SOCIAL NETWORK BASED ANALYSIS OF BEHAVIOR

A BACHELOR'S THESIS

Submitted in partial fulfillment of the requirements

for the degree

of

BACHELOR OF SCIENCE

in

STATISTICS

Submitted by

Chandrakana Nandi

(Roll No.09181SC103)

Under the Guidance of:

Dr. R. D Singh

Professor

Department of Statistics

BANARAS HINDU UNIVERSITY

VARANASI-221005, INDIA

April, 2012

DECLARATION BY THE CANDIDATE

I, Chandrakana Nandi, do hereby declare that the research work presented in this thesis is fully authentic and is my original work. All the results are conclusions have been independently derived by myself and they are true to the best of my knowledge.

Place: BHU Chandrakana Nandi

Date: Roll no.09181SC103

CERTIFICATE

This is to certify that the work of this project, being submitted in the partial fulfillment of the degree of Bachelor of Science in Statistics at Banaras Hindu University(BHU), Varanasi, is an authentic record of the original work of **Ms**. **Chandrakana Nandi (Roll no.09181SC103)**, carried out under my supervision. She has provided due acknowledgements to all the materials that have been used. The work is done in full compliance with the requirements and constraints of our prescribed curriculum.

Place: BHU

Date:

Dr. R. D Singh

Professor, BHU

ACKNOWLEDGEMENT

The author would like to express her sincere gratitude towards Prof. R.D Singh, for being her mentor for the final semester project. He has been very supportive and provided valuable inputs and suggestions. She would like to thank the head of the department, Prof. K. K. Singh for providing a very stimulating research environment and excellent facilities.

The author is grateful to Prof. S. K Singh who taught her statistical inference and motivated her to implement her ideas. Special thanks to Prof. Sushma Tripathi for her constant support and encouragement. She would like to thank all her teachers in BHU who have been a source of inspiration for her.

The author thanks the lab and library incharges of the Department of Statistics who were always very helpful and allowed her to avail the facilities whenever required.

Finally, she expresses her heartfelt gratitude towards God who gave her the strength to carry on with her work. She thanks her parents for all their support and valuable suggestions. They have always had faith in her and given her the courage to move forward in life.

It is only due to the cumulative support of all that this thesis work could be successfully compiled.

Chandrakana Nandi

B.Sc Final Semester, BHU.

ABSTRACT

Recent developments in social networking have tremendously influenced the behavior of the youth. The research work presented in this thesis has been undertaken in order to conduct a statistical analysis of such behavior of social network users. The purpose is to determine the effect of social networks on the social and academic life of the undergraduate students of Banaras Hindu University (BHU). The data has been collected by means of a sample survey conducted among the students. The **Karl Pearson's correlation coefficient (Product Moment Correlation Coefficient)** has been obtained between the duration of time spent on social networks by a student and his academic performance in terms of his SPGA in the last major examination and also between the activity of a student in the physical world and the virtual world (social networks) in terms of the number of friends he has in them respectively. Several tests of hypothesis have been conducted using the χ^2 **Goodness of Fit Test and the** χ^2 **Independence of Attributes Test** through which the effects of social background, medium of study in school and the level of knowledge in computer science of a student, on his behavior in social networks and the extent to which the users rely on the privacy policies of the social networks have been analyzed. Two **Ordinal Regression models** have been developed with **Complementary log-log link function** and **Cauchit link function** and they have been compared. This work enables us to analyze and demonstrate the effect of the use of social networks on the youth of our society. From the analysis of the data of a social network, we have detected some interesting changes in behavior of social network users which may be helpful in the anticipation of various disasters like psychological problems, aggressive feelings and destructive attitude towards professional career etc. It is very important to detect them at an early stage so that we may be able to take corrective measures.

TABLE OF CONTENTS:

1. INTRODUCTION

1.1 History of Social Networking

The ubiquity of internet has tremendously revolutionized the way we interact with each other. From the advent of email, bulletin board systems, to the current social networking sites, technology has been integrated with communication and ushered a new digital era. It has taken a very leading role in decreasing the digital divide between urban and rural populations. [1]The journey of this revolution began through the first email sent in 1971. Following that, the BBS (Bulletin Board systems) exchanged data over phone lines with other users for the first time in 1978. In the same year, USENET, an online bulletin board established and distributed their first copy of early web browsers. However, the first web based social networking site was not developed until 1994, when GEOCITIES was founded. The basic concept of GEOCITIES was to allow users to create their own websites, followed by THEGLOBE.COM in 1995 and the very famous AOL instant messenger in 1997. The same year, SIXDEGREES.COM developed facilities for online profile creation and creating friend lists. In the year 2002, FRIENDSTER was launched whose user-base grew to 3 million in the first three months! The next year witnessed the arrival of MYSPACE which was conceived as a clone of FRIENDSTER. Finally, in 2004, one of the most popular social networking sites of all times, FACEBOOK was born and very soon it overtook MYSPACE as the leading social networking site. In 2006, another very popular blogging cum social networking site came in to being, which is known as TWITTER. Recent activities such as the up rises in Libya , Egypt, Syria and even in our own country have proved that these social networking sites have become a very powerful forum for bringing together the common masses across the world for raising their voices against many social and political issues. However, a few years back, a group of computer scientists [2] at the Carnegie Mellon University, detected some peculiar behavior among some people in some social networking forums which they could correlate to terrorist activities. Thus it has become a matter of interest for scientists to predict the dynamic behavior of social network users. Such an analysis may facilitate detection of catastrophic incidents like terrorist attacks (9/11 in USA, 26/11 in India) through the use of social networks and also prevent the youth of the society from the bad effects of the excessive

use of social networks. This work is intended to make a statistical analysis of the behavior of the undergraduate students of the Science Faculty of BHU and how their academic and social lives have been affected by social networks. The analysis can be extended to all kinds of populations like corporate associations, government employees and even homemakers!

1.2 Analysis of Previous Work:

The field of research in predicting dynamic behaviors of the social network users is nascent. The scientists from the fields of statistics and computer science from different parts of the world are now actively involved in finding ways to predict the data obtained from the social networking media. The prediction is particularly challenging due to its heterogeneous nature. One of the most significant works has been done by Ian McCulloh [3] of School of Computer Science, CMU. His work is based on the use of the **Neyman –Pearson Lemma for testing of simple hypothesis** to detect the changes in a dynamic social network. He also implemented the **Statistical Process Control** approach for detecting anomalies in the behavior of a stochastic process over time. This approach is widely used for quality control in manufacturing industries and has been found to be very efficient in rapidly detecting changes in a longitudinal network data.

Several methods of analysis of network data over time have been presented in social science literatures since a long time, [4,5,6]. The various methods for analysis of social network data include Markov Chain models, multi- agent simulation models and statistical models. In 1999, K. M Carley [7,8], carried out research on the evolution of social and organizational networks. She has also worked on assessment of terrorist Groups and the impact of alternative courses of action by using a dynamic network approach. Others [9,10], have focused on different statistical models of network change.

1.3 Motivation:

I have been deeply motivated by reading the above research works and thus took up social networking data analysis as my research topic. I strongly believe the Mumbai attack in 26/11,

Parliament attack on 13 the December, 2001 and many more such organized crimes which require a lot of co ordinations and previous planning could have been averted had social network analysts monitored the social ,e-mail and phone networks of the notorious terrorist groups which would in turn facilitate military leaders to react prior to the actual occurrence of the disaster. Unfortunate as it may seem, it is the young mind which is most vulnerable to provocations which diverts them from their goals and causes indulgence in wrong deeds. Apart from the detection of terrorist activities, it is equally important to focus on the day-to-day social network activities of the youth of the major educational institutions of our nation. I believe that apriori monitoring of their behaviors could bring diverted minds back to the main stream and save their futures.

