MARCH 6, 2018

DNSC 6215 SOCIAL NETWORK ANALYTICS

Jingbo Zhang
Elvis Leng
Parul Jain

Abstract

Frequent contact among students themselves and students and teachers in elementary schools is critical for providing a personal and effective education. This report focus on the analysis of a contact network in one elementary school. The main finding of this project is about students' behavior patterns. First, we find that the students in low grade tend to communicate more with others. But when they get older, they tend to stay in their own circle more. Second, students are found to lean to make friends with students in same class. Third, it's relatively easy for students to form weak contact with heterosexual students. However, students tend to choose same-gender students as their close friends. Fourthly, students with close relationship may share 13 or more close friends together. In addition, we find that in this school, the communication between teachers and students is absolutely insufficient. Teachers may need to take corresponding action to improve this.

Video link: https://youtu.be/JO8TGH2VD58

1 Introduction

1.1 Motivation

Communication plays an important role in education, especially in elementary school education. Some previous research shows that interpersonal communication is related to students' academic performance and will definitely affect one's personality. Therefore, we're interested in analyzing a contact network in one elementary school, trying to discover some interesting patterns related to students' and teachers' behavior.

The initial objective for this project is to answer if there's sufficient student-student and teacher-student communication in this school. But when we dig deeper, we find more noteworthy things. After the entire analysis, we're not only able to answer the above question, but also gaining some idea with the motivation of generating close relationship among students.

The rest of the report is organized as follows: In section 2, we do some basic univariate analysis on both two networks. In section 3, we analyze contact only between teachers and students. We then develop two block models based on class and gender respectively in section 4. Section 5 covers a QAPq logistic regression in finding the relationship between close communication and number of common friends.

1.2 Dataset Introduction

2 Grade	No. of Students
1	48
2	49
3	45
4	44
5	46

Table 1. Students' grade distribution

The dataset we used is downloaded from http://www.sociopatterns.org. This dataset contains face-to-face communication records in an elementary school between students and teachers as well as among students themselves. In this dataset, each node represents one individual. Each edge between nodes represents face-to-face communication between individuals. This data set contains 242 nodes, including 232 students across 5 grades and 10 teachers. All nodes have attributes attached to them: anonymous ID, type (teacher or student), grade and class a student belongs to, and the gender of each student. The basic statistics of students' grade and gender are shown in table 1 and 2.

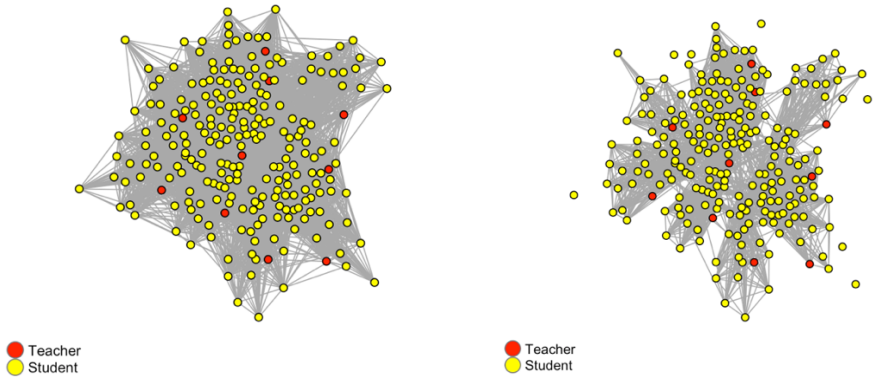
Gender	No. of Students
F	112
M	115
Unknown	5

