#JustDoIt

Tweets Analysis

TEAM 9
Shen Yang, Guanhang Chen, Wen Ya Shen,
Damla Erten, Pramodhini Somasekhar

BANA 277

Customer & Social Analytics

Sanjeev Dewan

Dec.14, 2018

Table of Contents

Executive Summary

For this project, we mainly focus on analyzing tweets collected involving the event of Nike endorse Colin Kaepernick on its 30th anniversary of its #JustDoIt campaign, which was also comment by president Donald Trump. The purpose of the project is to analyze relevant twitter user's general emotion and attitude towards Nike, Donald Trump, and Colin Kaepernick and their relevant social network connection. The data is obtained from the website Kaggle.

First of all, we did a sentimental analysis on the content of tweets in order to analyze people's reaction about this event. The result shows that are more positive tweets in the event. We would like to further analyze the content of tweets that reweeted the main influencers of the event. From the social network analysis, we found Nike, Donald Trump, and Colin Kaepernick are part of the main influencers. The result shows that there are still generally more positive tweets towards them. However, we were not able to identify tweets by their actual attitude towards these three main influencers. As people could mention both positive and negative comment towards all three people in one tweet, we were not able to know their exact attitude.

Then, we did a social network analysis to show the correlation. From the social network, we find out that Nike, Donald Trump as well as Colin Kaepernick are among the top three influential users. What's more, we did a regression analysis and we hope to know which kind of variables might have an impact on the result. The results shows that turning on the geographic location, the number of retweet will more than double. In addition, tweets from a new registered user are more likely to be retweeted.

Finally, a conclusion is included based on three types of analysis: sentimental analysis, social network analysis as well as regression analysis. We attached our R code and slides on the appendix.

I. Introduction

On Nike's 30th anniversary of its #JustDoIt campaign, the brand made an announcement to endorse Colin Kaepernick, an American football quarterback as the face of the campaign. A while back, Kaepernick made a controversial decision not to stand up during the national anthem, as a protest to racial injustice made by police. Kaepernick's decision has stirred a heated debate, especially when President Donald Trump commented on it via Twitter. Nike's Twitter advertisement partnership with Kaepernick is important, as Kaepernick's decision is very controversial. We are interested in exploring the data with all tweets that includes the hashtag #JustDoIt a few days after Kaepernick posted a campaign photo of his face with the caption "*Believe in something, even if it means sacrificing everything."* on September 3, 2018.

We would like to analyze the reaction of people on Nike's campaign from the tweets, and find out if people are expression a positive, negative, or neutral reaction to this campaign. In addition, we will performed a social network analysis to find any social connections between Twitter users. Furthermore, a regression analysis will be performed to further analyze which variables contribute effects to the number of retweets for Twitter users.

II. Data

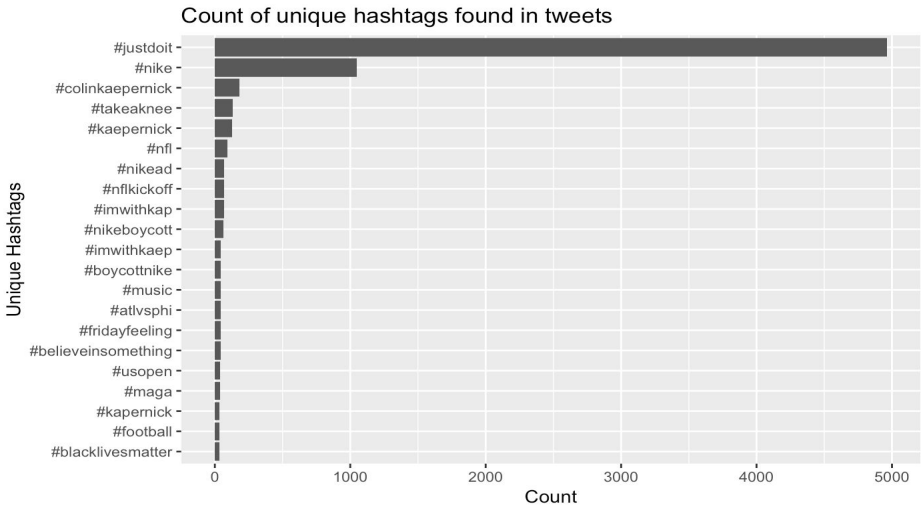
The dataset was obtained from the Kaggle website. This dataset contains 5089 observations and 75 variables. The variables are separated into two parts that include tweets attributes started with tweet_ and user attributes started with user_.

Preliminary analysis are performed to examine some basic information about the variable. First, we calculated a few basic statistics on the numeric variables.

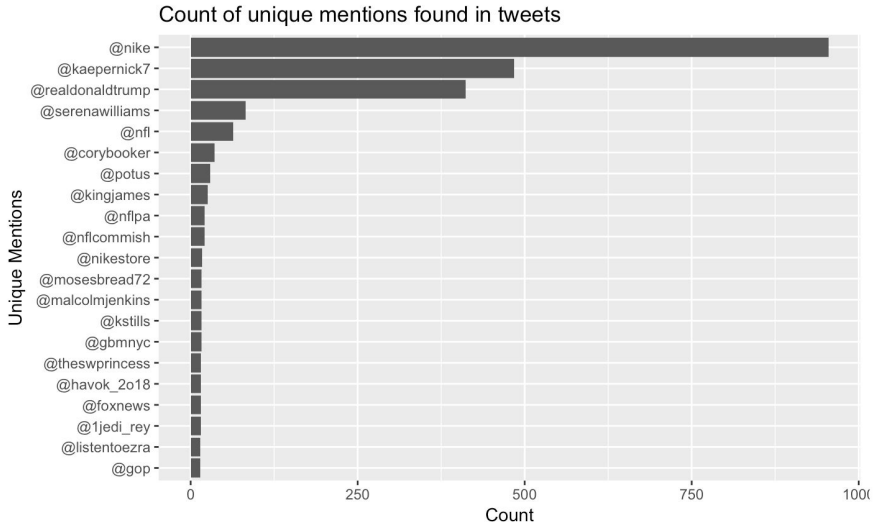
Table 1: Basic Statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
tweet_favorite_count	1	5089	4.47	76.14	0.0	0.47	0.00	0	4887	4887	53.94	3350.15	1.07
tweet_retweet_count	2	5089	1.23	15.60	0.0	0.06	0.00	0	748	748	31.75	1279.58	0.22
user_favourites_count	3	5088	10508.91	26031.94	1741.0	4421.37	2500.40	0	337135	337135	5.24	36.16	364.95
user_followers_count	4	5088	3325.23	44802.03	361.5	742.25	466.28	0	2896006	2896006	55.78	3475.13	628.09
user_friends_count	5	5088	1723.43	5710.91	532.0	809.56	597.49	0	129726	129726	11.80	183.83	80.06
user_listed_count	6	5088	41.89	222.46	6.0	13.07	8.90	0	11734	11734	31.77	1527.37	3.12
user_statuses_count	7	5088	18705.14	44584.12	4650.5	8693.93	6324.03	1	768609	768608	5.99	54.36	625.04

