Group 3: Haolin Yang

Jiaying Liu

Jamie Pan

Lingchen (Wade) Kong

DNSC 6215

Social Network Analytics

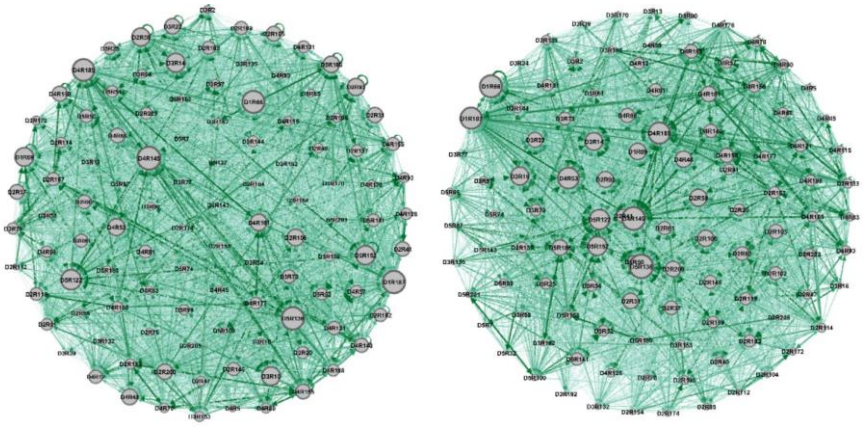
Project-A Report

I. Introduction

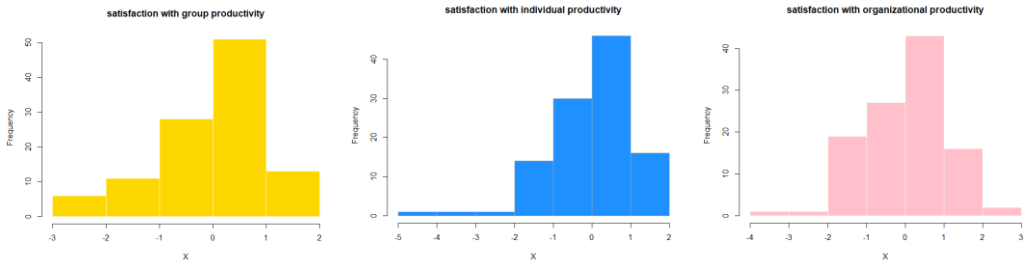
- **Describe the networks and attributes in general**

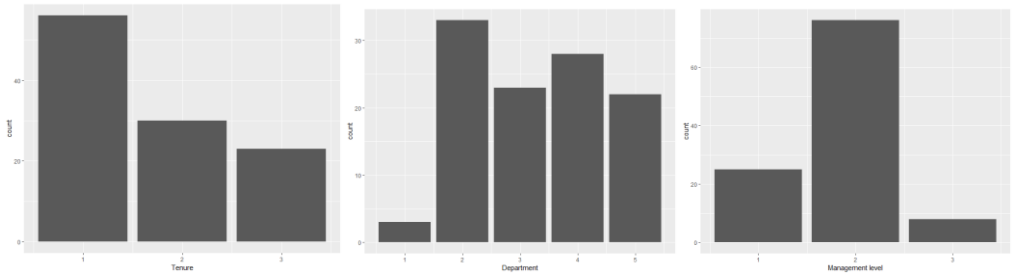
  Network datasets: Two networks. Collaboration network and Information Seeking network. The higher value in the network matrix indicates stronger relationship. Both networks contain 109 vertices and more than 2000 edges, which can be said two relatively large social networks.

Name of Network	Information Seeking Network	Collaboration Network
Number of Vertex	109	109
Number of Edge	4010	2029



Attribute dataset: The attribute dataset includes department, management level, tenure, and other information associated with each employee in the organization. By utilizing those attribute when we analyze the network we can get a closer look at how those different employees interact with each other.





- **Describe what kinds of questions can be asked and answered based on these data**

  Before we do all the analysis our group's research questions can be generalized as follows:

  a. What types of employee in the organization are more likely to become the center in the network?

  b. What is the structure of our network? What are some similarities or differences between collaboration network and information seeking network?

  c. Is there any connection between collaboration and seeking information?

  d. What specific attributes will have an influence on forming collaboration or seeking information ties?

