“Small World Phenomenon”

What is the average length of the chain of acquaintances in SCUT?

Authors：郭吉南 张新镭 杨赏纯 何泽安

Summary of research questions and results

(Tips: social radius measures the number of social connections between any two students)

Q1: What is the average number of social connections between any two of SCUT’s students?

Ans: The 95% confidence interval is (1.925744 , 3.717624).

Q2: Describe the distribution of students’ social radiuses in SCUT. Does it follow a nearly normal distribution?

Ans: The distribution centers at around 2.6 and spreads from 1.95 to 2.95 with a striking departure at 1.95. It doesn’t follow a normal distribution.

Q3: Evaluate the influence of some objective factors to student’s social radius.

1. major in natural science/humanities and social sciences
   There is little difference between students of different kinds of majors.
2. live in which campus (GZIC/HEMC/WuShan)
   A little. GZIC shows the longest social radius while HEMC shortest.
3. local/nonlocal students
   We have 65.73% of confidence to say that it is easier for local students to build a wider social circle at school. (which means smaller social radius)
4. the length of time in school
   The longer one has stayed in school, the smaller the social radius is.

Q4: Evaluate the influence of some subjective factors to student’s social radius.

1. whether joining a club or student organization
   The social radius of student joined a club or student organization is smaller.
2. whether joining the class committee
   The social radius of student joined class committee is smaller.
3. whether active in social activity
   The social active students have smaller social radii.
4. whether has a rich daily life
   Student who has a rich daily life has an average smaller social radius.

Motivation and background

Personal communication is of great importance in social life. We can get more information and resources in the society and also enjoy happiness in the process of communication. But in college, the social contact of students seems to be hindering. We may wonder how well the students in a college is connected. In this project, we will study the interconnection between the social networks behind college students. Many people have trouble making new friends since stuck in a fixed social circle. People feel urgent to expand their social circle but don’t know what to do. Thus, it is of great

significance to study what factors may influence the social condition, which may give some inspirations for expending the social circle.

We found several researches in this field. The first experiment conducted by Harvard psychologist Stanley Milgram[1] in 1960s revealed “The small world phenomenon”. The potential network structure behind human society was firstly probed experimentally, and the result coincided with the theoretical concept “six degrees of separation” originating from a 1929 short story by the Hungarian author, Frigyes Karinthy.[2]Recent years, popular social media platforms like Tencent QQ and Facebook have been used to study the social network behind online-society, 2.81 and 4.57 averaged degree of separation behind these two platforms were summarized by Lei Zhang.[3] In 2014, students from Cornell University conducted a similar experiment as Milgram’s in their campus and concluded with an average 1.25 degree of separation.[4] These results are different and less than the original 6 since people are closer in online-society and university is a greatly scaled-down version of human society.

Dataset

We design a questionnaire to get students’ information we concern and conduct a survey in South China University of Technology to collect data. At last we have collected 535 questionnaires. Next, we do data cleaning and generate a dataset containing 535 cases with 45 attributes shown as follows.

Table 1

Attribute	Value	Description
1.your_campus	GZIC/HEMC/WuShan	International campus/University town/WuShan campus
2.your_acade	A~AG	A~AG are mapping to 36 academies
3-38.A~AG	3/2/1.5	The number of your friends in X (A~AG) 3 : >=10friends 2 : (4,10)friends 1.5 : (1,3)friends
39.your_grade	Freshmen/sophomore/junior/senior/graduate	
40.club_stud_org	0/1	Have/haven’t joined a club or student organizations
41.class_commi	0/1	Haven’t/have joined a class committee
42.soci_acti	0/1	Non-active/active in social activities
43.dali_live	0/1	Poor/rich daily live
44.local	0/1	Nonlocal/local student
45.science	0/1	Major in humanities or social sciences/natural science

Methodology

1. Social network model

1) Basic element

The social network is modeled as an one-mode graph shown as figure 1 (a), which is composed of nodes, edges and length of edge.

Node : Each student is modeled as a node in the graph.

Edge : There is an edge between two acquainted students as a modeling of the social relationship. Thus one is connected with all students he/she knows.

Weight (Length of edge) : The length of edge measures the social distance between two acquainted students.

2) Assumptions

Assumption 1 : All students in the same academy are acquainted with each other.

Assumption 2 : Any two students in one sample don't share a same acquaintance in an arbitrary academy.

