FIT3152 Data Analytics
Assignment 1

Haw Xiao Ying || 29797918

Table of Contents

Question (a)

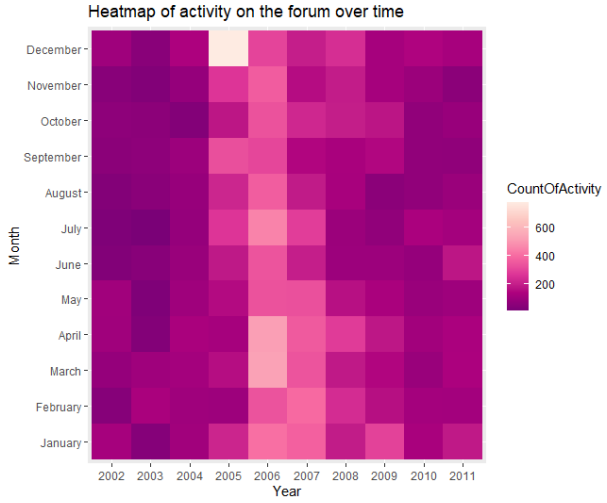

Diagram 1: Heatmap of activity on the forum over time

In Diagram 1, a brighter colour indicates the number of activities is high whereas a darker colour indicates the activeness of participants is low. According to Diagram 1, during year 2005, the activity started to increase as we can see that year 2005 had more counts of activity compared to previous year because it has a brighter column. The activeness of participant is also increasing gradually in year 2005 from January to December and had a huge increase in December 2005, which has the highest number of activities from 2002 to 2011. During year 2006, the forum has the most active participants since it has the brightest colour among all of the year. In year 2007, the participants became less active as we can see the heatmap shows that the colour became gradually dimmer from January 2007 to December 2007 which indicates less activities. After that, the activeness of the participants decreases. In conclusion, initially, there was not much activity on the forum, then it started to increase around year 2005 and is most active in year 2006. After that, the activity on the forum started to go downhill again. For the analysis of language on the forum over time, by looking at Diagram 2 which includes changes of Analytic, Clout, Authentic and Tone over time, we can observe from each of the heatmaps that they do not have a specific trend like Diagram 1 which we can clearly tell there is an increase in 2005 and decrease in 2007, the colours are quite random for each variable in Diagram 2. Other linguistic variables which are less important is not in Diagram 2 but can be seen when running the code in appendix. There are months where the variables suddenly became very high or very low. For example, during 2003, the linguistic variables varied a lot. Therefore, they did not have a specific trend but keep changing. Although there are some overlapping on the high or low percentage of different linguistic variables, the patterns of each heatmaps does not seem to be similar. By having a statistical correlation test which is shown in Diagram 3, the highlighted results are results that shown either weak or strong linear relationships of the variables. There are positive relationships which indicate both of the variables move in tandem whereas negative relationships indicate the variables change in opposing directions. The higher the value in Diagram 3, the stronger the relationship. To conclude, there are weak relationships between Clout and Authentic, Analytic and ‘i’, Clout and ‘i’, Authentic and ‘i’, Clout and ‘we’, Clout and ‘you’, Authentic and negemo, affect and negemo, anx and negemo; moderate relationship between Tone and posemo; and strong relationship between affect and posemo.