- **Describe what follows in your report (outline and approach)**
  Our group will display our analysis and findings in the following workflow:
  1. Descriptive analysis (Network Level & Node Level)
  2. Subgroup analysis and role analysis
  3. Statistical analysis (QAP & ERGM)
  4. Conclusion (overall findings and interesting aspects)

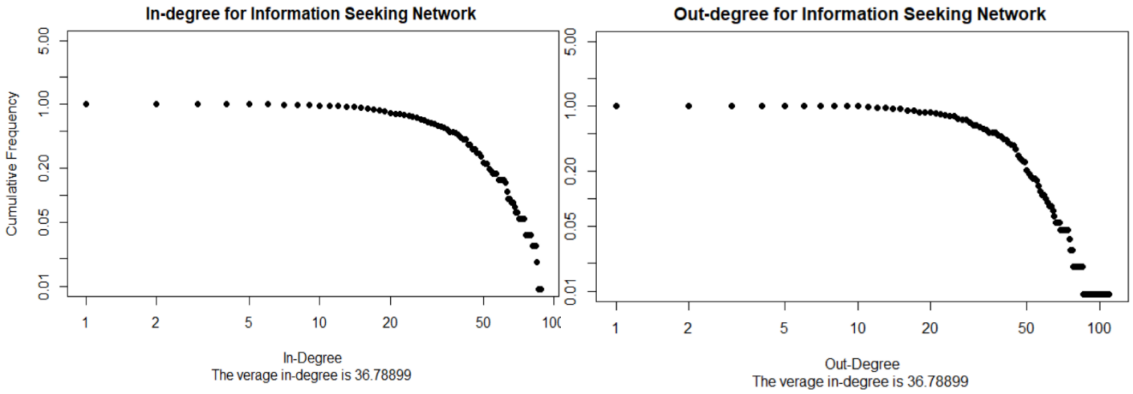
II. Descriptive Analysis

For the network level analysis, the basic parameters are shown in the following table

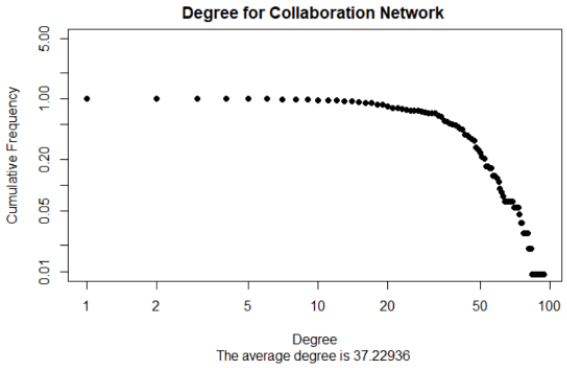
Name of Network	Information Seeking Network	Collaboration Network
Diameter	3	3
Mean Distance	1.6721	1.6692
Edge Density	0.3375	0.3384
Global Transitivity	0.5903	0.5504

As we can see from the table, almost all the parameters related to the two networks are very close, indicate their high similarity in structure. The longest geodesic distance in both network is 3; the mean of the shortest distance between each pair of nodes in both network is about 1.7; the portion of the potential connections in this two networks that are actual connected is about 34% and both networks have about 60% triangles.

The network of information seeking is directed and its in-degree and out-degree distributions is as followings:



The cumulative frequency is decreasing when the number of in-degree or out degree increases. The mean value of in-degree and out-degree are the same: 36.78899; while for collaboration network (undirected), its degree distribution is about 37.299 as the plot shown below. The same trend can be observed as the one of information seeking degree distribution. With the higher number of tie one node have with other nodes, its frequency in the network tend to be lower.



For the node level analysis, the basic parameters are presented as followings:

Name	Information seeking	Collaboration
Degree Centrality	36.78899	37.22936
Closeness Centrality	0.005607663	0.0056173
Betweeness Centrality	72.58716	36.13761

