**Social Network Analysis of Zhihu**

Harbin Institute of Technology (Shenzhen)

Computer Science and Technology

17S051048 冷 佳

Date: 05/03/2018

Contents

[1. Introduction 2](#_Toc513468438)

[2. Data Analysis 2](#_Toc513468439)

[2.1 Overview of Data 2](#_Toc513468440)

[2.2 Basic Statistics of Data 3](#_Toc513468441)

[2.2.1 User Average Characteristics 3](#_Toc513468442)

[2.2.2 Log-Log Distributions 4](#_Toc513468443)

[2.2.3 Agree and Follower Correlation 6](#_Toc513468444)

[3. Network Analysis 7](#_Toc513468445)

[3.1 User Subset Selection 7](#_Toc513468446)

[3.2 Statistics of Network 7](#_Toc513468447)

[3.3 Centrality Analysis 11](#_Toc513468448)

[3.4 Popular Topics Extraction 13](#_Toc513468449)

[3.5 Network Representation Learning 15](#_Toc513468450)

[4. Conclusion 17](#_Toc513468451)

# 1. Introduction

A Social network is defined as a network of relationships or interactions, where the nodes consist of people or actor, and the edges or archs consist of the relationships or interactions between these actors. Social networks and the techniques to analyze them existed since decades. There can be several type of social networks like email network, telephone network, collaboration network. But recently online social networks like Facebook, Twitter, Weibo, Zhihu etc have been developed which gained popularity within very short amount of time and gathered large number of users.

“Zhihu” is the biggest Chinese online community of “Asking-Answering”, till to September 2017, there are already more than 100 million registered users in Zhihu, and there are more than 13 million questions and 46 million answers generated. In such many users and information, there are certainly some inner links, which can be for us very valuable.

In our project, we have tried to analyze the data we downloaded online and find out what kind of form of social network may exists in Zhihu by analyzing the connections of its users. We also analyzed the social network which consists of user following relationship to find out whether the knowledge areas (topics of questions) have any contact with the interpersonal relationships, which is a special issue of Zhihu compared to common social websites like Twitter or Facebook.

# 2. Data Analysis

## 2.1 Overview of Data

The data we used in this project is downloaded online and stored in a SQLite database. The database file have 688MB and 26,000 users, 4.6 million following, 2.2 million questions, 1.7 million user questions, 5.4 million user topics. The schema of the database is shown as follows.

* User (user\_url, user\_id, answer\_num, followee\_num, follower\_num, agree\_num, thanks\_num, layer)
* Following (user\_url, followee\_url)
* Question (question\_id, topic)
* UserQuestion (user\_url, question\_id)
* UserTopic (user\_url, topic)

There are many relationships in Zhihu. Each user can follow or followed by other users, ask or answer questions. Each question can have several topics. Each answer can be agreed or thanked. In this project, we are interested in the user following relationships which would represented in social network in section 3.

## 2.2 Basic Statistics of Data

We use the data to analyze the relationships based on several features.

### 2.2.1 User Average Characteristics

We calculate the mean, the median and the standard deviation of number of followees, followers, answers, agrees and thanks of users. We used the seed user’s features values to be a comparision. The user average characteristics result is shown as follows.

Table User Average Characteristics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Median | Standard Deviation | Compared user |
| Followee | 176.3 | 67 | 565.9 | 149 |
| Follower | 3620.0 | 112 | 22978.5 | 306 |
| Answer | 68.9 | 17 | 225.9 | 120 |
| Agree | 3858.4 | 96 | 21951.4 | 837 |
| Thanks | 865.3 | 28 | 4627.6 | 293 |

We can find that the standard deviations of the five features numbers are so large that the mean values cannot reflect the true distributions of features numbers. We can see the median is much smaller than mean which means there are some concerned users and also not concerned users. This also leads to a hypothesis that the features cannot follow uniform or normal distributions.

### 2.2.2 Log-Log Distributions

We draw the log-log distributions of followee\_num, follower\_num, answer\_num, agree\_num and thanks\_num to user count. The results are shown as follows.

