Activity, Language Use and Social Interactions in An Online Community

Priscilla A. C. Tham

Monash University Malaysia

Author Note

Priscilla A. C. Tham is now at School of Information Technology, Monash University Malaysia

ptha0007@student.monash.edu

Activity, Language Use and Social Interactions in An Online Community

The existence of discussion forums provided the public with a new medium for communication adapted to the modern online community aside from physical interactions. It has proven useful as an information channel in the educational study, political protest, cross-country contact, etc. Following the movement control order (MCO) due to the fatal Coronavirus pandemic, the significance of online interactions increases in maintaining awareness of the latest situation by crowdsourcing data. Data tapped from collective intelligence contain a variety of information which when extracted, bring forward real-time knowledge of the community.

A sample size of 20,000 forum data throughout the year 2002 to 2011 is decided upon to analyze. Prior to analysis, data cleaning is performed to remove anonymous users for a much accurate result and to remove posts deduced as image posts where word count is zero. A few libraries are also involved to transform data into readable and interpretable format, namely:

a) ggplot2,
b) ggnet,
c) network and,
d) sna

A. Analyze Activity and Language on The Forum Over Time

How active are participants, and are there periods where this increase or decreases? Is there a trend over time?

A certain user's or thread's activity is determined using the metric post frequency. The algorithm includes: (1) Counting the occurrences of each author ID and thread ID in the data, (2) Group the occurrences of each author ID and thread ID in the data by year and (3) Plot a line graph

to show the trend. Table 1 records the all-time most active user and thread and most active user's most active thread from completing (1).

Table 1 All-time most active user and thread and most active user's most active thread IDs.

Most active user (AuthorID)	Most active thread (ThreadID)	Most active user's most active thread (ThreadID)
39170	252620	145223

Subsequently, the correlation (Figure I) between author ID and thread ID:

```
> cor(forum_data$AuthorID, forum_data$ThreadID)
[1] 0.6484128
```

Figure I Correlation between author ID and thread ID

proposes test on the hypothesis that the probability of the most active thread being the most active user's most active thread is ≤ 60%. The p-value states that if the probability of the most active thread being the most active user's most active thread is ≤ 60% to expect a difference as small as the one observed: 3.108739e-23% (< α = 0.05). Thus, the p-value suggests that there is enough evidence to reject the null hypothesis which renders the correlation as significant at the population level. The most active thread's thread ID may be the same as the most active user's most active thread's thread ID in the entire population.

Next, Figure II records the post frequency in each year from the end of computation (2) and (3). The trend is such that user activity rocketed until the year 2006 which is the maximum cut-off then dropped until at least the year 2010. It matches Google Trend on the term 'forum' (Google, Google Trend, 2020) whereas Google Trend shows increasing interest on the term 'Facebook' from the year 2006 onwards (Google, Google Trend, 2020). A deduction can be made that there was a cultural shift at the time from a traditional discussion forum to new social platforms where interactions are much more diversified with exciting games compared to just text posts.

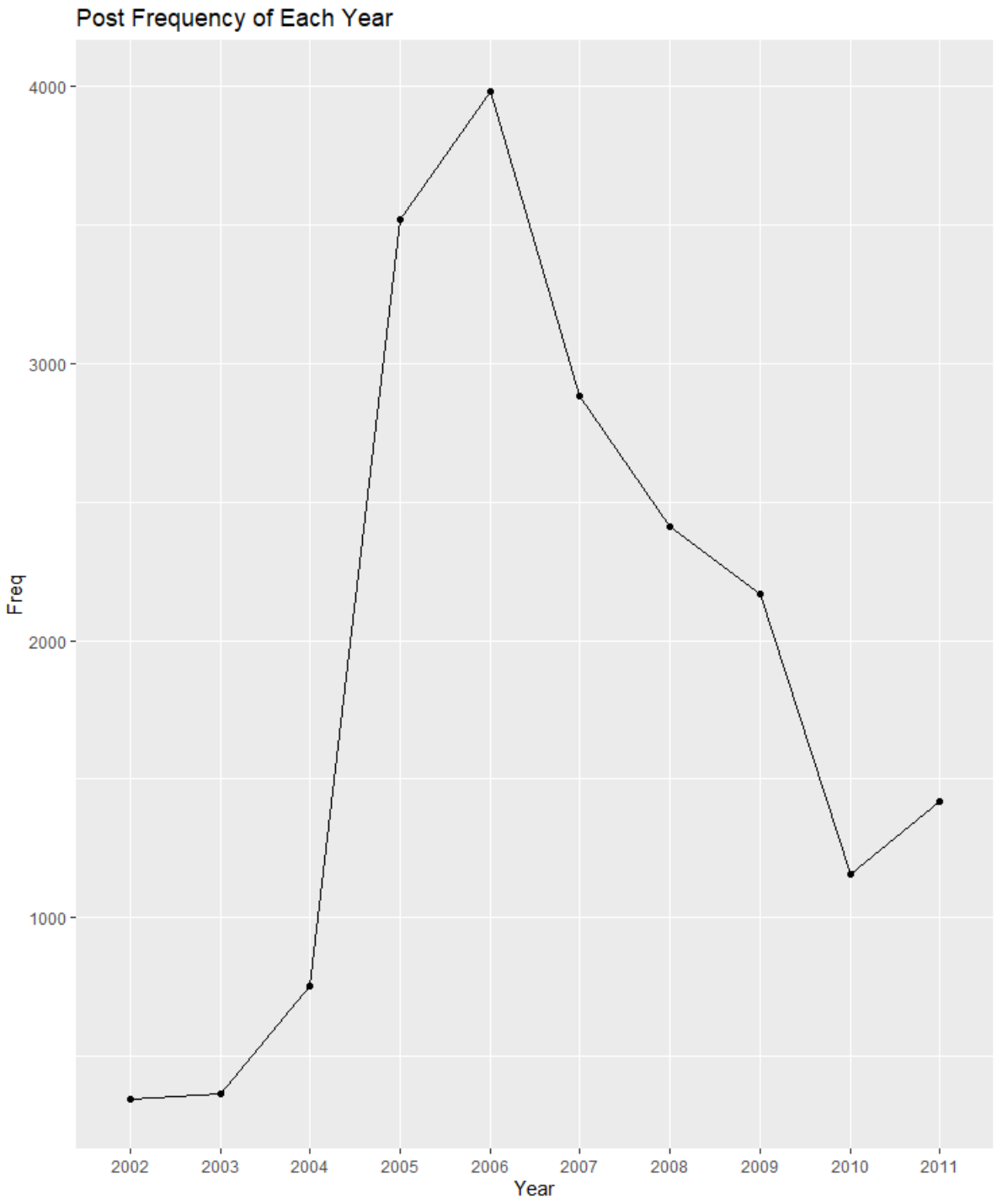


Figure II Post Frequency of Each Year

Looking at the linguistic variables, do these changes over time? Is there a relationship between them?

The data consists of the linguistic analysis based on Linguistic Inquiry and Word Count (LIWC) which values represents the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Table 2 details the data fields included in this paper for analysis.

Table 2 Description of data fields from the data included in this paper.

Column	**Description**
Analytic	LIWC Summary (analytical thinking)
Clout	LIWC Summary (power, force, impact)
Authentic	LIWC Summary (using an authentic tone of voice)
Tone	LIWC Summary (emotional tone)
ppron	LIWC (all personal pronouns)
i	LIWC ("I, me, mine" words) First person singular
we	LIWC ("We, us, our" words) First person plural
you	LIWC ("You" words) Second person
shehe	LIWC ("She, he, her, him" words) Third person singular
they	LIWC ("They" words) Third person plural
affect	LIWC (expressing sentiment)
posemo	LIWC (Positive emotions)
negemo	LIWC (Negative emotions)
anx	Words indicating anxiety
anger	Words indicating anger
social	Words referring to social processes
family	Words referring to family
friend	Words referring to friends/friendship
leisure	Words referring to leisure
money	Words referring to money
relig	Words referring to religion
swear	Swear words
QMark	Question Mark (Punctuation)

The data fields are divided into groups as such: (a) affect, posemo and negemo, (b) Analytic, Clout, Authentic and Tone, (c) ppron, i, we, you, shehe and they and (d) the rest in Table 2 to establish relationship easier.

```
> round(cor(forum_data[, c(17)],
+           forum_data[, 18:19]), digits=4)
     posemo negemo
[1,] 0.8269 0.4775
```

Figure III Correlation between affect and both posemo and negemo

Group (a) is such because positive and negative emotions describe the nature of the sentiments expressed. Moreover, the correlation between affect and posemo (Fig III) is high enough to propose test on the hypothesis that the probability of positive sentiment expressed is ≤ 80%. The p-value states that if the probability of positive sentiment expressed is ≤ 80% to expect a difference as small as the one observed: 7.135658e-80% (< $\alpha = 0.05$). Thus, the p-value suggests that there is enough evidence to reject the null hypothesis which renders the correlation as significant at the population level. The sentiment expressed may be positive emotions in the entire population.

This is elaborated in Figure IV of affect, posemo and negemo LIWC summary throughout the years. The curves of all three variables display similar local minimum or maximum of different LIWC values at the same point in time. Although so, the curve of negemo fluctuates stronger than affect and posemo which reflects the smaller correlation between affect and negemo. The trend is such that the act of expressing sentiment and positive emotions in the forum reduces from 2006 onwards while negative emotions continue fluctuating. Negative emotions did not rise as expected despite influenza A virus subtype H1N1 (A/H1N1) pandemic (Chan, 2009).