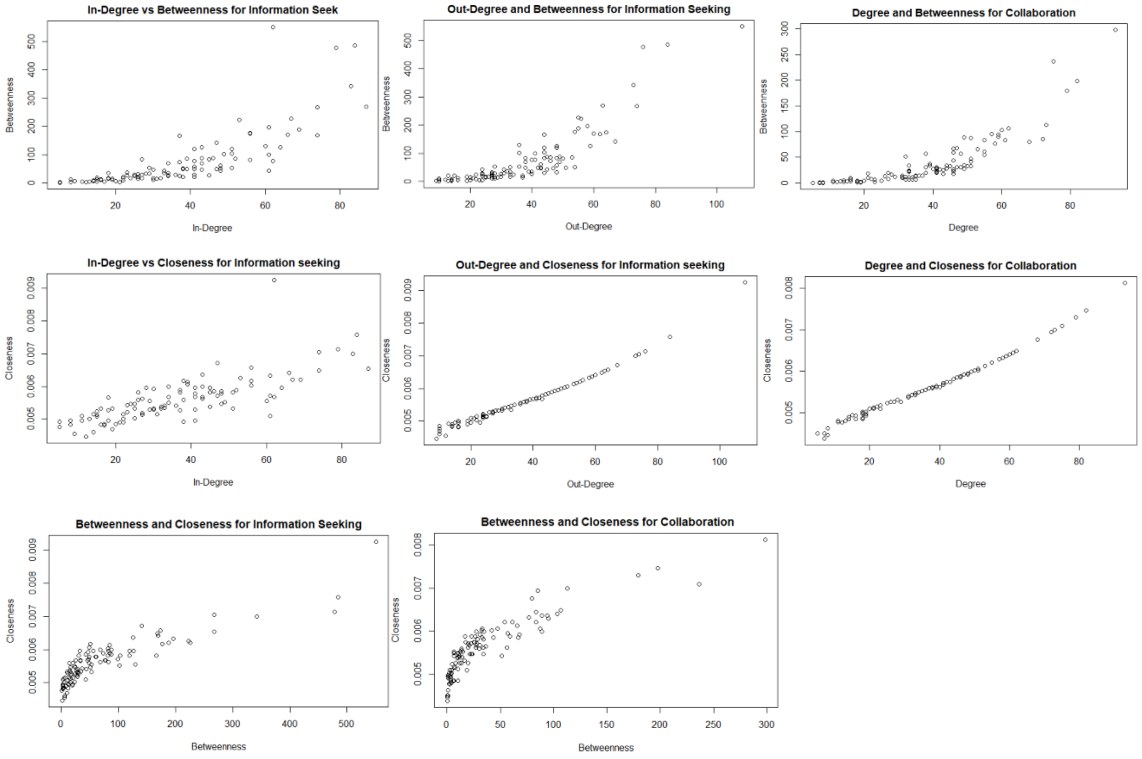
In fact, according to the parameters of closeness, betweenness and degree, two networks are highly centrality and there are 10 important nodes existing with high weights and values. They are D1R187, D1R66, D2R200, D2R61, D2R92, D4R149, D4R189, D4R53, D5R122, D5R136 and their information is concluded below:

ID	Department	Man-level	Tenure	Gender	Individual	Group	Organization
D1R187	1	1	2	2	-1.164	0.027	-1.742
D1R66	1	1	3	2	0.46	0.195	0.41
D2R200	2	1	2	1	-0.214	0.03	0.289
D2R61	2	1	2	2	0.123	0.325	0.136
D2R92	2	2	2	1	1.009	1.449	-0.38
D4R149	4	1	3	2	0.231	-0.369	0.144
D4R189	4	1	3	2	1.043	0.128	0.724
D4R53	4	2	1	2	1.564	0.219	0.995
D5R122	5	1	3	2	0.703	0.857	0.603
D5R136	5	1	3	2	1.265	0.599	0.231

There some interesting phenomenon can be observed:

- There are two or three centralities could be formed in each department, which means the distribution of the center nodes in each department is even. This might indicate that for each department there are two or three managers.
- Most management level of employees in the collaboration and information seeking network is 1 and their tenure are more than 2 years. In other words, the person with lower management and longer tenures would be found to be more likely to collaborate and seek information.
- People categorized as gender 2 are more popular and likely to be an important node in both networks.
- Most people appear as central nodes tend to have higher evaluation on the individual, Group and Organization.

The relationships between the degree centrality and betweenness centrality, the degree centrality and closeness centrality, the betweenness centrality and closeness centrality are all positive correlated. In fact, the conclusion that can be reached that the node that has many connections with other nodes will be regarded as a significant node with large possibility and other nodes are more likely to connect with this node, which means this node will have higher betweenness centrality. All the relationships are showing as followings:



III. Subgroup & Role Analysis

Communities

As we have applied Leading Eigen vector detection algorithms to the relationships in information to seek and relationships in collaborations we get a basic understanding of their structures.

Community of information seeking and collaboration network.

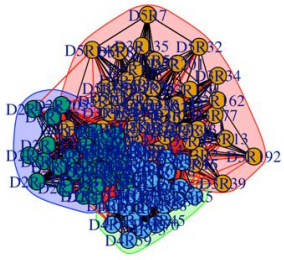
When it comes to employee social network, we can see three groups in here. According to the outputs, the majority of employees in group 1 are department 3 and 5. And almost all employees in department 4 are in the group 2. The group 3 has a lot of people in department 2.

