Social network study of IISER Mohali Students

A dissertation submitted for partial fulfilment of

BS-MS dual degree in Science

Aniket Gaur

MS10050

Indian Institute of Science Education and Research Mohali

November 2015

Dedicated to Manas

Since he dedicated his thesis to me.

Certificate of Examination

This is to certify that the dissertation titled "Social network study of IISER Mohali students" submitted by Aniket Gaur (MS10050) for the partial fulfilment of BS-MS dual degree program of IISER Mohali has been examined by the thesis committee duly appointed by the institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr Alok Maharana Dr. Shashi Bhushan Pandit Dr. Kuljeet Singh Sandhu

(Supervisor)

(Dated: November 30th 2015)

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Singh Sandhu at Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contribution of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in bibliography

Aniket Gaur

Dated: November 30th 2015

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to best of my knowledge.

Dr. Kuljeet Singh Sandhu

(Supervisor)

Acknowledgement

I offer my warmest gratitude to my supervisor Dr. Kuljeet Singh Sandhu , who gave me the opportunity to do MS thesis in his research group. I would like to thank him for his guidance, support, valuable comments and encouragement from the start until the end of my project.

My project would have been a complete failure if a lot of people hadn't helped me. All the people in Computational biology lab advertised my project not just among their friends but in entire IISER Mohali just so that I get more response in my online form. For that special thanks to Arashdeep, Keerthivasan, Meenakshi, Navjot mam ,Nitika, Preeti, Rivi, Vivek Sagar and all other members in my lab. All these people also tried to help me by giving advices or suggestions, and some of those suggestions turned out to be very helpful for my project. I would like to thank again Meenakshi and Nitika who helped me in distributing offline forms in Hostels.

I would like to thank Arul, Tj and Manas who also helped me in advertising my project in IISER Mohali. I also want to thank Kanupriya who helped in completing my thesis.

At the end I would like thank to all the people in IISER Mohali who filled my long boring online and offline forms.

Aniket Gaur

Contents

Certificate of Examination iii
Declaration iv
Acknowledgement v
Abstract vi
Chapter-1(Introduction) 1
Chapter-2(Data required) 3
Chapter-3(Methods for data acquisition) 4
Chapter 4(Methods for data analysis) 6
1) PCR 6
2) Markov Model 7
Chapter 5(Web app and data analysis) 9
1) Webapp 9
2) Data analysis 9
Chapter 6(Discussion) 21
1) Module analysis of networks 21
2) Comparison of values in network properties and individuals' attributes 22
3) PCR of Individual attributes and PCR with network properties 23
Chapter 7(Conclusions) 24
Appendix 25
1) Network Questions 25
2) Personality Questions 25
3) Rationality Questions 27
4) Quiz 27
5) Email text that was sent at first place 28
6) Snap-Shots of webapp 29
References 32

Abstract

We are all a part of social network so there can be a lot of factors which decides its properties. IISER Mohali comprises of students from different parts of the country. Some of them might have different social interacting circles. In our study we collected data of students' network through online survey form, where they also answered questions related to personality, rationality, etc. After data collection, we made the network of students in different categories (Friends, Academic support, financial support and emotional support network).

In friends network there was only one connected component and some modules (connected sub networks) and in other networks there were few small disconnected components. In all the cases it seems that out of the people who filled the online form, very few of them are disconnected from each other in IISER-M. Students network at IISER-M follows scale-free like degree distribution, hierarchy and dis-assortativity, validating the non randomness and reliability of our data.

We compared individuals' properties (like personality and rationality) with other network properties. Correlogram and PCR analyses suggested that different personality traits are associated with distinct network properties, highlighting the possible predictive power of the network attribute for students' personality prediction.

Chapter-1

(Introduction)

Our behaviour is largely dependent on our society in which we dwell. We always try to adapt to our surroundings and in this process our behaviour also changes. Organisation of the society is, therefore, intrinsically linked to our personality and performance. To understand the organizing principles of the society, we thought of taking IISER Mohali as one closed society and map different kinds of social interactions. Overlaying personality and performance traits of students onto their social networks might help uncovering the relational attributes shaping personality and performance of students.

IISER-M is a mixed kind of community, where people from almost all over India join as students. Since almost everyone lives in four buildings of hostels, go to the same respective mess to eat, same places to attend lectures, same place to play sports and many more activities people do which intersect with others. We thought it might be interesting to study how people make their network, what kind of people are there in one's network and how one's network affects his/her behaviour. Moreover, closed nature of the IISER-community allows for easy access to the network, personality and academic datasets.

In IISER Mohali, if we consider distribution of people's interaction group, it is dependent on their home town or mother tongue or maybe some other factors. Our initial aim is to look whether there are any specific things which make one interaction group different from the other.

Understanding the network may be very important for an Institute like IISER-M to function better. In this way one may be able to understand, what kind of interactions are good and what are not? What probable factors help students to perform better in academics and what are the reasons behind his/her bad performance? So if someone who hangs out with people, who are good in academic, are performing well academically or vice-versa. What personality traits of someone make or help him to hang out with people of some other or same personality trait? How someone's personality helps to make connections in a community and what are the factors which bind them?

We decided to collect data from students (including PhD and post-doc) by asking them questions about their friends network(which will decide what kind of people are there in their network) and other questions for personality, rationality, BMI etc. On the basis of individual's qualities (like personality, rationality, intelligence etc.), we can compare it

with people who are there in his/her network. So one can also try to see how connected people are among themselves and in entire network what are hubs present and what specific properties of one hub makes it different from other? Why is there a hub in a network and why only by some set of people and what are the different properties they have which makes them part of hub? What properties people have, who are not part of any hub but they just are at very out of network as an isolated node? How connected the whole network is and are there any disconnected components if yes what could be probable reasons behind it. Basically number of questions one can ask will be many. So we thought of designing an experiment which can try to answer some of these questions and allows us of come up with many such questions which we might not have been able to think of.

Chapter-2

(Data required)

Before data collection, we needed to decide, what are the questions we should ask them online. For carrying out this task, we decided to divide our form in two parts, one part consisted of questions related to networks and other comprised of questions related to an individual's properties like personality, rationality, BMI (Body mass index) etc.

For questions of first part refer to appendix[1]. In questionnaire we thought of asking questions like "In IISER Mohali who is your friend?", without explaining in depth what we meant by 'friend' and we asked to give a rating for each name on the scale of 1-10 on the basis of their interaction with that person. We had three more questions which were related to person's academic support network, emotional support network and financial support network.

Second part had questions related to personality and rationality (for details refer to appendix[2]). For personality questionnaire, we had some choices but we thought of taking the one which had comparatively lesser number of questions and at the same time questions were straight forward. So we took 50 sets of questionnaire from International Personality Item Pool(ipip[1]). It contains questions for five personality traits (1) Extraversion, (2) Agreeableness, (3) Conscientiousness, (4) Emotional Stability and (5) Intellect/Imagination, each question was in the form of statement and user needed to tell how well it agrees with his personality by moving slider on scale of strongly disagree, disagree, neutral, agree and strongly agree. After this we asked them to tell their height and weight and at the end we also decided to add rationality test in our form which we got from online rationality test[2]. From there we only took shortest version of questionnaire which had 10 questions which judged two traits: rational engagement and rational ability. After rationality test we also included few other questions, one was "what is your religious view?" and other questions were in the form of a quiz which tries to determine how well people think to solve given problems(see appendix[3]) which required logical thinking. After setting up all the questions we thought of fetching people's facebook information. For that we used Facebook-graph-api[3] which provides a platform by which one can ask for permission from facebook user for the access of information from facebook account. In that we only asked permission for access to the friend list.

Chapter-3

(Methods for data acquisition)

At the initial stage, we needed to setup a medium by which we can collect data from students. For that we thought of setting up online web-app through which people can fill online form. We didn't use other online survey websites like Google forms, Surveymonkey, etc because we wanted to customize our form according to our use and those online sites would not have given us that much freedom.

We setup an online web-app in which people can login so that every individual can be authentic, which will try to make sure that form is filled by that individual only not by somebody else. To make such a platform we needed to know some web developing framework. With a prior knowledge of programming in python, we chose Django[4] for the setup of web-app. It is an open source library with many modules to do web development.

How *django* works: Django is python based webframe work, and when it comes to making web-app and writing codes, it is very organized. To start making project (say "iisermwebapp") first thing one needs to do is to install django library and then run the command "django-admin.py startproject iisermwebapp", which will create a folder named iisermwebapp with some start settings files (like manage.py, settings.py) as described in Figure1.

Now we needed to create django-app which can be used by created settings file. In django project folder, you make new folder and give name of your app and start writing codes in there and in your settings file you need to add name of this new app so that you can use it in your djangoproject . An advantage of Django is that you can create separate file for each thing like mapping urls, writing views for each url, manage data base request and making tables in database, and other your own codes unrelated to django but can be supporting codes which you might need for project.

