Social Network Analysis of Zhihu

Group 9

Gu Zhijing, 1155073548

Huang Xianghai, 1155055168

Lyu Zishen, 1155071037

Zhao Che, 1155066626

Overview

- Motivation
- Obtain Data
- Basic Statistics
- Graph Analysis
- Summary

Motivation

- Zhihu (www.zhihu.com) is the biggest Chinese knowledge ask-answer web site in China (example in English world: Quora). It has more than 17 million registered users (over 10 million among them are monthly active) by March 2015

- We wonder in such a platform, what kind of form of social network may existsa, and we want to know whether the knowledge areas (topics) have any contact with the interpersonal relationships (which is a special issue of Zhihu compared to social websites like Twitter or Facebook).

Obtain Data

- Features We Care About
  - **User** can have many **followees** and **follower** - directed relationships;
  - **Question** (asked by users) can have several **topics** wiki-tagged;
  - User can have only one **answer** in one question;
  - Answer can get **agree** and **thanks** from other people;
- Database Design

  User (user_url, user_id, answer_num, followee_num, follower_num, agree_num, thanks_num, layer)

  Question (question_id, topic)

  Following (user_url, followee_url)

  UserQuestion (user_url, question_id)

  UserTopic (user_url, topic)

Obtain Data

- Crawler
  - Implemented in Python, based on an open source package called “zhihu-python”;
  - Apply multiprocessing to crawl data on a single machine due to resource limitation;
  - Automated crawling, with automatic error handling;
  - Data crawled are stored in csv files in the first place, and are transferred to Sqlite database afterwards for later processing;

Obtain Data

- Crawling Strategy
  - For one individual user: crawl porfolio information, followers and followees, question_ids of questions he or she has answered;
  - Crawl two followee layers from a pre-defined seed user (zhao-che);
  - Handle one list of distinct user_url to avoid the subtree replication of the same user (e.g. A->B, C->B);
  - Crawl the topics of the questions based on the question_ids we have got.
- Depth First v.s. Width First
  - DF can crawl all the users in the whole website but time is not allowed;
  - WF is more flexible when we only want to crawl several layers from the seed, without any lost.

Obtain Data

- Summary of obtained data

  SQLite .db file: 688MB

  User: 26K, Following: 4.6M, Question: 2.2M, UserQuestion: 1.7M

  UserTopic: 5.4M (Joined by UserQuestion and Question)

Basic Statistics

- User’s followee, follower, answer, agree, thanks

  Follower: mean 3620.0, median 112.0, standard deviation 22978.5 305(zhao-che)

  Followee: mean 176.3, median 67.0, standard deviation 565.9 149(zhao-che)

  Answer: mean 68.9, median 17.0, standard deviation 225.9 120(zhao-che)

  Agree: mean 3858.4, median 96.0, standard deviation 21951.4 828(zhao-che)

  Thanks: mean: 865.3, median 28.0, standard deviation 4627.6 290(zhao-che)