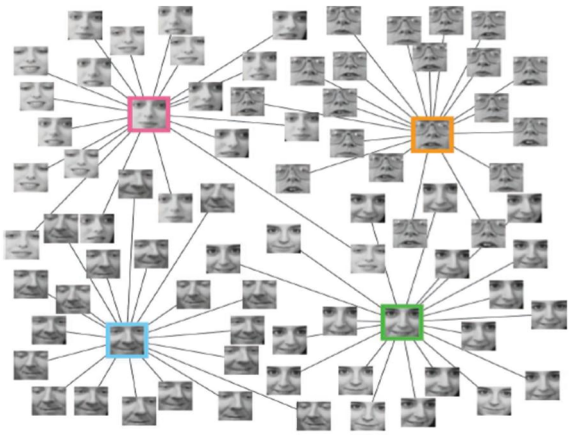


Fig 1. Links in a typical social network[2]

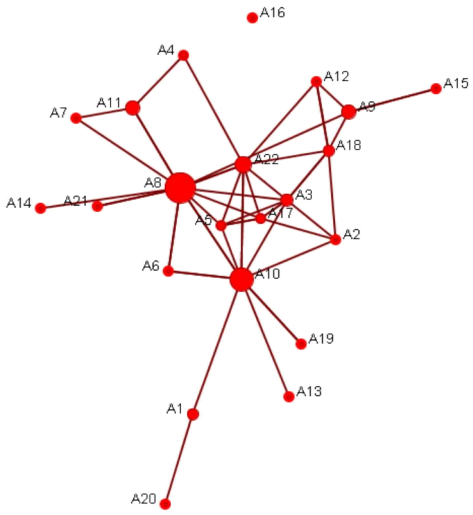


Fig 2. A typical Social Network in Nodes and Edges[2]

1.4 Statement of the problem:

The major challenge of analyzing the data from social networks is its availability and heterogeneity. Two typical social networks have been depicted in fig.1 and fig 2. The problem of converting raw communication data into graphical forms has been addressed and many software tools such as R are available for this purpose [11]. However, due to privacy and confidentiality issues, I faced a lot of trouble in collecting data. Thus, I decided to collect reliable data by means of an exhaustive sample survey conducted among the undergraduate students of the Faculty of Science of Banaras Hindu University (BHU). Based on this data the behaviors of the students have been statistically analyzed. The objective is to determine whether social networks are having a negative influence on the academics of the students and if they are losing their abilities to socialize in the physical world due to the immense popularity of social networks by means of correlation coefficients. In this thesis, the following problems have been addressed:

- Correlation between the SPGA and duration of time spent on social networks
- Correlation between the activity in the physical and virtual world
- Chi-square goodness of fit test for determining the purpose of using social networks

- Chi-square goodness of fit test for finding whether students trust the privacy policies of the social networks.
- Testing the independence of the following pairs of random variables:
    - Medium of study and social network activity
    - Knowledge of computers and social network activity
    - Awareness about social networks and activity in social networks
    - Expression of emotions in social networks and trusting their privacy policies.
- Ordinal regression models for factors affecting the activity of a student in a social network.

2. METHODOLOGY

2.1 Tools and Techniques used:

The major portion of any statistical project is collection of data, which is also one of the most difficult tasks. In my project, I have collected the data by means of a sample survey conducted in my university. The questionnaire has been included at the end of the thesis. My population is the entire group of undergraduate students of the Faculty of Science. Depending on the number of students in each year, I decided the proportion of the number of samples to be taken from each of them. The work in this project has been mainly done using **SPSS [12]**. It is a statistical software that enables treatment of the data by different statistical tools.

2.1.1 Variable types:

1. Nominal Variable: A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation

2. Ordinal variable: A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.

3. Scale Variable: A variable can be treated as scale when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars.

In the data of this project all the questions had about 2,3,4 or 5 options and each option represents a particular category. Thus, all the analysis is done based on the assumption that the data is ordinal.

2.1.2 Karl Pearson's Product Moment Correlation:

My basic objective was to find out the impact of using social networks on the academic performance of the students, for which I have obtained the **Karl Pearson's Product Moment Correlation** coefficient. I have also obtained the correlation between the activity of a student in the real world and in social networks where, activity is measured in terms if the number of friends a student has.

The formula for the correlation between two random variables X and Y is [12]:

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \quad \textbf{(1)}$$

where $\bar{x}$ and $\bar{y}$ denote the arithmetic mean of X and Y respectively.

2.1.3 Ordinal Regression Model:

When our objective is to conduct an analysis of ordinal responses, the linear regression models cannot be used efficiently. Those methods can work only by assuming that the outcome (dependent) variable is measured on an interval scale. This is not true for ordinal outcome variables and thus the simplifying assumptions of the linear regression model are not satisfied due to which the regression model may not accurately portray the relationships among the data. In particular, linear regression depends on the way we define categories of the target variable.

In case of an ordinal variable, ordering of the categories is very important. Thus, if we are collapsing two adjacent categories into one large category, the changes are not very significant, and models built on the basis of the old and new categories are expected to be quite similar. However, since linear regression models are sensitive to the categorization, models built before and after the merging of categories are going to be different. To overcome this problem, we use a generalization of linear regression called a **Generalized Linear Model** in order to predict

cumulative probabilities for categorical data. With this method, we can fit a separate equation for each category of the ordinal dependent variable. These equations give a predicted probability of the response falling in the corresponding category or any lower category. Generalized linear models are a very powerful class of models, which can be used for various types of statistical analysis. The basic form of a generalized linear model is shown in the following equation[12].

link (γ_{ij})=θ_j−[$\beta_1 x_{i1}$+$\beta_2 x_{i2}$+...+$\beta_p x_{iJ}$] **(2)**

where,

link( ) is the link function

γ_{ij} is the cumulative probability of the j^{th} category for the i^{th} case

θ_j is the threshold for the j^{th} category

p is the number of regression coefficients

$x_{i1}...x_{ip}$ are the values of the predictors for the i^{th} case

$\beta_1...\beta_p$ are regression coefficients

Ordinal Regression allows us to model the dependence of a polytomous ordinal response on a set of predictors. The predictors are factors or covariates. The design of Ordinal Regression is based on the methodology of McCullagh [12], and the procedure is referred to as PLUM in the syntax of SPSS.

There are three major components of an ordinal regression model:

1. Location component: The portion of the equation shown above which includes the coefficients and predictor variables, is called the location component of the model. It uses the predictor variables to calculate predicted probabilities of membership in the categories for each case.

2. Scale component: The scale component is an optional modification to the basic model to account for differences in variability for different values of the predictor variables. The model with a scale component follows the form shown in this equation.

$$\text{link}(\gamma_{ij})=\frac{\theta_j-[\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{iJ}]}{e^{\tau_1 z_{i1}+\tau_2 z_{i2}+\cdots+\tau_m z_{im}}} \quad \textbf{(3)}$$

where

$z_{i1}\ldots z_{im}$ are scale component predictors (a subset of the x's)

$\tau_1\ldots\tau_m$ are scale component coefficients

In order to use the ordinal regression model, we have to select a **Link Function.** The link function is a transformation of the cumulative probabilities that allows estimation of the model. In this project, I have chosen two link functions, namely the **Complementary log-log link function** and the **Cauchit link function** and then a comparison between the two has been made.

