Team Project

SHEN Anqi 54577992, ZHANG Yijun 54505695, HE Peiwei 54471952

1 Introduction and research question

1.1 Introduction of United Airlines Controversy

United Airlines, an American airline company which is the world's third-largest airline when measured by revenue, has passed a tough month. On Sunday, April 9.Their airline officials and airport security dragged a doctor named David Dao off of an overbooked flight. A bunch of passengers have been speaking up about their terrible United experiences, and finally. It became international news.

After things happened, Video of the incident recorded by passengers went viral on social media, resulting in outrage over the violent incident. Politicians expressed concern and called for an official investigation. United Airlines issued its first response to the incident via Twitter. Apologize for having to re-accommodate customers. The internet went wild over Munoz's use of the word "re-accommodate."

Involving this crisis and crisis dealing measures of the company. This report will gather tweets data and mass media news reports about this crisis(including NYTimes, WathingtonPost, BBC news) to show and analyze the general public opinions for this crisis. Based on those analyses, we provide some basic advice for the company to react this crisis in a better way.

1.2 Research Question

Based on linear transmission model of messages. We made some research questions for this crisis,it includes

1. What's the information source of this crisis? What people create for message? How they create?
2. What are differences between transmitters? How they transmit this crisis?
3. How users encoded this news? How they feel?
4. What's the noise in this processing?
5. What measures does the company take? Do they work?
6. All in all, What are advices we can provide for United Airlines.

2 Data Gathering

2.1 Twitter Data

Using API to scrape those keywords with hashtags in twitter.

```
library(twitteR);library(quanteda);library(corpustools);library(RColorBrewer);
library(ggplot2);library(igraph);library(RTextTools);library(slam);library(reshape2);
library(plyr)
```

Search the Twitter by using harshtags related to United Airlines from April 10th to April 18th, and finally got 60000 tweets in total. Here is the tags we use:unitedairlines, dontflyunited, boycottunited, boycottunitedairlines, NewUnitedAirlinesMottos, UnitedAirlinesmemes, UnitedAirlinesAssault.

```
setup_twitter_oauth(consumer_key, consumer_secret, token, token_secret)
tweets <- plyr::ldply(searchTwitter("#unitedairlines", n=10000, lang="en"), as.data.frame)
tweets0417=subset(tweets,language=="en")
```

```
......
total <- rbind(tweets0412, tweets0417,tweets0418,tweets0418a,tweets0418b,tweets0418c,
               tweets0418d,tweets0418e,tweets0418f)
```

Use the tweets, corpus the tweets text and convert into dfm. Remove stopwords. Finally we could get a clean dtm.

```
stopwords = c(stopwords("english"),"coa","apps","https","t.co","who","you","this","too","for","via","wi
tweet_corpus = corpus(total$text)
tweet_dfm = dfm(tweet_corpus,stemming = T,removePunct=T,removeNumbers=T, removeTwitter=T,minWordLength =
tweet_dfm = dfm_select(tweet_dfm, stopwords,selection=c("remove"),valuetype=c("fixed"))
tweet_dfm = dfm_select(tweet_dfm,stopwords("english"),"remove")
```

2.2 Mass Media Data Gathering

After using Selenium in python to scrape JS page, we got all news links from those websites.We saved it to a url lists(Selenium is a browser simulator which helps me to paginate and get JS content).