We also found the top unique hashtags and mentions from the variable tweet_full_text to look at the most popular topics in the campaign.



Plot 1. Top unique hashtags



Plot 2. Top unique mentions

III. Sentiment Analysis

Tweets Data Cleaning

For sentiment analysis, we would like to focus on the variable tweet_full_text in order to explore which tweets are positive, negative, or neutral. First, some data cleaning need to be performed in order to start analyzing the content of the tweets. Numbers, urls, punctuations, and special symbols such as hashtags, mentions, emojis need to be removed. After removing unnecessary symbols, all text need to be turned to lower case letters to avoid confusion on the word.

```
[1] "done is better than perfect  sheryl sandberg    "
[2] "shout out to the great fire department and the tour  much love to nyc            "
[3] "there are some amazingly hilarious nike ad memes happening on my newsfeed  soooo i decided to get a little
creative too     "
```

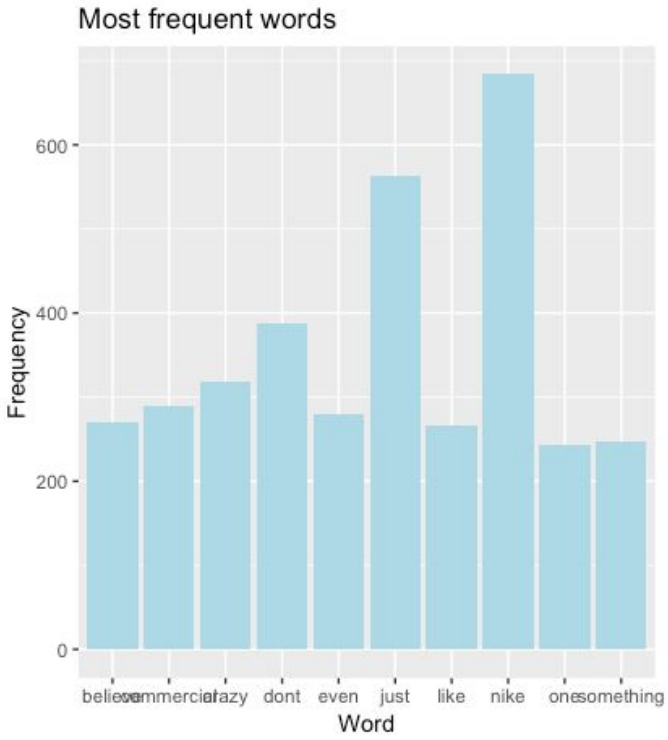
Plot 3. Example of cleaned tweets data

Visualizations

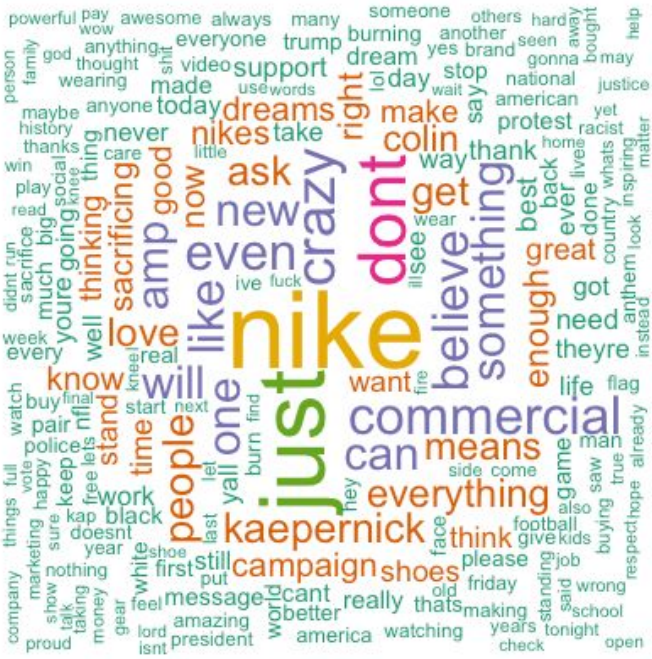
After we clean up punctuations, stopwords, white spaces and convert words to lowercase. Then, we use wordcloud package to visualize the frequent words that we cleaned before. We infer on the picture that the words with larger size, the frequency emerges more in tweets. Also, we head the top ten frequent words in a bar chart so that we could have a directly perception. The word cloud and the bar chart might be a little different from the pictures that we show on the slides. Since after presentation, we do some revise for our code and improve the accuracy for report. From those visualizations, we could get an information that people hold more positive attitudes towards this event.

Table 2. Most frequent word in tweet

	word	freq
nike	nike	684
just	just	564
dont	dont	388
crazy	crazy	319
commercial	commercial	289
even	even	279
believe	believe	270
like	like	266
something	something	247
one	one	244



Plot 4: Most frequent words in tweet



Plot 5: Most frequent words in tweet WordCloud

Sentiment Analysis on Tweets

After tweets data are cleaned, each word in the tweet is assigned a sentiment score by using the package “syuzhet”. Each tweet will be assigned the sum of all sentiment score from each word. When the tweet’s score is above 0, it means that the tweet have a positive emotional valence. When the tweet’s score is equal to 0, it would mean that the tweet have a neutral emotional valence. While when the tweet’s score is below 0, the tweet would express a negative emotional valence. After all tweets are evaluated, we found that overall there are more positive tweets.