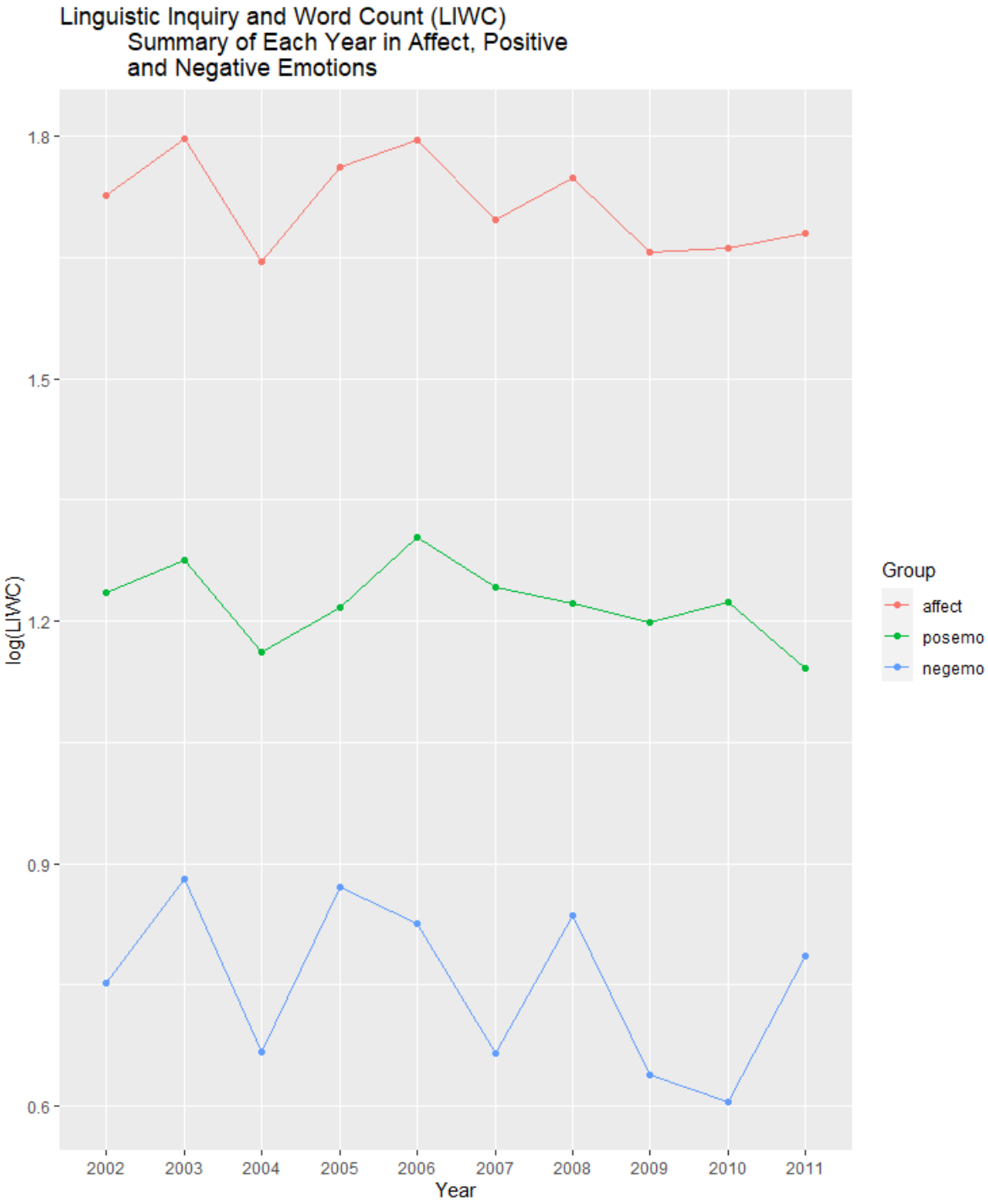


Figure IV LIWC Summary of Each Year in Affect, Positive and Negative Emotions

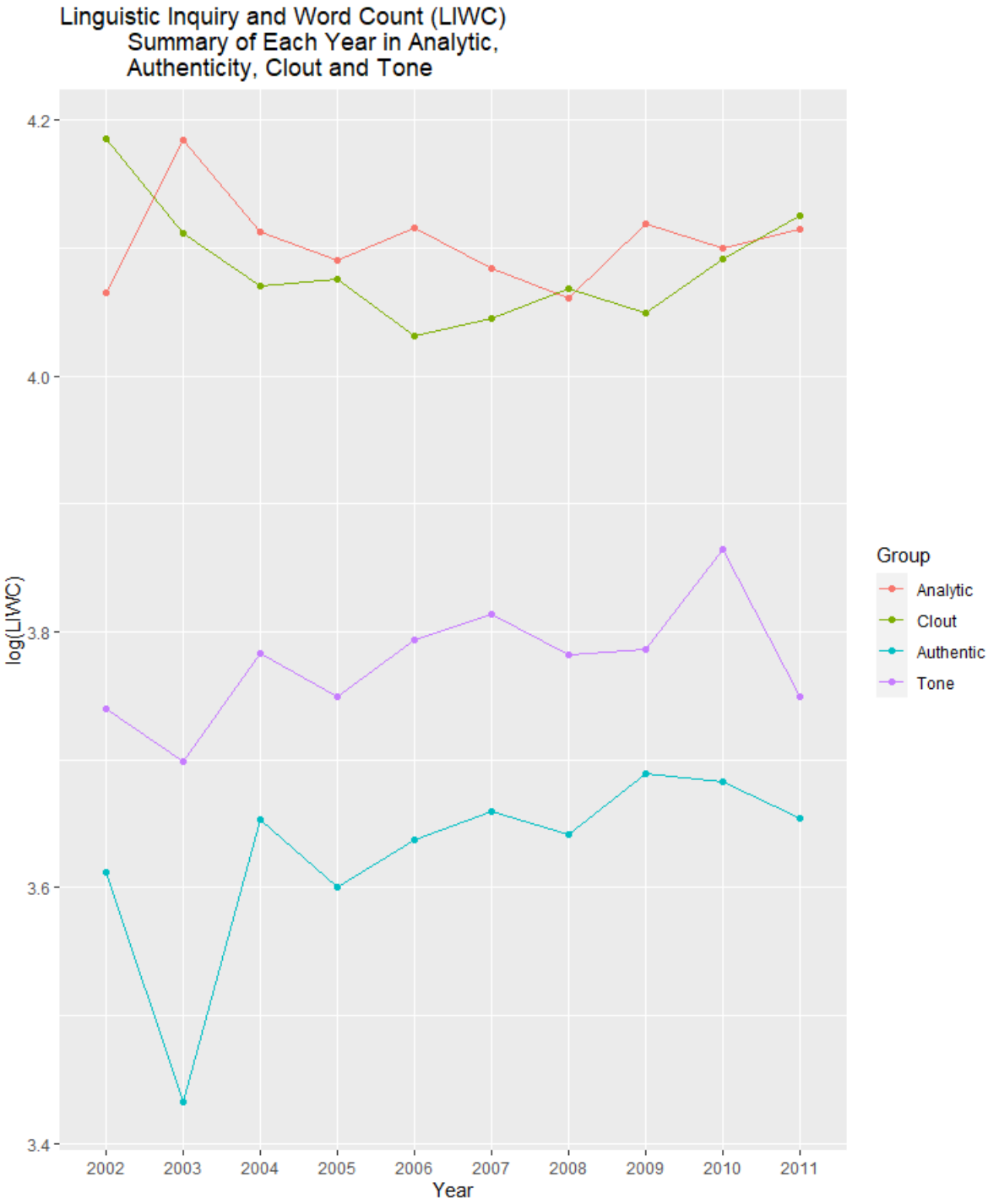


Figure V LIWC Summary of Each Year in Analytic, Authenticity, Clout and Tone

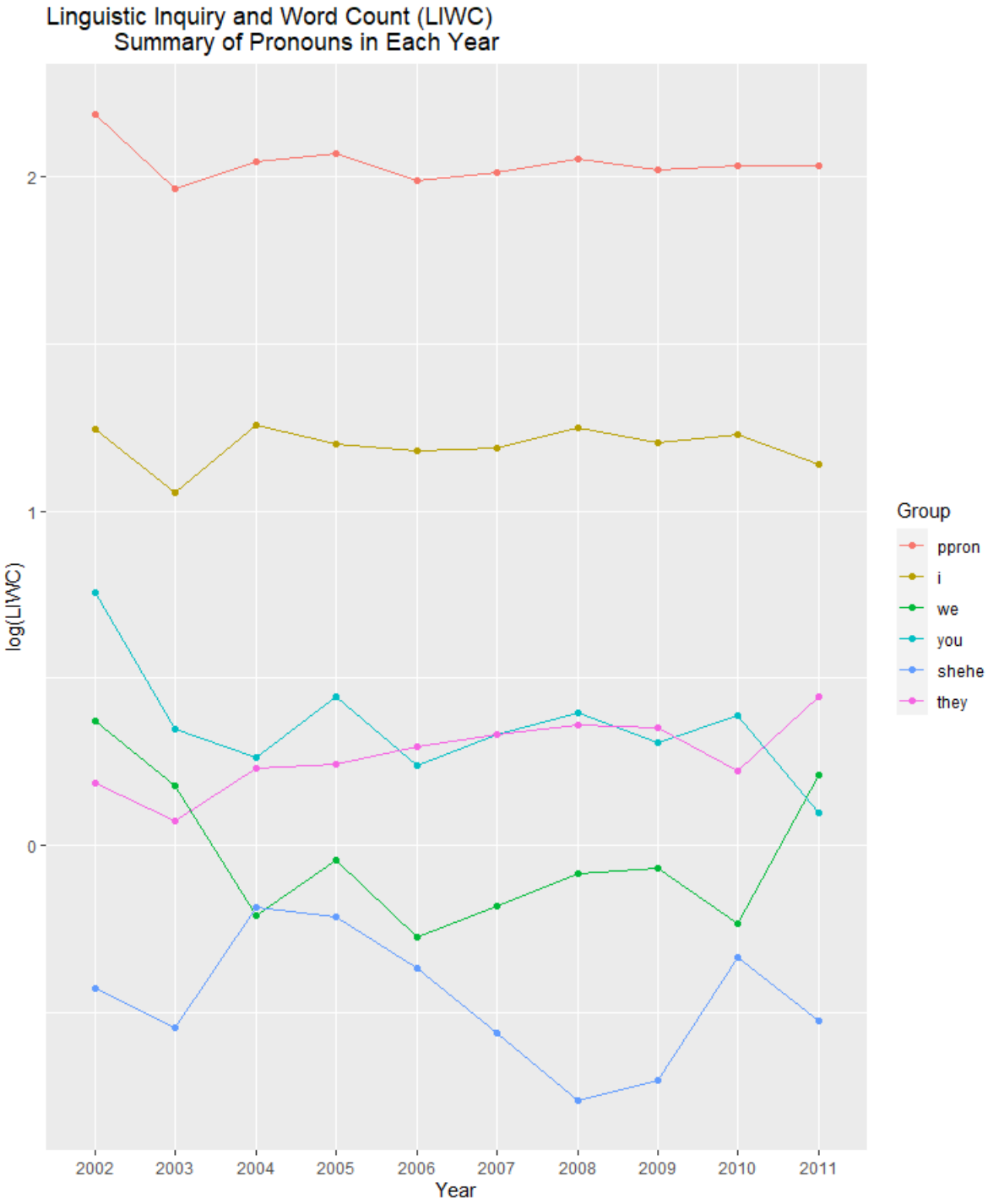


Figure VI LIWC Summary of Pronouns in Each Year

Additionally, LIWC summary of (b) and (c) throughout the years is also plotted. Group (b) is such because the variables describe the nature of the voice projected whereas group (c) is such because the variables describe the usage of pronouns. The trend is such that the users growingly instil analytical thinking combined with power, force and impact in text posts while authenticity and emotion kept escalating (Fig V). The subject of size 'they' is increasing in conversation as 'she', 'he', 'her', 'him' and 'you' decreases (Fig VI). Otherwise is quite constant for personal pronouns and 'i', and often fluctuates for 'we', 'us' and 'our' (Fig VI).

```
> cor(forum_data$social, forum_data$clout)
[1] 0.6605335
```

Figure VII Correlation between social and Clout

Group (d) is such because the variables are reference to the words used in text posts. Moreover, the correlation between social and Clout (Fig VII) is high enough to propose a test on the hypothesis that the probability of post content referring to social processes has higher Clout values is $\leq 60\%$. The p-value states that if the probability of post content referring to social processes has higher Clout values is $\leq 60\%$ to expect a difference as small as the one observed: 0% ($< \alpha = 0.05$). Thus, the p-value suggests that there is enough evidence to reject the null hypothesis which renders the correlation as significant at the population level. The social processes referred post content may have higher Clout values in the entire population.

