CHAPTER 3

c0015

Discovery of Emergent Issues and Controversies in Anthropology Using Text Mining, Topic Modeling, and Social Network Analysis of Microblog Content

Ben Marwick
Department of Anthropology, University of Washington, Seattle, USA

s0010 3.1 Introduction

p0010 The aim of this chapter is to show some basic methods using R to analyze text content to discover emergent issues and controversies in diverse corpora. As a specific case study, I investigate the culture of microblogging academics within the dynamics of a professional conference to gain insights into the key issues and debates emergent in this community and the transformative effects of using Twitter in academic contexts. Microblogging academics can be considered a type of online community which has its own norms, rules, and communicative behaviors (Gruzd et al., 2011) that can be analyzed with anthropological methods (cf. Boellstorff, 2011; Wilson and Peterson, 2002). My hypothesis is that data mining the publically available microblog text content generated in relation to the 109th Annual Meeting of the American Anthropological Association (AAA) in November 2011 can reveal the main issues and controversies that characterized the event as well as the community structure of the people generating the corpus. Although the duration of the meeting represents a narrow slice of Twitter content, it is ideal for looking at which academics are tweeting and why they tweet because academic meetings are a period of highly concentrated intellectual and social activity within the academic community. It is during these times that the distinctive patterns of shared learned knowledge, behaviors, and beliefs that characterize communities are most apparent (Egri, 1992). It is hoped that the methods presented will be suitable for the analysis of a wide variety of communities that generate large amounts of text content.

p0015 There are a number of unique and eventful characteristics of the 2011 meeting that make the related Twitter content especially worthy of investigation. These include organizational issues

Data Mining Applications with R

such as the session was convened in response to controversy surrounding the removal of the word "science" from the AAA's long-range plan statement in 2010 (Boellstorff, 2011; Lende, 2011), the AAA Presidential Address that discussed the 2010 final report of the Commission on Race and Racism in Anthropology, and the revision of the AAA code of ethics. Beyond these major organizational topics, other issues that were prominent at the time of the meeting were the Occupy movement and the future of scholarly publishing. Analysis of the Twitter messages relating to these issues gives insights into the behavior of microblogging anthropologists and their fit within the structure and culture of the discipline. Since Twitter postings are highly accessible to the public, this chapter also reveals the potential of evaluating how anthropologists use Twitter as a public face of the discipline.

p0020 Among academics in general, Twitter use is relatively rare with Priem et al. (2011) finding that about 2.5% of 8826 scholars at five U.K. and U.S. universities used Twitter weekly. Priem et al. found that no academic rank or discipline was significantly overrepresented in their sample. They also noted that although Twitter is popular as a scholarly medium for making announcements, linking to articles, and engaging in discussions about methods and literature, about 60% of the messages were personal. The use of Twitter at academic conferences has also been the subject of a number of systematic analyses, mostly aiming to identify how Twitter is used in this context and who benefits from it (Ebner, 2009; Ebner and Reinhardt, 2009; Ebner et al., 2010; Letierce et al., 2010a; McCarthy and Boyd, 2005; Reinhardt et al., 2009; Ross et al., 2011). These previous studies, summarized in Table 3.1, show that microblog content from conferences can be a corpus of substantial size comprising a large number of very short documents.

t0010 **Table 3.1 Summary of Related Previous Research on Microblogging of Academic Meetings**

Conference	Conference Attendees (*n*)	Authors (*n* [%])	Messages (*n*)	Source
EduCamp 2010	NA	272	2110	Ebner et al. (2010)
International Semantic Web Conference 2009	405	273 [67]	1444	Letierce et al. (2010a)
Digital Humanities 2009, That Camp 2009, Digital Resources in the Arts and Humanities 2009	542	379 [70]	4574	Ross et al. (2011)
ED-MEDIA 2009	1000	173 [17.3]	1595	Ebner and Reinhardt (2009)
ED-MEDIA 2008	1000	10 [10]	54	Ebner (2009)

The number of authors as a percentage of attendees is included in square brackets.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:13 Page Number: 72

s0015 3.2 How Many Messages and How Many Twitter-Users in the Sample?

p0025 To obtain raw data for this study, I searched the Twitter Web site (cf. Gentry, 2011) and downloaded 1500 messages that had been labeled by each message's author as relevant to the 109th Annual Meeting of the AAA (1500 messages is the maximum number of messages that the Twitter application programming interface (API) allows to download at one time). Authors of Twitter messages frequently use a shared system of notation for identifying the subject of their messages where a hash symbol is placed before the topic word or phrase (Kwak et al., 2010). In this case, the #aaa2011 hashtag was the subject identifier, so I extracted all messages containing this hashtag as follows

```
# get package with functions for interacting with Twitter.com
require(twitteR)
# get 1500 tweets with #aaa2011 tag, note that 1500 is the max, and it's subject to filtering and
other restrictions by Twitter
aaa2011 <- searchTwitter('#aaa2011', n=1500)
# convert to data frame
df <- do.call("rbind", lapply(aaa2011, as.data.frame))
# get column names to see structure of the data
names(df)
# look at the first three rows to check content
head(df,3)
# see how many unique Twitter accounts in the sample
length(unique(df$screenName))
```

p0090 In this sample, there are 307 authors whose messages span from 11:00 am EST on 17 until 6:00 pm EST on 20 November 2011. Although the #AAA2011 hashtag was in use in the weeks leading up to the meeting and continued to be used after the meeting concluded, I chose to limit the sample to those produced only during the course of the meeting. There are two reasons for this strategy. First, the scope of this study is limited to a synchronic analysis of Twitter use at the meeting as a time of intensive intellectual and social activity among anthropologists. This is a sampling strategy that has become standard in research on Twitter use in academic and professional contexts because it is a time when people are highly active on Twitter (Ebner, 2009; Ebner and Reinhardt, 2009; Ebner et al., 2010; Letierce et al., 2010a,b; Reinhardt et al., 2009). Second, the Twitter Web site is not explicit about how it makes messages publically available, so it is not always clear if Twitter reveals only a sample of messages matching the hashtag or the entire set of matching messages. During repeated sampling of the Twitter archives, I found I could only obtain a reproducible sample of messages for the period of the meeting, excluding the first day. For the days before the meeting, I was not confident that the archive was making available all the relevant messages. Furthermore, the number of messages in each sample declined with increased time after the event to the point where a few months after the event there were no Twitter messages with the #AAA2011 hashtag. This limitation unfortunately excludes the possibility of tracking which issues generated discussion beyond the meeting and whether Twitter posts during the meeting could predict the staying power of

particular topics (cf. Ebner and Reinhardt, 2009; Reinhardt et al., 2009). The unpredictable nature of the results obtained from the Twitter Web site necessitated the exclusion of days for which I could not obtain a reproducible number of messages and more broadly is a serious limitation on the reproducibility of analyses of Twitter corpora.

s0020 3.3 Who Is Writing All These Twitter Messages?

p0095 Although all the Twitter messages used in this study were publically available at the time the sample was collected, Twitter-users can hide all of their messages at any time, so for the rest of the analysis I have anonymized individual authors here to preserve their confidentiality. The authors include individual and corporate authors (such as the AAA, The Society for the Anthropology of Food and Nutrition, and Wiley-Blackwell). About half of all individual authors in the sample use pseudonyms. The degree of anonymity of the pseudonyms varied greatly. Some authors used a cryptic username unique to their Twitter account with no implied biographical information, giving absolute anonymity to the author. Some used a pseudonym on Twitter that was linked to their physical world self elsewhere on the Internet. Others used a username that could not be linked to a specific physical person, but implied a gender, academic status (e.g., graduate student, postdoctoral scholar, etc.), scholarly interests (e.g., bioanthropology, archeology, or medical anthropology) or some combination of the three.

```
# Create a new column of random numbers in place of the usernames and redraw the plots
# find out how many random numbers we need
n <- length(unique(df$screenName))
# generate a vector of random number to replace the names, we'll get four digits just for
convenience
randuser <- round(runif(n, 1000, 9999),0)
# match up a random number to a username
screenName <- unique(df$screenName)
screenName <- sapply(screenName, as.character)
randuser <- cbind(randuser, screenName)
# Now merge the random numbers with the rest of the Twitter data, and match up the correct random
numbers with multiple instances of the usernames...
rand.df <- merge(randuser, df, by="screenName")
```

p0155 The use of real names by some of the authors is notable because it links their professional identities as scholars to their authorship of their Twitter messages, giving them ownership of and accountability for their messages. This indicates a use of Twitter by some anthropologists as instrument of professional communication and makes these users visible as the informal public faces of the discipline to Twitter-users. The anonymity preferred by other anthropologists using Twitter is an indication of the heterogeneity of the Twitter-using community and the existence of individuals who prefer to maintain varying degrees of separation between their identity as a Twitter author and other dimensions of their identity (e.g., as a scholar or a student).

p0160 Of the 233 authors with a gender-identifying Twitter username (i.e., excluding unrevealing and ambiguous usernames), 128 (55%) identified as female and 104 (45%) as male. Among the 128 authors who provided enough information to determine their academic status, about half of these are graduate students (66, 49%). The next most represented group is faculty at the rank of assistant professors (23, 17%) followed by people with sessional, fixed term appointments, or nontenure track teaching appointments (14, 10%). The remainder is made up of associate professors (11, 8%), full professors (9, 7%), community college faculty (6, 5%), postdoctoral scholars (5, 3%), and undergraduates (2, 1%). In terms of academic status, it seems reasonable to conclude that more junior members of the discipline are most frequently represented on Twitter. This subset may be analogous to Prensky's (2001) "digital natives" or people whose upbringing was immersed in information and communication technologies, although the presence of more senior academics suggests a mixed group with a range of exposures to technology. Although specific ages for the authors are unavailable for this sample, the relatively small proportion of full professors relative to assistant professors and graduate students suggests that younger scholars are more often users of this form of virtual communication than older ones.

s0025 3.4 Who Are the Influential Twitter-Users in This Sample?

p0165 Figure 3.1 shows the frequency distribution of messages per author in this sample. The distribution approximately follows a power law, consistent with previous observations of Twitter usage and other online and real-life cultural phenomena (Bentley et al., 2004; Letierce et al., 2010a). Figure 3.1 shows that the majority of the messages were authored by about half a dozen individuals (most of whom used their real names, which are not given here). Figure 3.1 also shows that the most prolific authors also tend to have their messages most frequently repeated or cited by other authors. This behavior is known as retweeting and allows messages to spread beyond the network of the original message's author. Whereas the observed motivations for retweeting are numerous and difficult to disentangle (Boyd et al., 2010), the effect of retweeting is to increase the spread of the message and in turn, the author's influence on other authors. In this sample, 451 messages (30%) are retweets, a figure consistent with samples of Twitter messages from other academic conferences, but substantially higher than the 3% observed in general Twitter data (Letierce et al., 2010b). This indicates that these authors are reading and retweeting widely among their network.

```
# determine the frequency of tweets per account
counts <- table(rand.df$randuser)
# create an ordered data frame for further manipulation and plotting
countsSort <- data.frame(user = unlist(dimnames(counts)), count = sort(counts, decreasing =
TRUE), row.names = NULL)
# create a subset of those who tweeted at least 5 times or more
countsSortSubset <- subset(countsSort,countsSort$count > 5)
```