Table 3. Number of negative, neutral, and positive Tweets

Negative	Neutral	Positive
1246	1344	2499

From the result from table 2, we can see that overall people are expressing a positive emotion towards this campaign. However, we are uncertain about people’s attitude toward an individual such as Colin Kaepernick and Donald Trump. Therefore, we would like to further our analysis by focusing on the main voices of all the tweets.

Sentiment Analysis on Tweets that Mention Most Influential Users

From the social network analysis, five most influential users are identified. Please refer to social network analysis section for detailed information. From these five most influential, three of them are the Twitter account of Colin Kaepernick, Donald Trump, and Nike. Since these three people are the main focus of this campaign. We would like to further analyze the tweets that reweeted these three accounts. The same process of data cleaning is performed and the same package is used to assign each word with a sentiment score. The result still have more positive emotion tweets.

Table 4. Number of negative, neutral, and positive Tweets

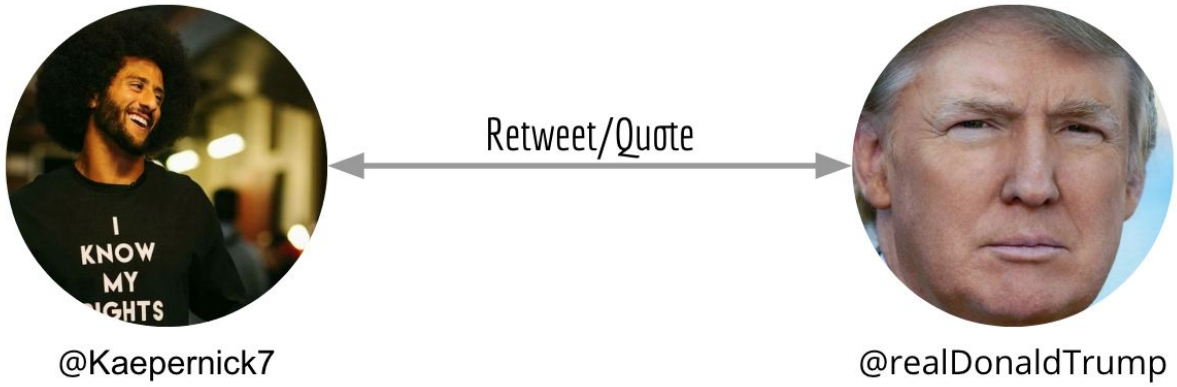
Negative	Neutral	Positive
145	100	259

We can infer that people overall have a positive emotion toward these three people. However, we still are uncertain about people's attitude toward the individual. Since people could mention all three users in one tweet and express opinions about all three at the same time, we have no method to analyze people's attitude toward a specific person. Therefore, we reached a limitation on analyzing people's real opinion about this campaign. We can not find whether these people support Nike or Donald Trump.

IV. Social Network Analysis

Network Building

When it comes to the social network analysis, the first step is to build the network. In our research, if two twitter users retweet or quote the other one, we will treat this behavior as connection. In other word, edge is defined to be retweet or quote behavior.



Plot 6. Edge definition

After building the network, some basic information of the network was calculated, including the number of nodes, the number of edges, edge density and diameter of the network. The results are shown in the following table.

Table 5. Basic information of the network

	Nodes	Edges	Edge density	Diameter
Value	1582	1180	0.00094	10

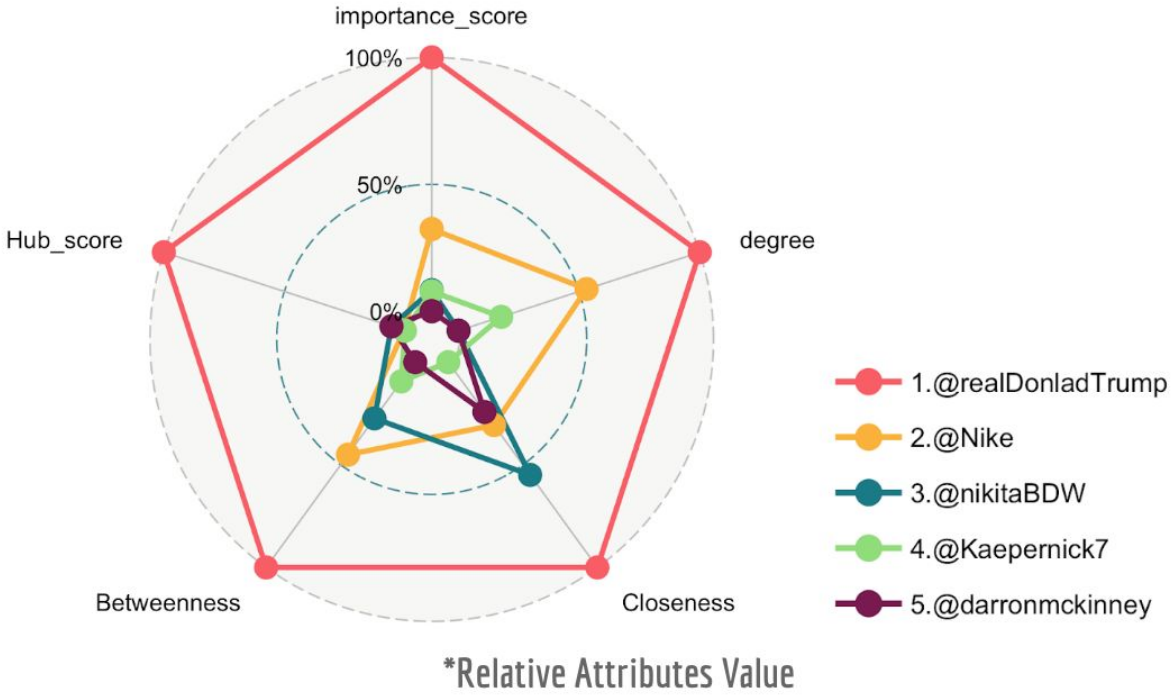
Identify the Most Influential Twitter Users

One of our research goals is to find the most influential Twitter users. So we need to have some kinds of standard. Basically, we choose four commonly used attributes in graph theory, including degree, betweenness, closeness and hub score. And then, normalization is performed because it

can better represent the relative importance of a node in each attributes. Finally, we create a new attribute, importance score. It equals to the sum of the normalised degree, betweenness, closeness and hub score.