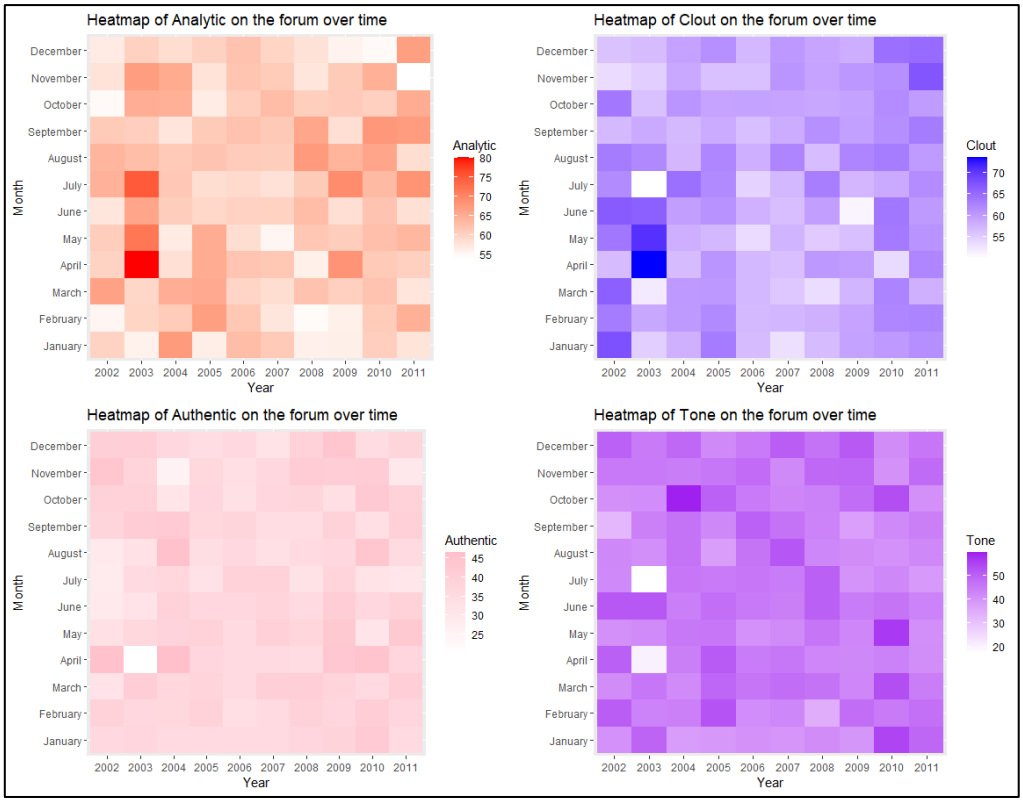

Diagram 2: Heatmaps of linguistic variables on the forum over time

	WC	Analytic	Clout	Authentic	Tone	WPS	i	we	you	they	number	affect	posemo	negemo	anx
WC	1.00	0.10	0.09	-0.04	-0.08	0.19	-0.08	0.04	-0.03	0.05	0.00	-0.08	-0.09	0.00	0.01
Analytic	0.10	1.00	0.11	-0.09	-0.01	0.07	-0.32	-0.11	-0.21	-0.16	0.13	-0.01	0.00	-0.04	-0.02
Clout	0.09	0.11	1.00	-0.37	0.01	0.06	-0.46	0.35	0.35	0.21	-0.03	0.02	0.03	-0.02	0.01
Authentic	-0.04	-0.09	-0.37	1.00	0.02	0.02	0.45	-0.06	-0.01	-0.05	-0.02	-0.11	-0.07	-0.09	-0.03
Tone	-0.08	-0.01	0.01	0.02	1.00	-0.07	0.07	-0.02	0.06	-0.09	-0.01	0.26	0.53	-0.45	-0.17
WPS	0.19	0.07	0.06	0.02	-0.07	1.00	-0.03	0.04	-0.04	0.07	0.00	-0.13	-0.14	0.00	0.01
i	-0.08	-0.32	-0.46	0.45	0.07	-0.03	1.00	-0.11	-0.03	-0.09	-0.03	-0.01	-0.01	-0.01	0.00
we	0.04	-0.11	0.35	-0.06	-0.02	0.04	-0.11	1.00	-0.05	0.02	-0.06	-0.02	-0.03	0.02	0.03
you	-0.03	-0.21	0.35	-0.01	0.06	-0.04	-0.03	-0.05	1.00	-0.08	-0.07	0.02	0.04	-0.03	-0.02
they	0.05	-0.16	0.21	-0.05	-0.09	0.07	-0.09	0.02	-0.08	1.00	-0.07	-0.03	-0.08	0.07	0.04
number	0.00	0.13	-0.03	-0.02	-0.01	0.00	-0.03	-0.06	-0.07	-0.07	1.00	-0.06	-0.04	-0.05	-0.03
affect	-0.08	-0.01	0.02	-0.11	0.26	-0.13	-0.01	-0.02	0.02	-0.03	-0.06	1.00	0.87	0.41	0.12
posemo	-0.09	0.00	0.03	-0.07	0.53	-0.14	-0.01	-0.03	0.04	-0.08	-0.04	0.87	1.00	-0.09	-0.04
negemo	0.00	-0.04	-0.02	-0.09	-0.45	0.00	-0.01	0.02	-0.03	0.07	-0.05	0.41	-0.09	1.00	0.32
anx	0.01	-0.02	0.01	-0.03	-0.17	0.01	0.00	0.03	-0.02	0.04	-0.03	0.12	-0.04	0.32	1.00

Diagram 3: Correlation test results

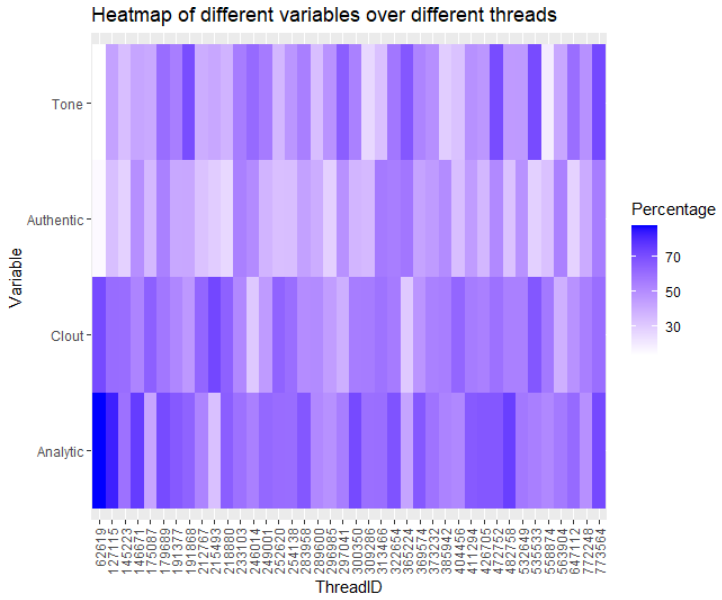

Diagram 4: Heatmap of variables (LIWC summary) over threads

Question (b)

Originally, there are too many threads to interpret if I am to analyse all of them as there are 4063 threads in total. Moreover, some of the threads may even have not much posts so there is not much information to obtain. Therefore, I have filtered the threads so that only threads with posts more than 30 will be analysed to see if there is a difference in language used by different threads. As we can see from Diagram 4, the percentage of each variable (Analytic, Clout, Authentic and Tone) varies from around 20 to 80 for different threads whereas for the variables that are less important such as I, we, you, they, number, affect, negemo, posemo and anx, we can also see obvious difference between different threads (Diagram 5). In addition, t-tests are done for thread 252620 and 283958 which have the most posts in a thread. We have moderate to strong evidence to reject the null hypothesis (mean of linguistic variables are the same for the two threads) as p-values are less than 0.05 except for affect and anx. This means that the language used for thread ID 252620 and 283958 are mostly different. Besides that, to compare language of different threads at similar times, I have created a function to roughly calculate the number of threads created in a same month using similar language and different language. Threads that consist of more than 30 posts are analysed and there are 7 threads using same language and 4 threads using different language. Therefore, I would say that it is possible to see different threads using different language but it does not guarantee that different threads must use different language. Next, only some of the threads are chosen to analyse if the language used within threads changes over time. In this case, threads with posts more than 100 are chosen. There are 4 threads which have number of posts more than 100, those are threads with ThreadID: 252620, 283958, 127115 and 472752. For these threads chosen, there are threads that has long active time and there are also short active time threads. So, we can take a look at the changes of language over time in both situations. Bar graphs of linguistic variables over time are plotted for each of the thread mentioned above (only thread 252620 and 283958 are shown here: Diagram 6 and Diagram 7). From the graphs, we can see that the language used within these threads are not consistent as the variables are sometimes high and sometimes low. T-tests are also done to have a better evidence for me to conclude. In conclusion, the language within a thread changes over time for most of the threads. Although it is said that authors discussed about the same topic for the same thread, I suppose that different author may have their own opinion and personality, therefore they are using a different language to discuss.