Meanwhile, Figure VIII displays the proportion of words referring to the variables in (d) throughout the years. The trend is such that incorporation of family, friend starting from 2003, leisure starting from 2003 and money starting from 2005 related words in text posts are improving. Discussions in the forum are obviously getting optimistic as negative related words such as anger and swearing are slowly declining.

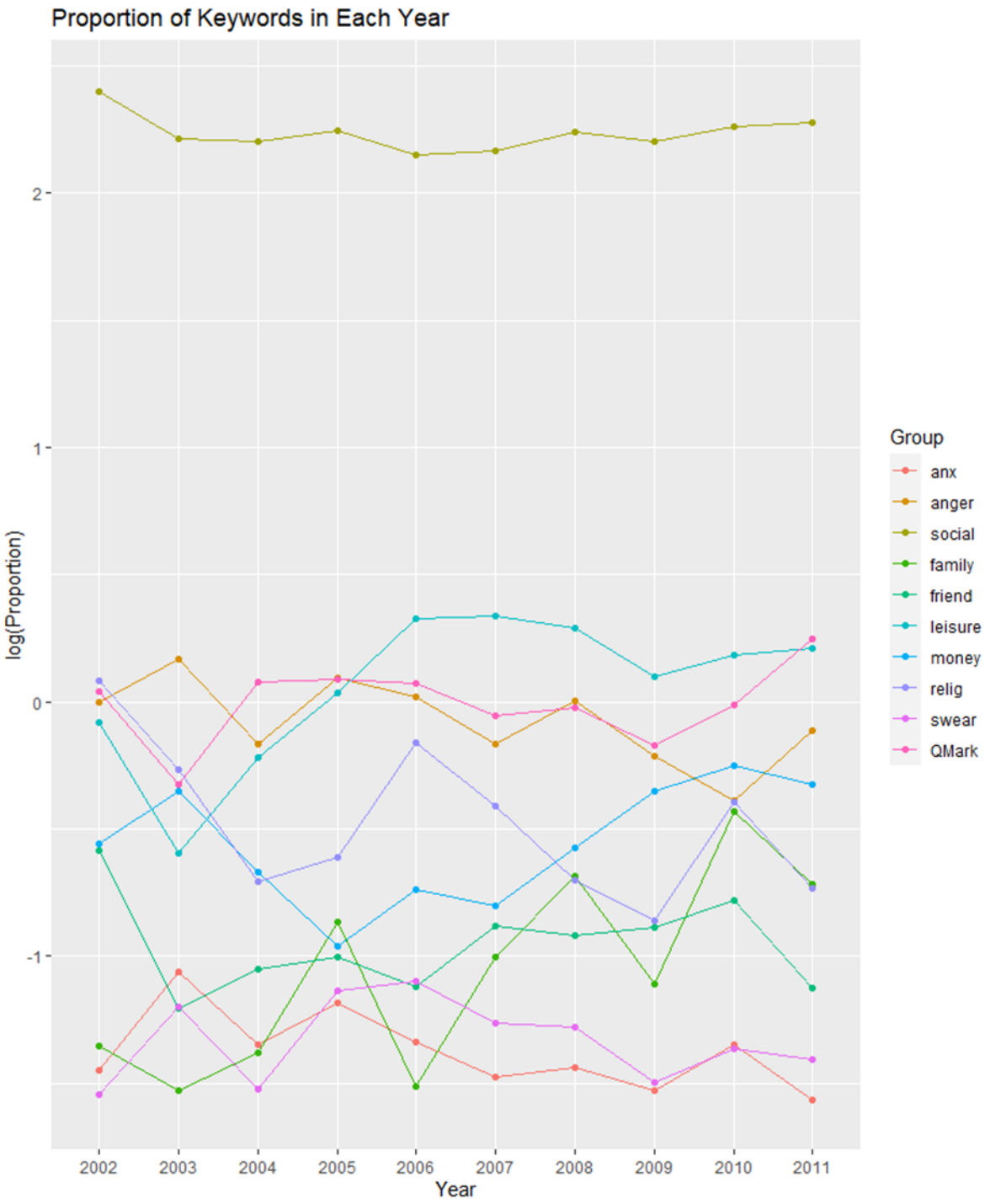


Figure VIII Proportion of Keywords in Each Year

```
> cor(forum_data$anger, forum_data$negemo)
[1] 0.6711201
```

Figure IX Correlation between anger and negemo

This is justified by the correlation between anger and negative emotions (Fig IX) which is high enough to propose a test on the hypothesis that the probability of anger referred words relating to negative emotions is ≤ 60%. The p-value states that if the probability anger referred words relating to negative emotions is ≤ 60% to expect a difference as small as the one observed: 0% (< $\alpha = 0.05$). Thus, the p-value suggests that there is enough evidence to reject the null hypothesis which renders the correlation as significant at the population level. The anger referred words may relate to negative emotions in the entire population.

$$Polarity[i,j] = posemo[i,j] - negemo[i,j]$$

Now, posemo and negemo are combined to a column called Polarity to normalize the two emotions rather than differentiating them using the formula above. A post is emotionally positive if Polarity is > 0, or emotionally negative, if Polarity is < 0. Figure X shows a surge of optimistic users enters the forum from 2004 to 2006 and 2009 to 2011 though there are more pessimistic users in 2010. It causes the trend in Figure IV on the emotions in posts throughout the years. An inference can be made that the trend in analytical thinking, power, force and impact, and authenticity in posts is due to the participation of these new users. The smaller points are seen to expand bigger later.

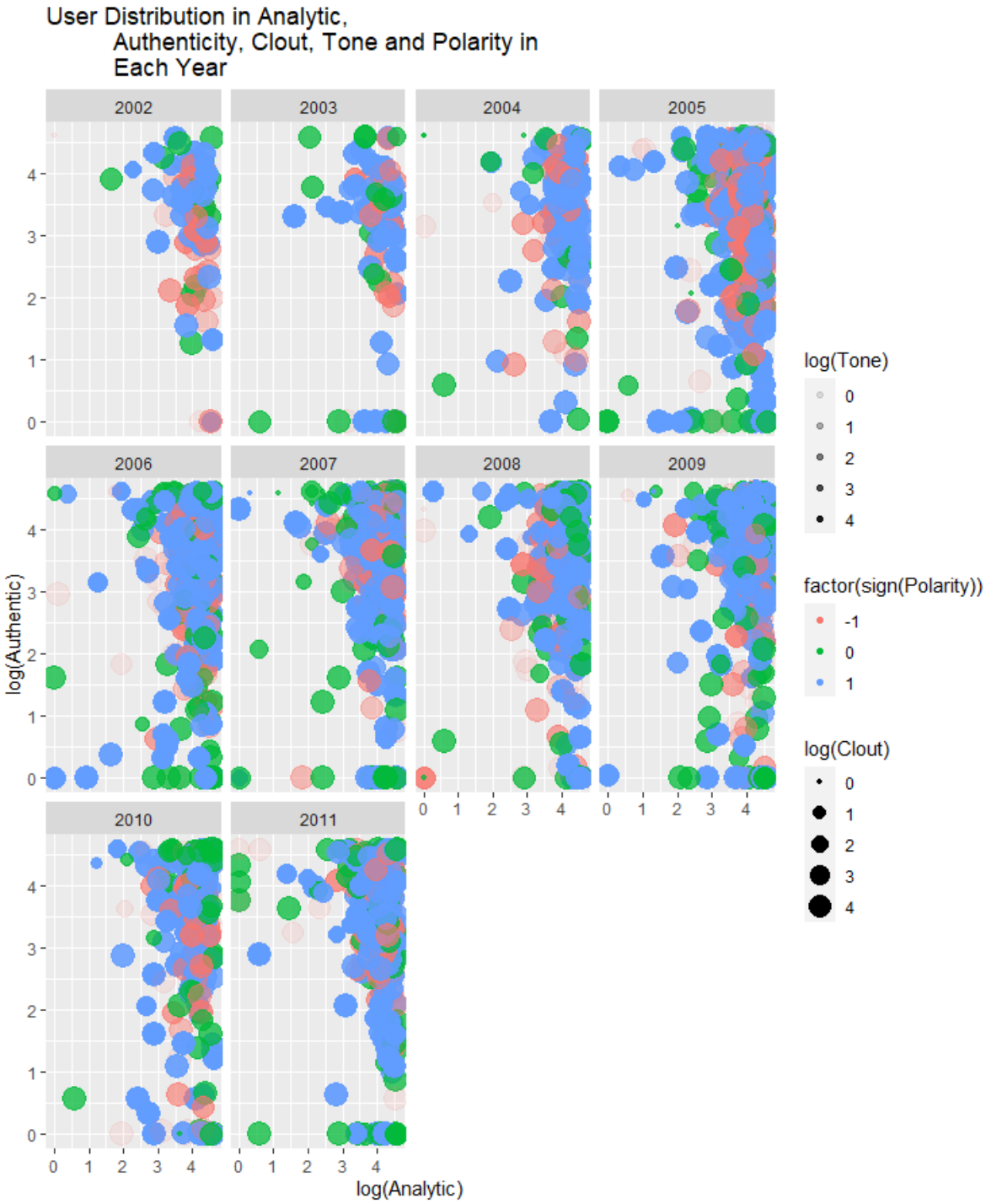


Figure X User Distribution in Analytic, Authenticity, Clout, Tone and Polarity in Each Year (Appendix II)

B. Analyze the Language Used by Groups

Describe the threads present in your data.

There are too many threads to consider, hence, nine random threads are chosen for analysis. The nine random threads must contain text posts in more than a year to allow trend observation. The thread IDs: 172784, 248834, 674017 are seen to plummet in the act of expressing sentiment whereas the thread IDs: 283958, 517685 have very similar pattern of sentiment expression and polarity (Fig XI). The thread IDs: 530558, 218736 are the exact reflection of each other (Fig XI). Clearly, there are differences between threads which is forwarded from the differences in linguistic variable involved while it is similar that every thread activity is not continuously active.