2.1.4 Tests of Hypothesis:

Some of the analysis required testing of hypothesis, for which I have used the **chi-square 'goodness of fit' test** and the **chi-square 'independence of attributes' test**.

2.1.4.1. χ^2goodness of fit :

The chi-square goodness of fit test is used to determine how well a statistical model fits a set of observed data. It is used to test if a sample of data came from a population with a particular probability distribution. In this test, we tabulate the data and find out whether there is any significant difference in the expected value and the observed value of a random variable. The test statistic is:

$$\chi^2=\sum_{i=1}^{n}\frac{(o_i-e_i)^2}{e_i} \quad \textbf{(4)}$$

where,

o_i- i^{th}observed frequency

e_i- i^{th} expected frequency

and χ^2 follows the chi-square distribution with (n-1) degrees of freedom.

The test procedure is: **For a two tailed test, reject the null hypothesis H_o at α% level of significance, if calculated $|\chi^2|$> tabulated χ^2_{n-1}**

2.1.4.2. χ^2 test for independence of attributes:

Let us suppose that we have two random variables, **X** and **Y**, with X having **m** levels and Y having **n** levels. Let us be testing:

$$H_o: X \text{ and } Y \text{ are independent}$$

against

$$H_1: X \text{ and } Y \text{ are not independent}$$

Then, the test statistic is:

$$\chi^2 = \sum_{r=1}^{m}\sum_{c=1}^{n}\frac{(o_{r,c}-e_{r,c})^2}{e_{r,c}} \quad \textbf{(5)}$$

where,

$o_{r,c}$- observed frequency of the r^{th}level of X and c^{th} level of Y

$e_{r,c}$- expected frequency of the r^{th}level of X and c^{th}level of Y

And χ^2 follows chi-square distribution with $(m-1)\times(n-1)$ degrees of freedom.

The test procedure is**: For a two tailed test, reject H_o at α% level of significance if the calculated | χ^2 | > tabulated $\chi^2_{(m-1)(n-1)}$**

2.2 Formulation of the present problem:

2.2.1 Collection of data:

The total population size (total number of undergraduate students in the Faculty of Science, BHU) =2083, with

number of students in the first year=917

number of students in the second year=671

number of students in the third year=495 and

Sample size=120

The proportions in which the samples are collected are:

First year: $\frac{917}{2083} \times 120$=52(approx)

Second year: $\frac{671}{2083} \times 120$=38(approx)

Third year: $\frac{495}{2083} \times 120$ =28(approx)

The data has been collected by conducting a sample survey in which the questionnaires were distributed approximately according to the above ratio. Initially, I wanted to gather the data from the history of the web browsers and from the social network profiles for the students individually which would have provided much more reliable information, but due to privacy issues, that could not be done, and thus I implemented the idea of conducting a survey. However, a survey has its own disadvantages such as missing data, inaccuracy on the part of the respondent and their lack of interest and awareness, which have some effect on the results. After the data has been collected, it has been fed in to the SPSS software for further analysis.

2.2.2 Analysis:

2.2.2.1 Correlations:

Students are the most frequent users of social networks. It is one of their favorite modes of entertainment and many –a- times, it is found to have a bad impact on their academic performance. To determine whether social networks actually have any influence on the studies of a student, I have found out a correlation between the SPGA of a student in his/her last major examination and the frequency with which the student uses social networks.
Also, a correlation between the activity of a student in the physical world and the activity in the virtual social networks has been calculated and some very interesting results were obtained. SPSS has an inbuilt function named **correlate** within which I chose the **bi-variate Pearson correlation coefficient** which uses formula **(1)**.

2.2.2.2 Modeling:

Two regression models have been developed for ordinal data using the concept of ordinal regression. The ordinal outcome variable or the dependent variable is the activeness of a student in a social network. The possible predictors are:
1. Sex
2. Medium of study in school, i.e. English, Hindi or any other regional language
3. The kind of background the student belongs to, i.e. rural or urban
4. Whether the student has ever undergone any professional training in core computer science.

Activeness in social networks has two categories: Yes and No. Using histogram we represent the distribution of the values of this ordinal variable. Since the maximum students gave positive response which here can be assumed to be a higher category, we use the complementary log-log link function. For finding out whether the Cauchit link function would have provided a better representation, two models are constructed. To perform the modeling, SPSS provides a function **regression** under which we select **ordinal.** The scale component has not been included in this research. We then analyze the model fitting information, the goodness of fit table, the **pseudo** $\boldsymbol{R^2}$ values to determine whether the model is good.

2.2.2.3 Testing:

Two chi-square 'independence of attributes' tests and two non-parametric chi-square 'goodness of fit' tests have been carried out for which there are inbuilt functions in SPSS: for goodness of fit, we choose non-parametric tests and within that select chi-square which is based on formula **(4).** For independence of attributes, we go to descriptive statistics and select chi-square statistics under the crosstabs option. We get a case processing summary with the total number of data in each category mentioned in a tabulated manner along with the number of missing values. Then a cross tabulation is generated by SPSS between the two variables whose independence is tested for. Finally the value of the chi-square statistic is calculated using formula **(5)** based on which we take our decision for a given level of significance.

The following tests are conducted:

1. Non-parametric chi-square goodness of fit test for determining the purpose of using social networks:

H_o : Students use social networks for staying in touch with friends.

H_1 : Students use social networks for staying in touch with friends and academic purposes as well.

2. Non-parametric chi-square goodness of fit test for finding out whether people trust the privacy policy of social networks.

H_o : Students trust the privacy policies of the social networks.

H_1 : Students do not trust the privacy policies of the social networks.

The test statistic that is used for both the above tests is given in (4) and the procedure is as mentioned in 2.1.4.1

3. Independence of the medium of study at school and activeness of a student in social networks.

H_o : Activity of a student in social networks is not affected by the medium of his study at school.
H_1 : Activity of a student in social networks is affected by the medium of his study at school.

4. Independence of the knowledge of computers and the activity of a student in social networks.

H_o : Activity of a student in social networks is not affected by his/her knowledge about computers.
H_1 : Activity of a student in social networks is not affected by his/her knowledge about computers.

5. Independence of the awareness about social networks of a student and his activeness in social networks

H_o : Activity of a student in social networks and his/her awareness about social networks are dependent on each other.
H_1 : Activity of a student in social networks and his/her awareness about social networks are independent of each other.

6. Independence of the trust of a student on the privacy policy of social networks and whether he/she feels comfortable in sharing emotions in social networks:

H_o : Expressing emotions in social networks depends on the trust a student has on social networks.
H_1 : Expressing emotions in social networks does not depend on the trust a student has on social networks.
The test statistic that is used for the above tests is given in (5) and the procedure is as mentioned in 2.1.4.2

2.2.2.4 Pictorial representation of the responses:

Histogram representations of the responses of the students for each of the items on the questionnaire are generated, which helps in an efficient interpretation of the results. In SPSS, pictorial representation of data is very simple. There is an icon called Graphs within which we go to the chart builder and choose our desired chart type. I have chosen the bar chart. X-axis has the variable to be analyzed and the Y-axis has the count.

3. RESULTS AND DISCUSSIONS

3.1 Correlation coefficients:

3.1.1. Association between SGPA and social network activity:

Correlations

		SGPA	frequency of using SN
SGPA	Pearson Correlation	1	-.034
	Sig. (2-tailed)		.716
	N	119	115
frequency of using SN	Pearson Correlation	-.034	1
	Sig. (2-tailed)	.716	
	N	115	116

Interpretation:

We can see from the above table that the correlation between SGPA and frequency of using social networks =**-.034**

The implication of a negative correlation is that use of social networks has a negative impact on the GPA of a student. Thus, students should spend less time on social networking sites in-order to improve their academic performance. However, we have a large p-value **(.716),** which implies that the correlation is not very significant.