Basic Statistics

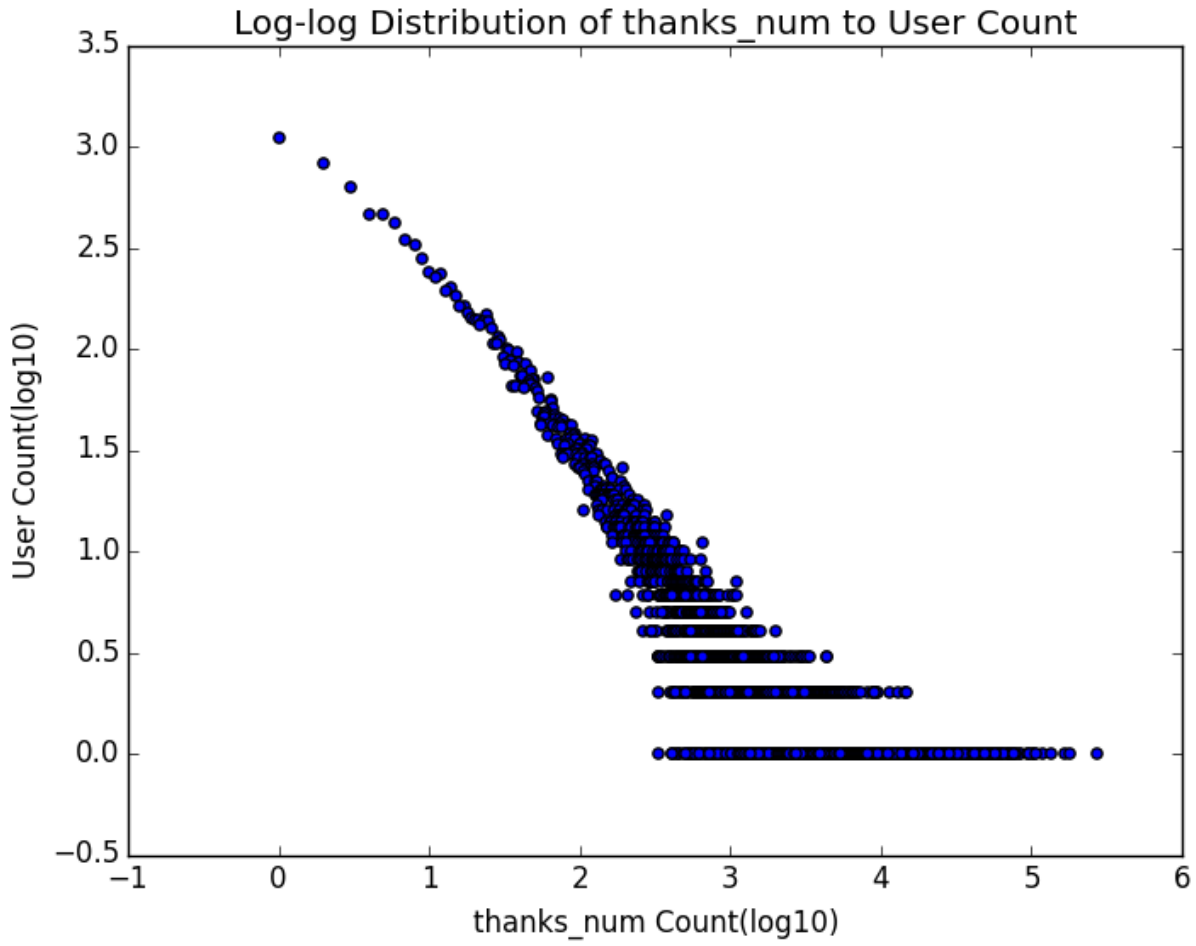
- Log-log Distribution

  Followee: Out Degree

  Follower: In Degree

  Answer, Agree, Thanks

  We can see significant power law in all five features of users.