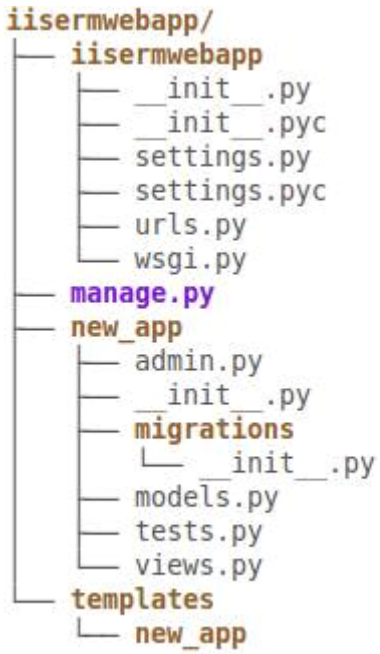


Figure 1: Django project(iiserwebapp) folder tree

What are the things we wanted in web-app?

We had few questionnaire in which user needed to give rating on some scale for which we used a jquery api which can work on all kind of browsers. Our jquery slider worked for almost every kind of pc based web. Other than that we had jquery for autocomplete so that people do not have to write full name of their friends, reduce the chances for someone to misspell somebody's name. Auto complete was needed because we wanted people to only mention people from IISER Mohali only and entries will have less chance of being fake. This autocomplete helped us a lot in mapping names of network.

We also needed to add facebook share option for people. For that we created facebook app using facebook-graph-api. In Graph api people can give permission to our fb-app to access information from their facebook account. In facebook we only asked permission for the access of facebook friendlist and there we also give option to user how much they want to share with us.(For further details on webapp, refer to chapter 5)

Chapter 4

(Methods for data analysis)

Properties of each network nodes may follow some distribution pattern which can determine nature of network. For this, one can do power law analysis in which we look for distribution of degree with number of nodes in network which can tell us whether a given network is scale free or not. In scale free network, degree distribution follows power law. We can also see whether network is hierarchical or not. For that, network needs to have high clustering and scale freeness. In hierarchical networks the clustering coefficient can be expressed as a function of the degree as $C(k) \sim k^{-\beta}$ (In scale-free networks the exponent β takes the value of 1.)[5]

On a given network there can be possibility of it being modular, which means network has sub network in a form of modules. For a network, module is a sub network which has more edges connecting to the nodes within itself and less to outside. In our case we used ModuLand a Cytoscape app which helped us to determine modules in our network. (for more details refer to chapter 6 section 1).

Since there will be a lot of parameters for analysis and some of those values can have co-linearity among themselves, having that in mind we thought of doing Principal component analysis(PCA)[6] of data. Other thing in data we can look for probabilistic dependence of variables among themselves for that there are models to study and apply, one of those models is Markov model[7].

1) PCR[6]:

In our data collection for each individual, we will have two kinds of properties one will be network properties (degree, clustering coefficient etc) and other will be individual's attributes (his personality, BMI etc). In each kind of properties, there will have some variables and few of them may have co linearity among themselves. In this kind scenario PCA can be very useful. It is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. In this we transform a number of correlated variables into a (smaller) number of uncorrelated variables called principal components.

In our case say we have a set vectors $X_i=[x_{1,},x_2, \ldots,x_n]$ of n dimension, and there is

corresponding vectors $Y_i=[y_1,,y_2, \ldots,y_m]$ which we want to predict. In usual multi linear regression Y = XB + E (where E is error term) B will be given by $(X^TX)^{-1}X^TY$. Since X might have co linearity among x_i's so we take

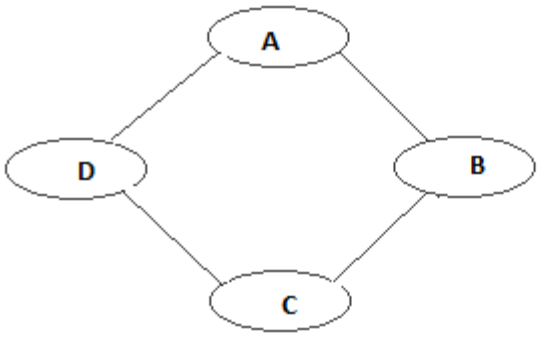
$$X=T*P \quad \text{(where T is orthogonal scores and P is loadings.)}$$

Next we regress Y on T*P by that get regression coefficient $B = P(T^TT)^{-1}T^TY$. For this purpose we used R package called "PLS".

2) Markov Model

Given a network, properties of one node might affect other nodes. In our network of students, one's academic performance or his personality may have effect on people who surround him. So we thought of learning Markov model which can help us to understand how these kinds of properties can have probabilistic correlations.

Markov network analysis[7] is used for undirected graph. Suppose we have a set of four people and edges represent (Figure 2) their interaction (by some standard way). Where $\phi[A,B]$ called affinity function for A,B where A and B can have two kind of values a^0 or a^1 and b^0 or b^1 respectively. Affinity function determines the likeness for particular combination of a^i and b^i.



φ[A,B]

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

Figure 2 Undirected graph example of affinity function ϕ[A,B]

Now for a given undirected network of A, B, C and D if affinity function for each node pair [A,B], [B,C], [D,C] and [A,D] are known, then we need to find network what value the a^i,b^i,c^i and d^i of these four people can have {where 'i' can be 0 or 1}.

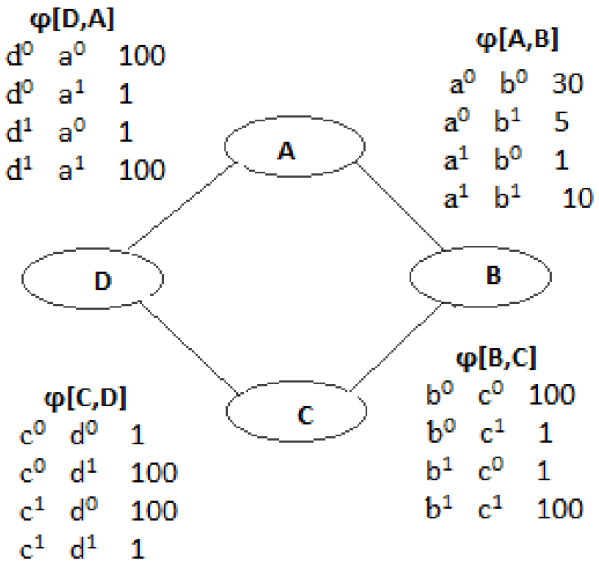


Figure 3 Affinity function for each edges

One can get probability of P[a^{i1},b^{i2},c^{i3} ,d^{i3}] for some combination which a, b, c and d as described in Figure3. P(A,B,C,D)= ϕ[A,B]* ϕ[B,C]* ϕ[C,D]* ϕ[D,A]*(1/z) where z used to normalized all values. Markov networks for probabilistic graph can help us to understand how one person's traits affect the people who are around him/her.

Markov model can be applied while taking people academic data with network data. With that one may be able to predict how person's surroundings affect that individual's gain or loss in academic in, probabilistically speaking.

Chapter 5

(Web app and data analysis)

1) **Webapp**

Our web app had admin interface in the background which was designed using django library for admin. In that, one can connect all tables, created in database, and monitor what is the status of user's entries. On other side we were also collecting facebook data. There, we were only saving facebook graph api access token and corresponding facebook username which have validity of 60 days. To get friendlist by using access token for a fb userid, we can do it by taking "https://graph.facebook.com/userid/friendlist?access_token=" on browser or parsing it in python json library by which we can store facebook friends network. This will only give names of people who have used our fb-app and total number of friends that userid has, given that the person has given us permission for the access of his facebook friends list.

We had our app ready and we tested it. We launched our website on IISER Mohali's intranet with different links for wifi and LAN. We also needed to send email to everyone which will have login details. To send mass email to everyone in IISER Mohali we wrote a program in python which uses python library named 'smtplib'(First email given in appendix[4]).