The python code snippet is like below:

```
driver = webdriver.Chrome()
driver.get("http://www.bbc.co.uk/search?q=United+Airlines&sa_f=search-product&scope=")
for i in range(20):
    elem = driver.find_element_by_xpath("//*[@id='search-content']/nav[1]/a")
    elem.click()
    time.sleep(10)
```

Then using Rvest to scrape the content, title and time of the single page.(Just Show how scrape nytimes articles as limited pages). Finally, We gather 100 news from NewYork Times, 114 news from BBC and 17 news from Washington Mail.

```
nyt_news = as.vector(readLines("list_nyt_urls.txt"))#nyt_news
text_list_nyt = c();time_nyt = c();title_nyt = c()
length(nyt_news)

for(url in nyt_news){
  r = GET(url,user_agent(user_agent))
  text = iconv(list(r$content))
  text_tree = read_html(text)
  text_raw = text_tree %>% html_nodes(css=".story-content") %>% html_text(" ")
  time = text_tree %>% html_node(css=".dateline") %>% html_text(" ")
  title = text_tree %>% html_node(css=".headline") %>% html_text(" ")
  text_clean = gsub("[\r\n]","", text_raw)
  article = paste(text_clean,collapse = "\n")
  text_list_nyt = c(text_list_nyt,article)
  time_nyt = c(time,time_nyt);title_nyt = c(title,title_nyt)
}
nyt_17_articles = data.frame(text_list_nyt,title_nyt,time_nyt)
saveRDS(nyt_17_articles,file = "nyt_17_articles.rds")
```

3 Data Analysis

3.1 Wordcloud Plot

```
#Plot a naive wordcloud of United Airlines to have a brief view.
total = readRDS("totaltweets.rds")
tweet_dfm = readRDS("tweet_dfm.rds")
```

```
topfeatures(tweet_dfm, 10)
textplot_wordcloud(tweet_dfm, max.words = 50, random.color = TRUE, rot.per = .25, colors =sample(colors
```

3.2 Topic Model

3.2.1 Bulid Topic Model

Build a topic model of tweets by using LDA, automatically assigns topics to text documents. We concluded 5 topics, including crisis, passenger, company, discussion and jokes.

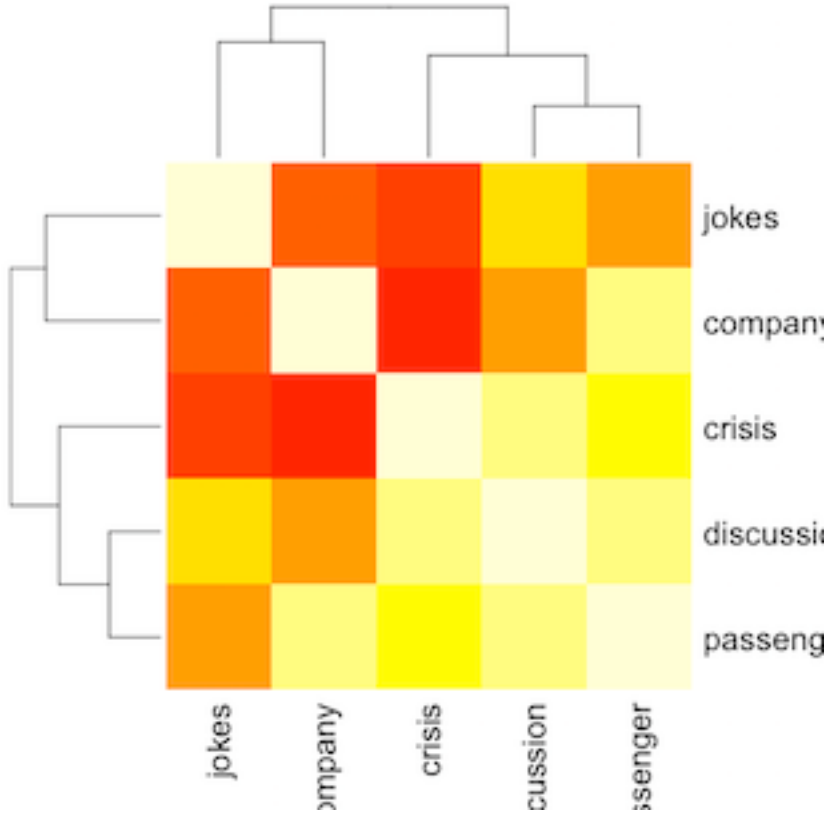
```
ignore = c(stopwords('english'),'*unitedairlines','united*','newunitedairlinesmottos','airlines','made'
d = dfm(tweet_dfm, remove=ignore, stem=T)
d = convert(d, to="topicmodels")
set.seed(123)
m = LDA(d, k = 5, method = "Gibbs", control=list(alpha=.1, iter=100))
topic=data.frame(terms(m, 10))
colnames(topic) = c("crisis", "passenger", "company","discussion","jokes")
topic
```

