IN4MATX 205 Project Report / Sushil Bajracharya / Spring 2007

Abstract

In this project report we perform several statistical tests and analysis to look into any interesting relation (if exists) between the political preferences and the use of Web 2.0 technology.

Organization of report

The report is divided into 4 sections and various sub-sections. In each section we look into some selected variables related to Web2.0 technology or political preference. We discuss all recodings and introduction of new variables before discussing a test. After setting this background and giving some exploratory observation, if any, individual tests are presented in the following format fulfilling the requirements for the project:

H0: <null hypothesis>
H1: <alternate hypothesis>

< Test Description and results >

< Conclusions >

Assumptions / Checks:
< Assumptions and checks/additional tests >

The assignment required 10 tests (5 distinct and 5 mixed). These 10 tests are conducted throughout various sections in the report. The hypotheses are stated before any test details. Test results and conclusions make up the main body for discussion inside the vertical bars, and assumptions and further tests performed are discussed after the main description. Plots are included where necessary.

Assumptions and Tests

Here are the list of assumptions held for various tests and how we check for it.

- For t-test we assume normality in data and no outliers in the group
  - Non-parametric tests are used as required
- For ANOVA we assume normality, independence, and equal variance
  - The assumption for independence always holds as we are looking at responses from different individuals. We are making the assumption that same person did not completed the survey twice.
- For linear regression we check on the normality in the residual plots and look at the plot of the fitted residuals
- For chi-square tests if the table is greater than 2X2 we check if the expected count is greater than 1 and if 80% of the expected counts is at least 5.

Depending on an individual test, some alternatives were tried out that are discussed in further detail in the report later.

Section1: Blog Readership and Political Awareness

The variable related to the major category of blogs read by the respondents (*MainBlogCategoryREAD*) is interesting in particular as it is the only variable where a use of Web 2.0 technology (blogs) shows up in connection to politics. The question asks "What is the main category of the blog(s) you read ?". The response includes "Political" blogs as one of the categories.

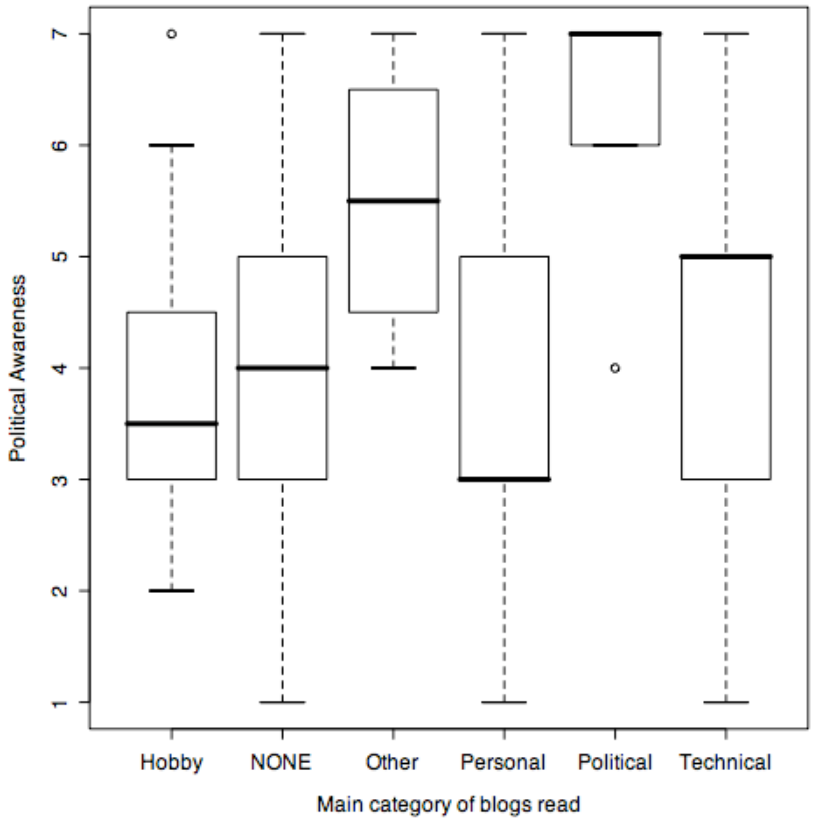

Figure 1 Boxplot for political awareness across different major topics of blog read

We start our observations by looking into some of the plots for these variables. Figure 1 shows the box plot of *PoliticalAwareness* grouped by major categories of the blogs read. Some major differences are easy to spot. Those who rate themselves politically aware seem to be chiefly reading political blogs. While those responding to be less politically aware are mainly reading personal blogs. The relation between blog readership and political awareness is interesting, as it might indicate whether more politically aware respondents read political blogs or not. Or, whether political awareness has nothing to do with respondents reading political blogs. We apply ANOVA test on the survey data to further investigate statistical significance of this relationship.

***H_0**: Political Awareness is same for all categories of respondents reading various major category of blogs.*
***H_1**: Political awareness varies at least between two groups of different main category of blogs read.*

ANOVA test on political awareness grouped by main category of blog read gives a p-value of less than 0.001 which shows a strong evidence against the null hypothesis that there is no difference in mean political awareness among these groups.

Using Tukey's multiple pairwise comparison between the groups we get a noticeable difference between means among few groups:

	Difference	*Confidence Interval*
Political-Hobby	2.5512821	[0.69, 4.41]
Political-NONE	2.4063545	[0.94, 3.87]
Political-Personal	2.6529081	[1.17, 4.13]

This difference is much clearer in the confidence interval plots (Figure 2) of the various groups.

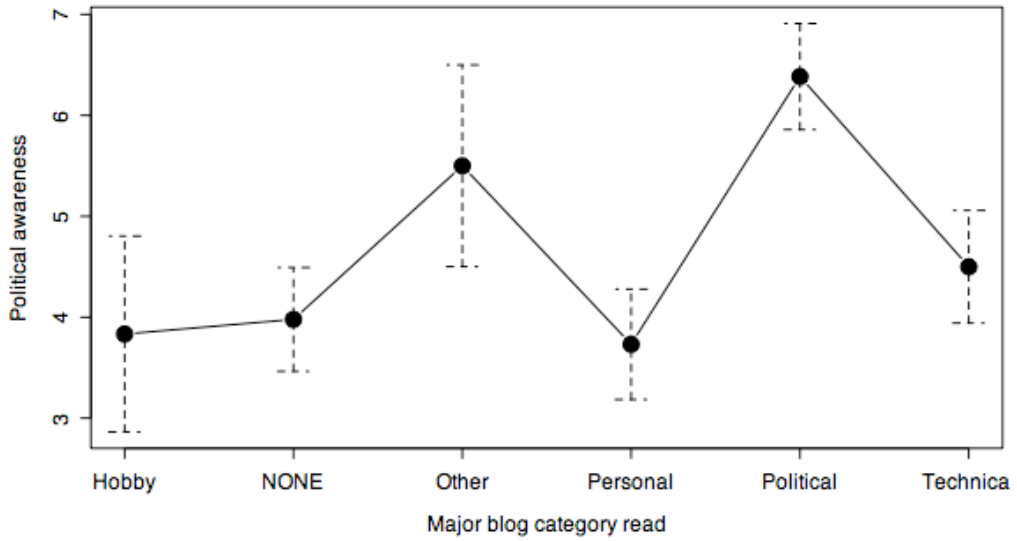

Figure 2 Plots showing confidence intervals and differnce in means

The conclusion from this test is that there is strong evidence that politically aware respondents are very likely to mainly read political blogs.

Assumptions:

The boxplot here shows some outliers and lack of normality in some groups. We performed Shapiro test for test of normality and obtained a p-value less than .001 indicating strong evidence of lack or normality in the data.

Trying to transform the data by taking square root of *PoliticalAwareness* got rid of the outlier in the 'Hobby' category of blog read.