2) Data analysis

After data collection, in total of 468 people filled the form online (including complete and incomplete) and we had total number of 1079 user, there were 611(around 56.6%) people who did not fill the form at all. And response distribution for people who started filling the online form is in the Table 1 where 1st column represents till where people completed their online form and Figure 4 is pie chart for percentage column of Table 1.

		percentage
complete_with_fb	272	58.11965812
complete_no_fb	100	21.36752137
BMI	42	8.974358974
4th_part	28	5.982905983
personality_incomplete	9	1.923076923
2nd_part	5	1.068376068
3rd_part	5	1.068376068
personality_complete	4	0.854700855
1st_part	3	0.641025641
total:	468	

Table 1Till where people completed the online form

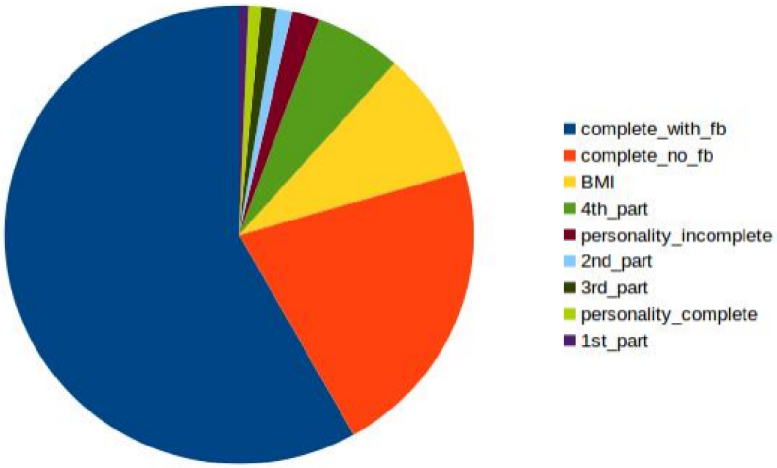


Figure 4 Pie plot for percentage column in table1

Discipline	total	people responded	percentage repononse
ms12	111	82	73.8738738739
ms13	115	71	61.7391304348
ms11	76	46	60.5263157895
ms10	19	11	57.8947368421
int phd	65	31	47.6923076923
phd	280	105	37.5
ms14	175	56	32
ms15	196	59	30.1020408163
others(post_doc,project students)	42	7	16.6666666667

Table 2 Response distribution from each batch or discipline

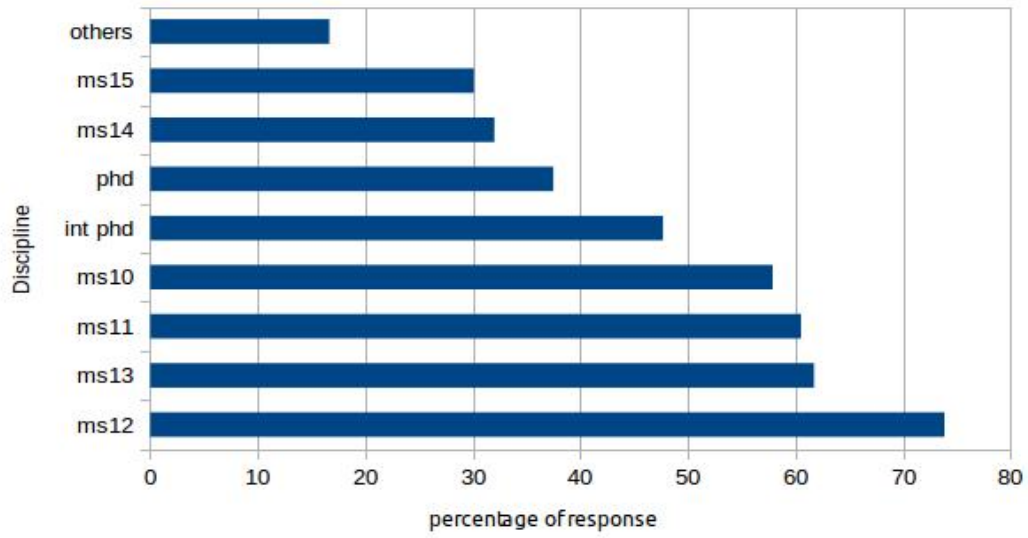


Figure 5 Bar plot batch response from each batch

Now in the collected data we needed to make networks for all the four kinds (friends network, Academic support network, financial support network and emotional support network). After fetching the data from database and loading the network data in cytoscape[8].

First thing we did was the identifying module in all the networks. So Figure 9 is friends network. And Figure 10 is academic support network, Figure11 is financial support network and Figure 12 is emotional support network. In all the four networks each colour represents one kind of module. For each network image there is table image attached to it that gives node representation for every discipline to each module (For discussion on module analysis of network refer in Section 1 of Chapter 6)

In Cytoscape, we also did basic data analysis of network attributes. Like node degree distribution plot in Figure 6, Average clustering coefficient and number of neighbour in Figure 7 and Figure 8 is number of neighbour versus average neighbourhood connectivity.