Assumption 1 is reasonable to some extent in SCUT since students from same academy usually take the same courses. Meanwhile, assumption 2 is guaranteed by the fact that the sample size is much less than population (the number of whole students in SUCT). So there is little chance to pick two students knowing actually the same student in one sample.

With these two assumptions, we can add an edge to all nodes from the same academy as the figure 1 (b) shows and assure that all the students from the same academy in one sample are different individuals.

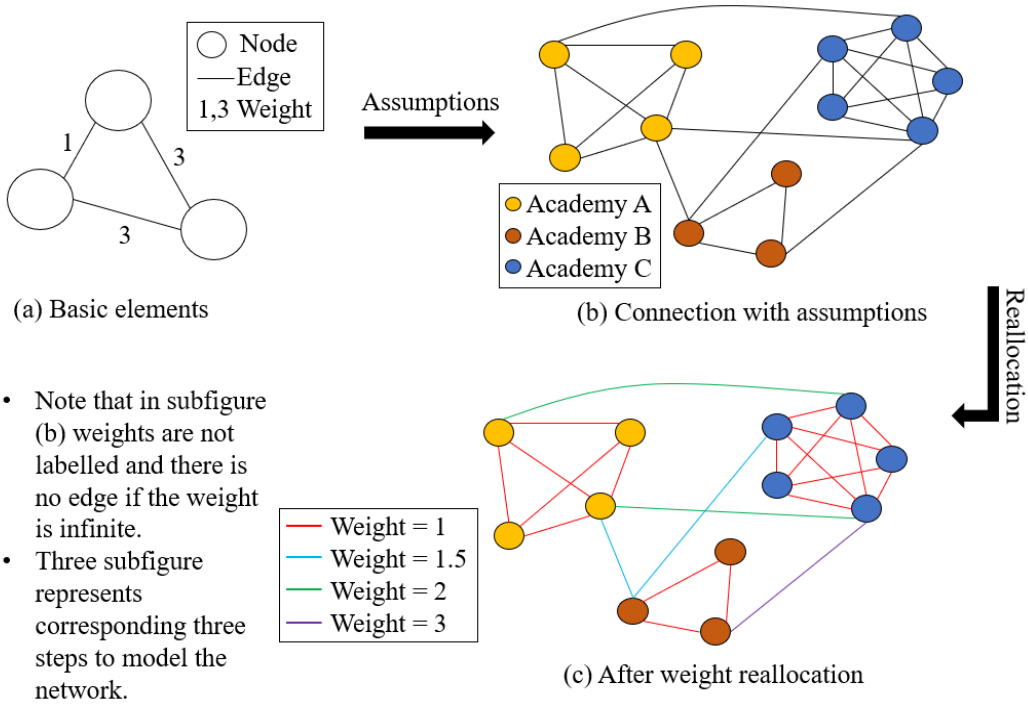


Figure 1 Network diagram for the modeling process

3) Reallocation of edge lengths

The lengths of edges can be classified into two types with different assignment criteria shown as table 2.

Type 1: According to assumption 1, length of all edges between two nodes from the same academy are 1.

Type 2: The length of edge from one node to nodes from different academies is classified into three levels according to the number of friends in the neighbor node's academy. These levels measure the social distance between one node and a group of nodes from that academy.

Num_frien_in_x	**0**	**1~3**	**4~10**	**>10**
Edge Length	**No edge**	**3**	**2**	**1.5**

Table 2 Mapping form attribute to edge length

2. Algorithm for sample statistics

1) Description of library: ***igraph***

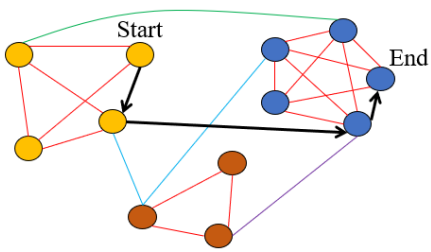
igraph is an open-free package in R language for graphs and network analysis. It can handle large graphs very well and provides functions for graph visualization and graph algorithms.[7] We use ***igraph*** to convert the social network into graph and compute shortest distances between any two node we are interested in.

2) Sample statistic d_1 : the average degrees of separation among students in one sample

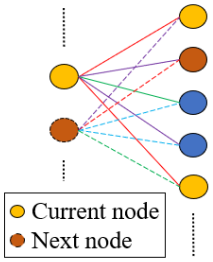
Use ***igraph*** to compute the shortest distance from any node to all the other nodes. Compute the average value to get d_1.