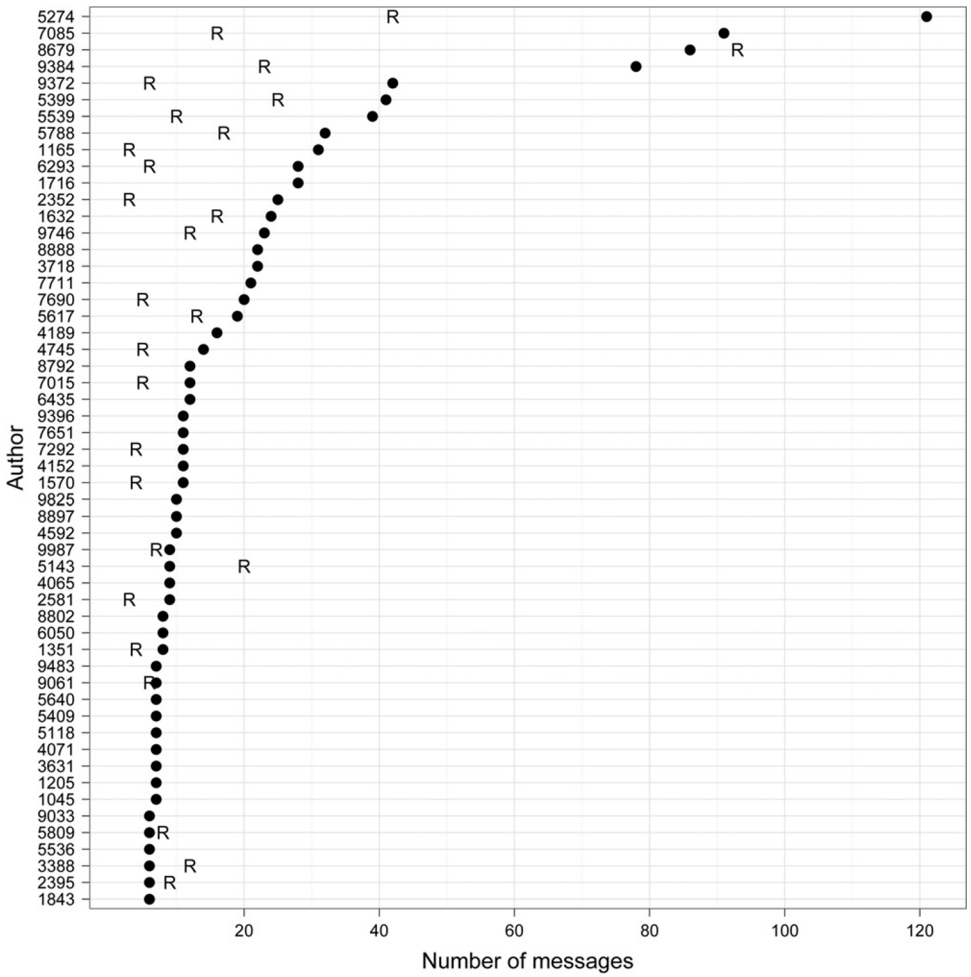

Figure 3.1

f0010 Number of messages by author, for all authors posting more than five messages (solid circle), and number of each author's messages that are repeated or cited by other authors, for all messages repeated or cited more than twice (indicated by "R").

p0200
```
# extract counts of how many tweets from each account were retweeted, this code is derived from
the excellent tutorial here http://heuristically.wordpress.com/2011/04/08/text-data-
mining-twitter-r/
# first clean the twitter messages by removing odd characters
rand.df$text=sapply(rand.df$text,function(row) iconv(row,to='UTF-8'))
# remove @ symbol from user names
trim <- function (x) sub('@','',x)
# pull out who the message is to
require(stringr)
rand.df$to <- sapply(rand.df$to,function(name) trim(name))
```

```
# extract who has been retweeted
rand.df$rt <- sapply(rand.df$text,function(tweet)
   trim(str_match(tweet,"^RT (@[[:alnum:]_]*)")[2]))
# replace names with corresponding anonymising number
randuser <- data.frame(randuser)
rand.df$rt.rand <
- as.character(randuser$randuser)[match(as.character(rand.df$rt),
                           as.character(randuser$screenName))]
# make a table with anonymised IDs and number of RTs for each account
countRT <- table(rand.df$rt.rand)
countRTSort <- sort(countRT)
# subset those people RT'd at least twice
countRTSortSubset <- subset(countRTSort,countRT>2)
# create a data frame for plotting
countRTSortSubset.df <-data.frame(people = as.factor(unlist(dimnames
(countRTSortSubset))), RT_count = as.numeric(unlist(countRTSortSubset)))
# combine tweet and retweet counts into one data frame
TweetRetweet <- merge(countsSortSubset, countRTSortSubset.df, all.x = TRUE)
# create a Cleveland dot plot of tweet counts and retweet counts per Twitter account
# solid data point = number of tweets, letter R = number of retweets
require(ggplot2)
require(grid)
ggplot() +
   geom_point(data = TweetRetweet, mapping = aes(reorder(people, count), count), size = 3) +
   geom_point(data = TweetRetweet, mapping = aes(people,
RT_count), size = 4, shape = "R") +
   xlab("Author") +
   ylab("Number of messages") +
   coord_flip() +
   theme_bw() +
   theme(axis.title.x = element_text vjust = -0.5, size = 14)) +
   theme(axis.title.y = element_text (size = 14, angle=90)) +
   theme(plot.margin = unit(c(1,1,2,2), "lines"))
```

p0390 Figure 3.2 shows that there are no clear correlations between number of messages, number of retweets, and number of followers (the number of other Twitter-users who subscribed to the messages of an author) but authors with high frequencies of messages and retweets tend to have a high number of followers (e.g., authors 8679, 5724, and 7085). There are two interesting implications from the follower data in Figure 3.2. First is that range in the number of followers is very wide, from 9 to 6979, indicating that the audiences of the authors vary from an audience of a small circle of close colleagues based on face-to-face relationships to an audience of a large number of people who might only know the author via Twitter messages. Second, the low correlation between the number of followers and the number of retweets (Kendall's $\tau = 0.28$, $p = 0.02$ from the R package "Kendall") suggests that the size of an author's Twitter audience is not a good predictor of how widely their messages are propagated.

```
# calculate the number of followers of each Twitter account
# extract the usernames from the non-anonymised dataset
users <- unique(df$screenName)
```

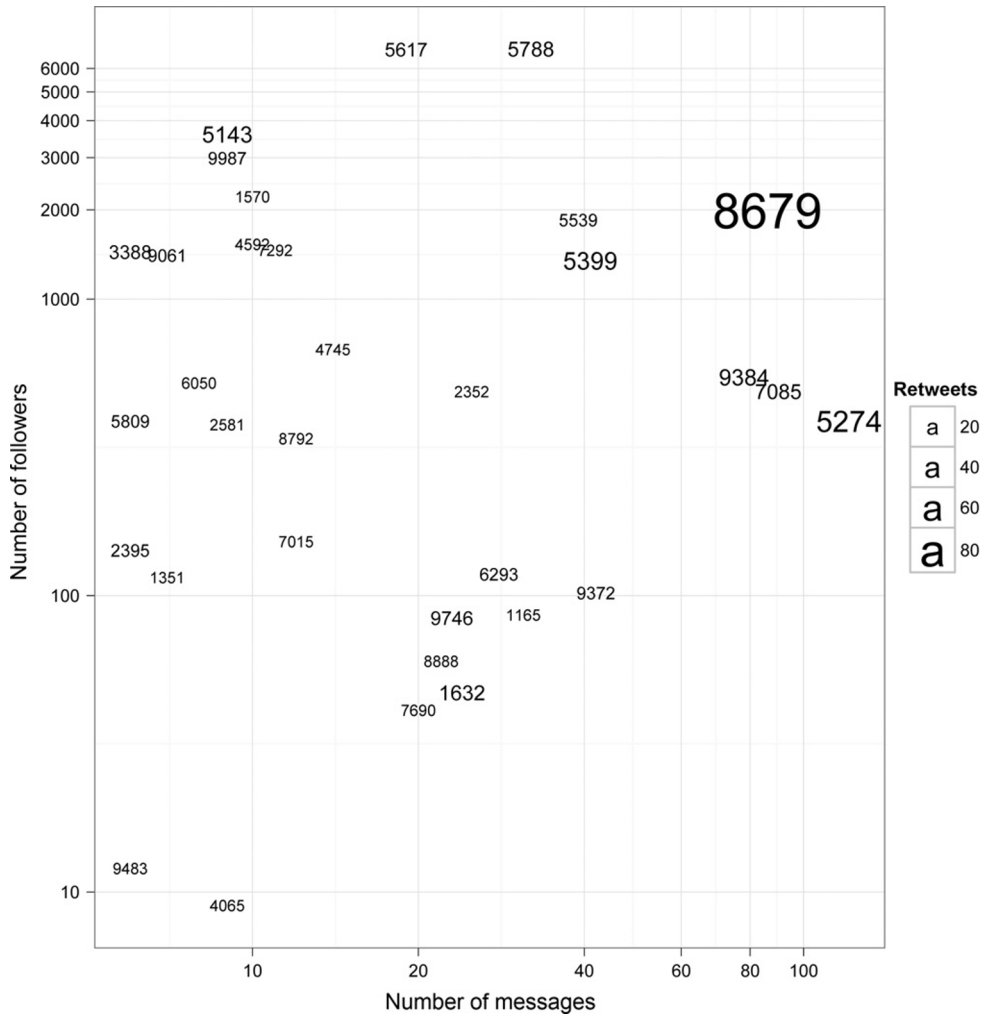

Figure 3.2

f0015 Plot of each author's number of followers on the Twitter network by the number of their messages in the corpus. The size of the author's identifying number indicates the frequency that their messages were retweeted.

```
users <- sapply(users, as.character)
# make a data frame for further manipulation
users.df <- data.frame(users = users, followers = "", stringsAsFactors = FALSE)
# loop to populate users$followers with a follower count obtained from Twitter API
for (i in 1:nrow(users.df))
{
   # tell the loop to skip a user if their account is protected
   # or some other error occurs
   result <- try(getUser(users.df$users[i])$followersCount,
silent = FALSE);
```


Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:14 Page Number: 78

```
   if(class(result) == "try-error") next;
   # get the number of followers for each user
   users.df$followers[i] <-
getUser(users.df$users[i])$followersCount
   # tell the loop to pause for 60 s between iterations to
   # avoid exceeding the Twitter API request limit
   # note that this is going to take a long
   # time if there are a lot of users the sample!
   print('Sleeping for 60 seconds...')
   Sys.sleep(60);
}
# merge follower count with number of tweets per author
followerCounts <- merge(TweetRetweet, users.df, by.x = "screenName", by.y = "users")
# convert to value to numeric for further analysis
followerCounts$followers <- as.numeric(followerCounts$followers)
followerCounts$counts <- as.numeric(followerCounts$counts)

# create a plot of number of followers by number of messages and number of retweets
ggplot(data = followerCounts, aes(count, followers)) +
   geom_text(aes(label = randuser, size = RT_count)) +
   scale_size(range=c(3,10)) +
   scale_x_log10(breaks = c(10,20,40,60,80,100)) +
   scale_y_log10(breaks = c(10,100,seq(1000,7000,1000))) +
   xlab("Number of Messages") +
   ylab("Number of Followers") +
   theme_bw() +
   theme(axis.title.x = element_text (vjust = -0.5, size = 14)) +
   theme(axis.title.y = element_text (size = 14, angle=90)) +
   theme(plot.margin = unit(c(1,1,2,2), "lines"))
```

p0590 Further insights into message propagation via retweets can be obtained from the ratio of retweets to original messages produced by each author (Figure 3.3). Authors with a retweet ratio greater than one had a higher number of their messages being retweeted by others than original messages, indicating a degree of influence and popularity that might not be predicted from the total number of messages they authored. It is remarkable to note that only one of the authors with a high retweet ratio is also among the most prolific (author 8679). Au2

```
# Make table with counts of tweets per person
t <- as.data.frame(table(rand.df$randuser))
# make table with counts of retweets per person
rt <- as.data.frame(table(rand.df$rt.rand))
# combine tweet count and retweet count per person
t.rt <- merge(t,rt, by = "Var1")
# creates new col and adds ratio tweet/retweet
t.rt["ratio"] <- t.rt$Freq.y / t.rt$Freq.x
# sort it to put names in order by ratio
sort.t.rt <- t.rt[order(t.rt$ratio), ]
# exclude those with 2 tweets or less
sort.t.rt.subset <- subset(sort.t.rt, sort.t.rt$Freq.x > 2)
#
```

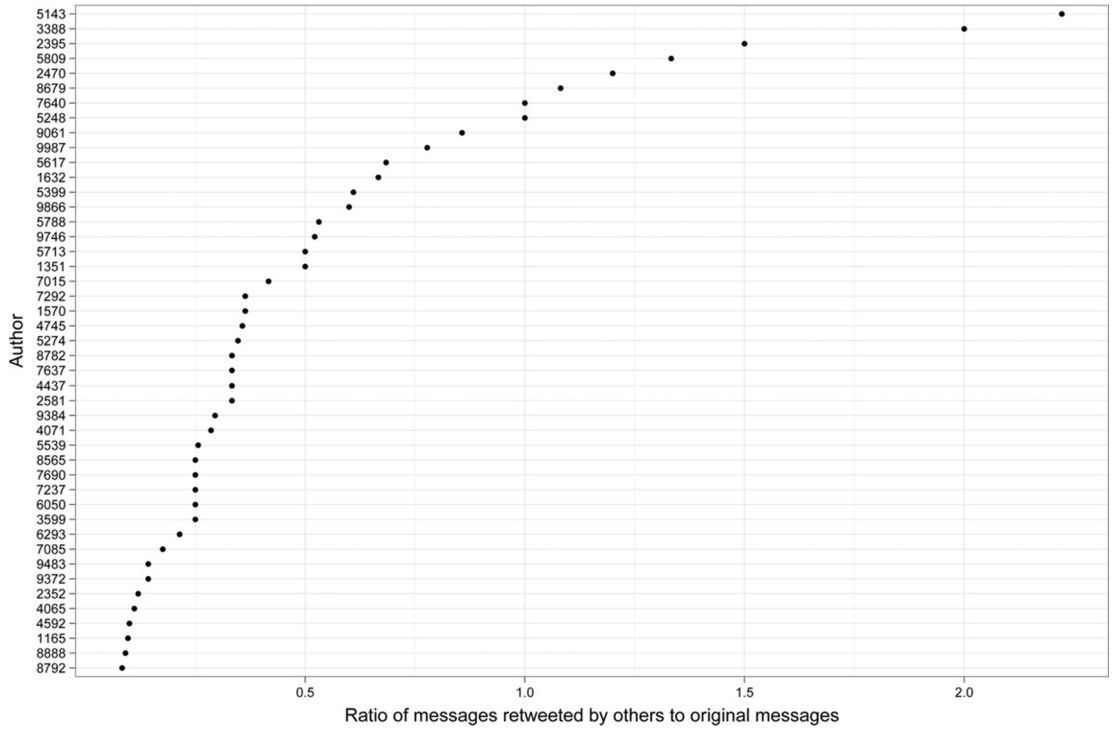

Figure 3.3

f0020 Ratio of retweeted messages to total messages by each author.

```
# drop unused levels leftover from subsetting
sort.t.rt.subset.drop <- droplevels(sort.t.rt.subset)
# plot nicely ordered counts of tweets by person for
# people > 5 tweets
ggplot(sort.t.rt.subset.drop, aes(reorder(Var1, ratio), ratio)) +
   xlab("Author") +
   ylab("Ratio of messages retweeted by others to original messages")+
   geom_point() +
   coord_flip() +
   theme_bw() +
   theme(axis.title.x = element_text(vjust = -0.5, size = 14)) +
   theme(axis.title.y = element_text(size = 14, angle=90)) +
   theme(plot.margin = unit(c(1,1,2,2), "lines"))
```

s0030 3.5 What Is the Community Structure of These Twitter-Users?

p0725 The degree of connectedness, and other related network properties of this community can be further explored by computing descriptive indices of the network graph resulting from the relationships contained in the retweet data (Butts, 2008a; Ye and Wu, 2010).

t0015 **Table 3.2 Summary of Graph-Level Social Network Indices**

Index	Value	Range of CUG Test Distribution	Interpretation
Density	0.012	0.492-0.508	Significantly fewer connections between community members than expected
Reciprocity	1.000	0.487-0.510	Significantly higher tendency of ties to be reciprocal rather than unidirectional
Transitivity	0.059	0.493-0.507	Significantly less instances of "a friend of a friend is a friend" than expected
Centralization	0.222	0.044-0.107	Significantly more centralized than expected

These indices provide succinct numerical summaries of the structure of the community that produced these messages. For each of these indices, a conditional uniform graph test can be undertaken to compare the observed index values against those which would be obtained by simulated data with known substantive properties similar to the data. The extent and direction of the deviation of the indices from their baseline distributions can create structural biases within the network, which may provide useful clues regarding the organization of the community (Butts, 2008b). Table 3.2 summarizes these graph indices and indicates that the community has distinctive structural properties, such as significantly fewer connections between individuals than expected, significantly higher tendency of reciprocal rather than unidirectional ties, significantly fewer triadic relationships than expected, and a significantly higher degree of centralization than expected. Some of these properties are also apparent in the network graph (Figure 3.4) which shows the network graph with a distinctive pattern of highly interconnected individuals in the center and many individuals who connect with only one highly connected individual.

```
# extract tweeter-retweeted pairs
rt <- data.frame(user=rand.df$randuser, rt=rand.df$rt.rand)
# omit pairs with NA and get only unique pairs
rt.u <- na.omit(unique(rt)) #
# begin social network analysis plotting
require(igraph)
require (sna)
degree <- sna::degree
g <- graph.data.frame(rt.u, directed = F)
g <- as.undirected(g)
g.adj <- get.adjacency(g)
# plot network graph
```

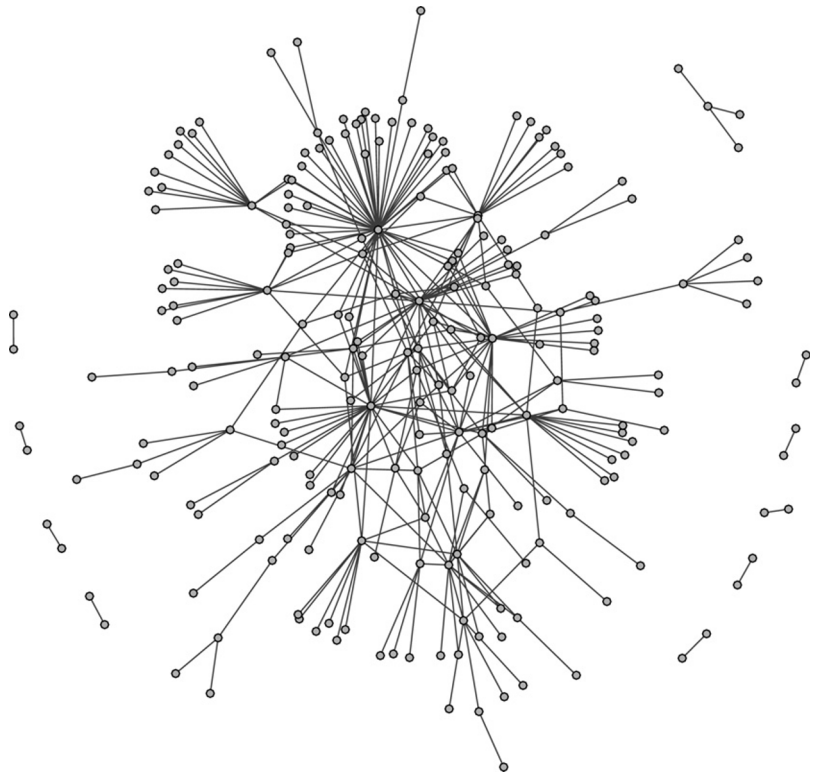

Figure 3.4

f0025 Visualization of the community of authors based on their retweeting behaviors.

```
gplot(g.adj, usearrows = FALSE,
        vertex.col = "grey",vertex.border = "black",
        displaylabels = FALSE, edge.lwd = 0.01, edge.col
         = "grey30", vertex.cex = 0.5)
# get some basic network attributes
gden(g.adj) # density
grecip(g.adj) # reciprocity
gtrans(g.adj) # transitivity
centralization(g.adj, degree)
# calculate Univariate Conditional Uniform Graph Tests
# density
print(cug.gden <- cug.test(g.adj, gden))
plot(cug.gden)
range(cug.gden$rep.stat)
# reciprocity
print(cug.recip <- cug.test(g.adj, grecip))
plot(cug.recip)
range(cug.recip$rep.stat)
# transistivity
```


Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:17 Page Number: 82

```
print(cug.gtrans <- cug.test(g.adj, gtrans))
plot(cug.gtrans)
range(cug.gtrans$rep.stat)
# centralisation
print(cug.cent <- cug.test(g.adj, centralization, FUN.arg=list(FUN=degree)))
plot(cug.cent)
range(cug.cent$rep.stat)

# find out how many communities exist in the network using the walktrap
g.wc <- walktrap.community(g, steps = 1000, modularity=TRUE, labels=TRUE)
plot(as.dendrogram(g.wc, use.modularity=TRUE))
max(g.wc$membership)+1
```

p0940 An inspection of the corpus of messages indicates that many of the highly connected individuals were broadcasting snippets of information from presentations they were attending. An example of a frequently retweeted message is "Would take two hours of moderate exercise daily to bring industrialized activity budget in line with subsistence [sic] activity budget," referring to a presentation in the "Scars of Evolution" session. These snippets were being retweeted by others who do not appear to have been at the same presentation but wanted to acknowledge their interest in the presentation and circulate the details through their network of Twitter contacts. Other types of frequently retweeted messages include links to weblog posts and news articles that comment on issues of the meeting (e.g., "What role should science play in #anthropology? http://t.co/4KylJaE1," referring to Jaschik (2011)) and observations and announcements about meeting events (e.g., "Undergrad student poster session! Come by and view the wonderful research by our future academics!"). Twenty-five cohesive groups of message writers and retweeters were identified in this sample using the walktrap community structure detection algorithm (Pons and Latapy, 2005).

s0035 3.6 What Were Twitter-Users Writing About During the Meeting?

p0945 Now I turn to some basic and widely used text mining techniques (Feinerer et al., 2008) to identify the issues that captured the attention of Twitter-using anthropologists during the meeting. To prepare for this analysis, I converted all text in the corpus to lower case, removed punctuation, numbers, and stopwords (words that occur very frequently due to their importance in sentence construction, for example, *is*, *and*, *the*) and stemmed words (removing morphological affixes such as -s, -ed, -ing, leaving only the stem of the word so that there is a single token that indicates different forms of the same word that have a common meanings). The result is that each document is a string of tokens, where a token is a sequence of characters that are grouped together as a useful semantic unit (but are not always immediately recognizable as words) for further processing. A document term matrix was then created where each column represents a token and each row represents a document. From here I identified the most frequently occurring tokens in the entire corpus

and repeated the stopword removal process to remove context specific but relatively uninformative high-frequency tokens such as *aaa*, *session*, *panel*.

```
require(tm)
a <- Corpus(VectorSource(df$text)) # create corpus object
a <- tm_map(a, tolower) # convert all text to lower case
a <- tm_map(a, removePunctuation)
a <- tm_map(a, removeNumbers)
a <- tm_map(a, removeWords, stopwords("english")) # this list needs to be edited and this
function repeated a few times to remove high frequency context specific words with no semantic
value
require(rJava) # needed for stemming function
library(Snowball) # also needed for stemming function
a <- tm_map(a, stemDocument, language = "english") # converts terms to tokens

a.dtm <- TermDocumentMatrix(a, control = list(minWordLength = 3)) # create a term document
matrix, keeping only tokens longer than three characters, since shorter tokens are very hard to
interpret
inspect(a.dtm[1:10,1:10]) # have a quick look at the term document matrix
findFreqTerms(a.dtm, lowfreq=30) # have a look at common words, in this case, those that appear
at least 30 times, good to get high freq words and add to stopword list and re-make the dtm, in
this case add aaa, panel, session
findAssocs(a.dtm, 'science', 0.30) # find associated words and strength of the common words. I
repeated this function for the ten most frequent words.
```

Term frequency and association analyses are simple but widely used methods in text mining because the results are relatively simple to calculate and interpret (Namey et al., 2007: 141; Ryan and Bernard, 2000: 776). Table 3.3 shows the 25 most frequently occurring tokens in the corpus. Table 3.4 shows the 10 tokens most strongly correlated with the 10most frequently occurring tokens. Once these tokens were identified, close reading of a sample of the full text was undertaken to investigate their meaning and context. The 10 most frequently occurring tokens reflect four topics of the meeting that Twitter-using anthropologists were responding to. The most frequent token, *scar*, is contained in messages referring to the session "The scars of human evolution" in which author 8679 was a presenter. The majority of messages containing this token are either messages by this author or other authors retweeting his messages. Author

Table 3.3 High-Frequency Tokens in the Corpus

Frequency	Tokens
100	scar
60	scar, scienc[e]
50	scar, scienc[e], digita[l]
40	scar, scienc[e], digita[l], people[e]
30	scar, scienc[e], digita[l], people[e], activit[y], evoluti[ion], male, publishin[g]
20	scar, scienc[e], digita[l], people[e], activit[y], evoluti[ion], male, publishin[g], birt[h], bra[in], bud[get], domingue[z], ethic[s], foo[d], future, industrialize, occup[y], primate, ris[k], rol[e], sout[h], theor[y], virgini[a], writin[g]

The characters in square brackets show the terms that the tokens most frequently derive from.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:21 Page Number: 84

t0025 **Table 3.4 Token Associations in the Corpus for the 10 Most Frequently Occurring Tokens (Limited to the 10 Tokens with the Strongest Associations with Correlations >0.2)**

Token	Associated Tokens (Strength of Correlation)
Scar	doctor (0.30), milfor (0.30), pond (0.30), wolp (0.30), birt (0.28), bra (0.26), lif (0.26), expenditure (0.24), compare (0.23), evoluti (0.23)
Scienc	humanis (0.63), scientific (0.63), rol (0.46), debat (0.45), nuance (0.33), educati (0.25), see (0.25), plac (0.25)
Digita	space (0.29), morp (0.27), archive (0.25), audi (0.25), collaboration (0.25), exhibition (0.25), includ (0.25), layere (0.25), repatriati (0.25), tur (0.25)
People	decad (0.37), pri (0.37), reproductiv (0.37), surviv (0.37), jer (0.31), wakin (0.29), archiva (0.24), destructiv (0.24), prett (0.24), sometime (0.24)
Activit	bud (0.95), exercis (0.91), subistenc (0.91), moderat (0.89), brin (0.87), dail (0.87), lin (0.78), industrialize (0.64), ironi (0.23), stretc (0.23)
Evoluti	anato (0.35), childbirth (0.35), litte (0.35), thrill (0.35), detail (0.31), stre (0.31), disabl (0.30), persona (0.28), treva (0.28), wend (0.28)
Male	subsistenc (0.80), weig (0.80), expenditure (0.64), energ (0.60), industrialize (0.51), female (0.42), competit (0.35), aggressivenes (0.34), favore (0.32), level (0.29)
publishin	futur (0.43), academi (0.40), sav (0.40), gree (0.35), brandin (0.30), speculation (0.30), vita (0.30), opportunitie (0.26), curatoria (0.24), pushin (0.24)
Birt	decad (0.59), pri (0.59), reproductive (0.58), surviv (0.58), amaz (0.50), suppor (0.49), measure (0.40), pas (0.37), los (0.36), weigh (0.36)
Bra	compare (0.67), rat (0.67), restin (0.67), mammal (0.64), metaboli (0.59), neonat (0.56), primte (0.56), siz (0.53), human (0.51), adul (0.49)

See the text for reconstruction of the terms from these tokens.

5399 also used this token in frequent messages referring to this session and was similarly retweeted. Associated with *scar* is the name of one of the discussants of the session, Milford Wolpoff, whose name mostly occurs in the context of messages noting his reference to the television series Dr. Who, indicating the mixture of scholarly and informal messages relating to this session. The terms *birth*, *brain*, *life*, etc., can be reconstructed from the tokens in Table 3.4 and come from highlights of scholarly content in the presentations, mostly in messages by author 8679. Among the other top 10 high-frequency tokens, *peopl*, *activit*, *evoluti*, *male*, *birt*, and *bra* (most frequently *brain*, though also resulting from the unrelated term *brand*) also relate to this session. The dominance of this session in the Twitter content appears to reflect the experience of a small number of people who participated in the session and the followers of these people who rebroadcast snippets of detail from the presentation most likely for the benefit of others who were not attending the session.

p1020 The second most frequent token is *scienc*. This token is more evenly distributed across the authors and, as can been seen from the associated tokens in Table 3.4, relates to the debate about whether anthropology is more of a humanistic or scientific discipline. Messages containing this token fall into two categories. First are direct observations on the session "Science in Anthropology: An Open Discussion," which was organized in response to controversy

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:22 Page Number: 85

surrounding the removal of the word *science* from the AAA's long-range plan statement in 2010. An interesting contrast between the importance of this issue among the community of Twitter-using anthropologists and the wider group of meeting attendants is revealed by this message: "#AAAsci is dominating #AAA2011 conversation, yet the room is less full than anticipated. 516 CD. Looks like there's much discussion ahead." This indicates that the science issue had been frequently mentioned by Twitter-users, but the low attendance at the session suggested to that author that it was not a high priority for the majority of participants.

p1025 The second category of messages, discussing science, contains links to articles in *The Chronicle of Higher Education* (Berrett, 2011) and *Inside Higher Education* (Jaschik, 2011). The link to the *Inside Higher Education* story on the science debate was the most frequently shared link in the corpus and indicates the importance of this issue to Twitter-using anthropologists (Figure 3.5). In this sample, 276 messages contained links, i.e., 18% of the sample, which is a substantially lower proportion than similar datasets (Weller et al., 2011). The cited articles were

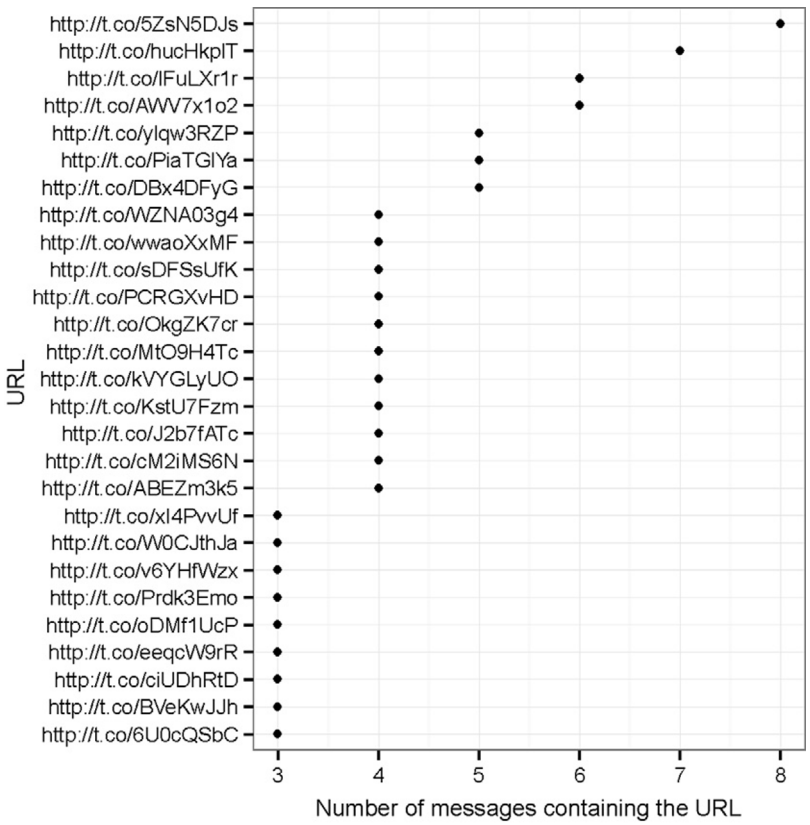

Figure 3.5

f0030 Frequency of URLs in the corpus.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:22 Page Number: 86