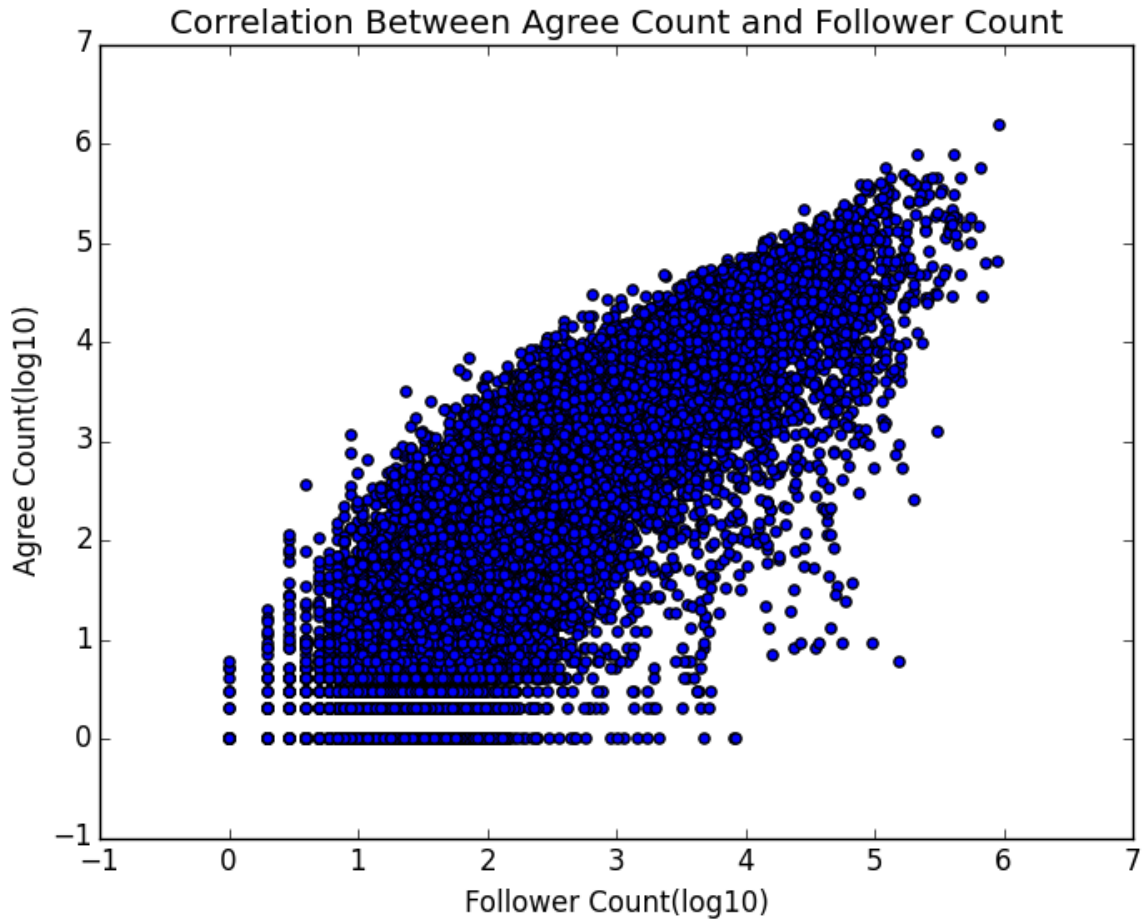


赞同数范围	人数	百分比
1~9	196647	63.96%
10~99	79714	25.93%
100~999	25343	8.24%
1000~9999	5173	1.68%
10000~99999	531	0.17%
100000以上	22	0.01%

Similar Statistics:
http://zhuanlan.zhihu.com/sulian/19781120 3.5M Users

Basic Statistics