3) Sample statistic $\overrightarrow{d_2}$: the radius of the social circle of students

Computed the mean of the shortest distances from one node to all the other nodes separately, store them into a vector $\overrightarrow{d_2}$.



(a) The process of finding the shortest from the start nodes to the end nodes



(b) Traverse all nodes

Pseudocode of the algorithm

For Node i in the set of all nodes:
  For Node j in the rest nodes:
    $\delta_{i,j}$ = *shortest distance* between (Node i, Node j)
  End
  The social radius δ_i of Node i $= mean(\delta_{i,j})$
End
$d_2 = [\delta_1, \delta_2, \delta_3, \ldots]^T$
$d_1 = mean(d_2)$

Figure 2 Illustrations and pseudocode for the algorithm

3. Statistical analysis

1) Examine the distribution of social radius d_2 in the sample

Details about these four distributions' properties and examine methods are listed in the Table 3.

Property	Description	Examine Method
Shape	Distribution curve's shape	Histogram
Spread	Intrinsic variability	Boxplot / Five-number Summary
Center	Average value or Middle value	Mean / Median
Normality	Similarity to Normal Distribution	Normal Quantile Plot

Table 3 Methods to examine the distribution of social radius

2) Inference

a) 95% confidence interval of the average degree of separation d_1

From the algorithm described in Section 2, we can obtain all the shortest distances among all nodes in a sample, which can be represented by an adjacent matrix, D. The matrix D is square and any row or column stores the shortest distances from one special node to all the other nodes in the sample. Thus, we can easily compute the mean and the standard deviation of all the shortest distances in a sample as follows.

$$Mean\ \bar{\boldsymbol{d}} = \frac{Summation\ of\ all\ distances}{Amount\ of\ distances} = \frac{\sum Elements(D)}{Row\ length(D) \times Column\ length(D)} \quad (1)$$

$$Standard\ Deviation\ \boldsymbol{s} = \frac{\sum[Elements(D) - \bar{d}]^2}{Row\ length(D) \times Column\ length(D) - 1} \quad (2)$$

Using these two quantities, we can calculate the 95% confidence interval of the average degree of separation as follows.

$$Lower\ bound = \bar{d} - 1.96 \times \frac{s}{\sqrt{Row\ length(D) \times Column\ length(D)}} \quad (3)$$

$$Upper\ bound = \bar{d} + 1.96 \times \frac{s}{\sqrt{Row\ length(D) \times Column\ length(D)}} \quad (4)$$

$$95\%\ Confidence\ Interval\ = [Lower\ bound, Upper\ bound] \quad (5)$$

b) Inference between $\overrightarrow{d_2}$ and other attributes

We can separate the population into two groups using different criterion according to the attributes. And we're to compare distributions of social radius in two populations. We explain ***the two-sample t significance test*** technique to compare the mean responses in the two population which we'll employ later.

Suppose that an SRS of size n_1 is drawn from a normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another normal population with unknown mean μ_2. To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values for the t(k) distribution, where the degrees of freedom k are the smaller of $n_1 - 1$ and $n_2 - 1$.[8]

Results

Q1: What is the average number of social connections between any two of SCUT's students?

The computation shows the 95% confidence interval is (1.925744 , 3.717624).

Q2: Describe the distribution of students' social radiuses in SCUT. Does it follow a nearly normal distribution?

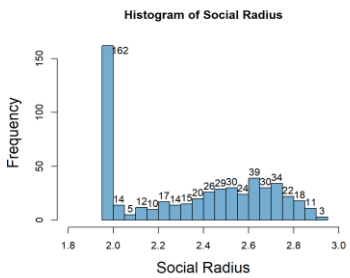
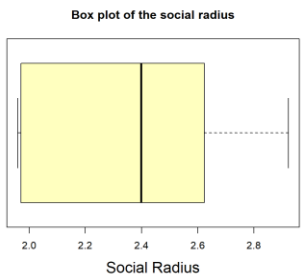


Figure 3 Histogram and box plot for population



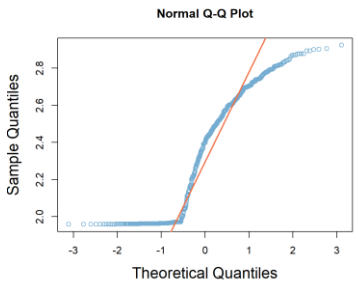


Figure 4 Q-Q plot for population

From histogram and box plot in figure 3, we can see the distribution centers at around 2.6 and spreads from 1.95 to 2.95 with a striking departure at 1.95. The Q-Q plot in figure 4 shows the points are quite departure from the line which indicates that the distribution of the students' social radiuses in SCUT is not normal. But we noticed that the data from 1.95 to 2.0 take the majority. So we eliminate the data larger than 2, and plot it again. The rest of data shows a nearly normal distribution from the plots in figure 5.