Table 2. Students' gender distribution

2 Network Metrics Analysis

2.1 Initial Exploration

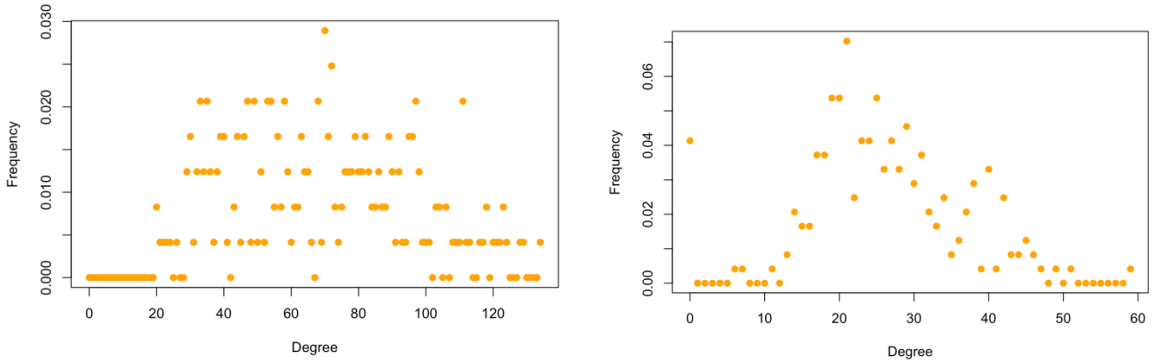
Since the dataset contains two undirected networks, each of which represents communication records in one day. We first merge the two networks to get the whole contact graph. Then we defined strong contact graph, which consist of same population and communications which are happened for both of the days. Contact graph is an aggregation of weak ties in this population while strong contact graph represents close relationship between individuals. By visualizing these two networks, we can see directly from figure 1 that contact graph is a completely connected network. However, we can find some obvious isolated nodes in strong contact graph, which means that there are some students that do not have close relationship with other people. Students with 0 degree in strong contact graph may be introverted and may need more participation in social activities.



(a) Contact graph (b) Strong contact graph

Figure 1. Visualization of both networks

2.2 Degree Distribution

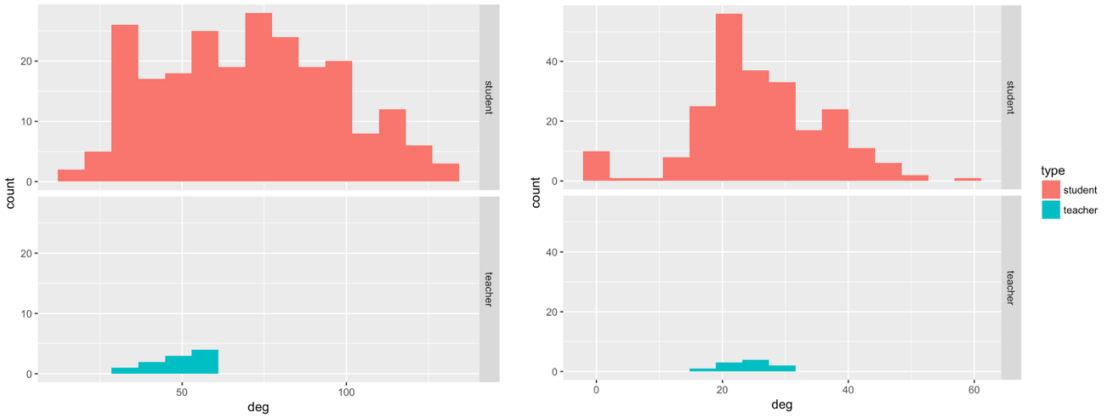


(a) Contact graph (b) Strong contact graph

Figure 2. Degree distribution

First, we look at the degree distribution in two networks among whole population. From figure 2, we can get the conclusion that strong contact network has more narrow degree distribution. In addition, distributions in two networks are close to normally distributed.