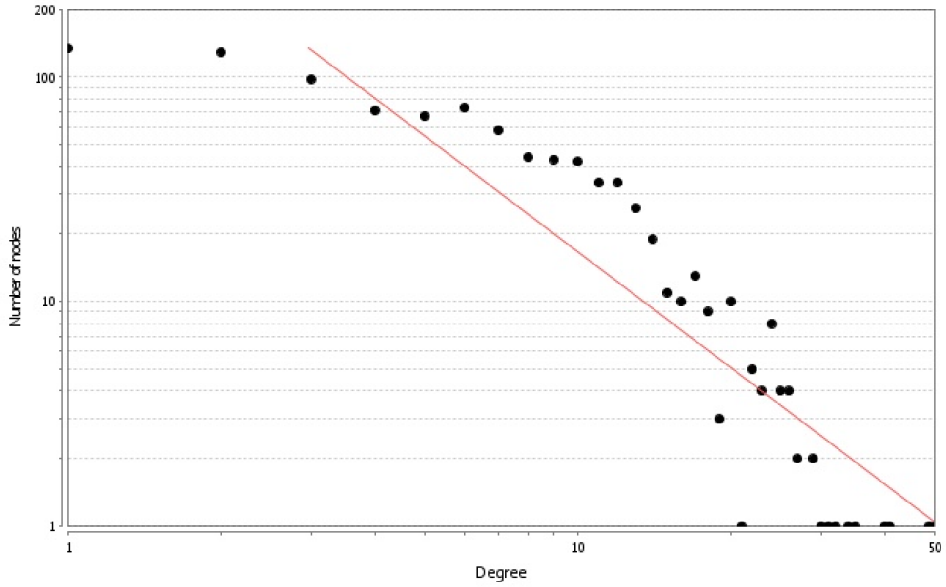


Figure 6 Degree vs Number of nodes

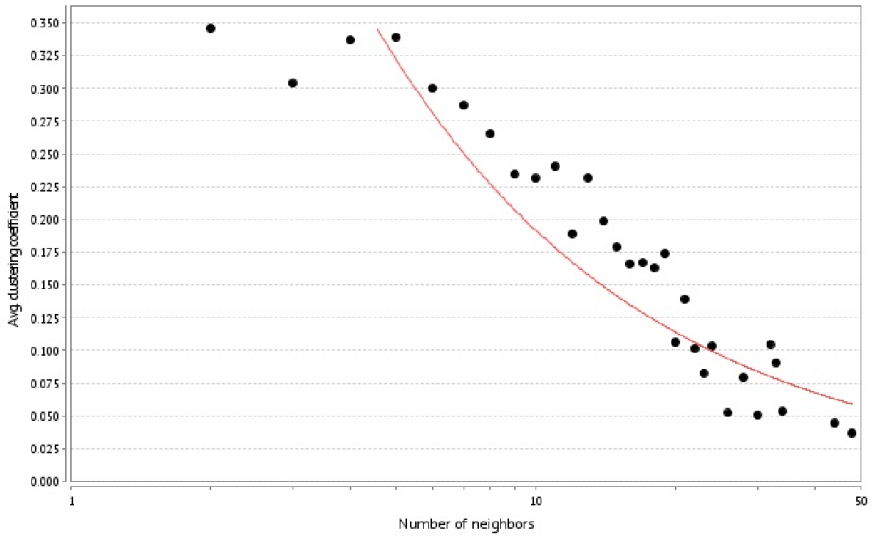


Figure 7 number of neighbour (degree) vs Avg clustering coefficient

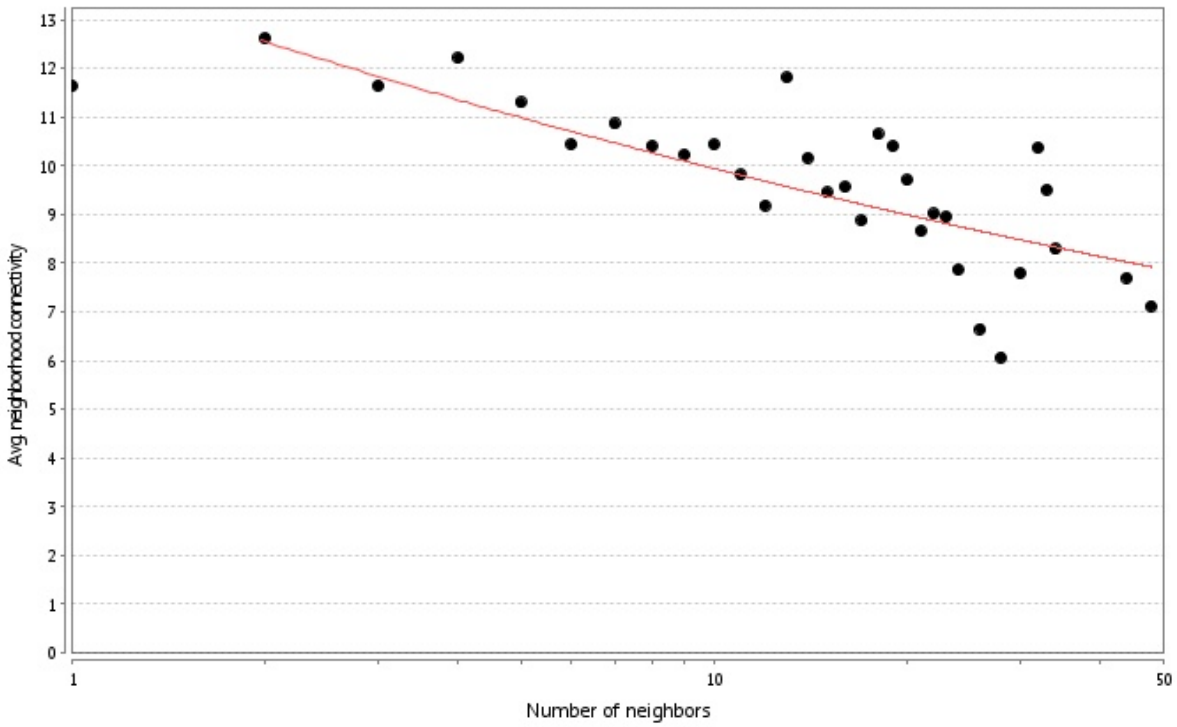


Figure 8 Number of neighbours (degree) vs avg. neighbourhood connectivity

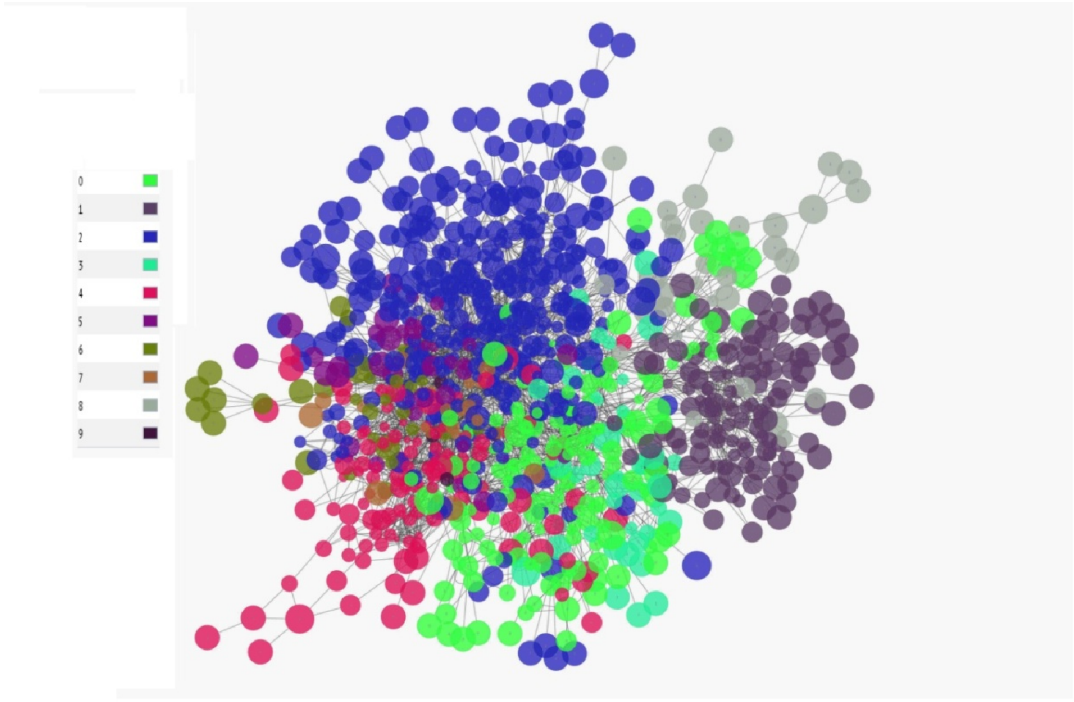


Figure 9 Friends network where each colour represents one kind of module

	ms10	ms11	ms12	ms13	ms14	ms15	mp	phBIO	phPHY	phCHM	phMTH	phNST	others	total	Shannon Entropy
0	0	17	9	3	107	14	5	1	1	1	0	0	2	160	1.753048984
1	0	0	1	0	0	125	0	0	0	0	0	0	0	126	0.0667796411
2	18	19	10	51	12	0	56	80	21	59	15	15	18	374	3.2424883286
3	0	1	0	0	7	0	0	0	25	0	0	4	3	40	1.6093586774
4	0	14	67	27	17	0	1	0	1	1	0	0	0	128	1.8625017537
5	0	0	0	38	3	0	0	1	0	0	0	0	0	42	0.5309809225
6	0	1	23	1	0	0	0	0	0	2	0	0	4	31	1.2754245366
7	0	21	0	0	0	0	0	3	0	0	0	0	0	24	0.5435644432
8	0	0	0	0	1	32	0	0	0	0	0	0	0	33	0.1959092709
9	0	0	0	4	0	0	0	0	0	0	0	0	0	4	0

Table 3 Distributions from all discipline of people in each module, where first column is module id.

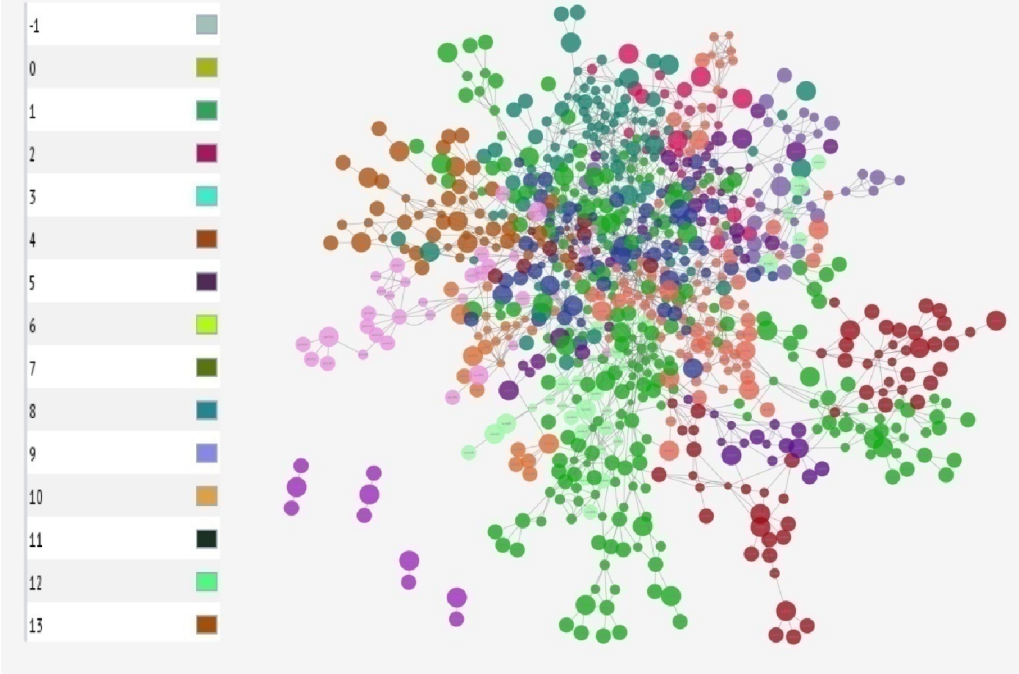


Figure 10 Academic Support network

	ms10	ms11	ms12	ms13	ms14	ms15	mp	phBIO	phPHY	phCHM	phMTH	phINST	others	total	Shannon Entropy
-1	0	0	0	0	5	0	0	0	3	2	0	0	0	10	1.4854752972
0	2	2	61	3	0	0	8	3	0	18	0	0	1	98	1.774471042
1	5	15	26	12	86	48	14	7	11	4	1	0	5	234	2.7985750073
2	0	0	0	30	0	0	0	0	0	4	0	0	0	34	0.5225593745
3	2	36	9	4	0	0	8	7	0	9	0	0	3	78	2.4189621928
4	0	1	0	0	0	5	5	0	0	0	12	0	0	23	1.6436127841
5	0	0	0	0	26	6	0	0	0	0	0	0	0	32	0.6962122601
6	3	2	3	16	0	0	7	55	0	1	0	0	1	88	1.7647038493
7	0	0	0	0	0	15	0	0	0	0	0	0	0	15	0
8	2	0	0	0	0	0	5	2	0	7	0	17	0	33	1.8702125063
9	2	1	0	0	0	0	0	5	34	0	1	0	10	53	1.5807004917
10	0	0	0	25	0	0	9	0	0	1	0	0	0	35	0.9971193096
11	0	0	0	0	0	58	0	0	0	0	0	0	0	58	0
12	0	10	0	0	0	0	2	0	0	0	0	0	0	12	0.6500224217
13	0	0	8	18	0	0	1	0	0	1	0	0	0	28	1.2695459925

Table 4 Module distribution for Academic support network where first column is module id.