- Correlation between Agree and Follower

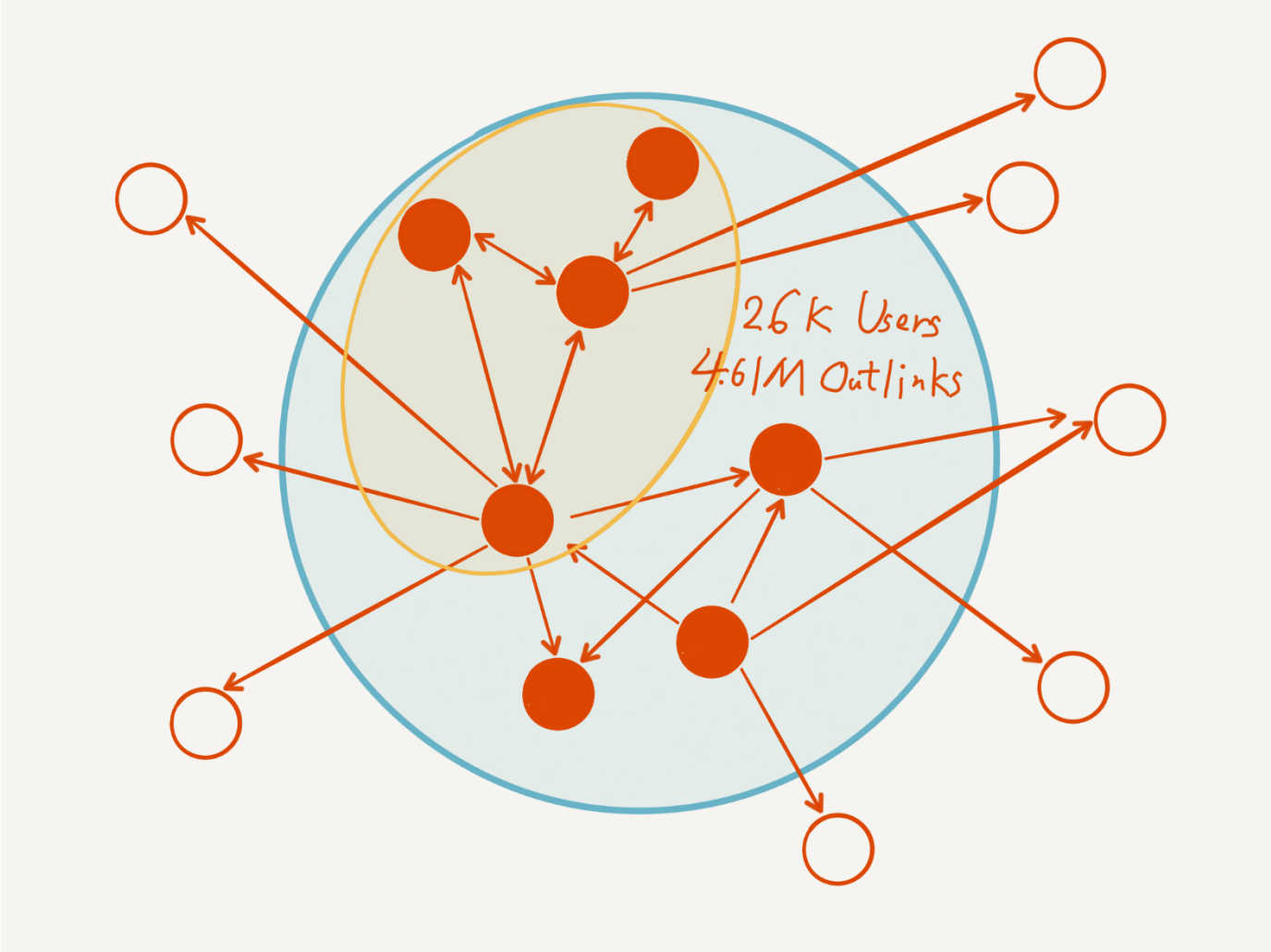


Graph Analysis

- Following Graph Overview
- Pick User Set
  - We should define a user set to analyse;
  - We should avoid using outlinks to users outside the user set.