published online while the meeting was in session and give history to the controversy as well as reporting on the 2011 "Science in Anthropology" session. There is no evaluation of the news articles in the messages, only broadcasting of the headline and a link to the online article. This behavior has previously been observed as a common use of Twitter during academic conferences (Weller et al., 2011). One interpretation of this behavior is that authors are trying to work around the 140-character limit of Twitter messages by linking to long-form writing that contains more complex and nuanced discussions.

```
# investigate the URLs contained in the Twitter messages
require(stringr)
require(ggplot2)
df$link <- sapply(df$text,function(tweet) str_extract(tweet,
("http[[:print:]]+"))) # creates new field and extracts the links contained in the tweet
df$link <- sapply(df$text,function(tweet)
str_extract(tweet,"http[[:print:]]{16}")) # limits to just 16 characters after http so I just
get the shortened link.
countlink <- data.frame(URL = as.character(unlist(dimnames(sort(table(df$link))))),
N = sort(table(df$link))) # get frequencies of each link and put in rank order
rownames(countlink) <- NULL # remove rownames
countlinkSub <- subset(countlink, N>2) # subset of just links appearing more than twice
# plot to see distribution of links
ggplot(countlinkSub, aes(reorder(URL, N), N)) +
  xlab("URL") +
  ylab("Number of messages containing the URL")+
  geom_point() +
  coord_flip() +
  theme_bw() +
  theme(axis.title.x = element_text(vjust = -0.5, size = 14)) +
  theme(axis.title.y = element_text(size = 14, angle=90)) +
  theme(plot.margin = unit(c(1,1,2,2), "lines"))
```

p1120 The third most frequent token, *digita*, refers to two sessions, mostly "Digital Anthropology: Projects and Projections" and to a lesser degree, "Coming of Age in the Digital Age: Youth Media Practices and Gendered Identities." Similar to the Scars session, many of the messages containing the *digita* token originate from a single author (5274), who was also a presenter in the "Digital Anthropology" session, and relate to the scholarly content of that session. A focus on digital topics is to be expected from a community of authors who exist because of their use of digital media such as Twitter.

p1125 The final theme that can be readily identified from these data relates to the token *publishin*. The associated tokens show that the authors were concerned with the future of academic publishing. Most of these tokens relate to two messages originally by author 5274 as comments on the "Digital Anthropology" session (in which this person was a presenter) that were frequently retweeted. In this prominence of the digital and scars session in the Twitter corpus, we see how the interests of a small number of authors have dominated the corpus. The high frequency of posts about these sessions, and about the science session, reflect a small but engaged

B978-0-12-411511-8.00003-7, 00003

community whose interests and experiences are not necessarily reflective of others involved in the meeting. For example, although the Society for Medical Anthropology is one of the largest AAA sections, content from its sessions are not prominent in the Twitter corpus.

s0040 3.7 What Do the Twitter Messages Reveal About the Opinions of Their Authors?

p1130 The token frequency and association data give some simple insights into the issues that dominate the messages emanating from the meetings. But they are not very effective at revealing whether they had a positive or negative opinion about the issues they write about. Some insights into this may be obtained using sentiment analysis, the computational study of the opinions that people have about entities and events (Thelwall et al., 2011). The number of occurrences of positive and negative words in each document was counted to determine the document's sentiment score. To calculate the document sentiment score, each positive word counts as +1 and each negative word as −1. Although the method I used has the advantage of being simple to use, it does not handle polysemy, for example, irony and sarcasm (inspection of the corpus indicated that these forms of expression were very rare). I used a list of 6789 positive and negative words created by Hu and Liu (2004) to calculate scores for all documents in the corpus (the list is available online at http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html).

p1135
```
# This is based on Jeffrey Breen's excellent tutorial at http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/
```

p1140
```
# download sentiment word list from here: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar un-rar and put somewhere logical on your computer
hu.liu.pos = scan('C:/...somewhere on your computer.../opinion-lexicon-English/positive-words.txt', what = 'character',comment.char=';') #load +ve sentiment word list
hu.liu.neg = scan('C:/...somewhere on your computer.../opinion-lexicon-English/negative-words.txt',what = 'character',comment.char= ';') #load -ve sentiment word list
pos.words = c(hu.liu.pos)
neg.words = c(hu.liu.neg)
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
    require(plyr)
    require(stringr)
# we got a vector of sentences. plyr will handle a list
# or a vector as an "l" for us
# we want a simple array ("a") of scores back, so we use
# "l" + "a" + "ply" = "laply":
   scores = laply(sentences, function(sentence, pos.words, neg.words) {
# clean up sentences with R's regex-driven global substitute,
gsub():
       sentence = gsub('[[:punct:]]', '', sentence)
       sentence = gsub('[[:cntrl:]]', '', sentence)
       sentence = gsub('\\d+', '', sentence)
       # and convert to lower case:
       sentence = tolower(sentence)
       # split into words. str_split is in the stringr package
```


Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:23 Page Number: 88

```
        word.list = str_split(sentence, '\\s+')
        # sometimes a list() is one level of hierarchy too much
        words = unlist(word.list)

        # compare our words to the dictionaries of positive & negative terms
        pos.matches = match(words, pos.words)
        neg.matches = match(words, neg.words)

        # match() returns the position of the matched term or NA
        # we just want a TRUE/FALSE:
        pos.matches = !is.na(pos.matches)
        neg.matches = !is.na(neg.matches)

        # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
        score = sum(pos.matches) - sum(neg.matches)

        return(score)
    }, pos.words, neg.words, .progress=.progress )

    scores.df = data.frame(score=scores, text=sentences)
    return(scores.df)
}
aaa.text <- laply(aaa2011, function(t)t$getText()) # draw on the original object of tweets
that we first got to extract just the text of the tweets
length(aaa.text) #check how many tweets, make sure it agrees with the original sample size
head(aaa.text, 5) #check content sample, see that it looks as expected, no weird
characters, etc.
aaa.scores <-
score.sentiment(aaa.text,pos.words,neg.words,.progress='text')
# get scores for the tweet text
# create a histogram of sentiment scores
ggplot(aaa.scores, aes(x=score)) +
  geom_histogram(binwidth=1) +
  xlab("Sentiment score") +
  ylab("Frequency") +
  theme_bw() +
  opts(axis.title.x = theme_text(vjust = -0.5, size = 14)) +
  opts(axis.title.y=theme_text(size = 14, angle=90, vjust = -0.25)) +
  opts(plot.margin = unit(c(1,1,2,2), "lines"))
aaa.pos <- subset(aaa.scores,aaa.scores$score >= 2) # get tweets with only very +ve scores
aaa.neg <- subset(aaa.scores,aaa.scores$score <= -2) # get tweets with only very -ve scores

# Now create subset based on tweets with certain words, such as the high frequency words
identified in the text mining. eg. science
scien <- subset(aaa.scores, regexpr("scien", aaa.scores$text) > 0) # extract tweets
containing only 'scien'
# plot histogram for this token,
ggplot(scien, aes(x = score)) + geom_histogram(binwidth = 1) +
  xlab("Sentiment score for the token 'scien'") +
  ylab("Frequency") + theme_bw() +
  theme(axis.title.x = element_text(vjust = -0.5, size = 14)) +
  theme(axis.title.y = element_text(size = 14, angle = 90, vjust = -0.25)) +
  theme(plot.margin = unit(c(1,1,2,2), "lines"))
# repeat this block with different high frequency words
```

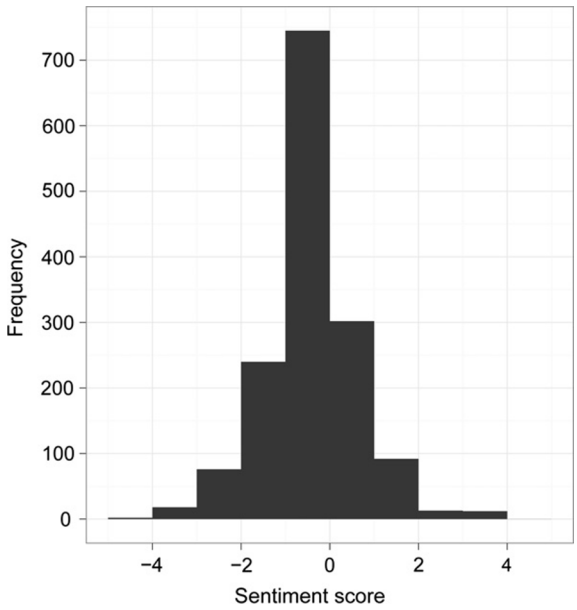

Figure 3.6

f0035 Histogram of sentiment scores for all documents.

p1455 The modal sentiment over the entire corpus is neutral to slightly negative (Figure 3.6). A small number of very positive scores (>2) counter the negative mode, resulting in a mean sentiment score of 0.08. The range of scores in this sample (−4 to 4) is smaller than a larger sample of general Twitter messages (−6 to 7, Breen, 2011). Taking the subset of documents ($n=65$) that contain the token *scien*, a similar slightly negative mode is evident in Figure 3.7 but there are no highly positive scores. This results in a significantly more negative sentiment about the science issue than overall sentiment about the meeting ($t=2.53$, df = 126.26, $p=0.01$). This is consistent with the weblog and news article commentaries produced during and shortly after the meeting report that meeting participants were frustrated with the discussion of the science issue (Antrosio, 2011; Berrett, 2011; Jaschik, 2011; Lende, 2011; Marks, 2011; Van Arsdale, 2011).

p1460 Sentiment surrounding the token *digita* (54 messages) was more positive, with a mean score of 0.37, a relatively high minimum score of −2 and a positive skew in the distribution of scores (no significant difference to overall sentiment: $t=-1.46$, df = 35.34, $p=0.15$) (Figure 3.8). These data convey the enthusiasm that some of the authors have for digital technologies in anthropology, which is evident in a sample of the full text, for example, "amazing opportunities for authors and readers," "great panel this morning," and "great papers." On a related issue, documents containing the token *publishin* had a mean score of exactly 0.0 and a range from −2 to 1 indicating a mix of

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:24 Page Number: 90

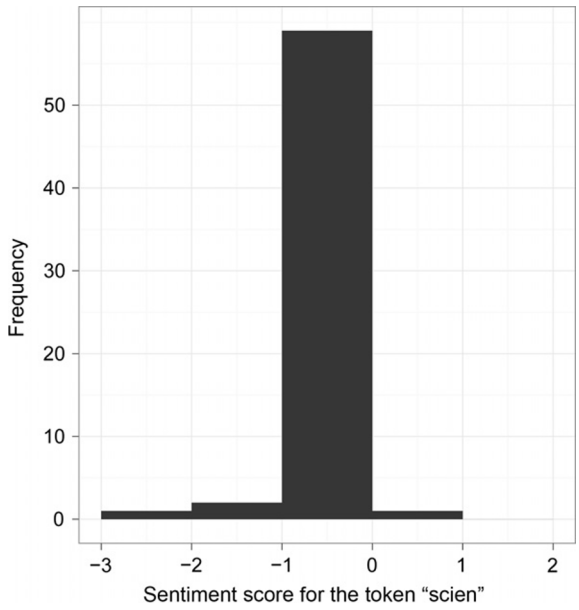

Figure 3.7

f0040 Histogram of sentiment scores for documents containing the token *scien*.