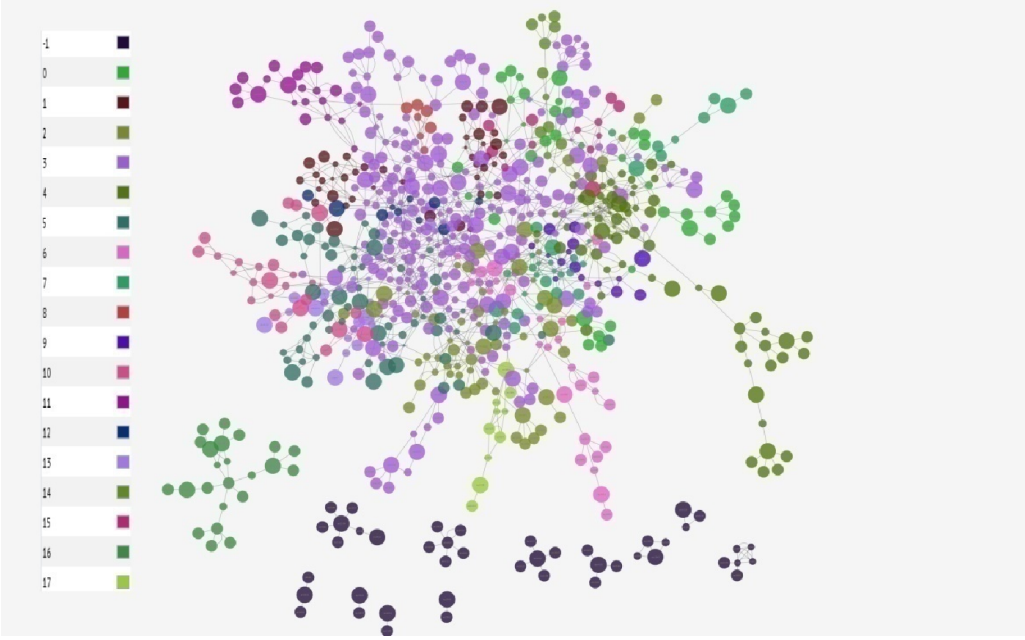


Figure 11 Financial support network

	ms10	ms11	ms12	ms13	ms14	ms15	mp	phBIO	phPHY	phCHM	phMTH	phINST	others	total	Shannon Entropy
-1	0	0	0	4	15	11	2	1	0	2	0	8	1	44	2.4444973745
0	0	0	0	2	24	13	0	0	0	0	0	0	0	39	1.1791248765
1	0	18	3	1	19	0	0	0	1	1	0	0	0	43	1.6931447482
2	2	1	37	11	1	0	4	1	1	5	1	0	1	65	2.1390883852
3	10	29	16	7	37	37	14	62	11	25	2	0	8	258	3.1822929297
4	0	3	38	26	0	0	3	0	1	1	0	0	0	72	1.5707242627
5	1	5	3	2	6	1	21	8	0	7	9	0	0	63	2.8297596327
6	0	0	0	19	0	0	6	0	0	2	0	0	0	27	1.1170949926
7	0	0	0	21	9	0	0	0	0	1	0	0	0	31	1.0584546687
8	0	0	0	7	0	0	0	0	0	0	0	0	0	7	0
9	0	0	0	15	0	0	0	0	0	0	0	0	0	15	0
10	0	1	0	0	0	2	0	0	18	0	1	0	4	26	1.4289531787
11	0	0	0	0	0	21	0	0	0	0	0	0	0	21	0
12	1	7	0	0	0	0	0	0	0	1	0	0	1	10	1.3567796495
13	0	0	0	0	1	0	0	0	11	0	0	2	3	17	1.4516607644
14	0	1	0	0	5	5	0	0	0	0	0	0	0	11	1.3485878959
15	2	0	8	1	0	0	1	0	0	0	0	0	0	12	1.4182958342
16	0	0	0	0	1	22	0	0	0	0	0	0	0	23	0.2580186687
17	0	0	0	0	0	10	0	0	0	0	0	0	0	10	0

Table 5 Module distribution for Financial support network, where first column is module id.

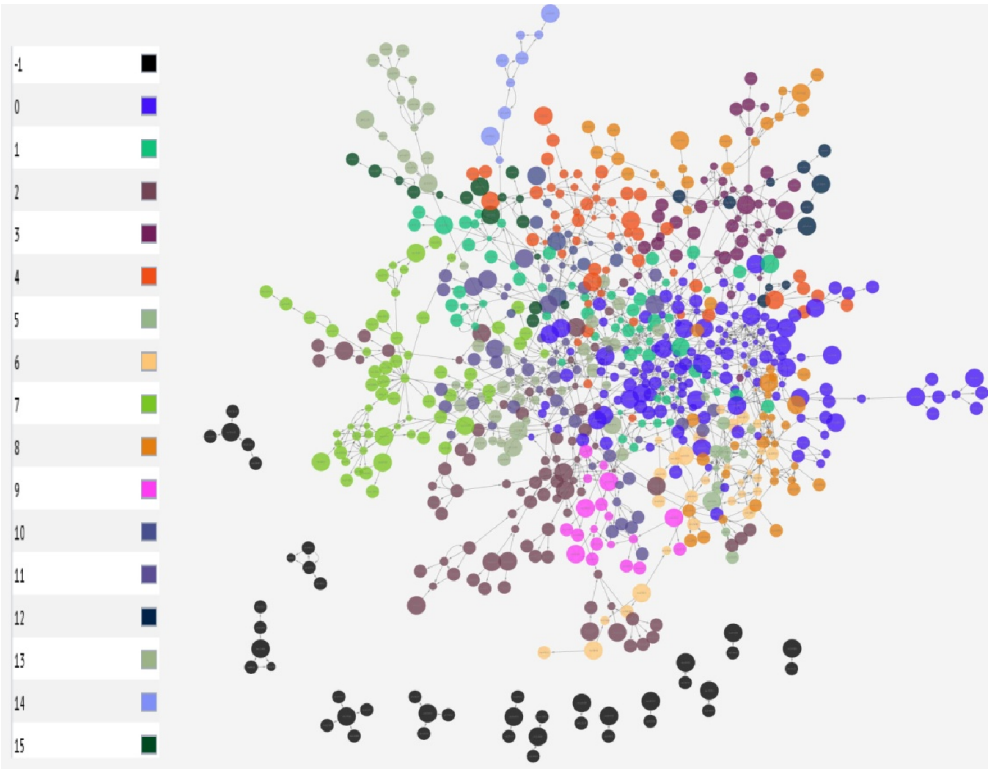


Figure 12 Emotional support network

	ms10	ms11	ms12	ms13	ms14	ms15	mp	phBIO	phPHY	phCHM	phMTH	phINST	others	total	shannon Entropy
-1	0	1	0	1	4	17	6	3	6	3	0	0	1	42	2.5824080584
0	5	29	54	8	5	22	8	4	2	6	0	0	3	146	2.7292984422
1	3	33	3	7	1	0	7	0	2	8	0	0	1	65	2.3104024306
2	0	1	1	18	37	9	4	1	1	2	1	0	5	80	2.3475857194
3	0	0	0	36	7	0	0	1	0	0	0	0	0	44	0.7828683343
4	1	1	0	0	8	6	17	9	4	9	0	1	0	56	2.6992246842
5	0	0	1	10	0	0	0	0	0	1	0	0	0	12	0.8166890884
6	0	0	14	15	0	0	0	0	0	0	0	0	0	29	0.999142104
7	0	0	0	1	0	53	1	0	4	1	0	0	1	61	0.8228720247
8	2	1	25	3	15	0	0	0	6	0	0	1	0	53	2.0116105264
9	0	0	0	2	17	0	0	0	0	0	0	0	0	19	0.4854607608
10	0	0	0	0	6	0	1	1	1	0	5	0	0	14	1.8702538164
11	4	1	0	3	2	8	5	18	1	3	2	0	3	50	2.9051971296
12	0	0	3	0	0	0	3	0	0	0	5	0	1	12	1.8250112109
13	0	0	0	0	6	0	0	39	3	2	0	0	0	50	1.0759492089
14	0	0	0	0	0	0	0	0	0	2	0	7	0	9	0.7642045065
15	0	0	0	0	0	0	0	1	10	3	0	0	1	15	1.3752793655

Table 6 Module distribution for Emotional support network, where first column is module id.