3.1.2. Association between activity in the physical world and the virtual world (social networks) :

Correlations

		friends in real world	friends in most used SN
friends in real world	Pearson Correlation	1	.194*
	Sig. (2-tailed)		.039
	N	119	113
friends in most used SN	Pearson Correlation	.194*	1
	Sig. (2-tailed)	.039	
	N	113	114

*. Correlation is significant at the 0.05 level (2-tailed).

Interpretation:

The positive correlation with value=**0.194** indicates that the activity of a student in the real world is directly proportional to the activity of a student in the virtual world. The more a student interacts in the real world, more will be his/her friends in social networks. This result is quite significant because it has become a trend to add ones acquaintances in the social networks in order to stay in touch with them even when people are physically far away. Here, since the p-value is quite less **(.039)**, the correlation is significant and interestingly, it implies that if a student wishes to have more friends in a social network, he/she should start going out and socializing in the real world!

3.2 Regression Models:

Case Processing Summary

		N	Marginal Percentage
activeness	n	42	35.6%
	y	76	64.4%
Medium	English	70	59.3%
	Hindi	48	40.7%
Background	rural	52	44.1%
	urban	66	55.9%
any course on CS	n	92	78.0%
	y	24	20.3%
	2	2	1.7%
Sex	female	49	41.5%
	male	69	58.5%
Valid		118	100.0%
Missing		2	
Total		120	

3.2.1. Complementary log-log link function:

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	51.451			
Final	41.165	10.286	5	.068

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	11.858	11	.374
Deviance	14.467	11	.208

Pseudo R-Square

Cox and Snell	.083
Nagelkerke	.115
McFadden	.067

Interpretation:

We can determine if this model gives adequate predictions from the model fitting information table. We have two values for the -**2log likelihood** and a chi-square value which is the

difference between the two -2log likelihood values. The value of the significance is **0.068** which is not very high indicating that the model gives a good prediction of the probability of the outcome category. The goodness of fit table gives chi-square values which are used to test if the observed data is inconsistent with the fitted model. The significance value for the **Pearson chi-square is .374** which is neither too high nor too low, which means that the data and the model predictions do not differ by large values. However, there is a small disadvantage because the chi-square is sensitive to empty cells and thus the significance values are not very reliable. The **pseudo R-square** values are the **coefficients of determination** which provide an idea about the proportion of the variance in the ordinal outcome variable associated with the predictors. The maximum value of R-square is **1** and larger the value of R-square, better is the model. Here, we can see that the values are not very large which means that there are several scopes to improve the model.

3.2.2. Cauchit Link function:

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	51.451			
Final	41.080	10.371	5	.065

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	11.734	11	.384
Deviance	14.382	11	.213

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	51.451			
Final	41.080	10.371	5	.065

Pseudo R-Square

Cox and Snell	.084
Nagelkerke	.116
McFadden	.067

Interpretation:

We can see that there is not much difference between the two models expect that the Pseudo R-square values for the Cauchit link function are slightly higher. So, we can say that this model is a little better than the complementary log-log link function model.

3.3 Testing of hypothesis:

3.3.1. Testing the Independence of 'medium of study' and 'activity in social networks':

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Medium * activeness	119	99.2%	1	.8%	120	100.0%

Medium * activeness Cross-tabulation

		Activity		
count		n	Y	Total
Medium	English	21	50	71
	Hindi	21	27	48
Total		42	77	119

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.519[a]	1	.112		
Continuity Correction[b]	1.936	1	.164		
Likelihood Ratio	2.503	1	.114		
Fisher's Exact Test				.123	.082
Linear-by-Linear Association	2.498	1	.114		
N of Valid Cases[b]	119				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.94 and b. computed only for a 2x2 table.

Interpretation:

H_o : Activity of a student in social networks is not affected by the medium of his study at school.

H_1 : Activity of a student in social networks is affected by the medium of his study at school.

In the above we are only interested in the Pearson Chi-square value.

The calculated value of χ_1^2 = 2.519, where χ_1^2 follows χ^2 with one degree of freedom.

The tabulated value of χ_1^2 at 5 % level of significance (α) = 3.841459

Since calculated | χ_1^2| < tabulated χ_1^2, thus for a two tailed test, H_0 is not rejected.

Therefore, we conclude that **medium of study of a student has no effect on his activity in social networks.**

3.3.2. Testing the independence of 'knowledge of a student in computers' and 'his activity in social networks ':

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
any course on CS * activeness	119	99.2%	1	.8%	120	100.0%

Knowledge of computers * activeness Cross-tabulation

count		Activeness			Total
		n	y	2	
any course on CS	n	34	58	1	93
	y	8	16	0	24
	2	0	2	0	2
Total		42	76	1	119

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.525[a]	4	.822
Likelihood Ratio	2.383	4	.666
Linear-by-Linear Association	.436	1	.509
N of Valid Cases	119		

a. 5 cells (55.6%) have expected count less than 5.
The minimum expected count is .02.

Interpretation:

H_o : Activity of a student in social networks is not affected by his/her knowledge about computers.

H_1 : Activity of a student in social networks is affected by his/her knowledge about computers.

In the above we are only interested in the Pearson Chi-square value.

The calculated value of χ_4^2 = 1.525, where χ_4^2 follows χ^2 with four degrees of freedom.

The tabulated value of χ_4^2 at 5 % level of significance (α) = 9.487729

Since calculated | χ_4^2| < tabulated χ_4^2, thus for a two tailed test, H_0 is not rejected.

Therefore, we conclude that the **activity of a student in social networks is not affected by his/her knowledge about computers.**

3.3.3. Testing the independence of 'awareness about social networks' and 'activeness in social networks':

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Awareness * activeness	120	100.0%	0	.0%	120	100.0%

Awareness * activeness Cross-tabulation

Count		activeness			Total
		n	y	2	
Awareness	n	4	1	0	5
	y	38	76	1	115
Total		42	77	1	120

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.648[a]	2	.098
Likelihood Ratio	4.477	2	.107
Linear-by-Linear Association	4.499	1	.034
N of Valid Cases	120		

a. 4 cells (66.7%) have expected count less than 5.
The minimum expected count is .04.

Interpretation:

H_o : Awareness and activeness are dependent variables

H_1 : Awareness and activeness are not dependent variables

In the above we are only interested in the Pearson Chi-square value.

The calculated value of χ_2^2 = 4.648, where χ_2^2 follows χ^2 with two degrees of freedom.

The tabulated value of χ_2^2 at 5 % level of significance (α) = 5.991465

Since calculated | χ_2^2| < tabulated χ_2^2, thus for a two tailed test, H_0 is not rejected.

Therefore, we conclude that the **activity of a student in social networks and his/her awareness about social networks are dependent random variables.**

3.3.4. Testing the independence of the 'trust of a student on the privacy policy of social networks' and 'whether he/she feels comfortable in sharing emotions in social networks'

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
trust privacy * expressing feelings	109	90.8%	11	9.2%	120	100.0%

trust privacy * expressing feelings Cross-tabulation

Count		expressing feelings			
		no, i don't think it safe	yes, sometimes	yes, quite often	Total
trust privacy	n	22	18	1	41
	y	26	33	9	68
Total		48	51	10	109

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.748[a]	2	.093
Likelihood Ratio	5.414	2	.067
Linear-by-Linear Association	4.243	1	.039
N of Valid Cases	109		

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.748[a]	2	.093
Likelihood Ratio	5.414	2	.067
Linear-by-Linear Association	4.243	1	.039

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 3.76.