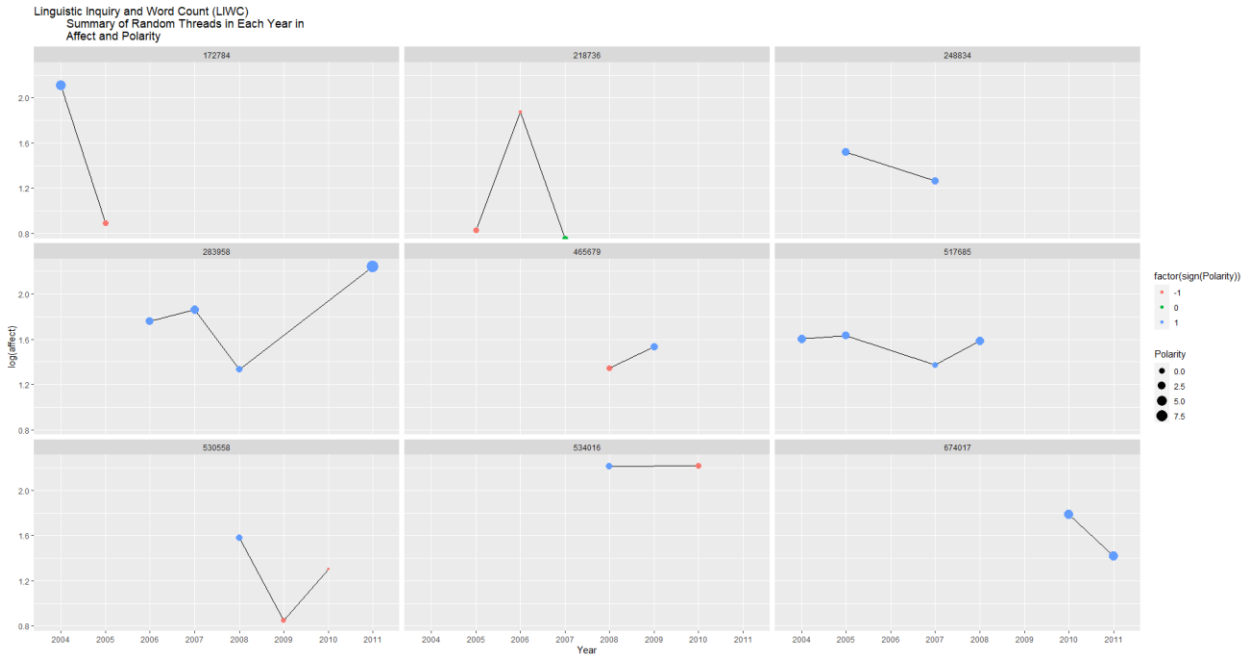


Figure XI LIWC Summary of Random Threads in Each Year in Affect and Polarity (Appendix III)

By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by these different groups? Does the language used within threads change over time?

Subsequently, it is safe to assume that the users in thread IDs: 465679, 517685, 674017 tend to mix in power, force and impact while being analytical. The authenticity and emotion in these threads are also soaring (Fig XII). The users in thread IDs: 283958, 534016 are superior in analytical thinking but are dropping in power, force and impact (Fig XII). As social is significantly related to Clout as mentioned in section A, the usage of words referring to social processes in the thread ID: 283958 also declined drastically (Fig XIII). The peak of anger of the thread IDs: 218736, 530558 at 2006 and 2010 respectively (Fig XIII) corresponds to the polarity of the threads at the same point of time (Fig XI) as well following the significant correlation in section A.

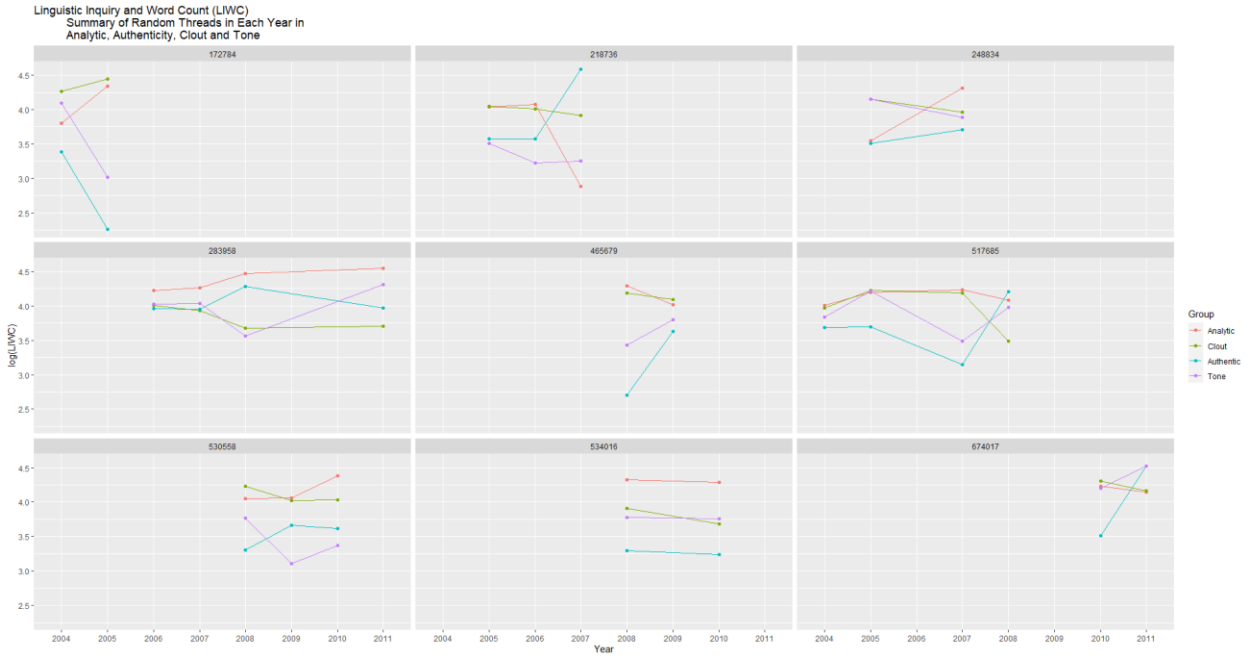


Figure VIIII LIWC Summary of Random Threads in Each Year in Analytic, Authenticity, Clout and Tone (Appendix IV)

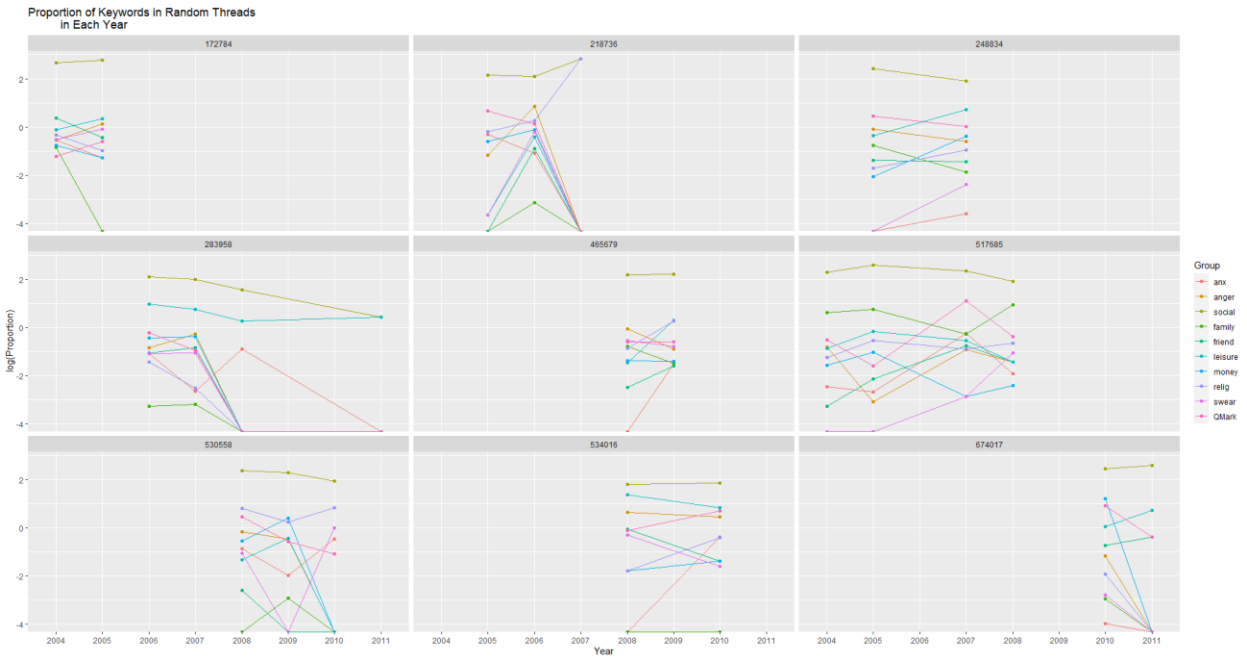


Figure IXIII LIWC Summary of Random Threads in Each Year in Keywords (Appendix V)

Furthermore, the observations in Figure XIV on post distribution of each thread from year 2004 to 2011 hints on the conversations in the thread IDs: 248834, 283958, 517685, 674017 as emotionally positive throughout its active period. The thread IDs: 218736, 530558 are mostly emotionally negative whereas the rest of the thread balances between positive and negative in overall throughout their active period. Also, all threads obviously have more users engaging at the start of activity but later decreases.