After getting all the properties for friends network we made two sets one was related to network and other was related to individual(personality, BMI, etc). There we only

considered the people for whom we have all the information, for that we looked for correlation among network properties and individual properties on which we made correlation Heat map see Figure 13 (for more discussion in section 2)

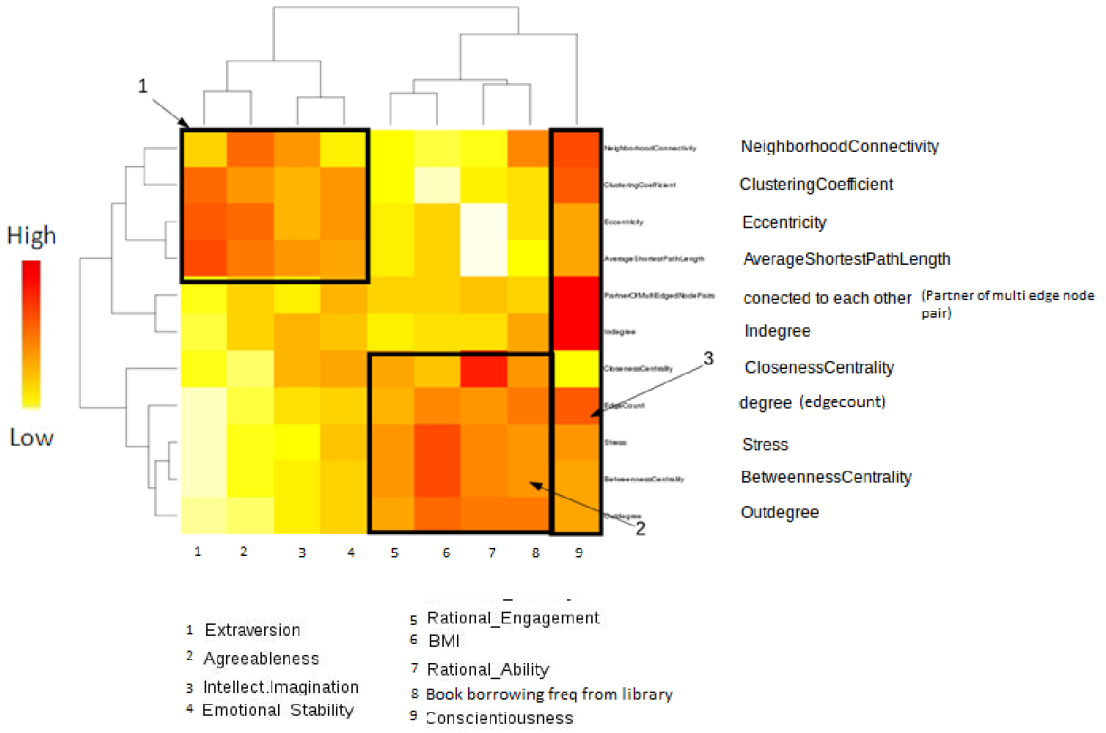


Figure 14 Heatmap of correlation between network properties[11] and individual attributes, where row is individual attributes and column is network attributes

Since there can be some values in networks properties which can be correlated to each other, we did Principal component analysis (PCA) of networks properties for all the nodes. PCA can help us to look for correlation between major components of networks properties and individual's attributes. Here we did PCA on individuals' attributes and Figure 16 shows percentage variance of network parameters explained by first two components because both components together were manage to explain 40% of variance see Figure 15 and taking two components is also easier to analyses. In that, it suggests that clustering related parameters like clustering coefficient are equally represented in Component-1, while component-2 is largely composed of centrality related parameters like betweenness, degree etc. Figure14 is Biplot which shows each attribute lies where in components 1 and 2. It depicts that component -1 & 2 together explain the attributes like conscientiousness, emotional stability, intellectual imagination and rational engagement. While component-1 alone explains Extraversion, agreeableness, and rational ability. BMI is largely explained by component-2 For more discussion on PCR results, see section 3 in Chapter 6.

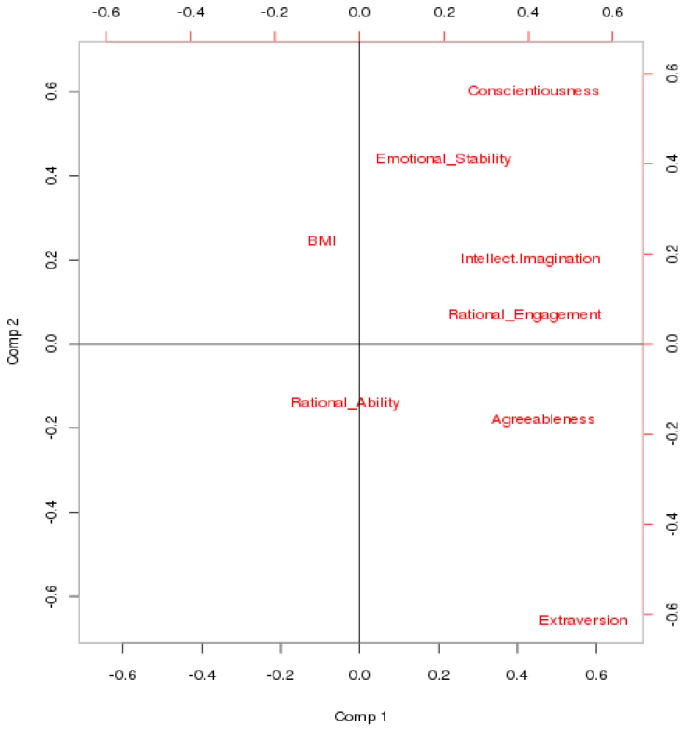


Figure 14 Biplot for first two components of Principal components analysis

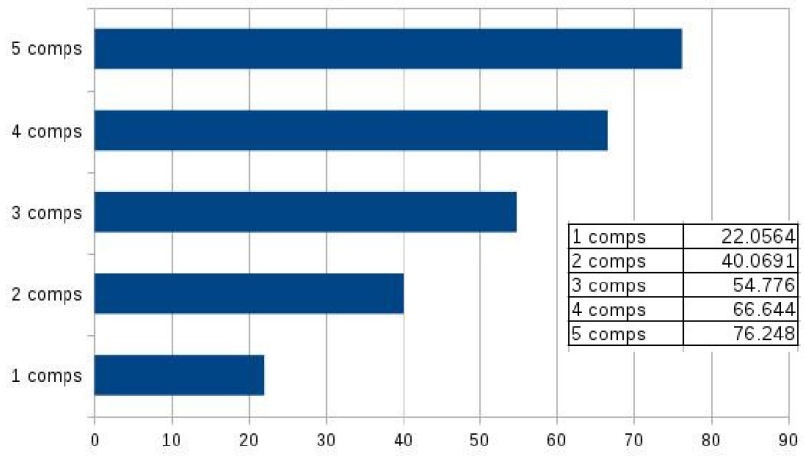


Figure 15 Percentage variance explained by each PCA components (cumulative)

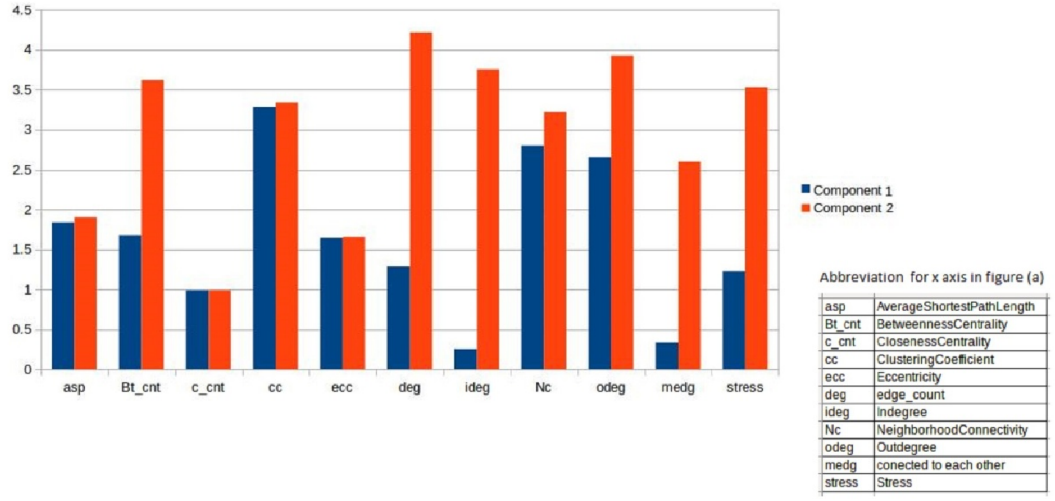


Figure 16 Distribution of explained variance with first two components by PCR

Chapter 6

(Discussion)

In all four networks one thing which is apparent is that there is always one big connected component and few isolated small components and only in friendship network there is one connected component. People who had completed first part of network question were around 43.34% of total strength and total number of nodes in friends network is 967 which is around 89.6% from all the nodes(1068). This implies that from only 43.34% of population manage to cover almost all the people, who most probably are not isolated nodes. Having only one connected component in friends network is also surprising that in spite of people from different parts of country, they are not completely isolated from each other.