Looking at the original data showed that *PoliticalAwareness* in the category of 'Political" blog read is very high most of the times. Removing the outlier in that group and performing ANOVA with the square root value for *PoliticalAwareness* still gave similar results to the original; ie low p-value (< 0.001)

A non-parametric test, Kruskal-Wallis test on the variables gives a Kruskal-Wallis $\chi^2_{df=5} = 29.7$ with a p-value of less than 0.001. This supports the conclusions we made based on ANOVA that there is some evidence for difference in political awareness among various major category of blogs read by respondents.

Section 2: Blog Readership and Liberal-Conservative Rating

We repeat the similar test as in Section 1 using the same categorical variable of main blog topic read but now against the ratings respondents gave themselves for being liberal or conservative.

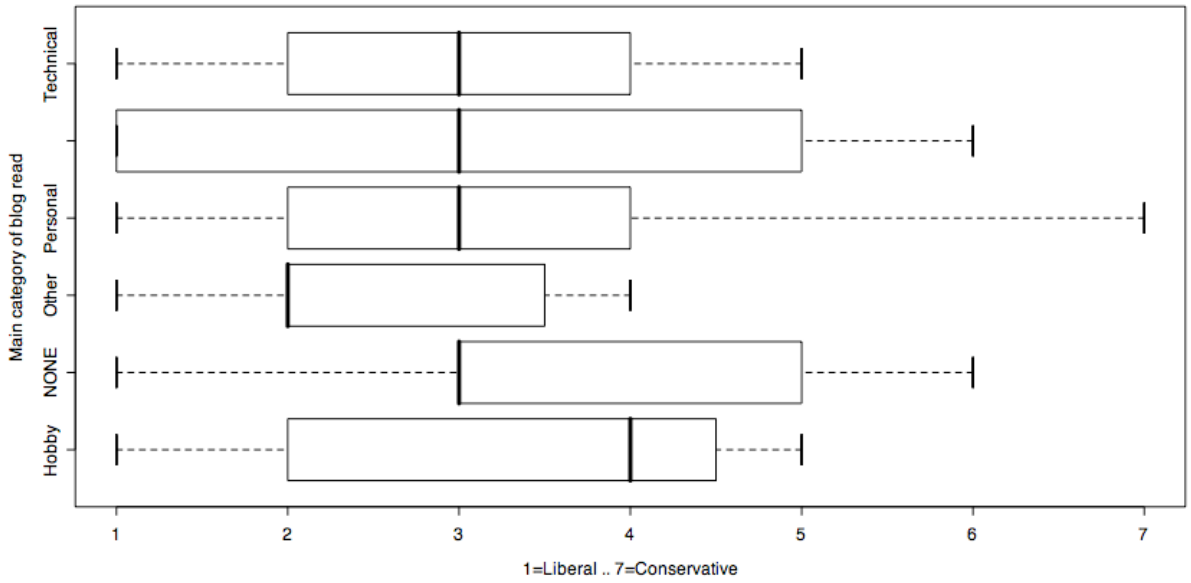

Figure 3 Liberal-Conservative plot for respondents reading various main category of blogs

The box plot of *LiberalConservative* rating grouped by main category of blog read (Figure 3) did not show as much distinct variation in the means of different groups. So we recode the Liberal-Conservative rating as follows:

LiberalConservative = "Liberal", if orignal rating in the survey = 1,2,3
"Neutral", if orignal rating in the survey = 3
"Conservative", if orignal rating in the survey = 4,5,6

H₀: The main category of blog read is independent of the Liberal-Conservative rating
H₁: The main category of the blog read is dependent on how Liberal or Conservative a respondent rate herself

We perform a chi-square test now to test the independence between the two groups we are comparing: *MainBlogCategoryREAD* and *LiberalConservative*. We get $\chi^2_{df=10} = 10.15$ with p-value = 0.43.

This shows that we do not have a good evidence that Liberal-conservative ratings distribution are different for various category of blog readers.

Figure 4 shows that the distribution from the cross tablulation of *LiberalConservative* with *MainBLogCategoryREAD*. Although the Chi square test did not give any significant result, it is interesting to note that there are no "neutral" rated respondents among those who mainly read political blogs. And among those who mainly read political blogs, there seem to be more Liberals than Conservatives.

Checks: 38% of expected counts are less than 5 and all expected counts are greater than zero.

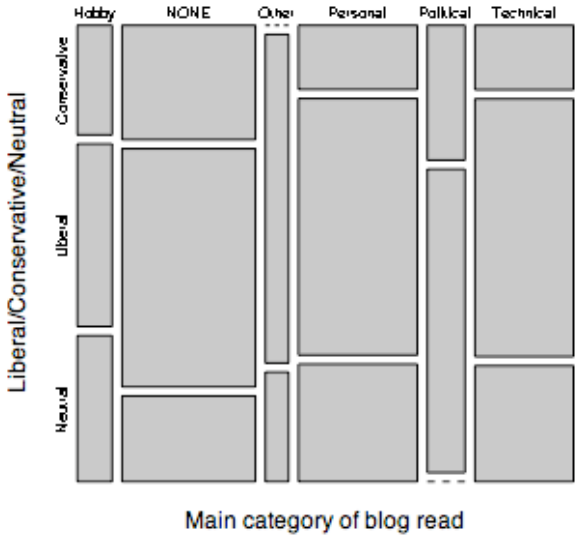

Figure 4 Mosiac plot for contingency table of (LiberalConservative X MainBlogCategoryREAD)

Section 3: Web 2.0 Technologies and Political Preferences/Characteristics

In this section we'll look into each of the major Web2.0 technologies that was considered in the survey and see if we can discern any relationship between them, in particular between various technologies and political preferences. We'll also look into if there is any interesting correspondence among the various technologies.

Section 3.1 Blogging Habits (Mostly Single Person created but publicly shared content)

We consider three blogging related questions that were asked in the survey; degree of involvement with blogs (*BlogInvolvement*), number of blogs owned (*BlogsOwned*), and the frequency of blogging (*BloggingFrequency*).

The variables are recoded as follows:

BlogInvolvement (BI) = 0 if, "Not Involved"
1 if, "write comments only" or "read only"
2 if, "All of blogging activity - Create, Manage, Post"

BlogsOwned (BO) = 0..4 for owning 0..4 blogs
(ie; value of 0 for recorded response of 0; 1 for 1 and so on), and
5 for "5 or more" blogs

BloggingFrequency (BF) = 5 if, "Several times a day"
4 if, "Daily"
3 if, "Weekly"
2 if, "Monthly"
1 if, "Not at all"
(Please refer to the appendix for further rationale behind this recoding)

Section 3.1.1 Blog Involvement and Political Awareness

We perform a chi-square test on the respondents using two variables: *BlogInvolvement* and *PoliticalAwarenessBinary*. The latter divides the respondents into two groups with the following recoding:

PoliticalAwarenessBinary = "AWARE" if score for *PoliticalAwareness* = 5-7 (from survey)
"NotAWARE" if score for *PoliticalAwareness* = 1-3
(So, this comparison discards the respondents with a response of 4 for *PoliticalAwareness* considering them "NEUTRAL" in terms of political awareness)

H₀: The degree of involvement in blogs is independent of political awareness
H₁: Blog Involvement is not independent of Political Awareness

The chi-square test for *BlogInvolvement* with *PoliticalAwarenessBinary* gave; $\chi^2_{df=2} = 3.46$ *with p-value* = 0.17. This shows no evidence against the null hypothesis of the two groups being independent.

However, Chi-square test using the original 5 categories of response for involvement with blogs with *PoliticalAwarenessBinary* gave us; $\chi^2_{df=4} = 7.0381$, p-value = 0.13. This still has no strong evidence against the independence of categories here.

Performing the chi-square test with variables *BlogInvolvement* and *PoliticalAwarenessThree* (in this case we include respondents with *PoliticalAwareness* = 4) gave us; $\chi^2_{df=4} = 7.19$ with p-value = 0.12. A mosaic plot for the contingency tables from this cross tabulation is shown in Figure 5.

These three chi-square tests lead to the conclusion that the degree of involvement with blogs is independent of political awareness among the respondents.

Check: All expected counts greater than zero and all expected counts greater than 5

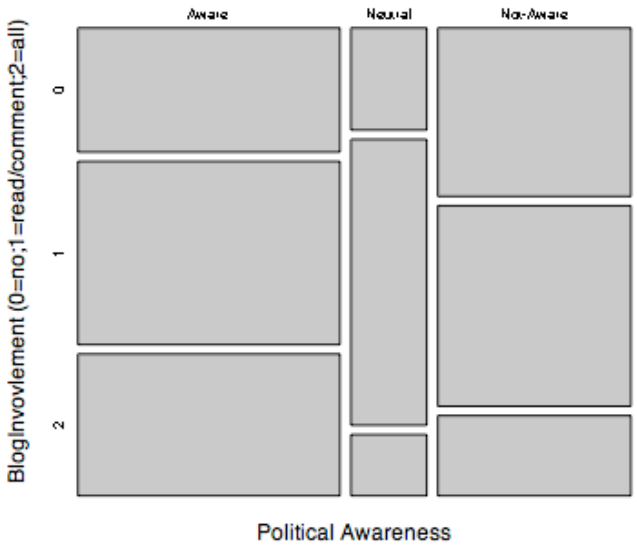

Figure 5 Mosiac Plot for contigency tables of (BlogInvolvement X PoliticalAwarenessThree)

The rest of the variables (BO, BF) did not gave any interesting results either when put together with political awareness.

Section 3.1.2 Blog Involvement and Liberal-Conservative Rating

A boxplot of Liberal-Conservative rating for three categories of blog involvement (*BlogInvolvement*) showed some possible variation in the means of the Liberal-Conservative rating. Those who are extremely involved in blogs ("All" in Figure 6) seem to have lower rating meaning (very liberal) compared to those who are not involved or involved to a less degree ("comments/read only").

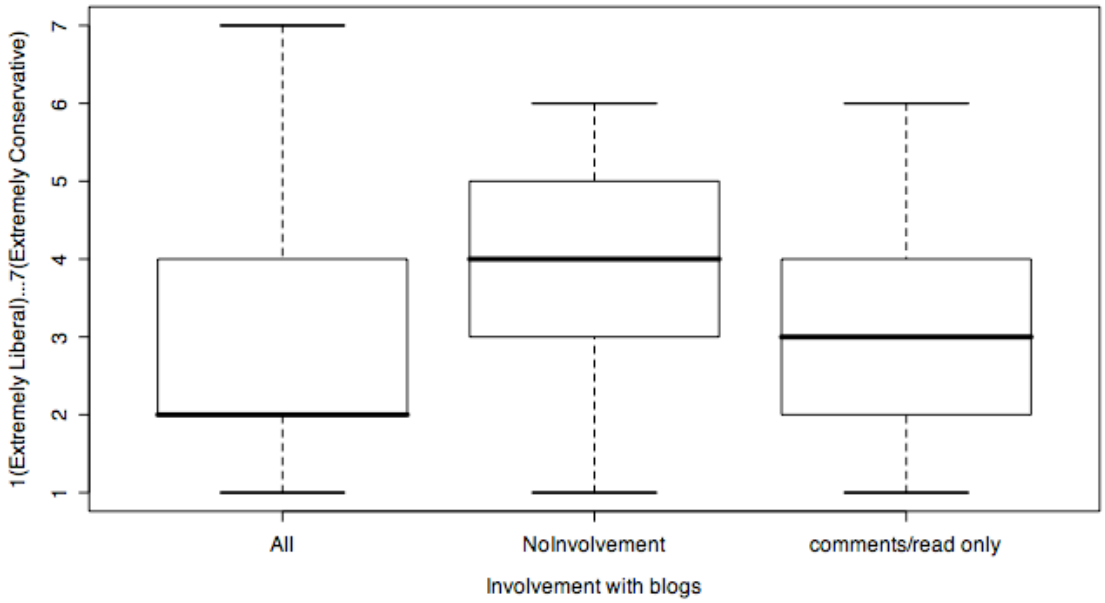

Figure 6 Liberal-Conservative rating for three categories of blog involvement from responents

H_0: There is no significant variation in the Liberal-Conservative rating across various groups of respondents with different degrees of involvement with blogs
H_1: There is a significant variation in the Liberal-Conservative rating of respondents between, at least, two groups with different degrees if involvement with blogs

An ANOVA test of Liberal-Conservative rating against the *BlogInvolvement* category gave a p-value of 0.03.

This provides strong evidence against the null hypothesis that the mean ratings among the groups are same.

Tukey's multiple pairwise comparison shows that the a small noticeable difference of 0.69 exists between those who are not involved with blogs at all, and, those who just comment or read. Other differences had the value 0 in the confidence interval.

This leads to the conclusion that there is a small noticeable difference between the Liberal-Conservative rating between two groups: those who are not involved in blogging at all Vs those who just comment or read only.

Assumptions/Checks:
Shapiro test for normality for the original score given to the Liberal-Conservative rating gave a very low p-value (< 0.001) showing a strong evidence for lack of normality.

Levene's test for homogeneity of variave gave F(df=2) = 1.94 with p-value of 0.15. This provides no evidence against the lack of homogeneity in the variance.

A q-q plot for normality however showed some indication of linearity with outliers at the extreme ends. (Figure 7)

Kruskal-walllis test gave a Kruskal-Wallis $\chi^2_{df=2}$ = 6.64 with p-value 0.03 providing a strong evidence against the null hypothesis.

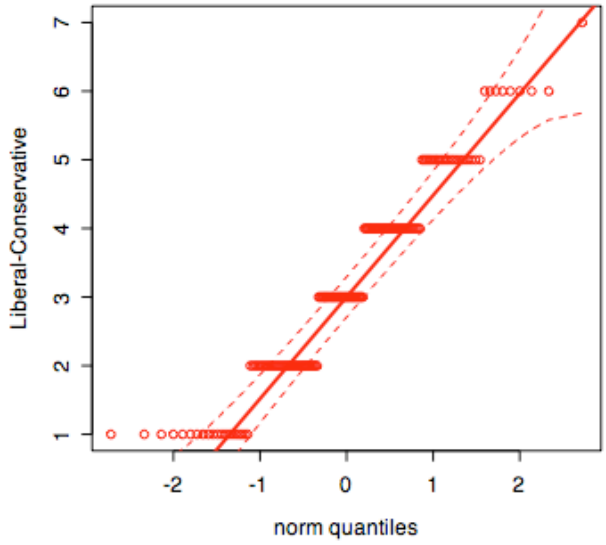

Figure 7 QQ plot for Liberal-Conservative ratings (1-7)

Section 3.1.3 Blogging Habit and Party Preference

The nature of the scores for variables BI, BO, BF makes it difficult to treat them as continuous variables when taken in isolation so we put them together in the following expression to compute a overall score of *BloggingHabit* (*BH*) for each respondent;

BloggingHabit = SQUARE_ROOT*(BlogInvolvement + BlogsOwned) * BloggingFrequency*

BH = SQUARE_ROOT*((BI + BO) * BF)*

(A box-plot of values for "(BI + BO) * BF" showed couple of outliers, taking a square root of the expression got rid of those outliers. Thus, the square root of the value of the expression is used instead of the actual value from the expression.)

The intuition behind the calculation of BH is to assess one's blogging habit: How much is one involved with blogs and how frequently ?.

We would like to see how this score of *BH* varies among three groups of political party preference. For party preference we use a categorical variable *PartyPreference* that had the original responses for political party preference from the survey. A box plot of *BloggingHabit* (*BH*) scores among various preferences for political party is shown below (Figure 8).

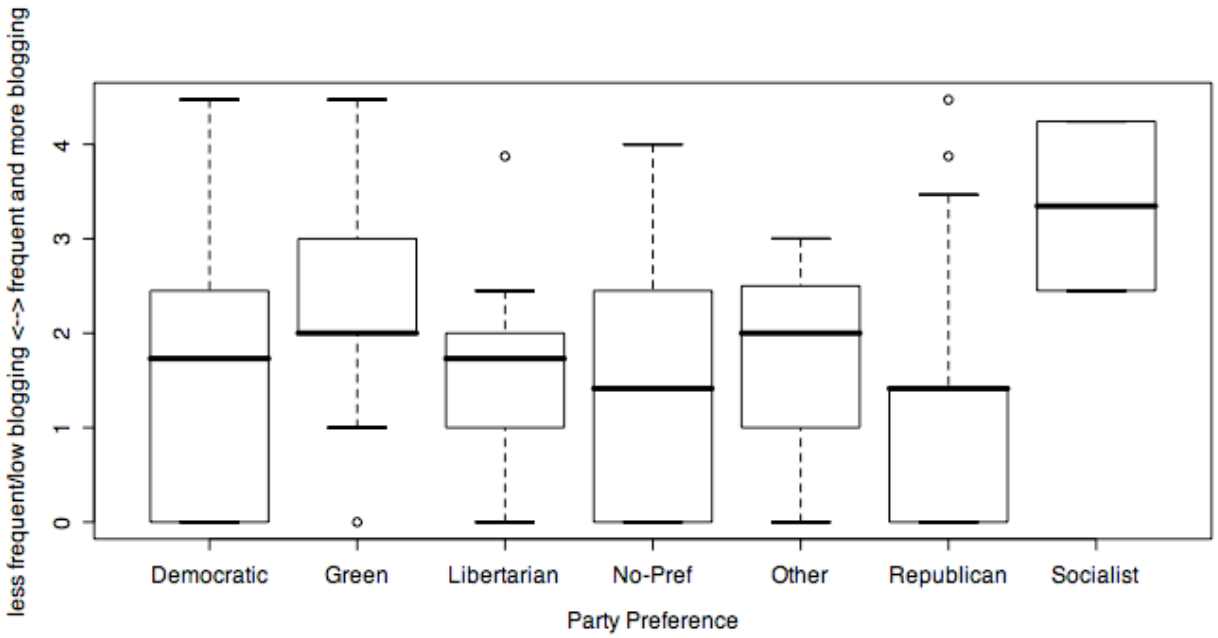

Figure 8 Box Plot for different scores of BloggingHabits (low value = low profile blogger; high value = high profile blogger) across different party preferences

H_0: Blogging habit do not vary among respondents with different party preferences H_1: Blogging habit vary at least between two groups of respondents with differing party preferences

An ANOVA test in this case gave a p-value of 0.33 indicating no evidence in the difference in the mean score of blogging habits for respondents grouped by their preference for political party.

Assumptions/Checks:

Figure 8 shows couple of outliers among various groups and some of the distributions are oddly skewed (for eg; 'Green', 'Republican'). The score BH already has a transformation of square root on it, other transformations could not remove the outliers either.

The QQ plot for *BloggingHabit* (Figure 9) shows deviation from normality.

Levene's test for homogeneity in variance gives p-value of 0.8, indicating no evidence against homogeneity of variance among groups.

Kruskal-walis rank sum test on BloggingHabit Vs PartyPreference gives Kruskal-Walis $\chi^2_{df=6} = 6.97$ with p-value = 0.32. This also provides no evidence against the null hypothesis.

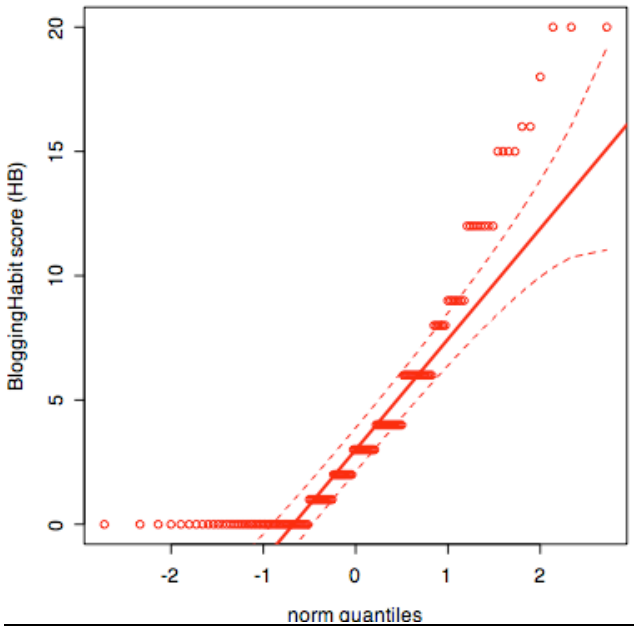

Figure 9 QQ plot for BloggingHabit Score (BH)

Section 3.2 Wiki Involvement (Multi-author public content)

In this section we look into the variables recorded for Wiki Invovement (Public and Private) Wikis and if they have any relationship with the political variables.

The two variables for involvement with public and private Wikis corresponding to the Likert scale values were respectively recoded as follows:

WikiInvolvement{Public|Private} = 0 if, No involvement
1 if, Create new entries Edit/modify entries
1 if, Read only
2 if, All of the above

Section 3.2.1 *WikiInvolvementPublic* Vs *WikiInvolvementPrivate*

We use Chi-square test to test the homogeneity among the two categories: Involvement in private Wikis and Involvement in public Wikis.

H₀: The distribution of involvement in public Wiki is same for every level of involvement in private Wikis (ie the groups are homogenous)
H₁: The distribution of involvement in public Wiki is not same for every level of involvement in private Wikis

The chi-square test gave a value of: $\chi^2_{df=4} = 32.65$ and p-value smaller than .001. This provides strong evidence against the homogeneity of the two categories. This leads to the conclusion that the respondents' involvement in public Wikis is different from that in private Wikis.

Mosaic plot for contingency tables from this test is shown in Figure 10.

Check: 55% of expected counts is smaller than 5, there are no counts smaller than 1.

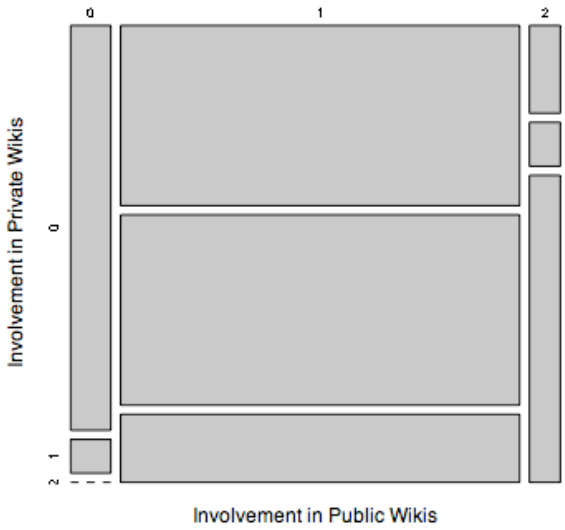

Figure 10 Mosaic plot for contingency table (Involvement in Pubic Wikis X Involvement in Private Wikis)

Section 3.2.1 Public Wiki Involvement

We define a new categorical variable *WikiInvolvementPublicBinary* as to be a binary variable:

WikiInvolvementPublicBinary = "NO" if not involved in public Wikis at all
"YES" if any kind of involvement exists in public Wikis

H₀: Political awareness does not vary among the two groups where one are involved in public Wikis and other are not involved in public Wikis
H₁: Political awareness varies depending on whether respondents are involved in public Wikis or not

We perform a t-test on the variable *PoliticalAwareness* with this new categorical variable *WikiInvolvementPublicBinary*. We get: $t_{df=14.67} = -2.24$ with p-value of 0.04. This provides strong evidence against the null hypothesis that political awareness is same across those who are involved in public Wikis and those who are not involved at all.

Assumptions/Check:
The box plot for *PoliticalAwareness* Vs *WikiInvolvementPublicBinary* show no outlier (Figure 11)

Shapiro test for *PolticalAwareness* shows lack of normality (p-value < 0.001)

Performing the Wilcoxon rank sum test (a non-parametric test) on *ConcernOnIdTheft* with *LibConsBinary* gave W = 596 with p-value = 0.04. This strengthens evidence against the null hypothesis and leads us to the conclusion that there is an evidence for difference in political awareness among those who are involved in public Wikis and those who are not involved at all.

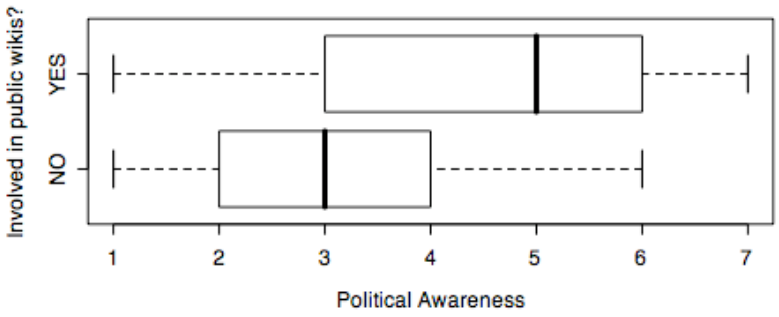

Figure 11 Political Awareness Against Involvement with public Wikis (yes/no)

Section 3.3 [*SN*] Social Networking (Community-based information sharing)

We define a compound score called a "Social Networking Score" as a new variable *SocialNetworking* (*SNet*) computed as follows:

$$SNet = CommunityBasedPopularity + OnlineTagging + SocialNetworking + (FeedbackForProducts * WhenFeedback)/ 2$$
$$= CBP + OT + SN + (FFP * WF)/ 2$$

(The recoding of the variables gives a maximum possible value of 6 for CBP, OT and SN, and a maximum possible value of 12 for (FFP * WF). To make the effects of these variables equal we divide (FFP * WF) by 2. See Appendix for further discussion)

Where the variables to the right correspond to the recoded values from the survey question as follows:

CommunityBasedPopularity recoded values for the responses to: "How often do you participate in community-based popularity websites such as Digg, Reddit, YouTube etc to vote for your favorite entries ?" as:

CommunityBasedPopularity = 0..6 for 1..7 of recorded response value in the Likert Scale

OnlineTagging recoded value for Degree of Involvement in Online Tagging systems as:

OnlineTagging = 0..6 for 1..7 of recorded response value in the Likert Scale

SocialNetworking recoded value of involvement in social networking sites as follows

SocialNetworking = 0..6 = 1..7 of recorded response value in the Likert Scale

FeedbackForProducts recoded the response value for the question: "How often do you provide feedback for the products (and/or sellers of these products) that you purchase online? (Never = 1, Always = 7)" as:

FeedbackForProducts = 0..6 for 1..7 of recorded response value in the Likert Scale

WhenFeedback recoded the response value for the question: "In what scenario do you provide feedback for the products (and/or sellers of these products) that you've purchase online ?"

WhenFeedback = 0, if response is "I never provide Feedback "
1, if response is "Typically provide feedback"
2, if response is "Both satisfied and dissatisfied " or
"Only when I am dissatisfied" or
"Only when I am satisfied"

We do not differentiate among those preferring to write dissatisfied or satisfied feedbacks for online purchases as we consider any form of writing responses as contributing with information. 'Typically provide feedback' is taken as a less confident response, thus is given a lower score than for the responses where the respondents are quite confident about them providing feedback for products (whether satisfied, not-satisfied or both).

Section 3.3.1 *BloggingHabit* Vs *SocialNetworking*

Since both blogging and social networking are key Web2.0 technologies, it would be interesting to see if the scores we have computed as indicators of blogging habits (*BH* – higher value implies strongly inclined towards blogging; from Section 3.1.3) and social networking (*SNet* – higher value implies strongly inclined towards social networking) have any relation at all.

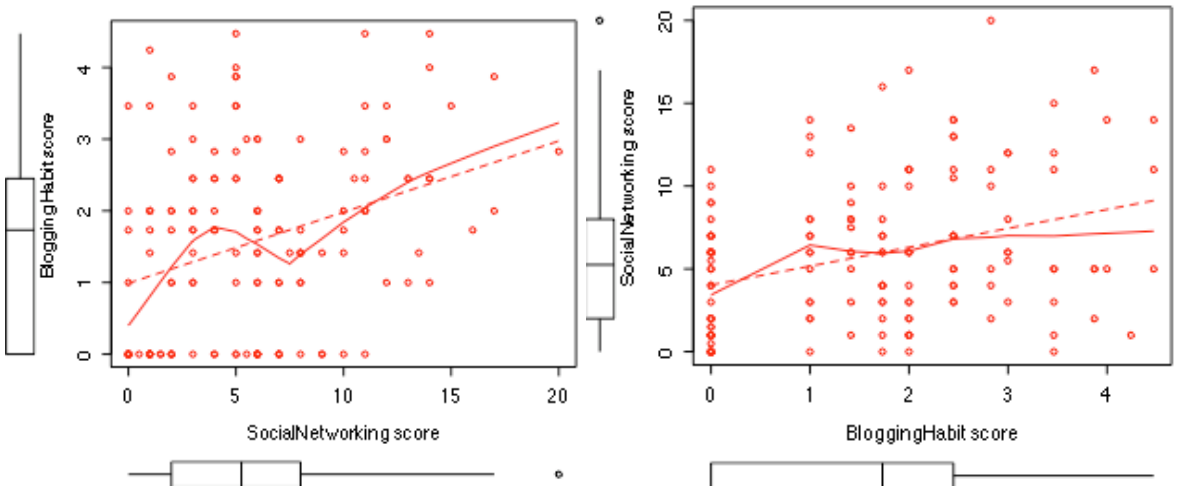

Figure 12 Plots of BloggingHabit score Vs SocialNetworking score and vice-versa

Figure 12 shows the scatterplots for the two scores (*SNet* Vs *BH*, and, *BH* Vs *SNet*). Both plots show a possibility of linear relationship between these two variables.

H₀: There is no linear relationship between blogging habits and social networking score (correlation coefficient = 0)
H₁: correlation coefficient is not equal to 0

Correlation test on these two variables give us $t_{(df=150)} = 4.73$ with p-value of less than .001. The correlation coefficient given is 0.34. This leads us to the conclusion that there is a strong evidence against the null hypothesis of correlation coefficient being zero indicating a possible linear relationship between *BH* and *SNet* scores.

Section 4. Concerns on using Web2.0

Several questions in the survey addressed concerns respondents might have with using Web2.0 technologies. The variables related to these questions are listed below:

Variable	***Question: How concerned are you of..***
ConcernOnContactInfoOnline	..having contact info online
ConcernOnIdTheft	..of identity theft
ConcernOnPicOnline	..of having your picture online
ConcernOnPutPalsInDanger	..of putting your friends/family in danger
ConcernOnRevealIdentity	..on revealing identity
FileSharingLiabilityConcern	..on liabilities of online file sharing

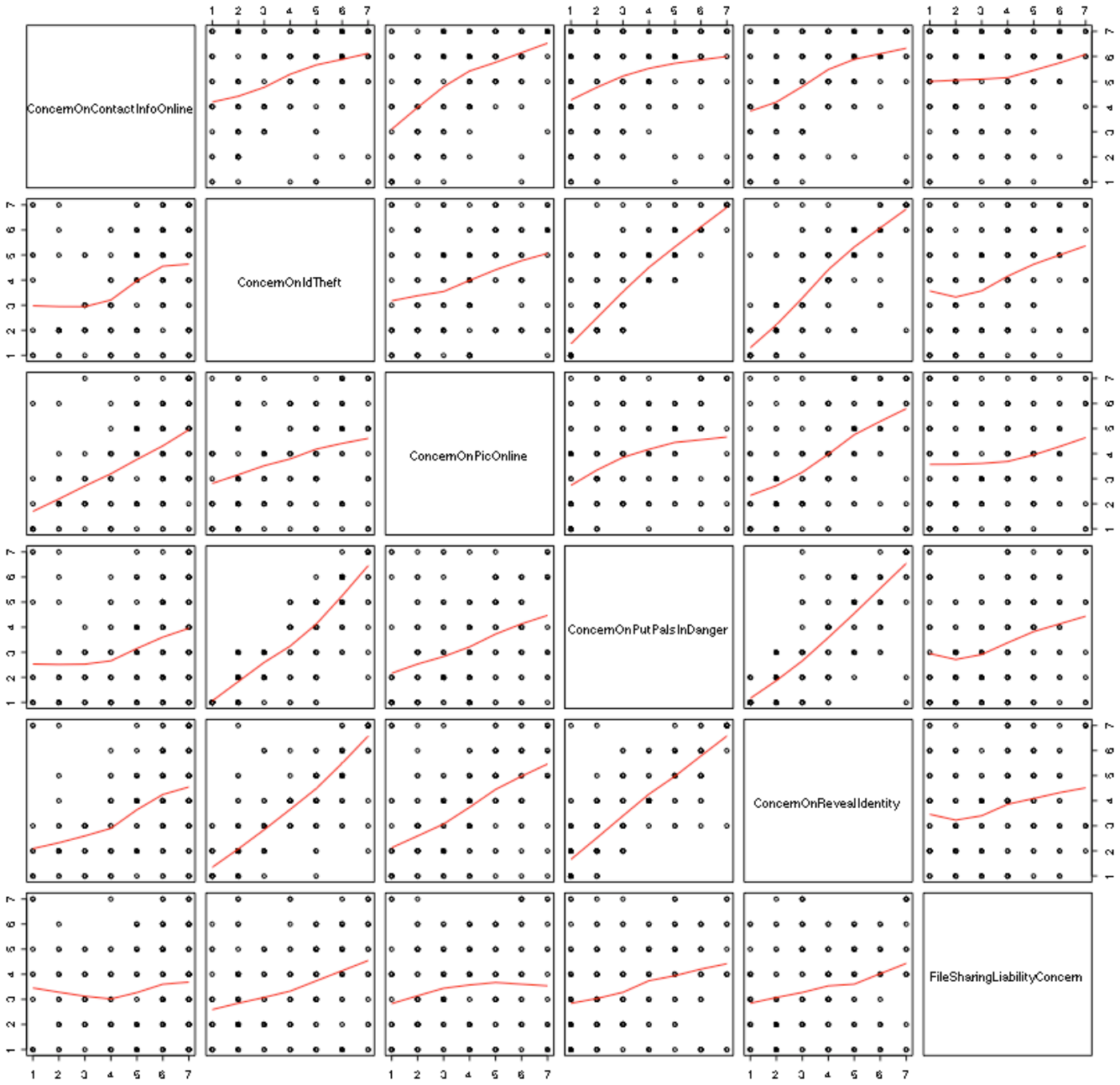

Figure 13 Pairwise plots of vaious concerns that respondents had with Web2.0 technologies

Section 4.1: *ConcernOnRevealIdentity* Vs *ConcernOnPutPalsInDanger*

Figure 13 shows a pair wise scatterplot of all the variables in the Table above. The plot helps us to see if any of these variables are strongly related to another. It can be seen that *ConcernOnRevealIdentity* seems to have a linear relationship with *ConcernOnPutPalsInDanger* and *ConcernOnIdTheft*.

This also follows a quite plausible thinking that revealing identity might expose greater risk of putting loved ones in danger. We'll thus perform a linear regression on the variables (*ConcernOnRevealIdentity* Vs *ConcernOnPutPalsInDanger*) to see how strong the relation is. Figure 14 shows the scatter plot.

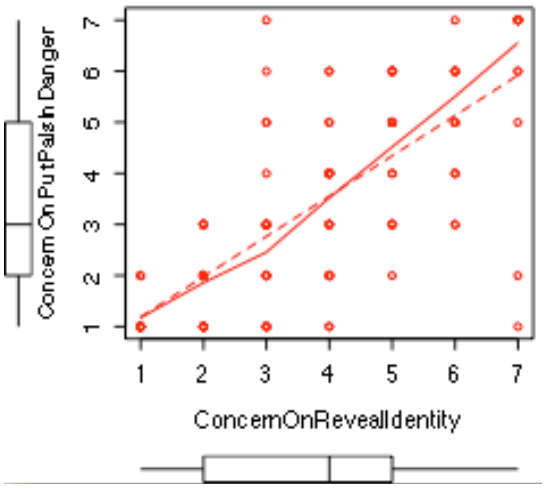

Figure 14 Scatter plot showing relation between concerns on putting loved ones in danger Vs concern on revealing identity

H₀: There is no linear relation between the concern on revealing identity and concern on putting dear ones in danger ie; (slope= 0)
H₁: There is a linear relation between the concern on revealing identity and concern on putting dear ones in danger ie; (slope not equal to 0)

Linear regression gives an intercept of 0.39 with slope 0.79.

p-value for the slope is less then 0.001 (p <2e-16). There is enough evidence against the null hypothesis of slope being zero.

The residuals have median of 0.02 (closer to zero); the min and max values of residuals are nearly equal (|~4|). All these indicate good presence of linear relation. We get more evidence by looking at two diagnostic plots in Figure15.

Assumptions/Checks:
Normal plot of residuals (Figure 15, right plot) show strong presence of normality and the residuals Vs fitted graph (Figure 15, left plot) does not show any strong pattern. This strengthens the evidence of linear relation between the variables:
ConcernOnRevealIdentity Vs *ConcernOnPutPalsInDanger*.

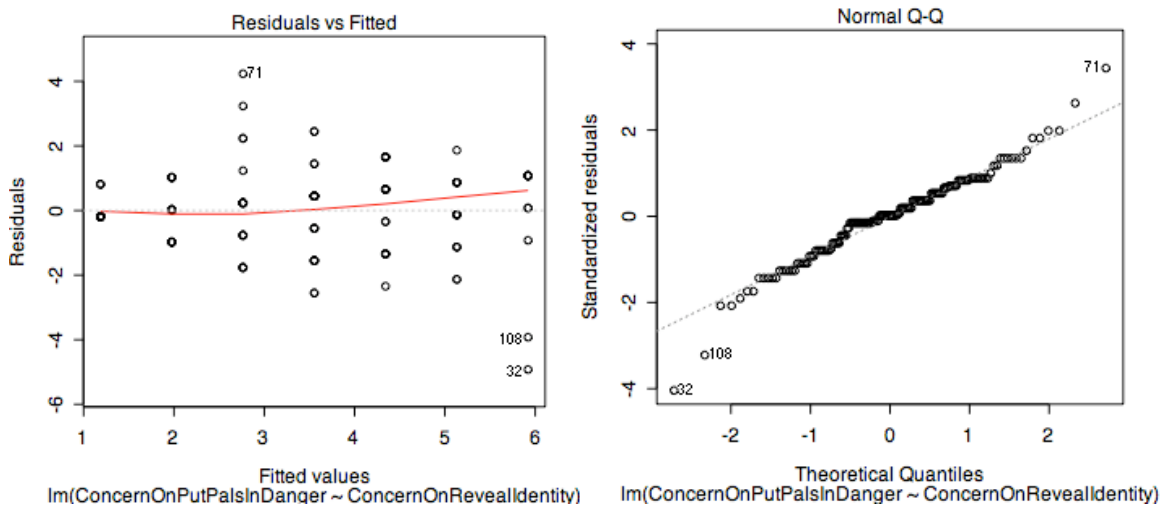

Figure 15 Diagnostic plots for the linear relation between the variables: ConcernOnRevealIdentity Vs ConcernOnPutPalsInDanger

Section 4.2 *ConcernOnIdTheft* and Liberal-Conservative rating (*LibConsBinary*)

Exploring the various concern-related variables with some political preferences among the respondents revealed an interesting variability.