Plot 7. Importance score definition

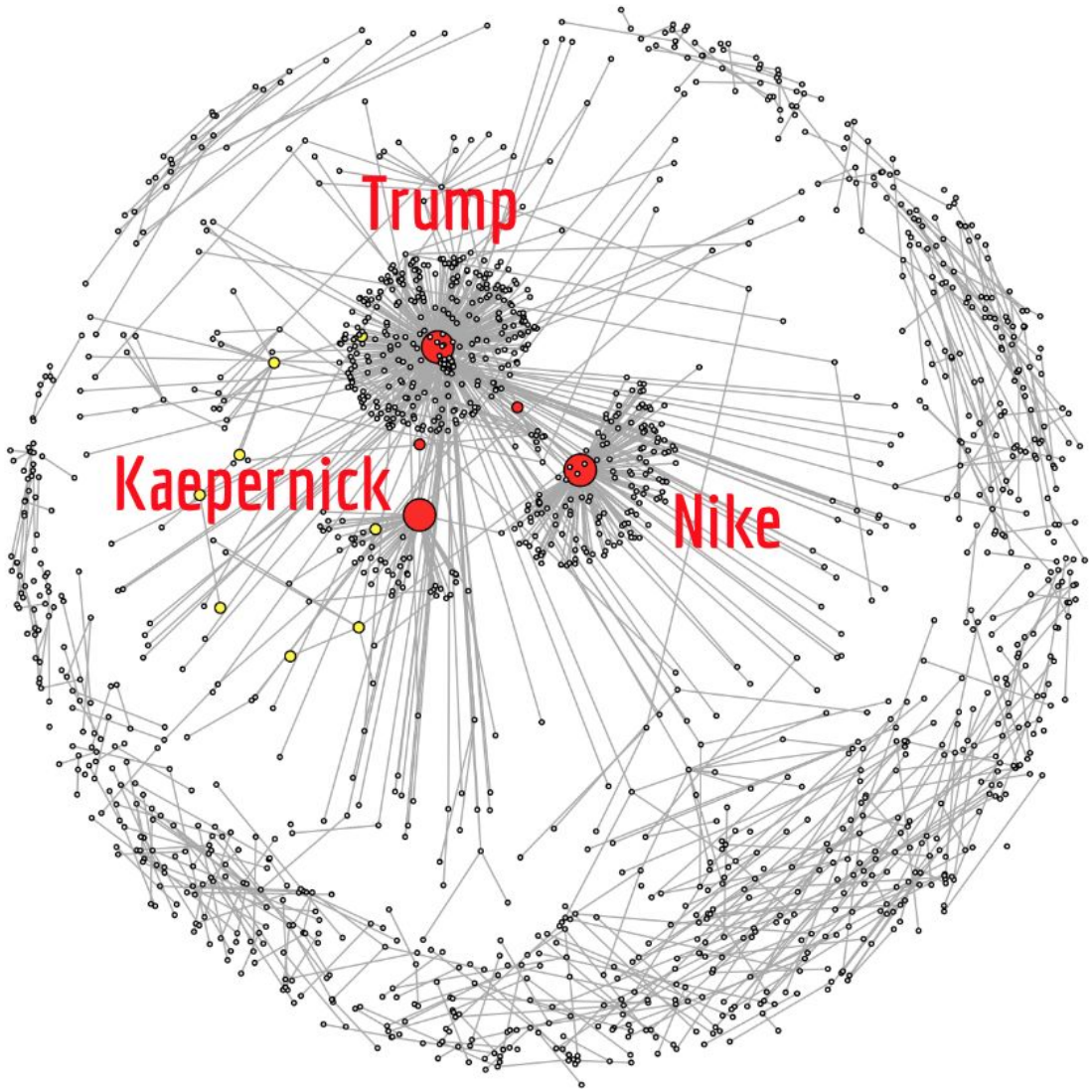
This radar chart shows the relative attributes value of top 5 most influential Twitter users. We can see that the president Trump is the most Influential Twitter User in this event, and he is far more Influential than the second one, nearly double. Besides, Nike Company and the leading figure, Colin Kaepernick, are also among the top 5 Most Influential Twitter Users.



Plot 8. Top 5 most influential Twitter users

Network Structure

To see the influence of the users in a more direct way, we visualised the whole network. Trump, Nike and Kaepernick have the largest degree, so they are three hubs of the network. Another interesting finding is the structure of the network. Obviously, Trump, Nike and Kaepernick are the center of the network. A large group of people interact with them. And the rest of people interact with each other. Although the number of people in this group is not small, but they can not form a new hub or center. So we can conclude that Trump, Nike and Kaepernick are the main voice of the event.



Plot 9. Network structure

V. Regression Analysis

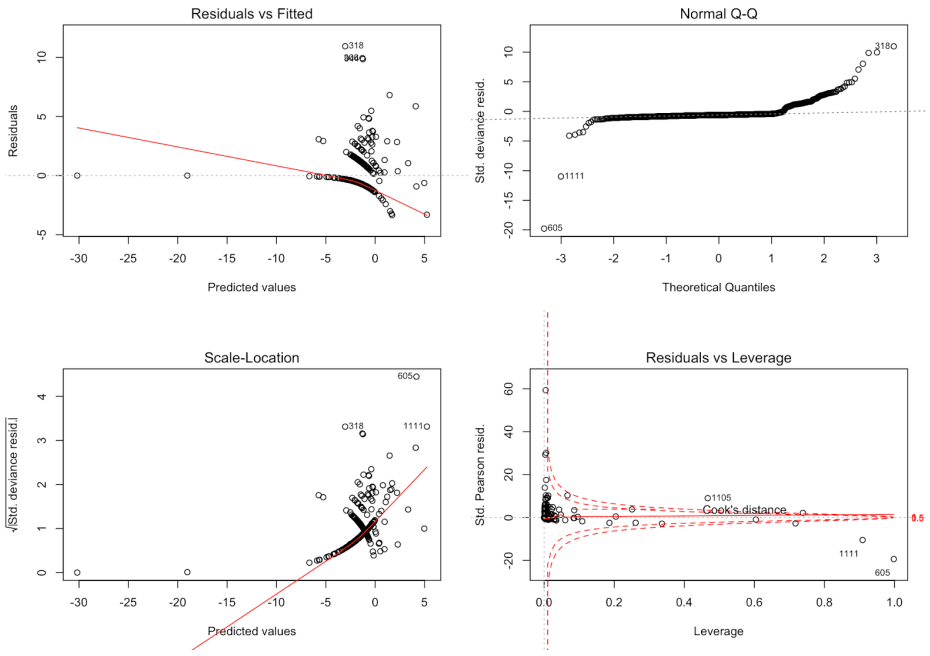
As we know, when people discuss and retweet something about it, the event will be keep pushing forward and attract more and more attention. This is power of social media. So we are very curious to know what kind of variables will have an impact on this process.