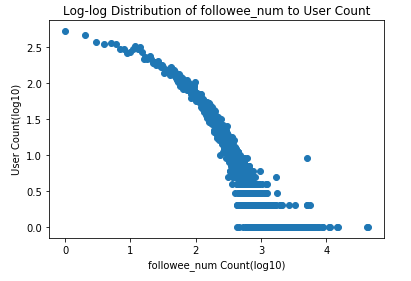


Figure The Log-Log Distribution of Followee\_num to User Count

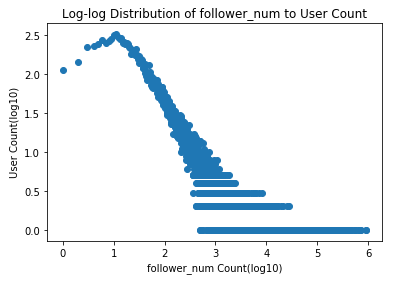


Figure The Log-Log Distribution of Follower\_num to User Count

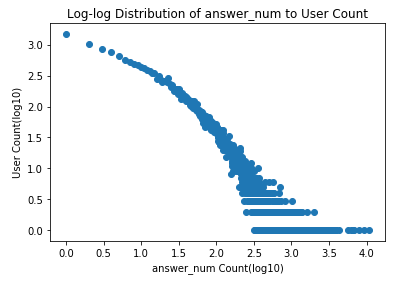


Figure The Log-Log Distribution of Answer\_num to User Count

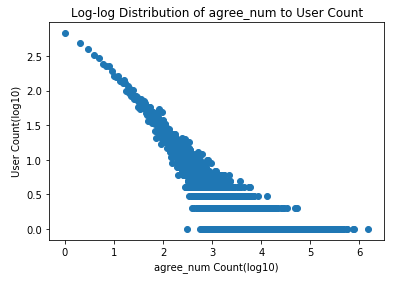


Figure The Log-Log Distribution of Agree\_num to User Count

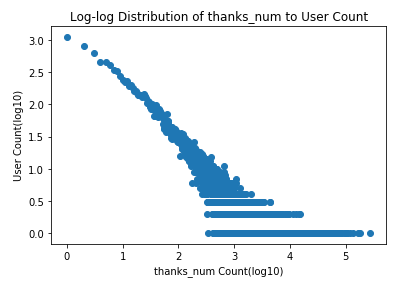


Figure The Log-Log Distribution of Thanks\_num to User Count

Log-log distributions of five features shows a power law which proves our hypothesis to be true. The distributions of followees and followers can be seen as out-dgreee distribution and in-degree distributions of the following relationship which can took as a social network like Twitter.

### 2.2.3 Agree and Follower Correlation

We draw the distribution of agree number to follower number. The result is shown as follows.

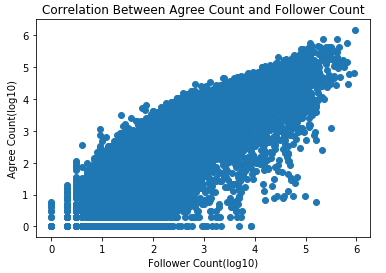


Figure The Crrelation Between Agree Count and Follower Count

We can find that agree number and follower number are proportional. The more agree number users get, the more followers users get. Vice versa.

# 3. Network Analysis

We select user following relationships to analyze the social network. The nodes in the network represent the users, and the edges represent the following relationship which are directed. The main tool we used is Networkx, a popular python graph computing library.

## 3.1 User Subset Selection

We select two subsets from the data for network analysis because of the limitation of time and machines. The subsets are the users who have more than 10 thousand agrees called Net10k and 50 thousand agrees called Net50k. We use the following SQL commands to select subsets.

*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)*

## 3.2 Statistics of Network

First, we draw the graphs to have a overview of user following network. The visualization of Net10k and Net50k are shown as follows.

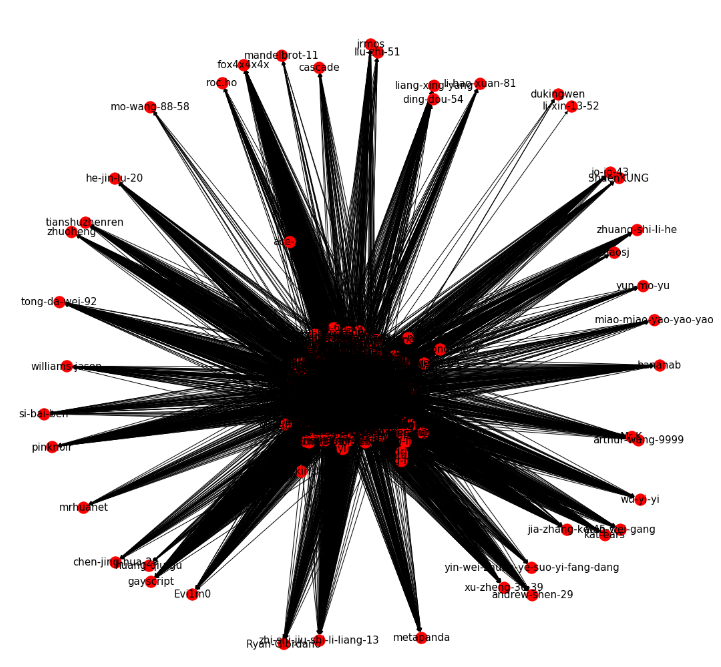


Figure The visualization of the Net10k

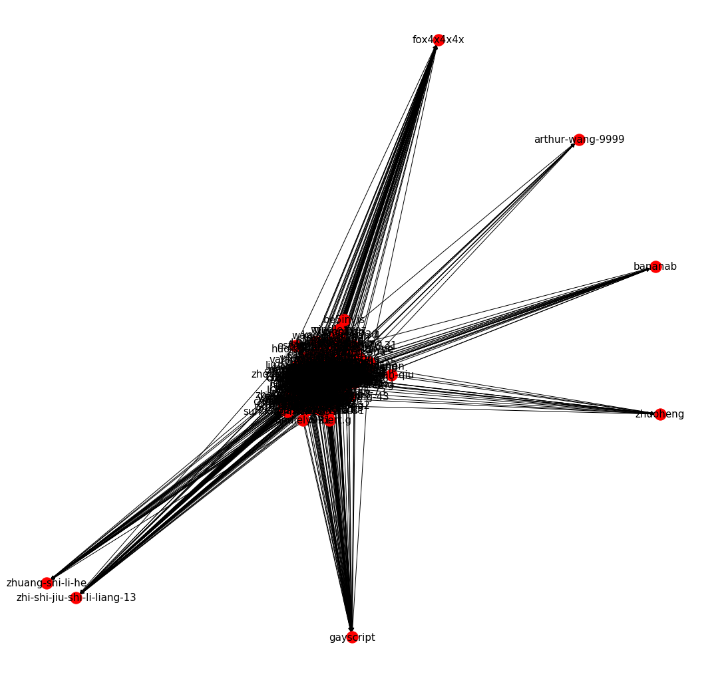


Figure The visualization of the Net50k

Then, we calculate the basic properties of the networks. There are 1896 nodes and 231416 edges in Net10k. The degree distribution of Net10k is shown as follows. We can find that most of nodes have 1.8 degrees, which means that most users only follow 1.8 individuals.

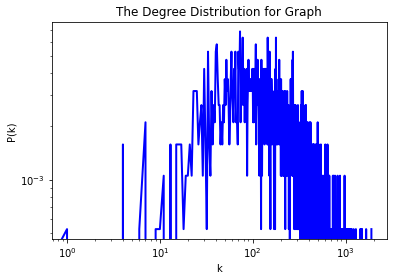


Figure The degree distribution of Net10k

There are 375 nodes and 27324 edges in Net50k. The degree distribution of Net50k is shown as follows. We can find that most of nodes have 2 degrees, which means that most users only follow 2 individuals. We can figure that the degree of Net50k is larger than Net10k which means hot users are more likely to follow more users.

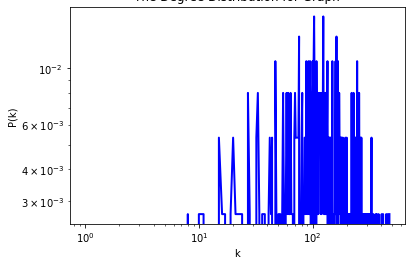


Figure The degree distribution of Net50k

The average lengths of shortest path of Net10k and Net50k are 2.07 and 1.82 which means that one user can know another user by about 2 hops. Because the network is directed, we cannot calculate the clustering coefficients.

The density of Net10k and Net50k are 0.064 and 0.195, which means that there is a larger tendency of clustering for users who have more agree number. The vertex closeness and the vertex betweenness distributions of Net10k are shown as follows.

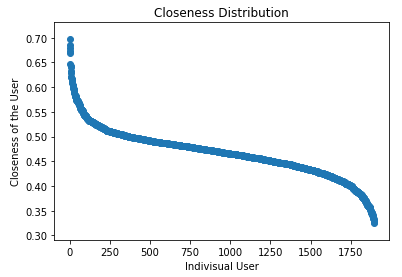


Figure The closeness distribution of Net10k

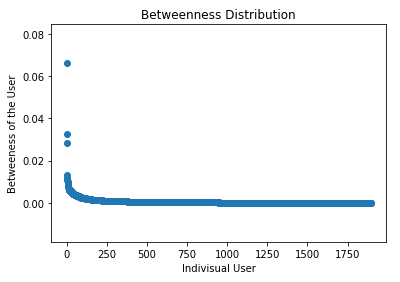


Figure The betweeness distribution of Net10k

Most of closeness of users in the networks is 0.50, which means users are equally important and they are quite close to each other. Most of closeness of users in the network is 0, even the largest betweeness is about 0.066, which means there are no important users that will make the network paralyzed without itself. That is to say, one user can reach another user through any of other users.

The vertex closeness and the vertex betweenness distributions of Net50k are shown as follows which are similar to Net10k.

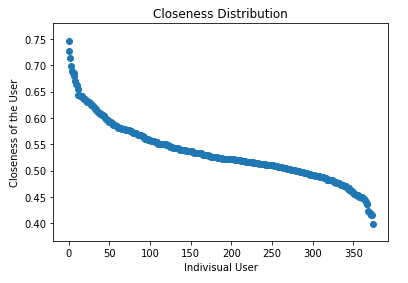


Figure The closeness distribution of Net50k

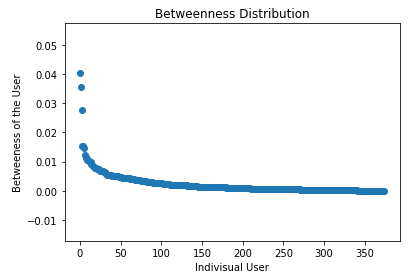


Figure The betweeness distribution of Net50k

## 3.3 Centrality Analysis

We use PageRank and HITS to do centrality analysis. PageRank is a way of measuring the importance of website pages. According to Google: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. HITS is a link analysis algorithm that rates Web pages. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

The top 5 of PageRank and HITS users in networks are shown as follows.

Table Top5 of PageRank in Net10k

|  |  |
| --- | --- |
| PageRank | PR Score |
| jixin | 0.00736 |
| ma-bo-yong | 0.00560 |
| zhang-jia-wei | 0.00551 |
| gejinyuban | 0.00510 |
| zhouyuan | 0.00503 |

Table Top5 of HITS in Net10k

|  |  |  |  |
| --- | --- | --- | --- |
| Hub | Hub Score | Authority | Authority Score |
| zhounuo | 0.00344 | zhang-jia-wei | 0.00345 |
| Namoamitabhaya | 0.00338 | liangbianyao | 0.00340 |
| jun-mo-52 | 0.00336 | gejinyuban | 0.00324 |
| qisini | 0.00325 | ma-bo-yong | 0.00319 |
| wang-wang-wang-08-18 | 0.00289 | jixin | 0.00315 |