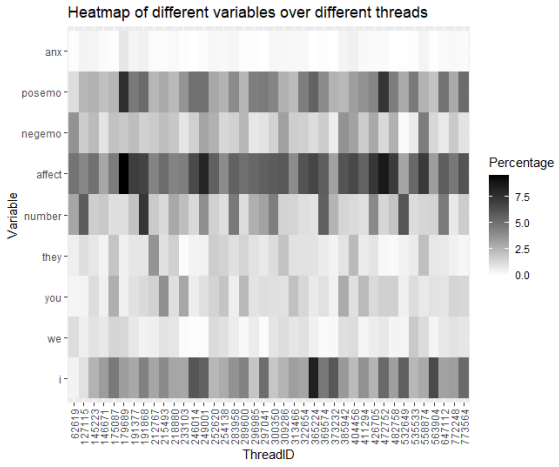

Diagram 5: Heatmap of variables (other than LIWC summary) over threads

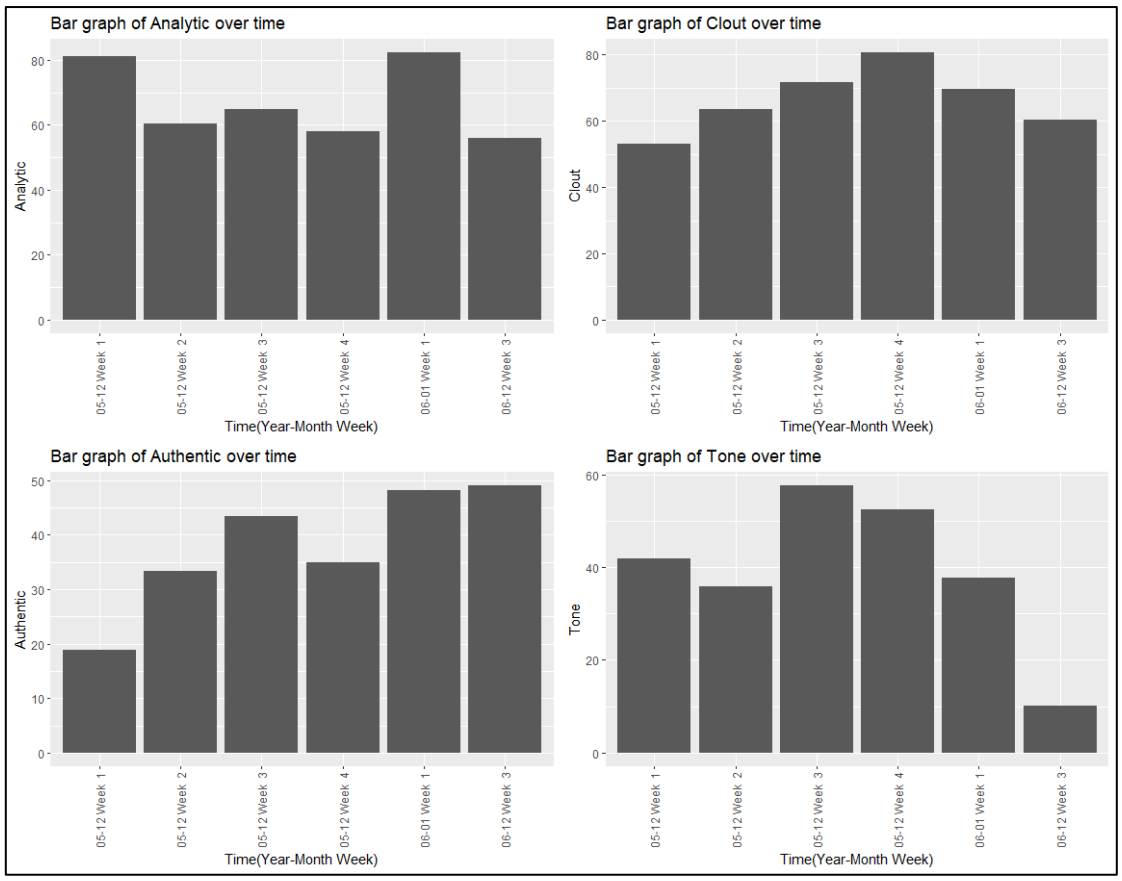

Diagram 6: Bar graphs of LIWC summary variables over time (ThreadID: 252620)