	crisis	passenger	company	discussion	jokes
1	drag	passeng	fli	genius	first
2	plane	seat	dontflyunit	whoever	new
3	passeng	overbook	never	dontflyunit	simpson
4	class	treat	ever	disgust	ascaniospread
5	guy	beat	airlin	absolut	beingfarhad
6	asian	want	custom	behaviour	footag
7	new	dontflyunit	like	internet	disturb
8	use	see	ceo	now	insid
9	ordinari	human	market	train	ceo
10	asian-look	drag	passeng	best	oscar

3.2.2 Topic Heatmap

Plot a heatmap of these 5 topics.

```
cm = cor(t(m@beta))
colnames(cm) = rownames(cm) = colnames(topic)
diag(cm) = 0
heatmap(cm, symm = T)
```



As the topic model shown, we can see that the most heated discussion is around the company and the crisis, while when people discuss this crisis on Twitter, they are likely to make jokes about United Airlines.

3.3 Sentiment Analysis

3.3.1 Applying Sentiment Dictionary to DTM

```
lexicon = readRDS("lexicon.rds")
#Read lexicon as dictionary to define positive and negative words.
pos_words = lexicon$word1[lexicon$priorpolarity == "positive"]
neg_words = lexicon$word1[lexicon$priorpolarity == "negative"]
```

```
#Read tweets texts, and apply the sentiment dictionary to create matrix of positive and negative words.
uatext = total$text
uadtm = create_matrix(uatext,language="english",removePunctuation=T)

total$npos = row_sums(uadtm[, colnames(uadtm) %in% pos_words])
total$nneg = row_sums(uadtm[, colnames(uadtm) %in% neg_words])
#Calculate the sentiment scores.
total$sent = (total$npos - total$nneg) / (total$npos + total$nneg)
total$sent[is.na(total$sent)] = 0
#Use quanteda to apply the sentiment dictionary to tweets.
```

```
dict = dictionary(list(pos=pos_words, neg=neg_words))
dfm2 = dfm(paste(total$text), dictionary=dict)
```

Convert it to a data.frame and combine it with the tweets, according to the sentiment classification and

retweetcount.

```
d2 = as.data.frame(dfm2)
total = cbind(total, d2)
tweetsent=subset(total,select=c("id","screenName","text","retweetCount","pos","neg","sent"))
tweetsent$gold = as.factor(ifelse(tweetsent$retweetCount > 1000, "pos", "neg"))
tweetsent$pred = ifelse(tweetsent$sent > 0, "pos", "neg")
mt = table(tweetsent$gold, tweetsent$pred)
mt
```

```
##
##         neg   pos
##   neg 30236 15868
##   pos  8233  5459
```

As the result shown, when the retweetCount is more than 1000, there seems to be more negative words in the text. So we guess more negative contents about United Airlines spreaded easily on Twitter.

3.3.2 Validating sentiment

Compute accuracy, precision, and recall on a dichotomized version of the predictions and manual sentiment. Firstly calculate overall prediction accuracy:

```
sum(diag(mt)) / sum(mt)
```

```
## [1] 0.5969463
```

```
pr = mt["pos", "pos"] / sum(mt[, "pos"])
re = mt["pos", "pos"] / sum(mt["pos", ])
f1 = 2*pr*re/(pr+re)
print("FScore:")
```

```
## [1] "FScore:"
```

```
c(Precision=pr, Recall=re, FScore=f1)
```

```
## Precision    Recall    FScore
## 0.2559666 0.3987000 0.3117736
```

As the FScore is not very high, the coorelation between retweetCount and words sentiment is not very significant. As the result shown, when the retweetCount is more than 1000, there seems to be more negative words in the text, which means more negative contents about United Airlines are more easily to spread on Twitter. To validate sentiment analysis result, as the FScore 0.34 is not very high, the coorelation between retweetCount and words sentiment is not very significant.