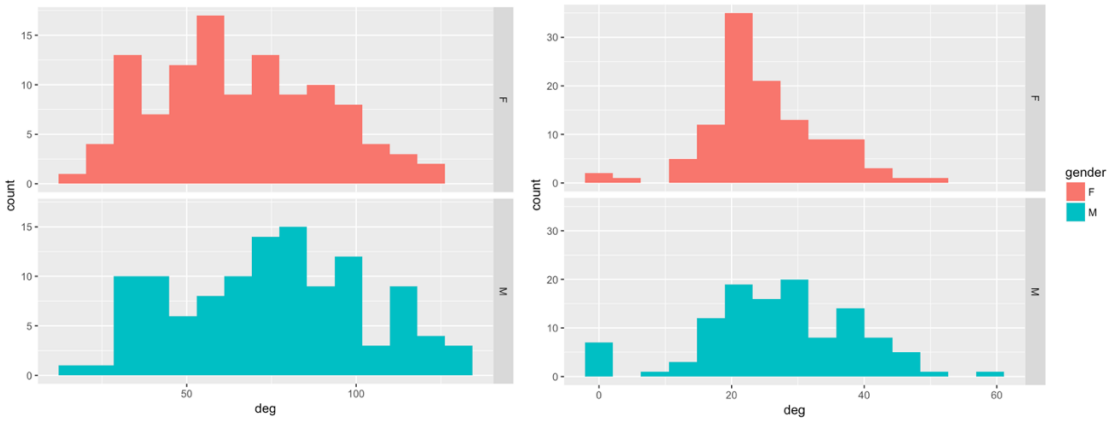
Next, we compare the degree distribution in students and teachers. From the degree of the strong contact graph, we can find that teachers tend to have lower degree than students on average in both graphs. In strong contact graph, highest frequency of the degree is around 20. there are almost no students who have degree below 10 (except for isolated students) or above 50, which implies that there are very less students that have no interactions.



(a) Contact graph (b) Strong contact graph

Figure 3. Degree distribution versus type

When we plot students' degree by gender, we can see that the degree distributions of boys or girls are almost identical in contact graph. But in strong contact graph, girls tend to have narrow distribution. In addition, all of the students that have highest degree are males. We can conclude that boys are probably more likely to become the 'star' in elementary school. However, most of the isolated students in strong contact graph are males, which indicates that girls, on the whole, are better at generating at least one close relationships with each other .

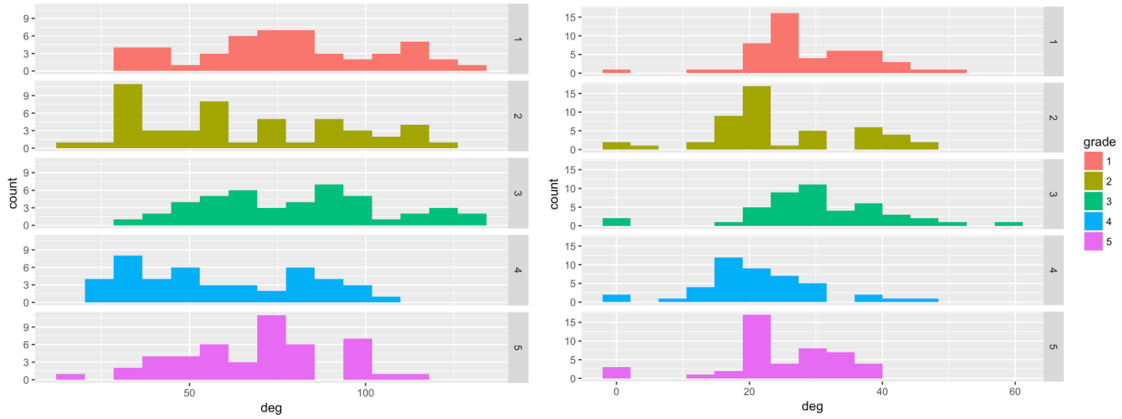


(a) Contact graph (b) Strong contact graph

Figure 4. Degree distribution versus gender

When we divide the contact degree to different grades, we find something strange. We can see that people in lower grades tend to have more contact with each other both in the weak contact network and the

strong contact network. This finding is not extremely shocking, but still surprised us. This implies that younger students tend to communicate with more people in order to make more new friends. But when students get older, they tend to stay in their own circle more.



(a) Contact graph (b) Strong contact graph

Figure 5. Degree distribution versus grade

2.3 Basic statistics

	Contact Graph	Strong Contact Graph
Average Degree	68.7273	25.8017
Reachability	1	0.9192
Mean Shortest Path	1.8505	4.6783
Density	0.2582	0.1071
Diameter	3	10
Transitivity	0.4798	0.5492

Table 3. Basic Statistics

Table 3 shows some basic statistic in two networks. Average degree in each network represent the average number of contacts one individual may have, taking students and teachers as a whole. The less-than-one reachability in strong contact graph indicates that there are some isolated nodes in the graph.