Thus, we can conclude that employees always seek information or other kinds of help to people that in the same department, such as group 2 and group 3. About group 1, we can assume that department 3 and 5 have the similar business or these two departments need a lot of cooperation and daily connection.

```
$`1`
  [1] "D3R2"   "D5R7"   "D3R10"  "D3R13"  "D3R14"  "D3R16"  "D3R22"  "D5R25"  "D2R31"  "D5R32"  "D3R34"
 [12] "D3R39"  "D5R50"  "D5R52"  "D4R53"  "D5R54"  "D5R55"  "D3R64"  "D5R65"  "D1R66"  "D3R73"  "D5R74"
 [23] "D3R77"  "D3R79"  "D5R87"  "D3R90"  "D2R92"  "D3R97"  "D5R99"  "D5R100" "D5R122" "D4R131" "D3R135"
 [34] "D2R137" "D3R139" "D5R141" "D5R143" "D3R144" "D5R152" "D3R153" "D3R162" "D5R168" "D3R170" "D5R180"
 [45] "D5R186" "D3R192" "D2R195" "D5R201"

$`2`
  [1] "D4R5"   "D4R12"  "D4R45"  "D4R48"  "D4R57"  "D4R59"  "D4R68"  "D4R70"  "D4R80"  "D4R81"  "D4R83"
 [12] "D4R86"  "D4R93"  "D4R115" "D4R121" "D2R133" "D4R140" "D4R149" "D4R155" "D4R156" "D4R158" "D3R164"
 [23] "D4R165" "D4R176" "D4R177" "D4R181" "D1R187" "D4R189" "D4R198"

$`3`
  [1] "D2R20"  "D2R37"  "D2R40"  "D2R41"  "D2R47"  "D2R58"  "D2R61"  "D1R69"  "D2R76"  "D2R82"  "D2R85"
 [12] "D2R91"  "D2R102" "D2R103" "D2R104" "D2R105" "D2R106" "D2R109" "D2R112" "D2R114" "D2R119" "D4R128"
 [23] "D3R132" "D5R136" "D2R146" "D2R154" "D2R167" "D2R172" "D2R174" "D2R200" "D2R203" "D2R205"
```



Community of collaborations

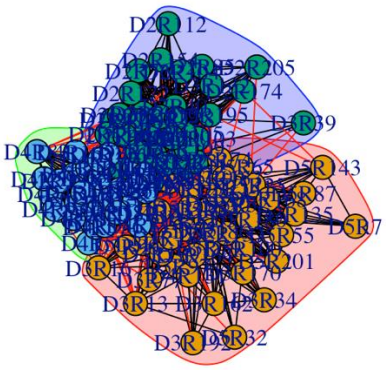
In this community, we still use the Leading Eigen vector detection algorithms to the relationships. The structure of groups and the plot are very similar to the community of information seeking.

They will choose those who are work in the same department or most cooperation departments to seek information or get collaboration. Therefore, we can get that employees have the solid connections with similar group at work.

```
$`1`
 [1] "D3R2"   "D5R7"   "D3R10"  "D3R13"  "D3R14"  "D3R16"  "D3R22"  "D5R25"  "D5R32"
"D3R34"  "D5R50"
[12] "D5R52"  "D4R53"  "D5R54"  "D5R55"  "D3R64"  "D5R65"  "D1R66"  "D3R73"  "D5R74"
"D3R77"  "D3R79"
[23] "D5R87"  "D3R90"  "D3R97"  "D5R99"  "D5R100" "D5R122" "D4R131" "D3R135" "D2R137"
"D3R139" "D5R141"
[34] "D5R143" "D3R144" "D5R152" "D3R153" "D3R162" "D3R164" "D5R168" "D3R170" "D5R180"
"D5R186" "D3R192"
[45] "D5R201"

$`2`
 [1] "D4R5"   "D4R12"  "D4R45"  "D4R48"  "D4R57"  "D4R59"  "D4R68"  "D4R70"  "D4R80"
"D4R81"  "D4R83"
[12] "D4R86"  "D4R93"  "D4R115" "D4R121" "D2R133" "D4R140" "D4R149" "D4R155" "D4R156"
"D4R158" "D4R165"
[23] "D4R176" "D4R177" "D4R181" "D1R187" "D4R189" "D4R198"

$`3`
 [1] "D2R20"  "D2R31"  "D2R37"  "D3R39"  "D2R40"  "D2R41"  "D2R47"  "D2R58"  "D2R61"
"D1R69"  "D2R76"
[12] "D2R82"  "D2R85"  "D2R91"  "D2R92"  "D2R102" "D2R103" "D2R104" "D2R105" "D2R106"
"D2R109" "D2R112"
[23] "D2R114" "D2R119" "D4R128" "D3R132" "D5R136" "D2R146" "D2R154" "D2R167" "D2R172"
"D2R174" "D2R195"
[34] "D2R200" "D2R203" "D2R205"
```