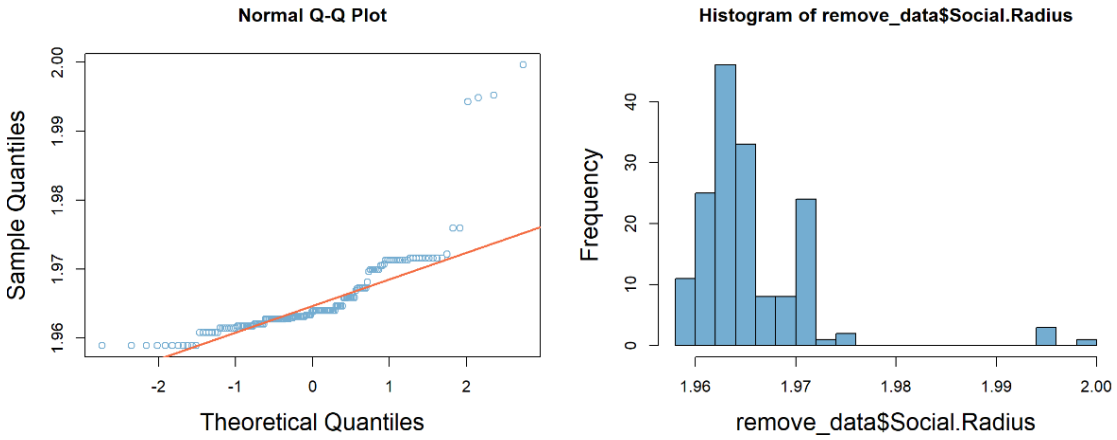


Figure 5 The histogram and Q-Q plot for the data below 2

Before we do the question 3 and 4, we made some plots to get more information about the dataset. Figure 6 shows the composition of the academies the students from. And figure 7 and 8 shows the division of some attributes.

