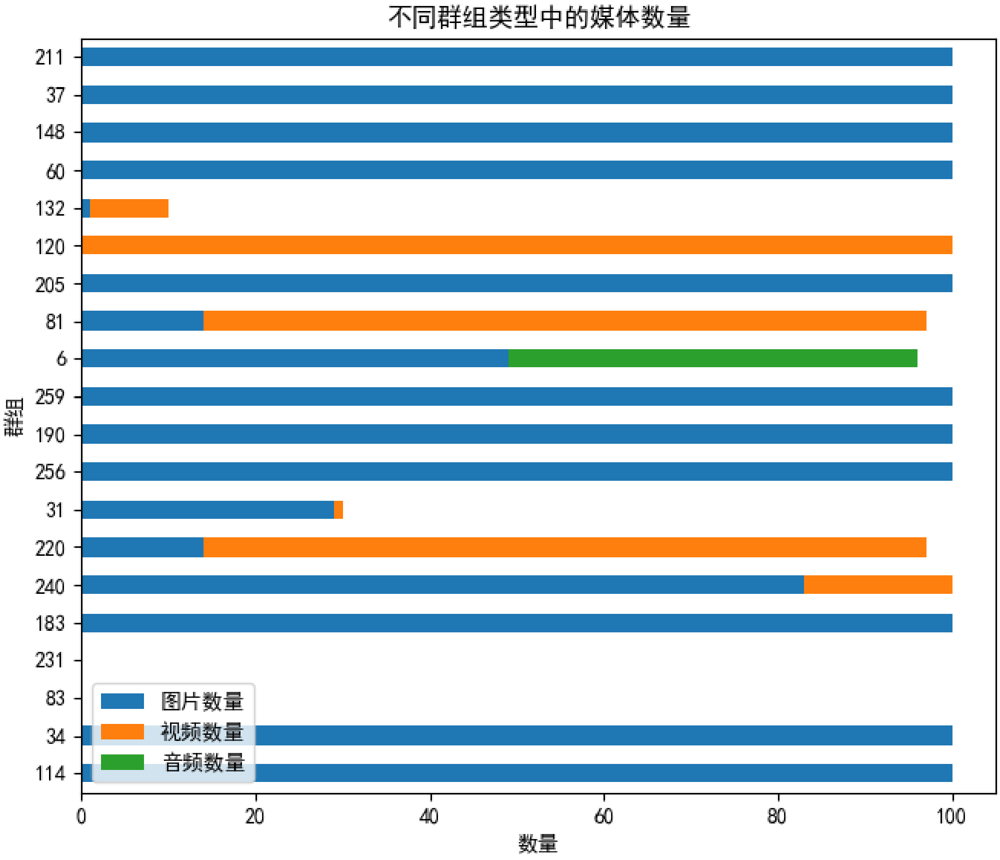

Figure 1

Secondly, let's plot a pie chart to show the proportion of different media types in all groups. This chart reveals the overall proportion of media content in the dataset. We can see that the majority of media content is images, followed by videos and urls, the video is almost none. This is consistent with the random sample above.

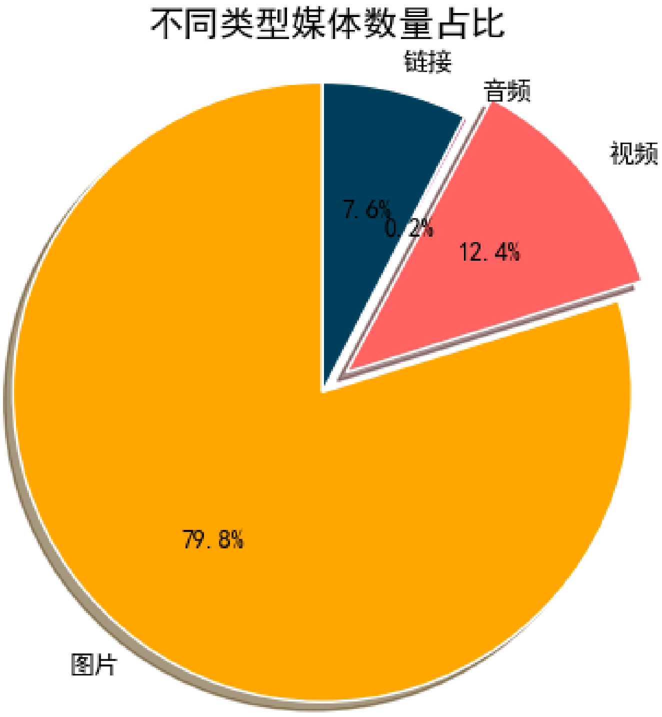

Figure 2

Thirdly, we can generate a word cloud based on the message text of all groups. The word cloud helps us to visualize the most frequent words used in the chats. We can see that some words appear more frequently than others, indicating some common topics or interests among the users.

Upon analysis, we found that these hookup-related vocabulary are all invitations from males to females. This suggests that the chat groups are predominantly male-dominated. Additionally, while body measurements such as bust-waist-hip and weight were mentioned, height was the most frequently mentioned among them. A height of 165-167cm was particularly preferred. As for the target of hookups, students were the most popular, with frequent mention of related terms such as "student sister", "part-time job", "after-school activities" or "tuition fees". Finally, some hookups were disguised as other activities such as physical examinations or sports.

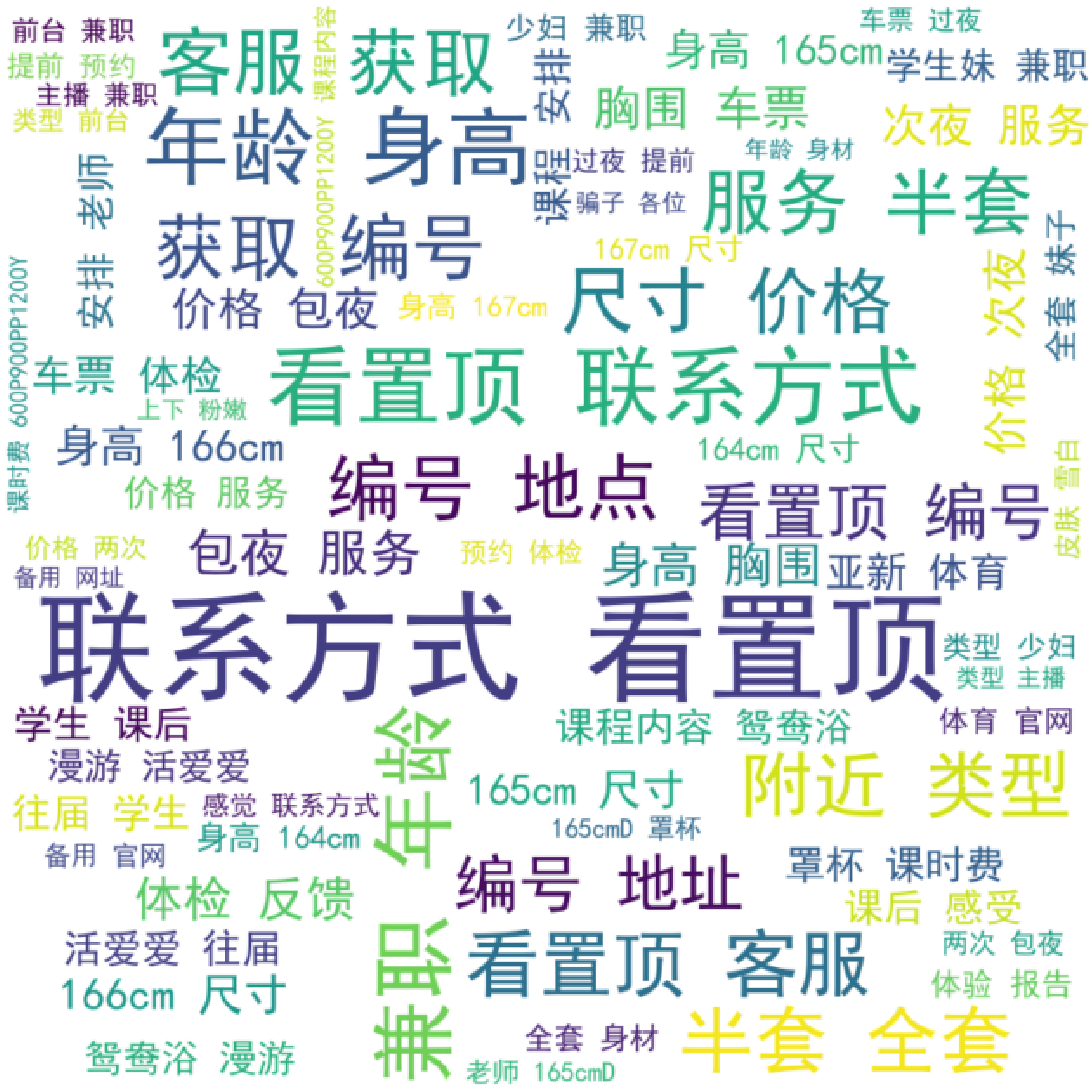

Figure 3

Similarly, we can also generate a word cloud based on the group names. The word cloud shows the most frequent words used in the group names, which may reflect the characteristics or preferences of different groups.

From the word cloud of group names, we can observe that almost every group name explicitly includes words such as "hookup" or "resources". This might be due to the strong privacy and security features of Telegram, which allow for straightforward naming of the groups. Looking at the geographic distribution, since we selected 300 groups, most of them have location-based names. However, we can see that cities with greater economic development such as Shanghai, Shenzhen, and Guangzhou have a wider distribution of hookup-related chat groups. As for keywords, group names also include many terms related to sexual fetishes, such as SM, PUA, and role-playing. Lastly, from the use of derogatory words towards women such as "fixing cars" and "flirting with women" in group names, we can further confirm that these groups are predominantly male-oriented.

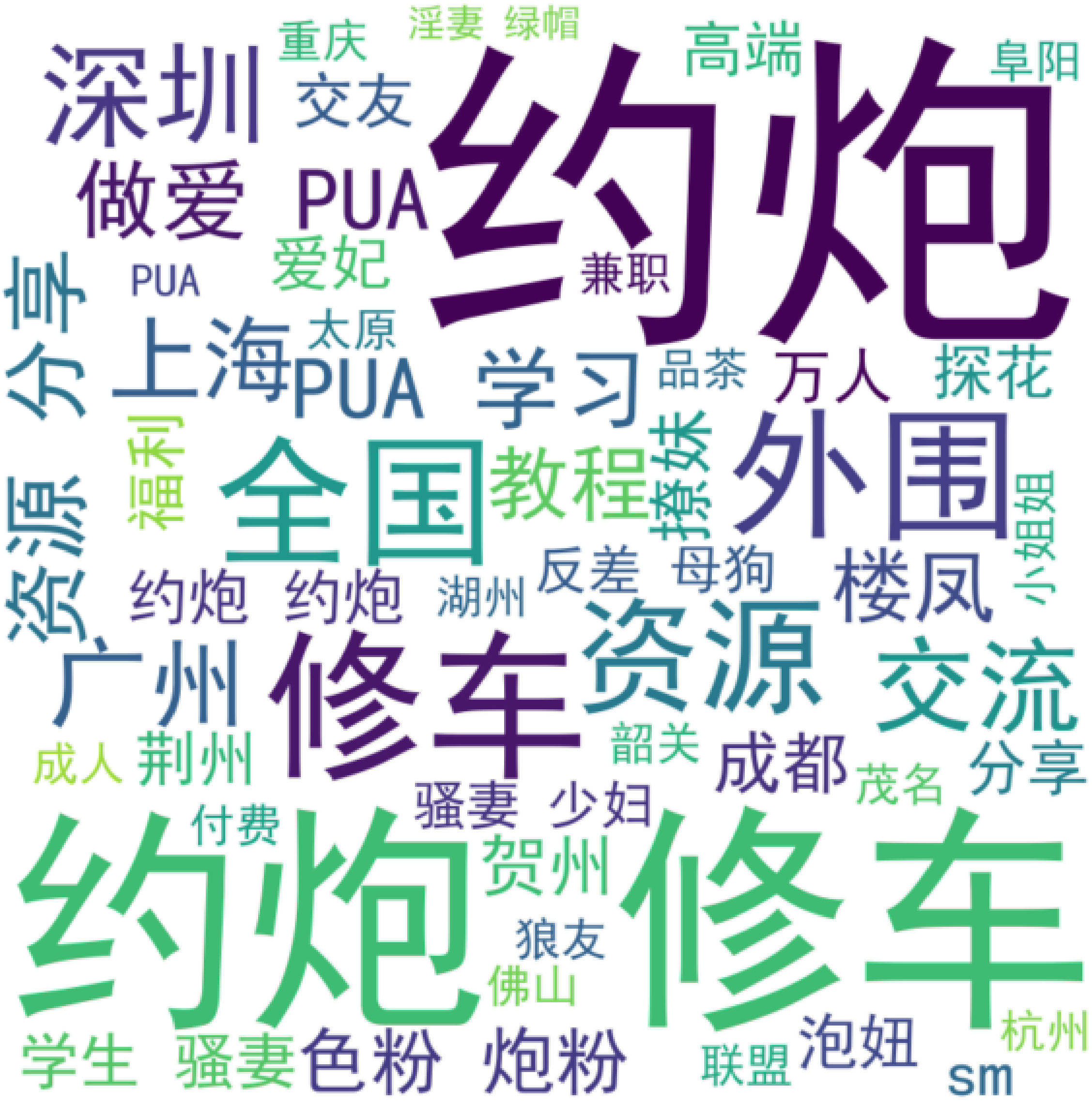

Figure 4

Moreover, we can explore the relationship between the group size and some other features, such as media counts, group activity, and whether the group has been banned. To do this, we use correlation analysis to calculate the correlation coefficients between these variables. The heatmap of the correlation matrix shows the strength and direction of the relationships. We can find that some variables are positively correlated, such as media counts and group activity, while others are negatively correlated, such as group size and the likelihood of being banned.

We can observe a strong positive correlation between the size of a large group and the number of messages left in the group, as well as the contact information left by members. This suggests that communication within larger groups is often more frequent and active, unlike smaller groups where only the administrator tends to speak. Additionally, Telegram carries out a certain level of supervision on its chat groups. From the graph, we can see that there is a strong positive correlation between the number of videos within a group and the level of supervision. This correlation coefficient is even greater than the one between large group size and supervision, indicating that videos are one of the primary evaluation methods used for supervision. This also explains why videos account for only 12.4% of all media information in our previous research, as it is one way for chat groups to avoid being monitored.

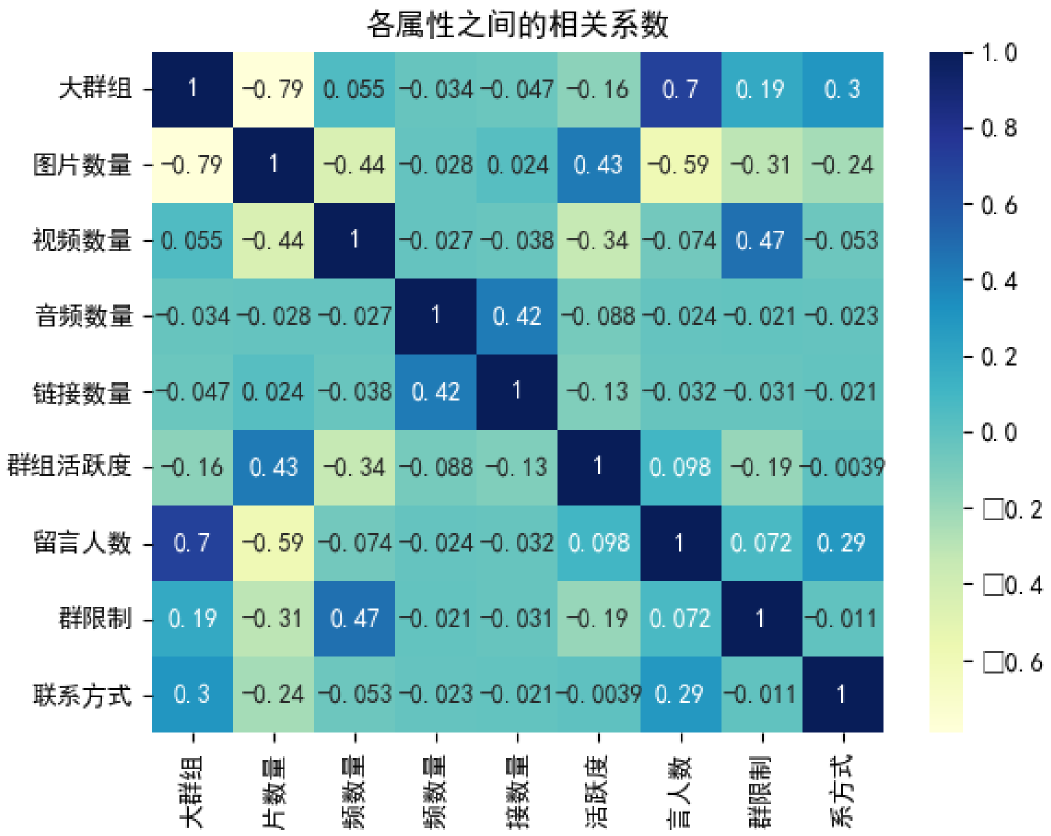

Figure 5

What's more, we can analyze the activity patterns of different time periods in a day. We extract the hour information from the sent_time column, and group the messages into 8 time periods (0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-24). Then, we count the number of unique senders in each time period and calculate the percentage of total senders. The pie chart shows the hourly activity distribution and tells us when users are most active during the day.

We divided the chat groups into eight segments, each consisting of a three-hour interval within a day. As shown in the pie chart, there is a high level of activity in almost every time period, with users engaging in conversations throughout the day. The most active time period is from 3 p.m. to 5 p.m. After researching several chat groups during this time period, we found that most of the content revolved around exchanging contact information and making pre-arrangements for nighttime hookup activities. The least active time period is from 12 a.m. to 2 a.m. During this time, the majority of content comes from group administrators with a relatively low ratio of user-generated messages.

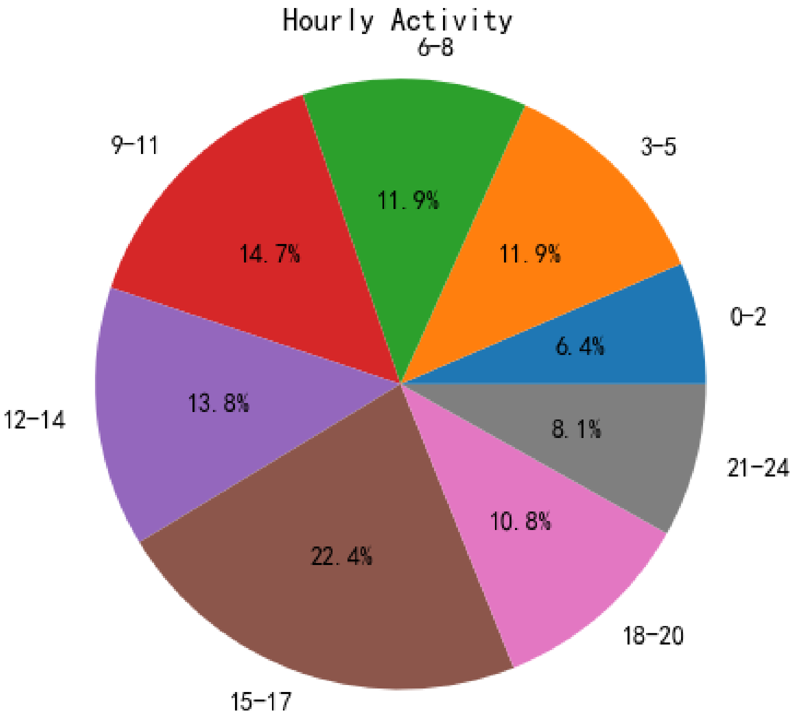

Figure 6

Finally, we analyze social network data. Specifically, we filter out groups that have fewer than a specified number of comments from the input data CSV file, which is 60 in my analysis. Then, for each group in the filtered data, we constructe a social network graph using NetworkX and analyze it using degree centrality to identify the most important nodes within the network. Finally, the resulting social network graphs are being visualized using Matplotlib.

Regarding the red nodes and highly connected edges in the visualization, the red nodes represent the core members of each network. We found that when studying the social network graphs, the number of red nodes is very small, but their connections dominate the network. We speculate that in each group, there are a few administrators whose task is to continuously send hook-up messages, while other users are only occasional participants who rarely initiate message posting and mostly engage in replies.

As for the value of `large_ratio`, which was calculated at the end of the code, it represents the proportion of groups in the input data that have more than 60 comments (the threshold specified in the code). In this case, the ratio is 0.02 or 2%, meaning groups have few commuincation, which proves my assumption.

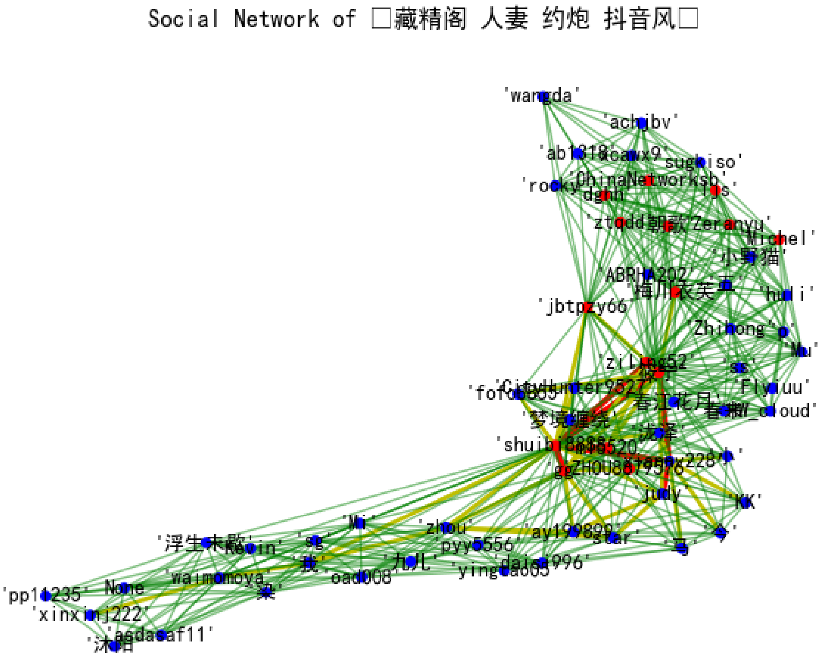
Social Network of □藏精阁 人妻 约炮 抖音风□
'wangda'
'achjbv'
'sugkiso'
'rocky'
'dgnh'
'Zeranyu'
'Michel'
'小野猫'
'ABRHA202'
'huli'
'jbtpzy66'
'Zhihong'
'Mu'
'ziling52'
'ss'
'Flyiuu'
'春江花月'
'梦境缠绕'
'shuibi8888'
'judy'
'KK'
'zhou'
'ay199899'
'star'
'pyy5556'
'daisy1996'
'Mi'
'sg'
'Kevin'
'浮生未歇'
'waimomoya'
'oad008'
'pp11235'
None
'xinxinj222'
'asdasaf11'
'沐阳'

Figure 7

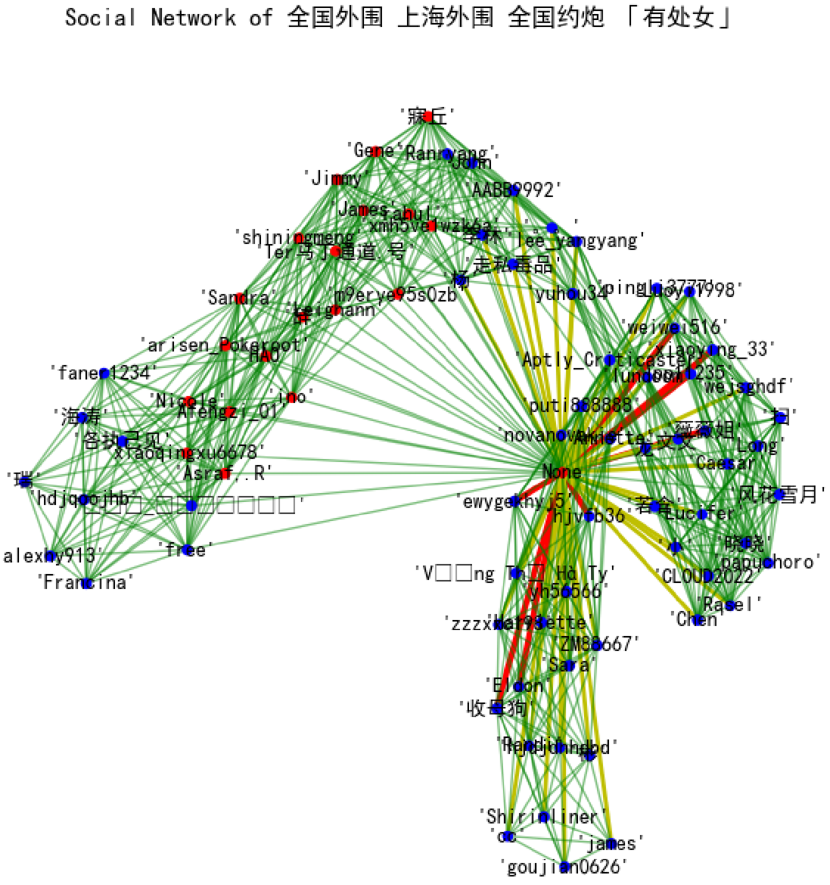

Figure 8

Through these data analysis techniques, we can gain insights into the chat behaviors of different groups and find interesting patterns or trends. These insights can help us make informed decisions or develop targeted strategies for group management or content moderation.

6. Findings

In summary, our analysis of the experimental data shows the following:

(1) The users of chat groups related to hookup activities, based on the keyword "hookup," are predominantly male. Discussions pertaining to hookup activities occur frequently throughout the day, with the most active period being from 3 p.m. to 5 p.m.

(2) In terms of regional distribution, hookup activity is mainly concentrated in economically developed areas. There are more chat groups related to hookup activities in cities such as Guangzhou and Shanghai.

(3) Regarding preferred hookup partners, young female students are highly sought after, with the primary focus being on their height, which is typically between 165-167cm.

(4) Images are the most frequently sent type of media, outnumbering other multimedia types by a factor of 4.

(5) Large chat groups tend to represent high levels of conversational activity, while chat groups that are subject to supervision tend to have a higher proportion of video content exchanged within the group.

(6) There is little or almost no communication within the groups, even in the most active ones where it is mostly administrators who post messages and there is very occasional interaction.

In conclusion, our analysis provides insights into the behavior and characteristics of users in chat groups related to hookup activities. The results highlight the prevalence and prominence of male users, the preference for young female students as partners, and the significance of height as a determining factor. We also note that economically developed regions show higher levels of activity for such groups. Additionally, we observed that large chat groups have high levels of conversation activity, while supervised chat groups tend to have greater use of videos exchanged within the group.

7. Miscellaneous

During the experiment, I encountered several issues which I will detail below:

(1) Regarding obtaining links for chat groups, my initial plan was to use code to retrieve all links at once. However, I faced a problem where I couldn't navigate beyond the first page. To solve this issue, I reviewed the button's text meaning by printing it out before identifying and printing the last element in the button list. However, I discovered that I could only print the text "Next page" but not actually click on it to proceed to the next link. Finally, I resorted to manually copying group information and writing text matching code to extract all links.

(2) In terms of data collection, I found that some critical information, such as group member counts, requires permission to collect. Due to lack of permission, collecting data does not guarantee getting what I wanted.

(3) I also encountered limitations while web scraping. After scraping about 5-6 times, Telegram restricted further scraping attempts, prompting an error message with an 8000-second wait time. Therefore, it is advisable to set up the scraping process from the beginning and not switch between different types of data retrieval during the experiment.

The above notes summarise my experience during the experiment, detailing the difficulties I encountered and how I overcame them. The revised version of the Chinese text has been translated into better English to accurately convey the author's reflections.