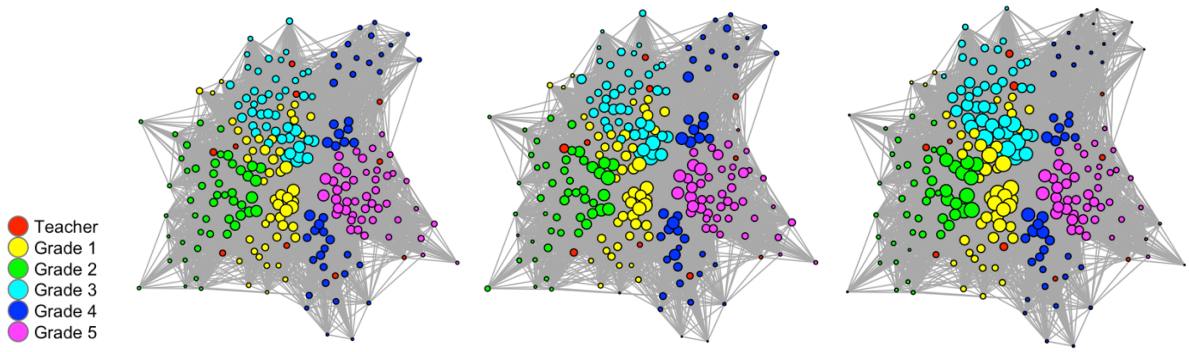
Some of these statistics correspond to the strength of weakness. The diameter and mean shortest path of contact network is less than these of strong contact graph, which indicates that it's better to spread information in weak relationship network because it takes less time to complete this spread to most individuals except for isolated nodes.

2.4 Centrality Analysis

With centrality analysis, teachers can find some 'star' students in this school. This may help to provide suggestion to question like which student should be assigned with tasks like planning social activities in school and which student should be choose for public speaking.

What's more, table 4 shows the correlation of all four centralities. We can find that degree centrality and eigenvector centrality have highest correlation. This means that nodes having high degree are also connected with important nodes, which may implies that students that have high degree, which can be seen as important nodes, get to communicate with other important students. This conform with the

character of interpersonal communication since in real life, outstanding person tends to make friends with other outstanding persons.



(a) closeness (b) betweenness (c) eigenvector

Figure 6. Centrality of nodes in contact graph

	Degree	**Closeness**	**Betweenness**	**Eigenvector**
Degree	1.0000	0.9421	0.8875	0.9712
Closeness	0.9421	1.0000	0.8616	0.8746
Betweenness	0.8875	0.8616	1.0000	0.5256
Eigenvector	0.9712	0.8746	0.5256	1.0000

Table 4. Correlation of Centralities

3 Teacher-student Communication Analysis

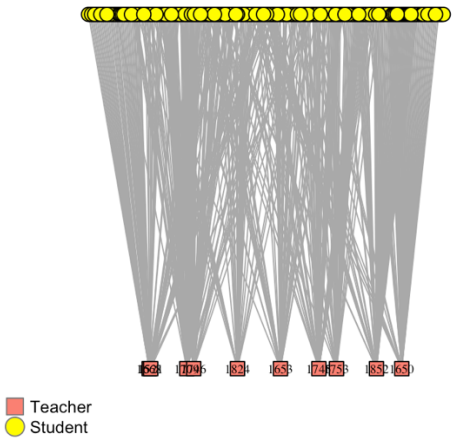


Figure 7. Teacher-student communication network

After seeing the school as a whole, we are going to dig deeper into the relationship between students and teachers. By deleting edges representing student-student and teacher-teacher communications, we get a subgraph of contact network, which only includes communication between students and teachers. This subgraph contains 432 edges. Analysis in this part will base on this subgraph.

The average degrees of teachers and students in this network are 43.2 and 1.8621 respectively. This means that on average, each one student is communicated less than 2 teachers. Then we rank the degree

by ascending order to see if there's anything under concern. By doing this, we find there're three students whose degree are 0. This is alarming because communication is an important segment for education, especially for elementary school education. Teachers need to be paying more attention to these 0-degree students.