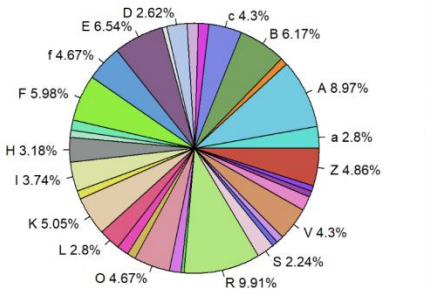


Figure 6
Pie chart of students from different academes

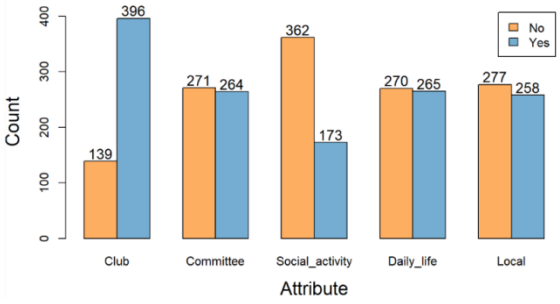


Figure 7
The division of attributes

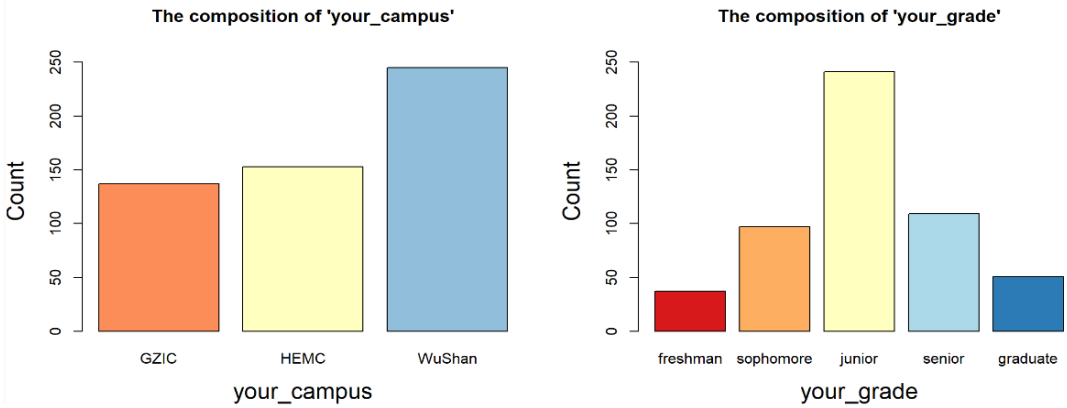


Figure 8 compositions of some attributes

Q3: Evaluate the influence of some objective factors to student's social radius.

1. major in natural science/humanities and social sciences

   It is convenient to verify the assumption by using the inference function. The p-value is quite large, which means we fail to reject Ho, so there is no difference in the social radius between students majoring in natural science and in other like humanities and social sciences, verified by figure 9.

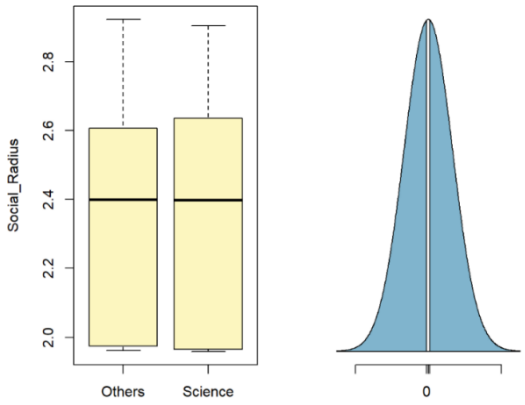


Figure 9 Inference about major

2. live in which campus (GZIC/HEMC/WuShan)

   Make a mosaic plot figure 10 about the social radius and the campus. We find that the most frequent "social Radius" is 2. Except for 2, the "Social Radius" are almost normally distributed with different centers. Among three, the social radius of students in GZIC centers at highest value, and the social radius of students in HEMC centers at lowest value.

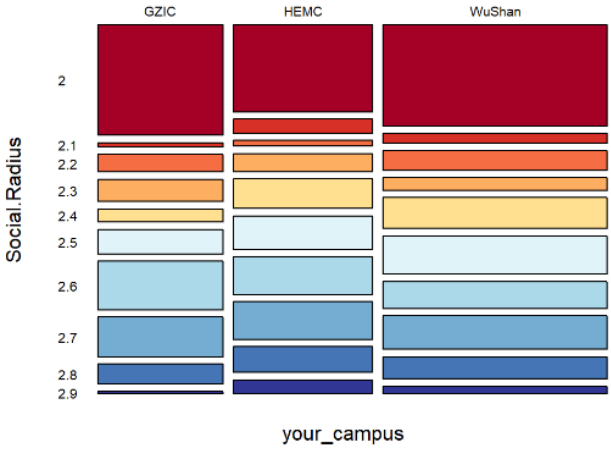


Figure10 Social radius distributions in three campuses

3. local/nonlocal students

   Conduct a test of significance to see if it is easier for local students to build a wider social circle at school than non-local students. The p-value = 0.3424. So we have 65.73% of confidence to say that it is easier for local students to build a wider social circle at school.

4. the length of time in school

   We use box plot figure 11 to show how the length of time in school relates to the average degree of separation. Discover that as the length of time in school increases, the main body of the social radius move downwards, which means it is easier for them to find an arbitrary

student in SCUT. You might have noticed that there is a rise in graduate, this is because some graduates are new to his school.