  SQL to handle:

  select user_url, followee_url from Following where followee_url in (select user_url from User where agree_num > 50000) and user_url in (select user_url from User where agree_num > 50000)



We know all the following and followed relationship within the circle, and the porfolio information of these 26k users. We also know all the other followee of these users, but for now we only know these followee's ID. 26k users have 4.61M followee in all.

Graph Analysis(agree>10k)

- Graph

  Diameter, Radius: Graph not connected, infinite path length

  Average Shortest Path Length: (Underestimate in networkx's DiGraph [(1,2),(2,4),(3,4)], 0.42)

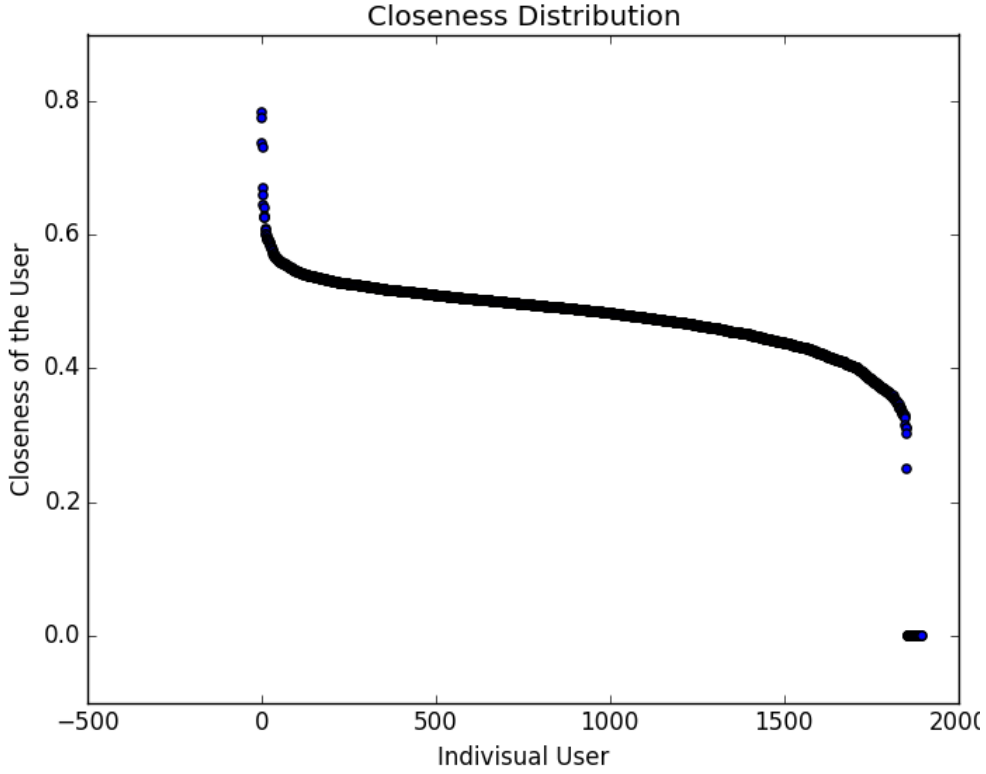
  Density: 0.064

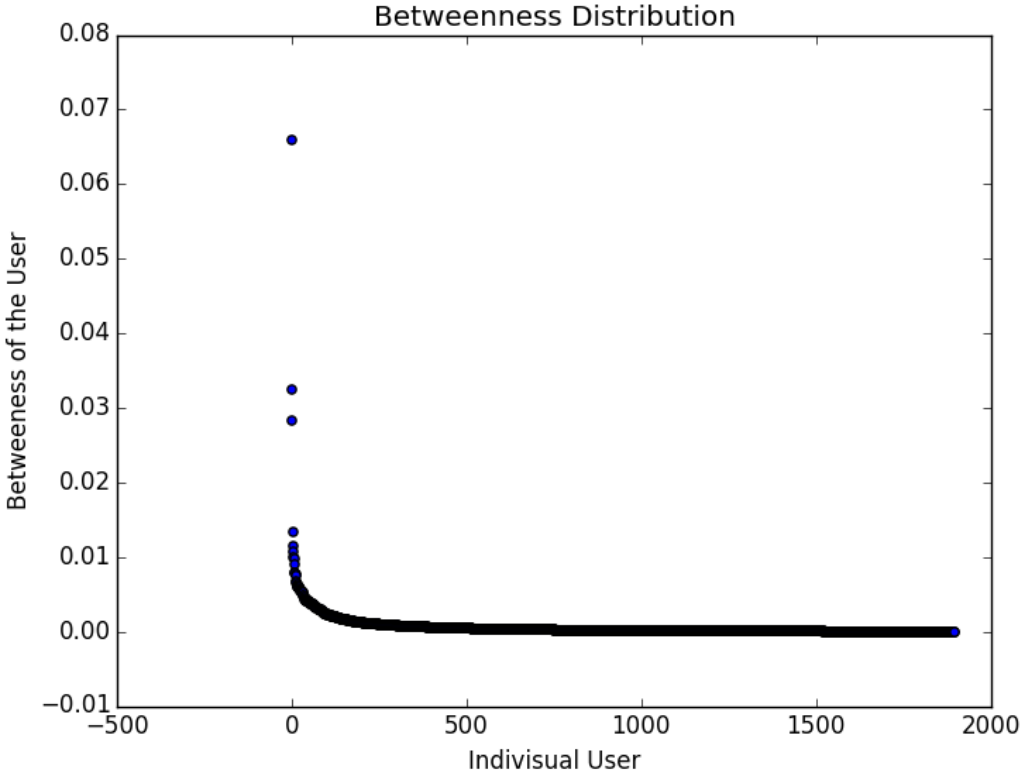
Graph Analysis(agree>10k)

- Nodes

  Betweeness: really low for every point, you don’t have to use a particular path.

  Closeness: nodes are mostly equally important.