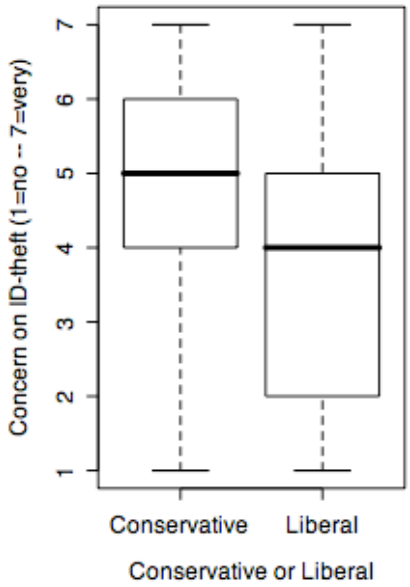

Figure 16 Concerns on Identity theft for Conservative vs Liberals

There seemed to be an interesting variation in the concerns about identity theft among those who were categorized as Conservatives or Liberals (Figure 16). For this we recoded the Liberal-Conservative scale into a categorical variable *ConsLibBinary* as follows:

LibConsBinary = "Liberal" if responded 1, 2, 3 for Liberal-Conservative rating
"Conservative" if responded 5,6,7 for Liberal-Conservative
(Note that we exclude the respondets who answered with score of 4 for this question)

H_0: Concerns on identity theft do not vary between Liberals and Conservatives H_1: Concerns on identity theft will vary between Liberals and Conservatives

We perform a t-test for the value of *ConcernOnIdTheft* with *LibConsBinary* as the binary categorical variable. The test resulted in t(df=45.31) = 2.22 and p-value of 0.03. The confidence interval for difference in means was [0.09, 1.71]. This shows strong evidence against the concerns for ID theft between the Conservative and Liberal groups being same.

Assumptions/Checks:

The box plot (Figure 16) shows no outlier in the data.

The qq plot for *ConcernOnIdTheft* (Figure 17) shows an overall linear plot with outliers at the extreme ends. Shapiro test shows absence of normality in the total data.

Performing the Wilcoxon rank sum test (a non-parametric test) on *ConcernOnIdTheft* with *LibConsBinary* gave W = 1563 with p-value = 0.03. This provided strong evidence against the null hypothesis and leads us to the conclusion that concerns on identity theft will vary between Liberals and Conservatives

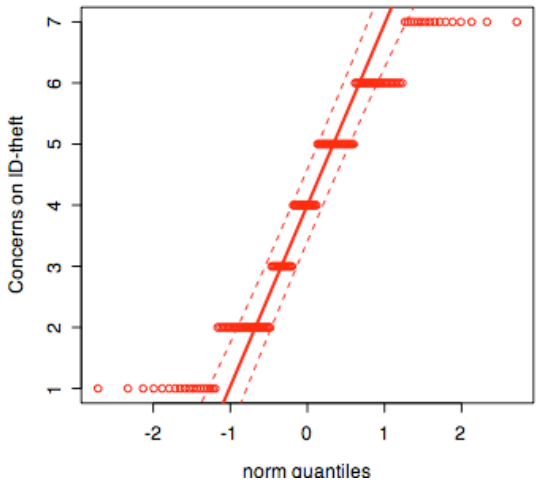

Figure 17 qq plot for Concerns on ID theft

Conclusion

Based on some statistically significant observations we made we can list the general conclusions at follows:

- Politically aware respondents choose political blogs as the main blog category they prefer to read
- Involvement with blogs is independent of political awareness
- Liberals are slightly more inclined to be involved in blogs
- Involvement in public Wiki is independent of involement in private Wikis
- Bloggers are more likely to be into Social networking and/or vice versa
- Conservatives are more likely to be concerned if identity theft

The basic story from the data is the Web2.0 technologies do not strongly categorize political preferences, but liberals and conservatives seem to have some difference in being involved with or concerned about the technology.

APPENDIX

Note on filtering data

Some of the respondents had many fields missing in their response. All the responses that had all information on the Web2.0 specific questions missing were removed from the data-set before the calculations. There were 5 such responses. Two more responses had even more data missing. In total 7 responses were removed for incomplete data.

Note on recoding the variable *BloggingFrequency*:

The variable *BloggingFrequency* seems to be quite ambiguous in the survey. Originally it meant to assess how frequently respondents are involved in any blogging activity (personal blogs or others). In the survey it was presented as; "How frequently do you read, edit, update, and/or comment on your own or others' blogs?". This implies those who respond "not involved in blogs" should have "Not at all" as a response to the question on blogging frequency. This is not the case as seen in the pie-charts below. There are people reading blogs who respond their blogging frequency as "Not at all".

It might be that the respondents interpreted the question of 'Blogging Frequency' as how often are they involved in their personal blog. So they might be reading others' blog but having a zero frequency in being involved in their own. But, this does not seem to be the case either as those who did not own any blog have given none-zero value for "Blogging frequencies".

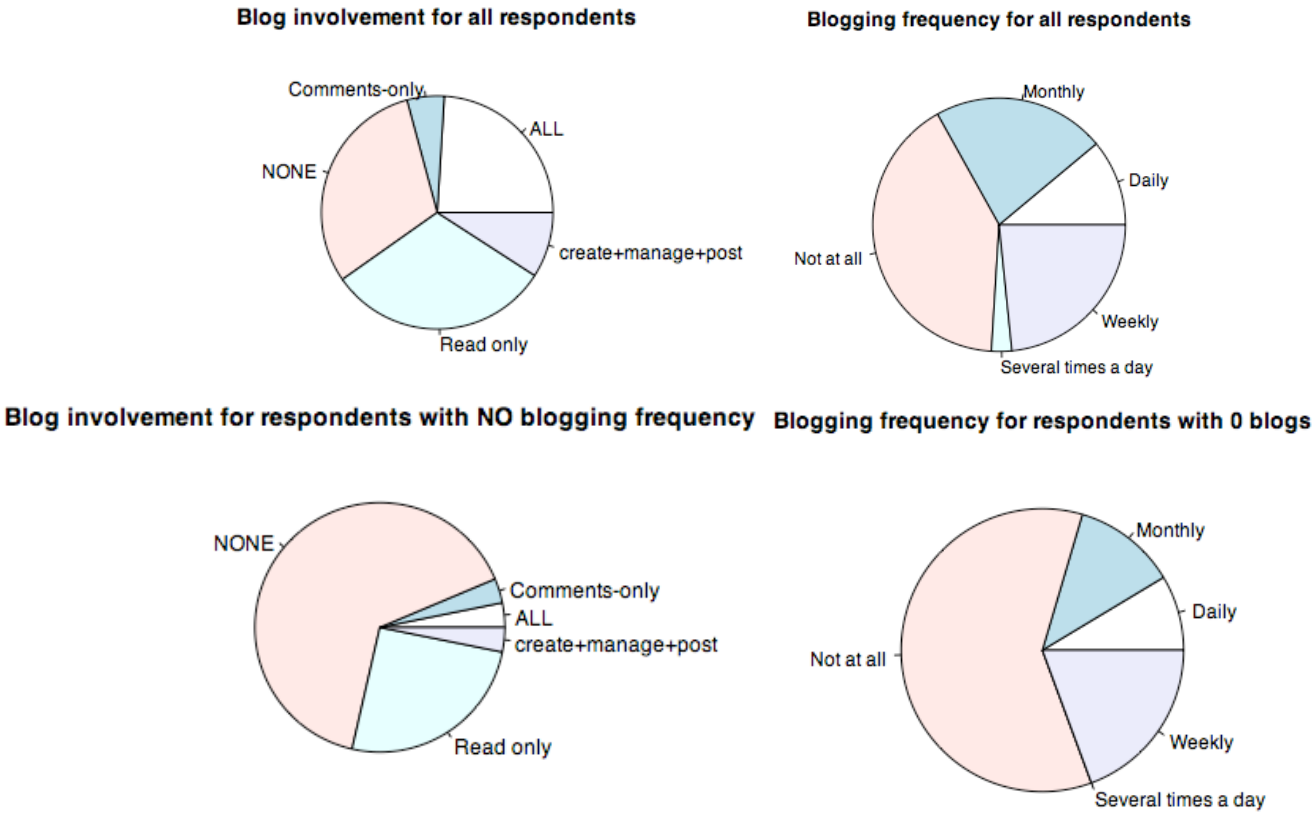

Figure 18 Pie charts showing possible ambiguity in the interpretation of variable BloggingFrequency

Based on this observation we make a slightly relaxed interpretation of what the respondents meant by "Blogging Frequency" and recode the variable with the least value of 1 instead of 0.

The way that BH has been expressed, it will give a score of zero BH to an individual who claims to be "Not Involved in Blogs" and also do not own any blog at all. In fact there were 6 respondents who gave a Weekly or Monthly *BlogFrequency* but claimed "No Involvement with blogs" and had zero blogs that they owned. This gave them a zero score for BH.

Notes on using values from Feedback information that respondents give for online purchases.

1. *FeedBackForProducts* and *WhenFeedback* variables are considered as indicators of one's Social Networking habits as these feedbacks contribute a great deal to some Web2.0 phenomena such as collaborative filtering (eg; reviews in Amazon).

2. The recoding of the variables gives a maximum possible value of 6 for CBP, OT and SN, and a maximum possible value of 12 for (FFP * WF). To make the effects of these variables equal we divide (FFP * WF) by 2

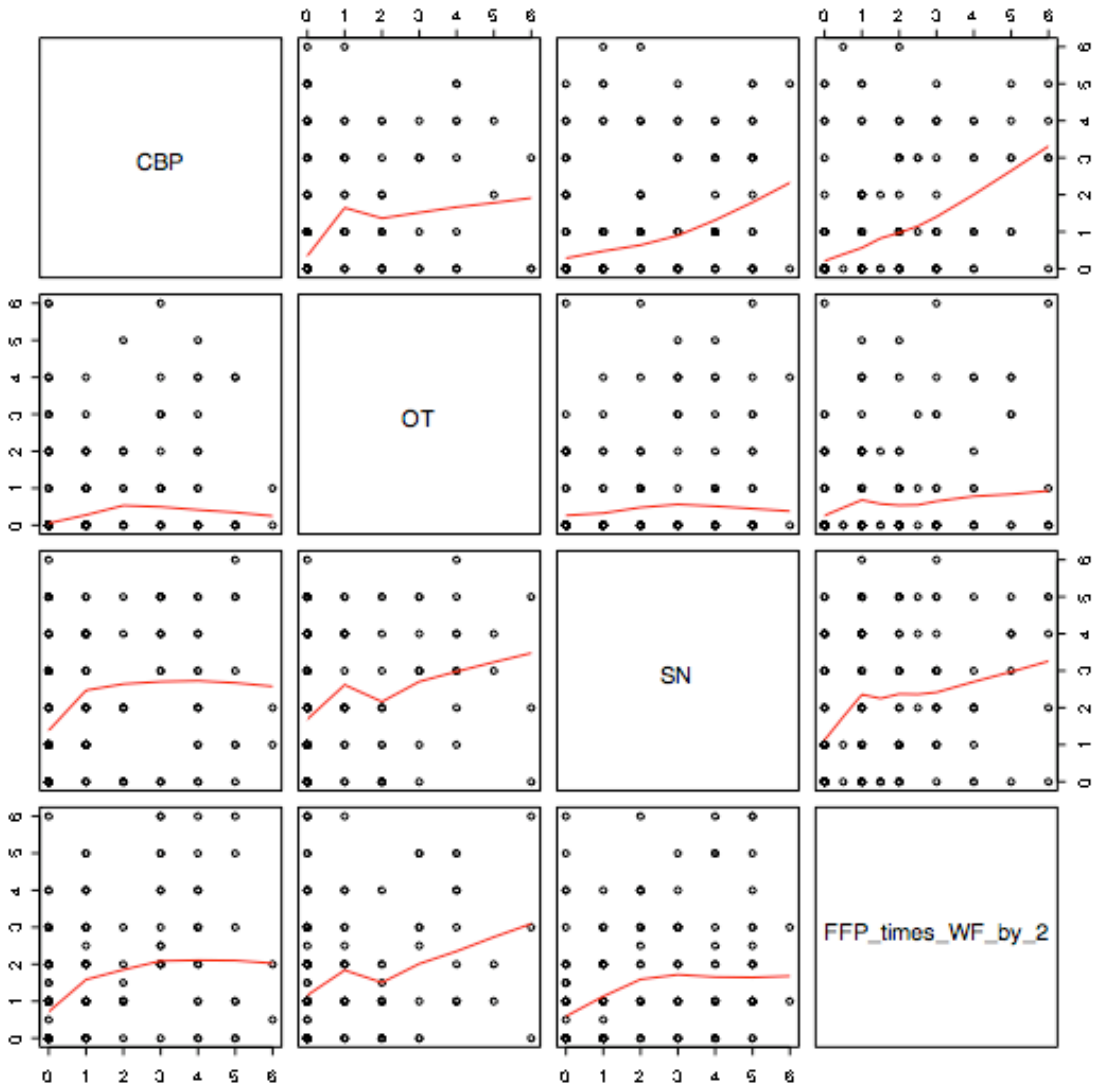

Figure 19 Pairwise scatter plot of variables used in calculating SNet (Social Network Score)

3. Looking at the pairwise plot of all the variables used in calculating the "Social Network Score" (*SNet*) for all the respondents reveals that most of the times these variables are positively correlated with each other. So, the formulation of *SNet* as the sum of these four variables is sensible.

4. These two variables suffer from the similar ambiguity as we had in the case of *BloggingFrequency* variable. Six people responded saying that they "Never' provide feedback for the *FeedbackForProducts* and yet respond with other responses than "I never provide feedback" for *WhenFeedback*. There seems to be some logical inconsistency here. Nine respondents, on the other hand responded "I never provide feedback" for *WhenFeedback* but responded various values other than 1 (ranging from 2 – 6), for *FeedbackForProducts*. This is more visible in the cross tabulation of these two variables as shown below.

	WhenFeedback				
FeedbackForProducts	***Both***	***Never***	***dissatisfied***	***satisfied***	***Typically***
1 (Never)	1	35	4	1	0
2	20	5	11	4	2
3	13	0	7	0	0
4	11	3	2	2	3
5	5	0	2	2	2
6	5	1	0	1	3
7 (Always)	4	0	1	0	2

The product formulation (*FeedbackForProducts* * *WhenFeedback*) will give a score of zero to any such user who seem to be inconsistent with the two variables *FeedbackForProducts* and *WhenFeedback*.