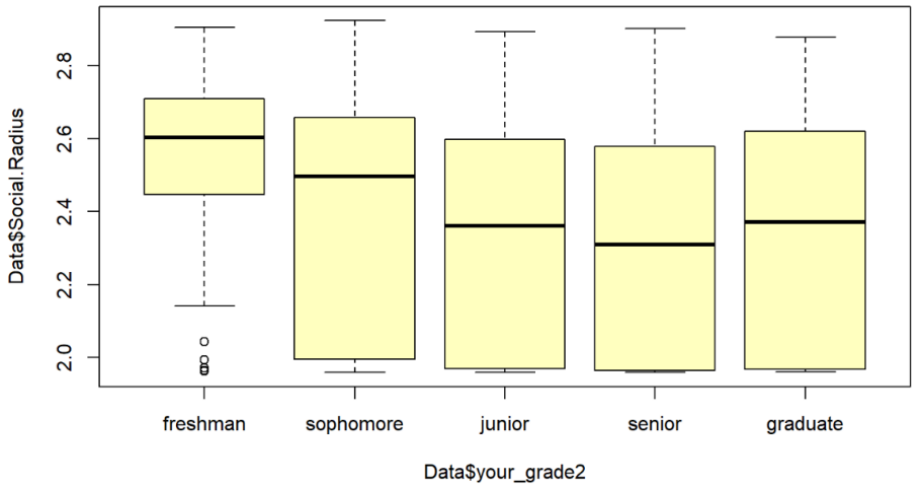


Figure 11 Box plot of 'Social.Radius' to 'your_grade'

Q4: Evaluate the influence of some subjective factors to student's social radius.

1. whether joining a club or student organization

   The result (Figure 12) shows that P-value is 5e-4 which is quite small so that can't show in the figure. So it is clearly to know that the social radius of students who joined a club or student organization is smaller than who did not.

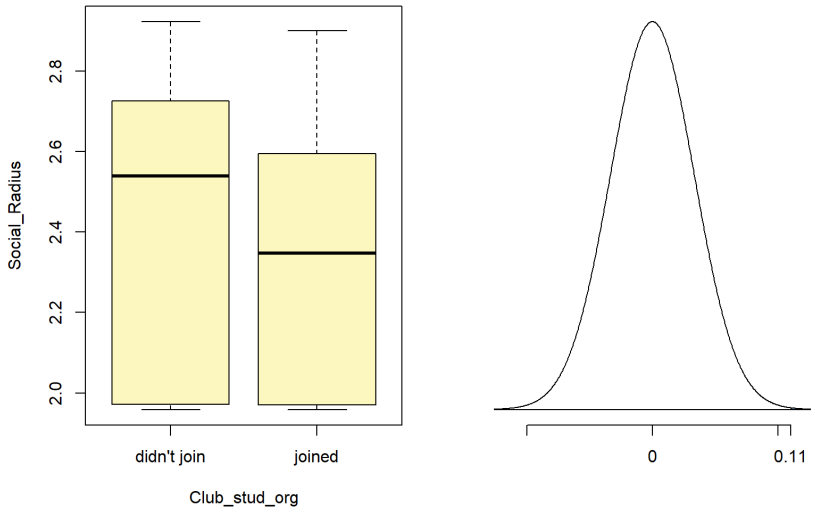


Figure 12 Inference about club and students organization

2. whether joining the class committee

   The 95% Confidence interval of the social radius of student joined class committee is (2.2823 , 2.3573), while (2.3396 , 2.415) for who didn't. And the social radius of student joined class committee is smaller.

3. whether active in social activity

   The significance test shows the P_value = 0 (very small). So we have enough evidence to reject Ho and accept Ha. From the data, it's clearly that social activity is definitely affecting the social radius. The social active students have smaller social radiuses.