Poisson Regression

We plan to use regression to analyze this question. The dependent variable that we interested in is the number of retweet. Since it is a count variable, poisson regression is selected to use in our research. The possible independent variable including tweet_favorite_count(int), user_default_profile(binary), user_default_profile_image(binary), user_favourites_count(int), user_followers_count(int), user_friends_count(int), user_geo_enabled(binary), user_listed_count(int), user_statuses_count(int), user_register_year(compared to 2018). The names of the above variables are self-explained.

Model Selection

The first model contains all the possible independent variable mentioned above. But one of the variables is not significant, which is user_default_profile_image. After deleting this non-significant variables, the second model is gotten. The result of Chi-squared test shows that there is no significant difference between these two model, which $p = 0.8706$, indicating that the second model is prefered for simplicity. Based on that, the second model is our final model. In addition, the residual plot is shown as follow.



Plot 10. Residual plot

Model Interpretation

The poisson regression is log-linear model. The results should be explained in exponential way. After transformation, the results are shown as follow. The variables in the following table are all significant. We can see that if you turn on the geographic location, the number of retweet will more than double. In addition, tweets from a new registered user are more likely to be retweeted. Except geographic location and register year, the results of the rest variables are just like what we expect.

Table 6. Results of poisson regression

Tweet retweet Count .~

	Tweet favorite count	Default profile1	User favourites count	Followers count	Friends count
Exp (coefficient)	1.0179915	1.4196929	1.0000040	1.0000431	1.0000406

	Geo enabled1	Statuses count	Listed count	**Register year**	(Intercept)
Exp (coefficient)	**2.2464115**	1.0000086	0.9854591	**0.9557869**	0.1431984

In addition, the Deviance Residuals has little bit of skewness because median is not exactly equal to 0. Therefore, a robust method for standard error is applied.

Table 7. Robust estimation of standard error

	Estimation	Robust Std Err	Pr(>|z|)
(Intercept)	-1.943524e+00	3.498800e-01	2.778828e-08
tweet_favorite_count	1.783159e-02	2.728338e-03	6.331290e-11
user_default_profile1	3.504406e-01	2.841663e-01	2.174923e-01
user_favourites_count	4.048700e-06	1.479795e-06	6.219360e-03
user_followers_count	4.308335e-05	8.111312e-05	5.953138e-01
user_friends_count	4.055177e-05	8.324001e-05	6.261401e-01
user_geo_enabled1	8.093341e-01	2.592983e-01	1.800866e-03
user_listed_count	-1.464764e-02	2.700346e-03	5.816434e-08
user_statuses_count	8.611856e-06	2.380033e-06	2.964566e-04
user_register_year	-4.522029e-02	4.434672e-02	3.078714e-01

VI. Conclusion

From the sentimental analysis, we concluded that there are more positive tweets in general from all twitter users. Furthermore, there are more positive tweets from users that retweeted from Nike, Donald Trump, and Colin Kaepernick. We were not able to further analyze the exact attitude of users towards the main influential users of the campaign due to unable to assign scores on people's opinion towards a specific person. From the social network analysis, we concluded that Trump, Nike and Kaepernick are the center of the network. They dominated the voice of the event and none of other people can form a new center. From the poisson regression analysis, the effect of most variables are just like what we expect. For example, more friend will lead to more retweet count.

Some limitations also exist in our research, one of which is the data set. The data set we used only contains tweets in one day, which is Sep 7. But the hot of event will last few days or even longer than one week. If we can a data set which contains more information, the result might be more comprehensive.

For the future work, since our research has three different kinds of analysis, we want to see if we can combine two of them to get some novel results. First, we plan to combine the sentimental and social network analysis. We want to study people's reaction on this event, like positive or negative. And then, we try to explore the influence of main voice in social network. For example, we want to know whether the tweet of Trump can change someone's attitude since he is the center of the network. Second, we plan to combine social network and regression analysis. To be specific, we want to add network attributes into poisson regression model. We think the relative importance in the social network could have an impact on the number of retweet. After combining different parts of our analysis, we believe we can have a more comprehensive conclusion.

Appendix A

Reference

Dabbas, E. (2018). *5,000 #JustDoIt! Tweets Dataset* (Version 3) [Data file]. Available from Kaggle Web site: https://www.kaggle.com/eliasdabbas/5000-justdoit-tweets-dataset.

Appendix B

R Code - Data Overview

```
twitter <- read.csv("justdoit_tweets_2018_09_07_2.csv", header = T, stringsAsFactors = FALSE)
library(dplyr)
twitter <- twitter%>%
  select(tweet_favorite_count,tweet_retweet_count,user_favourites_count,user_followers_count,user_friends_count,
         user_listed_count,user_statuses_count)
library(psych)
describe(twitter)
```

```
```{r}
install.packages("tidytext")
install.packages("stringr")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("readxl")
library(readxl)
library(ggplot2)
library(stringr)
library(dplyr)
library(tidytext)

#Read in data and clean urls, convert to lowercase letters and extract all hashtags #
twitter <- read.csv("justdoit_tweets_2018_09_07_2.csv", header = T, stringsAsFactors = FALSE)
head(twitter)
twitter$tweet_full_text <- gsub("http.*", "", twitter$tweet_full_text)
twitter$tweet_full_text <- gsub("https.*","", twitter$tweet_full_text)
twitter$tweet_full_text <- tolower(twitter$tweet_full_text)
twitter$hashtags <- str_extract_all(twitter$tweet_full_text, "#\\w+")
#find most frequent hashtags
hashtags <- twitter$hashtags
hashtags <- unlist(hashtags)
hashtags_freq<- as.data.frame(table(hashtags),stringsAsFactors = FALSE)
hashtags_freq <- hashtags_freq[order(hashtags_freq$Freq,decreasing = TRUE), ]
hashtags_freq%>%
  top_n(10)
```

```
#plot most frequent hashtags
hashtags_freq%>%
  top_n(20)%>%
  mutate(hashtags = reorder(hashtags, Freq))%>%
  ggplot(aes(x = hashtags, y = Freq)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
      labs(y = "Count",
      x = "Unique Hashtags",
      title = "Count of unique hashtags found in tweets")
```
```
```{r}
#Extract all mentions @ and find most frequent mentions
twitter$mentions <- str_extract_all(twitter$tweet_full_text, "@\\w+")
mentions <- twitter$mentions
mentions <- unlist(mentions)
mentions_freq <- as.data.frame(table(mentions),stringsAsFactors = FALSE)
mentions_freq <- mentions_freq[order(mentions_freq$Freq,decreasing = TRUE), ]
mentions_freq%>%
  top_n(10)
#Plot most frequent mentions
mentions_freq%>%
  top_n(20)%>%
  mutate(mentions = reorder(mentions, Freq))%>%
  ggplot(aes(x = mentions, y = Freq)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
      labs(y = "Count",
      x = "Unique Mentions",
      title = "Count of unique mentions found in tweets")
```

R Code - Sentiment Analysis

```r
install.packages("syuzhet")
library(syuzhet)
twitter_2 <- read.csv("justdoit_tweets_2018_09_07_2.csv", header = T)
#remove urls, hashtags, mentions, digits, punctuation, emojis from tweets and convert to lowercase letters
twitter_2$tweet_full_text <- gsub("http.*", "", twitter_2$tweet_full_text)
twitter_2$tweet_full_text <- gsub("https.*","", twitter_2$tweet_full_text)
twitter_2$tweet_full_text <- gsub("@\\w+","",twitter_2$tweet_full_text)
twitter_2$tweet_full_text <- gsub("#\\w+","", twitter_2$tweet_full_text)
twitter_2$tweet_full_text <-gsub("[^a-zA-Z ]", "", twitter_2$tweet_full_text)
twitter_2$tweet_full_text <- tolower(twitter_2$tweet_full_text)
twitter_2$tweet_full_text <- gsub("[[:punct:]]", "", twitter_2$tweet_full_text)
twitter_2$tweet_full_text <-iconv(twitter_2$tweet_full_text, "latin1", "ASCII", sub="")
#convert to vector
word_2 <- as.vector(twitter_2$tweet_full_text)
#assign sentiment score for each word
sentiment_score_2 <- get_sentiment(word_2)
#get positive, negative and neutral tweets
positive_tweets_2 <- word_2[sentiment > 0]
negative_tweets_2 <- word_2[sentiment < 0]
neutral_tweets_2 <- word_2[sentiment == 0]
#number of negative, neutral and positive
types_2 <- ifelse(sentiment_score_2< 0, "Negative", ifelse(sentiment_score_2 > 0, "Positive", "Neutral"))
table(types_2)
```

```r
#user_id of Trump, Nike, and Kaepernick
num <- c(25073877, 415859364, 45055696)
name <- twitter%>%
  filter(tweet_in_reply_to_user_id %in% num)
##remove urls, hashtags, mentions, digits, punctuation, emojis from tweets and convert to lowercase letters
name$tweet_full_text <- gsub("http.*", "", name$tweet_full_text)
name$tweet_full_text <- gsub("https.*","", name$tweet_full_text)
name$tweet_full_text <- gsub("@\\w+","",name$tweet_full_text)
name$tweet_full_text <- gsub("#\\w+","", name$tweet_full_text)
name$tweet_full_text <-gsub("[^a-zA-Z ]", "", name$tweet_full_text)
name$tweet_full_text <- tolower(name$tweet_full_text)
name$tweet_full_text <- gsub("[[:punct:]]", "", name$tweet_full_text)
name$tweet_full_text <-iconv(name$tweet_full_text, "latin1", "ASCII", sub="")
#convert to vector
word<- as.vector(name$tweet_full_text)
#assign sentiment score for each word
sentiment_score <- get_sentiment(word)
#get positive, negative and neutral tweets
positive_tweets <- word[sentiment > 0]
negative_tweets <- word[sentiment < 0]
neutral_tweets <- word[sentiment == 0]
#number of negative, neutral and positive
types <- ifelse(sentiment_score< 0, "Negative", ifelse(sentiment_score > 0, "Positive", "Neutral"))
table(types)
```

Visualization

```
library(ggplot2)
library(twitteR)
library(SnowballC)
library(wordcloud)
library(tm)
library(qdap)
library(stringr)

dataclean <- as.vector(twitter_2$tweet_full_text)

tweetscorpus <- Corpus(VectorSource(dataclean))
tweetscorpus = tm_map(tweetscorpus, removeWords, stopwords("english"))
a <- TermDocumentMatrix(tweetscorpus)
b <- as.matrix(a)
c <- sort(rowSums(b),decreasing=TRUE)
dataclean <- data.frame(word = names(c),freq=c)
head(dataclean, 10)
data1 <-head(dataclean,10)

#plot wordcoud
wordcloud(words = dataclean$word, freq = dataclean$freq, min.freq = 20,
          max.words=250, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(6, "Dark2"))

#Plot most frequent word
data2 <- data.frame(word = dataclean$word[1:10],freq = dataclean$freq[1:10])
ggplot(data2,aes(x=word,y=freq)) + geom_bar(stat = 'identity',fill = 'lightblue') +
  labs(title = 'Most frequent words',x = 'Word', y = 'Frequency')
```

R Code - Social Network Analysis

```
#title: "twitter"
#author: "Shen YANG"
#date: "11/24/2018"
#source: https://www.kaggle.com/eliasdabbas/5000-justdoit-tweets-dataset/home

setwd("~/Documents/UCI/Customer and Social Analytics/Twitter Project")

rm(list=ls())
tweets = read.csv("justdoit_tweets.csv", na.strings=c("NA","NaN", " ", "none") )
tweets$user_created_at = as.character(tweets$user_created_at)
tweets$user_default_profile_image = as.factor(tweets$user_default_profile_image)
tweets$user_default_profile = as.factor(tweets$user_default_profile)
tweets$user_geo_enabled = as.factor(tweets$user_geo_enabled)

for (i in 1:dim(tweets)[1]) {
  tweets$user_created_year[i] = as.numeric(tail(strsplit(tweets$user_created_at[i], split = " ")[[1]], n=1))
}

tweets =  mutate(tweets, user_register_year = (2018 - user_created_year))

################## Network Building ##################

#Create network
library(igraph)
library(dplyr)
#create edge_matrix
edge_matrix = data.frame(tweets$tweet_in_reply_to_user_id, tweets$user_id)

#generate network
net = graph.data.frame(edge_matrix, directed = F)
```

```
################## Attributes ##################

#degree_distribution
node_degree = degree(net, mode="all")
deg.dist = degree_distribution(net, cumulative = T, mode="all")
plot( x=0:max(node_degree), y=1-deg.dist, pch=19, cex=1.2, col="orange",
      xlab="Degree", ylab="Cumulative Frequency", xlim = c(0,15))
hist(node_degree, breaks=200, main="Histogram of node degree", xlim = c(0,15))

#density
edge_density(net, loops = FALSE)

#degree centrality
centr_degree_tmax(net, mode = "all", loops = FALSE)

#definition of attributes
#closeness: the inverse of the average length of the shortest paths to/from all the other vertices
#between: the number of geodesics (shortest paths) going through a vertex or an edge
#hub_scores: a good hub represented a page that pointed to many other pages
#authority_score: a good authority represented a page that was linked by many different hubs

#degree, closeness, between, and hub/authority scores for each node
network_attributes = data.frame(
  node_name = V(net)$name,
  all_degree=degree(net, mode = "all"),
  Closeness = closeness(net, mode = "all", weights = NA, normalized = FALSE),
  Betweenness = betweenness(net, directed = T, weights = NA),
  Hub_score = hub_score(net)$vector,
  Authority_score = authority.score(net)$vector
)

#Normalized network attributes
normalized_net_attr = data.frame(node_name = network_attributes$node_name,
                                 degree = scale(network_attributes$all_degree),
                                 Closeness = scale(network_attributes$Closeness),
                                 Betweenness = scale(network_attributes$Betweenness),
                                 Hub_score = scale(network_attributes$Hub_score))

#Define importance score (importance_score) = sum of nomalized (degree, Closeness, Betweenness and Hub_score)
normalized_net_attr = mutate(normalized_net_attr,
                             importance_score = degree + Closeness + Betweenness + Hub_score) %>%
  mutate(n_importance_score = scale(importance_score))
```

```
#radar map
library(ggplot2)
library(ggradar)

top5_node = normalized_net_attr[c(1,2,1237,18,920),1:6]%>%
  mutate_at(vars(-node_name),funs(rescale))

top5_node$node_name = as.character(top5_node$node_name)
top5_node$node_name[1] = "1.@realDonladTrump"
top5_node$node_name[2] = "2.@Nike"
top5_node$node_name[3] = "3.@nikitaBDW"
top5_node$node_name[4] = "4.@Kaepernick7"
top5_node$node_name[5] = "5.@darronmckinney"
top5_node = top5_node[,c(1,6,2,3,4,5)]

set.seed(111)
ggradar(top5_node, legend.title = "User_id",
        plot.title="                    Relative Attributes Value")
```

```
################## Plot the Network ##################

#compute diameter
D = diameter(net, directed = F, weights = NA)
#find the component vertex of diameter
node_along_diameter = get_diameter(net, directed = F, weights = NA)

#color all the vertex
V(net)$color = "gray92"
#colorthe vertex on diameter
V(net)[node_along_diameter]$color = "yellow"
#colorthe vertex has top5 highest importance score
V(net)[as.character(normalized_net_attr$node_name[1])]$color = "red"
V(net)[as.character(normalized_net_attr$node_name[2])]$color = "red"
V(net)[as.character(normalized_net_attr$node_name[1237])]$color = "red"
V(net)[as.character(normalized_net_attr$node_name[18])]$color = "red"
V(net)[as.character(normalized_net_attr$node_name[920])]$color = "red"


#give size to all the vertex
V(net)$size = 0.85
#give a larger size to the vertex on diameter
V(net)[node_along_diameter]$size = 2
#give largerest size to the vertex has highest degree
V(net)[as.character(normalized_net_attr$node_name[1])]$size = 6
V(net)[as.character(normalized_net_attr$node_name[2])]$size = 6
V(net)[as.character(normalized_net_attr$node_name[1237])]$size = 2
V(net)[as.character(normalized_net_attr$node_name[18])]$size = 6
V(net)[as.character(normalized_net_attr$node_name[920])]$size = 2

# Add some specific lables and legends.
V(net)$username = ""
V(net)[as.character(normalized_net_attr$node_name[1])]$username = "@Trump"
V(net)[as.character(normalized_net_attr$node_name[2])]$username = "@Nike"
V(net)[as.character(normalized_net_attr$node_name[1237])]$username = ""
V(net)[as.character(normalized_net_attr$node_name[18])]$username = "@Kaepernick"
V(net)[as.character(normalized_net_attr$node_name[920])]$username = ""
```