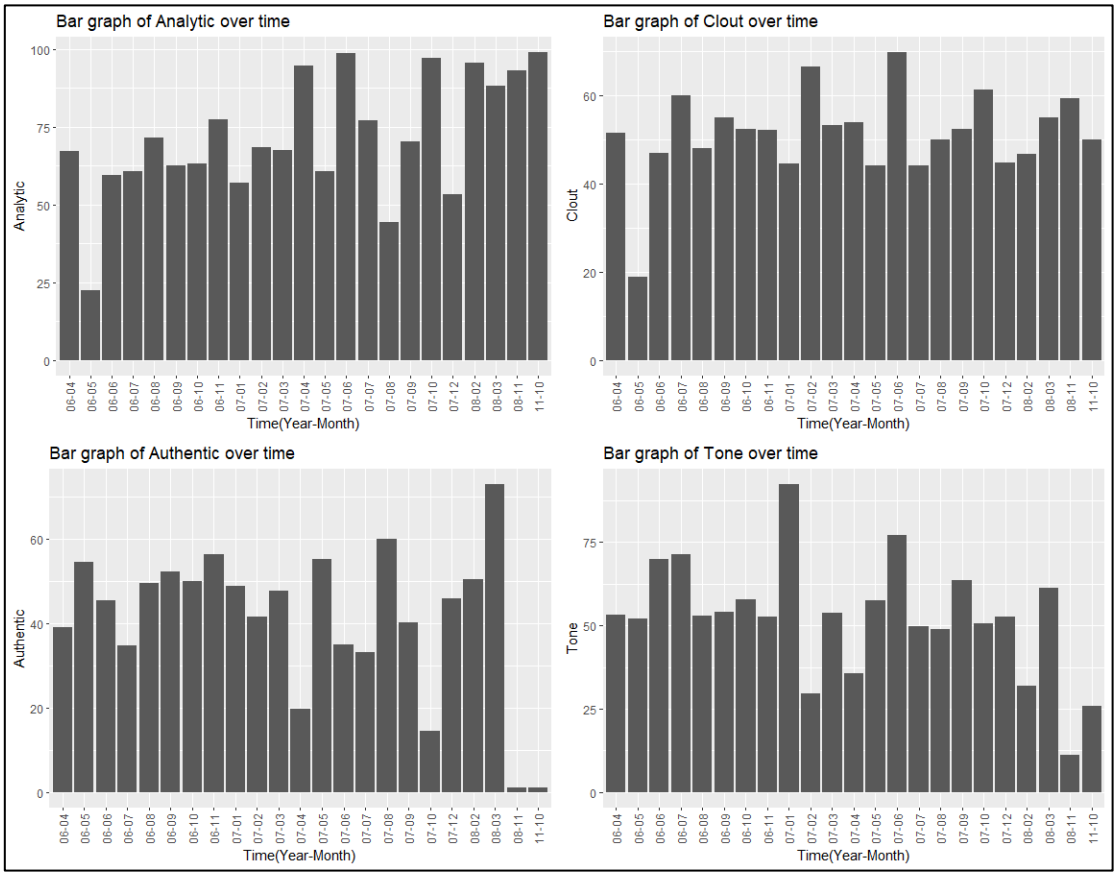

Diagram 7: Bar graphs of LIWC summary variables over time (ThreadID: 283958)

Question (c)

Social network of January 2009 and February 2009 are plotted and compared. Since authors who post to the same thread during the same month form a social network, the node in the social network are the authors (AuthorID) and the edge indicates the authors posting to the same thread. Diagram 8 shows the whole social network during January 2009 whereas Diagram 9 shows the whole social network during February 2009. A whole social network is plotted rather than just a small portion of it can ensure the result is not misleading. Just by observing the social network of both months, we can tell that Diagram 8 has a higher density of authors compared to Diagram 9 even though there are three clusters of disconnected authors in Diagram 8 while in Diagram 9, there is just two clusters of disconnected authors. Vertex count, edge count, diameter, average path length, graph density and transitivity of January 2009 and February 2009 are added beside the social networks respectively to have a better vision of them. Histogram of degree distribution for both months are also plotted (Diagram 12 and Diagram 13 respectively). We can see that there are more authors with higher degree distribution during January 2009 comparing Diagram 10 with Diagram 11. Referring to the factors such as diameter, average path length and degree distribution, we can clearly tell that January 2009 has a more connected and robust social network compared to February 2009. Therefore, we can deduce that participants are more active across different threads during January 2009 compared to February 2009.