Anonymous Name	Degree
1766	0
1775	0
1799	0

Table 5. Students that have 0 degree

We then looked at the contact frequency for each grade separately, and we discovered that at this school two teachers teach each grade. Taking grade 2 as an example, this subgraph is shown in figure 8. Only teacher 1650 and 1852 have tight communication between students in grade 2 while other teachers seldom communicate with students in grade 2. Graphs representing relationship in other grades can be found in the markdown file.

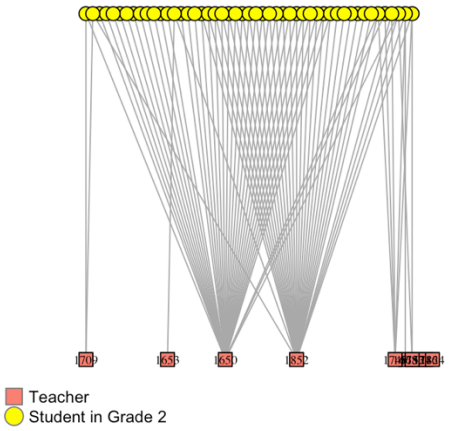


Figure 8. Subgraph of grade 2

Anonymous Name	Degree
1824	31
1521	33
1753	34

Table 6. Teachers that have low degree

Under this circumstance, since we can find the same correspondence relationship with students for every teacher, none of the teachers takes a different role in this school. Therefore, teachers' degree should be almost the same. By looking the degree of teachers in contact network with students in all grades, we can find three teachers having degree much lower than the average degree of teacher, which is 43.2. These three teachers may need to spend more time with students.

4 Block Model Analysis

4.1 Class Based Block Model

In this part, we consider block model based on classes to analyze the relation among different classes and grades. From figure 9, we can see directly from the graph that students' communication is extremely tight

within one group. This is relatively intuitive because students tend to have most chance to communicate with their classmates and most communications are happened in classrooms.

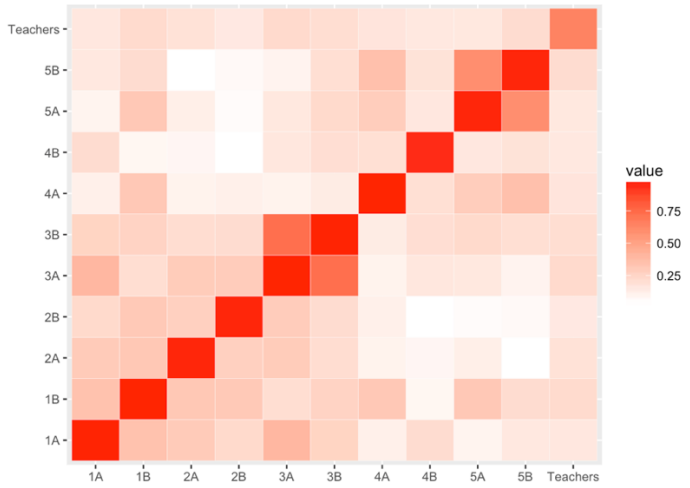


Figure 9. Communication density between different classes

Figure 10 shows relations that have weight exceed whole network density. There're two isolated blocks in this network – class 4B and group of teachers. Isolation of teachers indicates that the interaction between teachers and students are insufficient. Teachers need to spend more time with students. Unusually, class 1B looks like the 'central' class in this school, which can be derived from highest tendency of communicating in grade 1 students. What's more, most of relations between classes in the same grade are retained except for grade 2 and 4. Teachers responsible for these two grades, which are 1650 and 1852 for grade 2 and 1521 and 1653 for grade 4, should organize some grade-level activities to facilitate students' communication across classes.