Interpretation:

H_o : Expressing emotions in social networks depends on the trust a student has on social networks.

H_1 : Expressing emotions in social networks does not depend on the trust a student has on social networks.

In the above we are only interested in the Pearson Chi-square value.

The calculated value of χ_2^2 = 4.748, where χ_2^2 follows χ^2 with two degrees of freedom.

The tabulated value of χ_2^2 at 5 % level of significance (α) = 5.991465

Since calculated | χ_2^2| < tabulated χ_2^2, thus for a two tailed test, H_0 is not rejected.

Therefore, we conclude that the **students who trust the privacy policy of social networks tend to share their emotions with other in the social networks more.**

3.3.5. Goodness of fit test to determine the purpose of using social networks:

Purpose

	Observed N	Expected N	Residual
to stay in touch with old friends and make new ones	27	27.5	-.5
to update myself with the recent global events in academics	10	27.5	-17.5
both the above	71	27.5	43.5
Other	2	27.5	-25.5
Total	110		

Test Statistics

	Purpose
Chi-Square	1.036E2[a]
Df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 27.5.

Interpretation:

H_o : Students use social networks for staying in touch with friends.

H_1 : Students use social networks for staying in touch with friends and academic purposes as well.

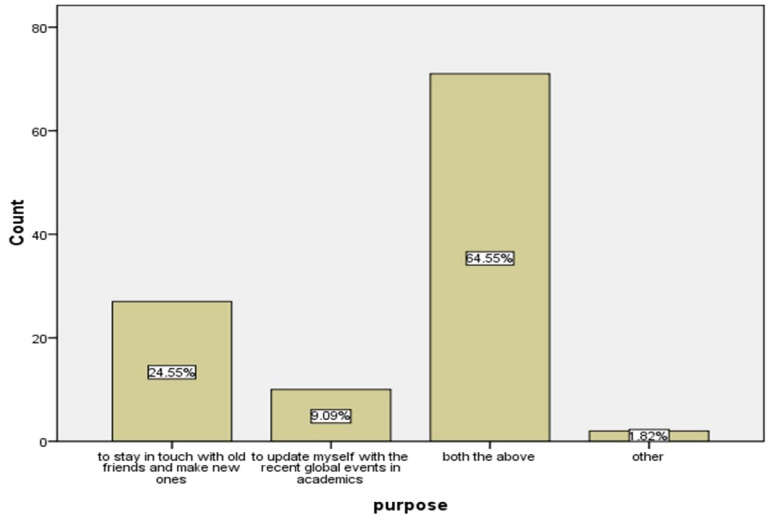
From the above table, we observe:

The calculated value of χ_3^2 $=1.036 \times 10^2$, where χ_3^2 follows χ^2 with three degrees of freedom.

The tabulated value of χ_3^2 at 5 % level of significance (α) = 7.814728

Since calculated | χ_3^2| > tabulatedχ_3^2, thus for a two tailed test, H_0 is rejected.

Therefore, we conclude that **the students use social networks for entertainment, i.e. socializing with friends and also for academic purposes.**



3.3.6. Goodness of fit test to determine whether students trust the privacy policies of the social networks:

trust privacy

	Observed N	Expected N	Residual
N	45	58.0	-13.0
Y	71	58.0	13.0
Total	116		

Test Statistics

	trust privacy
Chi-Square	5.828[a]
Df	1
Asymp. Sig.	.016

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 58.0.

Interpretation:

H_o : Students trust the privacy policies of the social networks.

H_1 : Students do not trust the privacy policies of the social networks.

From the above table, we observe:

The calculated value of χ_1^2 =5.828, where χ_1^2 follows χ^2 with one degree of freedom.

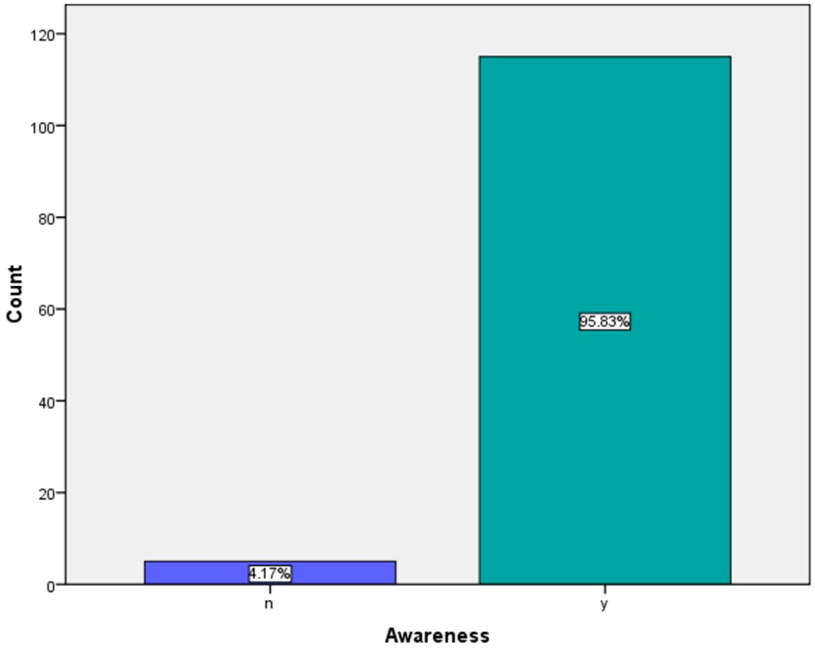
The tabulated value of χ_1^2 at 5 % level of significance (α) = 3.841459

Since calculated | χ_1^2| > tabulatedχ_1^2, thus for a two tailed test, H_0 is rejected.

Therefore, we conclude that **the students do not trust the privacy policies of social networks.**

3.4 Histograms for interpreting the responses of the students:

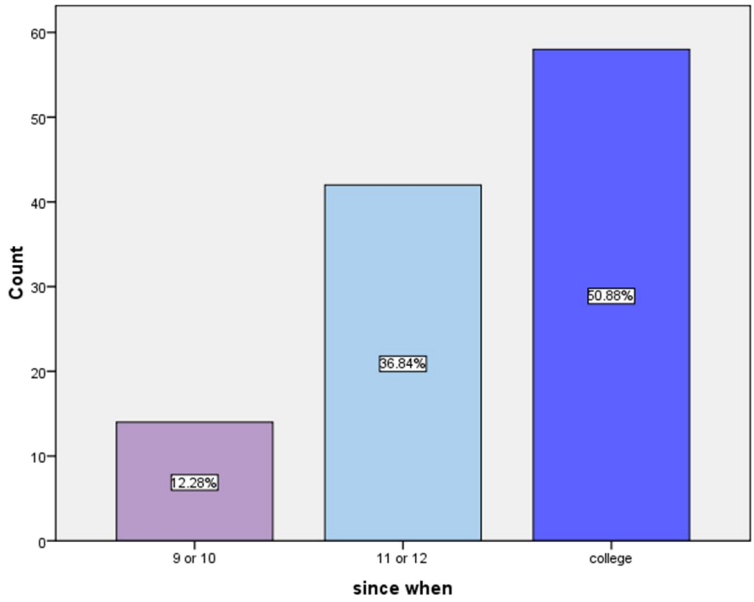
3.4.1. Awareness about social networks:



Interpretation:

From the above histogram we observe that **95.83%** of the students are aware of social networking, while **4.17%** of them are not. This is an indication of how fast, social networking is becoming one of the most sought after sources of entertainment.