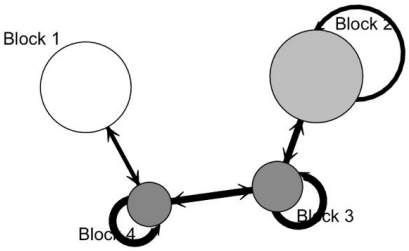
Block model

```
$`1`
 [1] "D3R2"   "D5R7"   "D3R10"  "D3R13"  "D3R16"  "D5R32"  "D3R34"  "D3R39"  "D5R55"
[10] "D3R64"  "D5R65"  "D5R74"  "D2R76"  "D3R77"  "D3R79"  "D2R85"  "D5R87"  "D3R90"
[19] "D3R97"  "D5R99"  "D5R100" "D2R104" "D2R112" "D4R131" "D3R132" "D3R135" "D3R139"
[28] "D5R141" "D5R143" "D3R144" "D3R153" "D2R154" "D3R162" "D3R164" "D5R168" "D3R170"
[37] "D2R174" "D5R180" "D3R192" "D2R195" "D5R201" "D2R205"

$`2`
 [1] "D4R5"   "D4R12"  "D2R20"  "D2R31"  "D2R37"  "D2R40"  "D4R45"  "D2R47"  "D4R57"
[10] "D2R58"  "D4R59"  "D2R61"  "D4R68"  "D1R69"  "D4R70"  "D4R80"  "D4R81"  "D2R82"
[19] "D4R83"  "D2R91"  "D4R93"  "D2R102" "D2R103" "D2R105" "D2R106" "D2R109" "D2R114"
[28] "D4R115" "D2R119" "D4R121" "D4R128" "D2R133" "D2R146" "D4R155" "D4R156" "D4R165"
[37] "D2R167" "D2R172" "D4R176" "D4R177" "D4R181" "D4R198" "D2R200" "D2R203"

$`3`
 [1] "D3R14"  "D2R41"  "D4R48"  "D4R53"  "D1R66"  "D4R86"  "D2R92"  "D5R136" "D4R140"
[10] "D4R149" "D4R158" "D1R187" "D4R189"

$`4`
 [1] "D3R22"  "D5R25"  "D5R50"  "D5R52"  "D5R54"  "D3R73"  "D5R122" "D2R137" "D5R152"
[10] "D5R186"
```



There is the structure in the left plot. We still can see the department 3 and 5 in the block 4, and department 2 and 4 in the block 2. They share the same role.

For information seeking and collaboration, there are four blocks based on collaboration or information seeking datasets in the right plot. There are many communications between 1 & 4, 4 & 3, and 3 & 2. But there is no communication between 1 and 2. Block 1 is the largest group but have connection to block 4. The block 3 and 4 are in the central location and have close relationships between each other. The block 3 also has close relationship to block 2. Block 2, 3, and 4 are not only seek information and get collaboration to other groups but also by themselves.

IV. Statistical analysis (QAP & ERGM)

The collaboration and information seeking adjacency matrices are translated to the un-weighted matrices first, where '1' in the new matrices indicates the previous level of rating above 4, and '0' indicates those below or including 4.

Then, the descriptive function provides the statistical description of two directed relationships.

- Collaboration Relationships
    - Sum indicates there are 3382 one-direction collaborations between two employees.
    - 'row-sum' has some extreme numbers, such as its number for D5R7 and D2R105 is 4, and D2R92 is 85. These illustrate that D5R7 and D2R105 rarely has asked for collaboration from others, and D2R92 has collaboration with other very often.
    - Similarly, extreme numbers in 'col-sum' for D1R187 (74) and D3R192 (3) tells that D1R187 is asked for collaboration the most, and D3R192 is not popular to be asked.
- Info Seeking Relationships
    - Sum indicates there are 4010 on-direction information seeking between two employees.
    - 'row-sum' has some extreme numbers, such as its number for D3R34 (9) and D2R92, which tells that D3R34 don't seek for information form others often, but D2R92 likes to seek information from others very much.
    - Similar, extreme numbers in 'col-sum' for D5R143 (8) and D3R39 (8) and D1R66 (84) tells that D1R66 is the most popular one to be asked for information, and D5R143 and D3R39 are not very often to be asked.