Table 4 Top5 of PageRank in Net50k

|  |  |
| --- | --- |
| PageRank | PR Score |
| jixin | 0.01120 |
| ma-bo-yong | 0.01031 |
| zhang-jia-wei | 0.01025 |
| liangbianyao | 0.00949 |
| commando | 0.00903 |

Table 5 Top5 of HITS in Net50k

|  |  |  |  |
| --- | --- | --- | --- |
| Hub | Hub Score | Authority | Authority Score |
| Namoamitabhaya | 0.00976 | liangbianyao | 0.00749 |
| jun-mo-52 | 0.00972 | zhang-jia-wei | 0.00721 |
| qisini | 0.00954 | ma-bo-yong | 0.00683 |
| edison-chen-8612 | 0.00885 | cai-tong | 0.00677 |
| miaomiaomiao | 0.00798 | xiepanda | 0.00670 |

There are some users in both top5 users of PageRank in Net10k and Net50k, which means they are most popular users in different user subsets. The highest PageRank scores of Net10k are smaller than Net50k, which means a larger equality in Net50k than in Net10k and more users will rate the PageRank scores 1.

## 3.4 Popular Topics Extraction

We extract popular topics in the networks. We assume that the topics of questions frequently answered by users are popular topics. In order to extract popular topics, we calculate the domain sets firstly, then we count the frequency of topics from questions answered by users from the domain ser. For Net10k, the dominant set contain 206 in 1896 users, and for Net50k 45 in 375. The top 20 topics of Net10k and Net50k are shown as follows.

Table Top 20 popular topics of Net10k

|  |  |
| --- | --- |
| Topic | Frequency |
| 调查类问题 | 2459 |
| 生活 | 2191 |
| 历史 | 1704 |
| 恋爱 | 1177 |
| 心理学 | 1159 |
| 电影 | 1135 |
| 人际交往 | 1039 |
| 社会 | 1010 |
| 政治 | 818 |
| 游戏 | 781 |
| 音乐 | 744 |
| 互联网 | 732 |
| 情感 | 732 |
| 医学 | 708 |
| 两性关系 | 698 |
| 相声 | 696 |
| 教育 | 689 |
| 中国 | 656 |
| 法律 | 651 |

Table 6 Top 20 popular topics of Net50k

|  |  |
| --- | --- |
| Topic | Frequency |
| 历史 | 1668 |
| 调查类问题 | 1519 |
| 生活 | 1399 |
| 社会 | 1135 |
| 互联网 | 1099 |
| 政治 | 865 |
| 心理学 | 829 |
| 知乎 | 791 |
| 足球 | 771 |
| 电影 | 720 |
| 人际交往 | 705 |
| 两性关系 | 703 |
| 恋爱 | 691 |
| 中国 | 575 |
| 情感 | 542 |
| 经济 | 493 |
| 编程 | 493 |
| 你如何评价 X | 487 |
| 日本 | 487 |

Compared with the results in two user subsets, we can find that the most popular topics in two subsets are 生活，调查类问题，生活.

## 3.5 Network Representation Learning

Network Representation Learning, also known as Network Embedding Method, uses a low-dimensional, dense, and real-valued vector to represent nodes in the network. The vectors contain semantic relationships that facilitate computational storage, eliminate the need to manually mention features, and the heterogeneous information is projected into the same low-dimensional space for downstream calculations. We use DeepWalk and Node2vec to do network embedding.

DeepWalk uses SkipGram method to learn the network representation. We use Random Walk to get the neighborhood of a Graph or Network node. Random walk randomly selects network nodes and generates a fixed-length random walk sequence. This sequence is analogous to a sentence in natural language. The skip-gram model is used to learn the distributed representation of nodes. The representation of nodes in Net50k by DeepWalk is shown as follows.



Figure The representation of nodes in Net50k by Node2vec

Node2vec is similar to DeepWalk, while the main innovation is to improve the random walk strategy. It defines two parameters p and q to achieve a balance in BFS and DFS. Futhermore, it takes into account local and macro information, and have a high degree of adaptation. We use DFS or BFS to sample a node's peripheral nodes. The representation of nodes in Net50k by Node2vec is shown as follows.