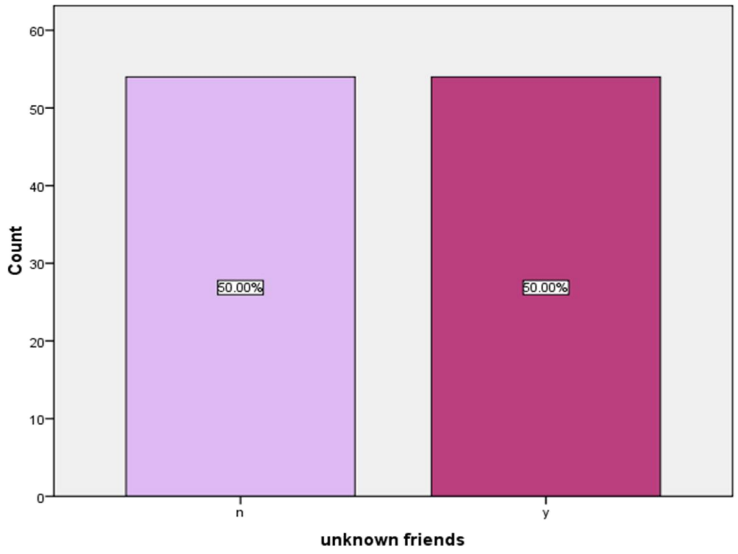
3.4.2. When do students usually begin using social networks?



Interpretation:

The above implies that most of the students start actively participating in social networking after they join college. In a sample of 120, **50.88%** of the students began social networking in college.**36.84%** of the students under test started using social networks in their higher secondary years at school and **12.28%** of them started social networking in ninth and tenth grade at school.

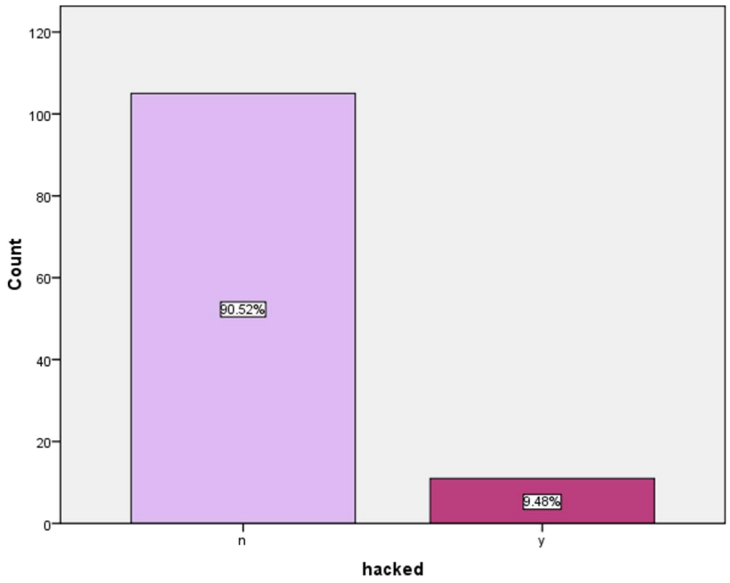
3.4.3. Do students have unknown friends in social networks?



Interpretation:

The above histogram tells us that **50%** of the sample students like to have unknown friends in social networks. This is a very interesting result and indicates that social networks have increased the outgoingness of the youth in our society. Many a times, they are shy to interact with people physically, but through social networks, they are even making friends with unknown people! However, this could also have an adverse effect on our society. There are people in social networks with wrong intentions who may lure the others into becoming their friends. In such cases, innocent people may get involved into illegal issues like terrorist activities, drug trafficking etc.

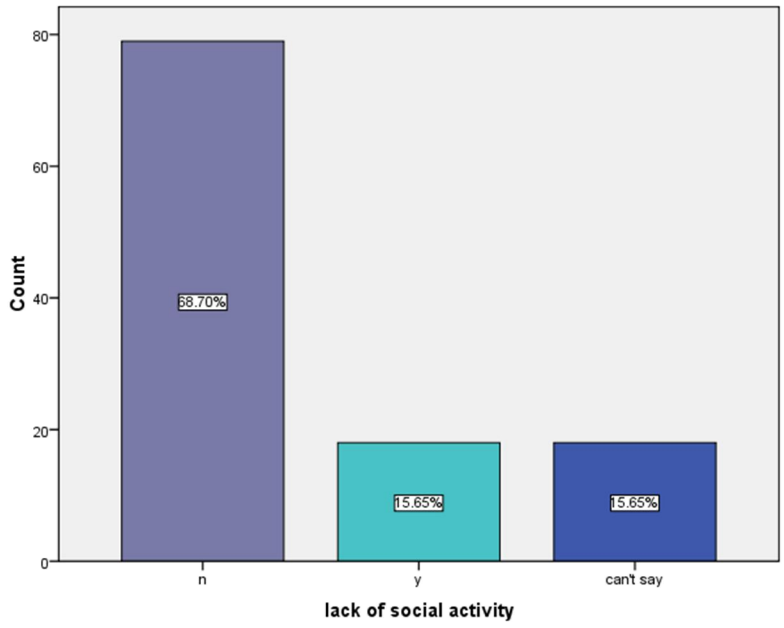
3.4.4. Do students experience the problem of their accounts getting hacked?



Interpretation:

We see that **90.52%** of the 120 students have never faced the problem of their accounts being hacked, while **9.48 %** students have experienced this problem.

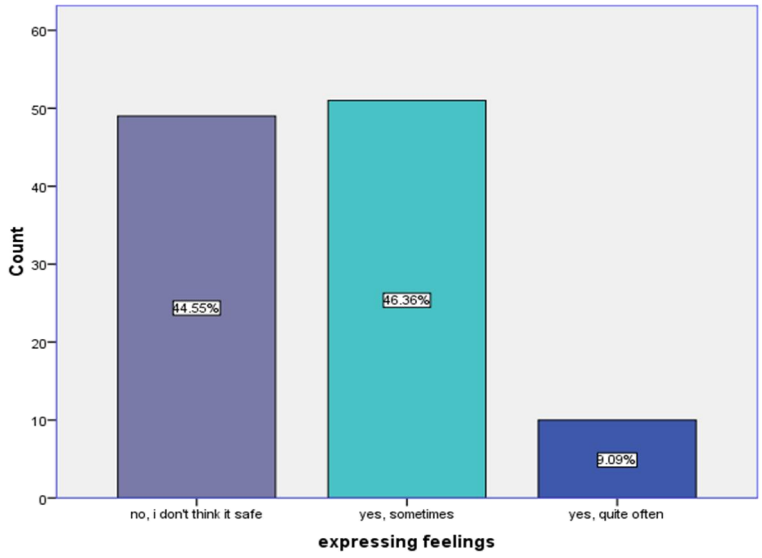
3.4.5. Do social networks affect the extent of interaction of a student in the physical world?



Interpretation:

We see that **68.70%** of the sample students do not feel that their social activities have been hampered by the modern 'broadband' social networks. **15.65%** of them feel otherwise. They think that due to the popularity of social networks, they have become lethargic and prefer to spend their leisure on social networks rather than going out and meeting people and making acquaintances physically. The other **15.65%** people were neutral.