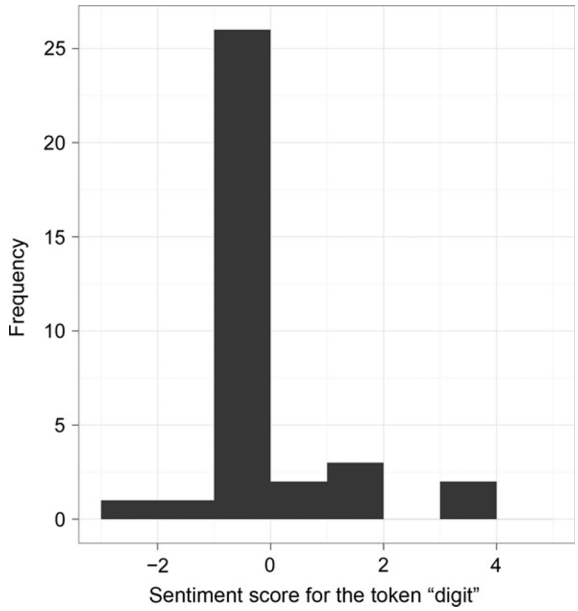

Figure 3.8

f0045 Histogram of sentiment scores for documents containing the token *digita*.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:24 Page Number: 91

positive and negative sentiment about the future of academic publishing. Sentiment scores were not calculated for documents containing the token *scar* because inspection of the full text indicated that they contain semantic terms such as *lower* and *fail* that relate to the scholarly content of the session rather than the opinion of the author and so would have given exaggerated negative sentiment scores. Direct inspection of the corpus reveals no obviously negative messages and three messages that are explicitly positive, noting the high attendance at the session, praising the humor of the presenters and creativity of the paper titles.

s0045 3.8 What Can Be Discovered in the Less Frequently Used Words in the Sample?

p1465 Although token frequency and association analyses are simple and revealing, their focus on the highest frequencies and strongest associations means they are not sensitive to less common patterns in the text. To investigate these rarer patterns, I used hierarchical clustering methods to produce a visualization of the distances between tokens in the corpus. This method takes the document term matrix and calculates distances between all of the tokens based on their frequencies in the documents and then classifies the tokens into nested groups (Suzuki and Shimodaira, 2006) (Figure 3.9). This method is useful because it reveals correlations between rarer tokens that do not appear in the frequency and association analysis, giving additional insights into what captured the attention of Twitter-using anthropologists.

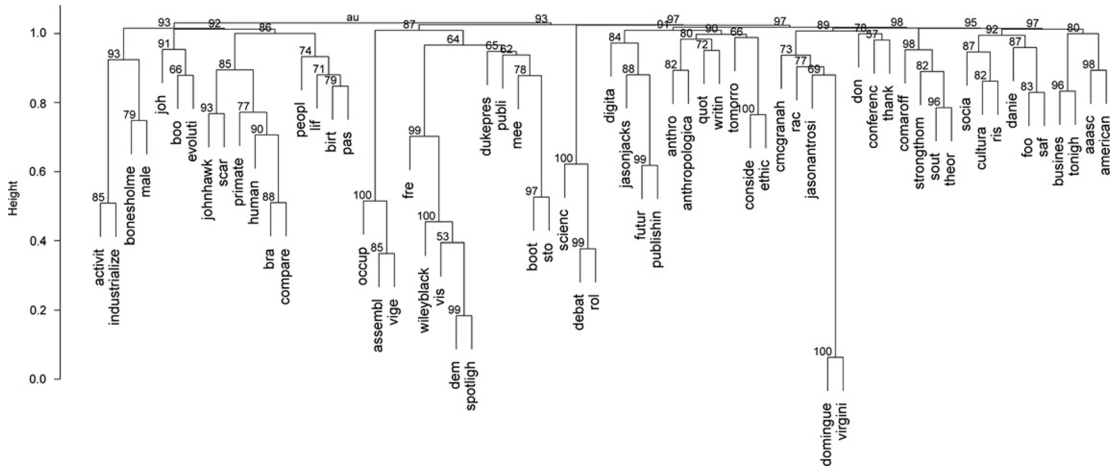

Figure 3.9

f0050 Cluster dendrogram of all documents with AU (approximately unbiased) *p*-values. For each cluster in the dendrogram, *p*-values between 0 and 1 were calculated by multiscale nonparametric bootstrap resampling (in this case, 5000 resamples). Clusters that are highly supported by the data have *p*-values closer to 1.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:24 Page Number: 92

```
a.dtm.sp <- removeSparseTerms(a.dtm, sparse=0.989) # I found I had to iterate over this to
ensure the dtm doesn't get too small... for example: 0.990 nrow=88, 0.989, nrow=67, 0.985,
nrow=37, 0.98 nrow=23, 0.95 nrow=6
a.dtm.sp.df <- as.data.frame(inspect(a.dtm.sp)) # convert document term matrix to data frame
nrow(a.dtm.sp.df) # check to see how many words we're left with after removing sparse terms
# this analysis is based on http://www.statmethods.net/advstats/cluster.html
# scale and transpose data for cluster analysis
a.dtm.sp.df.sc.t <- t(scale(a.dtm.sp.df))
require(pvclust)
fit <- pvclust(a.dtm.sp.df.sc.t, method.hclust = "average", method.dist = "correlation",
nboot = 10000) # this method may take a few hours the bootstraping, you can reduce the nboot value
for a quicker result
plot(fit, cex = 1.5, cex.pv = 1.2, col.pv = c(1,0,0), main="", xlab="", sub="") # draw the
dendrogram
```

Support for claims based on the token frequency and association data can been seen in the large left-most cluster that includes 17 tokens relating to the scars session. The debate about the role of science is clearly evident in a tight cluster near the center containing *scienc*, *debat*, and *rol*. The issue of digital media and the future of academic publishing are captured by a cluster just to the right of the science debate cluster.

In addition to this verification of the token frequency and association analysis, several further insights into the issues contained in the corpus may be derived from the cluster analysis. The cluster of *occup*, *asssembl*, and *vige* (Viger Hall, a location at the meeting) derives from messages encouraging people to participate in a general assembly in support of the Occupy protest movement. The clusters containing *wileyblack* and *dukepres* are readily identifiable as deriving from the stream of advertisements from these publishers.

The cluster containing the names Carole McGranahan, Jason Antrosio, and Virginia Dominguez (the current AAA president) refers to messages discussing the AAA Presidential Address. The focus of many of these messages is Dominguez's discussion of the 2010 final report of the Commission on Race and Racism in Anthropology, as indicated by the token *rac* in this cluster. Links to the PDF file of the report were also circulated in five messages, making it the third most frequently shared link in the corpus. Inspection of the full text reveals generally positive sentiment about the Presidential Address, for example, "an address worth thinking about" and "great presidential address." Moving further to the right of the dendrogram, a cluster including *sout*, *theor*, *comarof* derives from messages commenting on the session "Authors Meet Critics: Reading Jean and John Comaroff's 'Theory From The South: Or, How Euro-America is Evolving Toward Africa." The scholarly content of this session was reported in almost one hundred messages by a single author, whose messages were widely retweeted. The cluster containing *foo*, *saf*, and *danie* refers to 46 messages discussing papers presented in the 12 sessions (and evening reception) sponsored by the Society for the Anthropology of Food and Nutrition (identified by the hashtag #SAFN, from which the *saf* token derives). Inspection of the full text reveals that the token *danie* refers to Daniel Reichman's paper in the session "Ethnographic Approaches to Food Activism: Agency,

Democracy, and Economy" which contained the observation of a "new trend toward consumers" desire for absolute empirical knowledge about the provenance of their food', which was retweeted by a number of authors. Au5

s0050 3.9 What Are the Topics That Can Be Algorithmically Discovered in This Sample?

p1530 My final method for discovering what was important to Twitter-using anthropologists during the meeting is topic modeling. This method allows for a more complex subject analysis than the text mining and token distance techniques employed earlier (Newman and Block, 2011). Topic models are generative models that aim to discover the hidden thematic structures in large numbers of text documents. A topic is defined as a probability distribution over all words in the corpus that captures the salient themes that run through the corpus (Blei et al., 2010; Steyvers and Griffiths, 2007). Each document in the corpus is represented as a probability distribution over some of the topics. Topic modeling aims to infer the set of topics that were responsible for generating a collection of documents. The difference between topic modeling and text mining methods is that while text mining methods assume that each token is distinctive to a topic, topic models are mixed-membership models, meaning that each word or token may simultaneously belong to several topics, each document may contain several topics, and the distributions of these topics will vary over the documents in the corpus (Grün and Hornik, 2011). The unique contribution of this method is that it can identify a topic characterized by key words that may never appear next to each other in the same document.

```
require(slam)
a.dtm.sp.t <- t(a.dtm.sp) # transpose document term matrix, necessary for the next steps using
mean term frequency-inverse document frequency (tf-idf) to select the vocabulary for topic
modeling
summary(col_sums(a.dtm.sp.t)) # check median...
term_tfidf <- tapply(a.dtm.sp.t$v/row_sums(a.dtm.sp.t)[a.dtm.sp.t$i], a.dtm.sp.t$j,
mean) * log2(nDocs(a.dtm.sp.t)/col_sums(a.dtm.sp.t>0)) # calculate tf-idf values
summary(term_tfidf) # check median... note value for next line...
a.dtm.sp.t.tdif <- a.dtm.sp.t[,term_tfidf>=1.0] # keep only those terms that are slightly
less frequent that the median
a.dtm.sp.t.tdif <- a.dtm.sp.t[row_sums(a.dtm.sp.t) > 0, ]

summary(col_sums(a.dtm.sp.t.tdif)) # have a look
# Before going right into generating the topic model and analysing the output, we need to decide
on the number of topics that the model should use
# Here's a function to loop over different topic numbers, get the log likelihood of the model for
each topic number and plot it so we can pick the best one
# The best number of topics is the one with the highest log likelihood value.

require(topicmodels)
best.model <- lapply(seq(2, 50, by = 1), function(d){LDA(a.dtm.sp.t.tdif, d)}) # this will
make a topic model for every number of topics between 2 and 50... it will take some time!
```

```
best.model.logLik <- as.data.frame(as.matrix(lapply(best.model, logLik))) # this will
produce a list of logLiks for each model

# plot the distribution of log likelihoods by topic
best.model.logLik.df <- data.frame(topics=c(2:50), LL = as.numeric(as
.matrix(best.model.logLik)))
ggplot(best.model.logLik.df, aes(x = topics, y = LL)) +
   xlab("Number of topics") +
   ylab("Log likelihood of the model") +
   geom_line() +
   theme_bw() +
   theme(axis.title.x = element_text(vjust = -0.5, size = 14)) +
   theme(axis.title.y = element_text(size = 14, angle=90, vjust= -0.25)) +
   theme(plot.margin = unit(c(1,1,2,2), "lines")) # plot nicely
ggsave(file = "model_LL_per_topic_number.pdf") # export the plot to a PDF file
# it's not easy to see exactly which topic number has the highest LL, so let's look at the data...
best.model.logLik.df.sort <- best.model.logLik.df[order(-best.model.logLik.df$LL), ] #
sort to find out which number of topics has the highest loglik, in this case 23 topics.
best.model.logLik.df.sort # have a look to see what's at the top of the list, the one with the
highest score
```