1) **Module analysis of networks**

Next thing we did in all the networks is module analysis, which was done by using ModuLand[9] a Cytoscape app. ModuLand tries to determine modules in network and module is a sub network of network which has more edges among themselves and less to outside of sub network. Module distribution tables for all the networks are there with network images. Where Shannon entropy[10] (SE) (for a module in network $SE = \sum_{i=0}^{n} p_i \ log_2(p_i)$ where p_i is $p_i = \frac{\text{no of elements in i}^{\text{th}} \text{ discipline of that module}}{\text{total elements in that module}}$) tries to measure homogeneity in each module of network, for a module more the value of SE means it has more people from different backgrounds where values in our case will be in between 0 to 3.7. So in case of Friends network (Figure 9 and Table 3)module '2' has Shannon entropy of 3.2 and contains almost all phds and inphds(mp) students. Which could mean that these people are working as ties with other ms students and also among each other and their major field of work does not have much effect on their friendship network. In case of academic network (Figure 10 and Table 4) there isn't one big value for SE. But even there for module '1' SE value is maximum of 2.7 and it has more people from ms14 and ms15 and less people from other parts. This probably is because of people chose their major after 2nd year and until then all ms14 have same courses and same is true for ms15. So ms15 people might go to their first senior (ms14) to ask doubts, and in this module there are people from other disciplines who might be helping them (ms14 and ms15) where

people are less from phds and intphds but more from senior batch of ms11,ms12 and ms13. In module distribution of financial support network (Figure 11 and Table 5) is kind of same as friends network at least for case of phds and intphds but it has more number of modules compare to friends network and has more values of SE as zero as compare to others networks. This could be because people feel more comfortable to ask money to people who are in their batch. Only in the module analysis of emotional support network there isn't any module which has value of SE zero where its minimum is 0.4. So it is little surprising that people feel comfortable to talk about personal stuff with people who are not in there batch or share same major.

2) Comparison of values in network properties and individuals' attributes

Now we needed to compare networks properties of individual nodes in friends network with person's other attributes like personality, rationality, BMI, Book borrowing frequency from library etc. For that first we analyzed the network using Cytoscape which gives us values for the network properties (like clustering coefficient, eccentricity, degree, average neighborhood etc) of each node. Then we made heat map (Figure13) for correlation matrix between normalized data of individual's properties and network properties. This heat map is also clustered according to correlation values. One thing which is clear is that personality trait consciousness, which determine how organized or easy going a person is, has strong correlation with indegree and partner of Multi Edge Node Pairs(no of people mentioned back whom individual mentioned) (Figure 13,cluster 3). That could be because these people might be conscious while mentioning friends name and got mention back by those people and also mentioned by people whom he probably didn't mentioned. Other than that there were two more clusters named 1 and 2 in Figure 13 which has positive correlation with network properties. In cluster 1 extraversion shows positive correlation with eccentricity and average short path, and negative correlation with edgecount (degree) and partner of Multi Edge Node Pairs which we found to be interesting because one would expect that extrovert people have more number of friends. There it might be possible that extrovert people do not consider everyone (whom they talk to) as their friends and other people also did not mention back. It could also be possible that one's perception of being extrovert is incorrect. In cluster 2 Figure13, we see Rationality to be positively related to stress (number of shortest paths

pass-through that node), betweenness centrality and edgecount(degree). This could mean a person who tends to act rationally is the one makes network strong and he/she is quite popular compare to others. BMI and Book borrowing frequency from library show positive correlation with outdegree, stress and betweenness centrality which means they interact with lot of different people while in cafeteria or in library.

3) PCR of Individual attributes and PCR with network properties

PCR analysis confirms our observation through correlogram. Network attributes like clustering coefficient alone seem to explain personality traits like extraversion, agreeableness and rational ability. This is counter-intuitive since extraversion is generally expected to associate with number of interaction (degree) and not clustering coefficient. However, we see that degree is not associated with extraversion. We explain this contradiction as following: 1) It is likely that the extrovert students tend to very cautious while naming their friends because of their awareness of them being extrovert and intermingling nature. 2) It is equally likely that these students have incorrect self-perception of their personality.

Traits like conscientiousness, emotional stability, intellectual imagination and rational engagement were explained by both components with slightly differing proportion of loading from each component, suggesting that emotionally stable and conscientious students have relatively balanced network traits hinting at their cautious and balanced nature.

Interesting BMI was associated with the centrality related network attributes like degree. We interpret that students with high BMI might be hanging around more in canteens and that might allow them having more friends and hence higher centrality in the network. Importantly they also show high frequency of book borrowing in the library. Library visits might also help them making more friends.

In near future, we need to analyze our data more comprehensively by splitting the students into different batches, disciplines', gender, religion, ancestry, CGPA etc to get better insight of the data.

Chapter 7

(Conclusions)

In summary:

1. We successfully developed a web-platform and a databases to collect the network and personality attributes of IISER-Mohali students
2. We developed a facebook app to collect the facebook information of students that they permit us to access
3. We successfully performed preliminary statistical analyses of the data and learned that:

a) Students network at IISER-Mohali follows scale-free like degree distribution, hierarchy and dis-assortivity, validating the non randomness and reliability of our data.

b) Modules in the network are generally homogenous with an exception of one module wherein PhD and int-PhD students intermingled extensively with BS-MS students and possibly represent the core of the organization.

c) Correlogram and PCR analyses suggested that different personality traits are associated with distinct network properties, highlighting the possible predictive power of the network attribute for students' personality prediction.

Further analyses of the datasets would hopefully reveal more insights to students' organization viz-a-viz their personality and academic performance

(What are analyses still needs to be done)

1) Academic data analysis
2) Batchwise properties of students
3) Network percolation analysis
4) Relational learning (Markov model)
5) Comparison of face-to-face network with the facebook network.

Appendix

1) Network Questions

1) Give friends' name
2) People with whom you do group study or solve assignments or do discussion related to academic?
3) Suppose you need some money (urgent) to pay some bills, whom will you contact among IISERM students?
4) If you need someone to talk to for some kind of emotional support whom will you contact.

2) Personality Questions

1) Am the life of the party
2) Feel little concern for others
3) Am always prepared
4) Get stressed out easily
5) Have a rich vocabulary
6) Don't talk a lot
7) Am interested in people
8) Leave my belongings around
9) Am relaxed most of the time
10) Have difficulty understanding abstract ideas
11) Feel comfortable around people
12) Insult people
13) Pay attention to details
14) Worry about things
15) Have a vivid imagination
16) Keep in the background
17) Sympathize with others' feelings
18) Make a mess of things
19) Seldom feel blue(sad)

20) Am not interested in abstract ideas

21) Start conversations

22) Am not interested in other people's problems

23) Get chores(routine work) done right away

24) Am easily disturbed

25) Have excellent ideas

26) Have little to say

27) Have a soft heart

28) Often forget to put things back in their proper place

29) Get upset easily

30) Do not have a good imagination

31) Talk to a lot of different people at parties

32) Am not really interested in others

33) Like order

34) Change my mood a lot

35) Am quick to understand things

36) Don't like to draw attention to myself

37) Take time out for others

38) Shirk(avoid) my duties

39) Have frequent mood swings

40) Use difficult words

41) Don't mind being the center of attention

42) Feel others' emotions

43) Follow a schedule

44) Get irritated easily

45) Spend time reflecting on things

46) Am quiet around strangers

47) Make people feel at ease

48) Am exacting(attention) in my work

49) Often feel blue(sad)

50) Am full of ideas

3) Rationality Questions

1) I try to avoid situations that require thinking in depth about something
2) I don't like to have to do a lot of thinking
3) I enjoy solving problems that require hard thinking
4) I prefer complex problems to simple problems
5) Thinking hard and for a long time about something gives me little satisfaction
6) Using my gut feelings usually works well for me in figuring out problems in my life.
7) I believe in trusting my hunches
8) I trust my initial feelings about people
9) When it comes to trusting people, I can usually rely on my gut feelings
10) I can usually feel when a person is right or wrong, even if I can't explain how I know

4) Quiz

1) You have already given your answers to Quiz, you can do it only once.
A keyboard and mouse cost 1100 INR in total and price of keyboard is 1000 INR more than the mouse. what is the price of mouse?