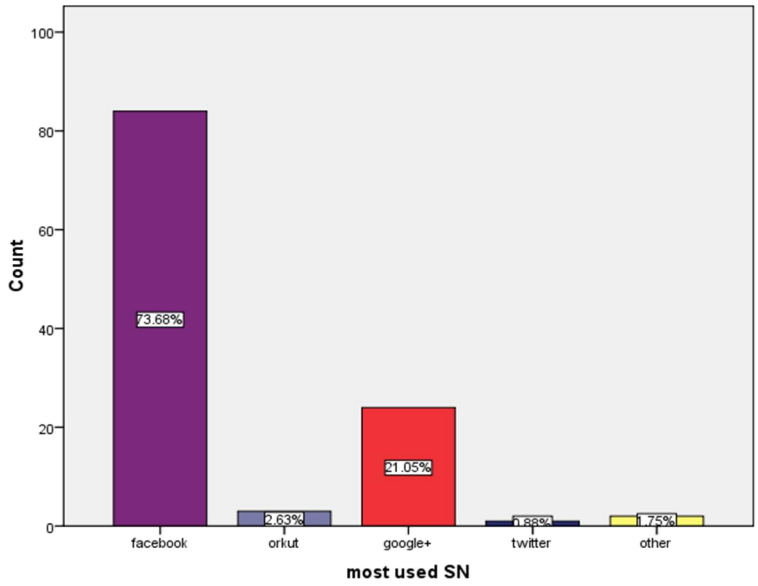
3.4.6. Do students express their emotions in social networks?



Interpretation:

About **44.55%** of the sample students feel it unsafe to share their emotions openly in social networks. They are afraid of their feeling getting leaked whereas **46.36%** students sometimes let out their sentiments. **9.09 %** of the group under test are very comfortable in expressing their feelings in social networks.

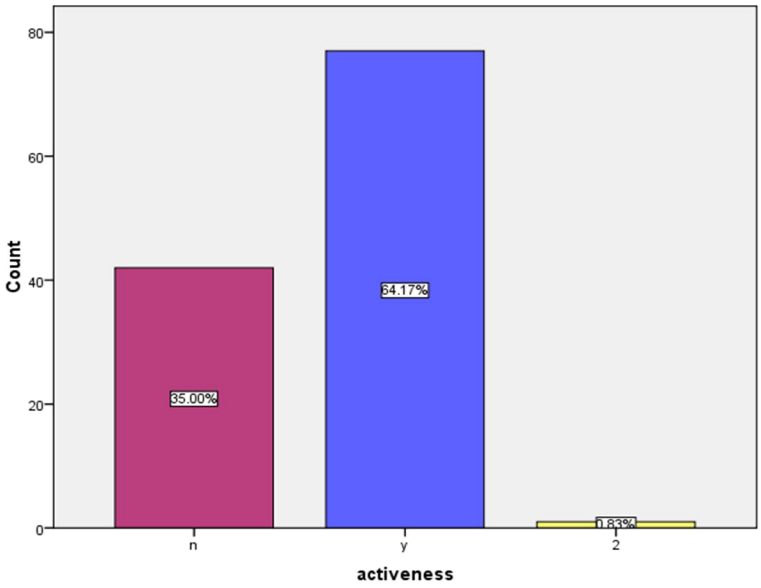
3.4.7. Social networks used by students:



Interpretation:

73.68% students under test prefer facebook over all other social networks.**2.63%** of them use orkut. Surprisingly, **21.05%** students use google+ while **0.88%** of them are active on twitter. The remaining **1.75%** students use other social networks. This indicates that Facebook is the most popular among all the prevailing social networks.

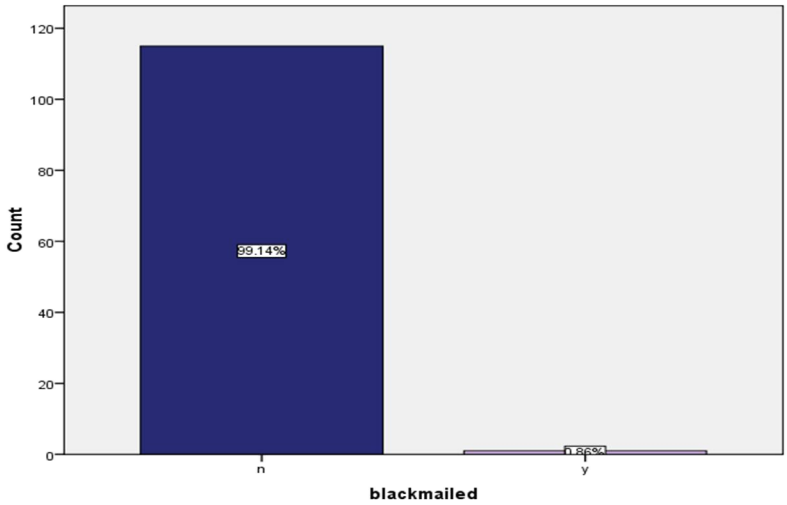
3.4.8. Active user of social networks?



Interpretation:

35% of the students are not very active users of social networks, whereas **64.17%** of them very active in social networks while the remaining **0.83%** did not provide any response.

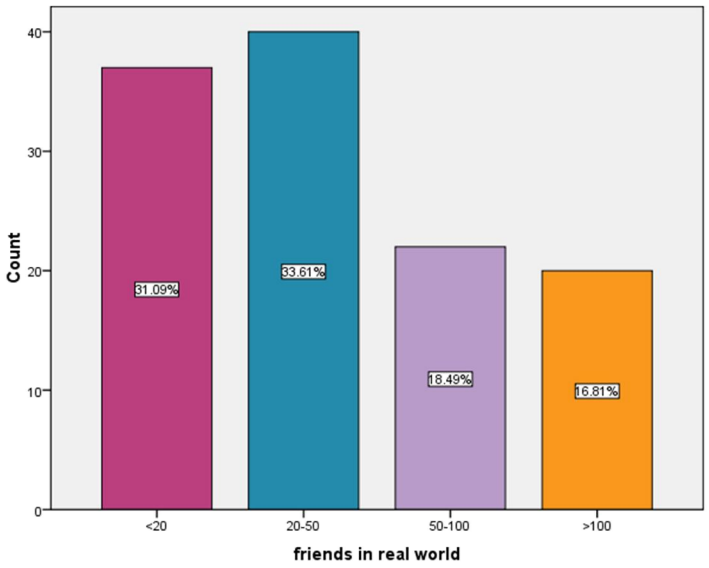
3.4.9. The experience of being blackmailed by any vested-interest group?



Interpretation:

99.14% of the sample of students were never blackmailed by any vested interest group while a very small portion , 0.86% of the sample students say that they have been blackmailed by people with wrong intentions at some point of time in social networks.