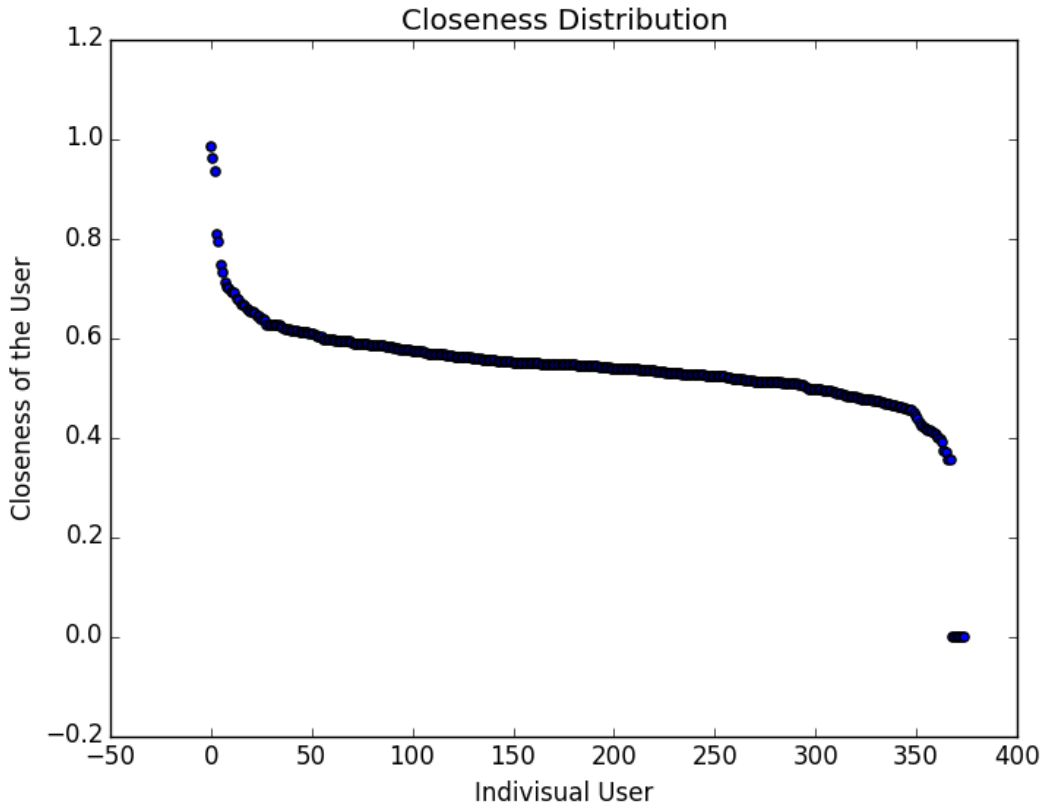


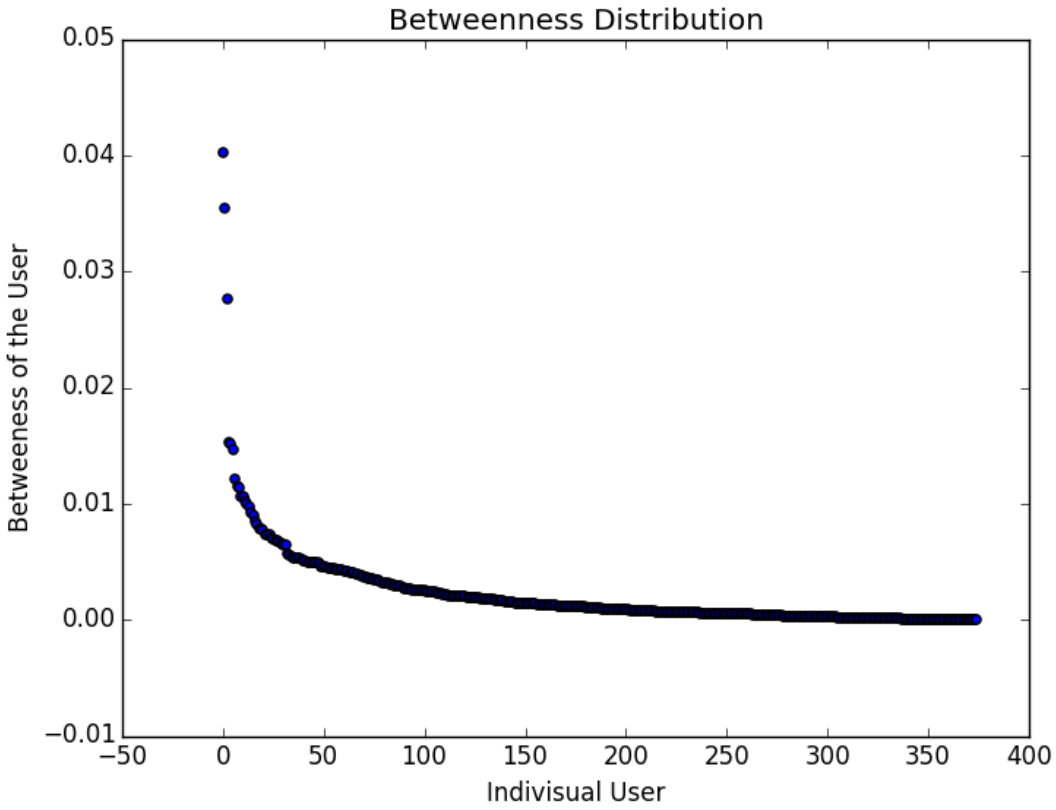


Graph Analysis(agree>50k)

Graph Density: 0.195(much larger than 0.064 in [agree>10k])

Nodes Centrality: similar with the [agree>10k]





Graph Analysis(agree>10k)

- PageRank and HITS

  PageRank top 3: jixin 0.0074, ma-bo-yong 0.0056, zhang-jia-wei 0.0055

  Hub top 3: zhounuo 0.0034, Namoamitabhaya 0.0034, jun-mo-52 0.0034

  Auth top 3: zhang-jia-wei 0.0034, liangbianyao 0.0034, gejinyuban 0.0032

  These link analysis scores can also be used to rank topics (yet to be tried due to time limit).

Graph Analysis(agree>50k)

- PageRank and HITS

  PageRank top 3: **jixin** 0.0112, **ma-bo-yong** 0.0103, **zhang-jia-wei** 0.0102