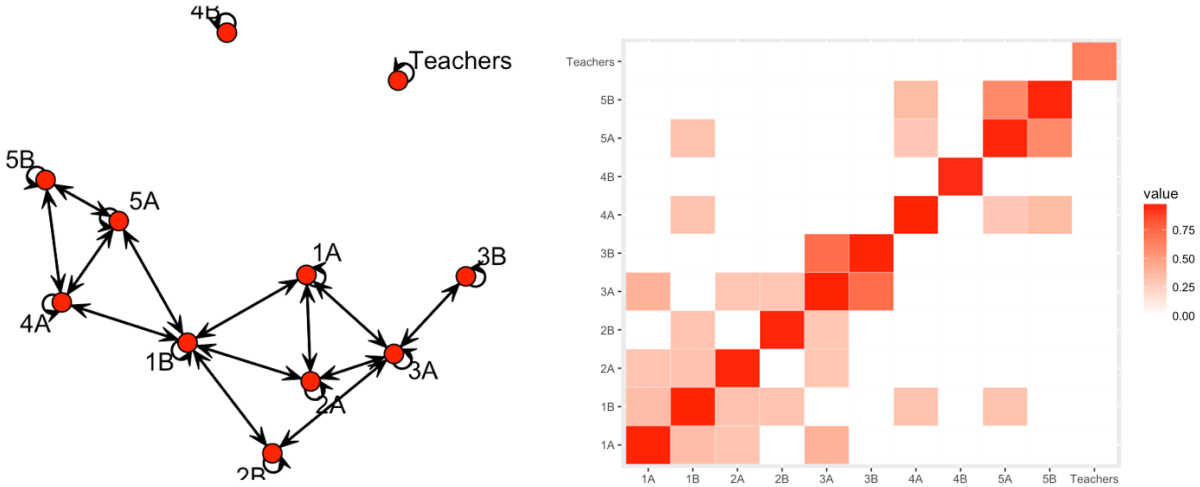


Figure 10. Block model on contact graph

Figure 11 shows the result of similar block model based on strong contact graph. In this model, group of teachers is the only block without an ego link. In addition, most blocks are isolated and do not have links to other blocks. This implies that close relationship is more likely to be built between students in same class.

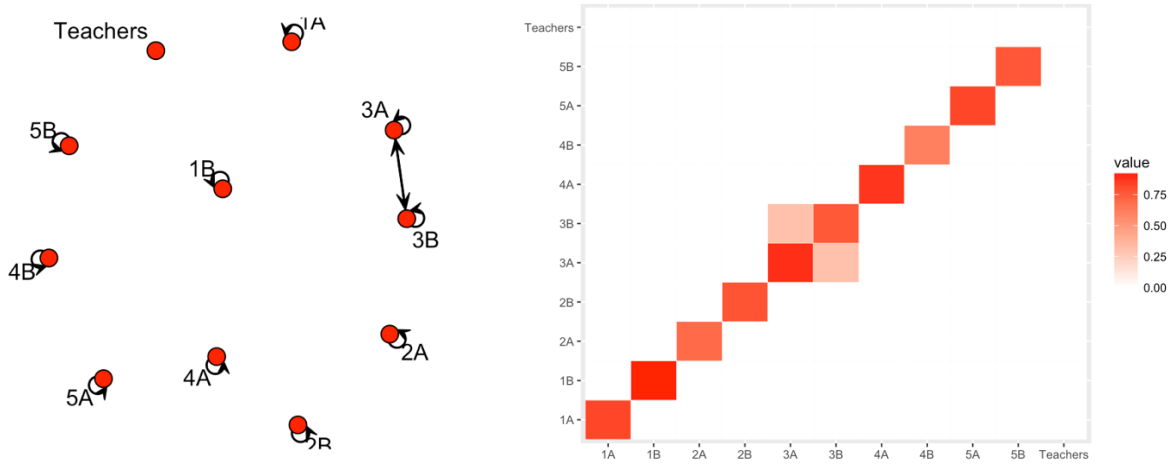
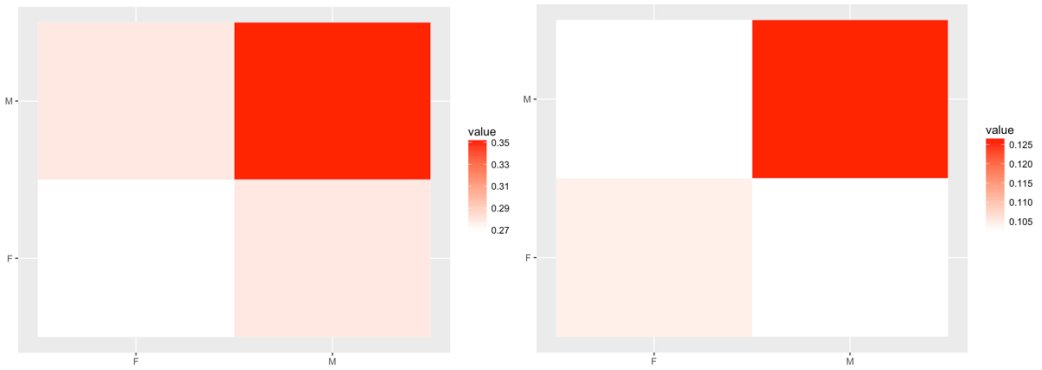


Figure 11. Block model on strong contact graph

4.2 Gender Based Block Model

In this part, we develop a block model in gender. The results are shown in figure 12. Obviously, male-male forms the tightest relation in both graphs compared with other gender combinations. For other combinations, we are surprised to get a different result in contact graph and strong contact graph. That is, in contact graph, girls tend to contact with boys. But when it comes to strong contact graph, female-female contact seems to be denser. This implies that girls tend to choose girls as their close friends.



(a) Contact graph (b) Strong contact graph

Figure 12. Contact density between different gender

5 Logistic Regression in Predicting Strong Contact

It's well known that two students that have close relationship with each other have high probability of sharing common friends. Hence we are interested in if the existence of strong relationship is correlated with number of common friends. We develop a QAP logistic regression, trying to predict strong contact with number of common friends. By doing this, we get a statistically significant model. The overall accuracy rate of this model is 95.62%. We also derive some interesting conclusion based on the model result.

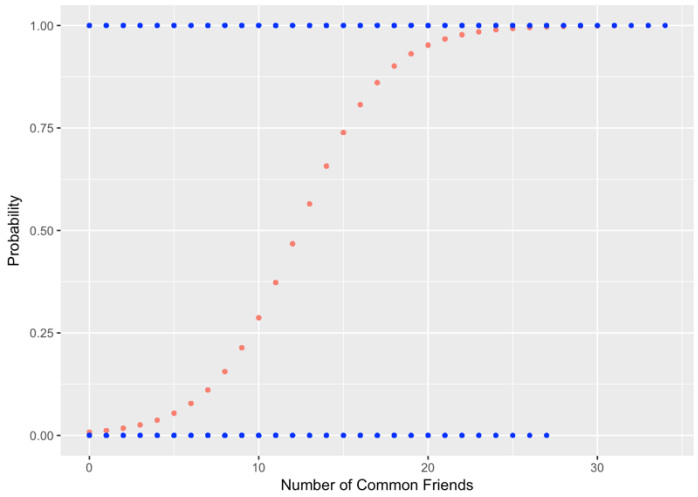


Figure 13. Logistic regression result

```
Network Logit Model

Coefficients:
            Estimate   Exp(b)     Pr(<=b) Pr(>=b) Pr(>=|b|)
(intercept) -4.8111309 0.00813865 1       0       1
x1           0.3900841 1.47710499 1       0       0

Goodness of Fit Statistics:

Null deviance: 80851.46 on 58322 degrees of freedom
Residual deviance: 14291.61 on 58320 degrees of freedom
Chi-Squared test of fit improvement:
  66559.85 on 2 degrees of freedom, p-value 0
AIC: 14295.61    BIC: 14313.55
Pseudo-R^2 Measures:
 (Dn-Dr)/(Dn-Dr+dfn): 0.5329826
 (Dn-Dr)/Dn: 0.8232363
Contingency Table (predicted (rows) x actual (cols)):

         Actual
Predicted     0     1
        0 51152  1626
        1   926  4618

 Total Fraction Correct: 0.9562429
 Fraction Predicted 1s Correct: 0.8329726
 Fraction Predicted 0s Correct: 0.9691917
 False Negative Rate: 0.26041
 False Positive Rate: 0.01778102
```

Figure 14. Logistic model summary

From the result above, we can find that number of common friends has positive relation with the probability of strong contact existence. If two students do not share common friends, the probability that they will have strong contact is only 0.008. But this probability will increase dramatically as the number of common friends increases. In addition, we can find that the cutoff of common friends is 13, which means that two students that have strong contact with each other need to share at least 13 common friends. We're surprised at getting such a huge number.