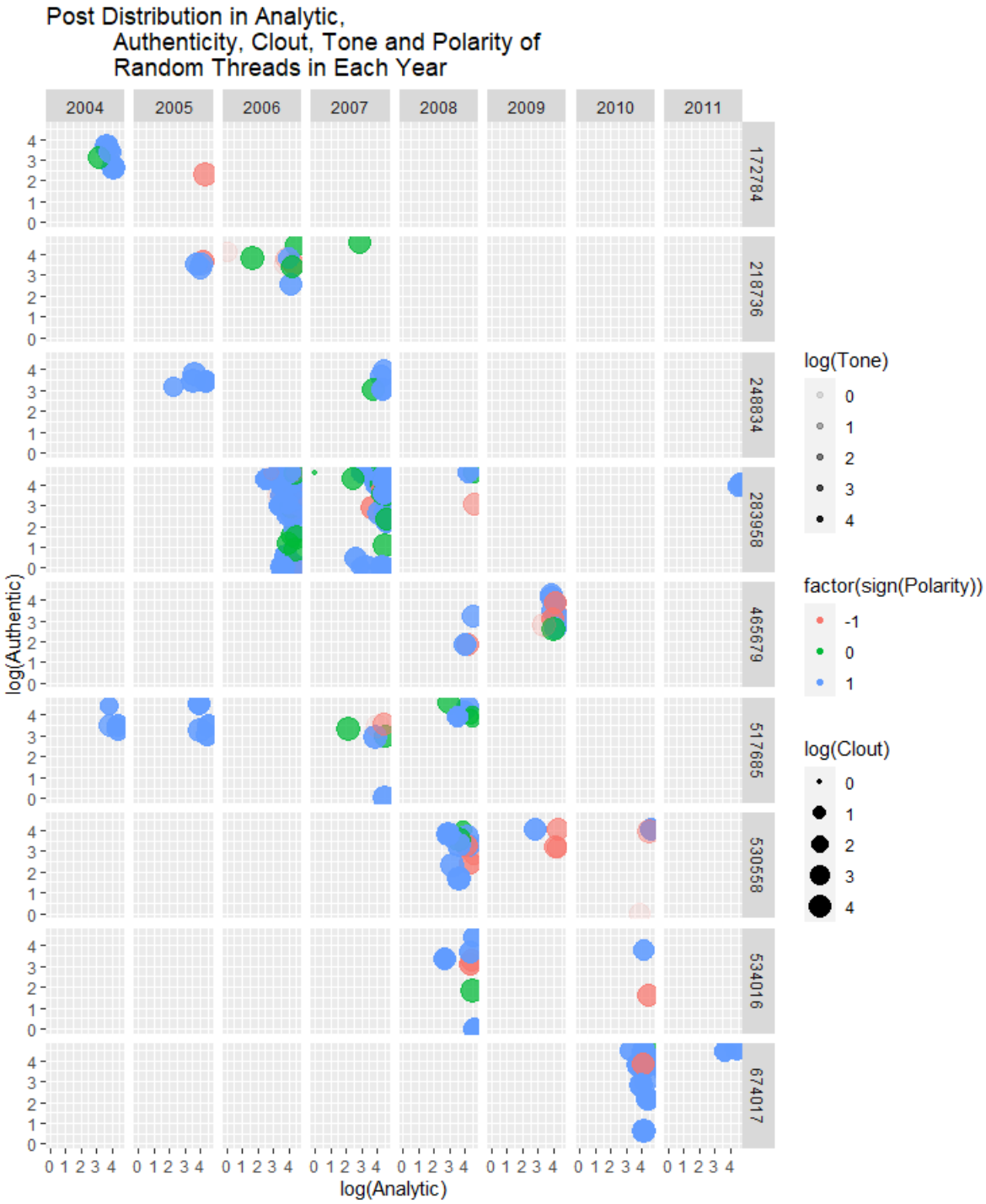


Figure XIV Post Distribution in Analytic, Authenticity, Clout and Tone and Polarity of Random Threads in Each Year (Appendix VI)

C. Social Networks Online

Can you define, graph and describe the social network that exists at a point in time, for example over one month? How does this change in the following months?

Finally, Figure XV and Figure XVI below maps the social network formed in the year 2007 and 2008 respectively. There are too many threads to consider in each year so, three random threads are chosen for analysis. Active users in 2007 only communicated in the same thread and formed a social network locally (Fig XV) whereas active users in 2008 improved. The user with the author ID: 118148 extended his/her size of social network by participating in other threads aside from his/her current thread resulting in the highest number of mutual acquaintances (Fig XVI). Notice that the size of a local social network also grows when the number of participants in a particular thread increases.

D. Conclusion

In conclusion, the analysis using the information extracted from 20,000 sample of the forum data has successfully found at least three significant relationships: the AuthorID and ThreadID variable, the affect and posemo variable, the social and Clout variable and the anger and negemo variable. Note that these correlations were taken into consideration due to its high positive correlation coefficient. Following that, when one rockets, the other also does. The threads then have their own unique pattern of LIWC summary and keywords proportion throughout the years. Take thread ID: 283958 – the most discussed thread – declines in words referring to social processes in accord to the drop in power, force and impact. Its peak of anger at 2006 also corresponds to its low polarity. Nevertheless, it is still emotionally positive in overall as there are more optimistic posts. The users though may have extended their social network in the later years but they are also less active.

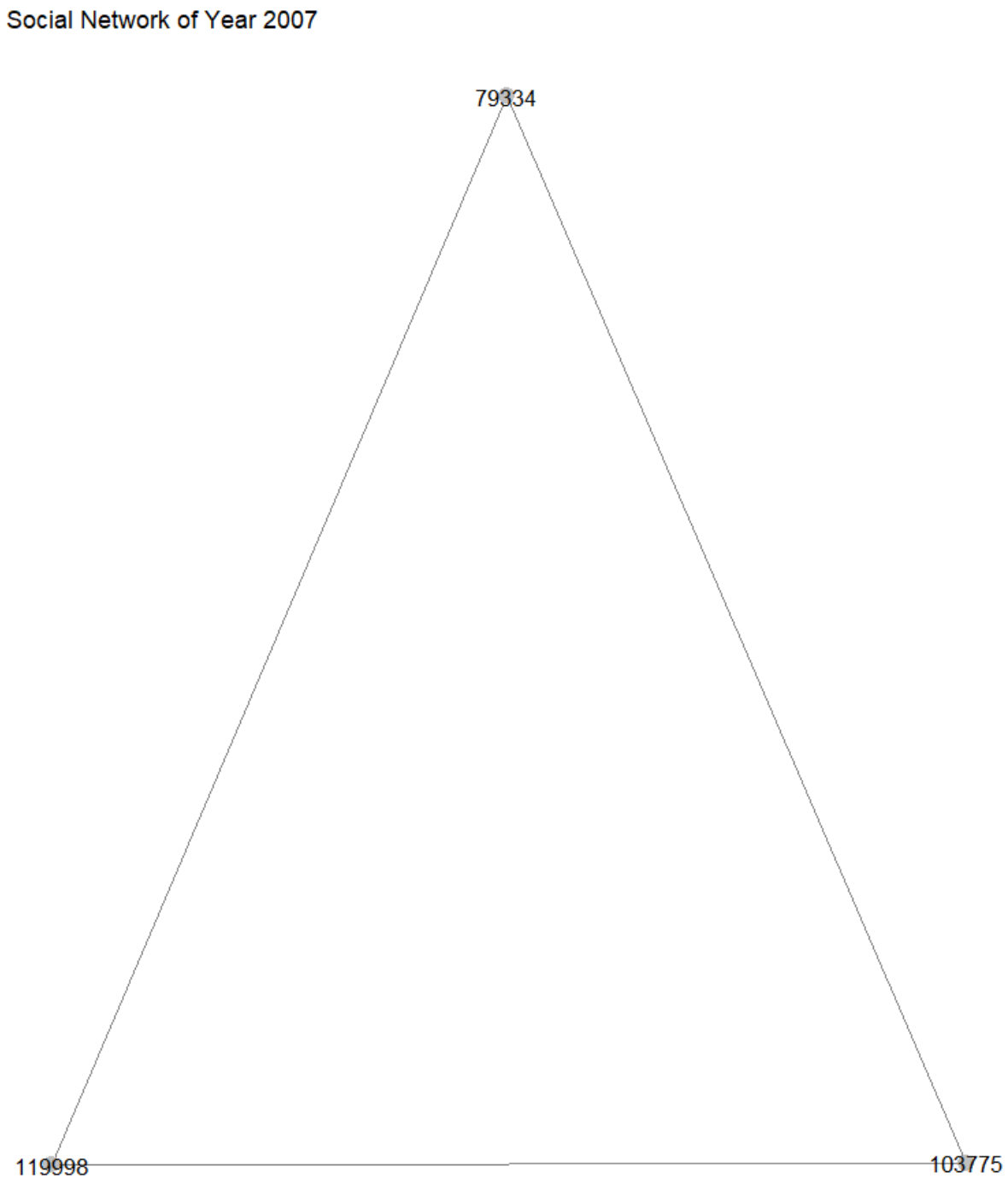


Figure XV Social Network of Year 2007

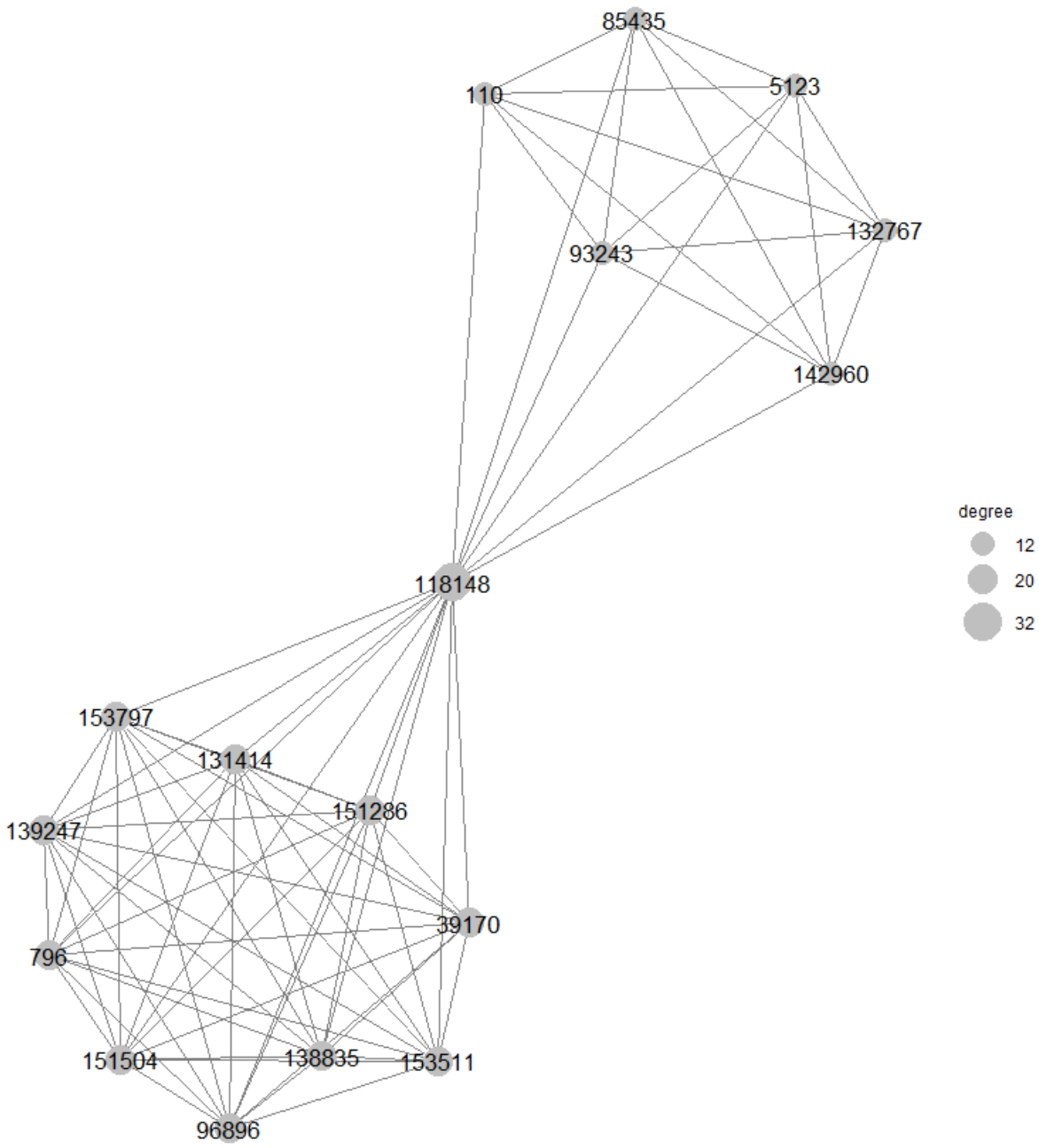


Figure XVI Social Network of Year 2008

References

Chan, D. M. (2009, June 11). *World now at the start of 2009 influenza pandemic*. Retrieved from World Health Organization: https://www.who.int/mediacentre/news/statements/2009/h1n1_pandemic_phase6_20090611/en/

Google. (2020, 5 10). *Google Trend*. Retrieved from Google Trend: https://trends.google.com/trends/explore?date=all&q=forum

Google. (2020, 5 10). *Google Trend*. Retrieved from Google Trend: https://trends.google.com/trends/explore?date=all&q=Facebook

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015.