  Hub top 3: Namoamitabhaya 0.0098, jun-mo-52 0.0097, qisini 0.0095

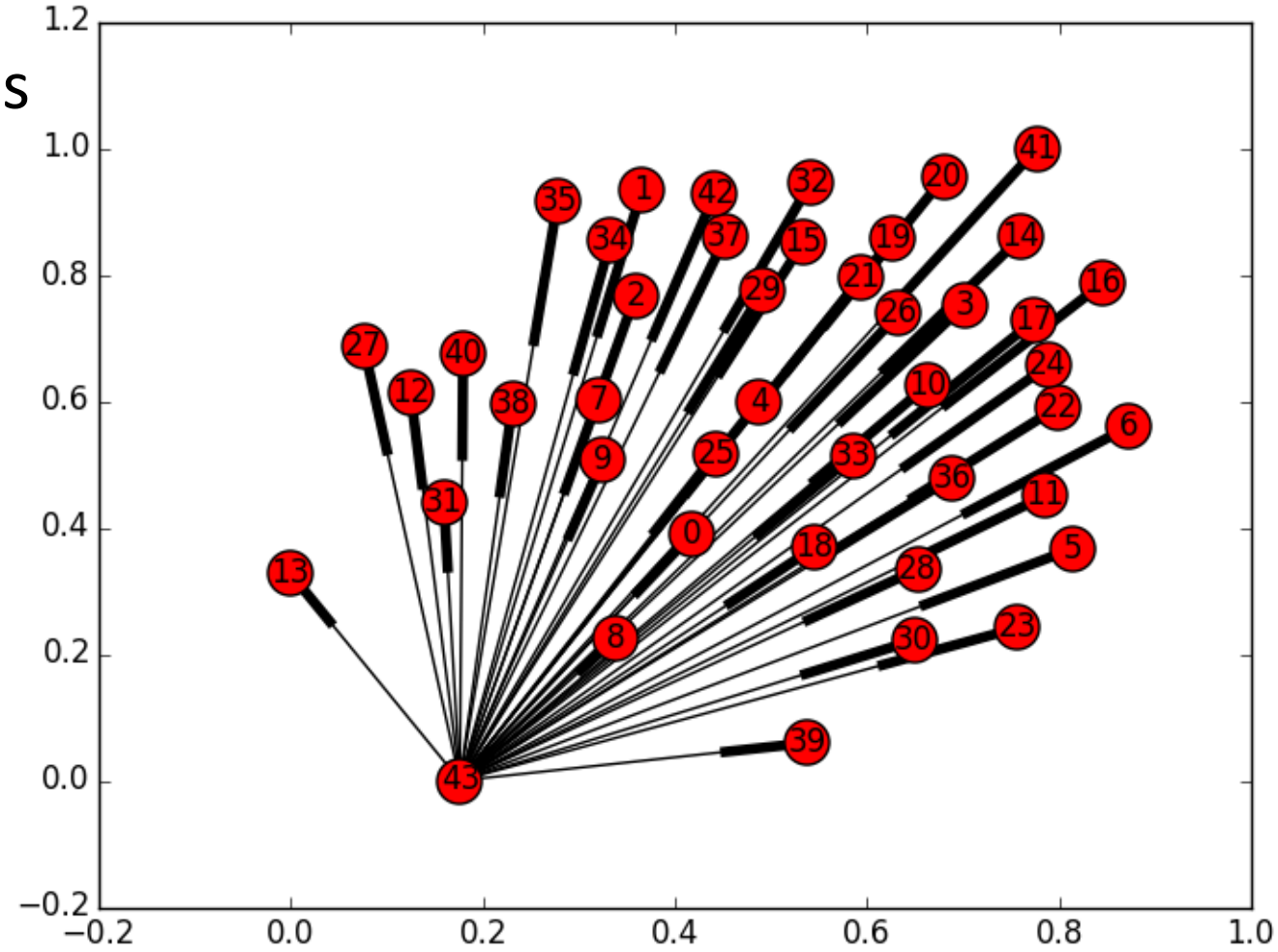
  Auth top 3: liangbianyao 0.0075, zhang-jia-wei 0.0072, ma-bo-yong 0.0068

  PageRank seems more stable than HITS, but this is just the Top 3 case.

Graph Analysis(agree>10k)

- Strongly Connected Components
  - links number: 231416
  - components nodes distribution:
    [1853, 1, 1, 1, 1, 1, 1, 1, 1, …, 1]
  - the first component:
    - average shortest path length 2.11
    - diameter 5
    - radius 2

  A really interesting distribution, like islands surrounding the continent.