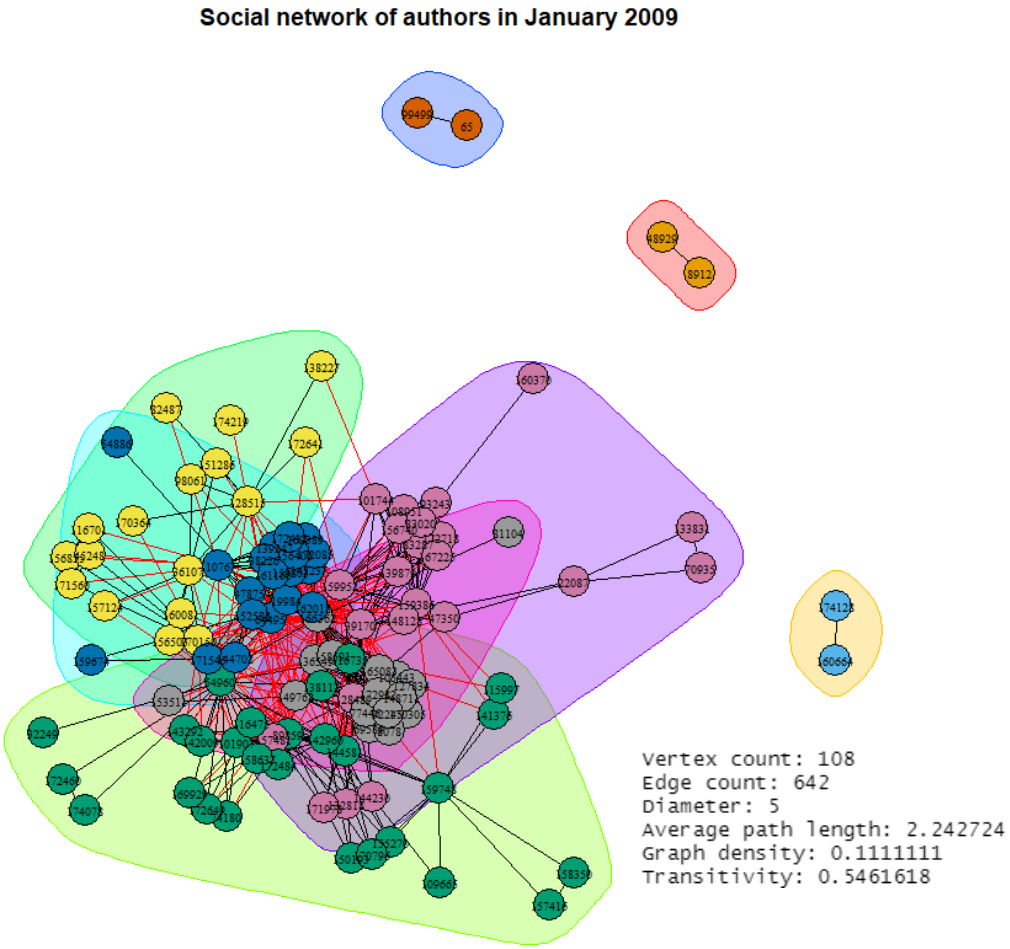

Diagram 8: Social Network during January 2009

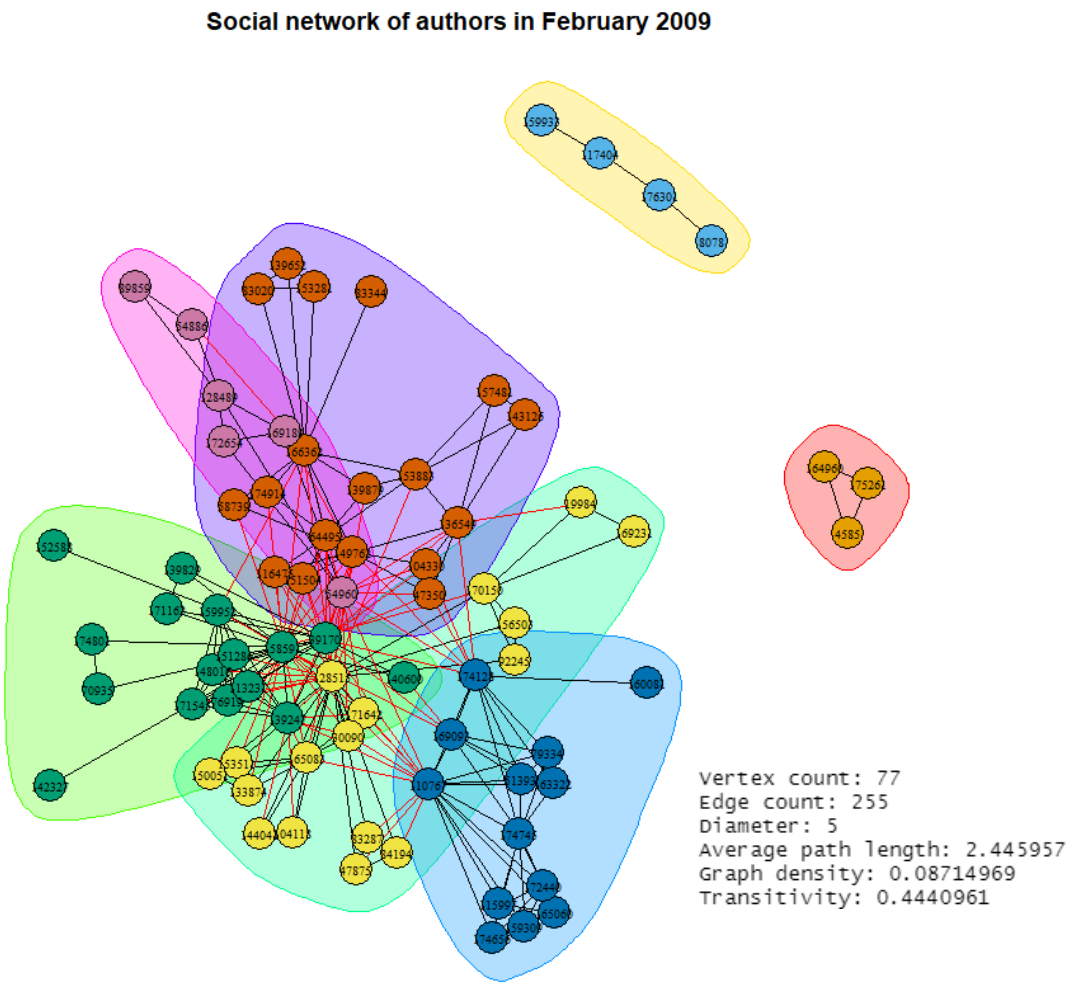

Diagram 9: Social Network during February 2009

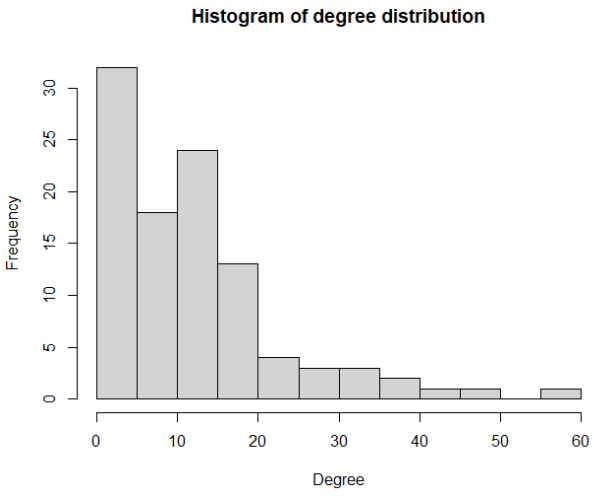

Diagram 10: Degree distribution of January

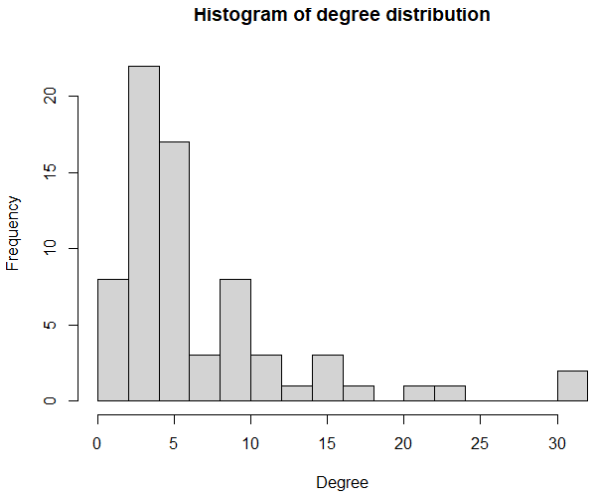

Diagram 11: Degree distribution of February

Question (d)

I used the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology to complete this assignment. First phase of CRISP-DM is business understanding, where in this assignment, it is to understand the requirements of the question. Therefore, I first focus on understanding the requirements of the question to build strong foundation for the following phase. Next, the second phase is data understanding. To understand the data, I visited the webpage provided to understand each meaning of the variables and the form of the variables are stored. After having a good understanding of the questions and data, I started to work on each of the questions. To have a helpful data for me to complete the requirements, I selected data which are useful for that specific question and also derived some new columns (attributes). For example, attributes such as "Year.month", "Year" and "Month" are created for me to plot graphs easily. Then, for the next phase, I chose the type of graphs which are suitable for me to analyse the data to meet the requirements of the question. Some statistical tests such as t-test or correlational test are also done during this phase. After the graphs are plotted and statistical tests are done, I evaluated them and obtained the conclusion out of them. For example, for question (a), after I read the documentation of the variables in the webpage, I noted down that there is a relationship in every different variable. But when I started to investigate on it by doing a correlational test, I found out that the relationships between variables are very weak or even have no relationship. Hence, I modified my research based on this investigation. After getting the conclusion, I reviewed the work accomplished to ensure that nothing important is overlooked. After all of the above are done, I compiled all of the useful graphs or data and documented it in this research report.