Figure The representation of nodes in Net50k by Node2vec

We randomly select some nodes to visualize them which is shown as follows.

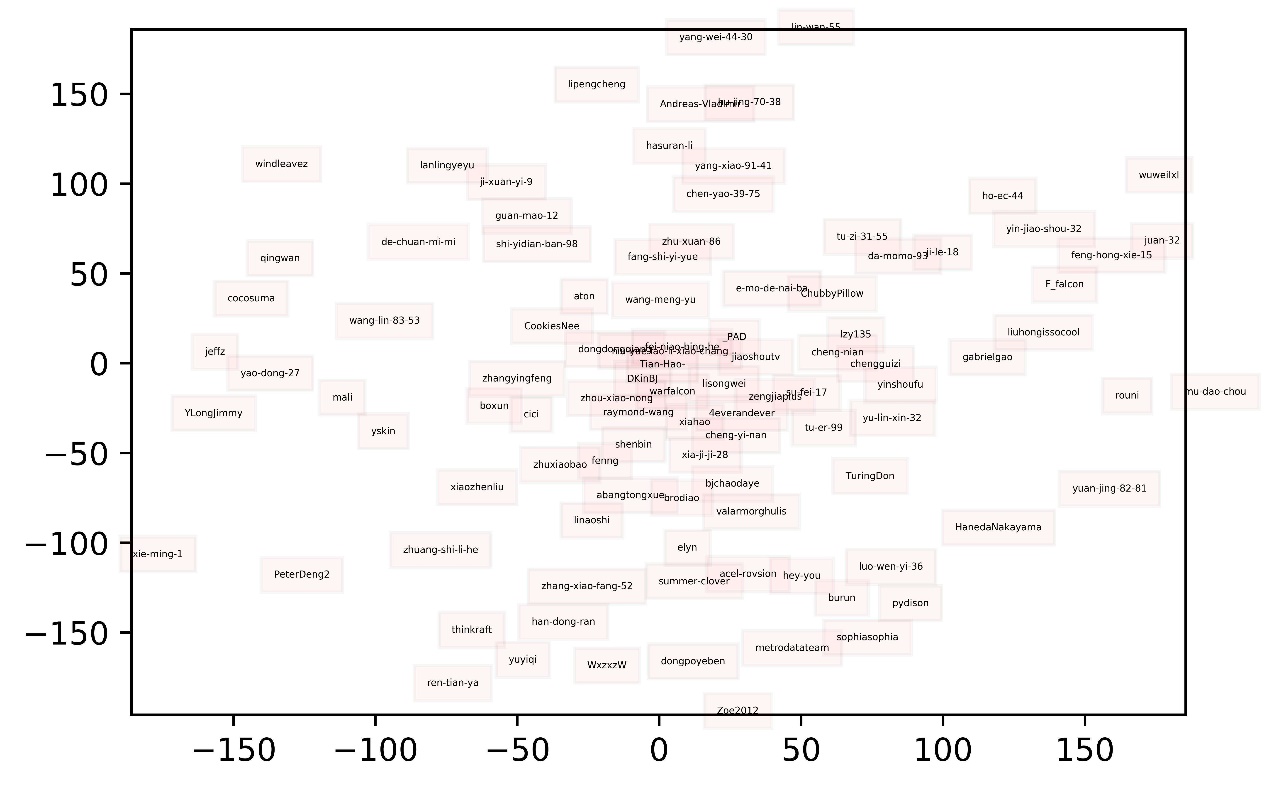


Figure The visualization of some nodes

We do the classification task based on the network embedding. We select one popular topic, and label whether the user interests or professional in this topic. We regard the network embeddings as feature to train the classifier. There are 10 percents of data regarded as test data. The classification result is

# 4. Conclusion

In our project, we have tried to analyze the data we downloaded online and find out the following relationship is a kind of form of social network exists in Zhihu by analyzing the connections of its users. We also analyzed the social network with statistics and centrality analysis. We apply the knowledge we learned in the class in the actual projects and draw lots of conclusions. It really make us supervised and fulfilled.