Appendix

```
# Author: Priscilla Tham Ai Ching
# Student ID: 28390121
# Date: 26/04/2020

library("ggplot2")
library("ggnet")
library("network")
library("sna")

# adding connections from one vector to another
# in adjacent matrix
makeAdjMat <- function(adj_Mat, by_data, v_data) {
 count = 0
 for (i in 2:length(by_data)) {
  if (by_data[i-1] !=
    by_data[i]) {
   count = 0

  } else {
   head = which(
    users==v_data[i-1])
   tail = which(
    users==v_data[i])
   adj_Mat[head, tail] = 1
   adj_Mat[tail, head] = 1
   count = count + 1

   if (count > 1) {
    # adding edges to all the other vectors
    # in the same thread
    for (j in (i - count):(i-2)) {
     head = which(
      users==v_data[j])
     tail = which(
      users==v_data[i])
     adj_Mat[head, tail] = 1
     adj_Mat[tail, head] = 1
    }
   }
  }
 }
 return(adj_Mat)
}

# count the number of occurrences of names
# and convert it to a data frame
getFreq <- function(names, renames) {
 return(as.data.frame(table(names, dnn=renames)))
}

# perform the function fn on rows of names
# then group them by the variable by
groupBy <- function(names, by, data, fn) {
 return(aggregate(names, by=by, data=data, FUN=fn))
}

# stack the names by the variable by
meltDf <- function(by, names) {
 return(cbind(row.names=NULL, by, stack(names)))
}
```

```
# rename columns in data frames
rename <- function(data, columns, names) {
 colnames(data)[columns] = names
 return(data)
}


# find the maximum in name and return its position
findMax <- function(name) {
 return(which.max(name))
}


set.seed(28390121) #XXXXXXXX = your student ID
forum_data = read.csv(file="webforum.csv",
                      header=TRUE)
forum_data = forum_data[
 sample(nrow(forum_data), 20000),
 ] # 20000 rows

# remove non-text post and anonymous user
forum_data = forum_data[
 forum_data$AuthorID!=-1 & forum_data$WC!=0, ]

# Question A
# count occurrences of each author ID
user_post_count = getFreq(forum_data$AuthorID,
                  c("AuthorID"))
most_active_user = user_post_count$AuthorID[
 findMax(user_post_count$Freq)
]

# count occurrences of each thread ID
thread_post_count = getFreq(forum_data$ThreadID,
                            c("ThreadID"))
most_active_thread = thread_post_count$ThreadID[
 findMax(thread_post_count$Freq)
]

# count occurrences of each author ID's
# each thread ID
each_user_thread_post_count = getFreq(
 list(forum_data$AuthorID, forum_data$ThreadID),
 c("AuthorID", "ThreadID"))

# threads of author ID = most active user author ID
most_active_user_threads =
 each_user_thread_post_count[
  which(each_user_thread_post_count$AuthorID==
        most_active_user), ]
most_active_user_most_active_thread =
 most_active_user_threads$ThreadID[
  findMax(most_active_user_threads$Freq)
]

cor(forum_data$AuthorID, forum_data$ThreadID)

### hypothesis test on H0: the probability of most
# active thread being the most active user's most
# active thread <= 60%
# H1: the probability of most active thread being
# the most active user's most active thread > 60%
other_users_threads =
 each_user_thread_post_count[
  which(each_user_thread_post_count$AuthorID!=
        most_active_user), ]
```

```
mean = mean(most_active_user_threads$Freq) -
 mean(other_users_threads$Freq)
approx_z_score = mean/sqrt(
 (var(most_active_user_threads$Freq)/
   length(most_active_user_threads$Freq)) +
  (var(other_users_threads$Freq)/
    length(other_users_threads$Freq)))
p_value = 2*pnorm(-abs(approx_z_score))
###

# adding the column Year to data
forum_data["Year"] = format(
 as.Date(forum_data$Date, format="%Y-%m-%d"), "%Y")

# count the occurrences of each year
yearly_post_count = getFreq(
 forum_data$Year, c("Year"))
ggplot(data=yearly_post_count, aes(
 x=Year, y=Freq, group=1)) +
 geom_line() +
 geom_point() +
 ggtitle("Post Frequency of Each Year")

LIWC = with(forum_data,
        cbind(Analytic, Clout, Authentic, Tone,
            ppron, i, we, you, shehe, they,
            affect, posemo, negemo))

yearly_LIWC = groupBy(
 LIWC, list(forum_data$Year), forum_data, "mean")

round(cor(forum_data[, c(17)],
       forum_data[, 18:19]), digits=4)

### hypothesis test on H0: the probability of
# positive sentiment expressed <= 80%
# H1: the probability of positive sentiment
# expressed is > 80%
mean = mean(yearly_LIWC$posemo) -
 mean(yearly_LIWC$affect)
approx_z_score = mean/sqrt(
 (var(yearly_LIWC$posemo)/
   length(yearly_LIWC$posemo)) +
  (var(yearly_LIWC$affect)/
    length(yearly_LIWC$affect)))
p_value = 2*pnorm(-abs(approx_z_score))
###

yearly_sentiment_LIWC = meltDf(
 yearly_LIWC[1], yearly_LIWC[12:14])
yearly_sentiment_LIWC = rename(
 yearly_sentiment_LIWC, 1:3, c("Year", "LIWC",
                "Group"))
ggplot(data=yearly_sentiment_LIWC, aes(
 x=Year, y=log(LIWC), group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 ggtitle("Linguistic Inquiry and Word Count (LIWC)
      Summary of Each Year in Affect, Positive
      and Negative Emotions")

yearly_voice_LIWC = meltDf(
 yearly_LIWC[1], yearly_LIWC[2:5])
yearly_voice_LIWC = rename(
 yearly_voice_LIWC, 1:3, c("Year", "LIWC",
                "Group"))
```

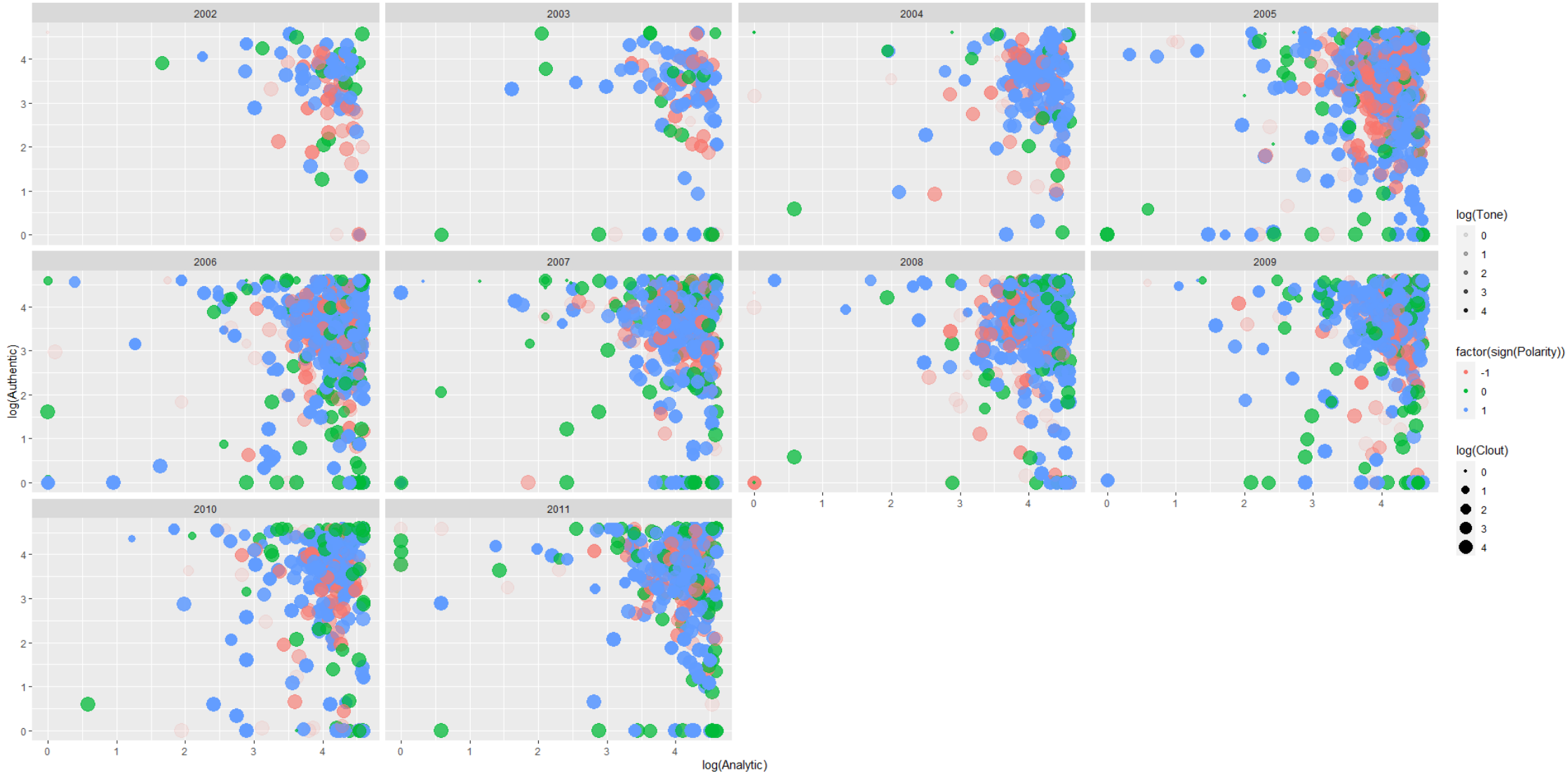
```
ggplot(data=yearly_voice_LIWC, aes(
 x=Year, y=log(LIWC), group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 ggtitle("Linguistic Inquiry and Word Count (LIWC)
       Summary of Each Year in Analytic,
       Authenticity, Clout and Tone")

yearly_pronouns_LIWC = meltDf(
 yearly_LIWC[1], yearly_LIWC[6:11])
yearly_pronouns_LIWC = rename(
 yearly_pronouns_LIWC, 1:3, c("Year", "LIWC",
                      "Group"))
ggplot(data=yearly_pronouns_LIWC, aes(
 x=Year, y=log(LIWC), group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 ggtitle("Linguistic Inquiry and Word Count (LIWC)
       Summary of Pronouns in Each Year")

keywords = with(forum_data,
           cbind(anx, anger, social, family,
               friend, leisure, money,
               relig, swear, QMark))

cor(forum_data$social, forum_data$Clout)

### hypothesis test on H0: the probability of a
# post content referring to social processes
# has higher Clout values <= 60%
# H1: the probability of post content referring to
# social processes has higher Clout values > 60%
mean = mean(forum_data$social) -
 mean(forum_data$Clout)
approx_z_score = mean/sqrt(
 (var(forum_data$social)/
   length(forum_data$social)) +
  (var(forum_data$Clout)/
    length(forum_data$Clout)))
p_value = 2*pnorm(-abs(approx_z_score))
###

cor(forum_data$anger, forum_data$negemo)

### hypothesis test on H0: the probability of
# anger referred words relating to negative
# emotions is <= 60%
# H1: the probability of anger referred words
# relating to negative emotions is > 60%
mean = mean(forum_data$anger) -
 mean(forum_data$negemo)
approx_z_score = mean/sqrt(
 (var(forum_data$anger)/
   length(forum_data$anger)) +
  (var(forum_data$negemo)/
    length(forum_data$negemo)))
p_value = 2*pnorm(-abs(approx_z_score))
###

yearly_keywords = groupBy(
 keywords, list(forum_data$Year),
 data=forum_data, "mean")
yearly_keywords = meltDf(
 yearly_keywords[1], yearly_keywords[-1])
yearly_keywords = rename(
 yearly_keywords, 1:3, c("Year", "Proportion",
```

```
                "Group"))
ggplot(data=yearly_keywords, aes(
 x=Year, y=log(Proportion),
 group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 ggtitle("Proportion of Keywords in Each Year")

# adding the column Polarity to data
forum_data["Polarity"] = forum_data$posemo -
 forum_data$negemo

user_yearly_LIWC = with(forum_data, groupBy(
 cbind(LIWC, Polarity), list(Year, AuthorID),
 forum_data, "mean"))
user_yearly_LIWC = rename(
 user_yearly_LIWC, 1:2, c("Year", "AuthorID"))
qplot(data=user_yearly_LIWC, facets=~Year,
    x=log(Analytic), y=log(Authentic),
    size=log(Clout), alpha=log(Tone),
    color=factor(sign(Polarity))) +
 ggtitle("User Distribution in Analytic,
      Authenticity, Clout, Tone and Polarity in
      Each Year")
#####

#Question B
# count the occurrences of each thread in each year
yearly_thread_post_count = getFreq(
 list(forum_data$ThreadID, forum_data$Year),
 c("ThreadID", "Year"))
yearly_thread_post_count = yearly_thread_post_count[
 yearly_thread_post_count$Freq!=0,]
# count the occurrences of each thread in
# yearly_thread_post_count
id_occurences = getFreq(
 yearly_thread_post_count$ThreadID, c("ThreadID"))

set.seed(28390121)
# sample 9 nine threads from
# yearly_thread_post_count that occur in more than
# one year
sample_ids = sample((id_occurences[
 id_occurences$Freq>1,])$ThreadID, 9)

thread_yearly_LIWC = with(forum_data, groupBy(
 cbind(LIWC, Polarity), list(Year, ThreadID),
 forum_data, "mean"))
thread_yearly_LIWC = rename(
 thread_yearly_LIWC, 1:2, c("Year", "ThreadID"))

# get the nine threads' sentiment LIWC in each year
thread_yearly_sentiment_LIWC = thread_yearly_LIWC[
 thread_yearly_LIWC$ThreadID %in% sample_ids, ]
ggplot(data=thread_yearly_sentiment_LIWC, aes(
 x=Year, y=log(affect), group=1)) +
 geom_line() +
 geom_point(aes(color=factor(sign(Polarity)),
          size=Polarity)) +
 facet_wrap(~ThreadID) +
 ggtitle("Linguistic Inquiry and Word Count (LIWC)
      Summary of Random Threads in Each Year in
      Affect and Polarity")

thread_yearly_voice_LIWC = meltDf(
```