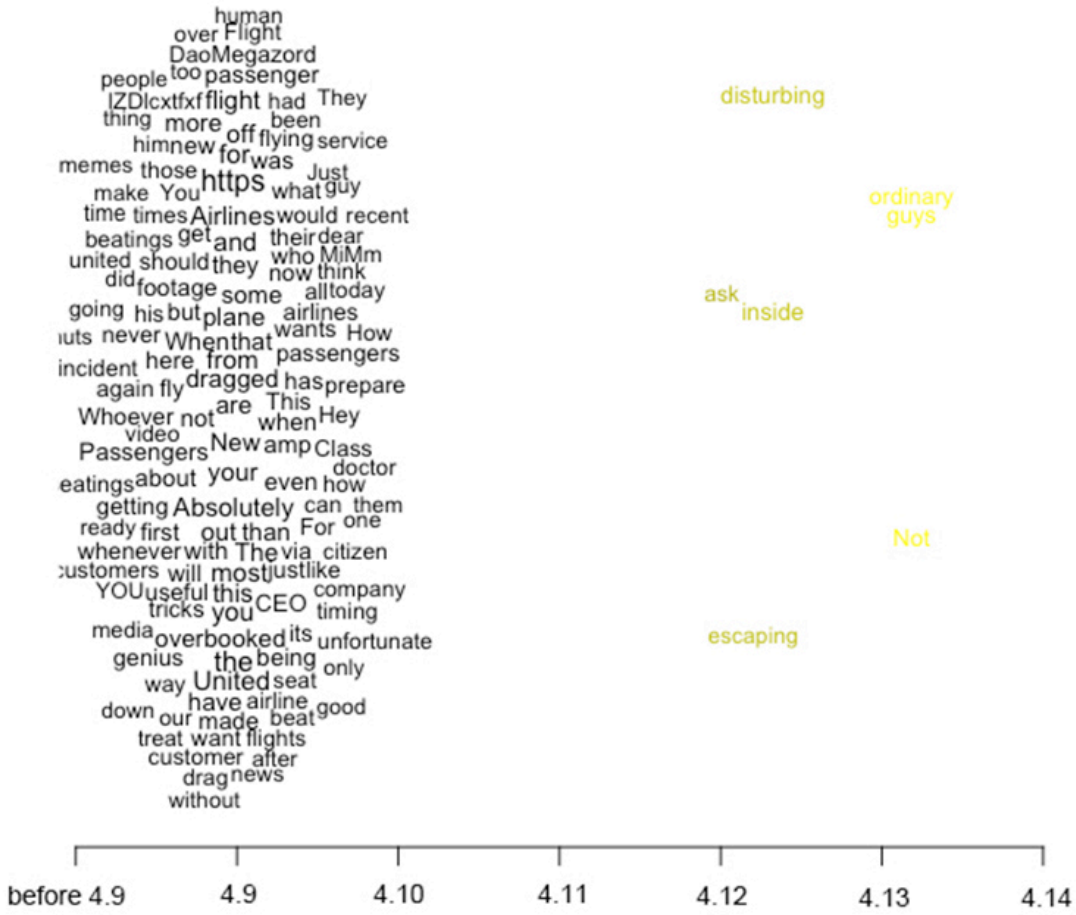
3.4 Time analysis(take twitter as example)

In this part, we analysis the change of keywords in twitter with time axis. It shows how opinion changes with time. As space-limited, we would just show some important codes.

```
total$date = as.Date(total$created)
total12 = total[total$date=='2017-04-12',];total13 = total[total$date=='2017-04-13',]
total_sample = rbind(total12,total13,total14,total15,total16)
text_from_tweets = total_sample$text
tweet_token = tokenize(as.character(text_from_tweets),removePunct = T)
ul1 = lapply(tweet_token, ldply)
for(i in seq(10000))
```

```
  {total_sample = data.frame(ul1[[i]])
  total_sample["tweet"] = i
  total_c = rbind(total_c,total_sample)}

total_sample["articles"] = as.numeric(rownames(total_sample))
total_sample$date = as.character(total_sample$date)
dtm = dtm.create(total_c$tweet, total_c$V1)
wordfreqs = dtm.to.df(dtm)
wordfreqs = merge(total_sample,wordfreqs,by.x="articles",by.y="doc")
mmode <- function(v){uniqv <- unique(v); uniqv[which.max(tabulate(match(v, uniqv)))]}
dates = aggregate(wordfreqs["created"], by=wordfreqs["term"], FUN=mmode)
cmp  = corpora.compare(dtm)
cmp  = arrange(merge(cmp, dates), -termfreq)
with(head(cmp, 150), plotWords(x = created, words = term, wordfreq = termfreq, random.y = T, scale=1))
```



We could see as time went on, the focus of opinion was changing. From the beginning, people focus on what's happened to David Dao and what was shown in that videos which recoding the violence. After that, people start to focus what CEO of United Airlines did in this crisis. And they thought themselves(like the word citizens). It cleanly revealed and divided stages of this crisis.

3.5 Network Analysis

```
#use regular expression to find repost(RT or via)
index = with(totaltweets,grepl("(RT|via)((?:\\b\\W*@\\w+)+)", totaltweets$text))
retweets = totaltweets[index,]
#find posters in repost text(information after RT @ or via @)
poster = str_extract_all(retweets$text,"(RT|via)((?:\\b\\W*@\\w+)+)")
poster = gsub(":", "", poster)
who_post = gsub("(RT @|via @)", "", poster, ignore.case=TRUE)
```

```
who_repost = retweets$screenName
reposts = as.data.frame(cbind(who_post,who_repost))#create a dataframe of repost edgelists
reposts$who_post = gsub("@.*", "",reposts$who_post)
reposts = rename(reposts, c("who_post"="Source", "who_repost"="Target"))
```