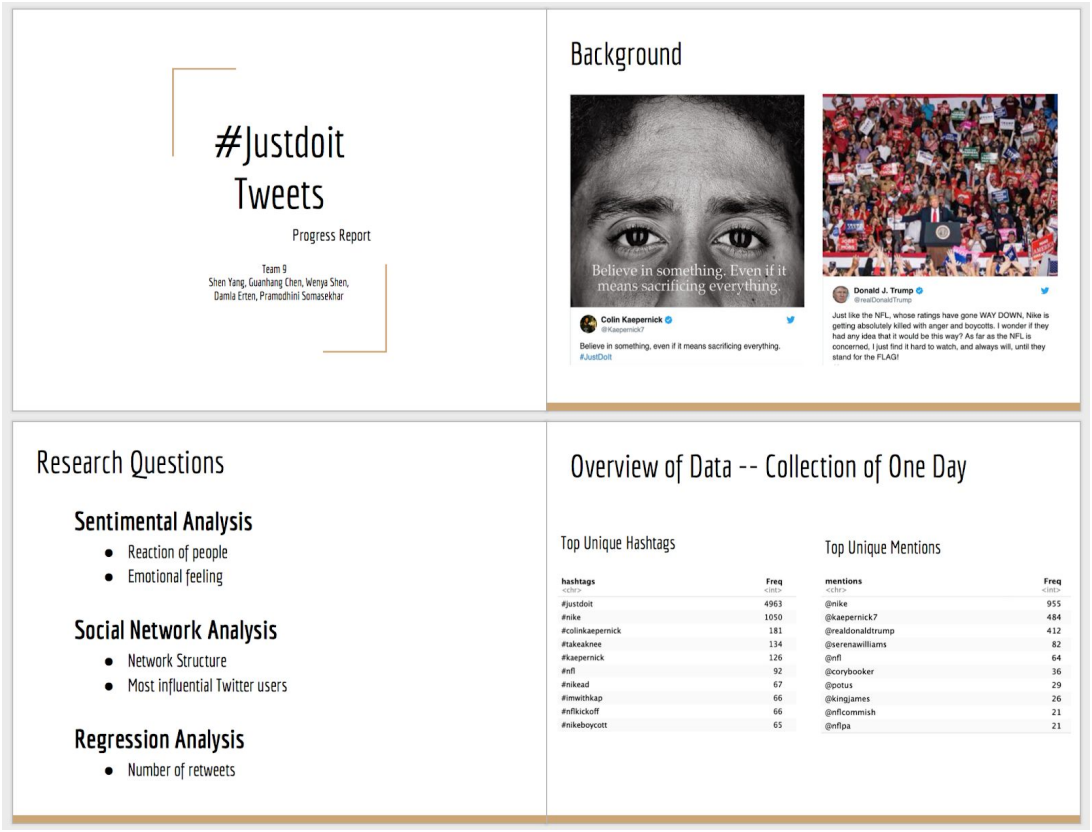
```
plot(simplify(net),
     vertex.color = V(net)$color,
     vertex.size = V(net)$size,
     edge.arrow.size = 0.05,
     vertex.label = V(net)$username,
     vertex.label.dist = 1.2,
     vertex.label.color = "firebrick1",
     #vertex.label.degree = pi,
     vertex.label.cex = 1.3,
     layout=layout.kamada.kawai,
     main = "Tweets Network")

legend(x = 0.7, y = -0.5,
       c("Nodes with highest importance score","Nodes along the diameter", "Normal nodes"), pch=21,
       col="#777777", pt.bg=c("red","yellow","gray"), pt.cex=2, cex = 0.75, bty="n", ncol=1, horiz = FALSE)
```

R Code - Regression Analysis

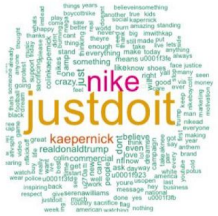
```
################## Possion regression ##################

#Put all interested variables into the model
fit1 = glm(tweet_retweet_count ~ tweet_favorite_count + user_default_profile + user_default_profile_image +
             user_favourites_count + user_followers_count + user_friends_count + user_geo_enabled +
             user_listed_count + user_statuses_count + user_register_year,
           data = tweets, family = poisson())
summary(fit1)

#delete non-significant variables
fit2 = glm(tweet_retweet_count ~ tweet_favorite_count + user_default_profile +
             user_favourites_count + user_followers_count + user_friends_count + user_geo_enabled +
             user_listed_count + user_statuses_count + user_register_year,
           data = tweets, family = poisson())
summary(fit2)

#Model comparison -> not significant -> fit2 is better for simplicity
anova(fit1, fit2, test="Chisq")

#Multicollinearity
library(car)
vif(fit2)

#Performing the deviance goodness of fit test
with(fit2, cbind(res.deviance = deviance, df = df.residual,
                 p = pchisq(fit2$deviance, df=fit2$df.residual, lower.tail=FALSE)))

#Residual plot
par(mfrow=c(2,2))
plot(fit2)
par(mfrow=c(1,1))

# robust
library(sandwich)
cov.output = vcovHC(fit2, type = "HC0")
std.err = sqrt(diag(cov.output))
robust = cbind(Estimation = coef(fit2), "Robust Std Err" = std.err,
               "Pr(>|z|)" = 2*pnorm(abs(coef(fit2)/std.err), lower.tail = FALSE))

#interpretation of coefficient
exp(coef(fit2))
```

Appendix C

PPT Slides

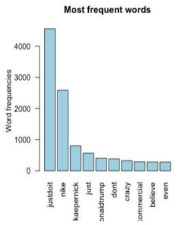


Sentimental Analysis



Sentimental Analysis

- No obvious attitude occur in top frequent words



word	freq
justdoit	4557
nike	2583
kaepernick	800
just	567
realdonaldtrump	403
dont	386
crazy	325
commercial	292
believe	282
even	281

Social Network Analysis

@Kaepernick7 @realDonaldTrump

Network Building

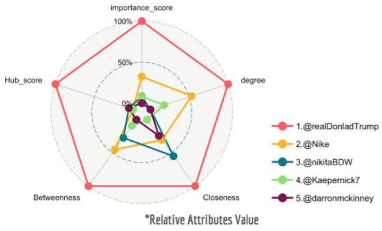
Social Network Analysis

Degree
Betweenness
Closeness
Hub_score

→ Normalised (Z score) → Importance_score = sum(.-)

Define Importance_score for each node

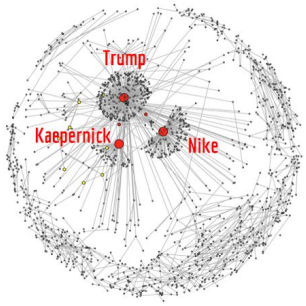
Social Network Analysis



*Relative Attributes Value

Top 5 Most Influential Twitter Users

Social Network Analysis



Regression Analysis

Tweet retweet Count .~

	Tweet favorite count	Default profile1	User favourites count	Followers count	Friends count
Exp (coefficient)	1.0179915	1.4196929	1.0000040	1.0000431	1.0000406

	Geo enabled1	Statuses count	Listed count	Register year	(Intercept)
Exp (coefficient)	2.2464115	1.0000086	0.9854591	0.9557869	0.1431984

Poisson Regression

Future Work

Further sentimental analysis:

- People's reaction on this event (Positive vs Negative)

Combine social network and regression analysis:

- Add network attributes into poisson regression model