2) Is the following conclusion logically valid?
Premise 1: All birds have wings.
Premise 2: Sparrows have wings.
Therefore, Sparrows are birds.

3) A particular mutation is present in one in every 1000 people. A biochemical assay always correctly tells that if mutation is present. The test has a false-positive rate of five per cent i.e, the test wrongly indicates that the mutation is present in five per cent of the cases in which the person does not have the mutation. What is the probability that an individual testing positive actually has the mutation?

4) Suppose you are given four cards by someone. Then he tells you that each card has a letter on one side and a number on the other. Your cards look like this:

G U 6 3

Now he says there is rule: If a card has a vowel on its letter side, it has an even number on its number side. Which card(s) must be turned over to find out if the rule is true or false?

5) In IISERD, there is a certain number of dogs. Every year, the number doubles. If it takes 10 years for the dogs to reach a number when they can protect entire IISERD, how long would it take for the dogs to reach their half of strength? (Assuming IISERD is place where dogs are immortal and won't leave IISERD in any circumstances and every dog has equal strength.)

5) **Email text that was sent at first place**

Hi!

We are doing social network study of IISER Mohali students. To collect the data from students, we have developed an intra-net web application. This is a humble appeal to you to kindly squeeze some time and fill-in and submit the required information on our application.

For wired connection, click: http://192.168.2.61:8000

For wifi connection, click: http://172.16.100.61:8000

Your login details are as following:

usr: rollno

pwd: password

Many thanks for your Co-operation.

Regards,

Aniket Gaur

ms10050@iisermohali.ac.in

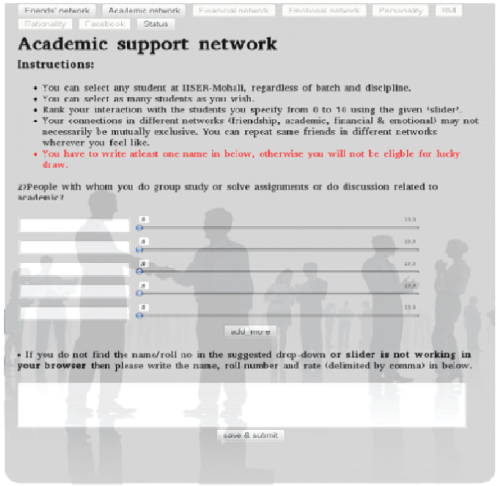
PS: Details about the project are given on the aforementioned page. If you

have specific queries, please write to me (ms10050@iisermohali.ac.in)
or Dr. Kuljeet Sandhu (sandhuks@iisermohali.ac.in).

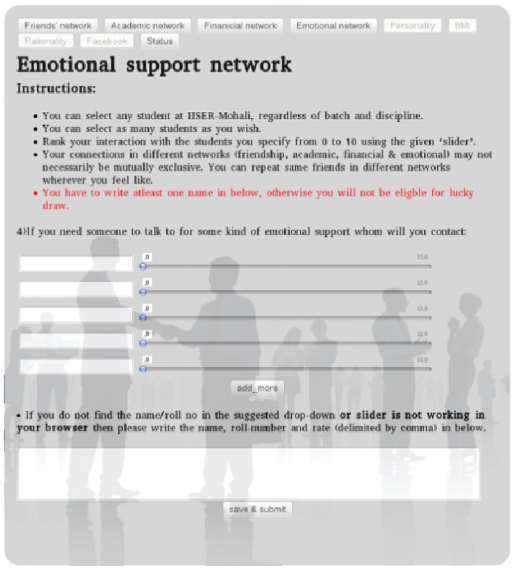
6) Snap-Shots of webapp

Index page of website

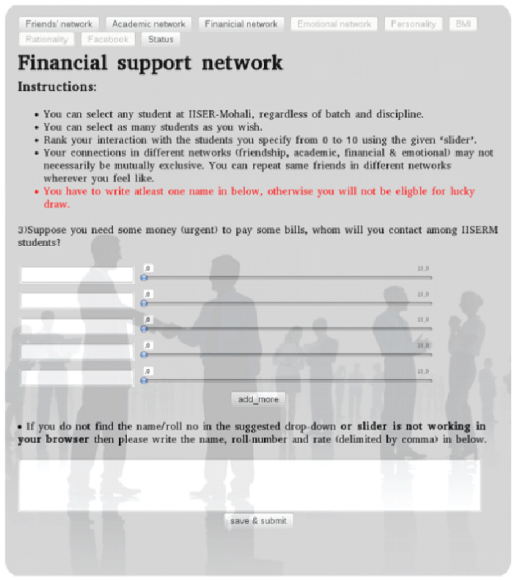
(a)



(b)

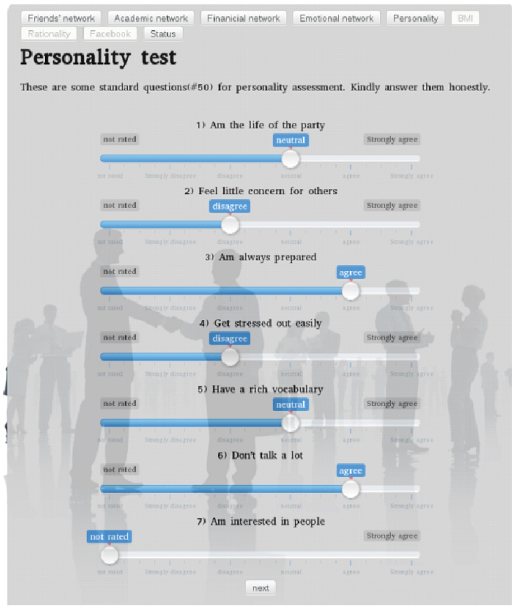


(c)

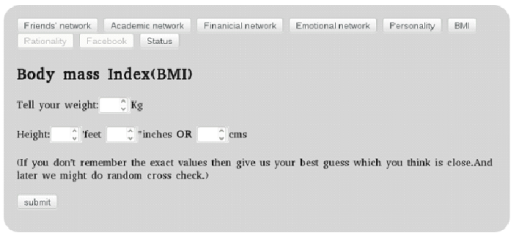


(d)

Where (a), (b), (c) and (d) are web page snapshots for network questions of friends, academic support, emotional support and financial support respectively

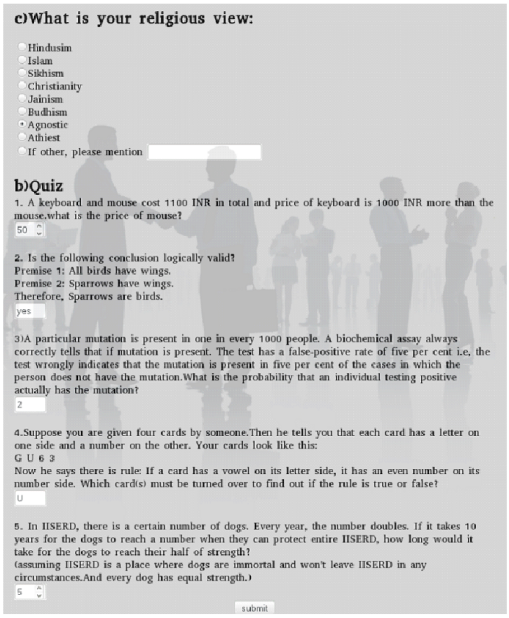


(a)



(b)

(c)



(d)

Where (a), (b),(c) and (d) are web page snapshots for personality page , BMI, Rationality, and quiz parts respectively

References

1. Ipip personality(http://ipip.ori.org/New_IPIP-50-item-scale.htm)
2. Rationality (Cognitive style and nonverbal sensitivity.Cognitive-experiential self theory validation by Loredana)
3. Facebook graph api(https://developers.facebook.com/)
4. Django(https://www.djangoproject.com/)
5. Pseudofractal Scale-free WebS.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes
6. The pls Package: Principal Component and Partial Least Squares Regression in RBjørn-Helge MevikNorwegian University of Life SciencesRon WehrensRadboud University Nijmegen
7. Probabilistic graphical model(Markov model) by Daphne koller and Nir Friedman
8. Cytoscape (http://www.cytoscape.org/)
9. ModuLand plugins for cytoscape by Máté Szalay-Bekő, Robin Palotai, Balázs Szappanos, István A. Kovács,Balázs Papp and Péter Csermely
10. Shannon entropy article by Annick Lesne
11. Cytoscape attributes (http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.6.1/#attributes)