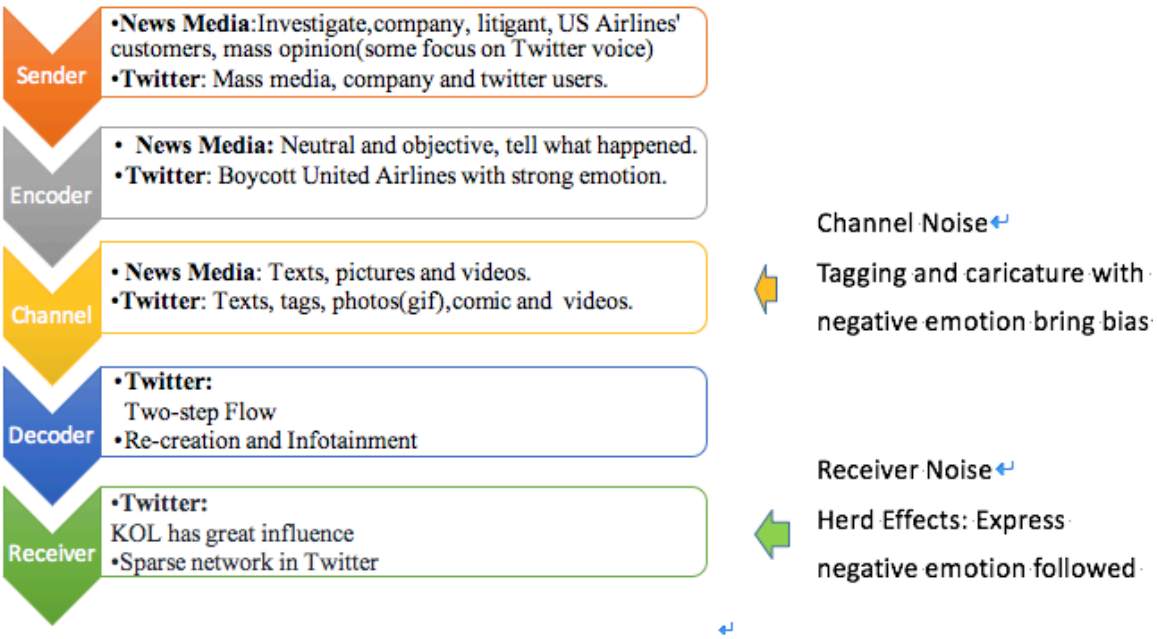
Analyze who became KOL in this crisis spreading in Twitter

```
#Find KOL: take users whose retweet count over 100000
KOL = cbind(totaltweets[totaltweets$retweetCount>10000,]$id,
            totaltweets[totaltweets$retweetCount>10000,]$screenName,
            totaltweets[totaltweets$retweetCount>10000,]$retweetCount)
colnames(KOL)=c("id","screenName","retweetCount")
#get these KOL's basic information of follower count and friends count
Get_info=function(name){
  message(name)
  u = twitteR::getUser(name);followerCount = u$getFollowersCount()
  friendCount = u$getFriendsCount()
  m = data.frame(cbind(name,followerCount,friendCount))
}
KOLinfo= data.frame(do.call(rbind,lapply(KOL],Get_info)))
high_repost = data.frame(cbind(KOL,KOLinfo))
```

4 conclusion

According to Fink's crisis model (1896), we find out this crisis has four stages:

- Prodromal period: We notice there are some negative news of United Airlines in mass media since 2016, but they took no measures to deal with these, regarding as little matters.
- Crisis breakout: The crisis broke out on April 9th with a video wide spreading, however, United Airlines treated it with excuse: the CEO denied their responsibility and shifted the blame to litigant.
- Chronic period: this is the stage where litigation occurs, media exposes are aired, internal investigations are launched—some staff were fired to take responsibility, government oversight investigations commence and so on. To notice, we can find unlike other crisis, United Airlines have large exposition and lasted a little long in social media especially on Twitter, proving as another dramatically change in counts and sentiment after 7 days of its breakout.
- Solution period: this is when things begin to return to normal.As prodromal period is hard to discover and take measures, we may focus on the second and third stages to find out how this crisis spread.



Here is the model we find in this crisis (Shannon&Weaver,1948), as we can not get the reader's behavior of mass media, so we only analyze decoder and receiver in Twitter.

Sender:News Media has various sources of information, some investigated by themselves, some interviewed company, litigant and other United Airlines' customers. As for New York Times, they even highlighted opinions in twitter, for we see twitter as an important topic in its topic model. This shows an interaction between mass media and social media: what say on social media would motivate news creation in mass media. As for social media like Twitter, mass media is a definitely crucial source, users as self-media also are information provider.

Graph1. Twitter Graph2.New York Times Graph3. BBC

Encoder As shown in word clouds above, we can find Twitter users sent messages with lots of hash tags with negative emotions. Some just showed their negative emotions, such as disgusting. Some thought it is worst and boycotted it. However, more people coded their tweets as irony, such as "genius", so in our sentiment analysis, we find it surprisingly that it is positive. And more tweets are, more positive the total sentiment is.