p1675 Topic modeling identifies subject categories without *a priori* subject definitions (Newman and Block, 2011). Instead, fast algorithms for computing with hierarchical mixture models find the underlying patterns of words that are embedded in the corpus. Latent Dirichlet allocation (LDA) has been shown to be a highly effective unsupervised probabilistic method for finding distinct topics in Twitter messages (e.g., Ramage et al., 2010; Zhao et al., 2011) and a variety of other collections of documents (e.g., Blei et al., 2003; Hall et al., 2008). In brief, specifying the LDA model consists of three steps: (1) draw K topics from a symmetric Dirichlet distribution, (2) for each document d, draw topic proportions from a symmetric Dirichlet distribution, and (3) for each word n in each document d, (3a) draw a topic assignment from the topic proportions and (3b) draw the word from a multinomial probability distribution conditioned on the topic (Grün and Hornik, 2011). I generated LDA models that decomposed the corpus into its salient topics, and determined the specific distributions over the tokens for each topic and distributions of topics over each document (cf. Blei et al., 2010). To fit the LDA model to the document-term matrix, the number of topics needs to be decided in advance. To identify the optimum number of topics for this corpus, I calculated the log-likelihood of the data for all models with between 2 and 50 topics. The model with the highest log-likelihood value indicates the number of topics that are the best fit for the data (Griffiths and Steyvers, 2004), in this case 23 topics (Figure 3.10).

```
lda <- LDA(a.dtm.sp.t.tdif,23) # generate a LDA model with 23 topics, as found to be optimum
get_terms(lda, 5) # get keywords for each topic, just for a quick look
get_topics(lda, 5) # gets topic numbers per document
lda_topics<-get_topics(lda, 5)
beta <- lda@beta # create object containing parameters of the word distribution for each topic
gamma <- lda@gamma # create object containing posterior topic distribution for each document
terms <- lda@terms # create object containing terms (words) that can be used to line up with beta
and gamma
```

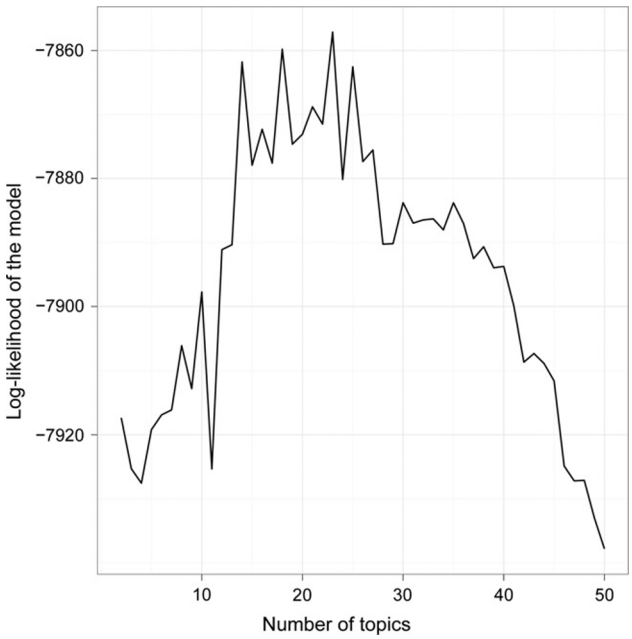

Figure 3.10

f0055 LDA model selection results showing the log-likelihood of the data for different numbers of topics.

```
colnames(beta) <- terms # puts the terms (or words) as the column names for the topic weights.
id <- t(apply(beta, 1, order)) # order the beta values
beta_ranked <- lapply(1:nrow(id),function(i)beta[i,id[i,]]) # gives table of words per
topic with words ranked in order of beta values. Useful for determining the most important words
per topic
```

p1730 Table 3.5 shows the top-ranked five tokens associated with each of the 23 topics. The topics automatically identified by the LDA model provide excellent verification of the issues identified by the token frequency and association analysis. Both methods identified the prominence of topics relating to the scars session, the "Theory from the South" session and the sessions on food, publishing, and Digital Anthropology. Other issues emerging from the topic model data include racism in anthropology, the role of science in anthropology, and changes to the AAA's code of ethics.

s0055 3.10 Conclusion

p1735 In summary, I have obtained a large number of short text messages written by participants of the 109th AAA meeting and used three methods of quantitative content analysis to discover the topical issues and controversies of the meeting according to the authors of these messages. I have also obtained some insights into the structure, rules, and practices of this community of authors. All three content analysis methods provide consistent results on the prominent topics, issues, and controversies of the meeting. Key issues for this community can be grouped

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:25 Page Number: 96

t0030 **Table 3.5 Topics and Their Five Top-Ranked Tokens Produced by the LDA Model**

1. [code of ethics]	2. [scars session]	3. [scars session]	4. mixed	5. mixed
Conside	scar	compare	publishin	conferenc
Ethic	cultura	bra	lif	don
Jasonjacks	johnhawk	primate	scar	peopl
Quot	quot	human	johnhawk	tomorro
Dukepres	bonesholme	johnhawk	comaroff	dukepres
6. [future of publishing]	**7. [SAFN sessions]**	**8. [scars session]**	**9. mixed**	**10. [race]**
Jasonjacks	foo	writin	peopl	virgini
Future	socia	ris	dukepres	domingue
Thank	saf	danie	scar	rac
Publishin	cultura	scar	johnhawk	cmcgranah
jasonantrosi	publi	publi	jasonantrosi	jasonantrosi
11. [scars session]	**12. [scars session]**	**13. [scars session]**	**14. [role of science]**	**15. [scars session]**
Scar	male	evoluti	scienc	activit
Birt	johnhawk	bonesholme	rol	industrialize
Pas	scar	johnhawk	debat	quot
Johnhawk	bonesholme	scar	jasonantrosi	johnhawk
jasonantrosi	jasonantrosi	joh	dukepres	scar
16. [adverts]	**17. [adverts]**	**18. [social]**	**19. [Digital Anthropology]]**	**20. [Occupy Montreal]**
Wileyblack	boot	quot	digita	occup
Vis	fre	boo	jasonantrosi	assembl
Dem	sto	mee	anthro	vige
Spotlight	publi	tomorro	quot	scar
Fre	wileyblack	anthro	joh	johnhawk
21. [theory from south session]		**22. mixed**		**23. mixed**
Theor		tonigh		johnhawk
Sout		anthropologica		american
strongthom		busines		aaasc
Comaroff		joh		scar
Quot		scar		quot

Tokens are ordered by the logarithmized parameters of the token distribution for each topic. I assigned the column labels in square brackets manually after inspecting the full set of topic-tokens (i.e., these column labels are not output from the model).

into scholarly concerns and concerns relating to policies specific to the AAA. Prominent scholarly concerns relate to papers presented in the scars session, the "Theory from the South" session, the Digital Anthropology session, the anthropology of food sessions, and the future of publishing forum. The concerns specific to the AAA corpus are the debate about the role of science in anthropology, racism in the discipline, and concern about revising the organization's code of ethics. Many of these issues were also represented in long-form weblog posts by anthropologists attending the meeting and the mainstream media, indicating a correlation between issues of interest to Twitter-using anthropologists, weblog authors (most of whom are

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:26 Page Number: 97

also highly active on Twitter), and the media. The most prominent controversy in the Twitter corpus, as measured by the sentiment analysis, was the role of science in anthropology. These messages were directed mainly to author peers participating in the conference, but there was limited dissent among authors, as indicated by the overall neutral and narrow range of sentiment scores. These observations are consistent with previous studies of the use of Twitter at academic meetings (Ebner, 2009; Ebner and Reinhardt, 2009; Ebner et al., 2010; Letierce et al., 2010a,b; Reinhardt et al., 2009; Ross et al., 2011).

p1740 Key attributes of the content of the corpus are the high proportion of retweeted messages and the circulation of links, indicating that sharing information and reporting news were common uses of Twitter by meeting participants. This distinctive content suggests that Twitter messages may have value for informing nonparticipants on the hot issues among Twitter-using anthropologists, contrary to previous work that found Twitter messages uninformative for nonparticipants (Ebner et al., 2010). Future research using interviews is needed to investigate the relationship between Twitter-using anthropologists and nonanthropologists. Institutional support for the use of Twitter by the AAA, such as a publically viewable projection of messages in a common space of the meeting venue, would likely stimulate more intensive use by attendees. This would result in a more complete record of the meeting in the Twitter corpus that would perhaps more credibly represent the diversity of the event to nonparticipants.

p1745 The structure of the community in this study is distinctive, with its demography biased toward more junior scholars and roughly equal representation of male and female authors. The relationship between gender and impact among Twitter-users (e.g., the number of followers and retweets) is an important issue for future investigation. A wide range of identity-signaling practices are employed with about half of the community using pseudonyms. The community has a small number of very highly interconnected individuals, and the majority of individuals are only connected to a small number of these highly connected individuals. One interpretation of this community structure is that Twitter-using anthropologists are comprised of many weakly connected groups composed of individuals sharing similar interests. For example, in several instances, we see one prolific individual broadcasting messages about the contents of a session and a group of dozen or so other individuals retweeting those messages. Among the different sessions where this occurred, few individuals appear to have been members of more than one group of retweeters.

p1750 This distinctive community structure is one of the most important emergent properties of the use of Twitter at the AAA meeting. The immediate nature of Twitter messages, compared to weblogs and other media, means that groups of individuals can rapidly and loosely self-assemble around specific events, such as conference presentations and specific people who are influential at these events. Similar phenomena have been described in the use of Twitter in political contexts (Holotescu et al., 2011). This is the transformative and emergent effect of Twitter in academia, to easily enable the spontaneous formation of information-sharing

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:27 Page Number: 98

communities bound by an interest in an event or topic. Twitter enables the kind of cross-cutting connectivity between groups of individuals that 19th-century sociologist Émile Durkheim (1893/1993) claimed was central to modern solidarity (Gruzd et al., 2011). The long-term stability of the membership and structure of these connections and communities formed by Twitter-users are important issues for future investigation.

p1755 A logical future extension of the methods presented here is for the analysis of longer texts such as weblog posts and journal articles. Furthermore, a corpus representing a longer period of time would also give insights into long-term community change and change in key issues and controversies. Although there are some pioneering examples of this kind of work (Blei and Lafferty, 2007; Griffiths and Steyvers, 2004; Hall et al., 2008; Mimno, 2012; Newman and Block, 2006, 2011), it remains for future work to take advantage of the reproducibility and accessibility that are key strengths of using R to make these methods more widely applicable.

st0070 ## *References*

Antrosio, J., 2011. Science in Anthropology: humanistic science and scientific humanism. Living Anthropologically Blog post 17-Nov-11. http://www.livinganthropologically.com/2011/11/17/science-in-anthropology/ (accessed 17.11.11).