QAP model

In the QAP test, we focused on studying the relationships between two networks, and we fit both networks into netlogit model.

- Both models have very similar outputs, except the estimate log-odds for intercept. The model with COL as DV has beta 0 as -3.69, which indicates that one would have log-odds of -3.69 to form a collaboration tie with absence of info-seeking ties. The model with INFO as dependent variable has log-odds of -2.23 to form an info-seeking ties with absence of collaboration ties.
- The log-odds for x1 in both models are the same with significant p-value of 0, and we can see the odds ratio for both is 156.60. It indicates that a new collaboration ties are 156.6 more likely to be formed in the presence of the info-seeking ties than in the absence of info-seeking ties. And vice versa.
- Since Chi-squared test of fit has p-value of 0, which we fail to reject the null hypothesis and conclude that there is no significant difference between observed and expected value. This means the model fits well.

ERGM for Collaboration Network

- The null model outputs the estimated log-odds for edges is -2.98006. After plugging this number into plogis code, the number tells that the probability of forming any tie in collaboration network is 4.83%.
- The estimated log-odds for mutual in model col.ergm.2 is 3.4229, which indicates that the log-odds of the collaboration tie is 3.4229 times greater if the reciprocal tie is present. The probability for forming a collaboration tie with presence of reciprocal tie is 46.16%.
- The modified model outputs AIC as 4008, which is the best try-out we had, and small AIC value indicates this model is the preferred one. Every factors in this model has significant effects on collaboration ties.
- From the original model, attribute 'Gender' and 'Organization' have shown no significant effect on collaboration ties, so they're removed from the model. Some other insignificant categories in some categorical attribute are removed as well. Only 'department.2', 'management.level.1' and 'tenure.1' and 'tenure.2' are remained, since they have significant effects on forming collaboration ties.

ERGM for Info-Seeking Network

- The null model outputs the estimated log-odds for edges is -2.6608. After plugging this number in to plogis code, the number tells that the probability of forming any tie in info-seeking network is 6.53%.
- The estimated log-odds for mutual in model info.ergm.2 is 3.35571, which indicates that the log-odds of the collaboration tie is 3.35571 times greater if the reciprocal tie is present. The probability for forming a collaboration tie with presence of reciprocal tie is 50.51%.
- The modified model outputs AIC as 4984, which is the best try-out we had, and small AIC value indicates this model is the preferred one. Every factors in this model has significant effects on info-seeking ties. * From the original model, attribute 'Gender' and 'Organization' have shown no significant effect on information seeking ties, so they're removed from the model. Some other insignificant categories in some categorical attribute are removed as well. Only 'department.3', and 'tenure.2' are remained, since they have significant effects on forming info-seeking ties.

V. Conclusions

Through a series of analysis of this two networks, we think it is time for us to conclude the findings and respond to our research questions.

For the first question we find that actually the first level managers, if we assume the first level managers are the lowest level in the organization, receive the most level of information seeking as well as collaboration. This is probably due to level 1 managers have more direct contact with other employees. In addition, people who have longer tenure years also become the center in both two networks, which means highly experienced employees will take most load of collaborating and providing information.

For both networks, they have a very centralized structure as we can see from low degree density figure, where people tend to collaborate or seeking information within the same department. Both network can be divided into three communities based on the same department or departments which have strong relationship bond. The high correlations between each category of centrality also indicate the fact that the node with many ties with other nodes will be regarded as a significant node (bridge).

Based on the analysis from QAP analysis we find that The log-odds for x1 in both models are the same with significant p-value of 0, and we can see the odds ratio for both is 156.60. It indicates that a new collaboration ties are 156.6 more likely to be formed in the presence of the info-seeking ties than in the absence of info-seeking ties. And vice versa.

Last but not the least, attribute ‘Gender’, ‘Organization’ and some other attributes have shown no significant effect on collaboration ties. Only ‘department.2’, ‘management.level.1’ and ‘tenure.1’ and ‘tenure.2’ are remained, since they have significant effects on forming collaboration ties. When it comes to information seeking network, attribute ‘Gender’ and ‘Organization’ also show no significant effect on collaboration ties while ‘department.3’, and ‘tenure.2’ are remained, since they have significant effects on forming info-seeking ties.