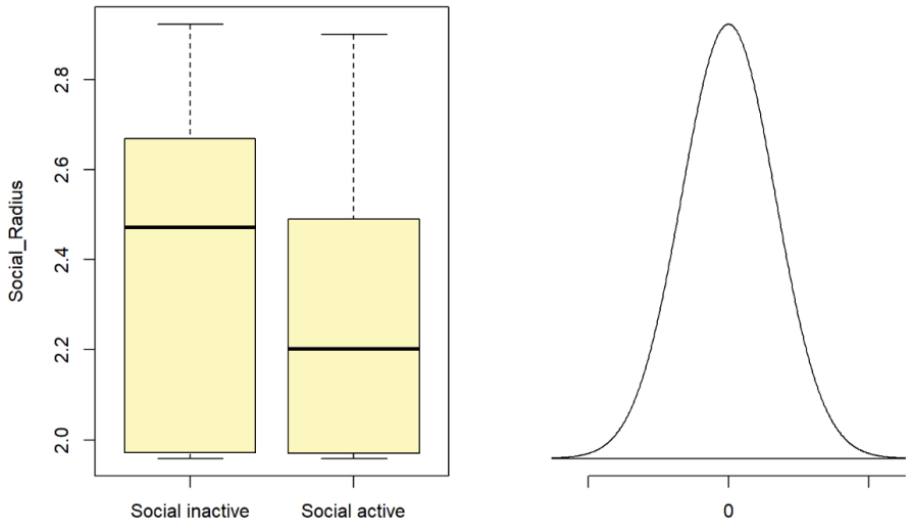


Figure 13 significance test for social activities

4. whether has a rich daily life

   From figure 14 we could find the mean value, median and 3rd quarter value of rich daily life students are smaller than students who have monotonous life. For the data larger than 2.0, the distribution of monotonous daily life students' social radius is unimodal whose center is within 2.7 to 2.8 and left-skewed. And the distribution of rich daily life students' social radius is unimodal whose center is within 2.6 to 2.7. The data change of monotonous daily life students sharper than rich daily life. Above all, we conclude that student who has a rich daily life has an average smaller social radius.

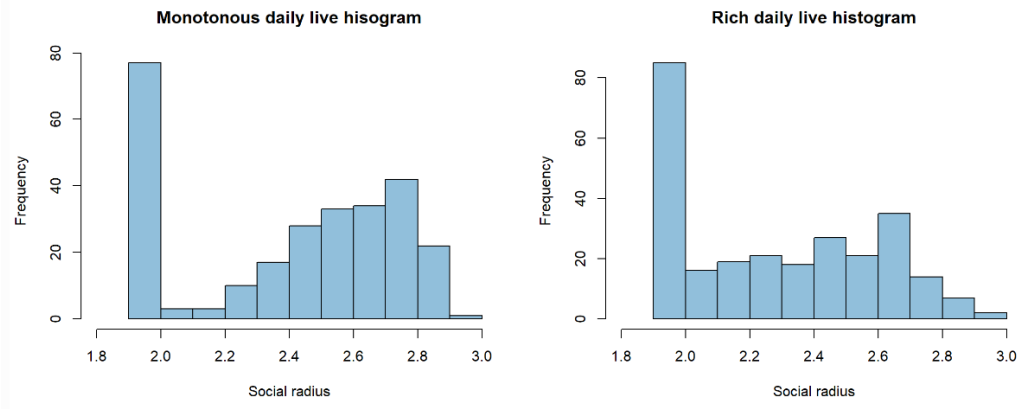


Figure 14 Distributions for comparing rich or boring daily life

Collaboration

Thanks for the students who have helped us in distributing questionnaires, listed below.

李鸿基 张泽鸿 施润晨 练方熙 金雨辰 梁荣辉 丁文华 毕帷帧 徐祥 梁瑶
叶灵辉 符恩豪 余华斌 王恺文 郑伟木 方凯荣 李艺豪 陈坚文 姜建强 郭俊伟
吴梓聪 白昕昱 黄立刚 姚泽坤

Reflection

1. **The evaluation of project results**

   1) Accuracy

   The final conclusion we draw about the average length of the chain of acquaintances in SCUT is around 2.84, which is close to 2.81 from the experiment conducted through Tencent. Its' also reasonable comparing to 6 in Milgram's experiment since university is a scale-down version of the human society.

   2) Discussion about shortages

   The amount of our dataset is 535 which may cause a bias in our final results, since our dataset is relatively small comparing to the total population in SCUT.

   The mathematical model is built with some reasonable assumptions, but the error in results caused by model is inevitable.

2. **Learn from the project**

   As for the implementation of our project, we employ statistical and analysis knowledges taught in theoretical courses to solve a practical problem, where we not only deepen our understanding of statistical theories, but also improve our ability to put statistical and analysis techniques into practice.

   Team collaboration is of great importance. Our project couldn't have been completed as we expected without the team leader and all team members' efforts. Every problem confronted was divided and conquered with at least two team members.

Reference

[1]: https://en.wikipedia.org/w/index.php?title=Stanley_Milgram&oldid=1114238500

[2]: https://en.wikipedia.org/w/index.php?title=Frigyes_Karinthy&oldid=1119380741

[3]: https://en.wikipedia.org/w/index.php?title=Six_Degrees_of_Kevin_Bacon&oldid=1116970868

[4]: MR: Collaboration Distance (ams.org)

[5]: Zhang, Lei & Tu, Wanqing. (2009). Six Degrees of Separation in Online Society.

[6]: http://www.analytictech.com/connections/v20(2)/smallworld.htm

[7]: igraph R package

[8]: Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. W H Freeman/Times Books/ Henry Holt & Co.

Appendix

Gantt chart of our project progress