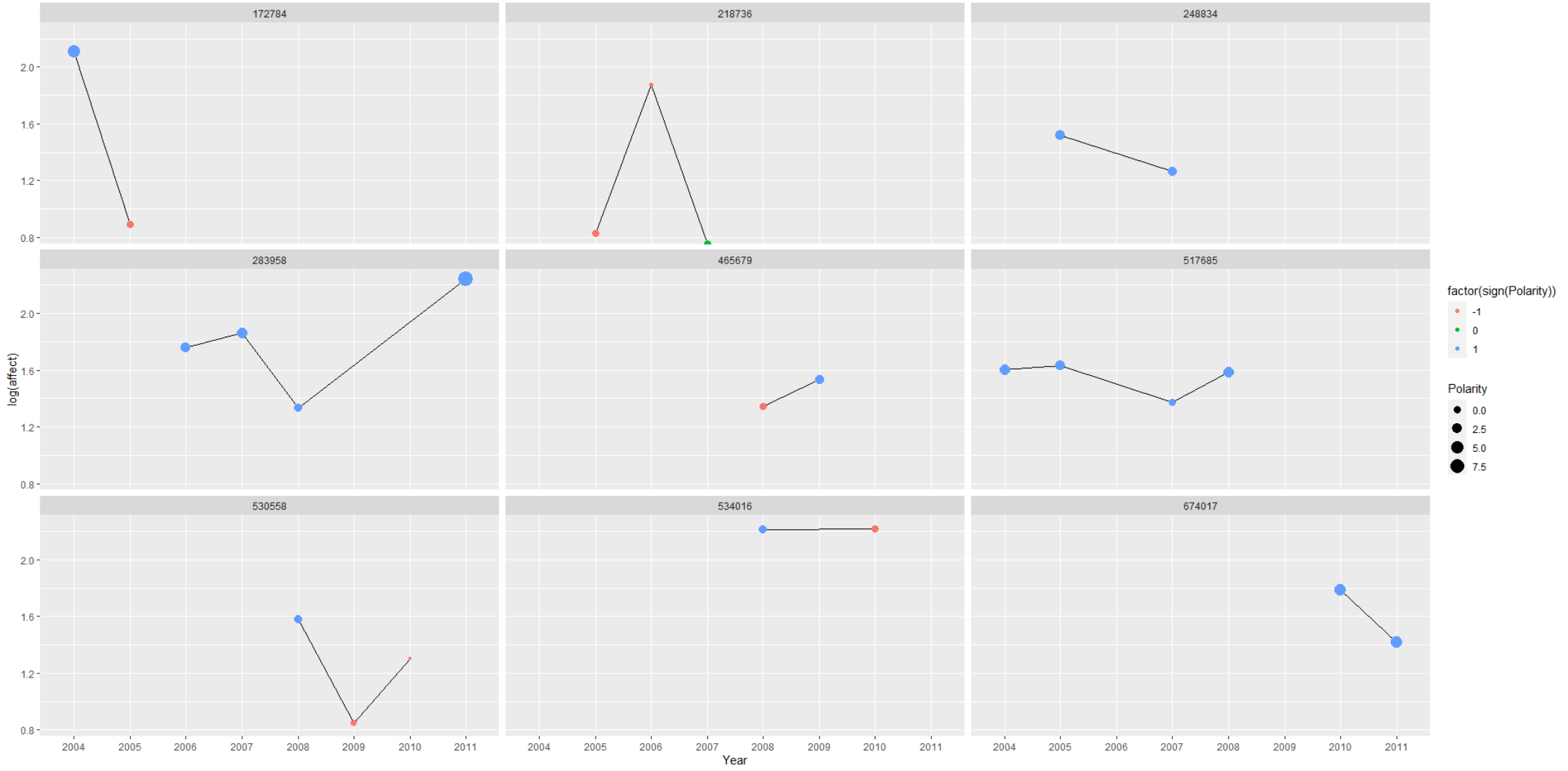
```
 thread_yearly_LIWC[1:2], thread_yearly_LIWC[3:6])
thread_yearly_voice_LIWC = rename(
 thread_yearly_voice_LIWC, 3:4, c("LIWC", "Group"))

# get the nine threads' voice LIWC in each year
thread_yearly_voice_LIWC = thread_yearly_voice_LIWC[
 thread_yearly_voice_LIWC$ThreadID %in%
   sample_ids, ]
ggplot(data=thread_yearly_voice_LIWC, aes(
 x=Year, y=log(LIWC), group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 facet_wrap(~ThreadID) +
 ggtitle("Linguistic Inquiry and Word Count (LIWC)
       Summary of Random Threads in Each Year in
       Analytic, Authenticity, Clout and Tone")

thread_yearly_keywords = with(forum_data, groupBy(
 keywords, list(Year, ThreadID),
 data=forum_data, "mean"))
thread_yearly_keywords = meltDf(
 thread_yearly_keywords[1:2],
 thread_yearly_keywords[
   3:ncol(thread_yearly_keywords)])
thread_yearly_keywords = rename(
 thread_yearly_keywords, 1:4,
 c("Year", "ThreadID", "Proportion", "Group"))

# get the nine threads' proportion of keywords in
# each year
thread_yearly_keywords =
 thread_yearly_keywords[
   thread_yearly_keywords$ThreadID %in%
     sample_ids, ]
ggplot(data=thread_yearly_keywords, aes(
 x=Year, y=log(Proportion),
 group=Group, color=Group)) +
 geom_line() +
 geom_point() +
 facet_wrap(~ThreadID) +
 ggtitle("Proportion of Keywords in Random Threads
       in Each Year")

thread_daily_LIWC = with(forum_data, groupBy(
 cbind(LIWC, Polarity),
 list(Year, ThreadID, Date), forum_data, "mean"))
thread_daily_LIWC = rename(
 thread_daily_LIWC, 1:3,
 c("Year", "ThreadID", "Date"))

# get the nine threads' LIWC in each day of
# each year
thread_daily_LIWC = thread_daily_LIWC[
 thread_daily_LIWC$ThreadID %in% sample_ids, ]
qplot(data=thread_daily_LIWC, facets=ThreadID~Year,
    x=log(Analytic), y=log(Authentic),
    size=log(Clout), alpha=log(Tone),
    color=factor(sign(Polarity))) +
 ggtitle("Post Distribution in Analytic,
       Authenticity, Clout, Tone and Polarity of
       Random Threads in Each Year")
#####

#Question C
# count the occurrences of each author ID in
# each thread ID in each year
```

```
yearly_thread_users = with(forum_data, getFreq(
 list(Year, ThreadID, AuthorID),
 c("Year", "ThreadID", "AuthorID")))
yearly_thread_users = yearly_thread_users[
 yearly_thread_users$Freq>1, ]

set.seed(28390121)
# sample 3 threads from yearly_thread_users
sample_threads = sample(
 yearly_thread_users[duplicated(
    yearly_thread_users$ThreadID), ]$ThreadID, 3)

# get the 3 threads' users in each year
yearly_thread_users = yearly_thread_users[
 yearly_thread_users$ThreadID %in%
   sample_threads, ]
# order them by thread ID so edges are added to
# users in the same thread only
yearly_thread_users = yearly_thread_users[
 order(yearly_thread_users$ThreadID), ]

# get the thread and users in 2007
thread_users_2007 = yearly_thread_users[
 yearly_thread_users$Year=="2007", ]
# get the unique users in thread_users_2007
# and order them so the author ID are arranged in
# the same order in the adjacent matrix. then
# the position of each author ID corresponds
# to its position in the adjacent matrix
users = unique((thread_users_2007[
 order(thread_users_2007$AuthorID), ])$AuthorID)
total_users = length(users)
adj_Mat = matrix(0L, nrow=total_users,
                 ncol=total_users)
adj_Mat = makeAdjMat(
 adj_Mat,
 thread_users_2007$ThreadID,
 thread_users_2007$AuthorID)

net_2007 = network(adj_Mat, matrix.type="adjacency")
ggnet2(net_2007, size="degree", label=users) +
 ggtitle("Social Network of Year 2007")

# get the thread and users in 2008
thread_users_2008 = yearly_thread_users[
 yearly_thread_users$Year=="2008", ]
# get the unique users in thread_users_2008
# and order them so the author ID are arranged in
# the same order in the adjacent matrix. then
# the position of each author ID corresponds
# to its position in the adjacent matrix
users = unique((thread_users_2008[
 order(thread_users_2008$AuthorID), ])$AuthorID)
total_users = length(users)
adj_Mat = matrix(0L, nrow=total_users,
                 ncol=total_users)
adj_Mat = makeAdjMat(
 adj_Mat,
 thread_users_2008$ThreadID,
 thread_users_2008$AuthorID)

net_2008 = network(adj_Mat, matrix.type="adjacency")
ggnet2(net_2008, size="degree", label=users) +
 ggtitle("Social Network of Year 2008")

#####
```