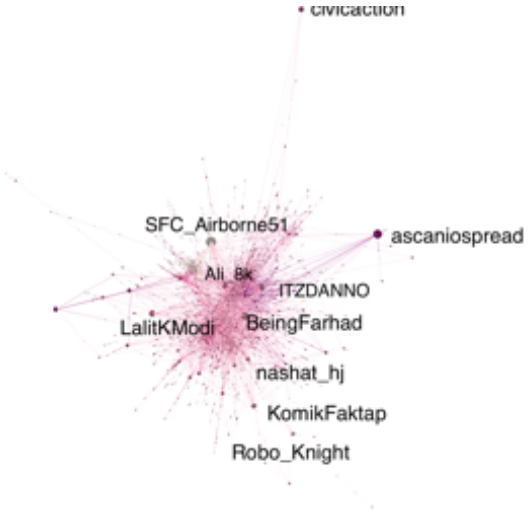
Mass media shows more objective and neutral attitude. New York Times emphasized the process of this controversy, and talked about racist issues. Washington Mail emphasized the litigants including CEO and David Dao and they paid more attention about human rights issues and company's issues. BBC just saw this as a foreign news on account of finding "US" and "twitter" as two highlighted words. They also traced

back some former faults of United Airlines made in UK. All medium avoided to use those words with Intense feeling. However, they described people are “outrange” and emphasized misdeed of United Airlines.

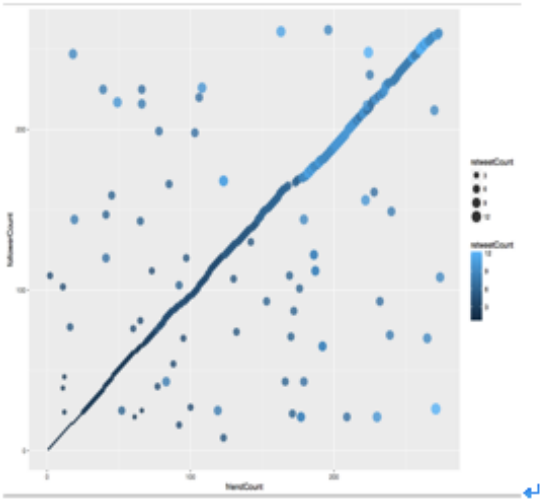
Channel Mass media treated this news with texts, pictures and videos, however, we just be able to analyze text one. There are some special spreading characteristics on Twitter in this crisis. Hundreds of users symbolized this crisis trough spoof videos and comics.

Decoder On account of two step flow in Twitter, it is possible to regenerate a crisis or amplify damage on company. As United Airlines we analyzed, we detect two high peaks of crisis communication, the first one is right after the outbreak stage, the second one, however, is 7 days after, due to the two step flow. And there is also more infotainment tendency with time passing by.

Receiver Key opinion leaders have great influence. As shown below in the spreading of hot tweets, the average Degree is 1.224. We can see the spreading network is sparse, and some opinion leaders have the power, such as Ascaniospread and LalitKModi. This shows some herd effects in this network, some tweets with less information just negative emotion were reposted widely. This is a noise from receiver self and this may cause neglect and voice losing of some users. And we also find one way to discover this KOL in Twitter: high influence users would have high friends count or high followers count or both. It may help crisis managing when crisis happen.



Graph 4. Repost network



Graph 5. Friend count and follower count of KOL

Improvement

- United Airlines should release news more timely, the video about policemen dragging that doctor was release at 9th April in Twitter. However, the apologetic letter was released at 10th April, it’s too late. Leave the time for negative opinion spreading.
- The content of first apologetic letter in breakout stage should be altered. They need to apologize more sincerely rather than “denying” it by using words like “re-accommodation”. We recommend a full responsibility taken as the first measures taking that may even gain good comments if dealing successfully with crisis.
- By using some communication strategies like apologize not only the victims in this crisis but also customers who had same situation, United Airlines could reduce the source of information.
- As there are much negative emotion tweets, the company can use drown method, posting much positive information to apologize and show clear attitudes. This may also reduce the source of information for mass media.
- We can use characteristics we have analyzed to find out opinion leaders, especially prevent the second damage, especially infotainment in chronic stage.