Appendix

```
library(ggplot2)
library(reshape2)
library(igraph)
library(igraphdata)
library(ggpubr)

rm(list = ls())

# We are going to use webforum for (a) till (c)
# Add a column so called Year.month, which will be frequently used
set.seed(29797918)
webforum <- read.csv("webforum.csv")
webforum <- webforum[sample(nrow(webforum),20000),] # 20000 rows
webforum$Year.month <- format(as.Date(webforum$Date), "%y-%m")


# ------------------------------------ Question(a) ------------------------------------ #

# Function to convert Year.month to month (January - Deecember)
convertToMonth <- function(yearmonth){
  list_of_months <-
factor(c("January","February","March","April","May","June","July","August","September","Oct
ober","November","December"),
levels=c("January","February","March","April","May","June","July","August","September","Oct
ober","November","December"))
  return(list_of_months[as.numeric(substr(yearmonth,4,5))])
}

# Function to convert Year.month to year (2002 - 2011)
convertToYear <- function(yearmonth){
  return(as.factor(paste("20", substr(yearmonth,1,2), sep="")))
}

### Analyze activity over time and see if there's a trend ###
# Create a data frame which is grouped by Year.month and the frequency of each month is in
the data frame too
# Add columns of Month and Year so that we are able to draw heatmap easily using these later
activity <- as.data.frame(table(webforum$Year.month))
colnames(activity) <- c("Year.month","CountOfActivity")
activity$Month <- convertToMonth(activity$Year.month)
activity$Year <- convertToYear(activity$Year.month)

# Heatmap of activity over time (the count of posts is the activeness of participants on
the forum)
ggplot(data=activity, aes(Year, Month)) + geom_tile(aes(fill=CountOfActivity)) +
scale_fill_distiller(palette = "RdPu") + ggtitle("Heatmap of activity on the forum over
time")

### Analyze language on the forum over time ###
# Create a data frame with the mean of all of the linguistic variables grouped by Year.month
# Add columns of Month and Year so that we are able to draw heatmap easily using these later
language <- aggregate(webforum[,5:19], list(webforum$Year.month), mean)
colnames(language)[1] <- c("Year.month")
language$Month <- convertToMonth(language$Year.month)
language$Year <- convertToYear(language$Year.month)
```

```
# Heatmap of linguistic variables:
#   - Heatmaps of LIWC Summary: Analytic, Clout, Authentic and Clout respectively
a <- ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=Analytic)) +
scale_fill_gradient(low="white", high="red") + ggtitle("Heatmap of Analytic on the forum
over time")
b <- ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=Clout)) +
scale_fill_gradient(low="white", high="blue") + ggtitle("Heatmap of Clout on the forum over
time")
c <- ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=Authentic)) +
scale_fill_gradient(low="white", high="pink") + ggtitle("Heatmap of Authentic on the forum
over time")
d <- ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=Tone)) +
scale_fill_gradient(low="white", high="purple") + ggtitle("Heatmap of Tone on the forum
over time")
ggarrange(a,b,c,d, ncol = 2, nrow = 2)

#   - Heatmaps of other linguistic variables: i, we, you, they, number, affect, negemo,
posemo, anx
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=i)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of 'i' on the forum over
time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=we)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of 'we' on the forum over
time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=you)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of 'you' on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=they)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of 'they' on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=number)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of number on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=affect)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of affect on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=negemo)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of negemo on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=posemo)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of posemo on the forum
over time")
ggplot(data=language, aes(Year, Month)) + geom_tile(aes(fill=anx)) +
scale_fill_gradient(low="white", high="black") + ggtitle("Heatmap of anx on the forum over
time")

# To see if there is a relationship between variables, a correlation test is done
round(cor(webforum[5:19]), digits=2)


# ------------------------------------ Question(b) ------------------------------------ #

### Analyze language used by different groups (threads) ###
# Create a data frame grouped by ThreadID and the number of posts in that ThreadID (Frequency)
threadFreq <- as.data.frame(table(webforum$ThreadID))
colnames(threadFreq) <- c("ThreadID","Frequency")
```

```
# Create a data frame grouped by ThreadID and the Year.month the ThreadID was created
mins <- aggregate(webforum$Date, list(webforum$ThreadID), min)
colnames(mins) <- c("ThreadID","Date")
mins$Year.month <- format(as.Date(mins$Date), "%y-%m")

# Create a data frame grouped by ThreadID with means of linguistic variables
groupedByThread <- aggregate(webforum[,5:19], list(webforum$ThreadID), mean)
groupedByThread <- as.data.frame(round(groupedByThread[,], digits=2))
colnames(groupedByThread)[1] <- "ThreadID"

# Merge data frames created above by their common column, ThreadID, to one single data frame
toBeMerged <- as.data.frame(merge(mins, groupedByThread, by="ThreadID"))
thread <- as.data.frame(merge(toBeMerged, threadFreq, by="ThreadID"))

# Total number of threads in the forum
nrow(thread)

# Since there is too many threads, only threads with posts more than 30 will be chosen
test <- thread[thread$Frequency > 30,]