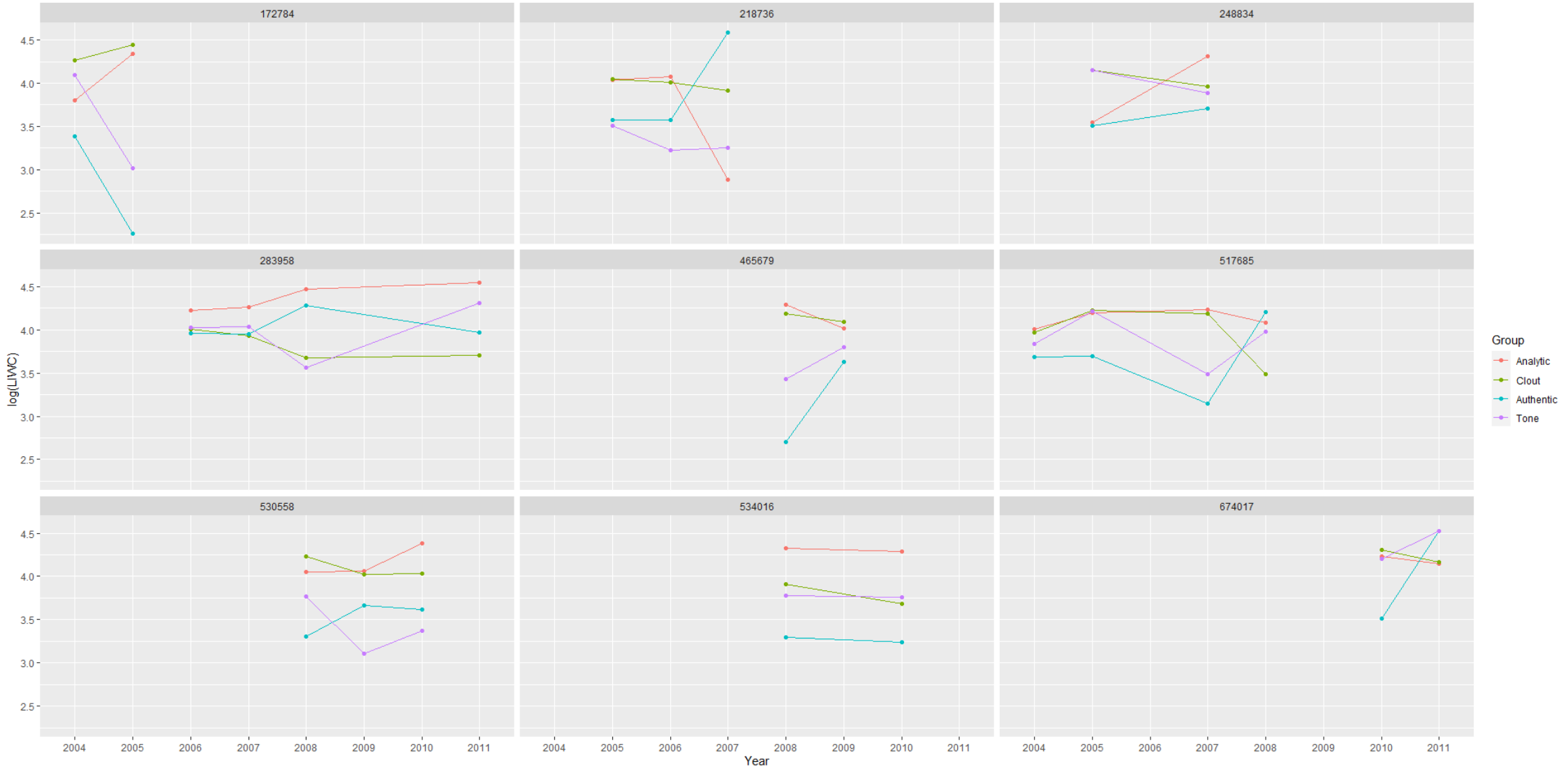
User Distribution in Analytic, Authenticity, Clout, Tone and Polarity in Each Year

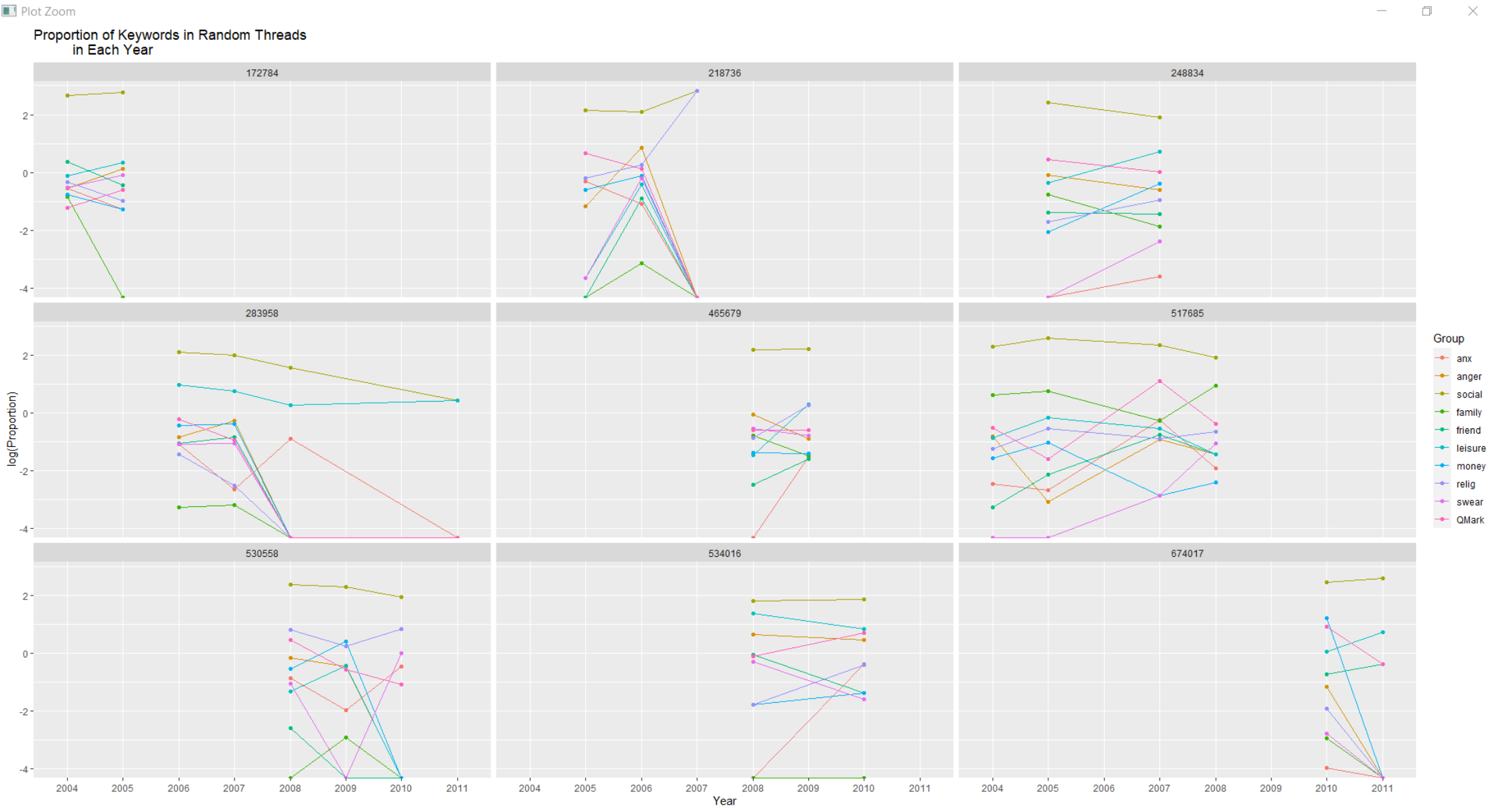


Linguistic Inquiry and Word Count (LIWC)
Summary of Random Threads in Each Year in
Affect and Polarity



Linguistic Inquiry and Word Count (LIWC)
Summary of Random Threads in Each Year in
Analytic, Authenticity, Clout and Tone




Plot Zoom
Proportion of Keywords in Random Threads
in Each Year
172784
218736
248834
283958
465679
517685
530558
534016
674017
log(Proportion)
Year
Group
anx
anger
social
family
friend
leisure
money
relig
swear
QMark


Plot Zoom
Post Distribution in Analytic,
Authenticity, Clout, Tone and Polarity of
Random Threads in Each Year
2004
2005
2006
2007
2008
2009
2010
2011
172784
218736
248834
283958
465679
517685
530558
534016
674017
log(Authentic)
log(Analytic)
log(Tone)
0
1
2
3
4
factor(sign(Polarity))
-1
0
1
log(Clout)
0
1
2
3
4