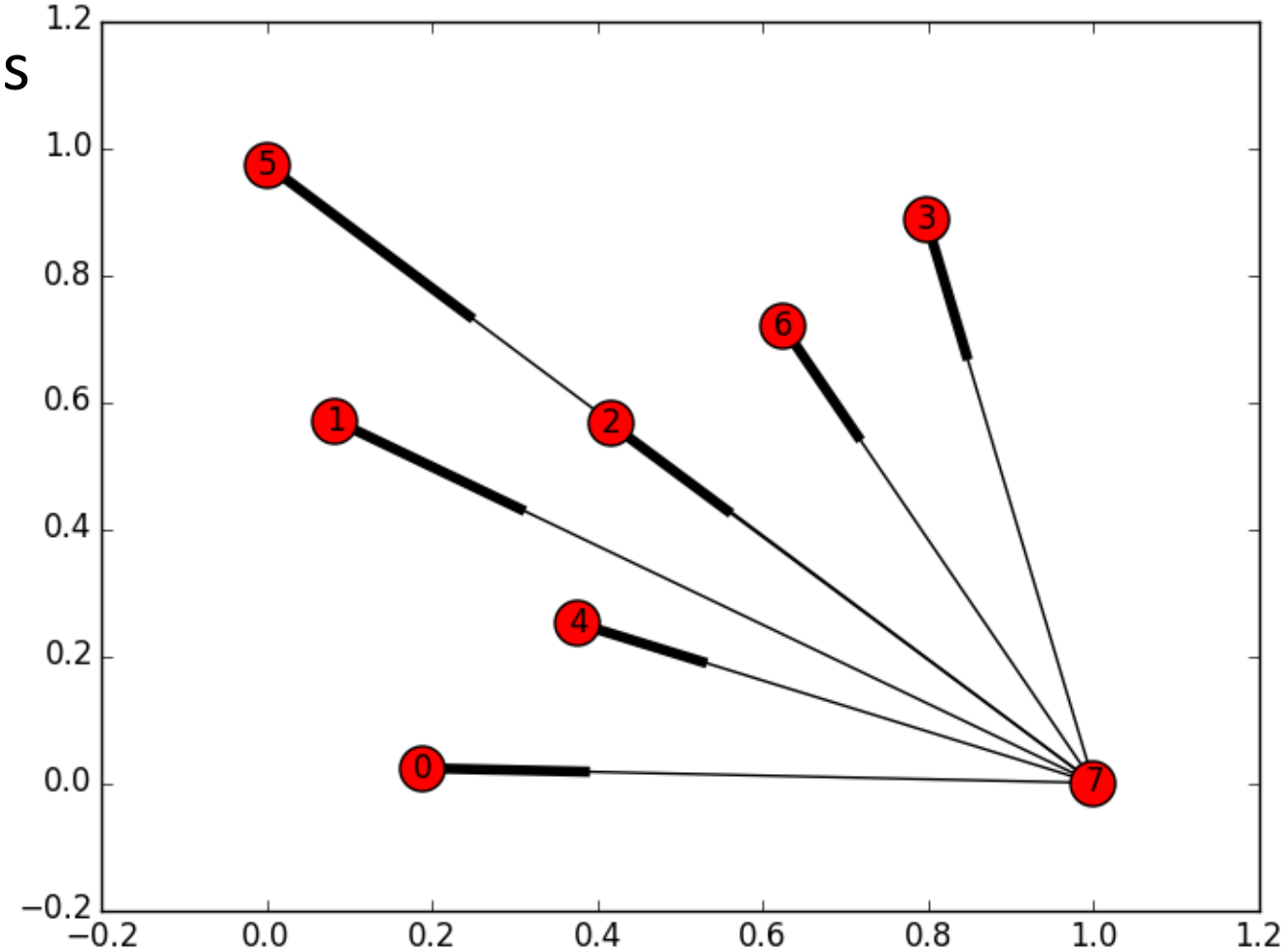


Graph Analysis(agree>50k)

- Strongly Connected Components
  - links number: 27324
  - components nodes distribution:

    [368, 1, 1, 1, 1, 1, 1, 1]
  - the first component:
    - average shortest path length 1.85
    - diameter 4
    - radius 2

  Less ASPL and Diameter.



粉丝数范围	人数	关注次数	理论最大值	关注率	平均关注人数	平均路径长度
10000以上	729	58,453	529,984	11.03%	80.18	1.50
1000以上	3190	332,158	10,169,721	3.27%	104.12	1.74
100以上	18517	2,072,572	342,842,256	0.60%	111.93	2.08

Similar Statistics:
http://zhuanlan.zhihu.com/sulian/19781120 3.5M Users

Graph Analysis(agree>10k)

- Following and Topic

  Dominant set(every user in the set follows at least one user in the group, 220 in 1896)

  We count topics of questions answered by users from the set

  Top 20:

  调查类问题 3792, 生活 3096, 历史 1713, 恋爱 1464, 心理学 1432

  电影 1419, 人际交往 1404, 社会 1332, 互联网 1214, 情感 1197
  政治 1028, 两性关系 994, 教育 897, 中国 823, 人生 815
  游戏 805, 文学 772, 知乎 772, 法律 750, 音乐 738

  爱情 699, 文化 659,创业 628, 大学 621, 程序员 619
  心理 617, 你如何评价 X 609, 女性 604, 编程 585, X 是种怎样的体验 582

Graph Analysis(agree>50k)

- Following and Topic

  Dominant set(45 in 375), count topics of questions answered by users from the set

  Top 20:

  生活 1435, 调查类问题 1365, 政治 1285, 历史 1204, 电影 1084

  健康 996, 社会 984, 医学 941, 恋爱 717, 中国 695

  两性关系 688, 英语 678, 人际交往 640, 心理学 634, 互联网 595

  法律 587, 微软（Microsoft） 555, 美国 552, 健身 538, 编程 511

Summary

- We have covered most of work we plan in the proposal, including: data crawling, basic statistics and graph analysis of both following relationship and topic related patterns of Zhihu.
- There remains loads of interesting issues of such as:

  analysis of the whole Zhihu data (may need months to crawl)

  following-topic network analysis

  topic extraction from the corpus of answers

  …
- We wish we could dig deeper in the future.


Q/A