Bentley, R., Hahn, M., Shennan, S., 2004. Random drift and culture change. Proc. R. Soc. B Biol. Sci. 271 (1547), 1443–1450.

Berrett, D., 2011. Anthropologists seek a more nuanced place for science. The chronicle of higher education. http://chronicle.com/article/Anthropologists-Seek-a-More/129823/ (accessed 17.11.11).

Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of science. Ann. Appl. Stat. 1, 17–35. Au6

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Blei, D., Carin, L., Dunson, D., 2010. Probabilistic topic models. Signal Process. Mag. 27 (6), 55–65.

Boellstorff, T., 2011. Three comments on anthropology and science. Am. Anthropologist. 113 (4), 541–544.

Boyd, D., Golder, S., Lotan, G., 2010. Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: Proceedings of the 43rd Hawaii International Conference on System Sciences, 5-8 January, Koloa, Kauai, HI, USA, 2010. Institute of Electrical and Electronics Engineers, pp. 1–10.

Breen, J., 2011. Sentiment analysis—a popular use of text mining in airlines' CRM. In: Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D. (Eds.), Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, New York, pp. 57–68.

Butts, C.T., 2008a. Social network analysis with sna. J. Stat. Software. 24 (6), 1–51.

Butts, C.T., 2008b. Social network analysis: a methodological introduction. Asian J. Soc. Psychol. 11 (1), 13–41.

Durkheim, É., 1893/1993. The Division of Labor in Society. Macmillan, New York, NY.

Ebner, M., 2009. Introducing live microblogging: how single presentations can be enhanced by the mass. J. Res. Innovative Teach. 2 (1), 91–100.

Ebner, M., Reinhardt, W., 2009. Social networking in scientific conferences—twitter as tool for strengthen a scientific community. Proceedings of the 1st European Conference on Technology Enhanced Learning, October 1-4, Nizza, Crete, Greece, 2009, vol. 2. Springer, pp. 1–8.

Ebner, M., et al., 2010. Getting Granular on twitter: tweets from a conference and their limited usefulness for non-participants. In: International Federation for Information Processing, World Computer Congress, Key Competencies in the Knowledge Society Conference, September 20-23, Brisbane, Australia, 2010. Springer, pp. 102–113.

Egri, C.P., 1992. Academic conferences as ceremonials: opportunities for organizational integration and socialization. J. Manag. Educ. 16 (1), 90–115.

Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in R. J. Stat. Software. 25 (5), 1–54.

Gentry, J., 2011. twitteR: R based Twitter client. R package version 0.99.15. http://cran.r-project.org/web/packages/twitteR/ (accessed 01.10.11).

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. 101 (Suppl. 1), 5228–5235.

Grün, B., Hornik, K., 2011. topicmodels: an R package for fitting topic models. J. Stat. Software. 40 (13), 1–30.

Gruzd, A., Wellman, B., Takhteyev, Y., 2011. Imagining twitter as an imagined community. Am. Behav. Sci. 55 (10), 1294–1318.

Hall, D., Jurafsky, D., Manning, C.D., 2008. Studying the history of ideas using topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 25–27 October Honolulu, Hawaii, USA, 2008. Association for Computational Linguistics, pp. 363–371.

Holotescu, C., et al., 2011. Microbolgging Meets Politics. The Influence of Communication in 140 Characters on Romanian Presidential Elections in 2009. Rom. J. Commun. Public Relations 13 (1), 37–50.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD (Association for Computing Machinery Special Interest Group for Knowledge Discovery and Data Mining) International Conference on Knowledge Discovery & Data Mining, 22-25 August, Seattle, USA, 2004. Association for Computing Machinery, pp. 168–177.

Jaschik, S., 2011. Not Feeling the Kinship. Inside Higher Education. http://www.insidehighered.com/news/2011/11/18/anthropologists-debate-role-science (accessed 18.11.11).

Kwak, H., et al., 2010. What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, 26-30 April. Association for Computing Machinery, Raleigh, North Carolina, USA, pp. 591–600.

Lende, D., 2011. The Montreal Anthropology Meetings—Recap of AAA Coverage. Neuroanthropology Blog post 22-Nov-11. http://blogs.plos.org/neuroanthropology/2011/11/22/the-montreal-anthropology-meetings-recap-of-aaa-coverage/ (accessed 22.11.11).

Letierce, J., et al., 2010a. Using Twitter during an Academic Conference: The iswc2009 Use-case. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 23-26 May, Washington, DC, USA, 2010a. Association for the Advancement of Artificial Intelligence, pp. 279–282.

Letierce, J., et al., 2010b. Understanding how Twitter is used to spread scientific messages. In: Proceedings of the Web Science Conference 2010: Extending the Frontiers of Society On-Line, 19th International World Wide Web Conference, 26-30 April, Raleigh, NC, USA, 2010b, pp. 1–8.

Marks, J., 2011. So, est-ce la science? Anthropomics Blog post 25-Nov-11. http://anthropomics.blogspot.com/2011/11/so-est-ce-la-science.html.

McCarthy, J.F., Boyd, D., 2005. Digital backchannels in shared physical spaces: experiences at an academic conference. In: Computer-Human Interaction 2005, Conference on Human Factors in Computing Systems, 5-7 April, Portland, Oregon, USA, 2005. Association for Computing Machinery, pp. 1641–1644.

Mimno, D., 2012. Computational historiography: data mining in a century of classics journals. J. Comput. Cult. Heritage. 5 (1), 1–19.

Namey, E., et al., 2007. Data reduction techniques for large qualitative data sets. In: Guest, G., MacQueen, K. (Eds.), Handbook for Team-based Qualitative Research. AltaMira Press, Lanham, MD, pp. 137–162.

Newman, D.J., Block, S., 2006. Probabilistic topic decomposition of an eighteenth-century American newspaper. J. Am. Soc. Inform. Sci. Technol. 57 (6), 753–767

Newman, D., Block, S., 2011. What, where, when, and sometimes why: data mining two decades of women's history abstracts. J. Women's Hist. 23 (1), 81–109

Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks. Comput. Inform. Sci. ISCIS. 2005, 284–293.

Prensky, M., 2001. Digital natives, digital immigrants Part 1. On the horizon 9 (5), 1–6.

Priem, J., Costello, K., Dzuba, T., 2011. Prevalence and use of Twitter among scholars. In: Metrics 2011: Symposium on Informetric and Scientometric Research., LA, USA, New Orleans, LA, USA, 2011.

Ramage, D., Dumais, S., Liebling, D., 2010. Characterizing microblogs with topic models. In: Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, 23-26 May Washington, DC, USA, 2010. The Association for the Advancement of Artificial Intelligence Press, pp. 1–10.

Reinhardt, W., et al., 2009. How people are using Twitter during conferences. In: Hornung-Prahauser, V., Luckmann, M. (Eds.), Kreativität und Innovationskompetenz im digitalen Netz: wie kommt das "Neue" mit Hilfe von Internettechnologien in die Welt?. Salzburg Research Forschungsgesellschaft, Salzburg, pp. 145–155.

Ross, C., et al., 2011. Enabled backchannel: conference Twitter use by digital humanists. J. Doc. 67 (2), 214–237

Ryan, G., Bernard, H.R., 2000. Data management and analysis methods. In: Denison, N., Lincoln, Y. (Eds.), Handbook of Qualitative Research. second ed. Sage Publications, Thousand Oaks, CA, pp. 769–802.

Steyvers, M., Griffiths, T., 2007. Probabilistic topic models. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (Eds.), Handbook of Latent Semantic Analysis. Psychology Press, London, pp. 424–440.

Suzuki, R., Shimodaira, H., 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22 (12), 1540–1542.

Thelwall, M., Buckley, K., Paltoglou, G., 2011. Sentiment in Twitter events. J. Am. Soc. Inform. Sci. Technol. 62 (2), 406–441

Van Arsdale, A., 2011. Science and the Ring Species of Anthropology. A.P. Van Arsdale Biological Anthropology Lab Blog post 21-Nov-11. http://blogs.wellesley.edu/vanarsdale/2011/11/21/anthropology/science-and-the-ring-species-of-anthropology/ (accessed 21.11.11).

Weller, K., Dröge, E., Puschmann, C., 2011. Citation analysis in Twitter. In: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences. Making Sense of Microposts (#MSM2011), Workshop at the Extended Semantic Web Conference 2011, 30 May, Crete, Greece, 2011. CEUR-WS.org, Tilburg University.

Wilson, S.M., Peterson, L.C., 2002. The anthropology of online communities. Ann. Rev. Anthropology 31, 449–467. Au7

Ye, S., Wu, S., 2010. Measuring Message propagation and social influence on Twitter.com social informatics. Soc. Inform. Lect. Notes Comput. Sci. 6430, 216–231.

Zhao, W., et al., 2011. Comparing Twitter and traditional media using topic models. Adv. Inform. Retrieval. 6611, 338–349.

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:27 Page Number: 101

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:28 Page Number: 102

Non-Print Items

Abstract

R is a convenient tool for analyzing text content to discover emergent issues and controversies in diverse corpora. In this case study, I investigate the use of Twitter at a major conference of professional and academic anthropologists. Using R I identify the demographics of the community, the structure of the community of Twitter-using anthropologists, and the topics that dominate the Twitter messages. I describe a series of statistical methods for handling a large corpus of Twitter messages that might otherwise be impractical to analyze. A key finding is that the transformative effect of Twitter in academia is to easily enable the spontaneous formation of information-sharing communities bound by an interest in an event or topic.

Keywords: *Twitter*, *Text mining*, *Topic modeling*, *Sentiment analysis*, *Social network analysis*, *Anthropology*

Comp. by: GAsokpandian Stage: Proof Chapter No.: 3 Title Name: Zhao
Date:17/8/13 Time:17:29:28 Page Number: 103