# Tidy the data to two sets: one contains variables of LIWC summary (test1), one contains
other linguistic variables (test2)
test1 <- melt(test, id.vars="ThreadID",
measure.vars=c("Analytic","Clout","Authentic","Tone"))
test2 <- melt(test, id.vars="ThreadID",
measure.vars=c("i","we","you","they","number","affect","negemo","posemo","anx"))
colnames(test1) <- c("ThreadID","Variable","Percentage")
colnames(test2) <- c("ThreadID","Variable","Percentage")
test1$ThreadID <- as.factor(test1$ThreadID)
test2$ThreadID <- as.factor(test2$ThreadID)

# Plot heatmap of both test1 and test2 to see the language of different threadID
ggplot(data=test1, aes(ThreadID, Variable)) + geom_tile(aes(fill=Percentage)) +
scale_fill_gradient(low="white", high="blue") + theme(axis.text.x = element_text(angle =
90, vjust = 0.5, hjust=1)) + ggtitle("Heatmap of different variables over different
threads")
ggplot(data=test2, aes(ThreadID, Variable)) + geom_tile(aes(fill=Percentage)) +
scale_fill_gradient(low="white", high="black") + theme(axis.text.x = element_text(angle =
90, vjust = 0.5, hjust=1)) + ggtitle("Heatmap of different variables over different
threads")

# Do t-test of language of two different threads to determine if they uses same language by
looking at p-value
# - Choose the top two threads with most posts as example
maxi <- thread[order(thread$Frequency, decreasing=TRUE),]
x <- webforum[webforum$ThreadID==maxi[1,1],]
y <- webforum[webforum$ThreadID==maxi[2,1],]
t.test(x$Analytic,y$Analytic)$p.value
t.test(x$Clout,y$Clout)$p.value
t.test(x$Authentic,y$Authentic)$p.value
t.test(x$Tone,y$Tone)$p.value
t.test(x$i,y$i)$p.value
t.test(x$we,y$we)$p.value
t.test(x$you,y$you)$p.value
t.test(x$they,y$they)$p.value
t.test(x$number,y$number)$p.value
```

```
t.test(x$affect,y$affect)$p.value
t.test(x$negemo,y$negemo)$p.value
t.test(x$posemo,y$posemo)$p.value
t.test(x$anx,y$anx)$p.value

# Function that uses t-test to analyze language of different threads at similar times (same
month)
analyseLanguageOfDiffThreadAtSimilarTime <- function(df, numOfPostsInThread){

  # Ensure no NA
  df <- df[complete.cases(df),]

  # Order the data frame by Date and ensure the number of posts in thread is more than
numOfPostsInThread
  df <- df[order(df$Date),]
  maxi <- df[df$Frequency>numOfPostsInThread,]

  # vector will be storing pairs of ThreadIDs created at similar time to be compared using
t-test
  vector <- vector()
  num <- nrow(maxi)-1
  for(i in c(1:num)){
    if(maxi[i,"Year.month"] == maxi[i+1,"Year.month"]){
      ans <- paste(maxi[i,"ThreadID"], maxi[i+1,"ThreadID"])
      vector <- c(vector, ans)
    }
  }

  # If there are no threads at similar time then return 0 for same and diff language
  if (length(vector)==0){
    return(list(same.language=0, diff.language=0))
  }

  # Create a table with two columns containing pairs of threads found above
  table <- strsplit(vector, split=" ")
  result <-  do.call(rbind, table)

  # Perform t-test for all linguistic variables and if there are more than half of the p-
values are larger than 0.1,
  # meaning we have weak/no evidence against the null hypothesis (mean of two groups are
the same), then it will be
  # the same language used by two different threads
  num <- nrow(result)

  same = 0
  diff = 0

  for (i in c(1:num)){

    x <- webforum[webforum$ThreadID==result[i,1],]
    y <- webforum[webforum$ThreadID==result[i,2],]
    lst <- list()
    counter = 0
    lst <- append(lst, t.test(x$Analytic,y$Analytic)$p.value)
    lst <- append(lst, t.test(x$Clout,y$Clout)$p.value)
    lst <- append(lst, t.test(x$Authentic,y$Authentic)$p.value)
    lst <- append(lst, t.test(x$Tone,y$Tone)$p.value)
```

```
    lst <- append(lst, t.test(x$i,y$i)$p.value)
    lst <- append(lst, t.test(x$we,y$we)$p.value)
    lst <- append(lst, t.test(x$you,y$you)$p.value)
    lst <- append(lst, t.test(x$they,y$they)$p.value)
    lst <- append(lst, t.test(x$number,y$number)$p.value)
    lst <- append(lst, t.test(x$affect,y$affect)$p.value)
    lst <- append(lst, t.test(x$negemo,y$negemo)$p.value)
    lst <- append(lst, t.test(x$posemo,y$posemo)$p.value)
    lst <- append(lst, t.test(x$anx,y$anx)$p.value)

    lst <- lst[!sapply(lst, is.nan)]

    for (j in lst){
      if (j > 0.1){
        counter = counter + 1
      }
    }

    if (counter > 7){
      same = same + 1
    } else{
      diff = diff + 1
    }
  }

  # Return the number of threads which uses same langauge and different language
  return(list(same.language=same, diff.language=diff))
}

# We can also analyze language of different threads at similar times (same month)
# Analyze threads that have more than 30 posts
analyseLanguageOfDiffThreadAtSimilarTime(thread, 30)


### Analyze language over time used Within a thread ###
# Function to find week of month of the date
convertToWeek <- function(date){
  return(ceiling(as.numeric(format(as.Date(date), "%d")) / 7))
}

# Create a data frame with number of posts in thread more than 100 then order it by its
number of posts
threadData <- thread[thread$Frequency > 100,]
threadData <- threadData[order(threadData$Frequency, decreasing=TRUE),]

# Number of threads which has number of posts more than 100
nrow(threadData)

# The ThreadID of those threads with number of posts more than 100
threadData$ThreadID

# We want to analyze the language within the threads mentioned above
for (i in c(1:nrow(threadData))) {

  cat("\n\nResults of t-test of ThreadID", threadData[i,1], ": (Analytic, Clout, Authentic
and Tone)\n")
```

```
  # Create a temporary data frame to store all of the posts of that ThreadID and order them
according to Date, Time
  temp <- webforum[webforum$ThreadID==threadData[i,1],]
  temp <- temp[order(as.Date(temp$Date), as.POSIXct(temp$Time, format="%H:%M")),]