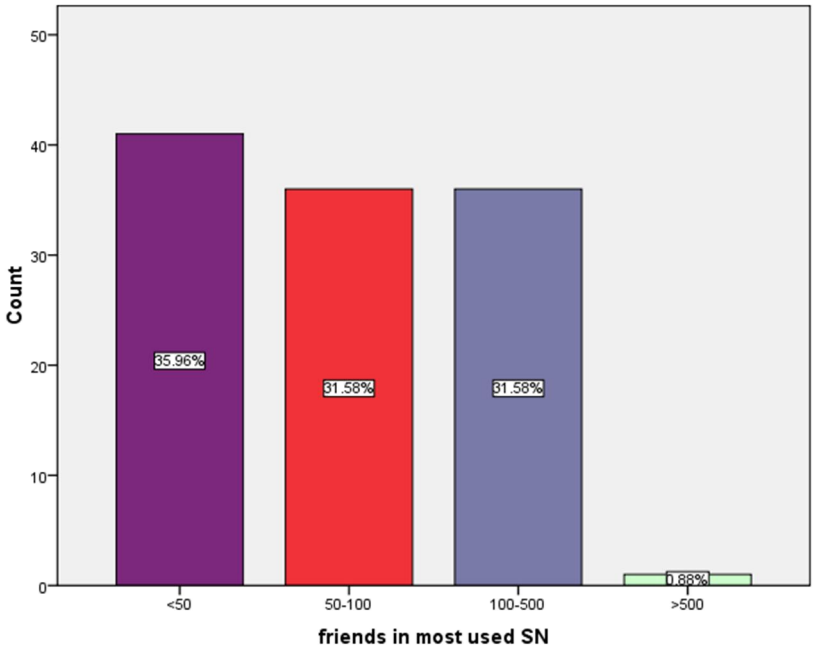
3.4.10. Number of friends in the real world:



Interpretation:

In this analysis, it has been found that **31.09%** students have less than **20** friends in the real world.**33.61%** of the people under test have **20-50** friends in the society, **18.49%** of them have **50-100** friends and **16.81%** of them have more than **100** friends in the real world. This indicates that usually people do not have too many friends in the physical world. Majority of them have very few friends in the real world.

3.4.11. Number of friends in the social network most used:



Interpretation:

We conclude that **35.96%** people have less than **50** friends in the most used social network. **31.58%** people under test have **50-100** and **100-500** friends and **.88%** people have more than **500** friends in social networks.

4. MAJOR CONTRIBUTIONS

Based on the above research, the following contributions in the field of social network data analysis have been made in this project:

- There is a negative correlation between the time spent on social networking sites and the academic performance of a student, which indicates that social networks have some bad impact on the lives of the younger generations of our society.
- There is a positive correlation between the activeness of a student in the real society and his activities in the virtual social networks in terms of the number of acquaintances.
- For ordinal regression models to find the effect of sex, background, medium of study at school and knowledge of computer science on the activity of a student in social networks, it was found that the Cauchit link function was a better than the complementary log-log link function
- The hypothesis that medium of study does not affect the activity of a student in social networks is not rejected
- The hypothesis that knowledge in computer science has no effect on the activity of a student in social networks is also not rejected
- The hypothesis that students trust the privacy policy of social networks is rejected
- The hypothesis that students use social networks for only socializing is rejected. It is very interesting to note that they use them for both academic purposes as well as for friendship and entertainment. However the accuracy of this test can be further checked if we can obtain real data from the social networks themselves.

5. RECOMMENDATIONS AND FUTURE WORK

There are several scopes of improvement in this research work. Due to the fact that the data has been collected by a sample survey, therefore there are cells with missing data. A **missing value analysis** could be done in order to predict the missing responses, find a pattern in the missing data and also obtain the mean, variance, covariance and correlation for the different methods for missing value analysis. This will greatly improve the results, lead to a better regression model and also provide better tests of hypothesis.

One may even introduce a scale component in the regression model so that the variability in the values of the predicted variable is taken in to consideration. It is an additional improvement of the ordinal regression model which yields a model that is more accurate for the given data.

Another factor that needs to be taken into account is the presence of outliers in the data. The Pearson's correlation coefficient gives best result when the data is normally distributed and outliers are absent. Thus, one can identify the outliers by a scatter plot and then eliminate those data from the analysis because they do not have much significance in the analysis.

I would also suggest, that though it is difficult, but one must try to obtain the real data from the social networking sites like facebook and twitter for more accurate results.

REFERENCES

[1] http://www.onlineschools.org/blog/history-of-social-networking/ (as on April, 2012)

[2] Kathleen M. Carley, et al., 2008, Behavioral Modeling and Simulation, From Individuals to Societies, Greg L. Zacharias, Jean MacMillan and Susan B. Van Hemel, editors, National Academies Press, Washington, DC

[3] Detecting changes in a dynamic social network, Ian Mc Culloh, Ph.D thesis ,March 31st , 2009, ISR,CMU,USA.

[4] Katz, L. and Proctor, C.H. (1959). The configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika,* 24, 317-327.

[5] Frank, O. (1991). Statistical analysis of change in networks, *Statistica Neerlandica* **45** (1991), 283–293.

[6] Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

[7] Carley, Kathleen & Behrens, Dean. (1999). Organizational and Individual Decision Making. *Handbook of Systems Engineering and Management*, 659-684. New York: John Wiley & Sons, Inc.. [abstract] [link] 2731kb

[8] Carley, Kathleen. (1999). Organizational Change and the Digital Economy: A Computational Organization Science Perspective. Cambridge, MA: MIT Press. [abstract] [link] 1532kb

[9] Sanil, A., Banks, D., and Carley, K.M. (1995). Models for evolving fixed node networks: Model fitting and model testing. *Social Networks,* 17, 1995.

[10] Feld, S. (1997). Structural embeddedness and stability of interpersonal relations. *Social Networks,* 19, 91-95.

[11] Statistical software –R

[12] Tutorial of SPSS

Analysis of Dynamic Social Networks and Their Effect on the Academic Performance of the Undergraduate Students of BHU.

2011-2012

Name:
Sex:
Age group: 1) 15-20 2) 20-25
3) 25-30 4)>30

1) What was your medium of study at school?
   i) English ii) Hindi
   iii) Other

2) What kind of a background do you hail from?
   i) Rural ii) Urban

3) What was your SGPA (out of 10) in your last major examination?
   i) >9 ii) 8-9
   iii) 7-8 iv) 6-7
   v) 5-6 vi) <5

4) Are you aware of the term "**Social Networking**"?
   i) Yes ii) No

5) Are you an active user of social networks?
   i) Yes ii) No

6) Which of the following social networks do you use the most?
   i) Facebook ii) Orkut
   iii) Google + iv) Twitter
   v) Other

7) About how many friends do you have in the **physical world**?
   i) <20 ii)20-50
   iii) 50-100 iv) >100

8) About how many friends do you have in the **social network** that you use the most?
   i) <50 ii) 50-100
   ii) 100-500 iv)>500

9) Did you ever undergo any professional course on computer science?
   i) Yes ii) No

10) How frequently do you use social networks?
   i) Every day ii) Once or twice in a week
   iii) Once in a fortnight iv) Once in a month or even less than that.

11) If you are a daily user of social networks, then how many hours do you spend on them (If not, you may skip this question)?
   i) <2 hrs ii) 2-5 hrs
   iii) >5 hours

12) Since when have you been using social networks?
   i) 9^{th} or 10^{th} grade ii) 11^{th} or 12^{th} grade
   iii) After I came to college

13) Why do you use social networks?

i) To stay in touch with old friends and make new ones

ii) To update myself with the recent global events both in academics as well as in other fields

iii) Both the above

iv)Other(Please mention):
……
……..

14) Do you have friends in social networks whom you are not acquainted with?
i) Yes ii) No

15) Do you trust the privacy policy of the social networks you use?
i) Yes ii) No

16) Has your social network account ever been hacked?
i) Yes ii) No

17) Have you ever been blackmailed by any vested-interest group in a social network?
i) Yes ii) No
If yes, please give a brief description: ……
……

18) Have social networks reduced you interaction in the real world to some extent?
i) Yes ii) No
iii) Can't say

19) Do you use social networks as a vent to let out your anger or frustration regarding various social issues?
i) No, I don't think it safe. ii) Yes, sometimes.
iii) Yes, quite often.

Thank you for your time!

Investigator's Signature

Respondent's signature

Date: