The difference between spam users and real users in social media by analyzing "fans" of online celebrity

Shihong Ling

1. Introduction

Online celebrities have become common but somehow "attracting" phenomena in social media such as Twitter, Weibo, TikTok, etc. Our net has many different views on them: some people love them, while others don't care or even hate them. But, overall, no matter what they have done, I think they do bring some "fun"to our life and have some properties which are worth researching. In this project, I will analyze the difference between spam (or fake) users and real users in social media by analyzing "fans" of these online celebrities. The online celebrity I chose for my project is Xunkun Cai, who is famous for incredible fan's increasing rate (almost 50000% after he debuted from the boy band "Nine Percent") and fan's group "iKun".

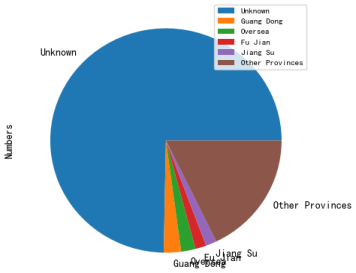
My project consists of three steps: first step, I create some crawl helper methods to collection user information and connections based on [1] and [2], and I choose one post from Cai's Weibo in 2019 (around 2K comments) as my target; second step, I analyze some features extracted from collected information by using Python statistical packages to show there are many suspicious users who should be counted as spam or fake users; third step, I use [3]'s algorithm and tool to generate social network graphs for both spam and real users, and find difference between them.

In addition, I want to make two clarifications: first, for the small change of title, in the midterm report, I took comments similarities into account however I found out this led me to mistake the definition of fans. For example, some users were just clout chasers, although their comments differed a lot but none of them were related to Cai, thus I decided not to use word "fans" because it was inaccurate; second, for the method how I collect connections for different groups, originally my plan was to use processed features to make a cluster model to separate users into two groups, and then I would crawl the connections among the users in two groups, however, due to my requests were too frequent, I got warned so I could only select several representatives from each groups and use them as roots to expand their social network.

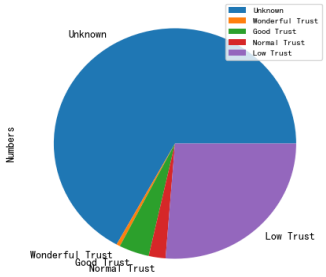
Thanks for reading my introduction and enjoy my sections 2 and 3!

2. Feature Analysis

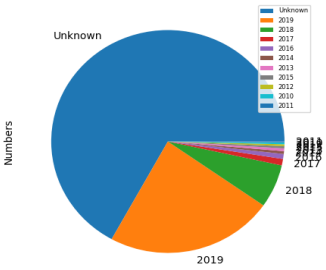
(Location)



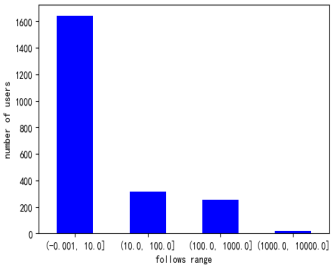
(Trust)



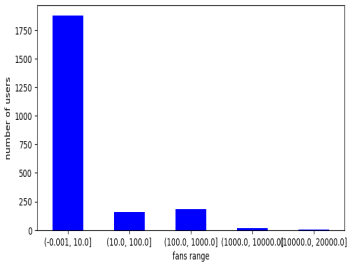
(Birth of Account)



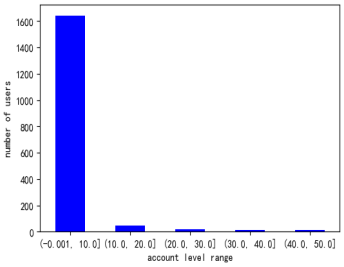
(Follows)



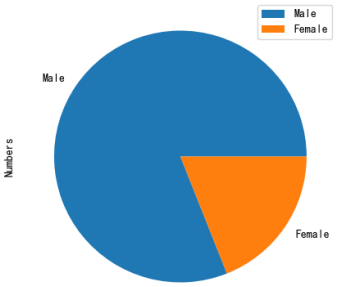
(Fans)



(Account Level)



(Gender)



In this section, I show some pie and bar charts based on some interesting features, and I will explain how I process them and why they are interesting.

Location: "Location" information contains province and city. To reduce variance, I delete "city" and summarize by "province" only. For display, I choose the top 5 locations which have the most users and regard the rest as "others". The interesting thing about this feature is that almost 3/4 users don't clarify their location.

Trust: “Trust” information is calculated by Weibo based on the user's activities. “Unknown” means the user doesn't have enough data to calculate trust and “Low” means the user is suspicious. The interesting thing about this feature is that most users are “Unknown” or “Low” trust.

Birthday of Account: “Birthday of Account” information contains year, month and day when the user’s account was created. To reduce variance, I delete “month” and “day”. The interesting thing about this feature is that most users have unknown birthdays or just “born”.

Follows: “Follows” information is the number of users that the selected account has followed. The interesting thing about this feature is that most users have very low followers (<10).

Fans: “Fans” information is the number of users who follow the selected account. The interesting thing about this feature is that most users have very low fans number (<10)

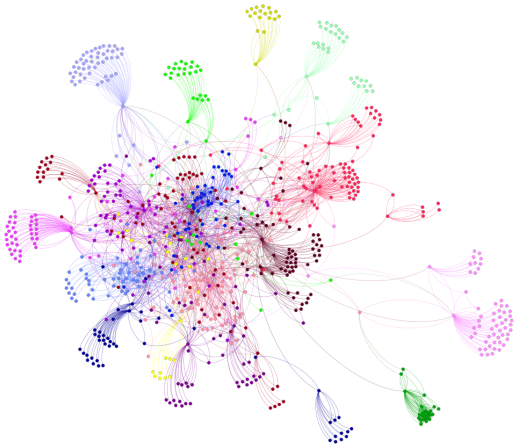
Account Level: “Account Level” information is the level of account, it is somehow like the “Birthday of Account” so it can be affected by the age of account and other issues. The interesting thing about this feature is that most users have very low levels (<10).

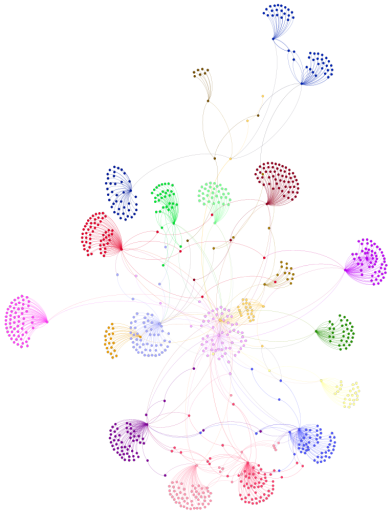
Gender: “Gender” information is the gender chosen by the account. The interesting thing about this feature is that most users are male while Cai debuted in a boy band whose target audience is female.

From these feature plots, we clearly find there is a big group of suspicous spam users inside Cai’s post’s comments and we can get some ideas about how to separate these spam users from real users.

3. **Social Network Analysis**

(Normal Social Network) **(Spam Social Network)**





The above plots are the social network built on two different types of users. From section 2, I can dig out some spam users and real users in my view, and use them as root to expand social networks. Here I borrow the method from [3] and apply different colors by calculating modularities. From these network plots, we can clearly see that real users connect more closely and have some small groups around, however spam users don't have a big center and only compose small groups (looks like they are controlled by someone).

These are my basic understandings about the differences between spam users and real users, and these findings may somehow help detect the spam users in social media.

4. **Reference**

[1] Weibo Spider: https://github.com/dataabc/weiboSpider

[2] Weibo Spider: https://github.com/dataabc/weibo-crawler

[3] Gephi (no date) *Yifanhu multilevel · gephi/gephi wiki*, *GitHub*. Available at: https://github.com/gephi/gephi/wiki/YifanHu-Multilevel (Accessed: November 29, 2022).

(Original paper: http://yifanhu.net/PUB/graph_draw_small.pdf)

5. **Github Link**

https://github.com/SeeonOwO/Data_Visualization_Final_Report