  # Create a data frame grouped by the Year.month with means of linguistics variables
  tempData <- aggregate(temp[,6:19], list(temp$Year.month), mean)
  colnames(tempData)[1] <- "Time(Year-Month)"

  # Do some t-test for the data in this thread, split the data to half according to date
  median_date <- median(as.Date(temp$Date))
  first <- temp[as.Date(temp$Date) > median_date,]
  second <- temp[as.Date(temp$Date) < median_date,]

  print(t.test(first$Analytic, second$Analytic, "greater", conf.level = 0.99))
  print(t.test(first$Clout, second$Clout, "greater", conf.level = 0.99))
  print(t.test(first$Authentic, second$Authentic, "greater", conf.level = 0.99))
  print(t.test(first$Tone, second$Tone, "greater",conf.level = 0.99))

  # If the x-axis has too little time, then split it to Year.month Week and draw the graph
of variables of LIWC summary into one graph, else just use the Year.month as x-axis to plot
the graph of variables of LIWC summary into one graph
  if (nrow(tempData) < 5){
    temp$Week <- paste(temp$Year.month, "Week ", convertToWeek(temp$Date))
    tempData <- aggregate(temp[,6:19], list(temp$Week), mean)
    colnames(tempData)[1] <- "Time(Year-Month Week)"
    a <- ggplot(data=tempData, aes(x=`Time(Year-Month Week)`, y=Analytic)) +
geom_bar(stat="identity")  + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Analytic over time")
    b <- ggplot(data=tempData, aes(x=`Time(Year-Month Week)`, y=Clout)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Clout over time")
    c <- ggplot(data=tempData, aes(x=`Time(Year-Month Week)`, y=Authentic)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Authentic over time")
    d <- ggplot(data=tempData, aes(x=`Time(Year-Month Week)`, y=Tone)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Tone over time")
    figure <- ggarrange(a,b,c,d, ncol = 2, nrow = 2)
    print(figure)
  } else{
    a <- ggplot(data=tempData, aes(x=`Time(Year-Month)`, y=Analytic)) +
geom_bar(stat="identity")  + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Analytic over time")
    b <- ggplot(data=tempData, aes(x=`Time(Year-Month)`, y=Clout)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Clout over time")
    c <- ggplot(data=tempData, aes(x=`Time(Year-Month)`, y=Authentic)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Authentic over time")
    d <- ggplot(data=tempData, aes(x=`Time(Year-Month)`, y=Tone)) +
geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1)) + ggtitle("Bar graph of Tone over time")
    figure <- ggarrange(a,b,c,d, ncol = 2, nrow = 2)
    print(figure)
  }
}
```

```
# ------------------------------------ Question(c) ------------------------------------ #

### Authors posting to the same thread at similar times (during the same month) as forming
a social network ###

# Function to draw a social network
drawSocialNetwork <- function(allPostWithinMonth, threadOfMonth, monthyear){

  # This list will contain AuthorIDs that post on the same thread during the same month
  lst <- list()

  # Loop to find all of the AuthorIDs that posted on the same ThreadID
  for (i in threadOfMonth[,1]){
    temp <- allPostWithinMonth[allPostWithinMonth$ThreadID==as.numeric(i),]
    u <- unique(temp["AuthorID"])
    if (nrow(u) > 1){
      lst <- append(lst, u)
    }
  }

  # Initialize merged network with the network of 1st thread
  merged = graph.full(length(lst[1]$AuthorID))

  # Rename vertex with AuthorIDs
  V(merged)$name = as.character(lst[1]$AuthorID)

  # For each thread
  for(i in 2:length(lst)){

    # Get id of author
    id_of_author = lst[i]$AuthorID

    # Get number of authors posting in this thread
    number_of_author = length(id_of_author)

    # Create a complete graph
    g = graph.full(number_of_author)

    # Rename vertex with AuthorIDs
    V(g)$name = as.character(id_of_author)

    # Merge with previously merged networks
    merged = union(merged, g)
  }

  # Do clustering for the social network
  network <- cluster_louvain(merged)

  # Place vertices on the plane using the force-directed layout algorithm by Fruchterman
and Reingold, merged as the graph to lay out
  LO <- layout_with_fr(merged)

  title <- paste("Social network of authors in", monthyear)
  plot(network, merged, vertex.label.cex=0.6, vertex.label.color="black", vertex.size=8,
edge.arrow.size=0.2, layout=LO, main=title)
```

```
  cat("Vertex count:", vcount(merged))
  cat("\nEdge count:", ecount(merged))
  cat("\nDiameter:", diameter(merged))
  cat("\nAverage path length:", average.path.length(merged))
  cat("\nGraph density:", graph.density(merged))
  cat("\nTransitivity:", transitivity(merged))

  # Plot histogram of degree distribution
  hist(degree(merged), breaks=15, xlab="Degree", main="Histogram of degree distribution")

}

# Choose a month to plot the social network, remember to omit author with AuthorID -1
first.month <- webforum[webforum$Year.month=="09-01",]
first.month <- first.month[!(first.month$AuthorID==-1),]

# Create data frame grouped by ThreadID of the data frame above (data frame of one month)
# If the frequency is 1, means there is only an author posting to that thread,
# drop that ThreadID because there won't be connection between the author himself/herself
threadOfFirstMonth <- as.data.frame(table(first.month$ThreadID))
threadOfFirstMonth <- threadOfFirstMonth[!(threadOfFirstMonth$Freq==1),]

# Draw the social network of this month
drawSocialNetwork(first.month, threadOfFirstMonth, "January 2009")

# Repeat the same as above for the next month
second.month <- webforum[webforum$Year.month=="09-02",]
second.month <- second.month[!(second.month$AuthorID==-1),]
threadOfSecondMonth <- as.data.frame(table(second.month$ThreadID))
threadOfSecondMonth <- threadOfSecondMonth[!(threadOfSecondMonth$Freq==1),]

drawSocialNetwork(second.month, threadOfSecondMonth, "February 2009")
```